

Visual Search and Analysis of Documents in the Intellectual Property Domain

Von der Fakultät Informatik, Elektrotechnik und
Informationstechnik der Universität Stuttgart
zur Erlangung der Würde eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Abhandlung

Vorgelegt von

Steffen Michael Koch

aus Stuttgart

Hauptberichter: Prof. Dr. Thomas Ertl
Mitberichterin: Univ.-Prof. Mag. Dr. Silvia Miksch
Tag der mündlichen Prüfung: 13. Dezember 2012

Institut für Visualisierung und Interaktive Systeme
der Universität Stuttgart

2012

Contents

Acknowledgment	xi
Danksagung	xiii
Abstract	xv
Zusammenfassung	xvii
1 Introduction	1
1.1 Problem Statement	2
1.2 Research Questions	3
1.3 Contribution	3
1.4 Thesis Structure	5
2 Foundations and Models	9
2.1 Visual Analytics	9
2.2 Information Visualization	11
2.3 Visualization Models	13
2.4 Visual Analytics Models	14
2.5 Document and Data Retrieval	17
2.5.1 Information Need	17
2.5.2 Text Document Retrieval	18
2.5.3 Machine Learning	20
2.5.4 Relational Databases and Other Data Sources	23
2.6 Visual Search Interfaces	23
2.6.1 Visual Query Definition	24
2.6.2 Visual Result Set Presentation	25
2.7 Sensemaking	26
2.8 Patent Data	29
2.9 Patent Characteristics	32
2.10 Patent Search Processes and Analysis Tasks	34
3 Visual Patent Analytics	39
3.1 Visual Interactive Support for Patent Search	43
3.1.1 Boolean Integration of Search Facilities	44
3.1.2 Visual Creation of Search Statements	46
3.1.3 Querying Metadata	52
3.1.4 Image Queries	54

Contents

3.1.5	Semantic Queries	55
3.1.6	Integrating Search Back-ends	56
3.2	Interactive Search Result Visualization	59
3.2.1	World Map	61
3.2.2	IPC views	62
3.2.3	Patent List	66
3.2.4	Patent Property Graph	68
3.2.5	Priority-Time View	69
3.2.6	Term Cloud	70
3.2.7	Legal Entity Chart	72
3.2.8	Detail Views	74
3.2.9	Selection Management	74
3.3	Feedback Loops and Insight Reintegration	78
4	Plug-In Visual Analytics	81
4.1	Visual Analytics on the Interaction Level	82
4.1.1	A Focus+Context technique for Edge Exploration	84
4.1.2	Feedback Loop	92
4.2	A Visual Analytics Approach to Classifier Creation	92
4.2.1	Background and Motivation	93
4.2.2	A Prototype for Visual Classifier Training	96
4.2.3	Feedback Loops and Workflows for Classifier Training	105
4.3	Integration	108
5	Scalability, Provenance and Reporting	111
5.1	Scalability Aspects of Patent Literature Analysis	113
5.1.1	Software and Data Scalability	114
5.1.2	Visual and Interaction Scalability	120
5.1.3	Platform Scalability	123
5.1.4	User, Task and Process Scalability	124
5.1.5	Scalability Conflicts	126
5.2	Collaboration, History Recording, and Analytic Provenance	127
6	Evaluation, Results and Discussion	137
6.1	Evaluation	137
6.1.1	The Difficulty of Evaluating Visual Analytics Approaches	138
6.1.2	Exploiting Analytic Provenance Data for Evaluation	141
6.1.3	Evaluation of the PatViz Approach	142
6.1.4	Advanced Focus+Context	145
6.1.5	Classifier Creation	147
6.2	Discussion	157

Contents

6.2.1	PatViz	157
6.2.2	EdgeAnalyzer	158
6.2.3	Classifier creation	158
6.2.4	General Considerations	159
7	Outlook	165
	Bibliography	169

List of Figures

2.1	The information visualization reference model according to Card et al.	13
2.2	The visual analytics process according to Keim et al.	14
2.3	Extension of Card et al. 's model for including visual analytics processes	15
2.4	Classification with support vector machines	22
2.5	The sensemaking process for intelligence tasks according to Pirolli and Card (simplified)	26
2.6	Notional sensemaking process in patent analysis	28
2.7	Front page of a European patent application.	30
2.8	Encoding scheme of the international patent classification (IPC) with an example from the field of optical recording.	31
2.9	An abstraction of the iterative patent search and analysis process	37
3.1	An overview of the PatViz desktop showing a variety of the available views for patent document search and analysis.	42
3.2	Boolean integration queries for back-end services	45
3.3	Visual keyword query	47
3.4	Multilingual keyword query	51
3.5	Combined metadata and keyword query	53
3.6	The back-end of PatViz	57
3.7	Parser/generator framework for visual query builder	58
3.8	World map	61
3.9	IPC treemap with selected and highlighted sections	62
3.10	IPC treemap	63
3.11	IPC after structural zooming operation	65
3.12	List view	66
3.13	Patent graph view	67
3.14	Zoomed sub structure in the patent graph view	69
3.15	Priority-time view	70
3.16	Term cloud	71
3.17	Legal entity chart	73
3.18	Visual selection management	76
4.1	The EdgeAnalyzer focus+context technique applied to the patent co-classification scenario as described in Section 3.2.2	84
4.2	The basic process for carrying out edge exploration with the EdgeAnalyzer approach.	85
4.3	Local edge de-bundling	87
4.4	EdgeAnalyzer views	87

4.5	Arc wheel	88
4.6	Dependencies and interfaces for EdgeAnalyzer	90
4.7	EdgeAnalyzer for parallel coordinates	91
4.8	Overview of user-steered classifier creation	96
4.9	The interface for user-steered classifier creation providing multiple coordinated views for inspecting a classifiers state.	97
4.10	Classification view	99
4.11	Term Chart depicting the classifier model’s most important dimensions	100
4.12	Cluster view for finding labeling candidates	101
4.13	Training data view	102
4.14	Labeling panel	103
4.15	Labeling actions and classification preview	104
4.16	Classifier history view	105
4.17	An abstract overview of the classifier creation process.	106
4.18	Integration of visual analytics approaches	109
5.1	For each query iteration a new tab is generated depicting the state of the corresponding query and result set views.	120
5.2	Overview microblog analysis	132
5.3	Automatically detected spatio-temporal term anomalies derived from the VAST Challenge 2011 microblog dataset	133
5.4	Selection graph for the analysis of VAST 2011 microblog scenario . .	134
5.5	Generated report	135
6.1	Basic method	150
6.2	Evaluation results of text classification user study for RCV1 corpus	153
6.3	Evaluation results of text classification user study for 20ng corpus .	154

List of Tables

6.1	Initial queries for evaluation	152
6.2	Best F_1 performance achieved during evaluation	155

List of Abbreviations and Acronyms

AL	Active Learning
CLEF	Cross Language Evaluation Forum
CLIR	Cross Language Information Retrieval
DFG	Deutsche Forschungsgemeinschaft
EPO	European Patent Office
HTML	Hypertext Markup Language
IEEE	Institute of Electrical and Electronics Engineers
InfoVis	Information Visualization
IPC	International Patent Classification
IR	Information Retrieval
ISVM	Linear Support Vector Machine
MCV	Multiple Coordinated Views
MDS	Multi Dimensional Scaling
NFS	National Science Foundation
NLP	Natural Language Processing
NTCIR	NII Test Collection for IR Systems
PCA	Principal Component Analysis
RDBMS	Relational Database Management Systems
SME	Small and Medium-sized Enterprises
SVM	Support Vector Machine
WIPO	World Intellectual Property Organization
VA	Visual Analytics
VAST	The IEEE Conference on Visual Analytics Systems and Technologies
XML	eXtensible Markup Language

Acknowledgment

I would like to thank my supervisor Thomas Ertl for his trust and support during the PhD process. For her interest in my work and her external report on the thesis I would like to thank Silvia Miksch. Many thanks go to my (former) colleagues Harald Bosch, Mark Giereth, Florian Heimerl, Charles Jochim, Robert Krüger, Christoph Müller, Alexandros Panagiotidis, Guido Reina, Martin Rotard, Dennis Thom, and Michael Wörner for the great collaboration, which formed the basis and stepping stones for the novel contributions presented in this thesis. Furthermore, I would like to thank the many colleagues and collaborators from the various projects in Germany and abroad for the great and successful cooperation. For their proofreading and helpful comments I am much obliged to my partner Julia Walther, as well as my colleagues Charles Jochim and Guido Reina; for the layout and his support with all L^AT_EX related questions I am greatly indebted to my colleague Martin Falk. Additionally, I would like to thank all colleagues and collaborators, especially those who are not explicitly mentioned, for the inspiring discussions that helped to develop my perspective on the research field as it is presented in this work. Last but not least, I would like to thank my partner and my family for their support over these past few years.

Danksagung

Bedanken möchte ich mich bei meinem Doktorvater Thomas Ertl für das mir entgegengebrachte Vertrauen und seine Unterstützung während der Promotion. Frau Silvia Miksch danke ich für ihr Interesse an meiner Arbeit und den Mitbericht. Meinen (ehemaligen) Kollegen Harald Bosch, Mark Giereth, Florian Heimerl, Charles Jochim, Robert Krüger, Christoph Müller, Alexandros Panagiotidis, Guido Reina, Martin Rotard, Dennis Thom und Michael Wörner danke ich für die tolle Zusammenarbeit, aus der wichtige Grundlagen und Bausteine für die in dieser Dissertation beschriebenen Neuerungen hervorgegangen sind. Zudem möchte ich mich bei den zahlreichen Kollegen und Kooperationspartnern aus dem In- und Ausland in den verschiedenen Projekten, an denen ich beteiligt war, für die gute und erfolgreiche Zusammenarbeit bedanken. Für das Korrekturlesen und Anregungen bedanke ich mich herzlich bei meiner Partnerin Julia Walther sowie meinen Kollegen Charles Jochim und Guido Reina; für das Layout, die Hilfe beim Satz der Arbeit und in allen L^AT_EX-Fragen gilt meinem Kollegen Martin Falk großer Dank. Weiterhin möchte ich mich bei allen Kollegen und Kooperationspartnern, insbesondere allen nicht explizit genannten, für die inspirierenden Diskussionen und Gespräche bedanken, ohne die ich meine in dieser Arbeit dargestellte Sicht auf die Dinge nicht so hätte entwickeln können. Zu guter Letzt will ich mich bei meiner Partnerin und bei meiner Familie für das Verständnis und die Unterstützung über die Jahre meiner Promotion bedanken.

Abstract

Today's society generates and stores digital information in enormous amounts and at rapidly increasing rates. This trend affects all parts of modern society, such as commerce and economy, politics and governments, health and medicine, science in general, media and entertainment, the private sector, etc. The stored information comprises text documents, images, audio files, videos, structured data from a variety of sources, as well as multimodal combinations of them, and is available in a multitude of electronic formats and flavors. As a consequence, the need for automated and interactive tools supporting tasks, such as searching, exploring, monitoring, sorting, and making sense of this information at different levels of abstraction and within different but steadily converging domains, increases at the same pace as the data is generated and represents one of the biggest challenges for current computer science.

A relatively young approach to tackle these tasks by exploiting human analytic power in synergetic combination with advanced computerized techniques has emerged with the research field of visual analytics. Visual analytics aims at combining automated methods, visualization techniques, and approaches from the field of human computer interaction in order to equip analysts with more powerful tools, tailored to domains, where large amounts of data must be analyzed. In this work, visual analytics methods and concepts play a central role. They are used to search and analyze texts or multimodal documents containing a considerable amount of textual content. The presented approaches are primarily employed for analyzing a very special type of document from the intellectual property domain, namely patents. Since the retrieval and analysis tasks carried out in the patent domain differ greatly from standard search and analysis tasks regarding rigorous requirements, high costs, and the involved risks, new, more effective, efficient, and more reliable methods need to be developed.

Accordingly, this thesis focuses on researching the combination of automatic methods and information visualization by using advanced interaction techniques in order to improve upon the state of the art in patent literature retrieval. Such integration is achieved and exemplified through different visual analytics prototypes, aiming at creating support for real-world tasks and processes. The main contributions presented in this thesis encompass enhancements for all stages of patent literature analysis processes. This includes improving patent search by presenting techniques for interactive visual query building, which helps analysts to formulate complex information needs, the development of a technique that allows users to build their own precise search mechanism in the form of binary classifiers, and advanced approaches for making sense of a retrieved result set through visual analysis. The latter builds

the base to let users generate insights needed for judging and improving previous query formulations. Interaction methods facilitating forward analysis as well as feedback loops, which constitute a critical part of visual analytics approaches, are discussed afterwards. These methods are the key to integrating all stages of the patent analysis process in a seamless visual manner. Another contribution is the discussion of scalability issues in context of the described visual analytics approaches. Especially interaction scalability, the recording of analytic provenance, insight management, the visualization of analytic reporting, and collaborative approaches are addressed.

Although the described approaches are exemplified by applying them to the field of intellectual property analysis, the developments regarding search and analysis have the potential to be adapted to complicated text document retrieval and analysis tasks in various domains. The general ideas regarding the facilitation of low-level feedback loops, user-steered machine classification, and technical solutions for diminishing negative scalability effects can be directly transferred to other visual analytics scenarios.

Zusammenfassung

Unsere heutige Gesellschaft erzeugt enorme Mengen digitaler Informationen, und das in rasant steigender Geschwindigkeit. Dieser Trend zeichnet sich in allen Bereichen der modernen Gesellschaft ab, sei es in Handel und Wirtschaft, in der Politik und der öffentlichen Hand, im Gesundheitswesen und der Medizin, in der Wissenschaft, den Medien, der Unterhaltungsbranche oder im privaten Umfeld. Die gespeicherten Informationen umfassen unter anderem Textdokumente, digitale Bilder, Tonaufnahmen, Videos, strukturierte Daten aus unterschiedlichen Quellen, sowie multimodale Kombinationen aus den verschiedenen Sparten. Sie alle liegen in einer Fülle unterschiedlicher elektronischer Formate und Varianten vor. Entsprechend wächst der Bedarf an automatisierten und interaktiven Werkzeugen, die Unterstützung für eine ganze Reihe von Aufgaben bieten - wie z.B. der Suche, der Exploration, der Überwachung, der Einordnung, und der Analyse gespeicherter digitaler Informationen, auf unterschiedlichen Abstraktionsebenen und in unterschiedlichen aber stetig konvergierenden Fachgebieten. Die Entwicklung von Werkzeugen um diese Aufgaben unter Berücksichtigung ständig wachsender Datenmengen zu bewältigen stellt dabei eine der größten Herausforderungen für die heutige Informatik dar.

Ein vergleichsweise neuer Ansatz zur Lösung dieser Probleme wurde mit dem Forschungsgebiet "Visual Analytics" geschaffen, der Synergieeffekte aus der Verbindung von analytischen Fähigkeiten des Menschen mit fortschrittlichen Informationsverarbeitungstechniken nutzt. Visual Analytics kombiniert dabei automatische Verfahren, Visualisierungstechniken und Ansätze aus der Mensch-Computer-Interaktion, um Analysten mit mächtigeren Werkzeugen für die Analyse großer Datensätze auszurüsten. In dieser Arbeit spielen Visual-Analytics-Ansätze eine tragende Rolle. Sie werden für die Suche nach und die Analyse von Texten und multimodalen Dokumenten, die einen großen Textanteil aufweisen eingesetzt. Die vorgestellten Ansätze, werden hauptsächlich auf die Analyse von Patenten als eine besondere Art von Textdokumenten angewandt. Da sich die Such- und Analyseaufgaben innerhalb des Patentumfelds deutlich von Standardsuche und klassischen Analysen unterscheiden, was die strengen Anforderungen, hohen Aufwand und Kosten und die damit verbundenen Risiken betrifft, müssen neue, effektivere, effizientere und verlässliche Methoden entwickelt werden.

Die vorliegende Arbeit beschäftigt sich deshalb mit der Kombination von automatischen Methoden und Methoden der Informationsvisualisierung unter Nutzung moderner Interaktionstechniken, um die Patentsuche über den aktuellen Stand der Technik hinaus zu verbessern. Die Integration dieser Methoden wird mittels einer Reihe von Visual-Analytics-Prototypen, welche Unterstützung für reale Prozesse

und Aufgaben bieten, erreicht und exemplarisch aufgezeigt. Die vorgestellten Neuerungen umfassen Verbesserungen für sämtliche Schritte des Patentanalyseprozesses. Dazu gehört die Optimierung der Patentsuche durch die vorgestellten Techniken zur interaktiven visuellen Anfrageerstellung, die Analysten dabei helfen einen komplexen Informationsbedarf zu formulieren. Des Weiteren wird ein Verfahren erläutert, das es Benutzern erlaubt, ihre eigenen, präzisen Suchmechanismen in der Form binärer Klassifikatoren zu erstellen. Außerdem werden moderne Ansätze präsentiert, wie Ergebnismengen mit Unterstützung von visueller Analyse interpretiert und verstanden werden können. Daraus wiederum können Benutzer sodann Erkenntnisse gewinnen, die für die Bewertung und Verbesserung vorhergehender Anfrageformulierungen notwendig sind. Im Anschluss folgt die Auseinandersetzung mit interaktiven Methoden, die sowohl eine Vorwärtsanalyse als auch Feedback Loops ermöglichen, die einen wesentlich Bestandteil von Visual-Analytics-Ansätze darstellen. Diese Methoden bilden die Grundlage, auf deren Basis alle Phasen des Patentanalyseprozesses auf nahtlose Weise visuell miteinander verknüpft werden können. Ein weiterer Forschungsbeitrag besteht in der Untersuchung von Skalierbarkeitsaspekten im Zusammenhang mit den verwendeten Visual-Analytics-Ansätzen. Eine wesentliche Rolle spielen hier vor allem die Skalierbarkeit von Interaktionstechniken, die Aufzeichnung analytischer Prozesse, die Kombination und weitere Nutzung von gewonnenen Erkenntnissen, die Erzeugung von Analyseberichten, sowie kooperative Ansätze.

Auch wenn die hier vorgestellten Ansätze anhand konkreter Beispiele für spezielle Fachgebiete beschrieben werden, verfügen einige der Entwicklungen über das Potential, auf andere komplexe Textdokumentsuch- und Analyseaufgaben übertragen werden zu können. Insbesondere die Ideen zur Optimierung von Feedback Loops und benutzergesteuerter Klassifikation, sowie technische Lösungen zur Verbesserung der Skalierbarkeit lassen sich direkt auf andere Visual-Analytics-Szenarien übertragen.

Introduction

Today's society generates and stores digital information in enormous amounts, and at strongly increasing rates [Gantz and Reinsel, 2011]. This trend affects all parts of modern society such as commerce and economy, politics and governments, health and medicine, science in general, media and entertainment, the private sector, etc. The stored information comprises text documents, images, videos, structured data as well as multimodal combinations of them, and is available in a manifold of electronic formats and flavors. As a consequence, the need for supporting tasks such as searching, exploring, monitoring, sorting, and making sense of this information at different levels of abstraction and within different but steadily converging domains, increases at the same pace.

A relatively young approach to tackle these tasks has been defined by the research field of visual analytics. Here, automated methods, visualization techniques, and approaches from the field of human computer interaction are combined in order to equip analysts with more powerful tools, tailored to domains, where large amounts of data should be analyzed. In this thesis the visual analytics approaches and ideas play a central role. They are applied to search and analysis tasks for text documents or multimodal documents containing a considerable amount of textual content. The presented approaches are primarily employed for analyzing a very special type of document from the intellectual property domain, namely patents.

1.1 Problem Statement

Many professionals have to deal with patents today, aiming at a variety of analytic goals including patentability search, freedom to operate analysis, validity search, portfolio analysis, as well as finding trends, monitoring competitors and many more. However, a variety of problems make patent analysis a very complex and time-consuming task. Patent documents are digitally stored in patent databases and repositories, and are freely available from patent offices. The stock of patent information, however, is increasing rapidly. For example, the repository maintained by the European Patent Office (EPO)¹ (accessible through the esp@cenet² service) holds more than 60 million patent documents. In 2010, an all-time high of 1.98 million filed patent applications has been reported by the World Intellectual Property Organization (WIPO)³, while 7.3 patents million were in force, worldwide.

Not only the large amount of patent documents poses a challenge, but also the complexity and heterogeneity of patent information, such as multimodal content, bibliographic information and other metadata, the ‘patentesque’ they are written in, and several other aspects complicate the tasks mentioned above. For obvious reasons applicants are trying to produce patent applications that still follow the rules of patentability, but they also aim to phrase them as general as possible to achieve a maximum of coverage for their patents. Furthermore, some patent applications are multi-lingual, others are only accessible in the language of the country where they have been applied for.

Some patent analysis tasks have to be carried out very thoroughly, since failure in finding all relevant documents can result in a high risk of litigation and probably have severe economic consequences. Even if a company does not intend to apply for patents, the patent landscape of the domain(s) a company is involved in has to be tracked closely.

With the large numbers of patents applied for today, there is an increasing ‘backlog’⁴ of unprocessed patent applications causing enormous costs. Patents are also a concern for small and medium-sized enterprises (SMEs), who do not maintain their own legal departments and therefore depend on external service providers. It would be beneficial to equip SMEs with the techniques, necessary to carry out certain patent analysis tasks on their own. As a consequence, there is a need for approaches that speed up patent analysis, make it available to a wider group of stakeholders, make it more reliable, easier to handle, reusable, and to work on other enhancements while taking into account the abovementioned problems. Visual

¹ <http://www.epo.org/>

² www.espacenet.com

³ www.wipo.int/

⁴ <http://www.ipo.gov.uk/pro-types/pro-patent/p-policy/p-policy-backlog.htm>

analytics approaches offer the chance to increase effectiveness and efficiency to improve this situation.

1.2 Research Questions

In this context the following research questions arise:

- How can information visualization models be amended or updated to acknowledge the requirements of visual analytics methods?
- Can the problems faced in patent search and analysis be alleviated by visual analytics techniques?
- How can the important issue of scalability be addressed by switching from traditional to visual analytics approaches?
- Is it possible to derive generic methods from the techniques developed for patent analysis tasks?

1.3 Contribution

This thesis introduces an approach for patent search and analysis tasks called ‘PatViz’. PatViz, which was developed as part of the EC-Project PatExpert and the DFG priority program ‘Scalable Visual Analytics’, can be seen as a visual analytics system for patent search and analysis. Its main contribution is a visual, interactive interface that spans all phases of patent search and analysis tasks. It facilitates visual query creation, visual inspection of result sets, and the combination assessment of findings. Since multiple patent repositories containing patent documents, bibliographic data, semantic information, and image data, can be accessed with PatViz, emphasis is put on their integration in one coherent interface. The integration is realized for visual query definition as well as result set presentation. Additionally, interactive means are provided on the basis of this integration that allow patent analysts to include found insights directly in subsequent query definitions, which directly supports the iterative nature of patent search and analysis tasks. Furthermore, a selection management and filtering approach is employed that enables analysts to construct and filter visually and interactively with a graph-based visualization. Through combinatory analysis of result sets or subsets of them, hypotheses can be tested, found insights externalized, and complex restrictions reintegrated into subsequent query refinements.

In order to explain the architecture that has been employed to create PatViz, an extended variation of the information visualization reference model is introduced.

Furthermore, a process model for visual patent analysis is proposed and aligned with an adapted version of the sensemaking model as has been suggested by [Pirolli and Card \[2005\]](#). Especially, feedback loops that are provided on different levels of abstraction are discussed in the context of the PatViz system, since these are necessary if analytic tasks are to be carried out in a seamless, visual, and interactive manner. One contribution of this work is to highlight and describe in detail those aspects and concepts of the information visualization reference model that play an important role in visual analytics approaches.

The PatViz system has been extended with a technique to enhance explorative tasks, and an approach is suggested that enables analysts to create classifiers for enhancing text document retrieval tasks. Both techniques are visual analytics approaches themselves, but are designed in a way to be integrated into large visual analytics systems. The first method is ‘EdgeAnalyzer’ providing a focus+context technique for the exploration of link and edge-based views. It facilitates iterative drill-down operations based on metadata and geometric characteristics of the edges or links under exploration. Different automatic grouping and visualization methods are employed in order to increase the scalability of the method in situations where many links are explored at once. In addition, it is possible to use multiple dependent and independent lenses in one view, which further increases analytic capabilities. In the context of patent analysis, the technique is used for patent co-classification analysis, but it is designed in a flexible fashion that makes its employment in other edge-based visualizations possible.

The approach for visual, user-steered classifier creation presented in this thesis is generic as well, and can be adapted to other text retrieval scenarios. It is intended as an additional method to keyword-based retrieval approaches and can be especially helpful in situations where analysts have problems to define good (sub)queries for specific retrieval tasks. In the proposed method, binary classifiers can be trained by labeling documents as relevant and non-relevant according to the analyst’s information need. In order to speed up the process, analysts are provided with a visual interface enabling them to carry out selective labeling operations with a high impact on the classifier training. Linear support vector machines are used as a classification for this approach. The technique aims at hiding the complexities of the classification model by translating it into comprehensible visual abstractions. Thus analysts who have no previous knowledge of the employed machine learning technique, are able to build and assess high-quality classifiers.

Both, EdgeAnalyzer as well as user-steered classifier creation were built based on specific analytic process models, which are presented in this thesis. They are used to depict the integration of these systems into larger approaches. Scalability aspects play another important part in the design of visual analytics systems. This is not only the case because visual analytics aims at finding solutions for scenarios where

large amounts of information have to be analyzed, but has many other potentially conflicting dimensions.

Because analytic processes, whether they are carried out in the intellectual property domain or in other fields, are not finished after some findings were made, collaboration, provenance recording, and analytic reporting are discussed in the context of the presented approaches as well. They are crucial for sharing, assessing, and informing others about performed analyses and should therefore be an integral part of visual analytics systems. Recording analytic provenance data can be seen as one important building block for collaboration and reporting. If the important analytic steps taken during a task are stored explicitly, they can later be exploited to explain analyses to others in collaborative scenarios or to present the results of an analysis to decision makers. This is shown by example within a scenario for analyzing microblog data, but using the selection management and filtering approach developed in PatViz.

Results of the evaluation of all the proposed approaches are presented, and negative aspects as well as identified advantages are discussed in detail. In the case of user-steered classifier creation a new evaluation approach is suggested that combines classic evaluation methods of information retrieval with a user evaluation in order to assess the value of this visual analytics approach.

1.4 Thesis Structure

This thesis is structured as follows: Chapter 2 briefly introduces the fields and terminology necessary understanding the subsequent parts. This includes a detailed description of the field of visual analytics and its most important research goals, information visualizations, visualization models, information and document retrieval, search user interfaces, sensemaking, as well as a closer look at the data properties of patent documents and common tasks in the process of patent analysis. With PatViz, Chapter 3 presents a software prototype for visual patent analysis, its views, and basic interaction facilities. This comprises the introduction of methods for integrating the different views and query facilities through advanced interactive methods, as well as one for selection and insight management. Chapter 4 depicts two approaches that can be seen as plug-ins for larger analytics approaches. The first one, EdgeAnalyzer, offers a focus and context technique for edge exploration, while the second one introduces a method for visual user-steered classifier creation which can be used to forge task-specific tools for document retrieval. Chapter 5 highlights scalability issues important in the context of visual analytics and examines how the presented approaches acknowledge these. Chapter 6 covers the evaluation of the proposed techniques and discusses the results of the methods

depicted in this work. An outlook to future developments in visual patent document analysis and how the suggested methods might influence other developments in the field of visual analytics is given in Chapter 7.

Parts of the work presented in this thesis have already been disseminated in the form of conference papers, journal articles, and a book chapter, as can be seen from the list at the end of this chapter. The work described subsequently is part of the joint effort of many researchers, who are either co-authors of the papers listed below, or who were collaborators in one of the projects this work has been funded by. These projects include PatExpert⁵, financed by the European Commission in the context of Framework Programs 6, as well as ‘Scalable Visual Patent Analysis’, which has been funded by the German Science Foundation (DFG) as part of the priority program ‘Scalable Visual Analytics’⁶. Additional funding has been provided by the Universität Stuttgart.

This thesis is partly based on the following publications:

M. Giereth, S. Koch, M. Rotard, and T. Ertl. Web Based Visual Exploration of Patent Information. In *International Conference on Information Visualization (IV 2007)*, pages 150–155, 2007b

M. Giereth, S. Koch, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, and L. Waner. *A Modular Framework for Ontology-Based Representation of Patent Information*, pages 49–59. IOS Press, 2007a

S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative Integration of Visual Insights during Patent Search and Analysis. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*, pages 203–210, 2009

H. Bosch, J. Heinrich, C. Müller, B. Höferlin, G. Reina, M. Höferlin, M. Wörner, and S. Koch. Innovative filtering techniques and customized analytics tools. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*, pages 269–270, 2009

C. Rohrdantz, S. Koch, C. Jochim, G. Heyer, G. Scheuermann, T. Ertl, H. Schütze, and D. A. Keim. Visuelle Textanalyse. *Informatik-Spektrum*, 33:601–611, 2010

⁵ <http://www.patexpert.org/>

⁶ <http://www.visualanalytics.de/>

continued...

A. Panagiotidis, H. Bosch, S. Koch, and T. Ertl. EdgeAnalyzer: Exploratory Analysis through Advanced Edge Interaction. In *Hawaii International Conference on System Sciences (HICSS 2011)*, pages 1–10, 2011

H. Bosch, D. Thom, M. Wörner, S. Koch, E. Püttmann, D. Jäckle, and T. Ertl. ScatterBlogs: Geo-spatial document analysis. In *IEEE Conference on Visual Analytics Science and Technology (VAST 2011)*, pages 309–310, 2011

S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative Integration of Visual Insights during Scalable Patent Search and Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):557–569, 2011

S. Koch and H. Bosch. From Static Textual Display of Patents to Graphical Interactions. In M. Lupu, K. Mayer, J. Tait, A. J. Trippe, and W. B. Croft, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Kluwer International Series on Information Retrieval*, pages 217–235. Springer Berlin Heidelberg, 2011

D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 41–48, 2012

F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual Classifier Training for Text Document Retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012

R. Krüger, H. Bosch, S. Koch, C. Müller, G. Reina, D. Thom, and T. Ertl. HIVEBEAT - A Highly Interactive Visualization Environment for Broad-Scale Exploratory Analysis and Tracing. In *IEEE Conference on Visual Analytics Science and Technology (VAST 2012)*, pages 177–178, 2012

Foundations and Models

Visual analytics is a multidisciplinary field and the techniques for searching and analyzing patent literature presented in this thesis follow visual analytics approaches. As a consequence, a broad spectrum of research areas, such as information visualization, information retrieval, and machine learning are touched in this work. Important foundations and aspects that are part of the approaches discussed later are presented in this chapter accordingly. In particular models, for describing information visualization and visual analytics approaches, play an important role, since they provide the frame for developing domain specific approaches on the one hand and are valuable means for generalizing new developments to other fields, on the other. Furthermore, the peculiarities of patent documents that are interesting during analysis and the metadata attached to them are briefly presented. Finally, search and analysis approaches and processes as they are currently employed in patent analysis are discussed.

2.1 Visual Analytics

The term *visual analytics* was introduced by [Wong and Thomas \[2004\]](#). Visual analytics as a research direction became a prominent topic after the book ‘Illuminating the Path: Research and Development agenda for Visual Analytics’ was published by [Thomas and Cook \[2005\]](#). At this point in time, visual analytics was suggested as an approach to analyze huge amounts of heterogeneous and conflicting data in order to prevent terrorist threats and to react adequately to disastrous

events. However, it was obvious from the beginning that the general idea of visual analytics may be beneficial to many other domains.

The class of problems that can benefit from visual analytics methods can be roughly characterized as arising in situations where huge amounts of dynamically changing, heterogeneous, multidimensional, ambiguous, uncertain, or incomplete data, have to be analyzed and where analytic goals and tasks are either complex or cannot be specified clearly a priori. In such situations neither purely automatic approaches can be applied, nor human effort alone will lead to satisfactory results in a reasonable period of time. Visual analytics therefore proposes the usage of automatic methods in order to support human analysts in their reasoning tasks. Visualization, which exploits the exceptional properties of human visual perception, can help to make large amounts of data and their context quickly accessible to human analysts. In order to support analytic feedback loops required for sensemaking, the introduction of interactive methods is inevitable. Thus, interaction techniques build the glue between the analyst and the computer-implemented systems for triggering automatic analyses, changing visual perspectives, combining selection and filtering to validate or invalidate hypotheses. They embed human analysts in the sensemaking process (see Section 2.7), which makes them an important prerequisite for analytic discourse. At the same time interaction methods are one of the most delicate parts of a visual analytics application that can either greatly increase its analytic power or cause major issues if they are not carefully integrated.

Regardless of the introduction of visual analytics as a research direction, there already existed ideas, tools and systems, which could be seen as following the visual analytics idea, before it was introduced. However, the growth of visual analytics research during the last years is remarkable. The fields of application for visual analytics broadened quickly from the domains suggested in [Thomas and Cook \[2005\]](#). In 2009, [Thomas and Kielman \[2009\]](#) list further potential sectors and applications for visual analytics: security, health, energy, commerce, transportation, food/agriculture, economy, insurance, cyber security, knowledge workers, and personal use. And what visual analytics achieved is that researchers start thinking of tightly integrating the three mentioned aspects visualization, automated approaches, and human computer interaction, thereby boosting the development of new, more scalable and holistic approaches.

Visual analytics is the offspring of the field of visualization, in particular *scientific visualization* and *information visualization*. While the research discipline of scientific visualization¹ develops methods for visualizing measured or simulated

¹ The term ‘scientific visualization’ was coined from ‘visualization for scientific computing’ which emerged the first time in the ViSC report [[McCormick, 1988](#)] of the NSF.

data, and typically aims at depicting data that correlate to a spatial representation, information visualization aims at visualizing abstract concepts and data (cf. [Tory and Möller \[2004\]](#)). There is a subset of visualization approaches and scenarios from both disciplines, which also satisfy the definition of visual analytics. In the context of this work only information visualization techniques are considered, since intellectual property documents such as patents represent abstract information.

2.2 Information Visualization

Visualization can be a means to let users gain insights into large amounts of information quickly. It is therefore a valuable instrument to increase scalability for the analysis of abstract data. However, this is only possible if the information to be transported is visually prepared in a suitable way, regarding the type of data as well as the task that should be carried out. In the case of patent documents, which are the primary object of investigation in this thesis, a variety of metadata is available, covering almost every data type discussed in [Shneiderman \[1996\]](#). This includes hierarchical data, categorical data, time-based information, and many more as depicted below in [Section 2.8](#). A broad discussion on the benefits of information visualization and situations where it can be exploited successfully is given in [Fekete et al. \[2008\]](#). [Amar and Stasko \[2004\]](#) discuss analytic gaps that hinder analytical reasoning and decision making by employing information visualization. By aiming at process- and task-tailored information visualization, the approaches described in this thesis try to fill these gaps.

A multitude of information visualization approaches exist for representing data having different types, and in the context of this work a variety of visualization techniques are used for presenting patent information as described in [Chapter 3](#). While visual representations are a good means for providing an overview of data items to be analyzed, their effectiveness can be increased greatly by introducing interaction techniques letting users explore details, relate visible data, filter information, and select certain aspects to be inspected more closely or to facilitate further analytic steps. The information visualization mantra stated by [Shneiderman \[1996\]](#) emphasizes these interaction aspects and suggests how tasks can be supported through interactive visualization:

“Overview first, zoom and filter, then details-on-demand”

Apart from providing different information visualization perspectives, the approaches proposed in this work make extensive use of interaction techniques as well, which are required to facilitate in particular analytic tasks based on such

views. Basic interaction techniques that are realized as view transformation, such as zooming and panning, help users to focus on specific regions or data items depicted in a view. Apart from zooming and panning, which is supported by many of the discussed views, overview+detail, focus+context [Cockburn et al., 2009], and brushing & linking in context of multiple coordinated views [Roberts, 2007] are used. Advanced filtering techniques and visual query definition also play an important role for the approaches presented in this thesis. They are presented in the next section in the context of search user interfaces. A variety of focus+context have been described in publications. Prominent examples are ‘Fisheye Views’ as presented by Furnas [1986], ‘Magic Lenses’ as described in Bier et al. [1993], or the ‘Table Lens’ introduced by Rao and Card [1994]. An overview of such approaches can be found in Card et al. [1999]. Focus+context techniques are deemed to be superior to overview+detail approaches in certain situations, since they do not break with visual workflow, resulting in a lower memory load of users and better visual search performance for explorative tasks. As part of the work presented in this thesis, a focus+context technique called ‘EdgeAnalyzer’ is described in Section 4.1.

However, focus+context techniques take effect within one single view. While different information can be integrated into one view, there is certainly a limit before a visualization gets very difficult to comprehend and too visually overloaded to let users get an overview of the shown data. In such a case it is often preferable to have differently detailed perspectives on data aspects. With overview+detail methods, users are supported in coordinated drill-down tasks. In order to visually analyze different data aspects at once, brushing&linking techniques can be applied as often available with multiple coordinated views (MCV) (see Roberts [2007] for an overview on the topic). Here different aspects can be visually related by constraining one aspect interactively, and observing the characteristics of other aspects under this constraint. MCVs are a key technique of the approaches described in this thesis.

There is a plethora of related work in the field of information visualization that addresses visualization variants for different data types and structures, as well as suitable interaction techniques to let users interactively explore and exploit the presented information. A historical overview of the development and employment of early examples of information visualization can be found in Tufte [1986]. Card et al. [1999] provide a selection of computer aided approaches in the field, addressing information visualization in general and models for information visualization as described below in more detail. A comprehensive work dealing with visualization and perception aspects is available with Ware [2004] and Aigner et al. [2011] describe visualization approaches that specifically consider time-related data.

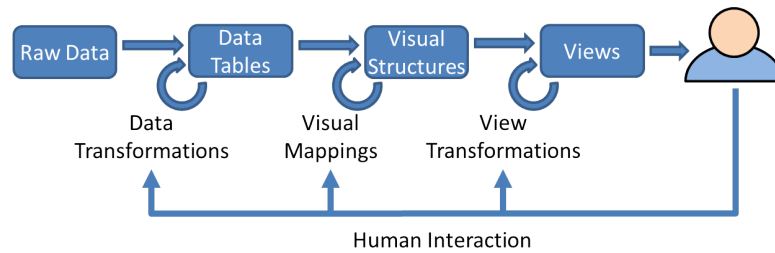


Figure 2.1 — The information visualization reference model according to Card et al.

2.3 Visualization Models

A variety of abstract models for designing information visualization approaches have been suggested. Among the most well known are the state chart model as proposed by Chi [2000] and Card et al.’s [1999]’s information visualization reference model. Haber and McNabb [1990] previously introduced a model for the visualization pipeline, which was extended later by dos Santos and Brodlie [2004]. The latter two give an abstract view on the visualization process in general. Since this work deals mainly with the analysis of abstract data, especially those explicitly addressing information visualization are of importance in this work’s context. Figure 2.1 shows the information visualization model according to Card et al. [1999].

Card et al.’s model is divided into several stages starting with raw data that is subsequently transformed into data tables. These data tables are enriched to visual structures by mapping them to visual attributes such as position, spatial extent, etc. [Bertin, 1967]. Finally, the visual data gets rendered into a view that is perceived by a user. In interactive environments the user can interact with systems following Card et al.’s approach in different ways. User interaction such as zooming and panning can be realized by changing view transformations in the rendering step. Interactions for modifying the visual mapping, such as switching to a different color schemes or changing the layout of a view, apply directly to the mapping step. Interactions that trigger changes in data filtering and aggregation functions affect the data tables.

Visualization toolkits, systems and products, such as Improvise [Weaver, 2004], Polaris [Stolte and Hanrahan, 2000], and Prefuse [Heer et al., 2005], just to name a few, adhere to the abstract scheme of the information visualization models. The most obvious reason for their lasting popularity lies in the models’ flexibility and, taking a software engineering perspective, in the separation of concerns they provide. This separation of concerns guarantees flexibility regarding the integration of different data sources and different visual perspectives. Tang et al. [2004] provide

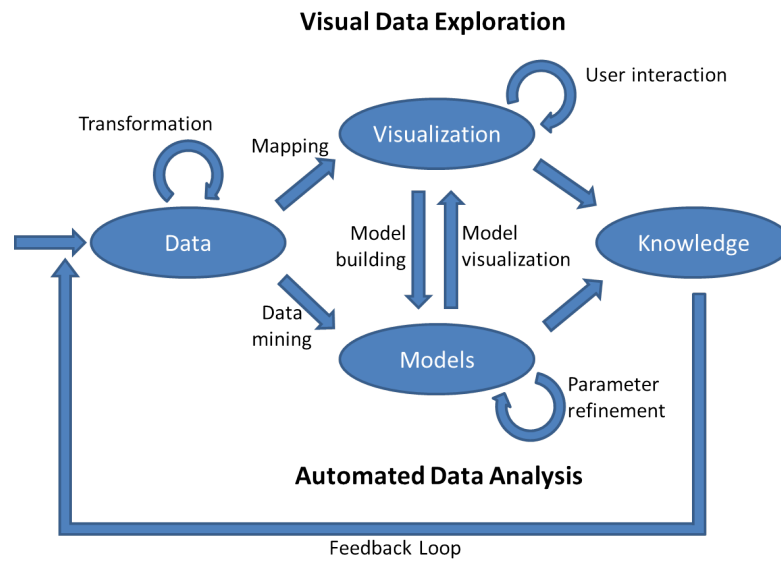


Figure 2.2 — The visual analytics process according to Keim et al. [2008]

an interesting discussion of this separation and distinct implementation strategies of the stages. They also come to the important conclusion that integrating data sources and means to access them within the same visual tool can improve the analytic process as a whole. This is a detail which is important in this work.

As a result of the abovementioned separation it is possible to branch models at different stages in order to support different usage scenarios. Splitting the data flow at the data tables stage allows for the creation of multiple visual perspectives on the same data. Such an approach can be used to build a system supporting MCV. Branching at the data source and raw data level makes it possible to visualize and explore different data sources or different filtered perspective of the same data source in parallel. Splits that occur at stages ‘Visual Structures’ and ‘Views’ are not so common but have also been exploited as part of collaborative approaches, e.g., for working on the same view of a data set in parallel at different locations or to show the same information in different views (see, for example, Tobiasz et al. [2009]).

2.4 Visual Analytics Models

With the introduction of visual analytics by Thomas and Cook, some new orthogonal aspects have to be addressed, at least more explicitly than they were stated with Card et al.’s model. A schematic view of the visual analytics process has been

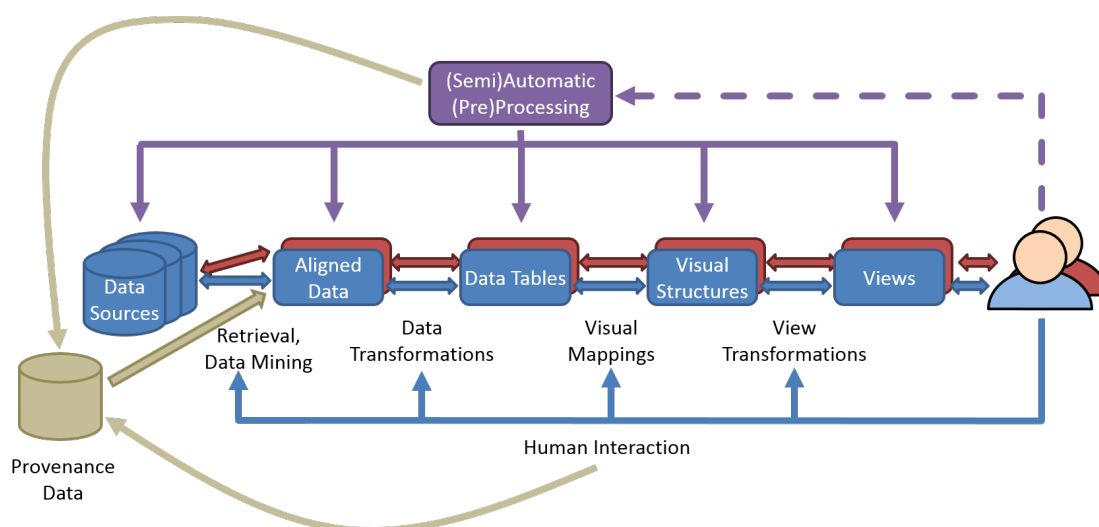


Figure 2.3 — Extension of Card et al.’s model for including visual analytics processes

described by Keim et al. [2008, 2010] emphasizing data model creation and data mining explicitly (see Figure 2.2). Keim et al. [2006] also adapted Shneiderman’s mantra to fit visual analytics approaches into:

“Analyse First – Show the Important – Zoom, Filter and Analyse Further – Details on Demand”

The information visualization reference model can be seen as an abstract model that does not restrict its usage to information visualization, but can be utilized as well to base visual analytics approaches on it. However, since visual analytics approaches define an additional set of typical characteristics, it is helpful to depict them by extending or rather concretizing these models.

Card et al.’s reference model was chosen for this purpose, since most of the visualization approaches presented in subsequent chapters adhere to the reference model. It is a good basis for depicting visual analytics approaches, since the users or analysts and the ways they interact with the system is important in the subsequently described work and should therefore be explicitly modeled within it (see Figure 2.3). The proposed extension for the reference model can be mapped to Keim et al.’s visual analytics process model easily and vice versa.

In many cases visual analytics scenarios have to deal with a variety of different data sources. This can be abstracted as seen with Card et al.’s model and with the alternative model for visual analytics processes of Keim et al.. However, taking them

into account explicitly for describing visual analytics approaches seems justifiable, if considering that many visual analytics applications do not start with a raw set of data that only has to be filtered or transformed. Typically, there are back-end systems, such as databases, repositories, or streaming interfaces involved that should be considered as an integral part of an VA approach. Without integrating them seamlessly into visual analytics processes, analytic scalability can hardly be achieved. The integration will get even tighter, if tasks-tailored retrieval strategies are going to be exploited and user-created tools can be directly applied at this very first stage of data production/recording. *Data sources* are therefore explicitly represented in the proposed model extension.

Because visual analytics approaches aim at solving real world problems, it is more than likely that an application specific data model exists or has to be created, which does not adhere to the idea of data tables (with exception to those working directly on relational information such as described by [Keim and Kriegel \[1994\]](#) or [Stolte and Hanrahan \[2000\]](#)). Accordingly, this has to be acknowledged by introducing another stage in the pipeline that represents the collected and derived data from potentially different sources as *aligned data* (see also [dos Santos and Brodlie \[2004\]](#)).

The rest of the model remains unchanged with respect to the stages proposed by [Card et al.](#). However, collaborative scenarios and the usage of different perspectives on the data to be analyzed are symbolized by the splitting of the pipeline into branches. It has been argued that the visualization pipeline in its proposed form does not meet the requirement of providing suitable back channels for data [[North et al., 2010](#)], since it represents a data-driven approach to information analysis and does not consider feeding back insight or semantics learned during analysis into the data model. They are, of course, right with their judgment that the visualization pipeline represents a data driven perspective, but a very abstract one.

Even if not formulated or depicted explicitly by [Card et al.](#)'s model, there is no reason why interaction should not feed back information into a data model, be it model updates, analytical insight/semantics, or provenance data about the analytic task in general. The proposed extension into a visual analytics model as shown in [Figure 2.3](#) considers these back channels with bi-directional arrows between the stages.

Furthermore, *(semi)automatic* processing potentially taking effect at every stage of the process is introduced. These methods might be applied without involvement of users, but can also be triggered, parametrized, or even created on analysts' interaction as well. As a final enrichment the recording of provenance information, either generated from automatic procedures or captured from user's interaction with a visual analytics system is indicated.

The information visualization pipeline as well as the sensemaking process described by Pirolli and Card [2005] (see Section 2.7) represent two sides of the same coin – a data driven or architectural view and a task-tailored or process-based perspective that can be brought together in visual analytics approaches as described in this thesis by the example of patent search and analysis.

2.5 Document and Data Retrieval

Methods for retrieving information from large databases and repositories have been developed since the very beginning of the digital age. A related research discipline, *information retrieval*, has evolved during the years. For the domain of retrieving text documents Manning et al. [2008] suggest the following definition for this area of research:

“Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).”

But not only text documents fall into the category of being unstructured in terms of data characteristics, images, audio, as well as video data exhibit the same properties, meaning that their semantic structures are typically not available for computational processing explicitly. Of course text documents do have structure such as title, headlines, paragraphs, etc., which is acknowledged by employing formalisms for creating semi-structured documents, such as XML formats. Any syntactic or even semantic structure however is not available directly for machine processing. Closely related to the field of text document retrieval is the domain of natural language processing (NLP). Both fields have some overlap regarding employed methods and data models.

In this work, information retrieval approaches are the base technology applied to searching and analyzing intellectual property documents. As described in more detail in Section 2.10 such documents are most often not only monolithic text documents but include images, formulae, etc. which makes their characterization as being multimodal or multimedial suitable and might require to take into account other unstructured data as well.

2.5.1 Information Need

As has been mentioned above retrieval tasks are performed as a result of a user’s *information need*. Manning et al. describe it as follows:

“An information need is the topic about which the user desires to know more, and is differentiated from a query, which is what the user conveys to the computer in an attempt to communicate the information need.”

This description already points out the discrepancy between what a user wants to retrieve and how this need is formulated. But there is also a qualitative aspect to information needs that has to be considered, in particular in context of analytic tasks. Information needs are not always clear from the very beginning of an analysis, moreover information needs might shift or new information needs may arise during the analysis of search results. As a consequence, methods that take these issues into account and let analysts change their focus during search and analysis, e.g. in form of providing explorative techniques, are required. A more detailed description of this topic in context of visual search interfaces can be found in [Hearst \[2009\]](#) (Chapter 3).

Besides their unstructured contents, a variety of structured bibliographic metadata is attached to documents such as patents and scientific articles. Such structured data is usually stored in traditional database systems for example relational database management systems (RDBMS). The search for, or better, accessing structured information from database systems are not considered as being part of information retrieval. However, both kinds of information play an important role searching and analyzing intellectual property documents, and, as a consequence, different mechanisms have to be foreseen to search for or manage them. Therefore, the terms *document retrieval* and *data retrieval* are used in the following to cover these two aspects.

2.5.2 Text Document Retrieval

With respect to the topics touched in this thesis, Boolean retrieval of documents but also vector space models play an important role (see [Baeza-Yates et al. \[1999\]](#) and [Manning et al. \[2008\]](#) for details of the topics touched in this section). Text documents are often processed as a bag-of-words model, meaning that in such a model the sequence of words within the text document is not taken into account in the model representation. After tokenizing documents into single words, stemming is often applied in order to abstract word forms that have the same stem but different suffixes as a result of declination and conjugation. Often, bag-of-word models are represented as vectors, which are typically high-dimensional but sparsely populated, since each word, or stem is represented as one dimension in the vector.

Such vectors can, for example, find application in the vector space model, on which a variety of text document retrieval approaches are based on. Since documents have different lengths (with respect to the terms or words they contain), these

vectors are typically length normalized. For retrieval it is also helpful to consider the importance of a term with respect to its occurrences within a document and its distribution over the corpus². One idea here is to give words that are widely distributed over many documents in the corpus less emphasis than words that occur more rarely, since the latter are potentially better for discriminating documents than frequent ones. This can be achieved using *inverse document frequency* (idf_t). Document frequency df_t describes the number of documents in a corpus containing a specific term t . Inverse document frequency is defined as

$$idf_t = \log \frac{N}{df_t},$$

whereby N is the number of all documents in the corpus.

Moreover, terms that occur often within a document are obviously better descriptors of its content than less frequent ones. This can be described using *term frequency* $tf_{t,d}$ which accordingly specifies the occurrence of one term within a document d . The combination of both led to the idea of the *Term Frequency - Inverse Document Frequency* ($tf-idf$) weighting scheme of terms which is exploited for increasing retrieval effectiveness and for a better ranking of results:

$$tf-idf_{t,d} = tf_{t,d} \cdot idf_t$$

In the vector space model such term-weights are added as values with the corresponding dimension of the documents' vectors.

The basic idea of retrieval approaches employing the vector space model is that similarity of documents can be determined with a distance measure, defined for these document vectors. For information retrieval tasks, the vectors of a document corpus are stored in the index structure of a text repository. Keyword queries to such a system are simply transformed into (weighted) vectors as well and the most similar documents in terms of the abovementioned distance measure can be returned. One of its biggest benefits over strictly Boolean approaches is that the vector space model facilitates ranking of results, typically taking into account the similarity of documents to a given query.

For evaluating the effectiveness of retrieval approaches a variety of measures have been proposed. Among the most commonly used, as within this thesis, are *recall* and *precision*. For patent retrieval in particular, recall is important, since missing relevant documents is not acceptable for most patent search tasks. Recall is defined as the number of relevant documents returned as the response to a query in relation to all relevant documents in the corpus:

$$Recall = \frac{\text{retrieved relevant documents}}{\text{all relevant documents}}.$$

² Collections of texts are referred to as 'text corpora' or 'copora' for short in the NLP domain

However, recall does not account for irrelevant documents (false positives) that are likely to be returned as well. As a consequence, precision has to be taken into account, since nobody wants to browse through masses of irrelevant documents. Here the result set's quality regarding contained relevant documents is measured:

$$\textit{Precision} = \frac{\textit{retrieved relevant document}}{\textit{all retrieved documents}}$$

In order to create one score for measuring both the F -score was proposed and is now widely used. The F -score depicts the harmonic mean of precision and recall:

$$F = \frac{2 \cdot \textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

More specifically this score is usually termed F_1 score, indicating that precision and recall are weighted equally for the computation of F . It is also possible to apply different weights to recall and precision respectively, but for comparability, F_1 will be used in this thesis.

2.5.3 Machine Learning

Machine learning techniques can be exploited as well and for enhancing document retrieval. This section provides related work and background information on those techniques, which are employed for classifier creation in Section 4.2. Machine learning approaches are categorized into *supervised* and *unsupervised* methods according to whether they rely on *labeled* data or not. Thereby labeled data means that typically a human annotator has labeled data which is exploited to train the machine learning algorithm. In context of retrieval tasks both variants find application. Unsupervised methods are, for example, used for clustering documents automatically, while supervised machine learning techniques need labeled training examples in order to be created.

In order to enable analysts to understand and perceive clustering of data, visual representations can be a suitable means. If the clustered data to be shown is high-dimensional, as is the case with documents represented in the bag-of-words model, methods have to be applied to represent the results in two- or three-dimensional space. A broad variety of projection and down-scaling techniques exist, either independent of the clustering algorithm itself or integrating them with projection methods. Linear, e.g. principal component analysis (PCA) [Jolliffe, 2005] as well as non-linear, such as multidimensional scaling (MDS) [Cox and Cox, 2000], approaches, have been employed for projecting high-dimensional spaces. Systems and approaches such as InfoSky [Andrews et al., 2002], IN-SPIRE [Wong et al., 2004], and [Wise et al., 1995] make use of such clustering and projection

techniques in order to visualize clustered documents. A relatively new approach combining linear and non-linear computation methods for accomplishing precise and fast dimension reduction has been suggested by [Paulovich et al. \[2008\]](#). Another direction was followed by [Kohonen et al. \[2000\]](#), who use a neural network approach for creating a map from patent data. [Börner et al. \[2003\]](#) provide a survey on techniques that can be applied for dimensionality reduction in context of text document visualization.

The employment of machine learning techniques for document retrieval also depends on the task and how concrete the information need can be specified. Clustering is typically employed in situations where no specific information need is available a priori, since it can provide users with automatic grouping of the data/documents to be inspected. Also variants that let users influence the clustering process actively have been proposed in recent visual analytics approaches (cf. [Joia et al. \[2011\]](#)) In general, the visualizations representing the results of clustering techniques are good for presenting overviews as well as leveraging explorative scenarios. Clustering methods have the benefit of being cheap in terms of effort required by analysts using them, since no labeling is required. Despite these enhanced clustering approaches that can be influenced by an analyst, they are not a good choice in situations where a specific information need exists, since they hardly represent the idea of grouping or separation that matches an analyst's information need.

Classification instead relies on labeled data and users can quite directly express their information need through their labeling actions when annotating examples according to their class membership. In context of this thesis, methods are described for letting analysts create and assess their own classifiers quickly. Here linear support vector machines (LSVMs) are used as the classification framework, which were proposed by [Vapnik \[1998\]](#). This choice was made since support vector machines are known to work well on text classification tasks and they are very fast [[Joachims, 1998](#)]. LSVMs are binary classifiers that aim at linear separation of two classes of a data set. This separation can be achieved with a *hyperplane* also referred to as *decision border* in the following, which is placed in high-dimensional space in such a way that it separates two classes in the training data. In case of document classification, where documents are represented as sparse vectors, it is typically possible to find a linear hyperplane for separating labeled data. But LSVMs do not employ an arbitrary separating hyperplane (\vec{w}, b) , but aim at finding one that maximizes the margin between positive and negative examples (see [Figure 2.4](#)) [[Burges, 1998](#); [Cristianini and Shawe-Taylor, 2000](#)]. Class membership is determined with the following decision function:

$$f(\vec{x}) = \text{sgn}(\vec{w}^T \cdot \vec{x} + b).$$

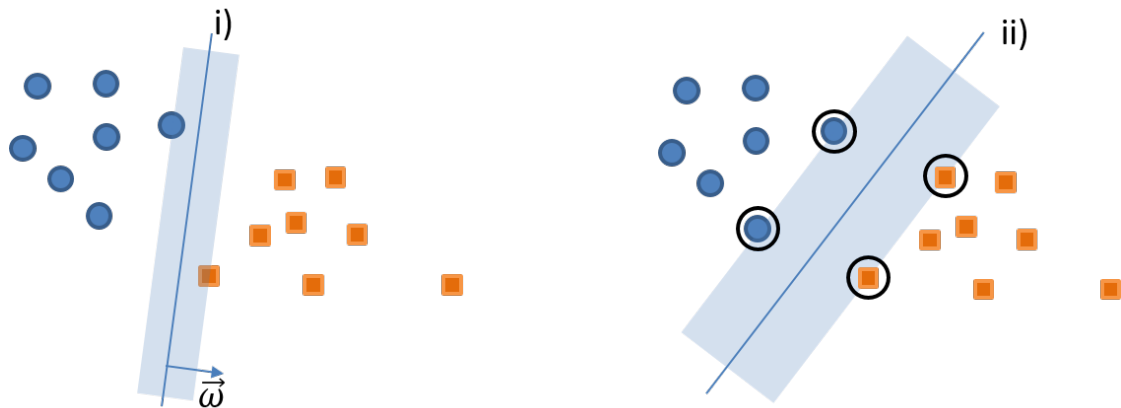


Figure 2.4 — Two-dimensional depiction of binary classification situations with separating ‘hyperplanes’ and corresponding margin. In i) the classes are separated correctly, but the hyperplane is not placed according to an SVM’s optimization criterion of maximizing the margin between the classes to separate as shown in ii). The support vectors are shown in ii) with black circles.

Thereby, \vec{w} describes the normal vector of the hyperplane, \vec{x} is the data item (or document vector in this context) to test and b is the bias to the coordinate system’s origin. As a consequence of maximizing the margin, only those examples that lie close to the class boundary influence the position of the hyperplane and are called *support vectors*. Details for solving the problem of finding an optimal hyperplane can be found in Vapnik [1998].

However, employing classifiers such as ISVMs comes at the cost of labeling effort. Active learning (AL) methods (see Settles [2009]; Olsson [2009]) can greatly speed up the labeling process and reduce this drawback. AL establishes a training/labeling loop, where (typically human) annotators are requested to label training examples and the classifier is subsequently trained with these labeled examples. The basic idea is to choose and automatically present those instances to annotators for labeling, which promise the highest benefit for classifier training, thereby reducing the number of iterations and the annotators’ labeling effort. AL can be applied in situations where a very small initial training set and a much larger set of unlabeled examples are available. Uncertainty sampling [Lewis and Gale, 1994] is one strategy for AL, which picks the training example as the most informative one that receives the lowest confidence (or probability) rating for the label assigned by the classifier. With respect to a support vector machine employed as classification method, this is the unlabeled example that lies closest to the decision border. Due to their

property of only choosing a subset of all training instances that influence the model, SVMs fit well with the concept of AL, because they only rely on a subset of training examples defining the hyperplane [Campbell et al., 2000].

2.5.4 Relational Databases and Other Data Sources

Relational database systems (RDBMS) are very common today. They are based on a relational model which was first described by Codd [1970]. RDBMS store information flexibly in a table oriented manner and provide access to this data in a controlled and managed way. Tables can be connected implicitly via attribute values associated with data items stored in the table's rows. Besides supporting set operations, other specific operations for constraining and constructing sets such as *selection*, *projection*, and *joins* are supported. In context of the work presented in this thesis, relational databases have been used to store especially the metadata (bibliographic) of patent literature. However, the fact that all operations on relational databases deliver sets of entities plays a role for designing the search interface. In case of RDBMS, data is accessed through the well-known Structured Query Language (SQL). For several reasons, the software prototype as described with PatViz in Chapter 3 does not provide the full expressiveness of a relational algebra as SQL does (cf. Section 3.1.1), but enforces implicit joins on patent documents as the primary object of investigation considering the restrictions on the documents attributes. This topic as well as related fields such as data warehouses, data mining, and data integration are discussed in context of visual analytics in Keim et al. [2010]. Since RDBMS are very commonly used, technical details are omitted here for brevity, detailed information on this topic can be found in [Elmasri and Navathe, 2003].

In the context of the PatViz approach, semantic repositories were available as back-end data sources as well. The semantic information was handled according to the model proposed with the Resource Description Framework (RDF)³ – one of the building blocks of the semantic web. In contrast to relational models the information is here represented explicitly through a graph structure. Different repositories for storing such RDF based information are available and with SPARQL, a structured query language for RDF data is provided.

2.6 Visual Search Interfaces

Using information visualization as means for querying textual information from digital libraries as well as representing the returned documents visually is chal-

³ <http://www.w3.org/RDF/>

lenging. One reason might be different processing of verbal information (including written text) and visual information in the human brain as Paivio [1986] suggests with the dual coding theory. Moreover, Ware [2004] states that

“[Written natural language] is by far the most elaborate, complete, and widely shared system of symbols that we have available.”

This does not apply in the same way to visually depicted, abstract information with respect to sharing and understanding.

Text is a perfect means to communicate complex processes but has to be processed sequentially to be understood, while images and visualization can be perceived in parallel and certain details can be obtained much quicker, and, e.g., structure can be represented often better with visualization than with words. It is therefore a valid question to ask why anyone should want to employ visual means for retrieving and representing textual information. And the answer is that no one should, if there is no need for it. If web search is taken as an example, which is certainly one of the most popular and successful applications of an IR technique today, it is obvious that typically no information visualization technique is employed in corresponding user interfaces. This has much to do with the type of information need that exists when a typical web search is performed. Most often a few precise hits that can be found on the first page of search results, satisfy a user’s information need.

However, the situation would be rather different if it is required to find *all* documents talking about a very specific aspect, as is often the case with patent search. It is impossible to read through thousands of documents quickly, since reading takes time and for analytic tasks it can be worth the effort of learning to deal with interactive, visual approaches, if the process can be sped up, can be made more reliable, or can be improved in another way. Furthermore metadata that is available with documents to be retrieved can be visualized more directly and exploited for enhancing search tasks as well.

A comprehensive overview on information visualization approaches for search interfaces can be found in Chapter 10 of Hearst [2009]. In the context of this thesis basically two applications of information visualization for search interfaces are interesting – techniques for visual query definition and visual result set presentation.

2.6.1 Visual Query Definition

Ahlberg and Shneiderman [1994]; Jones [1998]; Spoerri [1993] present approaches for query definition and filtering, however, not all of them are scalable enough or are too difficult to interpret to employ them in analytic tasks. An interesting

approach for advanced filtering has been proposed by [Shneiderman \[1994\]](#) with the filter-flow method, which facilitates direct querying (cf. [Ahlberg and Shneiderman](#)). The latter are commonly applied in an overview+detail manner, letting users define query or filter operations in one view, while the effects of these operations are shown in another one. A similar approach was taken with the selection management facility, which is part of the PatViz system (see Section 3.2.9).

Visual query specification is in particular useful in analytic scenarios, where queries tend to get large or complex, and iterative procedures are applied to refine it. Here, visual representations of the query structure provide overview and help users to quickly understand queries again after intermediate analysis tasks. With the visual query builder as presented in Section 3.1.2 such an analytic, task-tailored technique was created.

2.6.2 Visual Result Set Presentation

The visualization of large amounts of (heterogeneous) data is one of the strengths of information visualization. It is therefore not surprising that many approaches exist for visualizing query results from digital libraries as well as from databases. Depending on the data type(s) a broad variety of information visualization techniques are potentially useful and as has been mentioned before, patent documents are associated with a large variety of bibliographic data of different types. Additionally, methods for aggregating and representing textual information visually are of interest. This section can therefore give only a very coarse overview of related visualization approaches and addresses specifically those that were applied in similar contexts. [Stolte et al. \[2002\]](#) presented a system for representing the results of relational database queries with ‘Polaris’. TRIST by [Jonker et al. \[2005\]](#) enables analysts’ to explore a contrasting view of the results of several queries in parallel [[Proulx et al., 2006](#)]. [Baeza-Yates \[1996\]](#) suggest the usage of visualization specifically for large answers from text repositories, while [Börner et al. \[2003\]](#) provide a good overview on visualization of knowledge domains.

A subtopic of visual results set presentation is the explanation of query results. Typically the terms or constraints that were addressed through a query are highlighted in the result set views. A prominent example of such a technique was provided with ‘TileBars’ by [Hearst \[1995\]](#). Here, search term distribution in result documents is indicated through visual metaphors that present an aggregated overview of corresponding term occurrences in the documents. SeeSoft realizes an alternative approach for depicting query hits in large documents [Eick \[1994\]](#). [Collins et al. \[2009\]](#) proposed a visualization method for showing concepts and their relations to subconcepts in a document set with ‘DocuBurst’.

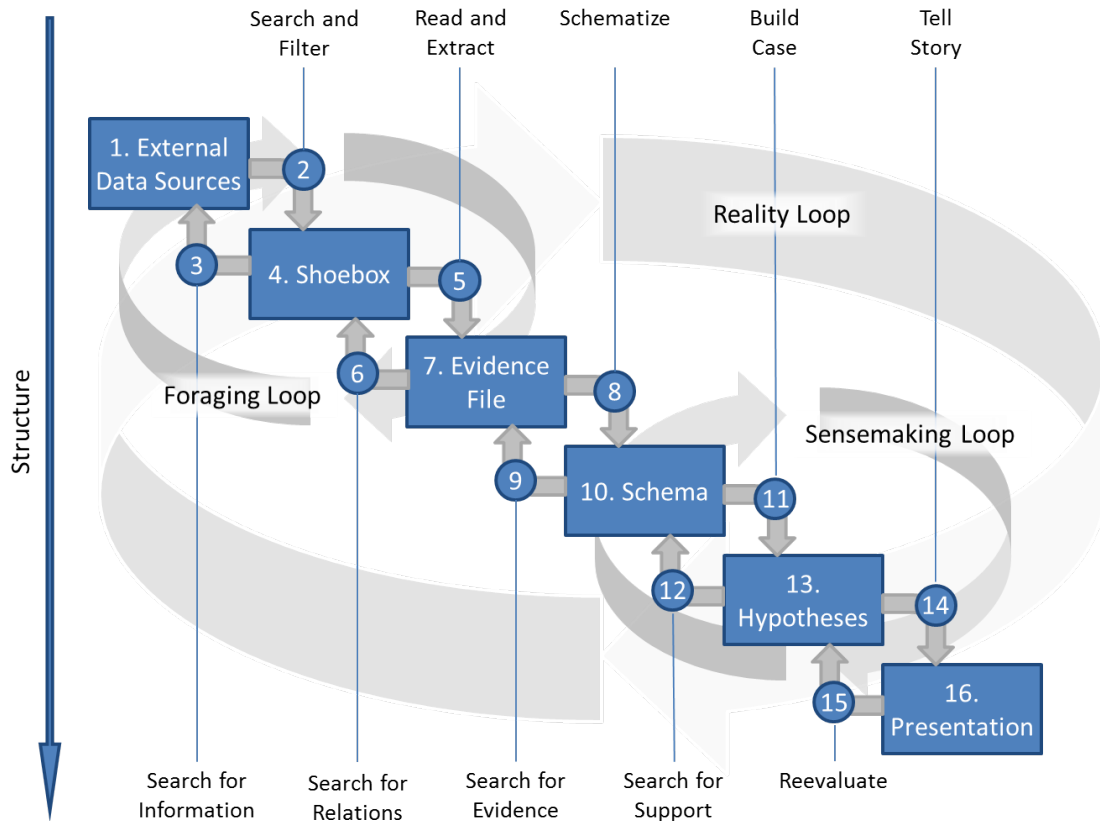


Figure 2.5 — The sensemaking process for intelligence tasks according to Pirolli and Card (simplified)

2.7 Sensemaking

The sensemaking model according to Pirolli and Card [2005], which has been derived from cognitive task analysis of intelligence tasks, depicts several stages of information foraging and sensemaking and the different ways to proceed from one stage to another. Again, it is important that the model does not describe a one-way process. Many different feedback loops on different levels of abstraction are pointed out to move back to earlier stages, or to iteratively carry out the subtask of one specific stage. The two main loops are termed foraging and sensemaking loop respectively. The foraging loop relates to ‘information foraging’ as described by Pirolli and Card [1995], covering all subtasks that are required for searching, collecting, filtering, and preparing the information to be used later in the sensemaking loop. In the sensemaking loop, this information is analyzed, hypotheses are built and tested based on the previously collected data, conclusions

are derived from the information under investigation, before it is finally exploited to take according action.

Pirolli and Card [2005] depict the sensemaking process for intelligence tasks as shown in Figure 2.5. However, the principal idea can be mapped to other analytic tasks as well. From a very abstract point of view, certain tasks in the patent domain can be seen as business intelligence tasks, or at least being part of it. It is therefore not too far-fetched to see similarities with the leverage points Pirolli and Card identify for the foraging loop:

- Exploration-enrichment-exploitation trade-off
- Scanning, recognizing, selecting items for further attention
- Shifting attentional control
- Follow-up searching

All the mentioned aspects incur costs in terms of effort and time spent on them and they can be found in patent search as well. The methods suggested in this theses focus on diminishing the costs of these specific problems.

In order to make the mapping of the sensemaking process easier to comprehend, some of the described activities and stages are adapted to an example scenario of patent analysis in Figure 2.6. Typically, patent analysis processes as described in Section 2.10 have a strong focus on the foraging part, however, this can shift according to the type of analysis that should be carried out. There are however several parallels as well.

In both cases the process starts with searching and filtering external data sources to collect important information into a ‘local’ storage facility called ‘shoebox’ in Pirolli and Card’s model and ‘result sets’ in the suggested model for patent search and analysis. Information that has been found can also trigger new information needs (see Section 2.5) and require additional search tasks or adaptations of the original search respectively. This stored information is screened and filtered further to extract the relevant information from it, which build the basis for the following sensemaking and reasoning subtasks – Pirolli and Card name this the creation of an evidence file, while in the patent process description it is termed (meta)data perspectives. Again, the filtered information might pose the need for identifying additional relevant information that has to be taken into account, either through additional filtering or updating the search. While the steps described so far make up the foraging loop for intelligence tasks, patent search tasks might put even stronger focus on foraging by introducing means for letting users compare and combine patent sets as an another explicit step, e.g., using advanced set management.

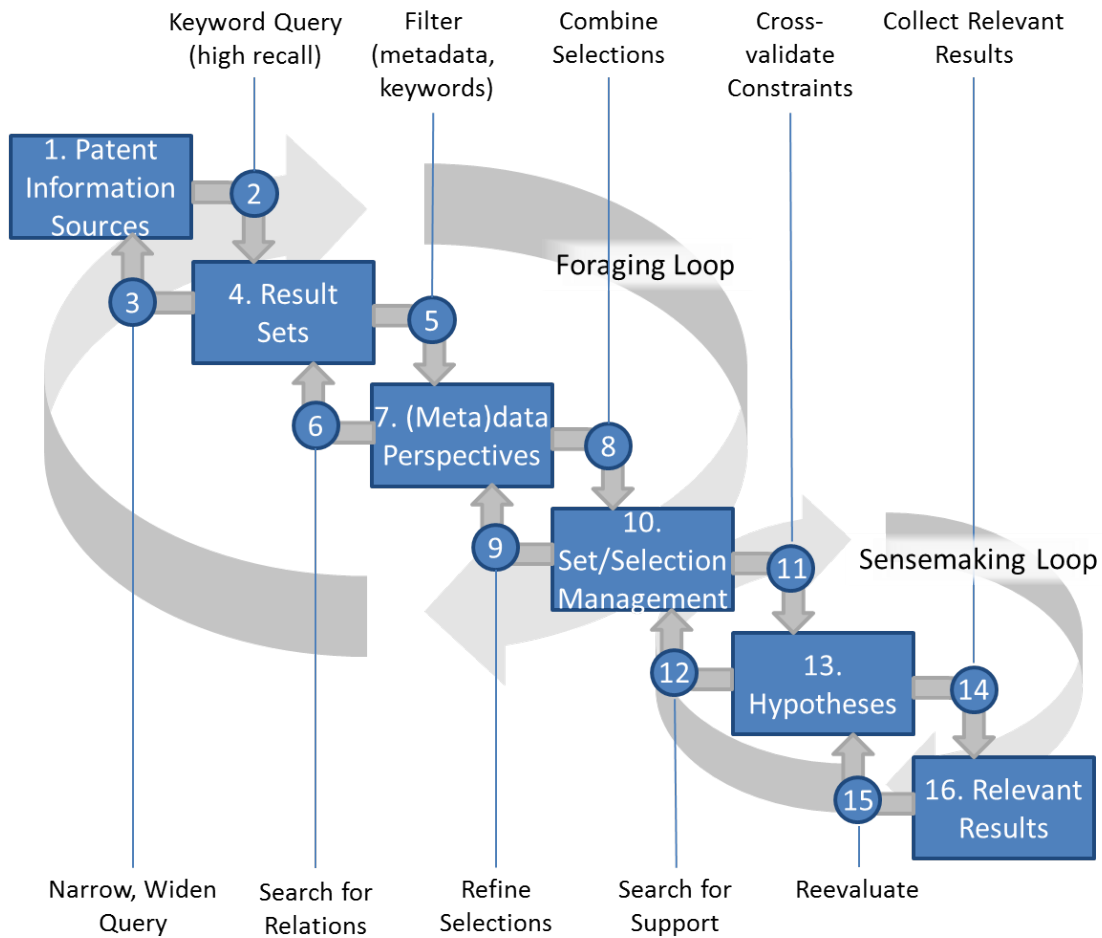


Figure 2.6 — Notional sensemaking process in patent analysis

Sensemaking instead is more characteristic for intelligence tasks. The sensemaking part of Pirolli and Card’s model starts with the creation of schema from the evidence file. Ideally, this schema should represent the analyst’s mental model. Because one contribution of this thesis is to support and integrate all steps visually and interactively, in the model for patent analysis the selection or set management can be seen as the first stage for sensemaking. With the schema/set management as a base, hypotheses can be created either for confirming or invalidating assumptions on the underlying information. Finally, the result of an analysis has to be summarized in order to present it to decision makers or for other further exploitation. Both models can be enriched iteratively with further information extracted from previous steps, which poses the need for feedback loops into the foraging stage.

Even if not always mentioned or depicted explicitly in the description as well as in the diagram, backward loops might not only reach back to direct preceding phases but to any previous stage in the process. As can be seen from the description of the process stages and its subtasks, information foraging as well as sensemaking can be a highly iterative process; this is the case with patent analysis tasks as well.

In contrast to the original sensemaking model, which identifies constant increase of effort when traversing the stages of the process from ‘data sources’ to ‘presentation’, this can be different for patent analysis, since here, the effort put into the activities of the foraging phase can already be very high. The increase in structure as depicted in the model is related to finding connections between information, schema creation, and hypothesis building, which lead to an increase of insight in the end. If the sensemaking process as a whole is supported through computer-based means these derived structures have to become integral part of the underlying data model, or at least the outcome of an analysis should be stored adequately. Otherwise no sustainable use of analytic results is possible.

2.8 Patent Data

Patents documents can be seen as multivariate, multimodal, heterogeneous, high-dimensional data objects. Apart from their multimodal content, comprising unstructured information such as text and image data, a considerable amount of bibliographic information and metadata is assigned to a patent, that can be exploited for search and analysis tasks as well. Bibliographic data that is important in the context of this work are shortly introduced in the following.



Patent publication number: Patent documents are assigned a serial number by the issuing authority on publication in order to make them uniquely identifiable.

Title: A required textual description of the patent document that can be exploited for scanning large numbers of patent documents quickly. Unfortunately, some patent titles are not very distinctive.

Legal entities: The most important legal entities mentioned in a patent are applicants/assignees and inventors. The inventor can also be the applicant of a patent application.

Designated states: The states for which the patent application is effective.

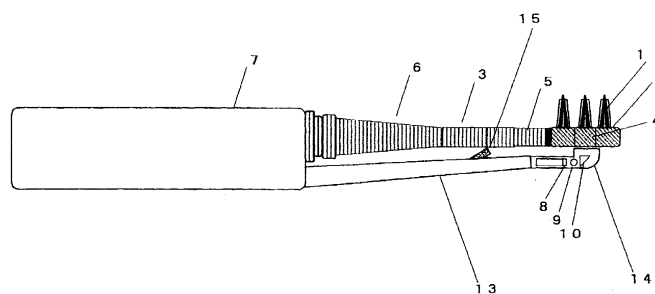
Priority: Applicants can file applications that are based on the same invention in other states by claiming priority for the invention’s first patent application. This is possible within one year after the initial application’s filing. The *Priority Date* is the filing date of the application priority is claimed for.

	Europäisches Patentamt European Patent Office Office européen des brevets	
(19)		(11) EP 1 101 436 A2
(12)	EUROPEAN PATENT APPLICATION	
(43) Date of publication: 23.05.2001 Bulletin 2001/21		(51) Int Cl.7: A61B 1/24, A46B 15/00
(21) Application number: 00125024.0		
(22) Date of filing: 16.11.2000		
(84) Designated Contracting States: AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR Designated Extension States: AL LT LV MK RO SI	(72) Inventors: • Kawamura, Taturou Kyotanabe-shi, Kyoto 610-0351 (JP) • Nakayama, Hiroshi Hirakata-shi, Osaka 573-1114 (JP) • Ooshima, Kiyoko Shijonawata-shi, Osaka 575-0024 (JP)	
(30) Priority: 19.11.1999 JP 33052499 16.12.1999 JP 35812099	(74) Representative: Grünecker, Kinkeldey, Stockmair & Schwanhäusser Anwaltssozietät Maximilianstrasse 58 80538 München (DE)	
(71) Applicant: MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD. Kadoma-shi, Osaka 571-8501 (JP)		
(54) Tooth brush with video device		

(57) A tooth brushing device with a video scope including a video scope having a prism mirror 10 and an objective lens 9 that form light from an object into an image, a CCD unit 8 that converts the light formed into the image, into an electric signal, and a handle section 7 held by an operator, a tooth brush having a brush section having a cavity portion 4 formed therein and a handle section held by the operator, and a display section for displaying an image picked up by the video scope,

in which the handle section of the toothbrush is integrated with the handle section 7 and the prism mirror 10 is arranged where a clear image of a tooth to be brushed by the operator is picked up by the video scope. When images are picked up through the cavity portion 4 formed in the brush section and a toothbrush section and an image pickup section are independently installed in the handle section 7, images can be picked up even during a tooth brushing operation.

Fig. 6



EP 1 101 436 A2

Printed by Jouve, 75001 PARIS (FR)

Figure 2.7 — Front page of a European patent application.

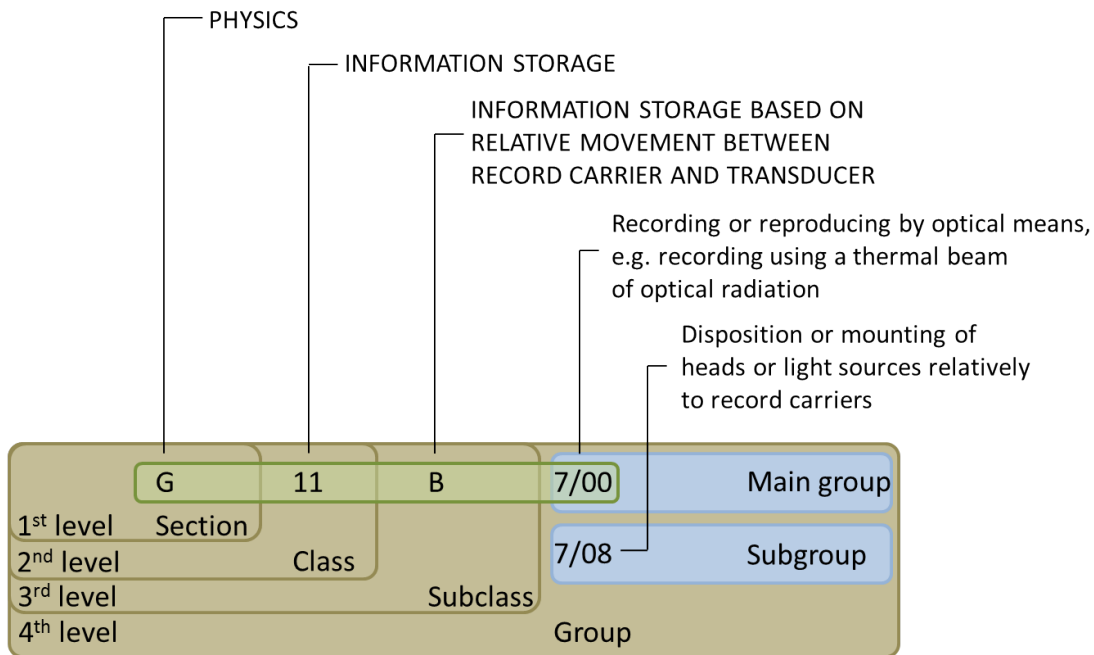


Figure 2.8 — Encoding scheme of the international patent classification (IPC) with an example from the field of optical recording.

Patent family: There are several definitions for patent families. Two of them are commonly used. i) a patent family consists of those patents having exactly the same combination of priorities. ii) a patent family comprises all patents that share at least one common priority.

Patent classification Various schemes for classifying patents exist. among the most widely used are the International Patent Classification⁴ (IPC), the United States Patent Classification⁵ (USPC), the European Classification⁶ (ECLA), and the FI and F-term classification schemes⁷ available from the Japanese Patent Office⁸ (JPO). Throughout the work presented in this thesis the IPC has been used as the main classification scheme and is therefore described in more detail in the following. The IPC is a hierarchical classification and comprises in its current version 8 sections that contain roughly 70,000 subdivisions. The scheme of IPC classification and its encoding scheme down to the group level is depicted in Figure 2.8.

⁴ <http://www.wipo.int/classifications/ipc/en/>

⁵ <http://www.uspto.gov/web/patents/classification/>

⁶ <http://www.epo.org/searching/essentials/classification/ecla.html>

⁷ http://www.intellogist.com/wiki/Japanese_F-Index_and_F-Terms

⁸ <http://www.jpo.go.jp/index.htm>

2.9 Patent Characteristics

Patents are typically conferred to applicants for a maximum time span of twenty years. Inventions have to meet certain criteria such as novelty, description of an inventive step, and they must be of practical use in order to be issued by patent offices. While in force, they protect the owners' invention, thereby granting them the exclusive right to decide who is allowed to make commercial use of the invention and who is not. In exchange for this protection, patent owners have to disseminate their inventions to the public. This means that other parties can access the information that is contained within the patent document. After its expiration, the patent is 'in the public domain' and the formerly protected invention can be used by anyone. As a consequence, the need to analyze patent information is high, even if patents have expired.

The search for and the analysis of intellectual property (IP) rights such as patents is nowadays a common and inevitable task for analysts from a broad range of fields. It is evident that patents are an important factor in today's globalized economy and the amount of patent applications increases at fast a pace [WIP, 2011]. As a result, the effort required for searching, analyzing, and keeping track of patent documents increases accordingly.

Patent specialists in companies are involved in a variety of different tasks. Prior art search, monitoring of competitors, trend recognition, technology assessment, freedom to operate analysis, and objecting to infringing/trivial patents comprise just an excerpt of typical tasks in business life [Joho et al., 2010]. This need to analyze IP-related document does not only emerge for large companies, but also for small and middle-sized enterprises (SME), who are not in the position to maintain their own legal departments. In both cases computational support is required, but especially SMEs rely on external service providers and/or their software products. Thus, software tools and systems that can be used by informed personnel having no extensive legal education, can help to reduce this dependence and decrease costs consequently. Even companies that do not aim at applying for patents themselves, have to stay aware of the development of intellectual property rights in their market segments and domains of interest, if they do not want to put their economic success at risk, especially with respect to globalizing markets. Apart from intellectual property specialists who are involved with the patent strategies of a company and reviewers from patent offices, many other parties are interested in patent information, since patents are an invaluable source of technical knowledge. These include experts from the finance sector, patent lawyers, scientists and many more.

Depending on the task, different aspects of the patent search and analysis process as described later in this section gain more or less focus and also the scope of

data to be analyzed changes. If carrying out, for example, novelty search not only patents and patent applications, but also scientific literature and other sources of IP have to be screened. For freedom to operate search, the patents in force in a specific country or region have to be analyzed. Search, however, is almost for all scenarios an inevitable part of the patent analysis process. Quite often the terms *patent search* and *patent analysis* are used interchangeably when mentioned in context of the tasks described above. This is not surprising, since search is typically performed iteratively and analyzing intermediate results is an integral part to drive the search process. Accordingly, in this work the terms *patent search* and *patent analysis* are also used exchangeably for all IP-related search problems including intermediate analysis as well as to describe more high-level and strategic tasks that rely on previous search steps.

Patent search is a very hard task [Atkinson \[2008\]](#). Unfortunately, not only the rapidly increasing amount of new patent applications and the already available mass of patent information makes patent analysis a tedious task, but also the complexity of available patent material hinders straightforward access to the information needed. For obvious reasons applicants are trying to produce patent applications that still follow the rules of patentability, but they also aim to phrase them as widely as possible to achieve a maximum of coverage for their patents. Occasionally, there also seem to be tendencies to obfuscate patent texts intentionally, e.g. by using terms that are not typical for the corresponding technical field, probably in order to prevent competitors from obtaining easy access to these patents. Because patent applications can address a large variety of technical fields, even without obfuscation, the language used within these sectors might differ greatly due to terminology and phrases that are commonly used within one of these fields. Especially the claims section of patent documents makes use of legalese, which further deteriorates their readability and as a consequence reduces their retrievability for inexperienced patent searchers. This led to the coining of the term *patentesse* [[Singer and Smith, 1967](#)] to name the very special language used in patent documents. Furthermore, some patent applications are multi-lingual, others are only accessible in the language of the country where they have been applied for. In addition to the difficulties described above, all problems common to the retrieval of natural language texts (e.g. ambiguities) further increase the complexity. But patents do not only have textual content. They can contain tables, figures of different nature, such as schematic drawings, electronic circuits, diagrams, chemical formulas, gene sequences, etc. Additionally, there is a plethora of metadata associated with patents like title, abstract, applicants, inventors, classification information, dates for publication, applying, granting them, as well as legal events, designated countries patents are in force in or applied for, citations of other IP documents, etc. Aside from patent

documents, other sources of IP have to be considered and it is quite common that multiple databases and repositories have to be accessed during search.

Several scalability issues have to be considered during the development of patent search and analysis approaches as described in Chapter 5. Large repositories of patent literature pose one challenge. The data contained in patent documents can be characterized as high-dimensional, multi-modal, heterogeneous, and ambiguous. Typical tasks are cost extensive in terms of required time, effort, and required expertise.

In practice many other issues have to be accounted for and cannot be described in detail in this thesis. For example application fields, such as the biochemical domain, require specific search tools, for example, to search for chemical formulas or gene sequences, and different tasks influence the ways and scopes of searching and analysis, which are not covered in this work. Furthermore, only patent documents (with exception to the proposed classifier creation approach in Section 4.2) are taken into account as part of the addressed research problems presented in this work.

To conclude, patents are, aside from their sheer amount, difficult to understand for human users and patent documents exhibit data characteristics, making them hard to process with computational methods. These properties makes patent analysis a field that fits a visual analytics approaches in general quite well, since the problem can neither be solved by human effort alone, nor fully automatically.

A good overview of patent search tasks, strategies, corresponding search scopes, and the problems patent practitioners are faced with, can be found in [Alberts et al. \[2011\]](#). All characteristics of patent search and analysis that are important for understanding the design decisions made during the development of the approaches presented in this thesis are discussed subsequently.

2.10 Patent Search Processes and Analysis Tasks

The effort put in searching for patent literature deviates from task to task, but in contrast to common search problems, such as web search, patent search is typically more exhaustive. Accordingly, patent search is almost always performed in an iterative manner. An initial query is formulated and sent to some repository, the results are inspected, in detail if needed, and the query is reformulated or updated in order to improve the results' quality. The predominant usage of Boolean query approaches in patent search increases this requirements for iterations.

This dominance of Boolean search approaches can be explained from several angles. One reason is the level of control and transparency Boolean search provides. The

effects of a query are directly understandable with respect to matched documents returned as result set. The basic set of operations used in Boolean search are AND, OR, and NOT. Restrictions can be stated with the binary AND operator, which narrows the results to a set of documents containing both expressions. Widening of a query can instead be achieved through connecting expressions with the binary OR operator, resulting in a broader set of result documents fulfilling either the constraints of the first or the second operand. Unary NOT operators invert the results of an expression, which is generally used to exclude a non-relevant set of documents from the results. Typically, Boolean search systems provide extensions such as wildcards and proximity operators for convenience and in order to reduce the verbosity of query formulation. Adding a restriction or widening to an existing query immediately effects the number of returned documents, which gives searchers an idea on the quality of their update. Removing it again, results in exactly the same situation as before. Boolean queries are literally constructed iteratively until the result suffices the searcher's expectation. Even if Boolean queries are straightforward to control, it requires a lot of experience to formulate adequate queries, returning all relevant results and as few as possible irrelevant ones. Being able to control precisely, however, might be a further, psychological reason for trusting such an approach better than others offering less possibilities to influence at least the amount of results.

As has been mentioned, searching legal documents collections is one field where Boolean search is still employed by professionals. Accordingly, Boolean search has been, and still is, the de facto standard for retrieving patent literature and the patent analysts' knowledge in applying this search strategy is comprehensive. As a consequence patent analysts are reluctant to switch or change to other search approaches. However, experience is not only an important factor with respect to patent searchers' familiarity with the Boolean search paradigm, but even more with regard to having expertise within a technical field and its patent-specific terminology. Patent specialists are typically well-trained professionals, and without their profound knowledge which terms to include and the expected number of relevant documents, the construction of high-quality Boolean search queries would become even more time-consuming. Nevertheless, [Joho et al. \[2010\]](#) have identified through a survey patent search tasks to take about 12 hours for completion on average, indicating them to be the predominant cost factor of patent analysis tasks.

Another problem hindering the usage of new automatic techniques is the use of patentese. There have been great advances in Natural Language Processing (NLP) in the last decades, visible especially from the evolution of web search engines, machine translation, and several other applications. Many of these advances are based on exploiting statistical properties of texts, e.g., for creating and training

machine learning approaches. The training data used for accomplishing this are typically taken from web documents, news articles, and other broadly available examples of written text. Unfortunately, such approaches do not always perform particularly well in the patent domain, since the use of terminology of, e.g., training documents extracted from the web and patent documents differs substantially. However, there are first approaches addressing this problem. The European Patent Office (EPO) provides a new machine translation service that has been developed in cooperation with Google. Here, the parallel corpora of European patent documents were exploited during training in order to achieve better translation quality⁹.

Depending on the task the necessity to identify a considerable amount of relevant patent documents can vary according to Trippe and Ruthven [2011]. But most of them require very high, up to full coverage of relevant documents. This means high recall is crucial for patent search. While in other fields than in the legal domain, it is consent that ranked retrieval models such as the vector space model exhibit advantages over Boolean approaches, patent professionals tend to prefer the Boolean model for another reason. The missing of relevant documents during patent retrieval might lead to severe economic consequences. At the point of writing this thesis the fight for securing advantages in the mobile devices market is preformed with unprecedented force and the values in litigation reach the billion dollar range^{10,11}. As a consequence, patent searchers want to be as sure as possible that no relevant documents are missed during search. Or, to put in other terms, they want to have a high level of trust in their searches and the found results. However, it is very difficult do build trust into mechanisms that cannot be fully controlled. Manual control and construction of queries, however, can create this perception of trust even if this can be misleading, when important search terms are missed.

Figure 2.9 depicts the patent search process as a rather abstract model, that might also be suitable for describing other extensive search tasks. In this work this process description will serve as a hook for more detailed explanations in subsequent sections as well as a bridge for describing the integration of the visualization pipeline and the sensemaking process.

By looking at the patent search process and by taking into account the specific properties of patent literature, of patent practitioners' tasks, as well as practitioners' expertise, possible areas for improvement can be identified. In the search stage several enhancements are thinkable. Ideally, the technical search mechanism itself could be improved while considering the specific needs of patent searchers: to find

⁹ <http://www.epo.org/news-issues/news/2012/20120229.html>
and <http://www.epo.org/searching/free/patent-translate.html>

¹⁰ <http://cand.uscourts.gov/lhk/applevsamsung>

¹¹ http://en.wikipedia.org/wiki/Apple_Inc._v._Samsung_Electronics_Co.,_Ltd.

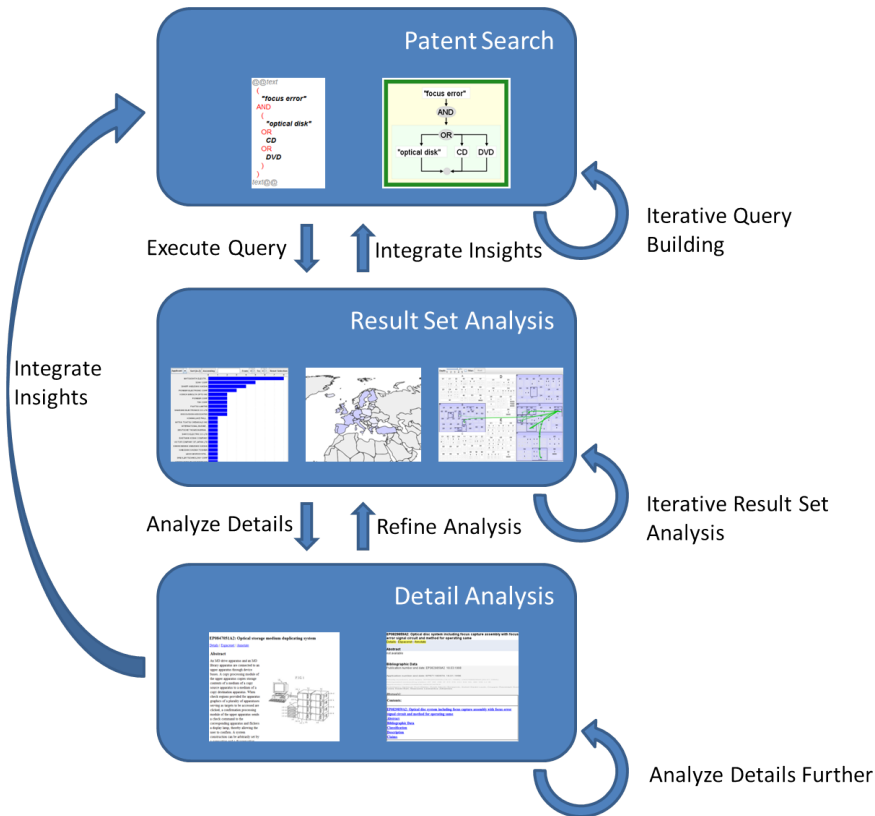


Figure 2.9 — An abstraction of the iterative patent search and analysis process

all relevant documents, while not being overwhelmed with a large number of false positives. Since human patent searchers are involved, supporting them poses a good chance for helping to improve the results. Speeding up the whole search process is one potential improvement, another one provides means that help users to create better queries another one. Ways to make search and analysis methods more easy to learn and to apply is another direction for finding enhancements. All of these aspects are closely related to the leverage points [Pirolli and Card, 2005] identified for the foraging loop.

Visual Patent Analytics

Various enhancements to patent analysis are presented in this chapter. These enhancements are shown based on the PatViz prototype implementation of a patent search and analysis approach. A couple of the approaches and software prototypes discussed in the following were developed as part of the EC project PatExpert¹ [Wanner et al., 2008]. Many of them were extended and refined in the project ‘Scalable Visual Patent Analysis’, which was funded in the first period of the priority program Scalable Visual Analytics² by the German Science Foundation (DFG). Within PatViz, the visual interactive creation of patent search queries forms the basis for these enhancements. It aims at simplifying the creation of complex queries for lay users³, while at the same time improving the overview of the logical structure of the employed Boolean queries, thereby supporting IP specialists as well. Additionally, the visual query representation is an important prerequisite for the implementation of interactive feedback loops, as will be discussed in detail in Section 3.3. To simplify and potentially speed up iterative query refinement, such feedback loops have to be integrated seamlessly, meaning that a system for visual analysis of patent documents has to support them, ideally through direct interaction embedded in an entirely visual environment. A structured visual representation can also serve as the foundation for creating parallel, multilingual search queries, without forcing users to build up similar logical structures redundantly from

¹ <http://www.patexpert.org/>

² <http://www.visualanalytics.de/>

³ Lay users in terms of users having little knowledge regarding the application of Boolean search strategies to patent search, but having the required domain expertise.

scratch. Finally, a flexible visual query-building mechanism not only allows for the integration of multiple patent information sources, but can serve as a basis for integrating advanced search mechanisms, such as user-steered classification.

Multiple visual perspectives on the result set of a particular search iteration enable users to assess the quality of a previous query more quickly and to derive additional insights that are important during strategic tasks. Smaller feedback loops facilitating sensemaking tasks within result sets increase the usefulness of single views. If designed adequately, the analytic possibilities in such an interlinked approach exceeds the expressive power of its single components. Again, suitable interaction mechanisms are needed to establish the relation between different views, test hypotheses, and to keep track of the analysis process as a whole. Interaction within one single view can be seen as the lowest level of feedback loops. Even in these cases, automatic methods can be exploited to let analysts identify properties of interest more quickly by creating plug-in visual analytics approaches on a fine grained level (see Section 4.1).

As described in the previous chapter, support for iterative refinement of search requests comprises an important component of patent search software. Thereby, query modification typically aims at either narrowing the request to remove noise from the results, or at widening the search because the user wants to find additional relevant patents. In either case, the insight that generates a user's desire for refinement is normally gained through the display of the previous result set data. Hence, patent search software needs an efficient feedback loop to transfer insights from result set exploration to the query formulation. PatViz facilitates such a feedback loop, enabling patent analysts to either integrate findings directly from results set views or to integrate more complex findings that have previously been constructed using the selection management technique described later in this chapter.

Current patent analysis products often do not directly support such closed transfer loops - at least not on the larger scale, and through facilitating visual perspectives. Software suites and patent libraries are, for example, available from Thomson Reuters (i.e., Derwent⁴, Delphion⁵, Micropatent⁶, Questel Orbit⁷, Lexis Nexis⁸, and several others). An extensive description and analysis of their use is provided by [Dou et al., 2005] and Hunt et al. [2007]. Freely available search engines include

⁴ http://thomsonreuters.com/products_services/legal/legal_products/a-z/derwent_world_patents_index/

⁵ <http://www.delphion.com/>

⁶ www.micropatent.com

⁷ http://www.questel.com/Prodsandservices/IP_Portal.htm

⁸ <http://www.lexisnexis.com/en-us/products/total-patent.page>

Esp@cenet⁹ from EPO, patent text and image databases¹⁰ from USPTO, the search library¹¹ maintained by WIPO, Google Patent Search¹², etc. Hence, providing additional visual perspectives of patent data has become more and more common during the last number years as can, for example, be seen with the approach offered by Questel. Trippe provides an overview of commercial tools to tackle common tasks for patent search and analysis in Trippe [2003]. A more recent survey can be found in Yang et al. [2008]. Moehrle et al. [2010] also contribute a current outline of commercial systems and relate them to a taxonomy based on a business process model.

Flexible systems for the analysis of relational data offering such feedback loops exist outside the patent domain. An example thereof is Polaris [Stolte and Hanrahan, 2000]¹³, which facilitates the integration of feedback loops and query creation as well. PatViz differs from this approach by integrating multiple back-end search services which are not solely based on relational data and cannot guarantee the completeness of a request's result. Because completeness of results is of high priority in patent search, users are forced to build trust in their retrieved results by carrying out the iterative query process described in Section 2.10.

A variety of other visual analytics approaches for analyzing documents were proposed during the last years. These include approaches for intelligence analysis such as [Görg et al., 2007], scientific literature analysis [Oelke et al., 2010; Dunne et al., 2011; Correll et al., 2011], as well as more generic approaches for dealing with text collections [Wise et al., 1995; Xu et al., 2011; Endert et al., 2012]. NetLens by Kang et al. [2006] enables users to analyze content actor relations as they are available with e.g., authors and corresponding scientific publications, employing an iterative visual approach for query specification as well. Alsakran et al. [2012] proposed an interesting system suitable for real-time analysis and visualization of text streams. An overview of visualizing textual information in digital libraries and knowledge domains can be found in Börner and Chen [2002] and Börner et al. [2003]. Most of them employ multiple coordinated views, as does PatViz, in order to analyze documents under different perspectives, yet they differ from PatViz with respect to the application domain, search approaches, text processing *and* metadata handling, and analytic focus.

PatViz was originally built as a graphical front-end for a set of different search engines and patent document analysis services created in PatExpert. However, the

⁹ <http://www.epo.org/searching/free/espacenet.html>

¹⁰ <http://www.uspto.gov/patents/process/search/>

¹¹ <http://patentscope.wipo.int/search/en/search.jsf>

¹² <http://www.google.com/patents>

¹³ The approach was successfully commercialized by Tableau Software <http://www.tableausoftware.com/>

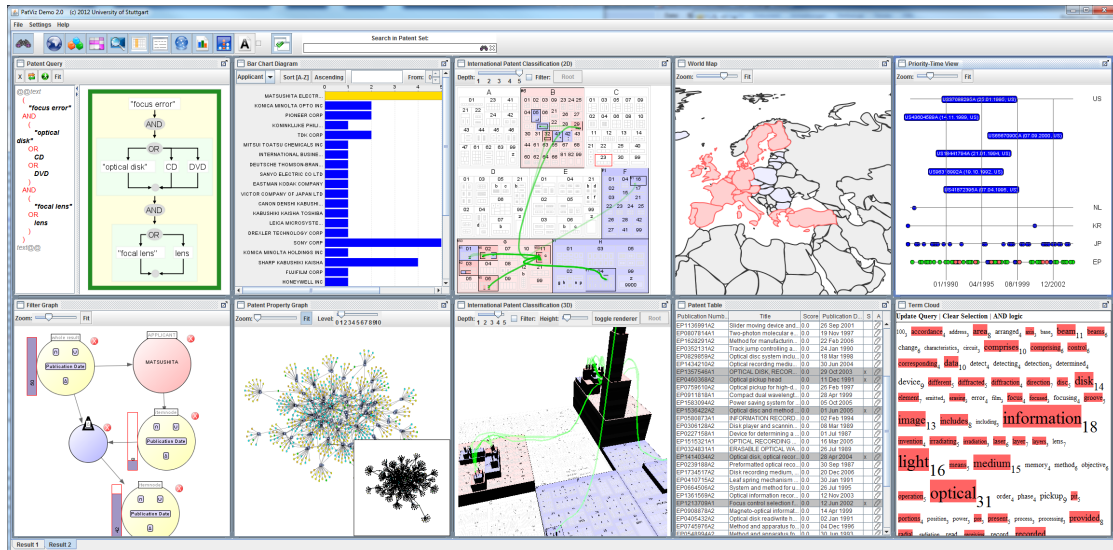


Figure 3.1 — An overview of the PatViz desktop showing a variety of the available views for patent document search and analysis.

patent domains accessible by PatExpert’s back-end systems were restricted¹⁴ to the IPC main classes ‘optical recording’ and ‘machine tools’, reducing the amount of patent documents to about 160,000.

The essential components of PatViz comprise a querying system, a multitude of visual result set representations, and the linkage between them. All these components are bundled in a desktop application that handles the data management and event propagation between components. The visual interface which allows users to build classifiers has also been developed as part of the DFG program Scalable Visual Analytics, albeit in its second funding period, where the scope of the project was extended to the analysis of scientific literature. Figure 3.1 shows an overview of the PatViz visual interface. The implementation of PatViz was accomplished using the Java programming language¹⁵ and makes use of a variety of third-party libraries, including Prefuse [Heer et al., 2005], JFreeChart¹⁶, and Lucene¹⁷.

¹⁴This restriction was necessary to develop the natural language preprocessing mechanisms on the patent material, because these mechanisms needed to be optimized for domain-specific vocabulary. This was a prerequisite for the fulfillment of other scientific objectives in PatExpert and does not reflect any scalability restriction of the graphical front-end.

¹⁵<http://www.java.com/>

¹⁶<http://www.jfree.org/jfreechart/>

¹⁷<http://lucene.apache.org/>

This chapter is partly based on the following publications:

M. Giereth, S. Koch, M. Rotard, and T. Ertl. Web Based Visual Exploration of Patent Information. In *International Conference on Information Visualization (IV 2007)*, pages 150–155, 2007b

M. Giereth, S. Koch, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, and L. Wanner. *A Modular Framework for Ontology-Based Representation of Patent Information*, pages 49–59. IOS Press, 2007a

S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative Integration of Visual Insights during Patent Search and Analysis. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*, pages 203–210, 2009

C. Rohrdantz, S. Koch, C. Jochim, G. Heyer, G. Scheuermann, T. Ertl, H. Schütze, and D. A. Keim. Visuelle Textanalyse. *Informatik-Spektrum*, 33:601–611, 2010

S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative Integration of Visual Insights during Scalable Patent Search and Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):557–569, 2011

S. Koch and H. Bosch. From Static Textual Display of Patents to Graphical Interactions. In M. Lupu, K. Mayer, J. Tait, A. J. Trippe, and W. B. Croft, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Kluwer International Series on Information Retrieval*, pages 217–235. Springer Berlin Heidelberg, 2011

3.1 Visual Interactive Support for Patent Search

Querying the retrieval system is the initial task that has to be performed when working with a patent information system. This is true for almost every use case independent of the analysts' specific goals. In certain situations this step might not be obvious, e.g., during monitoring, when patent applications in a specific technical domain should be observed or a competitor's patent applications are to be tracked, and when working with a 'predefined' patent portfolio. However, in these cases the search task has been carried out beforehand. Similarly, in the case of monitoring, preassigned queries have been defined limiting the patents the user

will be informed about, and any collection in a patent portfolio has very likely been created from one or several previously formulated queries.

As described in Section 2.10, queries in the patent domain tend to get complex and large. In PatExpert, different search facilities and data sources were integrated using one query mechanism. These include *full text search*, *metadata store*, *image similarity search*, and a *semantic repository*. The full-text search engine provides conventional keyword search for the patent analysis systems. Patent full texts as well as all metadata are stored within a relational database. Image similarity search is accomplished by a system based on a vector space model. Thereby, feature vectors are computed from the images through several preprocessing steps. Additionally, semantic information extracted from the patents' section describing the images is used for increasing this mechanism's effectiveness. The semantic information extracted from the patent documents is stored in the semantic web format RDF¹⁸, which is accessible through the mentioned semantic repository. Details on these back-end systems and their integration can be found in Codina et al. [2008].

Each query subsystem has its own formal query language. To facilitate the usage of all query subsystems in one common interface, a method to integrate them as well as their query languages had to be developed first. The combination of different search expressions from different search facilities was realized through a Boolean integration language. While the back-end services were mainly created by partners in the PatExpert consortium, their integration, including the definition of formal languages, was part of PatViz's development as presented within this thesis.

There are also other approaches to tackle visual query definition for search problems, either as part of searching data sets available in a visual interface or, as described here, for querying external data sources. An overview of them is provided in Section 2.6.1

3.1.1 Boolean Integration of Search Facilities

Providing different search engines that can be combined through a Boolean integration language allows for stating complex and powerful queries, but also makes query creation a sophisticated task for the user. To compensate the complexity of the new, combined query language, a visual query editor has been developed that is directly linked to a conventional textual interface. As a requirement, the editor has to provide a clear view of the logical structure of the whole query and an interactive way to create search expressions for each of the different facilities. The result of this integration can be seen in Figure 3.2. In order to create an appropriate

¹⁸<http://www.w3.org/RDF/>

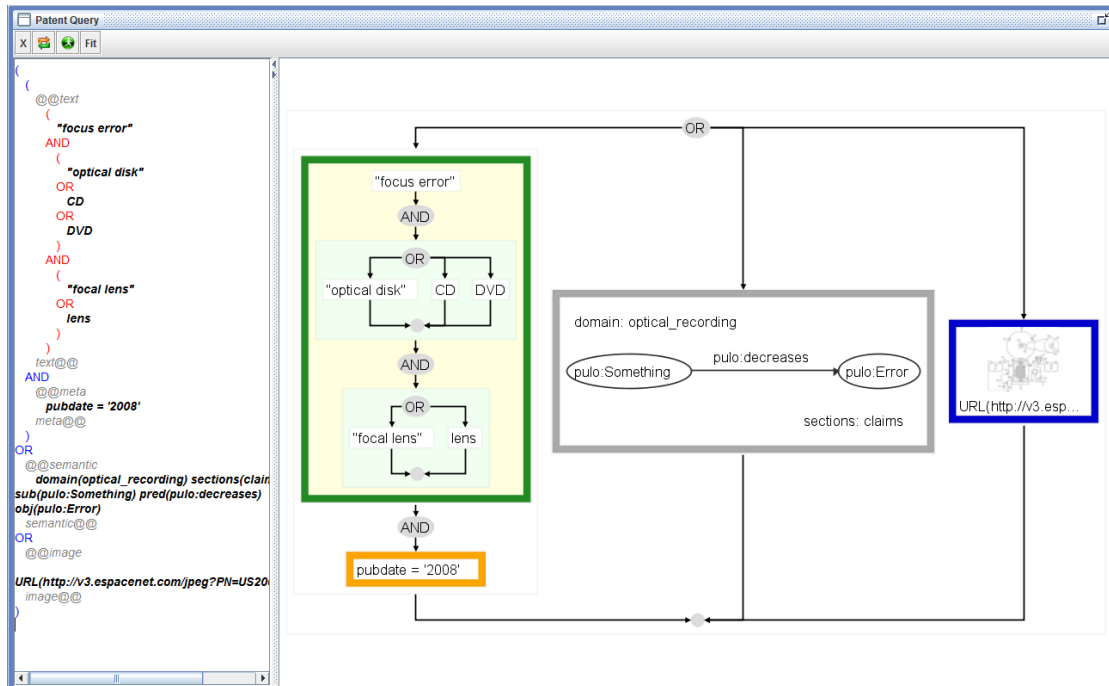


Figure 3.2 — Overview of a query that addresses the four back-end services. The window on the left shows the textual query representation, while the right window depicts the corresponding visual representation. Here, a complex keyword query is shown inside a green box, a metadata query restricting in the orange box, a semantic query in the grey box and an image similarity query in the blue box. Results matching the query either contain the shown Boolean combination of keywords and are published in 2008 as well, or they contain a concept that ‘can decrease an error’, or they contain images similar to the one shown in blue.

visual metaphor for the Boolean integration language, it was decided to use one that is related to the very common Syntax Diagrams according to Wirth [1973]. Therefore, our approach uses node-link diagrams with an orthogonalized circuit-like graph layout as displayed in the right half of Figure 3.2. The set of operators for the Boolean integration language is limited to ‘AND’, ‘OR’ and ‘NOT’.

Directed links describe a combination of these constraints correlating directly with the binary operators ‘AND’ and ‘OR’ in the visualization. A sequential link between two nodes always expresses the ‘AND’ relation and has the semantic meaning that both constraints represented by the connected nodes have to be fulfilled by a patent to pass the filter function. A branching link on the other hand represents an alternative (‘OR’) and has its semantic equivalent in a conjunction of filtered results of every branch that belongs to the same junction. All operator scopes

are represented by boxes in the visual representation. The ‘NOT’ operator, for example, is represented by a box, which encloses the negated constraint as can be seen in red in Figure 3.3. Users can identify which terms are part of which Boolean operation, where the scope of an operator ends, and they can spot the positions where they might want to alter the query easily. Boolean operators can be applied recursively in order to define more complex queries. This is possible locally, e.g., within a keyword query, or globally for designing complex queries to multiple back-ends. Therefore, two kinds of nodes can be distinguished – Boolean operator nodes represented by colored, filled boxes and constraint nodes shown as inside these boxes with a white background.

3.1.2 Visual Creation of Search Statements

The creation and modification of query constraints for all specialized sub-queries as well as their combination is possible by direct interaction with the visual representation. This ensures that only syntactically valid search terms can be created, and it frees the user from the cognitive task of remembering possible filter operations and values by representing them visually. New query parts can be added, through context menus. The location where the context menu is activated within the query visualization is thereby taken into account. Consequently, the offered manipulations are adapted to the corresponding query scope. If activated within the AND block of a keyword sub-query, for example, operations such as ‘surround with OR’, ‘add AND branch’, ‘negate block’, and ‘delete block’ are offered (as shown in Figure 3.3 for a keyword query). If the context menu is activated on a constraint node such as “focus error” in Figure 3.3, a form for editing the constraint is shown. In case constraints for other back-end facilities are to be included, the context menu has to be activated outside the key word query box.

Further interaction functionality allows to zoom and pan the graphical representation of the constructed query. Within the hierarchical graph representation, complex nodes can be collapsed/expanded to further enhance comprehensibility of the query graph.

In the textual view, the logical structure of the query is accentuated by reformatting the input with line breaks and indentation, like in the example in the left half of Figure 3.3. On the one hand, the textual query view ensures that users who are familiar with the query languages of the different search engines can still enter queries directly. This is an important aspect, since professional patent analysts want to create and test their queries quickly, and, mastery of the formal language presumed, entering textual queries is faster than interactive creation. On the other hand, having both representations available can help inexperienced users learn the

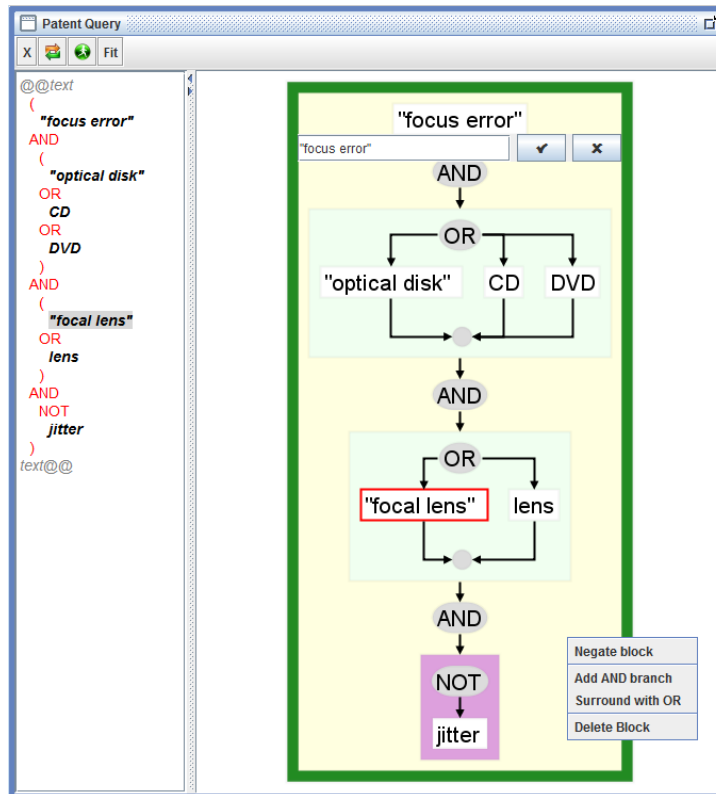


Figure 3.3 — A complex keyword query for retrieving documents that contain the term ‘focus error’ as well as any of the terms ‘optical disk’, ‘CD’, or ‘DVD’ and either the term ‘focal lens’ or ‘lens’ but not the term ‘jitter’. In the left part the textual form of this query is depicted, while the right area shows its visual representation. In the depicted situation the edit form for the node containing the term ‘focus error’ is shown in the visual representation. This dialog can be used to change or update the term. In the lower right region the context menu for modifying the visual query is shown. If a node in the visual representation is hovered (‘focal lens’ in this case) the corresponding textual part is highlighted. Any valid modification of the query in either representation immediately updates the other view.

query language vocabulary and supports all analysts with a structural overview of the logical composition of a query.

Additionally, a highlighting mechanism that serves two purposes was developed. Hovering the mouse pointer over a block in the visual query representation automatically highlights the corresponding parts of the textual representation. On the one hand, this helps to find the corresponding part in the textual representation of a query quickly, if a user intends to apply updates or changes textually, and, on the other hand, users who want to learn the formal language can easily identify textual and visual counterparts. The following sections briefly describes the query mechanisms for the different back-ends, starting with keyword queries and its (possible) extensions considering multilingual queries and templates, metadata search, semantic querying, and image search.

Keyword Queries

The creation of keyword queries is quite straightforward (see Figure 3.3). Keywords can be combined in an arbitrary manner using Boolean operators to create more complex constraints, as has been described above for the combination of queries for different search back-ends. The Boolean keyword search is realized by employing a Lucene¹⁹ repository as a back-end data source. Therefore, a variety of additional features and operations can be employed for searching. These include wildcards, proximity search, and term boosting.

Besides these extensions, Lucene allows for Boolean ranked retrieval by applying Boolean search mechanisms and employing the vector space model for ranking afterwards (see Section 2.5.2). So far, this ranking can only be exploited in PatViz if no other search back-ends are employed, since not all of them provide scoring values for the retrieved results, and even if they did, it would be problematic to achieve alignment of different scoring models to come up with a reasonable ordering of the results. However, ranking methods for other back-ends, such as for relational databases are theoretically possible. Fuhr and Rölleke [1997] presented an approach for probabilistic models in databases that could handle uncertainty of relations as well as ranking of the retrieved relations. In contrast to the approach of Fuhr and Rölleke, experiments that introduce ‘fuzziness’ on the operator level were undertaken in PatExpert [Codina et al., 2008]. However, this rather late development is not reflected in the visual interface and all its views for result set inspection. A variety of tools from the legal domain that employ Boolean search (without ranking) show the results in chronological order, newest first, with the assumption that new results might be more relevant than older ones. Through the various result set perspectives available in PatViz, chronological ordering is

¹⁹<http://lucene.apache.org/core/>

therefore one besides many other possibilities to depict a patent result set (see Manning et al. [2008] for details on ranking).

Multilingual Patent Search and Query Extension

An important problem in patent search is the multilingual nature of patent literature. As a consequence, it is often necessary to search for documents in several different languages. Jochim et al. [2010, 2011] made several suggestions how patent queries can be expanded with multilingual translations. Exploiting the fact that parts of European patents are available in English, French, and German, makes the creation of aligned, domain-specific translation dictionaries possible. Several initiatives have addressed multilingual patent retrieval in recent years. Since 2009, the Cross Language Evaluation Forum (CLEF)²⁰ sponsors an Intellectual Property track with different subtasks dedicated to crosslingual information retrieval (CLIR). Other initiatives, such as NTCIR²¹, organize separate workshops for both CLIR and patent retrieval since 2002. The benefits of using domain-specific translations over general-purpose translation lexica become obvious from receiving higher recall for the domain-specific variants when applied to the CLEF-IP 2010²² dataset. Such aligned translation dictionaries not only allow for multiple different synonymous translation, but also have a value for translation probability attached to each entry pair. This is especially interesting since translation probabilities are not bidirectionally equal, meaning that, e.g., a German term's probability to be translated into a certain English term might be different from translating the English target term back into German. Patent experts are typically proficient in several languages; however, it is likely that their foreign language skills differ from language to language.

Patent queries are often created according to a certain strategy. This applies in particular to the construction of the keyword query part, which constitutes the main area being affected by multilingual query creation, while other aspects, such as constraining metadata, semantic concepts, and images, are less language dependent. Here, several aspects important to a patent might be addressed, for example, when searching in disciplines such as mechanical engineering, the problem to be solved and the solution to be applied can be of interest. Accordingly, a query is constructed that tries to capture these aspects by different search terms and phrases. Synonyms for both aspects are likely to be introduced either ad hoc or during subsequent iterations to optimize the queries' coverage.

²⁰<http://www.clef-initiative.eu>

²¹<http://research.nii.ac.jp/ntcir/>

²²<http://www.ir-facility.org/clef-ip>

When searching for these aspects in multiple languages, similar logical structures can be applied for each language²³ and integrated within one combined search structure by the Boolean OR operation. Users can be supported by providing semi-automatic translation of a query formulated in a single language. For these translations domain-specific dictionaries can be exploited. By relying on those, not only the structure of the initial monolingual query can be replicated, but also the most likely translations of terms and synonyms for the problem at hand can be suggested automatically. Such an approach might be especially beneficial in the abovementioned situation, when analysts are not equally well-skilled in different languages. An initial suggestion for a translated query can be generated based on extracted dictionaries without the need for user interaction. Based on a first translation, improvements, such as selecting additional or different synonyms, can be accomplished in a straightforward manner. Currently, the described mechanism is not realized as part of the PatViz interface, but future work will aim at its incorporation for searching scientific literature and patent material. While from a logical point of view the creation of query translations is already possible (see Figure 3.4), some small adaptations are required to support analysts in exploiting it. One prerequisite is that a searcher formulates a Boolean keyword query using only one language and the translation operation is initialized explicitly afterwards. This is important, since it might be difficult to identify the language of single query terms if the original query already mixes terms from different languages.

If this requirement is fulfilled, automatic detection of the query language might be possible or can be provided explicitly by the analyst. Afterwards, suggested translations can be presented in a copied logical structure of the initial query. Another requirement is to keep track of the language affiliation of copied structures, which is not tracked by the system in its current version. If this is provided, users can change terms in one structure while having updated others automatically, or at least be presented with new suggestions on translations. As a consequence, even suggestions for enriching the original query with further synonyms becomes possible by using the opposite translation direction. Later changes in the logical structure of a multilingual query are more difficult, since an automatic procedure will not be able to distinguish between logical changes and the introduction of additional synonyms, for example. However, adding synonyms in translation will also not harm automatically-created translations as long as the scope of introduced synonyms is acknowledged correctly. A method for turning off the automatic enrichment of translated queries has to be provided in situations where such a behavior is not intended.

²³ Depending on the relatedness of languages and the desired keyword restrictions the logical structure might have to be adapted for very different languages such as English and Chinese for example.

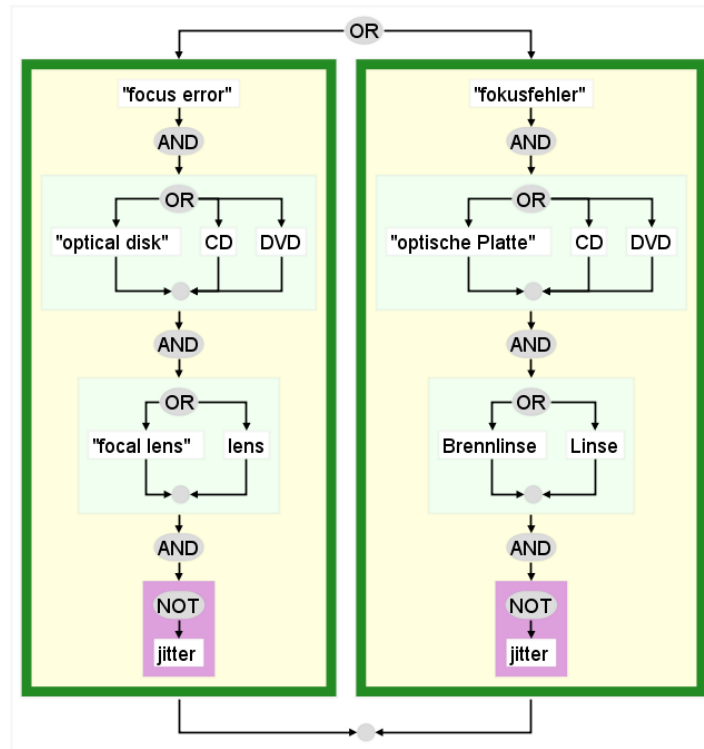


Figure 3.4 — Multilingual keyword query. The left branch depicts the original English query while the right one shows a ‘derived’ German query. In the shown case both query parts were constructed manually, but mechanisms for providing translation suggestions in both directions (semi-)automatically could be integrated to support patent searchers in query formulation.

Query Templates

Much knowledge and effort is invested in the costly process of query construction. Accordingly, the idea of saving and reusing this investment is appealing. This applies to the expensive analytic process in general and is discussed on the process level with focus on implicit exploitation later in this thesis (see Section 5.2). However, additional exploitation can be also achieved explicitly on the query formulation level. As has been described in the previous section, Boolean query construction follows the construction of a searcher’s assumptions and knowledge in the domain of search. For specific domains and tasks best practices for designing a query might be available a priori, as part of an experienced searcher’s knowledge. On a very coarse level it might therefore be interesting to work with templates of patent queries as a starting point [Alberts et al., 2011]. A very direct way to exploit

previous queries is saving them to disk. This possibility is integrated in the query definition view of the PatViz query system.

Such templates might also be used to provide inexperienced users with a general plan on how to construct such a query. Depending on the field and task, certain metadata information, such as the IPC classes, can already be taken into account to offer a good start for an initial search. However, this approach has never been tested for real users, and it might also pose risks if applied in the wrong scenario.

In order to emphasize the template characteristics of a stored query, a method for integrating variables into predefined queries is available in PatViz. Variables are marked with a leading ‘?’ and shown with a yellow background in the visual representation of predefined queries. When executing a query containing such variables to the full-text back-end, the values that should be applied are requested explicitly before it is sent. This prevents unintentional execution of stored queries with inappropriate constraints in the structures important for the task while not limiting flexibility. A second benefit of the described mechanism can, of course, be drawn from it by the creator. If similar search objectives within a domain have to be addressed frequently, analysts can build up and maintain a predefined set of template queries to get started more quickly with new searches meeting these conditions.

However, template queries alone are an insufficient means to learn and comprehend how to search for and analyze patent documents iteratively. Much more can be gained from understanding the search process itself and the methods of controlled widening and narrowing of search queries in order to achieve the required recall while maintaining acceptable precision. Approaches for tracking and comprehending such analytic processes in PatViz as well as in other developments are discussed in Section 5.2.

3.1.3 Querying Metadata

Constraining the metadata (bibliographic data) of patent documents is another important aspect of patent searching. In technical domains, such as mechanical and electronic engineering, restrictions of the search domain are often made by constraining the query using the IPC, but can also be a good starting point for creating initial result sets with good precision (see Section 2.10 and [Alberts et al. \[2011\]](#)). The prototype system developed during the PatExpert project employs a relational database for providing metadata. Such databases can be accessed using SQL, a formal query language that meets the requirements of a relational algebra. For several reasons the expressiveness of the query language for accessing the relational database in PatViz was restricted to subset of operations and constraints.

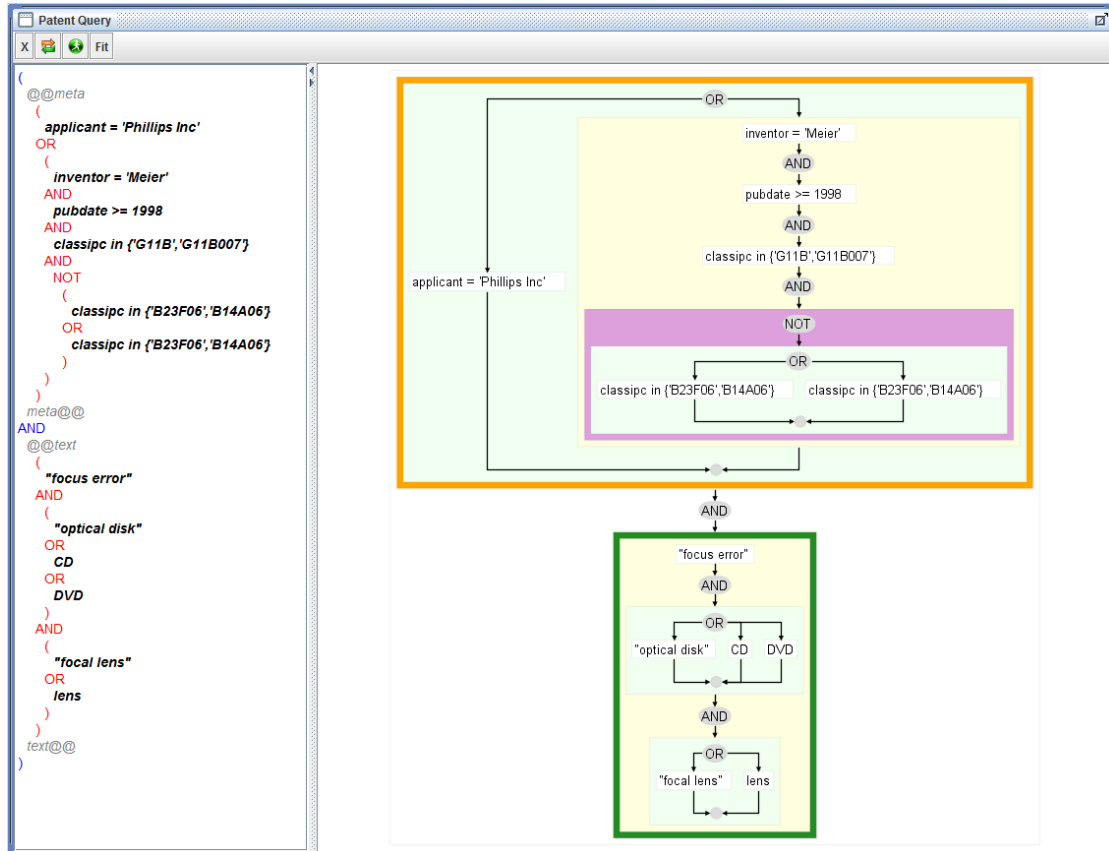


Figure 3.5 — Combination of a complex metadata query with a complex keyword query using the Boolean AND operator.

The most important restriction comprises the implicit *joins* on patent documents, meaning that always sets of patent documents are returned, even if the restriction applies to different entities associated with patent documents such as applicants. This is reasonable, since patent documents themselves present the primary target of analysis for patent researchers. There are not that many entities apart from the patent documents themselves, which makes an abstraction of the data as one plain table containing all entities' attributes, or at least a shallow hierarchical structure, possible. Relations between patent documents, such as citations, priority or family membership, can be propagated into visual perspectives directly. Since all other data sources provide only patent documents as results as well, the straightforward Boolean logic applied to combine different sources requires this implicit join strategy. Otherwise, explicit join mechanisms would be needed in the query language, at least for querying metadata entities of patent documents.

Another great benefit is that brushing & linking for patent-centric visual result set representations can be equipped with a much clearer interaction semantic, making document sets more easily comprehensible. Creating visual query languages for accessing databases are getting more common in certain domain specific approaches, but not yet for patent search interfaces. Here, formal query languages and form-based search interfaces are still the standard. The approach taken for querying metadata in PatViz can be seen as a pragmatic one, simplifying the model to a level that can be understood easily, while still maintaining a powerful query language. For constraining metadata aspects, a set of the most common attributes is offered to users in a context menu, as well as a set of appropriate relational operators. The value to constrain the attribute with has to be provided by the user. Figure 3.5 depicts the additional restriction of a keyword query in order to retrieve only documents that have either ‘Philips Inc’ as applicant, or that were invented by somebody with the name ‘Meier’, that were published in 1998 or later, and that are classified in a certain set of IPC classes but not co-classified in others, etc.

3.1.4 Image Queries

Alberts et al. [2011] report that descriptive images are often a crucial means for quick identification of the patent’s invention since they are not affected by patentese (see Section 2.10). Although patent images are frequently provided in bad quality and with handwritten comments, they constitute an important aid that allows patent analysts to screen documents more quickly. With the image similarity search developed by partners of the PatExpert consortium, analysts are not only able to evaluate the relevance in result set assessment, but can also request documents containing certain types and occurrences of images directly. The mechanism for image similarity search not only takes into account the images themselves but also the text describing them, which is available from the description section of patents. This additional textual information is preprocessed, important semantic aspects are extracted and stored in semantic web format RDF²⁴ (see 3.1.5). Details on the extraction process, the execution of image similarity search, and a retrieval framework for patent image search are available from Vrochidis et al. [2010].

In terms of image similarity search, a query by example variant is available in PatViz’s visual query building facility. Here, arbitrary images can be fed to the query interface either by identifying them within an available patent document or by providing a URL where the example image is stored. Image similarity is best applied in situations where recall should be improved by including it in existing queries through Boolean OR operation. However, this poses the risk of losing

²⁴<http://www.w3.org/RDF/>

precision. An adequate countermeasure is its restriction by classification as can be seen in Figure 3.2.

3.1.5 Semantic Queries

Extracting semantic concepts from textual material has the potential to reduce the effort put into search greatly. Actually, if it were possible to represent all aspects, their relations, and their categorization conveyed in a natural text in a unique, formal way, searching for them would be superfluous, since they could be accessed through automatic means directly, as long as the concepts are known to the searcher. The idea of the semantic web is based on introducing such computer-interpretable semantics [Berners-Lee et al., 2001], and a variety of mechanisms and standards for representing it formally have been developed and are maintained through the World Wide Web Consortium (W3C)²⁵.

Unfortunately, in practice it is not very likely to have natural texts annotated with such formal semantic information, because this would require either a human annotator introducing this information or an automatic procedure that attaches it. The first variant is very complex and time-consuming, while the latter is unreliable. However, certain aspects can be extracted and formally represented almost automatically, also for patent documents as has been shown by Potrich and Pianta [2008]; Cunningham et al. [2011]. Giereth [2012] proposes a semantic-web-based framework to represent patent information, which has been exploited in the PatExpert project in accordance to the project's main scientific objectives. In PatExpert, the extraction of semantic concepts and their relations is carried out as a preprocessing step on the available corpus of patent material. As a basis for this step, a variety of ontologies have been hand-crafted and, on a more detailed level, were created semi-automatically using the IPC and their description. In a final step, instances of concepts available in patent documents were extracted and aligned to these ontologies.

A rich semantic repository has thus been established for PatExpert aside from the already described data sources. As the name implies, semantic web data is organized in network structures, and, similar to relational databases, a formal query language (SPARQL²⁶) exists for retrieving semantic information from it. For PatViz's query creation mechanism, simplifications had to be made in order to conform to the Boolean paradigm and the restriction to retrieve only sets of patent documents. Therefore, the semantic query part allows for specifying constraints by defining the existence of concepts in the documents to be retrieved. For example, it is possible to define the restriction that a concept which occurs in a patent has a

²⁵<http://www.w3.org/standards/semanticweb/>

²⁶<http://www.w3.org/TR/rdf-sparql-query/>

specifiable relation to another concept in the same document. Or to put it more simply, one can define small patterns of a semantic graph that must be present in the documents to be retrieved. In order to tweak the query for better recall or higher precision, hyponyms²⁷ and hyperonyms²⁸ of the concepts required to be contained in a document can also be defined.

In the future the approach could be extended to let users define more complex semantic patterns within documents, because currently it is not possible to request chains of semantic patterns of arbitrary size. Again, the way it has been developed so far had its origin in pragmatic considerations: problems as well as solutions that are represented semantically and which might be of interest to a patent searcher can be requested and combined by using available Boolean operators. As mentioned before, the preprocessing steps needed to extract semantic concepts are far from perfect; as a consequence, semantic search can be an interesting addition for retrieving relevant documents at high precision and expanding queries beyond known keywords and metadata restrictions, but the other search mechanisms are still needed for assuring good recall.

3.1.6 Integrating Search Back-ends

Figure 3.6 depicts the integration of multiple search back-ends as well as their orchestration through a query broker service and the query builder in the PatViz front-end. The architecture employs a three-tier client-server approach, wherein the single tiers are connected asynchronously through web services. Such a separation is advantageous for a variety of reasons. Firstly, it decouples the front-end from the rest of the services, enabling multiple client systems to use the same back-end. Furthermore, local object oriented models can be established within each client that are necessary to feed and drive the visualization of the query as well as the result set visualizations after a query has been executed. This relieves the back-end services from the costly computational generation of views, which are only needed on the system's clients (see Section 5.1.1 for a detailed discussion of scalability issues). Secondly, updates of the data sources in terms of including new patent documents is possible without the need to apply changes on the client side as well. This reflects the common approach to how patent search systems are organized today. Thirdly, additional patent data sources can be added and integrated with manageable effort. However, integrating sources that provide new data types require changes on the client side. First, a new sub-query language has

²⁷ More concrete semantic concepts that could be summarized under a common semantic concept, e.g. 'optical lens' is a hyponym of the semantic class of 'lens'.

²⁸ Describes a more abstract semantic concept to given one and describes the relation opposite to a hyponym.

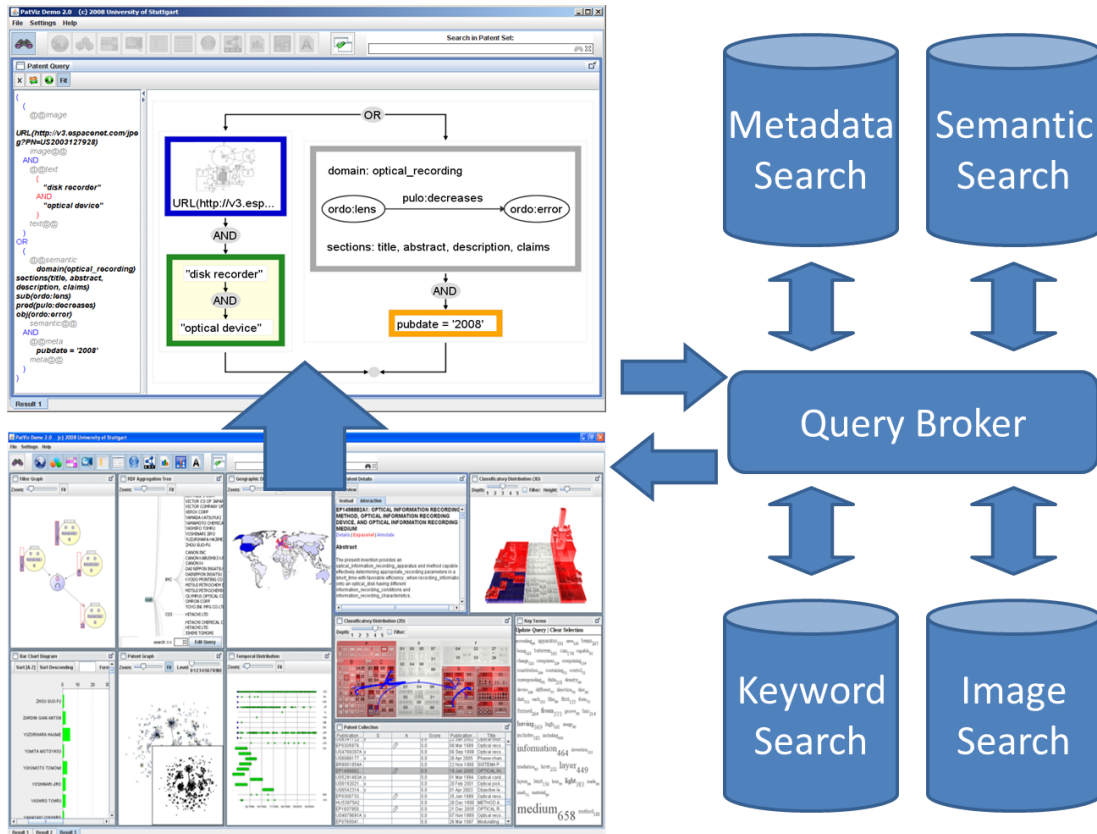


Figure 3.6 — General communication paths between front-end and back-end components. If a query is executed by a user, it is first sent to a broker mechanism that takes care of splitting it into its back-end specific parts. After this splitting step these parts are sent to the corresponding search services and all returned results are collected. Before transferring the results back to the client, all duplicates are removed by the broker. Subsequently the results can be inspected by users in the various result set views, suitable restrictions and constraints can be derived for improving the previous query to start another iteration.

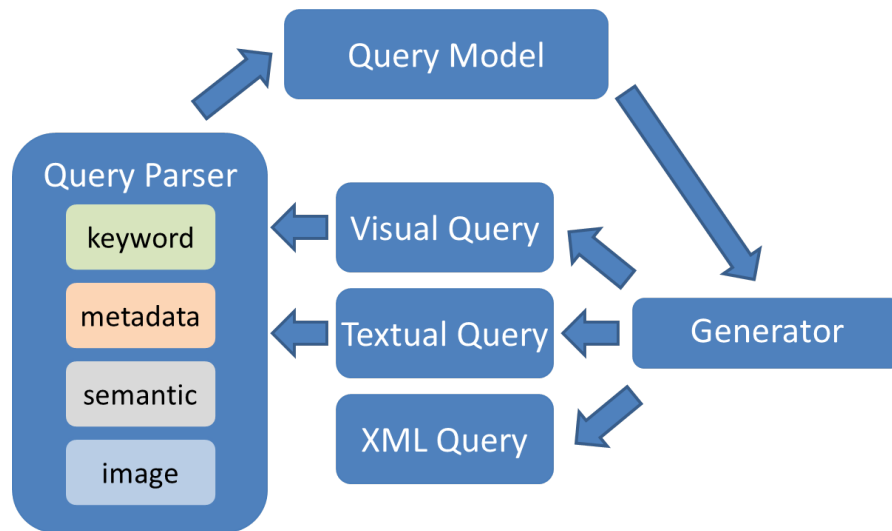


Figure 3.7 — Overview of the parser/generator framework for the query perspectives. The parser component translates visual and well-formed textual queries into an internal query model after changes are made by users. Based on the query model a generator component is responsible for creating visual and textual representations of the model to be shown in the views. If a query should be sent to the back-end services, the generator takes care of creating the corresponding XML query as well.

to be introduced. The inclusion of new sub query languages is accomplished by providing a corresponding grammar to create the language using a parser generator as shown in Figure 3.7. In the PatViz system this is realized through the javacc library²⁹. The creation of the visual counterpart, which must be linked to generated parse trees, has to be created manually. Additionally, the query broker has to be updated to accept query requests to the new data source and to route them accordingly.

The query broker depicted in Figure 3.6 plays an important role in the described setup. It takes query requests from clients, splits them into source-specific parts, possibly translating them into source-specific queries, and sends them to data sources. The answers from the data sources comprise the patent numbers of matching hits, and – in case of key word query, image similarity query, and semantic query – the position of the hits within each patent document. Afterwards, the broker collects the responses from the data sources, merges them, removes duplicates, and sends the results back to the PatViz front-end. PatViz maintains a sparse, object-oriented model of the returned patent documents. As soon as

²⁹<http://javacc.java.net/>

result sets are visualized for inspection by the patent analysts, metadata as well as additional content information are directly requested (on demand) from the repositories.

3.2 Interactive Search Result Visualization

A central idea of the described approach is the tight linkage between query (re)formulation and result representation. As mentioned above, one important aspect of applying the Boolean search paradigm in patent search is trust building through testing subsequent widening and narrowing operations against the analyst's information need. Visual result set representations are a good means to leverage the generation of insight on how complete and precise a result set is, at least if a patent searcher is knowledgeable in the corresponding technical field. In order to support users in evaluating and making sense of the results, it is beneficial to provide visual perspectives for all aspects that can be constrained during query building. Otherwise, the effects of changing a query, except for the change in the amount of returned results, are hard to understand and interpret. Accordingly, this section describes the group of visual components relevant to interactive query refinement tasks in the representation of the query's result set.

PatViz provides ten different views of the result set, which are shown in Figure 3.1. For their integration into the PatViz environment, all views must provide interfaces for basic brushing and linking operations. This means they must be capable to accept and react on highlight and selection events. Every brushing operation in one view results in the selection of a subset of patent documents. This subset is encapsulated in a selection event and broadcast through the PatViz environment. When other views receive such an event, they have to highlight the selection within their view correspondingly.

Care has to be taken in order to provide comprehensible and consistent selection semantics, despite different coordinated views being used within the system. Otherwise, the time needed to understand and use the system increases significantly, since interaction for each visual component has to be learned separately.

Consistency regarding selection has been realized in PatViz through two arrangements. Firstly, similar to the visual query builder, where constraints affect the restriction to patent documents, all selections possible within result set views lead to the selection of patent document sets. This means that performing the selection interaction on a bibliographic aspect is translated into the selection of all patent documents that fulfill this criteria. This can be seen as an interactive, visual pendant to faceted browsing or filtering but without removing the context as realized in traditional faceted browsing interfaces. Both filtering and faceted

browsing always affect the same types of objects (patent documents) in PatViz, while restrictions are defined based on the objects' properties. Internally, the selection mechanism not only collects the documents meeting a certain property characteristic, but rather stores them with the property restriction that determines the semantics of the selection operation as well. This can already be seen as an important feedback loop, enabling patent analysts to obtain an idea of a result set's properties and dependencies.

Storing the constraining aspects of a selection is not necessary for realizing brushing & linking in the multiple coordinated view display, but it constitutes an important base for advanced filtering or selection management and for realizing important feedback loops as described in Section 2.10. Secondly, multi-selection is only possible within single views, meaning that applying several constraining criteria from different views in parallel is prevented. The reason for this is again semantic clarity. As soon as multi-selection should be realized, multiple interesting questions arise: how are different selection restrictions combined – through Boolean OR-logic, or through Boolean AND-logic? And how are secondary selection and highlighting effects handled? Even if the combination of multiple selections from different views could be configured explicitly to a specific set operator, it would be quite difficult to reconstruct and comprehend such operations after some steps. As a consequence, multi-selection with clear semantics is only allowed within one view; nevertheless, additional means for explicit, constructive selection management were added as discussed in 3.3. When a selection interaction is triggered in another view, all former selections and highlights are dropped.

Even within one view, similar questions regarding multi-selection have to be addressed. Therefore, all multi-selection operations in result set views in PatViz are per default Boolean AND-combinations. Color-coding for selected and highlighted patent documents, or rather their aspects and facets, is consistent as well throughout the views of the system. The selected aspect is displayed in yellow, all affected, or respectively highlighted, facets are depicted in red (see for example Figure 3.8). This secondary highlighting effect can also affect the visualization where the initial selection is made as, for example, described in the next section for the map view.

Most of the views that are subsequently presented as part of the PatViz interface for result set analysis employ information visualization techniques, thereby deviating from typical patent search interfaces. Their combination, as well as the means for interaction realized on them, have been tailored to integrate them in a sophisticated visual analytics system that supports exploration and refinement of these sets, and that goes beyond current state of the art techniques.

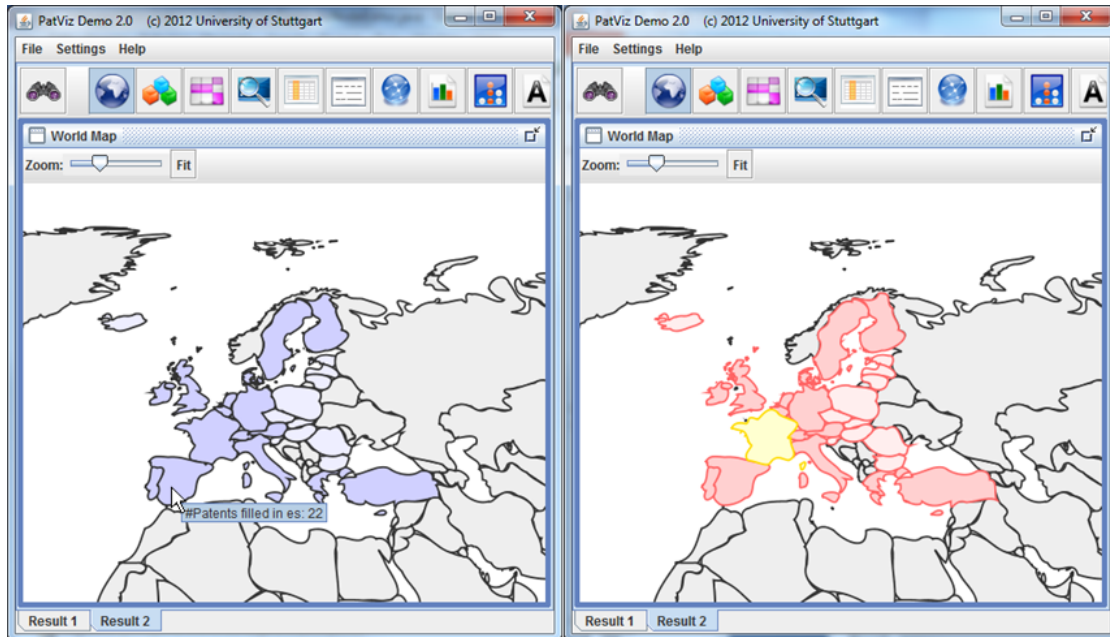


Figure 3.8 — The designated states of a patent document set. Color saturation depicts the number of documents that include the country as designated state. Spain is hovered by the user in the left case. The right view depicts the selection of all patents that have been applied for in France, resulting in secondary highlighting of all other countries these patents list as designated states.

3.2.1 World Map

Figure 3.8 depicts a map showing the countries where the patents in the result set are in force. The color saturation of the countries indicates the number of valid patent documents. Details on how many of them are in force are displayed if the user hovers the mouse pointer over a specific country. The patents of a country can be selected by clicking on it. As a result the selected country is displayed in yellow, and all facets that are linked to the selected documents in other views are shown in red. This may also include other countries in the same view, where selected patent documents are in force as well. Zoom and pan interaction is available to enable users to see more details and to concentrate on countries and regions they are particularly interested in. Information on the patents' designated states can become an important aspect during strategic tasks, e.g., finding cooperation partners, licensing, etc.

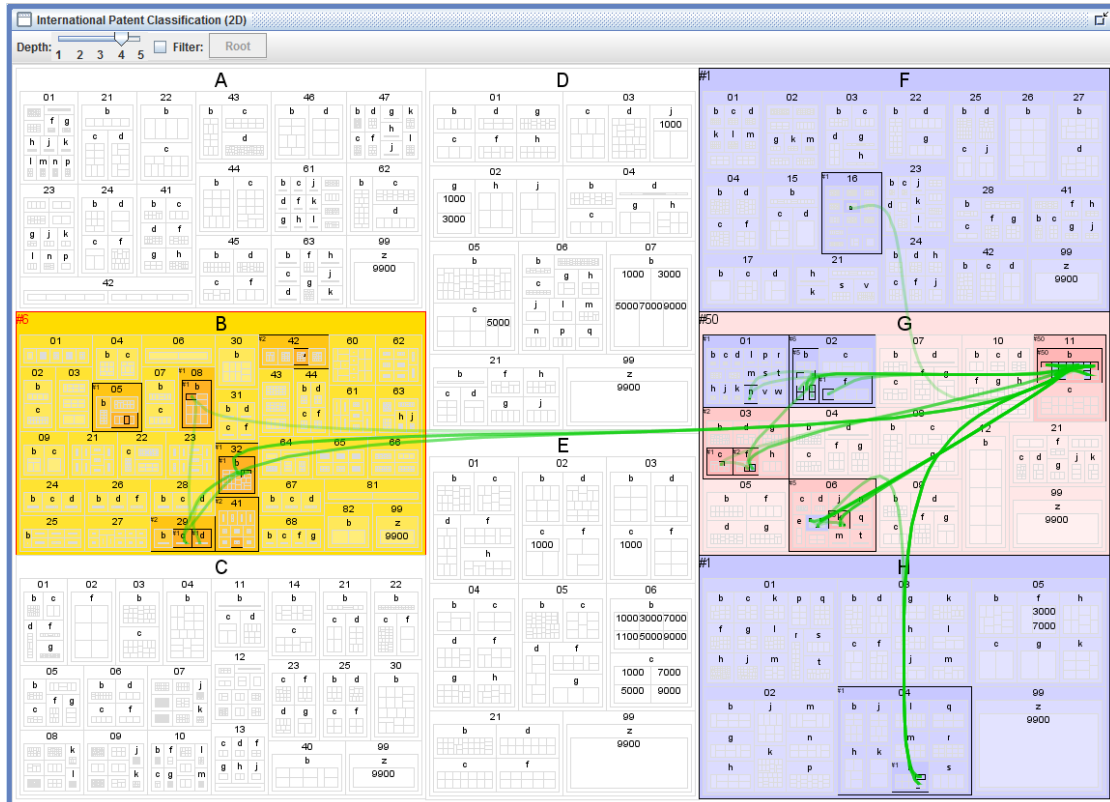


Figure 3.9 — The IPC view showing a patent result set at the main group level. Areas containing patent results are depicted in blue. Section ‘B’ has been selected (yellow) and as a consequence parts of section ‘G’ were highlighted (red). This effect results from patents co-classified in both sections as is indicated by the green co-classification edges.

3.2.2 IPC views

The integration of the treemap views in PatViz strongly builds on the work of Giereth et al. [2007b]; Giereth [2012], who employed this technique for representing patent classification information. Figure 3.9 shows the IPC (see Section 2.8) in the form of an ordered treemap as proposed by Shneiderman and Wattenberg [2001]. Patents in a result set are shown by coloring the corresponding treemap areas starting from the sections level in blue. Saturation of the map areas, again relates to the number of patent documents that are classified to be within them. Using color to map the number of result documents instead of an area’s size as is often done with treemaps has two reasons. Since the IPC holds quite a huge number of categories, it seemed beneficial to show the whole context of the IPC

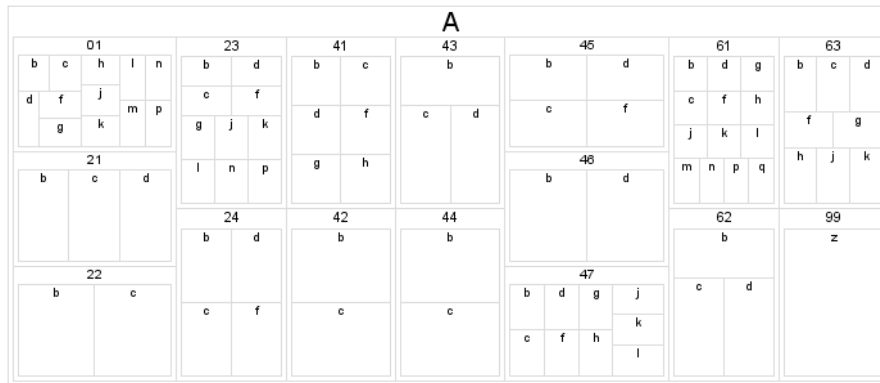


Figure 3.10 — The main section A of the IPC, depicted down to the subclass level. All nodes that are not leaf nodes have a border showing the corresponding IPC label as described in Section 2.8. The border also enables users to interact with such intermediate nodes by clicking on it.

in one overview perspective and a treemap is a perfect means to provide such an overview. The IPC view equips users with the possibility to browse and explore IPC categories, even if no result document has been classified in such an area.

Furthermore, labels for sections, classes, sub classes, and groups were introduced to provide users with orientation within the classification system and to facilitate interaction, e.g., clicking on a area of the treemap. This would not be feasible without sufficiently wide borders between parent and child areas in the treemap as can be seen in Figure 3.10. Maintaining borders intentionally, however, might lead to wrong interpretation of areas where no or just a few patents are classified, which forbids the mapping of patent number to treemap area size. Another consideration that led to this decision is the easy identification of source and target area of co-classification links as subsequently described.

In order to emphasize the exact class, subclass, or group membership of retrieved patent documents, the borders of parent structures of the IPC are colored and the most detailed category is filled. With this coloring method, it is possible to provide an overview of the coarser levels containing patents as well as showing specific details on the fine-grained levels. In these fine-grained levels, map areas where patents are categorized in are emphasized stronger, while sibling areas having no patents categorized in are depicted with a less saturated color. This means, membership in the IPC structure is propagated and accumulated along the path to the parent structures stopping at the IPC section level. Selection interaction is possible at each level by clicking the corresponding area and results in the restriction of the document set to all patents classified in the corresponding

sub-hierarchy of the IPC. The selection semantics can therefore be described as ‘select all patent documents that are classified in the clicked IPC node, or any of its child classifications’ down to the group level. Again, the color scheme for selection interaction is applied showing selected IPC parts in yellow and highlighted ones in red (see Figure 3.9).

The IPC view provides a slider in order to show more or fewer details in the classification view, which can be used by patent analysts to determine the granularity of the IPC level they are interested in. If set to 1, only IPC sections are shown. Level 2 and 3 additionally show classes and subclasses respectively. Level 4 adds main groups in combination with patent co-classification, whereas level 5 also takes into account IPC sub groups.

The co-classification information, shown as green edges between groups in the detail levels 4 and 5, enables analysts to understand if patents are categorized in multiple different IPC groups. Since connecting IPC groups directly with straight edges would cause visual clutter and would also make it more difficult for analysts to identify strong co-classification relations in the result set, hierarchical edge bundling is applied, as has been proposed by [Holten \[2006\]](#). Here, the IPC’s hierarchical structure is exploited for routing the edges by using the centers of parent areas within the IPC as additional points on the path. Through the employment of splines instead of drawing straight lines, links that follow similar directions within the hierarchy share parts of their path. As a consequence, fewer edge-crossings occur, and more space of the underlying treemap is kept clear of edges, providing patent analysts with a better overview of a result set’s co-classification characteristics. If multiple edges are drawn along the same path, they are visually aggregated using alpha blending which leads to higher saturation of the affected links, signaling that relations with high saturation are more common in a result set than ones with low saturation. However, the advantage of providing a good overview is decreased by the loss of details through the bundling process. A focus+context interaction mechanism to overcome this drawback and which conforms to the properties of visual analytics enhanced workflow itself, is discussed in Chapter 4, Section 4.1.

In addition to the described interaction possibilities, structural zooming, which can be seen as a focus+context technique, has been introduced. If users double-click an area of the IPC, the area is enlarged showing more detailed levels as before, while other areas are reduced accordingly, showing fewer details. This form of interaction is possible on each level of the IPC and always affects only the sibling areas of the zoomed one. Figure 3.11 depicts the IPC treemap after IPC section ‘B’ has been structurally zoomed. The reverse zoom-out interaction is achieved through another double-click on the same area. A similar approach for interacting with treemaps is described by [Blanch and Lecolinet \[2007\]](#).

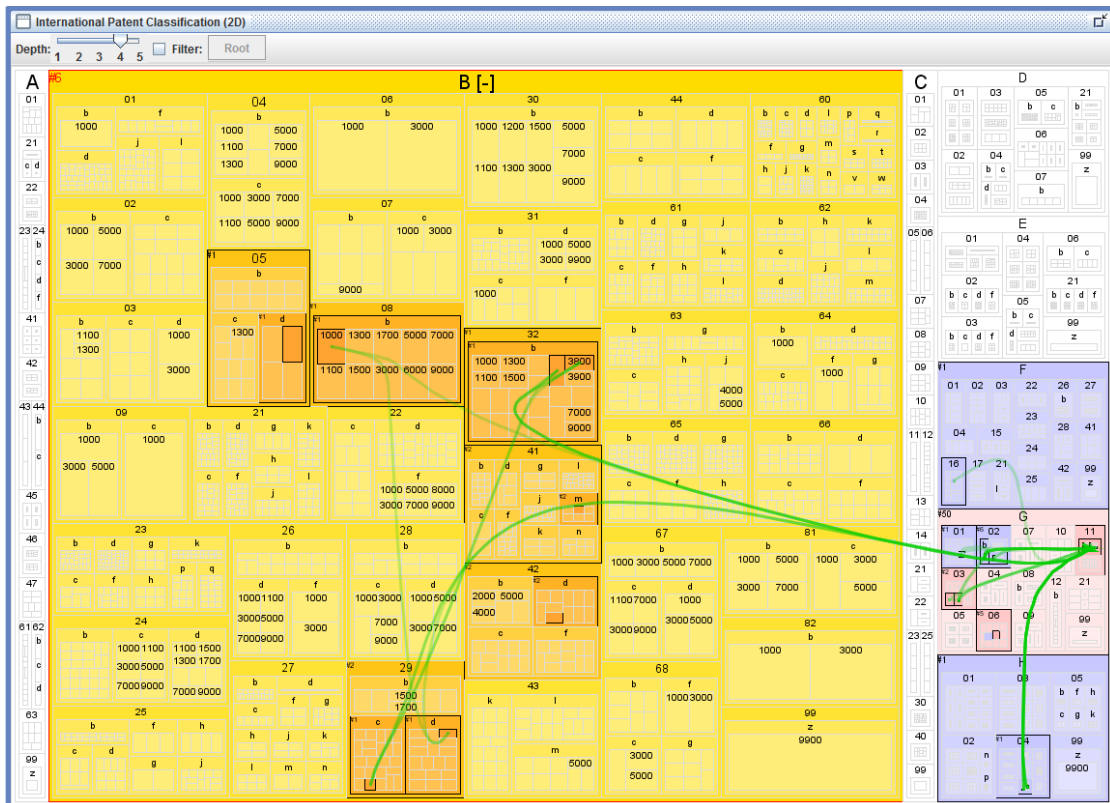


Figure 3.11 — Here the same situation as in Figure 3.9 is displayed after a structural zooming operation has been applied to section ‘B’.

The third view in the lower row of Figure 3.1 shows a 3-dimensional version of the IPC treemap. This is an experimental view that aims at depicting additional properties, e.g., the number of documents classified within an area, on the introduced third dimension of the treemap. As mentioned above, area size is not used to represent properties of the patent result set, a disadvantage that can be diminished by introducing a third dimension. However, showing additional details using 3-dimensional visualization comes at the cost of occlusion problems, which then have to be addressed through an increased complexity in interaction mechanisms, such as pan, tilt, and visual zoom. Details about this concept can be found in [Giereth et al. \[2008a\]](#).

The categorization of patent documents according to classifications systems such as the IPC (Section 2.8) plays an important role in the construction of queries, especially for broadening and increasing recall. Besides showing IPC classes in which patents of the result set are classified, explicit co-classification information is

Publication Numb..	Title	Publication Date
EP1734517A2	Disk recording medium, disk drive apparatus, reproducing method, and disk manufacturing method	20 Dec 2006
EP1628291A2	Method for manufacturing optical disk media of high-to-low and low-to-high reflectance types	22 Feb 2006
EP1617424A1	DIFFRACTION OPTICAL DEVICE AND OPTICAL PICKUP DEVICE USING THE SAME	18 Jan 2006
EP1583094A2	Power saving system for optical disc recording/reproducing apparatus	05 Oct 2005
EP1536422A2	Optical disc and method for recording additional information to an optical disc	01 Jun 2005
EP1521247A2	Optical pick-up device and optical information recording reproducing apparatus	06 Apr 2005
EP1515321A1	OPTICAL RECORDING MEDIUM	16 Mar 2005
EP1465171A2	Optical pickup device	06 Oct 2004
EP1434210A2	Optical recording m	30 Jun 2004
EP1414034A2	Optical disk, optical	28 Apr 2004
EP1361569A2	Optical information	12 Nov 2003
EP1357546A1	OPTICAL DISK, REC	29 Oct 2003
EP1333431A1	OPTICAL DISC	06 Aug 2003
EP1213709A1	Focus control selec	12 Jun 2002
EP1195755A2	Optical recording m	10 Apr 2002
EP1168315A2	Optical recording m	02 Jan 2002
EP1136991A2	Slider moving devic	26 Sep 2001
EP1124227A2	Optical pickup, tilt d	16 Aug 2001
EP1102258A2	Controller for data r	23 May 2001
EP0992988A2	Optical disk apparatus	12 Apr 2000
EP0986055A1	Information recording apparatus	15 Mar 2000
EP0911818A1	Compact dual wavelength optical pickup head	28 Apr 1999

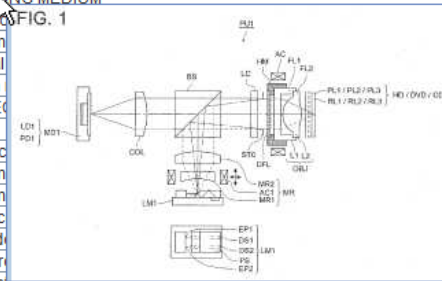


Figure 3.12 — List view showing a patent result set that has been sorted by publication date. For entries that are hovered over, the representative image contained in a patent’s abstract is shown.

provided. Such co-classification is suitable to create additional insight about which IPC classes are promising candidates to take into account during subsequent query refinements as well.

3.2.3 Patent List

Figure 3.12, showing a table of patents contained in the result set, comes closest to result set views typically presented by available patent search systems such as Esp@cenet³⁰. It shows several fields such as the patents’ publication numbers, their titles, a score for patent ranking if available from the search context, their publication dates, and the previous selection operation. The last field informs about the availability of semantic annotations extracted through preprocessing steps. The availability of such a table view is important, because it helps immensely in making the transition from rather text-based interfaces to interactive, visual ones easier. Even if patent practitioners are unexperienced in visual patent analysis, it guarantees that they can immediately use the system.

The table can be sorted by each field through clicking the corresponding column header. Sort order can be inverted by clicking again. If the mouse is hovered over one patent entry the representative image contained in a patent’s abstract

³⁰ <http://www.epo.org/searching/free/espacenet.html>

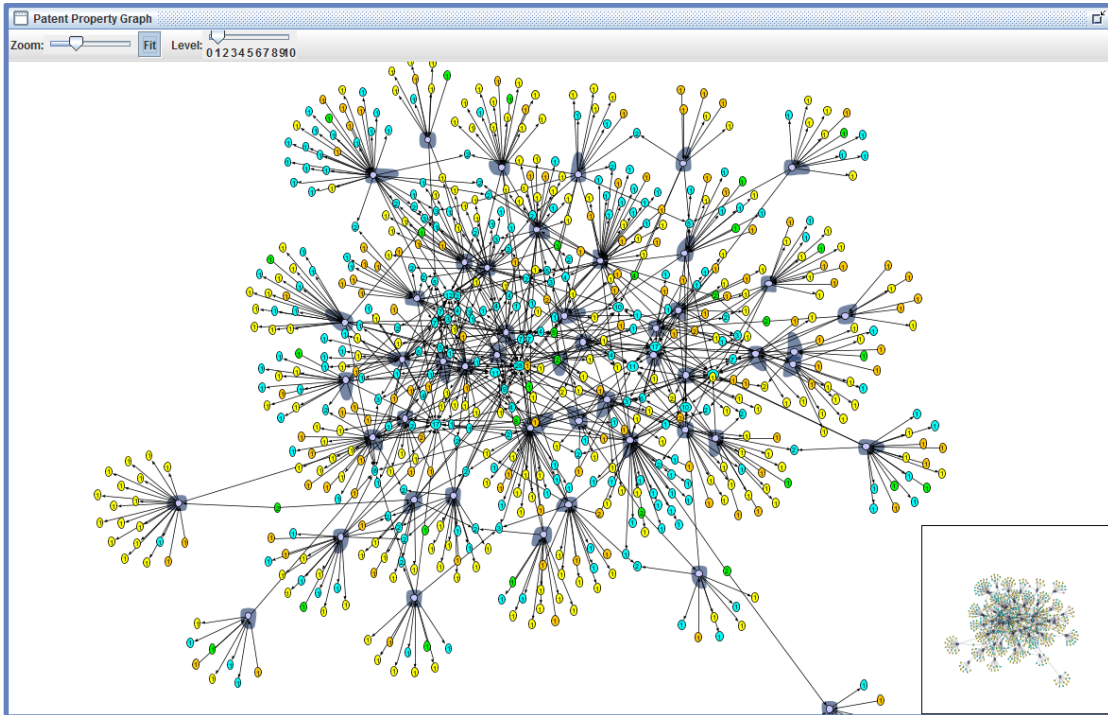


Figure 3.13 — The patent graph view depicts patents in the result sets as well as their relations and attributes. Depending on the chosen structural zoom level only relations connecting more or less attributes and patent documents are shown.

is displayed in order to disclose more details on the corresponding invention for quick result set browsing. The table also serves as the primary means to trigger the detail views, which can be accessed by double-clicking a specific entry. Multiple selection is possible like in other list interfaces with the modifier keys ‘ctrl’ for adding single documents to the selection and ‘shift’ for range selection. Next to the table, only two other views allow for the direct selection of a single patent document – the patent property graph, and the priority-time plot. The selection mechanism for this view therefore slightly differs from the other views, because selection is explicit, meaning that specific patent documents are selected directly through their patent numbers and not implicitly via a restriction of some content-related or metadata property. From a logical perspective this is consistent with the overall approach.

3.2.4 Patent Property Graph

Figure 3.13 shows a graph of the patent result set. While most of the other views depict certain facets of the patent set to be explored, this view provides a good overview of implicit connections between shared properties in the result set, which to some extent reduces the negative effects of coordinated views as opposed to integrated visualization. However, the usefulness of this view highly depends on the amount of patents to be displayed. Even for relatively small result sets, the number of relations and links can quickly exceed a problematic level resulting in heavy clutter caused by too much interconnectivity and a multitude of nodes. In such cases, reduction of the details is advisable. Similar to the IPC treemap, properties can be hidden or added through a slider that enables users to choose from different levels of detail. With all details activated, a graph of the currently available patent set with property relations is shown. These properties include applicants, inventors, cited patent documents, and IPC subgroups. Patents themselves are shown in dark blue. In addition, members of the same patent family are indicated by showing a colored convex hull around them.

The graph's layout is generated employing a n-body force simulation [Barnes and Hut, 1986] provided by the prefuse visualization toolkit³¹ [Heer et al., 2005]. A focus+context method for structural zooming is available, which makes more details of a clicked node visible, i.e., the node itself and all directly connected nodes are zoomed in and the labels of the nodes are expanded as shown in Figure 3.14. At the same time, the rest of the graph is greyed out and reduced in size. The described interaction also selects the node of interest. Depending on the node type, patent documents are either selected directly or restricted for patent sets analogously to the procedure described for other views above. Multi-selection is not allowed in this view, since it is possible to restrict different properties during selection, which does not comply with the general selection model (restrict only one facet) of the PatViz interface. In order to support users during selection interactions, the properties related to patents are labeled according their connectivity in zoomed out perspective. This means that properties which are shared amongst patent nodes will be labeled accordingly, e.g., an applicant holding 4 patents in the result set will be labeled with a 4 and exhibit 4 links to these patents. Such an indicator supports patent analysts in choosing promising nodes before the interaction itself is performed, because they can directly assess the selectivity of a subsequent selection operation. Zooming and panning interactions are available to give users the possibility to explore details of the graph. In order support users with a better overview during high zoom levels, an additional mini map of the graph is provided in the lower right-hand corner of the view.

³¹<http://www.prefuse.org/>

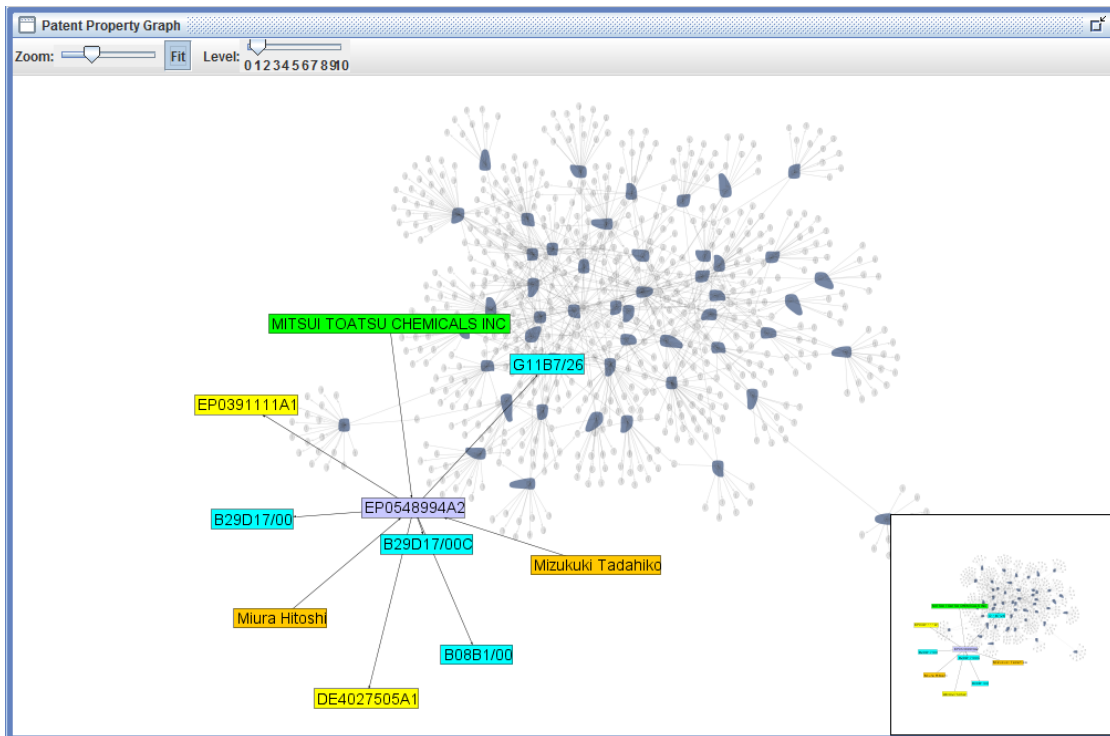


Figure 3.14 — Structural zooming operation applied to a European patent in the property graph view. Related attributes such as inventors (orange), applicant (green), IPC subgroups (light blue), and cited patent documents (yellow) are zoomed, while the rest of the graph is greyed out.

The patent property graph serves as an auxiliary view depicting patents and their properties in a common context. Many of the selection operations that are possible in this view can also be achieved with other perspectives.

3.2.5 Priority-Time View

The priority-time view (Figure 3.15) depicts patents and their corresponding priorities (see Section 2.8). On the x-axis time is shown: all patents and priorities are placed on it according to their application date. This view can optionally be switched to the publication date. In y-direction patents and priorities are ordered according to the country or regional office where the patent has been applied for.

If a patent document is hovered over, all its priorities are indicated through directed edges as depicted in Figure 3.15. Clicking on a country label expands the view in

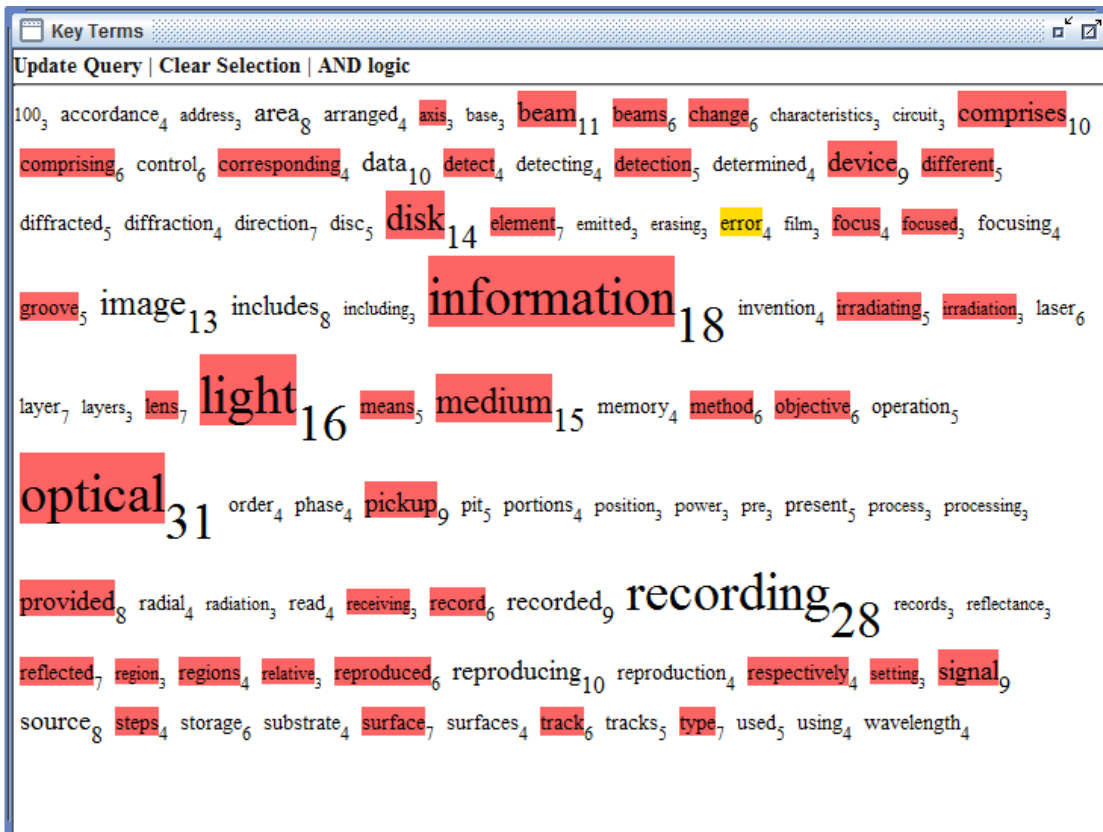


Figure 3.16 — Term cloud of a smaller patent result set. Label size and the attached, subscript numbers relate to the terms’ document frequency in the result set. The term ‘error’ has been selected in the shown situation and others are highlighted as a consequence of the selection, since they occur in one of the documents that contain this term.

measures for importance are applied. To underline the fact that terms contained in patents are provided in this scenario, the name ‘term cloud’ seemed to be more appropriate.

As provided in the PatViz interface, the terms’ local³² document frequency df (see Section 2.5.2) is mapped to importance and accordingly to the font size in which they are shown. Additionally, local document frequency is denoted explicitly with every shown term. The choice to use document frequency as the measure for importance was made for two reasons. Similar to the property graph view, df indicates the number of affected documents if a selection interaction is triggered.

³²Local in the sense of being computed on the loaded result set documents.

Another good measure could be the *tf-idf* scores of the result set's contained terms, which, however, would be only meaningful if the whole patent collection could be taken into account for *idf* computation (see, again, Section 2.5.2 on this topic). As a result the most valuable terms for discriminating the result patent set to the overall patent set could be extracted. However, this would have been a rather expensive operation as tests showed, because every term contained in the patent documents result set would have required a web service lookup to the repository, in order to retrieve document frequencies and the separate computation of each term's *tf-idf* value and their sorting. Thus, this option was discarded due to PatExpert's system design, but nevertheless it constitutes an interesting means to depict important terms for systems with tightly integrated back-ends.

In order to maintain a manageable number of patent terms, only the top 100 terms with highest local *df* are shown in the map. Unlike other views, the term map provides two modes for multi-selection: one for applying Boolean OR logic and another one for AND logic. The decision to allow two modes was again a pragmatic one, since this view was considered to be very important for query refinement. Selections made in this view are regularly reintegrated into previous queries, which motivated the introduction of an additional 'update query' button, to simplify reintegration.

A variety of layouts for tag clouds have been suggested and tested in related work, e.g., in Lohmann et al. [2009]. In the PatViz implementation an alphabetical ordering was chosen, since this was the most obvious method that enables analysts to scan for a specific term in the cloud.

3.2.7 Legal Entity Chart

As already mentioned in Section 2.8, legal entities, such as the applicants and inventors of a patent, constitute another interesting facet that can be exploited during patent search and analysis. The legal entity chart (Figure 3.17) enables analysts to explore these facets. Patent applicants and inventors are depicted in the form of a bar chart, showing the names of legal entities on the y-axis and the numbers of the current result set's patents related to them on the x-axis. The type of legal entity can be set by users as required. The chart can be sorted by the number of patents where a legal entity is involved in increasing and decreasing order. This way, users can quickly identify important players in the patent set under analysis. Additionally, it is possible to sort the chart according to legal entities' names alphabetically, which is especially helpful when specific legal entities are to be identified in the set.

Selection interaction is possible in several ways. Clicking on a bar or the name of a legal entity is the simplest way to accomplish it. Accordingly, the result set

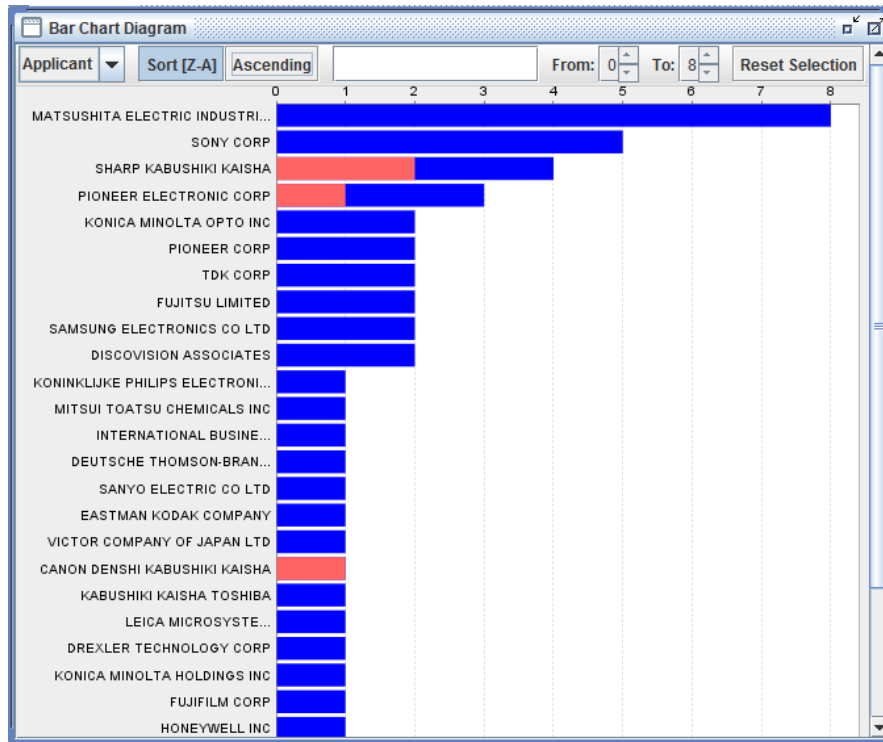


Figure 3.17 — Bar chart depicting the legal entities associated with a patent document such as applicants and inventors. In the depicted case applicants are shown and the chart was sorted by the number of patents they applied for. To handle large numbers of legal entities, a text field for reducing the shown applicants to those containing a user-specified substring is provided. Additionally, the shown entities can be restricted to those associated with a user-specified number of patent documents. In the depicted scenario fractions of the patents are highlighted (red) as a result of a selection in a different view.

is restricted to the chosen legal entity, and affected aspects in other views are highlighted. Selection of multiple legal entities is possible through brushing over multiple bars, by clicking several of them subsequently while pressing ‘shift’, or by applying a filter regarding the minimum and maximum number of patents a legal entity should be connected with.

Highlighting as a consequence of selections made in other views, may result in highlighting only a fraction of corresponding bars in the chart. It might also occur as a side-effect of selection performed in the chart itself, if more than one legal entity is registered in a patent’s metadata.

3.2.8 Detail Views

There are two views that allow for the inspection of a single patent’s content. While the first one depicts the whole patent document with its most important metadata aspects, as is common with most patent search interfaces, the interactive view additionally shows extracted semantic annotations. Apart from the possibility to explore details of these annotations by hovering the mouse pointer over them, a focus+context method can be applied to help users concentrate on those parts of the document where semantic concepts and relations can be found. A detailed description of the interaction technique is given in [Giereth et al. \[2008b\]](#). The interactive view can also be used to select important semantic concepts for integrating them into previous query statements.

3.2.9 Selection Management

Simply selecting a subset of documents, as facilitated through the interaction mechanisms of the views described in the previous section, is not expressive enough to reflect interesting subset definitions, which can then be used as part of advanced filtering, as a means to test hypotheses on, or to be exploited in a later step to improve the search query. Therefore, the views also need to maintain a description of the filter operation used to create the subset. In the example of the world map, this could be ‘filing-country = Sweden’ if users mark Sweden, or a concatenation of such statements if selecting a multitude of countries. The goal of this approach is to derive and preserve users’ intention from their interaction instead of applying direct selection. With the increased self-descriptiveness of selections it is possible to enrich the selection management with interactive adjustments of the underlying filter mechanisms, as well as to create appropriate filter definitions for the search query reformulation. However, if the selection operation should extend on separated sub-selections in multiple linked views of aggregated data, the selection of a particular data set may be difficult or even impossible, for the reasons discussed in the first

paragraphs of Section 3.2. Therefore, PatViz contains a graph-based technique for visual selection management allowing for the combination of data subsets by applying set operations on them. This technique provides increased expressiveness over classical approaches by utilizing them as building blocks for more complex extraction strategies.

The method itself employs a node-link-based graph view (see Figure 3.18) that provides nodes serving as interactive widgets. The directed graph, which is built in a user-steered manner, comprises three different types of nodes: *content nodes*, *filter nodes*, and *operator nodes*. In its initial state the technique displays a single (root) content node representing the entire set of patent documents contained in the current result set. Content nodes have a vertical bar attached to them symbolizing the size of the set they represent in relation to the whole set. Additionally, the bar is labeled with the exact size of a (sub)set. One method to create new nodes is given through a context menu, enabling analysts to represent an arbitrary selection made in one of the result set views as a new content node in the graph. As soon as such a node is added, it is automatically attached to the root node representing the whole set of available patents via a filter node which describes the selection restrictions made in the view. The other way to build new nodes is through direct interaction on existing graph nodes. Content nodes can be connected to filter nodes, which constrain one of the set's attributes, in order to restrict a content node's set of documents. The result of the restriction is another content node with the reduced document set. Set operator nodes constitute the third type of nodes. These nodes allow for the combination of different content nodes and thus can have an arbitrary number of incoming edges and one outgoing edge, each again connected to corresponding content nodes. In order to exploit the visual selection management successfully, typically all mechanisms have to be combined.

Set operators, i.e., union, intersection, and symmetric difference can be set explicitly via additional nodes, apart from the implicit combination contained in the graph structure itself, i.e., sequences (AND) and branches (OR). In contrast to the filter/flow metaphor described by Shneiderman [1994] these operators facilitate the combination of arbitrary sets of data objects without the need to generate multiple instances of a particular filter just to apply it in different combinations. DataMeadow [Elmqvist et al., 2008] describes a related, also network-based approach, to combine filters on multidimensional data, whereby different aspects can be filtered at once using interactive visual metaphors called DataRoses.

The construction of the graph itself is performed completely by the users. Guiding users when they interact with graph widgets prevents the occurrence of illegal graph configurations. For example, if a content node is dragged by a user it can be only attached to set operator nodes. During the drag interaction valid targets are highlighted in green to help users in identifying valid connection points in the graph.

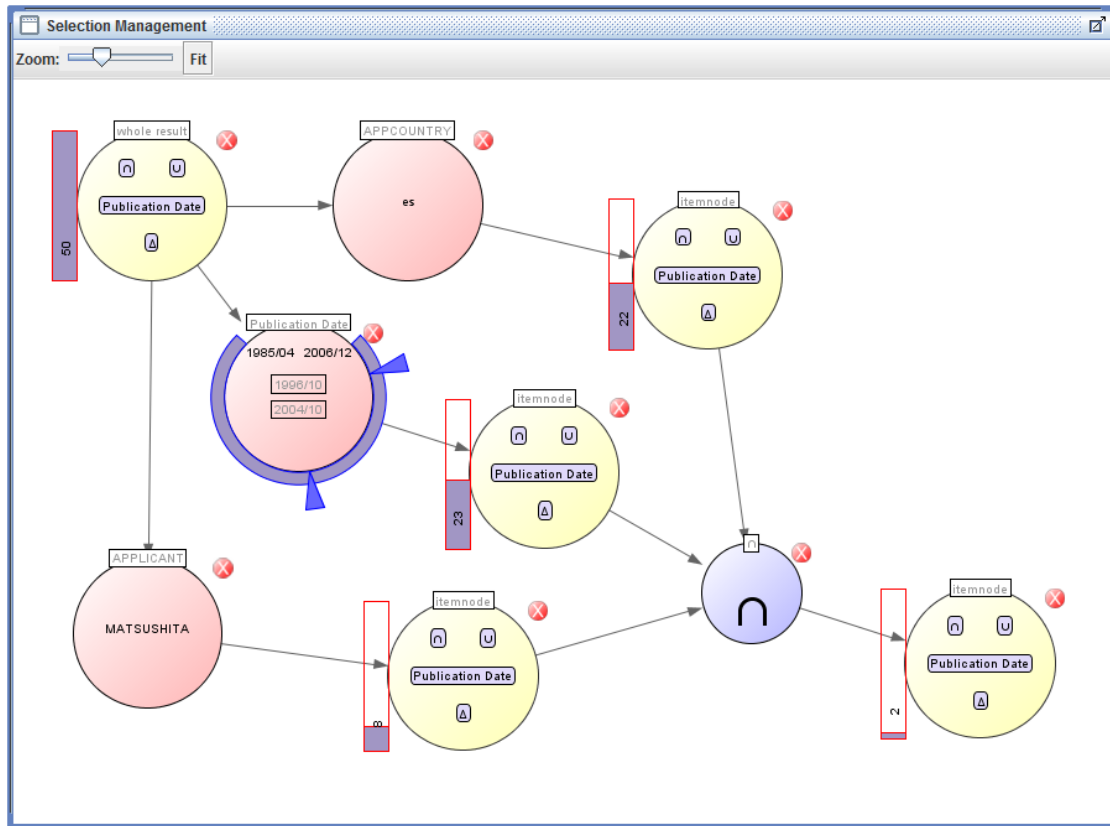


Figure 3.18 — The selection management view. In the depicted situation three content nodes (yellow) were derived from the available results set of 50 patent documents, which is represented by the upper left content node. These subsets are connected via filter nodes shown in red. The subsets were created (top to bottom) by selecting Spain in the world view and representing this selection in the management view, by directly applying a filter that restricts the publication date to the time period between April 1985 and December 2006, and by selecting all patents having ‘Matsushita’ as an applicant. Finally, all filtered/selected sets are combined through an intersection operation (blue node). With the bars (attached to a content node) indicating number of patent documents associated with a content node, it becomes quickly obvious how restrictive previous filtering operations are.

Different filter nodes are created with respect to the data type of the property that should be constrained. After combining different sets and parameterizing filter nodes, every document subset can be reflected back on the result set visualization by selecting an arbitrary content node. In this way, the visual selection management expands result set exploration facilities with an important, additional feedback loop.

Visual selection management also serves another important purpose. It is a means for extending user working memory, by providing a workspace to externalize findings. Pirolli and Card state that

“Techniques aimed at expanding the working memory capacity of analysts by offloading information patterns onto external memory (e.g., visual displays) may ameliorate [the problem that only a limited number of hypotheses, amount of insight, and found relations can be heeded by humans at the same time]”.

Selection management facility acknowledges the limits of humans’ attention span, and supports patent searchers to include and consider previous findings as well as their integration at any time in their analytic workflow.

It is possible to highlight document sets represented by content nodes in the selection management technique in the result set views using the nodes’ context menu. While this still makes the testing of hypotheses – e.g., ‘does a certain applicant dominate others with respect to patented solutions in a specific domain within a certain country’ – possible, comparing different sets can only be performed in a sequential manner. For the latter case it would be necessary to either duplicate views, to compare different result sets side by side, or to extend the available ones to show two or more subsets in parallel, emphasizing their differences and their common characteristics. Both solutions, however, would increase the complexity of the visual interface considerably and would conflict with the idea of a single, consistent, centralized selection mechanism.

Since the selection management technique does not depend on the type of documents in the sets, and because filter options are derived from the underlying data model, the selection management and its insight integration facility can be applied to other application domains without great difficulty. In cooperation with others this has been successfully demonstrated in different VAST-Challenge³³ submissions [Bosch et al., 2009, 2011; Krüger et al., 2012].

In 2011, the filtering and selection management technique was employed in the context of analyzing microblogging messages in order to gain situational awareness

³³<http://hcil.cs.umd.edu/localphp/hcil/vast/archive/index.php>

of a fictional epidemic outbreak, its means of transmission, and its cause [Bosch et al., 2011]. In combination with other coordinated views, the method turned out to be especially helpful to test a variety of hypotheses, and was significant for finding a meaningful interpretation of the given scenario. In 2012, the same technique was exploited as part of a toolkit for analyzing a large computer network [Krüger et al., 2012] and helped exclude certain events that might have caused some of the problems encountered affecting the network.

Besides having an explicit means for insight management, the mechanism for selection management also describes parts of the analytic process itself, including aspects such as invalidated hypotheses and other analytic steps that probably cannot be used further, but are still important for increasing an analyst's trust in the validity of the analysis. Together with the changes applied to previous queries, analytic processes are formally represented in the system without any explicit recording triggered by the user. However, users can provide descriptive information to made selections and combinations of them in order to identify and remember specific analytic findings more easily.

One distinctive aspect of the approach described above is its user-directed constructive nature. The formalized user-steered construction of an analysis and its explicit representation creates potential for exploiting a variety of synergetic effects. It is, for example, a suitable base for representing analytic provenance, which can be exploited in later steps to support collaboration and analytic reporting. Those aspects are discussed in detail in Chapter 5.

3.3 Feedback Loops and Insight Reintegration

Facilitating views and interaction methods to extract insights from visual perspectives is one key aspect of visual analytics. But without providing additional means for exploiting these insights successfully afterwards, e.g, during subsequent stages, analysts are forced either to keep these insights in mind, or to externalize them by recording them manually or by exporting them if possible. Support for interactive feedback loops increases analytic possibilities. It is therefore desirable to make insights exploitable directly within a visual analytics framework to provide seamless integration for larger sensemaking and feedback loops.

For this purpose, several levels for insight integration have been realized in the PatViz system. As discussed in Section 2.10, reading documents in the patent domain tends to be rather laborious, but the documents provide a variety of metadata that can be refined into aggregations, relations, and statistics. Thus, it is possible to create a rich set of views on result sets depicting metadata as well as summarized content. Without integration, the interactions provided by

an individual view are restricted to the adjustment of view-dependent parameters like sorting, filtering, highlighting, zooming, and panning. The user can only gain insights by exploiting the set's (meta)data which is related to the current view. The first level of integration is therefore realized through brushing and linking between the views to make, for example, connections in the result set visible. By cross-highlighting, the user can answer questions about the frequent filing countries of the applicant with the highest number of patents in the set. While being a powerful method, brushing and linking can only show connections between the selection in one view and its representation in the other views, but does not take into account their combination.

The second level of integration is therefore the saving and recombining of selections. Multiple views can now be used to define subsets and to combine them employing set operators, allowing the user to answer the same type of questions as above but with additional restrictions from other views, e.g., 'who is the applicant with the highest number of patent documents valid in Spain within my result set?'. This question could also be formulated as a new query, but this would make the combination of the answer with other subsets of the result set more complicated.

Up to this point patent analysts do not leave the phase of exploring the result set. While this phase is important for creating insights regarding the problem domain, it interferes with the patent domain's need for high relevance of result sets. Therefore, query widening has to come into play. The third level of integration addresses this requirement in the form of a query refinement by result set interaction. The views are aware of the type of data they are displaying and are capable of providing a search expression based on the user's selection in the corresponding view. The selection management component, in turn, is capable of combining the selections and their attached search term description into complex queries. Finally, the visual query editor allows for the direct incorporation of (combined) selections, to find more or exclude documents of the specified kind. This aspect cannot be achieved by a single component, but only by the whole system.

It is important that the last step of integrating findings into previous query formulations is steerable by the analyst, since the system cannot decide automatically *how* the integration into previous queries should be realized. The semantics regarding the correct scope inside the query and the intended Boolean operation to be applied for the integration have to be provided by the human analyst. The feedback loop to exploit insights from the analysis for query refinement is realized in PatViz in such a steerable manner, e.g., by dragging content nodes of the selection management facility directly into the query view or by adding it via context menu. A description of a use case, exemplifying iterative patent search and analysis with the proposed set of techniques, can be found in Koch et al. [2009].

For patent search and analysis the described composition of the visual front-end also opens up new search strategies which can hardly be followed using traditional approaches. A direct and formal query approach is intelligently connected with views for explorative proceedings. This allows for a seamless combination of an analyst's previous knowledge with berrypicking strategies [Bates, 1989; Hearst, 2009], which can be applied as a secondary means for increasing relevance. While the query approach should accommodate patent searchers in their established routines, visual berrypicking introduces new strategies alongside these familiar search patterns.

Instead of applying a search plan starting with a high-precision query as described by Alberts et al. [2011] followed by subsequent systematic broadening of the search, patent analysts can start with high-relevance approaches taking into account all factors at the very beginning. With the selection management system and filter techniques they can test and compare the different aspects against each other in order to increase precision for reducing the effort of a subsequent detailed patent inspection. Furthermore, crosschecking of patents and the (in)validation of hypotheses subsets becomes available through selection management. This establishes trust in the relevant subsets of a broad search, again without the cost of additional query formulation and the need to store intermediate results for later comparison. Such an approach can reduce the number of required iterations, as opposed to currently available systems for patent search, and is suitable for speeding up the search process.

This chapter presented an approach that covers the iterative patent search and analysis process. While the approach clearly addresses a specific domain, it is still very flexible regarding the analytic paths that can be followed by analysts. Moreover, it can be adapted for analyzing and searching scientific literature, which is also one aspect in patent searching, but has not yet been addressed. Domain adaption is important to optimally meet the needs of domain experts; for larger analytic approaches as those presented here, it is therefore imperative. Smaller analytic subtasks, which can also be supported employing visual analytics approaches, however, have the potential for broader application. The next chapter presents two of these approaches, also in context of patent analysis, but emphasizing characteristics that make it easier to generalize them.

Plug-In Visual Analytics

While the previous chapter covered the application of visual analytics concepts on the task and process level, this chapter describes two methods at a much more fine-grained level, focusing on subtasks as part of visual analytics approaches at a larger scale. The first of those is an interaction technique for exploring large and dense node-link-based views, the second one presents a method for creating task-tailored text retrieval mechanisms in the form of binary classifiers.

What makes these approaches interesting is that their design also follows visual analytics principles of combining visualization, interaction and automated techniques, inclusion of feedback loops on different levels, etc. Their application within larger systems can be seen as a recursive stacking of visual analytics approaches. If carefully designed, such approaches have the potential to be adapted to and employed in other contexts as well.

Even if their integration into larger analytic loops is optional and might only be suitable for specific groups of users or for specific tasks, the fact that they can be integrated seamlessly into larger visual analytics frameworks and tasks further increases the power of available methods. As a side effect of their focus on specific aspects of an analysis, they are less domain-dependent and can be transferred to other analytic scenarios where similar problems have to be addressed. By providing corresponding interfaces, even orthogonal aspects, such as provenance, collaboration, and presentation (see Chapter 5), can be preserved and made available to the larger visual analytics tasks.

This chapter is partly based on the following publications:

A. Panagiotidis, H. Bosch, S. Koch, and T. Ertl. EdgeAnalyzer: Exploratory Analysis through Advanced Edge Interaction. In *Hawaii International Conference on System Sciences (HICSS 2011)*, pages 1–10, 2011

F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual Classifier Training for Text Document Retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012

4.1 Visual Analytics on the Interaction Level

With ‘EdgeAnalyzer’ [Panagiotidis et al., 2011] a focus+context technique was developed that features visual analytics characteristics. It can therefore be seen as the recursive application of visual analytics principles and feedback loops within one interaction mechanism eligible to be used in the context of larger visual analytics approaches such as PatViz.

In information visualization the integration of several data aspects within one interactive view that provided details and context at the same time by using focus+context techniques [Card et al., 1999], is often considered to be preferable to providing different linked views showing these aspects separately, since the cognitive load for users can be lower for such integrated views compared to separated ones. Integrated views using focus+context interaction are particularly suggested for explorative tasks and scenarios, since switching between displays or windows can have negative effects on visual search performance [Larkin and Simon, 1987].

One example of such a visual combination of different data types is the integration of relational, i.e., graph-based, data properties on top of other spatially represented information, e.g. in the form of a two-dimensional map. A broad variety of scenarios can benefit from this combination. Examples are the visualization of air traffic, migration information, and other relational information on a map, the visual representation of call-graphs within hierarchically organized software packages, or, as addressed in the following, relational patent co-classification information depicted on the classification hierarchy represented as a treemap.

Graph-based data is often visualized in the form of node-link diagrams. In order to help users understand and interpret node-link diagrams, a large variety of layout algorithms are employed, which try to provide a suitable layout that takes into account a good trade-off between a variety of potentially contradicting criteria

[Di Battista, 1999]. If the number of nodes and their connectivity exceeds a certain level, even sophisticated and carefully designed graph layouts cannot prevent them from appearing visually cluttered due to overdraw effects of nodes and edges. Unfortunately, the combination of graph data on top of spatially represented data exacerbates the problem of overdraw, because not only the graph itself becomes cluttered but the spatially represented information is also occluded.

In recent years, various solutions for reducing visual clutter in such situations were proposed. A general overview of clutter reduction techniques, not limited to graph visualization, can be found in Ellis and Dix [2007], while von Landesberger et al. [2011] provide an overview of the state of the art in analyzing large graphs including interaction techniques. Specific methods to decrease visual clutter for graph drawings include graph drawing algorithms [Di Battista, 1999], node clustering [Kaufmann and Wagner, 2001] or summarization of nodes with other visualization techniques [Henry et al., 2007], as well as edge bundling techniques [Holten, 2006; Cui et al., 2008; Holten and van Wijk, 2009; Lambert et al., 2010; Gansner et al., 2011; Selassie et al., 2011]. The latter aim at presenting users with the gist of relational connectivity without occluding too much information of the underlying spatial visualization. Besides these constructive approaches, interaction techniques, as described in Wong et al. [2003]; Wong and Carpendale [2007]; Hurter et al. [2009] for local reduction of visual clutter in dense graphs were developed.

While maintaining good overview, edge bundling techniques come at the cost of obfuscating details of the depicted relational information. To some extent this problem can be diminished by using techniques such as emphasizing the strength of edge aggregation by color saturation; however, especially if edge properties are of interest, additional techniques are necessary to let analysts inspect details. Techniques for exploring edge properties interactively can help in situations like that. Indirect inspection of edges can be realized by enabling users to interact with nodes, which is suitable for tasks where analysts know the important nodes beforehand and are interested in their connectivity.

This approach, however, might result in many tedious interaction steps with nodes in order to understand the properties of aggregated edge bundles during explorative tasks. Particularly if multiple edges from different locations are bundled together ending in another region but pointing at spatially separate targets there, indirect exploration gets laborious because, potentially, every pair of nodes has to be explored separately. A solution to solve this problem is direct interaction with edges or bundles of them. The technique proposed with EdgeAnalyzer falls into this category of edge-interaction mechanisms and aims at explorative tasks (see Figure 4.1). Apart from exploring edges or bundles of them, it additionally provides mechanisms for drilling down into specific edge properties in order to select exactly those needed for the larger task to be solved. Alternative approaches for exploring

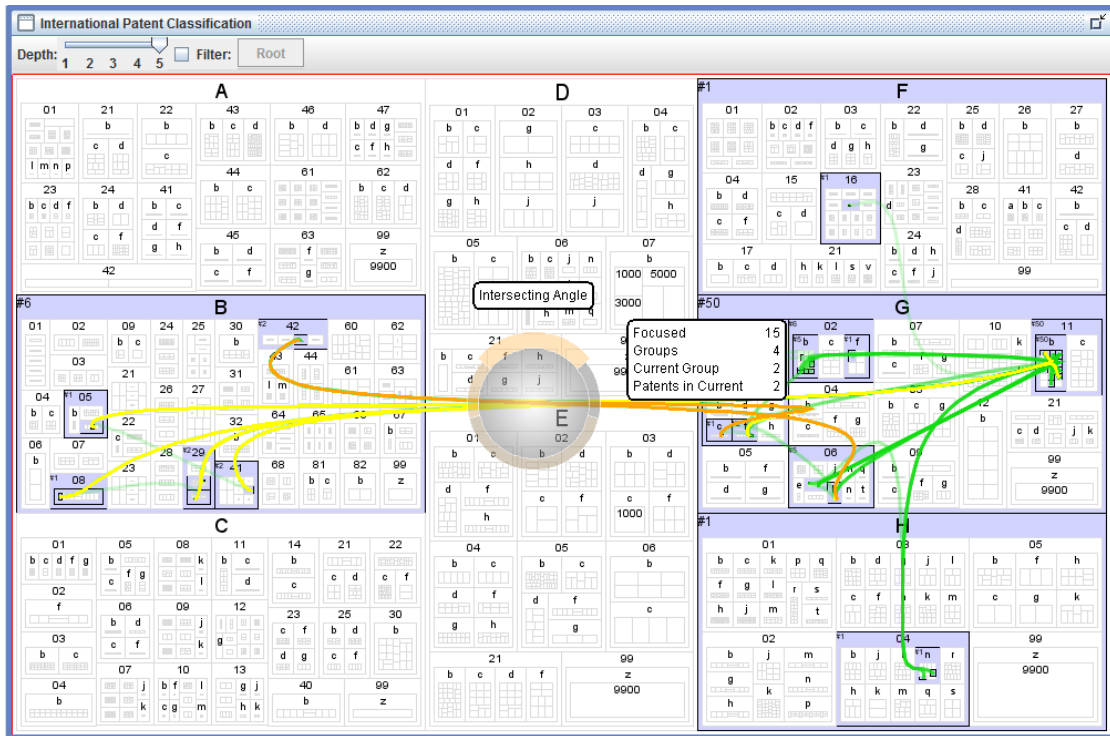


Figure 4.1 — The EdgeAnalyzer focus+context technique applied to the patent co-classification scenario as described in Section 3.2.2

the properties of large graphs, such as NodeTrix suggested by Henry et al. [2007], are available, but they are not developed to be applied on top of already available views with fixed spatial layout.

4.1.1 A Focus+Context technique for Edge Exploration

EdgeAnalyzer provides a lens-based visual metaphor for inspecting edges within a region. The lens can be resized for narrowing or widening the focus of interest. Theoretically, arbitrary shapes are possible for this interaction mechanism, but throughout this section circular lens shapes are considered in text and images.

EdgeAnalyzer provides a three-stage process for detecting edges of a hovered area in a first step, optionally grouping edges in a second step, and providing alternative views of the second step's outcome in the third step (see Figure 4.2). Corresponding to these steps, separate modules with well-defined interfaces are provided in order to make the mechanisms for grouping and visualization exchangeable for users and extensible for new methods if they should be required.

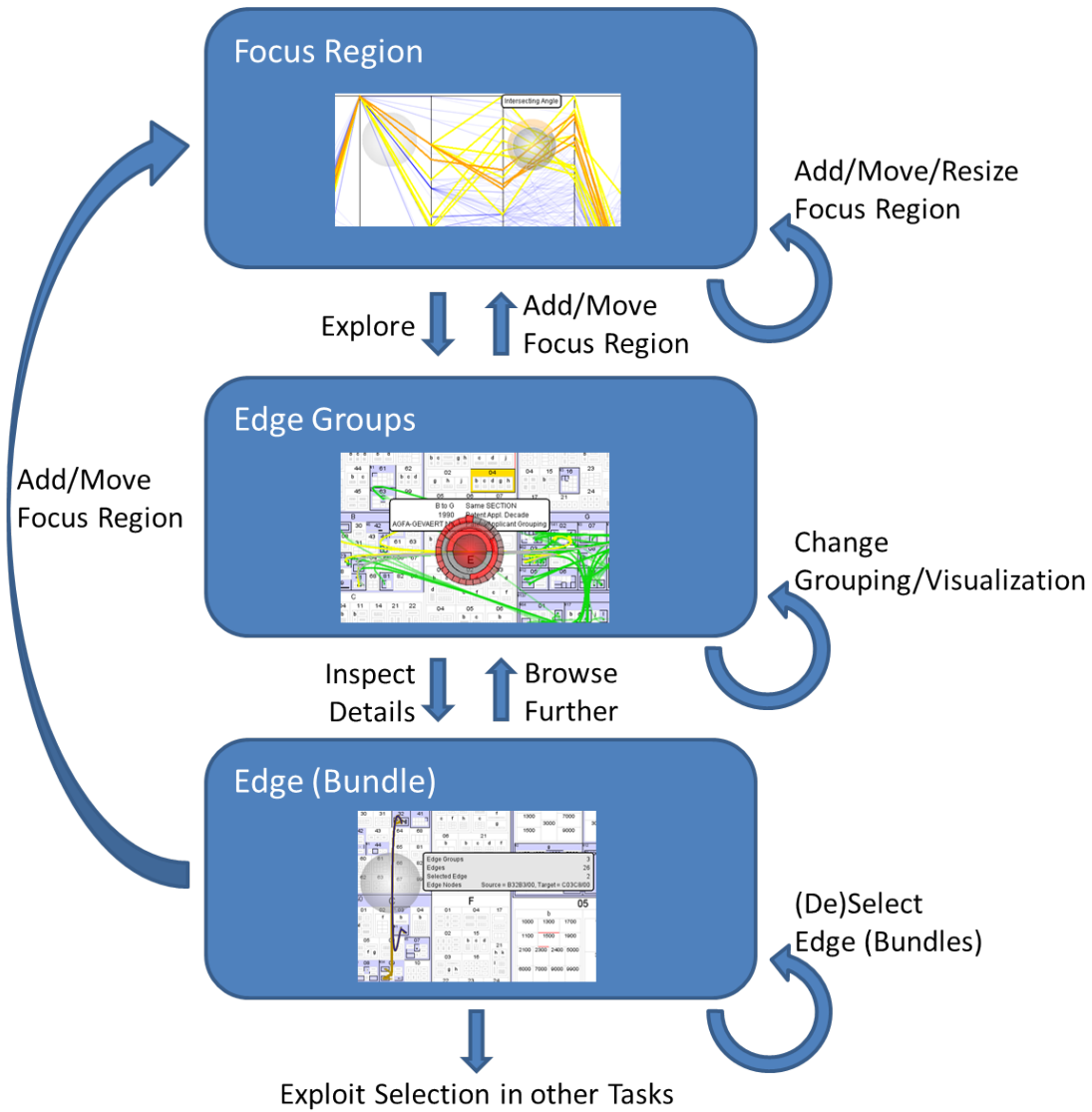


Figure 4.2 — The basic process for carrying out edge exploration with the EdgeAnalyzer approach.

Edge Exploration

In order to explore edges, analysts simply have to move the lens which is shown semi-transparently over an existing visualization that employs edges for depicting relational aspects. During movement as well as resizing of the lens, hovered edges are automatically and dynamically detected, grouped and visualized according to the user's preferences. The mechanism takes into account edges intersecting the lens as well as fully covered edges. The detection mechanism itself is generic, given that underlying visualizations are able to provide geometric information on their edges via a predefined interface.

If many edges or bigger edge bundles are inspected in the lens area, showing details for all edges from the beginning makes it difficult to understand the situation quickly. In order to make larger sets of edges manageable, ways of grouping them according to a user's needs are required. In the EdgeAnalyzer approach this is realized through organizing the edges supplied by the parent visualization in a flexible, internal data structure. After the detection of hovered edges is finished, internal grouping of edges is enabled through the corresponding module.

The grouping can be realized based on various edge properties. These properties include *geometric aspects*, such as intersection angle of edges with the lens, intersection points, global edge direction, etc., and *metadata properties of edges*, which, in case of patent co-classification, for example, comprise years of application, applicants, designated countries, and so forth. Thereby, characteristics visualized by the parent view EdgeAnalyzer is applied to can be taken into account as well, if available. Naturally, this comprises geometrical aspects *and* metadata. If metadata is to be included in the grouping process, another interface for accessing the parent visualization or the data it is based on has to be provided.

Grouping itself is realized through aggregation and clustering. The grouping can therefore be seen as a user-steered automated step in the process of exploration and selection, which is typical for visual analytics approaches. Using automated methods such as clustering (e.g. k-Means clustering) is especially helpful if large amounts of edges are explored and the user does not know beforehand, which criteria might be well suited for aggregating them. In a second step, the clustering might be refined by choosing different clustering parameters or by switching to an alternative grouping based on metadata characteristics of edges.

Depending on the task, different views for depicting the situation in focus can be beneficial. In order to get an idea on the number of edges and bundles of them, (local) de-bundling strategies as shown in Figure 4.3 can be conducive. The different views can take into account the edges' context outside the focused area, or ignore edges' paths and present them in an abstract way, independent from the parent view. In either case, the visualization of edges relies on the selected grouping

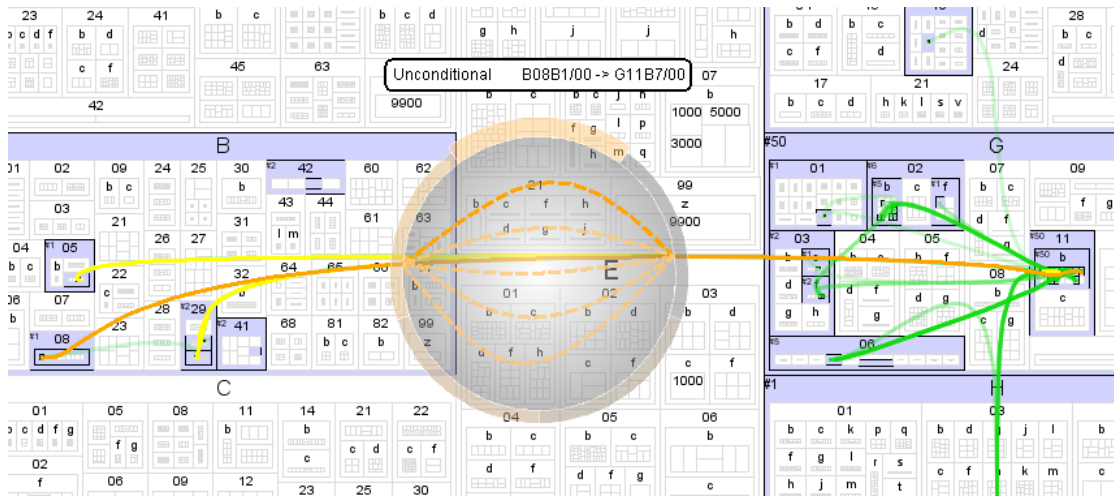


Figure 4.3 — EdgeAnalyzer’s lens applied to a patent co-classification edge bundle. No grouping is activated and the visualization mode is set to de-bundling.

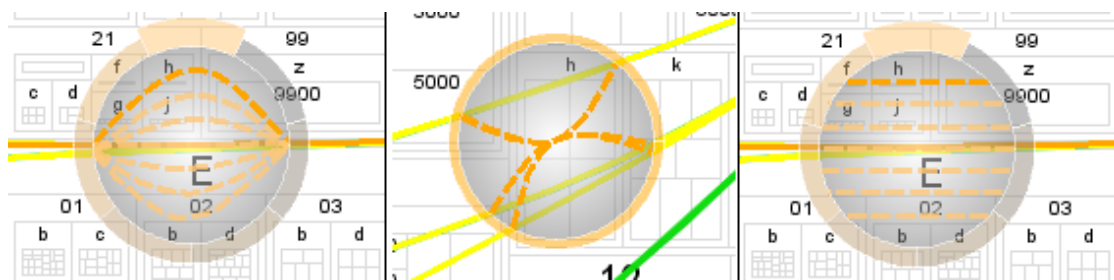


Figure 4.4 — Different visualization modes are available for users to choose from. In the left variant focused (groups) of edges are ‘de-bundled’. Similar with the right visualization. Here edges are shown as horizontal dashed lines instead. In the perspective in the middle, all edges are bundled together.

of the previous step. In order to keep users informed about the active grouping mechanism, an optional tooltip is provided. Another switchable tooltip summarizes the characteristics of the focused region by showing the number of edges, number of groups, as well as the id/label and additional information about the currently highlighted edge group (see Figure 4.5).

Browsing Groups of Edges

Browsing (groups of) edges is possible using the mouse wheel. On each wheel tick the next edge (group) is highlighted and details are depicted in the corresponding tooltip. By combining grouping and visual inspection of these groups, the browsing

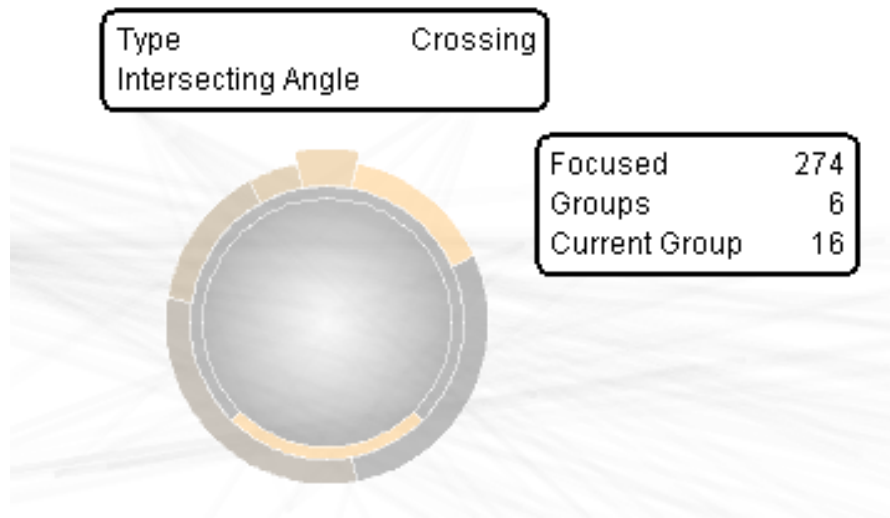


Figure 4.5 — Components of an EdgeAnalyzer lens. Two arc wheels for iterative grouping/drill down are shown around EdgeAnalyzer’s lens, as well as tooltips, telling

mechanism enables analysts to filter those groups that are of specific interest to their tasks. In case of employing EdgeAnalyzer in the co-classification view of PatViz, this, for example, can be exploited to determine IPC groups where the patents under analysis are frequently co-classified in, because edges always relate to patent documents. If such a frequent co-classification exists, this insight enables analysts to broaden their search to IPC subclasses or groups that were not taken into account in previous analyses.

Although specialized views for depicting and browsing edge groups are available, depicting groups’ sizes to these views is problematic. Mapping size to edge width or color, for example, makes depicting edges’ contexts, e.g. in form of aligning them to their paths outside the focused area, difficult. Additionally the space inside the lens is limited, and views are shown on top of existing visualizations. Depicting various properties in parallel inside the lens can quickly result in additional visual clutter. In order to avoid this issue, the lens was extended with a visual mechanism that has been termed ‘arc wheel’.

The arc wheel is a circular structure displayed around the lens, which is partitioned into arcs representing the currently explored edge groups and sizes (see Figure 4.5). An arc’s size thus depends on the number of edges within its group relative to the total number of focused edges. When users browse through edge groups,

the correspondingly selected arc is always located at the top of the wheel, while the wheel is rotated during browsing. All segments are colored according to an interpolated palette that fades from the edge group color to a medium gray, in order to indicate the wheel's current position to the users. Furthermore, if a subgroup is selected by an analyst, grouping can again be applied to the subgroup according to a user's needs. As a result, multiple arc wheels are stacked outwards, making complex filtering and iterative drill-down possible. In the patent co-classification scenario such a drill-down operation could, for example, consist of selecting a prominent bundle of edges connecting to specific IPC groups first, then exploring this bundle according to the patents' applicants showing immediately the most important players who applied for patents classified in both IPC groups.

The approach of the arc wheel shows some similarities to the 'Details Outside' method described by [Stasko and Zhang \[2000\]](#), who suggest a focus+context interaction technique for exploring subhierarchies in sunburst visualizations. With the Details Outside method the focused subhierarchy is drawn around a sunburst overview depicted with reduced size in the visualization's center. However, there are a number of significant differences to EdgeAnalyzer's arc wheel. Firstly, the arc wheel does not directly reflect the hierarchical nature of any underlying space-filling view, but an edge grouping hierarchy instead. Secondly, in the case of stacked arc wheels, every sub-group is shown as a circle of arcs, representing the partitioning of the selected parent group as a whole and not as a fraction, as it is done with child nodes of the selection in the Details Outside method. And thirdly, the arc wheel is always shown relative to EdgeAnalyzer's movable lens and not at a predefined location as in [Stasko and Zhang's](#) approach.

Selection of single edges and edge groups is possible throughout the browsing process. In the case of EdgeAnalyzer's integration into the PatViz interface, this results in the constrained selection of (sets of) patent documents, that can be used both for highlighting and within the selection management technique. Selections are handled analogously to PatViz. They store the selections' constraints with the selection of patent documents and can be directly used for query widening. However, selection criteria are derived from the chosen grouping method and storing selection constraint is only possible if grouping is metadata based. In those cases where grouping is based on geometric constraints, only the patent documents are marked as selected.

Advanced Filtering

EdgeAnalyzer also facilitates the combination of multiple lenses. Either independent or dependent child lenses can be spawned by the user in order to address more complex analyses. In this context, *independent* means that restrictions of the first

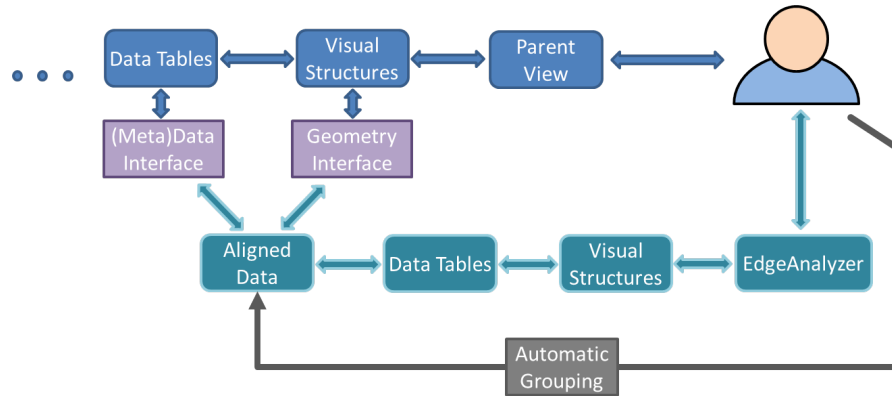


Figure 4.6 — The architectural dependencies of EdgeAnalyzer, showing a part of the parent VA model in blue (see Section 2.4), EdgeAnalyzer’s model (turquoise), and the required interfaces (purple).

lenses are not considered in grouping and filtering of the second lens. If independent lenses are used in a larger application context, both selections are combined accordingly with the Boolean OR operation. In the case of *dependent* lenses, the filter and drill-down operations of a child lens only apply to the selection of the parent lens. This is of particular benefit if the effects of one lens need to be observed in a spatially disjunct location, or if a target region is heavily cluttered with edges and pre-selection in a less cluttered region can overcome this issue. Dependent lenses can be seen as describing Boolean AND relations regarding the combination of their constraints. As mentioned above, the described focus+context interaction techniques require certain information from the visualization it is to be applied to. Figure 4.6 depicts these dependencies schematically. If the described interfaces can be provided, the technique can be flexibly employed with any edge or link based view.

Shortly after the publication of the EdgeAnalyzer approach a similar approach was presented with ‘MoleView’ by Hurter et al. [2011]. In contrast to EdgeAnalyzer, MoleView has been applied to a broader spectrum of visual primitives in addition to edges, including points, and image’s pixels. Similar to EdgeAnalyzer, it takes into account geometric properties as well as additional domain-related attributes of visually depicted data in focused regions. MoleView also facilitates mouse-wheel interaction resulting in changing the range of attribute values used as a constraint for filtering the underlying data. While MoleView is presented as a purely explorative method, EdgeAnalyzer has been designed to facilitate complex filtering and selection tasks required within larger application contexts. In addition,

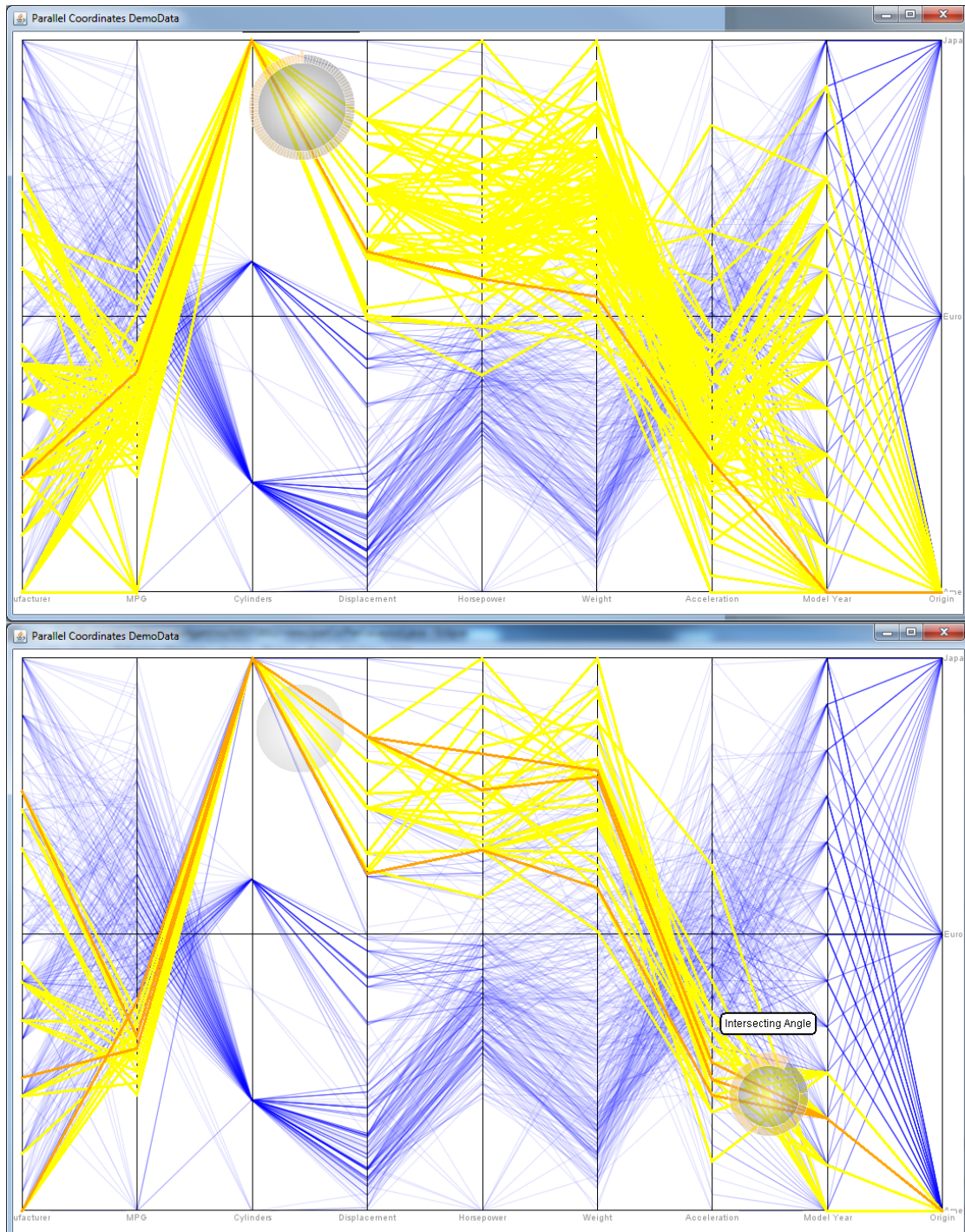


Figure 4.7 — Employing dependent lenses in the context of a parallel coordinates [Inselberg and Dimsdale, 1990] view. The upper image shows the situation with one applied lens, while the lower image depicts the situation after adding a dependent lens, restricting the edges to those intersecting both lenses.

EdgeAnalyzer enables analysts to drill down into underlying edges hierarchically, to step through (groups of) them, to choose from a set of automatic grouping methods, to switch visualization modes within the lens, and to apply multiple (in)dependent lenses for advanced explorative analysis.

4.1.2 Feedback Loop

A variety of low-level feedback loops are facilitated by EdgeAnalyzer. As with all focus+context techniques, focused regions can be explored while context is available which could be seen as the technique-immanent basic feedback loop. In the case of EdgeAnalyzer, the accentuation of edges under inspection is also visible outside the focus region enhancing this feedback loop for the specific analysis of links and edges. Furthermore, analysts are provided with direct feedback on edge group size and number through the arc wheel and optional tooltips. The tight integration of the arc wheel into edge-browsing activities keeps this information always up to date with regard to the currently selected edge (group), while still providing an overview of all groups under inspection. Through using multiple lenses, intra-visualization brushing & linking is supported which only applies to specific visual items within a view, in this case edges.

4.2 A Visual Analytics Approach to Classifier Creation

The combination of (semi-)automatic methods, such as from the field of machine learning, with interactive visualization techniques is one central aspect of visual analytics. Visual analytics methods can also be used to facilitate the creation of such automatic methods, e.g., for enhancing text retrieval tasks. This is especially interesting regarding the application of automated techniques tailored to specific subtasks which cannot be easily anticipated before they have to be addressed. One of the problems of employing task-tailored machine learning techniques is that some of them need extensive parameterization or training in order to be used effectively. Without any previous knowledge about machine learning, the creation of high-quality tools is hard or even impossible to achieve for an analyst. Supervised machine learning techniques have the potential to be used in such task-tailored tools, since they are normally trained with examples that are manually *labeled*. For scenarios described in related work, text labeling effort is often externalized by assigning the labeling of data to people, who are not interested in the classification task itself. Typically the training of a machine learning technique's model is then performed separately at a later point in time.

Apart from this aspect, the described approach differs from others that directly integrate an already available, pre-configured machine learning technique in domain-tailored systems, because it facilitates the visual interactive creation of a tool through an analyst, which can afterwards be employed fully automatically as part of retrieval tasks. A comparison with the search facilities included in the PatViz front-end illustrates this point. Here, general purpose techniques, such as keyword search on a document retrieval index, or approaches for image retrieval, and semantic search created in domain-specific preprocessing steps by specialists, were employed. These cannot be configured and adjusted by analysts themselves.

The following section presents an approach in which analysts can create and assess a classifier interactively and visually, for exploiting it afterwards in text retrieval tasks. In the context of this chapter, classification should not be mistaken for searching predefined classifications schemes like the IPC, which relates to a search for patents using classification codes from one of the patent classification systems (IPC, ECLA, US classification, and Japanese F-Terms classification). Instead, classification here refers to an automated technique that can be assigned to the field of machine learning.

4.2.1 Background and Motivation

Many domains and scenarios include the subtask of searching or filtering the data to be analyzed, especially if textual documents are involved. This is even the case if tasks are unknown at the beginning of the analysis, or if it starts with relatively vague objectives. Eventually, a more concrete information need (cf. Manning et al. [2008]) will manifest itself during exploration in said analysis processes and lead to these search and filter tasks. As has been discussed before, information need in the patent domain is often recall-biased (see Section 2.10, meaning that as many relevant objects as possible should be retrieved, while still maintaining good precision (see Section 2.5.2)). This restriction does not only apply to the patent domain, but is typical for all scenarios where missing relevant documents during search is not acceptable.

In common retrieval scenarios, analysts have to translate their information need into a keyword search query or a combination of filtering constraints. This implies the ability to derive such explicit queries or constraints either from a set of examples small enough to be manageable for an analyst – i.e., to read through them – or the skill to guess useful keywords and metadata restrictions. Generally, the creation of queries and constraints works well, especially if the analysts are experienced with respect to selecting keywords relevant to their tasks and domain. However, it can still be difficult to achieve good coverage of relevant documents if not all important aspects are considered.

In situations like that, the interactive creation of binary classifiers can be an additional, complementary method for improving information retrieval during analytic tasks. Here, the classifier's purpose is to separate a corpus into relevant and non-relevant documents and to improve recall through generalization. To demonstrate the applicability of the approach, linear support vector machine (LSVM) classification [Vapnik, 1998] has been applied in text retrieval scenarios intending its integration with keyword search based methods. This choice was motivated by the fact that LSVMs have been successfully applied to text classification tasks [Joachims, 1998, 1999] and they are known to achieve good classification performance with the vector space model, which is used for the approach. SVMs can also be used for multi-classification by stacking several of them (see Seifert and Granitzer [2010]). For their employment as part of retrieval tasks binary classification into relevant and non-relevant results is sufficient. The prototype system developed for visual classifier creation has been tested on different document sets, including a corpus of news groups postings, a set of news articles, and a corpus compiled from the abstracts of scientific papers. Apart from the latter corpus, the others are only of subordinate interest in the field of patent analysis. The choice was mainly based on the possibility to create gold standards needed to assess analysts' classification performance during a comparative user evaluation. As part of a user study, the approach was compared to other interfaces through which users created classifiers with less interaction and visualization support (see Section 6.1.4).

Two prerequisites have to be considered, if classification is to be used during search tasks. On the one hand, the classifier must be adaptable, letting analysts represent their information need as accurately as possible. Thus, analysts must be equipped with techniques to perform this adaptation, ideally without the obligation to become machine learning experts before. Approaches that enable analysts to build classifiers primarily on the observation level [Endert et al., 2011] meet this requirement. Another term under which such approaches are summarized is the black box model [Bertini and Lalanne, 2009]. Since the application presented below can be seen as an instance of the black box model, one needs views for presenting text documents on different levels of abstraction, as well as interaction techniques to let analysts carry out informed labeling. Furthermore, the interaction mechanisms should help analysts in controlling the creation of classifiers by observing and judging their quality. On the other hand, analysts must be enabled to assess the creation of a classifier by observing and judging its quality, e.g. via providing suitable interaction techniques on these views. This is important for detecting classification problems, such as overfitting, or too broad generalization, while adapting the classifier to the analytic task. In order to allow users judge the

quality, the same views as for labeling plus a preview feature are provided, thereby facilitating an interactive, visual analytics feedback loop.

Additional requirements have to be met for creating classifiers efficiently. The task itself requires a labeling effort that can only be justified if an important information need exists. The employment of classifiers for search should therefore be an optional step, and the decision to build one has to be made by analysts themselves. Additionally, the creation of the classifier should be possible in an efficient manner, or else the problem of creating traditional complex search queries is just shifted to the other potentially laborious task of labeling examples to train the classifier with.

Significant speedup in training a classifier can be achieved by exploiting *active learning* (AL) approaches (see Section 2.5.3). AL aims at helping users in labeling those instances that provide the highest impact on improving the classifier's performance during the next training iteration. Often, the instances that are classified as most uncertain, i.e., those closest to the classifier's decision border in high-dimensional space, are chosen for this purpose. While normally AL provides users directly with these most uncertain instances, the method proposed here loosens this restriction and gives analysts the freedom to choose from a visually presented set of uncertain instances. This transfers the labeling initiative to the analyst, while still giving hints on selecting good labeling candidates. In order to support this selection, analysts are provided with feedback on the informativeness of their current labeling choice. Apart from picking the most informative documents, analysts can select multiple instances for labeling at once.

An integration of binary classifiers with key term search can be exploited in two ways. First, the key term query can be used to bootstrap an initial classifier in order to relieve analysts of the burden to build it from scratch. Second, the built classifier can be applied in conjunction with classic search and filter methods analysts have a considerable expertise in.

The approach presented here is meant as a visual analytics technique that can be employed as part of larger visual text analysis tasks, but also for creating dedicated classifiers that can be exploited in batch mode processing of text document collections. It is scalable with respect to the size of the document collection. Additionally, a mechanism for capturing provenance information of a classifier's evolution is available. This information is made available in the form of the interactive history graph enabling users to start over at any intermediate state of the classifier in case subsequent labeling actions lead to unsatisfactory results. The same mechanism can be used to visually inspect, understand, and retrain built classifiers at a later point in time if the need arises, e.g., when dealing with dynamic data sources (see Section 5.1.1 for details on scalability).

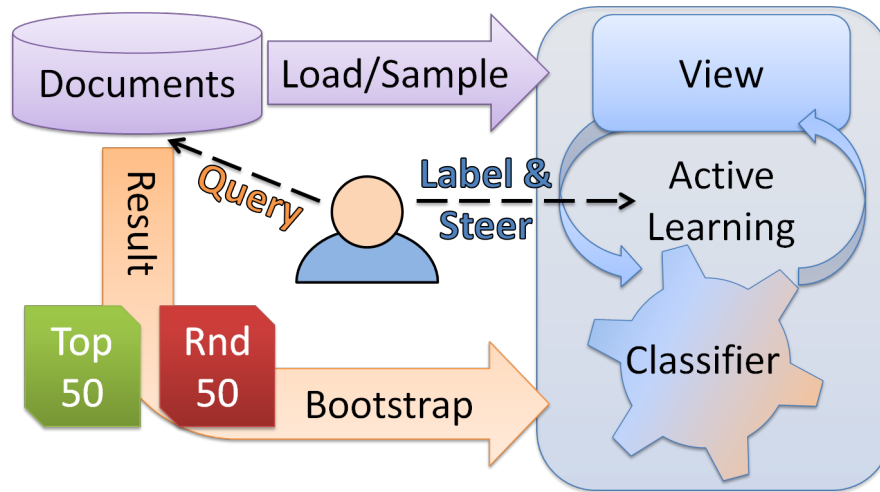


Figure 4.8 — 50 positive and 50 negative results of an user-formulates initial query are used to bootstrap an initial classifier. Afterwards the users train the classifier interactively through iterative labeling and training rounds.

4.2.2 A Prototype for Visual Classifier Training

Figure 4.8 provides a schematic overview of the system for individual text classifier creation. After an analyst provides an initial keyword query, the search is executed and the top 50 relevant documents are extracted using a ranked retrieval mechanism. In addition, 50 random, non-relevant documents are selected from the corpus as well, in order to have a set of positive and negative examples for bootstrapping the initial classifier. Once the classifier is created, its current state is visualized in a multiple coordinated view environment for inspection by the analyst. Various perspectives and interaction methods support analysts in identifying and selecting promising labeling candidates for adapting and refining the classifier during subsequent training iterations.

For the prototype implementation the Apache Lucene¹ framework was used again as text search engine for the initial bootstrapping step and as the base text document repository. The LibLinear² library [Fan et al., 2008], providing a very efficient linear support vector machine implementation for large data vectors, such as given for text representations with ‘bag of word’ models (see Section 2.5.2), has been employed for the classification tasks.

¹ <http://lucene.apache.org/>

² <http://www.csie.ntu.edu.tw/~cjlin/liblinear/>

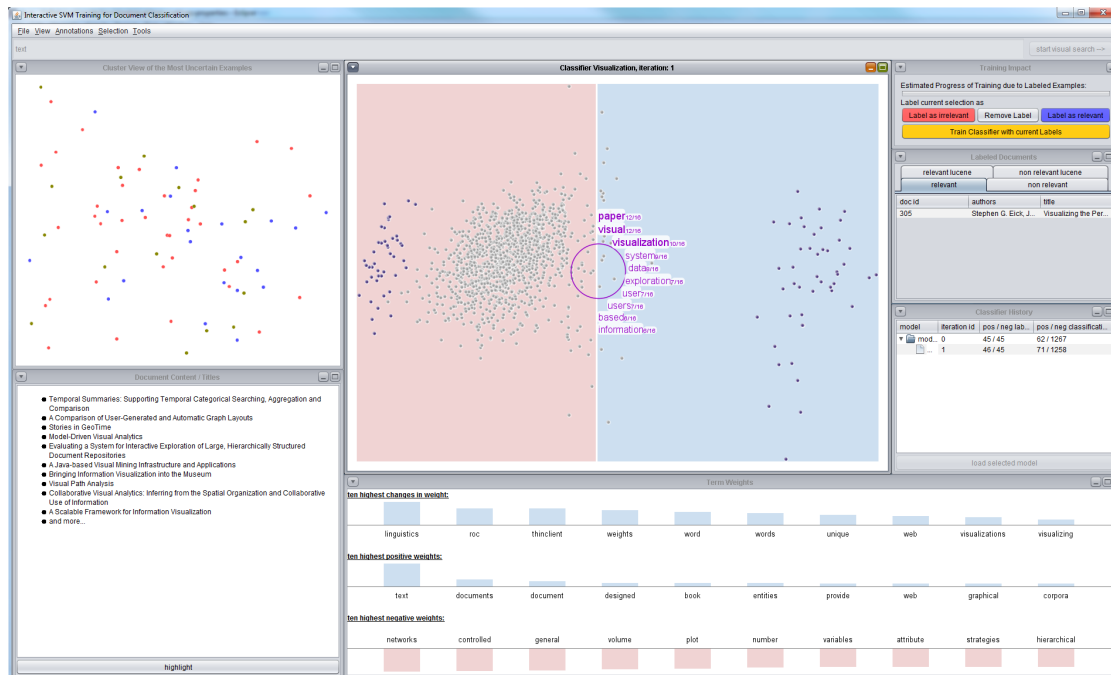


Figure 4.9 — The interface for user-steered classifier creation providing multiple coordinated views for inspecting a classifier's state.

Classification View

Figure 4.9 shows the views available for classifier training and refinement. The central idea behind the classification view (4.9) is to approximate the SVM classifier's current state in high-dimensional space as good as possible with a 2-dimensional analogy. As described in Section 2.5.3, an LSVM tries to find a hyperplane dissecting high-dimensional space in such a way that all training examples are separated according to their class membership, and that the margin between the documents closest to this decision border is maximized.

Accordingly, the classification view has been designed as a scatter plot to depict the two classes as two regions. The left region, shown with a light-red background, contains all non-relevant documents while the right, light-blue region holds the relevant documents. The white space between both areas represents the decision border or hyperplane of the LSVM. The documents are depicted as dots, which are either classified as relevant or non-relevant symbolized through their placement in one of the depicted colored regions. Training data, i.e., documents that have been labeled as relevant or non-relevant, either by the analyst or through bootstrapping, in previous steps, are shown in purple (see Figure 4.9, in the upper middle), whereas the gray dots are classified according to the classifier's state. The layout

of the dots in x-direction is solely based on their distance from the hyperplane in high-dimensional space, thereby representing the confidence or uncertainty of classification, showing the uncertain ones close to the decision border and the more confidently classified ones farther away. The distribution on the scatter plot's y-axis reflects inter-document similarity.

As mentioned above, especially those document close to the hyperplane are of interest to the analyst since they provide the potentially biggest impact during subsequent training steps. Accordingly, the set of the 100 documents U closest to the decision border are projected in y-direction according to their vectors' first principle component. This results in a good spatial distribution and reduces overdraw for these important documents, with the goal to make it easier for analysts inspect them. All other documents d_i are positioned in y-direction according to the ten documents closest to them $U_{10(i)}$ in set U . Similarity of the documents, or their vector representation respectively, is thereby computed using the cosine distance \cos . Their exact placement on the y-axis $y(d_i)$ is computed using the following weighted sum:

$$y(d_i) = \frac{\sum_{d \in U_{10(i)}} \cos(d_i, d) \cdot y(d)}{\sum_{d \in U_{10(i)}} \cos(d_i, d)}$$

Apart from helping analysts in concentrating on the uncertain documents near the hyperplane this approach also reduces the costs for computing the first principal component significantly and thus speeds up the creation of the view after a training step.

The classification view can be zoomed and panned to the the analyst's region of interest for closer inspection. Hovering document dots with the mouse or selecting them by clicking results in highlighting them in all other views. Additionally, selections can be made through rectangular brushing interaction or by using the term lens as described below.

The term lens visible in Figure 4.10 can be activated by pressing the 'shift' key, and its size is adjustable through using the mouse wheel. If activated, the term lens shows at most the top ten terms shared by the documents covered by the lens, mapping a term's document frequency to font size, and annotates them with the explicit frequency information. Additionally, the document frequency affects the ordering of terms from high frequency shown at the top to lower frequency shown at the lower part of the lens. The idea is completely analogous to the term cloud employed in PatViz, although it is used as part of a focus+context technique in this case and not in a separate view as in PatViz. The term lens facilitates quick browsing of the regions of interest, e.g. along the hyperplane, offering a coarse

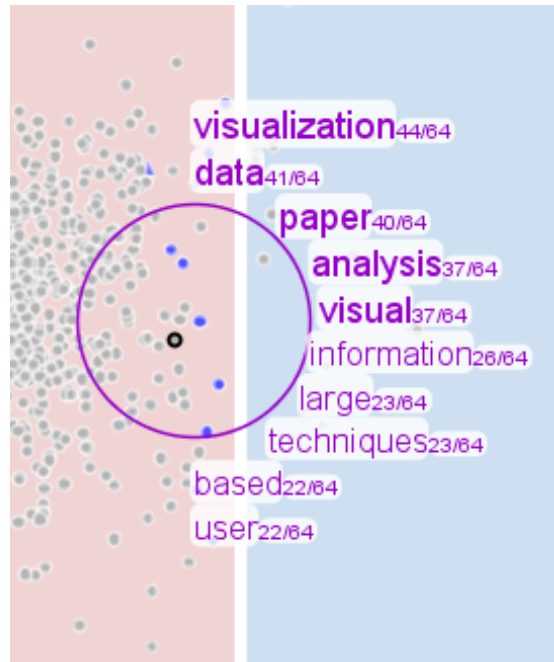


Figure 4.10 — The term lens applied to the classification view and the region near the hyperplane. The 10 most frequent terms are depicted around the lens (according to document frequency).

form of gist for the inspected documents to the analyst. By clicking during the usage of the lens, all covered documents are selected accordingly.

Detail View

The detail view depicts the textual contents of a document, if it is hovered or selected in one of the other views. This is of high importance since informed labeling decisions can often only be made if the unabridged contents can be accessed or at least scanned quickly. If multiple documents are selected, their titles are shown as a list of interactive links, facilitating detailed content inspection by clicking them.

Term Chart

Since documents are modeled as weighted term vectors on which the classification mechanism is also based, the term chart provides the user with direct information about the terms' importance with respect to the classification model. An LSVM's normal vector (see Section 2.5.3) representing the current hyperplane is a good

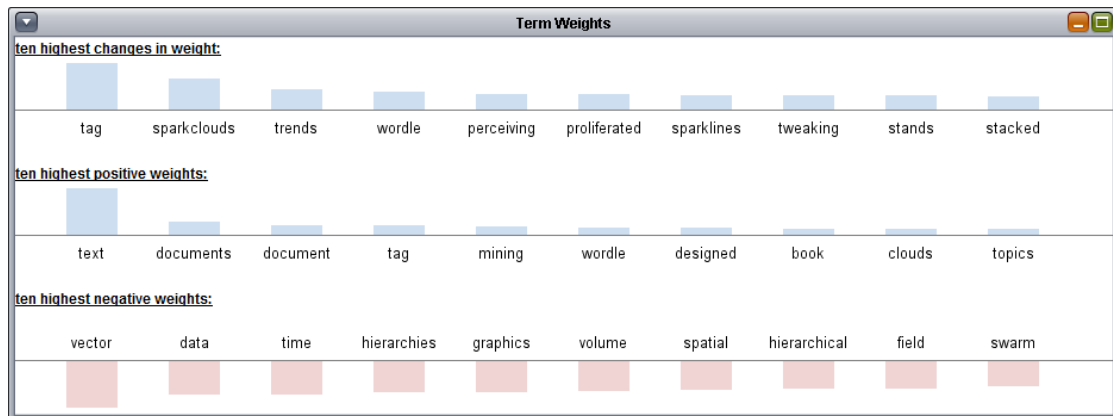


Figure 4.11 — The term chart depicting the changes in weight during the last classifier training in the first row, most positively weighted terms in the second row and the model’s ten most negatively weighted terms in the third row

indicator for providing analysts with additional insight on the model’s state and state changes. Accordingly, the top row of bar charts shows the ten terms with the highest changes compared to the classification model of the previous iteration. The middle row displays the ten terms that have the highest positive weights in the current model, and the bottom row shows the ten terms with the highest negative weights in the current model. This view gives analysts an idea of what the training algorithm has learned from the training data so far in summarized form. Each of the bars can be hovered by the mouse resulting in highlighting the documents containing the respective term in the other views. Selection works correspondingly by clicking a bar.

Cluster View

The cluster view (Figure 4.12) shows the 100 most uncertain documents (U) clustered solely by their similarity. In contrast to the classification view, this view does not take into account the hyperplane; it ignores classification uncertainty and uses both dimensions to depict document similarity. The clustering is computed using the bisecting k-means algorithm and accomplishes the subsequent projection into two-dimensional space using the LSP algorithm [Paulovich et al., 2008]. The respective implementations of the Projection Explorer (PEX) project [Paulovich et al., 2007] were used to realize this. Different class membership of documents is indicated by the corresponding colors. The basic idea of this view is to guide the user’s attention to potential candidates for labeling actions but not the clustering aspect per se. This view is supposed to show how very similarly documents are classified by the current model. An interesting observation in this view that would

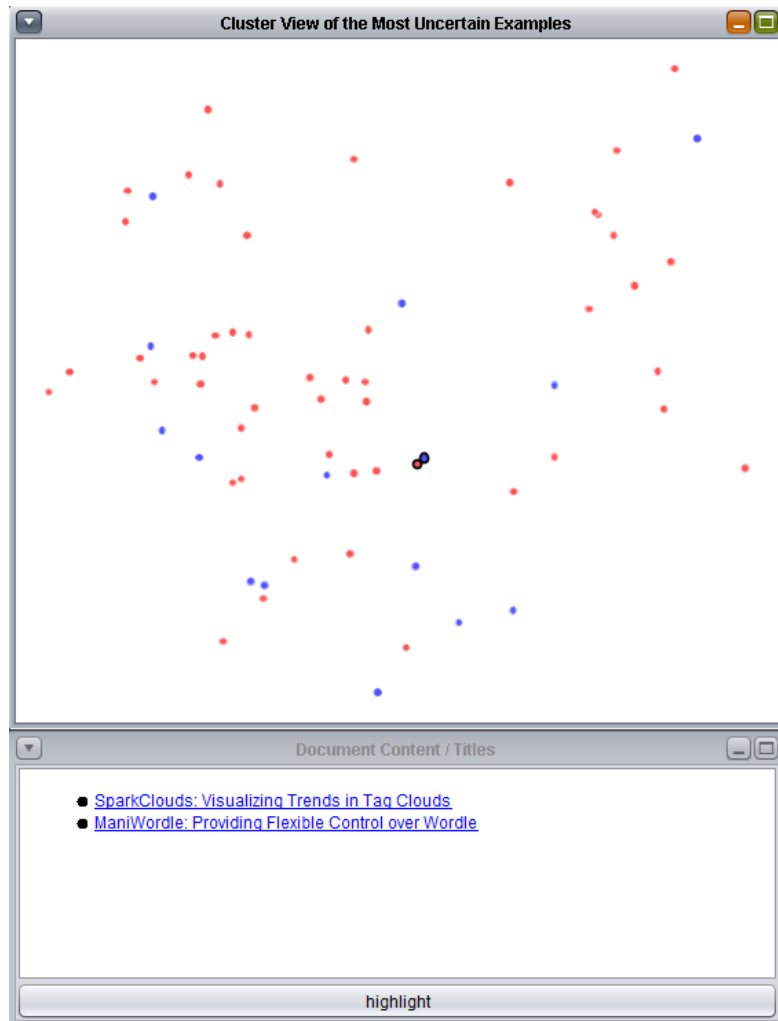
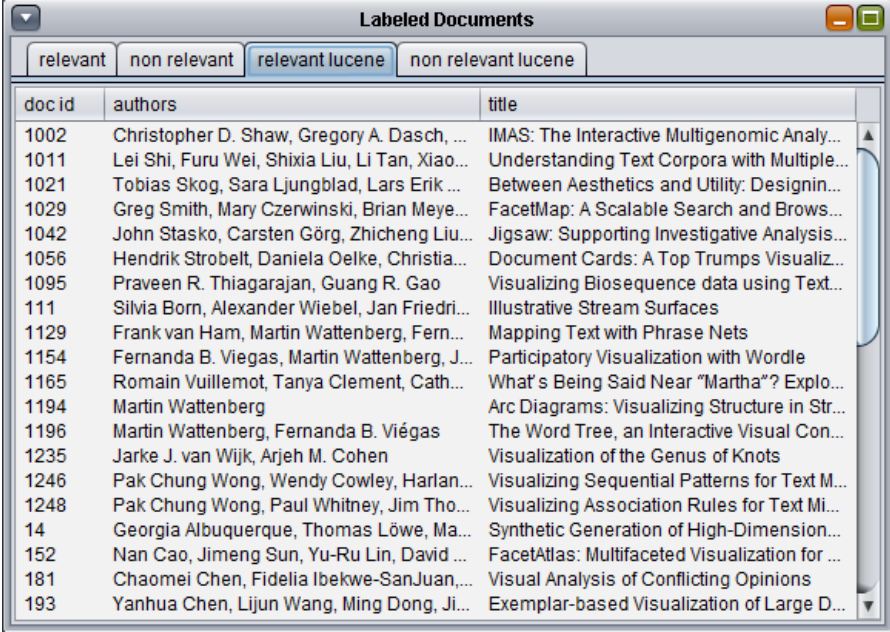


Figure 4.12 — The top 100 most uncertainly classified documents clustered according to document similarity. Documents placed close to each other having different class membership are potential labeling candidates. In the shown example, documents talking about text visualization should be separated from others. Two, according to the clustering, similar documents showing heterogeneous class membership have been selected. From the documents' titles it already becomes apparent that both documents are related to text visualization. As a consequence the incorrectly classified document can be annotated with the correct label.



doc id	authors	title
1002	Christopher D. Shaw, Gregory A. Dasch, ...	IMAS: The Interactive Multigenomic Analy...
1011	Lei Shi, Furu Wei, Shixia Liu, Li Tan, Xiao...	Understanding Text Corpora with Multiple...
1021	Tobias Skog, Sara Ljungblad, Lars Erik ...	Between Aesthetics and Utility: Designin...
1029	Greg Smith, Mary Czerwinski, Brian Meye...	FacetMap: A Scalable Search and Brows...
1042	John Stasko, Carsten Görg, Zhicheng Liu...	Jigsaw: Supporting Investigative Analysis...
1056	Hendrik Strobelt, Daniela Oelke, Christia...	Document Cards: A Top Trumps Visualiz...
1095	Praveen R. Thiagarajan, Guang R. Gao	Visualizing Biosequence data using Text...
111	Silvia Born, Alexander Wiebel, Jan Friedri...	Illustrative Stream Surfaces
1129	Frank van Ham, Martin Wattenberg, Fern...	Mapping Text with Phrase Nets
1154	Fernanda B. Viegas, Martin Wattenberg, J...	Participatory Visualization with Wordle
1165	Romain Vuillemot, Tanya Clement, Cath...	What's Being Said Near "Martha"? Explo...
1194	Martin Wattenberg	Arc Diagrams: Visualizing Structure in Str...
1196	Martin Wattenberg, Fernanda B. Viégas	The Word Tree, an Interactive Visual Con...
1235	Jarke J. van Wijk, Arjeh M. Cohen	Visualization of the Genus of Knots
1246	Pak Chung Wong, Wendy Cowley, Harlan...	Visualizing Sequential Patterns for Text M...
1248	Pak Chung Wong, Paul Whitney, Jim Tho...	Visualizing Association Rules for Text Mi...
14	Georgia Albuquerque, Thomas Löwe, Ma...	Synthetic Generation of High-Dimension...
152	Nan Cao, Jimeng Sun, Yu-Ru Lin, David ...	FacetAtlas: Multifaceted Visualization for ...
181	Chaomei Chen, Fidelia Ibekwe-SanJuan,...	Visual Analysis of Conflicting Opinions
193	Yanhua Chen, Lijun Wang, Ming Dong, Ji...	Exemplar-based Visualization of Large D...

Figure 4.13 — Overview on previous labeling actions (separated into non-relevant and relevant labels), including those from the bootstrapping steps.

deserve closer scrutiny, are, for example, heterogeneously classified clusters of documents. Such heterogeneously classified clusters identify suitable regions for detailed inspection since the chance that some of them are classified incorrectly is high. The term lens is also available in this view and works in the same way as in the classification view.

Training Data View

The training data view as can be seen in Figure 4.13 contains four different lists of the documents that have been assigned a label so far – either through the bootstrapping step or during iterative classifier training. The latter are accessible via the tabs ‘relevant’ and ‘non-relevant’, while the tabs ‘relevant lucene’ and ‘non-relevant lucene’ contain the documents added to the training set due to the bootstrapping. The documents in the lists can be highlighted in the views by clicking on them. This view is especially useful for inspecting the bootstrapping results, if an analyst suspects that the initial query definition might not have been precise enough.

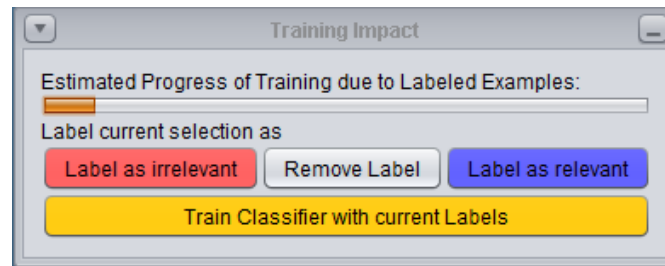


Figure 4.14 — The labeling panel showing the impact of currently labeled documents on a subsequent training step as well as the buttons for labeling actions and the removal of set labels.

Labeling Panel

Selected documents can be marked as relevant or non-relevant according to the current classification task by using the labeling panel shown in Figure 4.14. The panel offers two buttons for both labeling actions. In addition, labeling made during the current iteration can be revoked using the ‘Remove Labels’ button. The ‘Train Classifier with current Labels’ button triggers the training of a new SVM model considering all applied labeling actions. It can be useful to label wrongly classified documents on both sides, as well as confirming correct classification of yet unlabeled documents. Several selected documents can be labeled at once as well. As a direct result of any labeling action the changes according to document classification are shown as a preview in the classification view, without yet persisting the training step. The bar shown above the buttons of the labeling panel displays the impact of the currently labeled documents during the next training iteration. The shown impact is computed using a heuristic which takes into account that the reduction of the margin size of the new SVM model compared to the old model based on the fact that the margin of the classification model can be maximally reduced to half of its previous size by selecting one example during each iteration.

Labeling actions can be applied on selected documents. In the classification view and the cluster view, newly labeled documents are shown with triangle shapes, while other documents that would be affected in a subsequent training step are colored according to their anticipated change in class. Figure 4.15 depicts such a situation. Blue triangle-shaped glyphs, with one vertex pointing upwards, are labeled as being relevant. Red triangles pointing downwards represent documents labeled as non-relevant. The classification view additionally provides an automatic preview of the changes that apply regarding the current labeling situation. Red dots depict documents that will change their class to non-relevant, while blue ones will change to the relevant class respectively. The adaptation of the classifier by

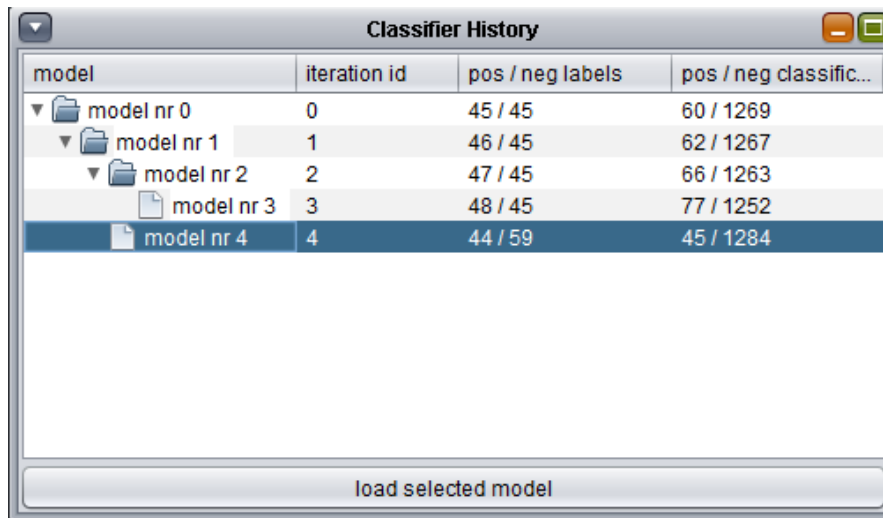


Figure 4.15 — The document represented by the blue triangle was labeled as being relevant resulting in the document represented by the blue dot being classified as relevant as well. A red triangle represents a document that has been labeled as being non-relevant.

training the model with the currently labeled documents introduces a new iteration and updates all views according to the new model’s properties. The described ‘preview’ mechanism is realized by computing the new classifier model through carrying out a training step each time a labeling action takes place. However, the effects of the new classifier are shown in the current visualization and they are not persisted until a user explicitly requests training. Afterwards the new situation is shown with an updated visual layout.

Classifier History

A classifier’s history is captured by preserving its state for each training iteration. This means that all intermediate states of a classifier are accessible at any point of the classifier creation process. In those cases where users are not satisfied with the results of a training step, e.g., because for some reason many obviously irrelevant documents are classified as relevant, it is possible to go back and reload a previous version of the classifier to start over with other labeling actions. The states of the classifier under development are depicted as a tree table as shown in Figure 4.16, whereby each iteration is assigned a unique ascending number to make the classifier’s evolution traceable. Analysts can also provide their own labels for identifying classifier states more easily at a later point in time. Furthermore, the number of positive and negative labels provided during a training iteration, as well as the number of positive and negative classified documents, are shown for



model	iteration id	pos / neg labels	pos / neg classific...
model nr 0	0	45 / 45	60 / 1269
model nr 1	1	46 / 45	62 / 1267
model nr 2	2	47 / 45	66 / 1263
model nr 3	3	48 / 45	77 / 1252
model nr 4	4	44 / 59	45 / 1284

load selected model

Figure 4.16 — The history of classifier creation during several training rounds. Here, arbitrary models can be loaded to start over with training iterations from a previous model.

each saved model. These numbers provide a coarse overview of the changes applied during one iteration and are intended to help analysts to judge the impact of the corresponding iteration. If users decide to go back to a previous classifier state and restart training from this iteration, a new branch is introduced to document this step. The mechanism represented by this view guarantees that no trained classifier model is lost.

4.2.3 Feedback Loops and Workflows for Classifier Training

Again a variety of different feedback loops are facilitated during user-steered classifier creation. The main analytic feedback loop is shown in 4.17. As with the interaction method presented in the first section of this chapter, visual classifier building is intended as an instrument that can be applied in addition to other search/retrieval tasks and not necessarily as a standalone method. If this is the case, it can be seen as a visual analytics method integrated into some larger visual analytics system.

Apart from the main feedback loop, a variety of smaller feedback loops are available through the orchestration and coordination of available views. While the system does not enforce a specific working strategy, a variety of promising approaches and

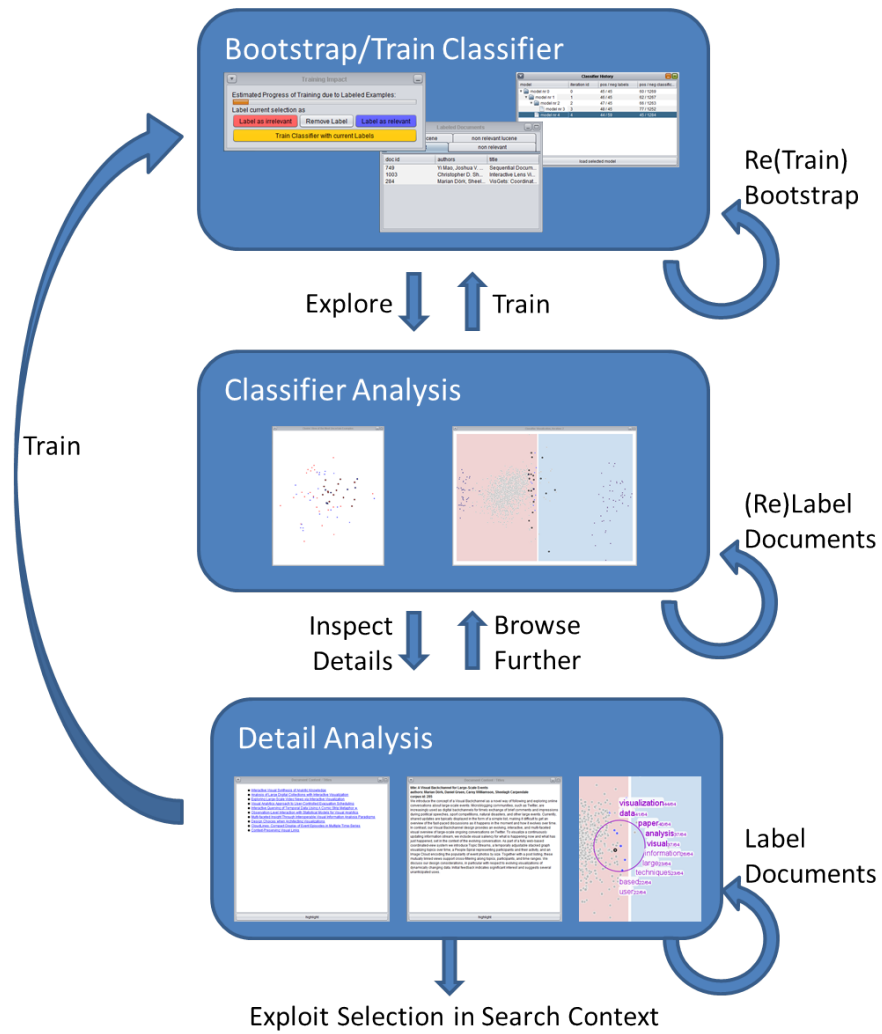


Figure 4.17 — An abstract overview of the classifier creation process.

usage patterns exist that exploit these micro-feedback loops. Some of them were intentionally integrated while others could be derived from the behavior of test subjects observed during the user study as well as from the comments received from the questionnaires and the discussion with participants (see Section 6.1.5).

The multiple coordinated views provide users with different perspectives on the state of the classifier. The most commonly used strategy during the evaluation is, as was expected, the selection or highlighting of documents in the classification view as well as from the cluster view and their subsequent inspection in the detail view. Especially the automatic preview in the classification view turned out to drive iterative refinements within one labeling session. As described above, labeling one or several documents, makes others change their class membership, as can be seen from the preview. Accordingly, users tend to check especially the documents with anticipated changes in class and immediately corrected unintended side effects of previous labeling actions, which led to high-quality labeling actions with considerable impact during follow-up classifier training.

The cluster view is typically employed as a secondary means for inspecting heterogeneously classified regions, as intended, and usually after exploring and labeling in the classification view. The term lens is used in the classification view as well as in the cluster view, and selections were made based on the displayed term frequency information. However, users refrain from labeling these selections as a whole; rather, they inspected the list of titles in the detail views and the document's content respectively before labeling.

The term chart facilitates explicit crosschecking and is typically applied in advanced stages of classifier training. Especially terms exhibiting an increase in importance according to the last training round and those shown as being generally important for the model are hovered to inspect the distribution of documents containing these terms in the classification view.

The approach also provides undo functionality on different levels of abstraction. Local undo operations within one iterative step are available through the remove labels functionality provided with the labeling panel. A global undo/redo mechanism is provided with the classifier history.

As mentioned before, active learning methods for labeling tasks actively request labels from annotators by employing a selection strategy for the instance to be labeled that promises the biggest chance for maximal training progress during subsequent training steps. As a consequence, fewer instances have to be annotated and classifier creation can be accomplished much more quickly. The most tentatively classified documents are good selection candidates as part of training an LSVM.

For the described approach, one goal was to transfer more labeling control to the analyst while still providing clues which labeling operations are likely to have high

impact. Three methods for speeding up user-steered classifier building, which could be seen as inspired by active learning, were integrated to support analysts with feedback on good labeling choices. As explained above for the classifier view, uncertainty directly relates to distance from the decision border. The closer instances selected for labeling are to this border, the larger the expected value for the next training iteration. The first and most important measure is therefore the representation of the decision border in the classification view offering a prominent clue where to label documents. The classification view's layout acknowledges the diversity of the documents closest to the decision boundary in order to support analysts in exploring labeling candidates in its vicinity. As a second measure, the cluster view only depicts the 100 most uncertain documents. This ensures that any labeling action in this view has considerable impact on the classifier's evolution during the next training step. The third measure informs analysts about the potential training progress of their labeling actions by providing a bar displayed above the labeling panel that shows expected impact. The labeling impact is computed from the reduction of the margin size of the new SVM model as compared to the old model, which is likely to decrease during the training process. As a result analysts also get an idea when to stop training.

4.3 Integration

Both presented approaches are intended to be integrated into larger systems that might facilitate visual analytics approaches themselves. However, a variety of constraints must be considered to achieve this. Figure 4.18 depicts both analytic processes in context of their integration into the PatViz process. In order to make additional visual analytics approaches available within a parent system, the parent system itself must offer a set of well defined interfaces. As with EdgeAnalyzer, one or several views in the system become the data source of the add-on technique, and accordingly have to provide interfaces for the data they depict, for geometrical information, and for their alignment. Integrating techniques for classifier creation might be easier with respect to the visual front-end. Here the difficulties arise from providing scalable back-end services where trained classifiers can be stored to be (re)used during subsequent analyses.

Theoretically, it would be possible to realize these integrations as plug-ins that can be dynamically added. To take effect on a broader scale, however, it would be necessary that current visualization toolkits are developed into visual analytics toolkits with well-defined interfaces. These interfaces not only comprise those for bi-directional data exchange, but also mechanisms for provenance recording, generalization modularization of automatic methods, and interfaces for scalable back-end integration. Considering the large number of different application areas

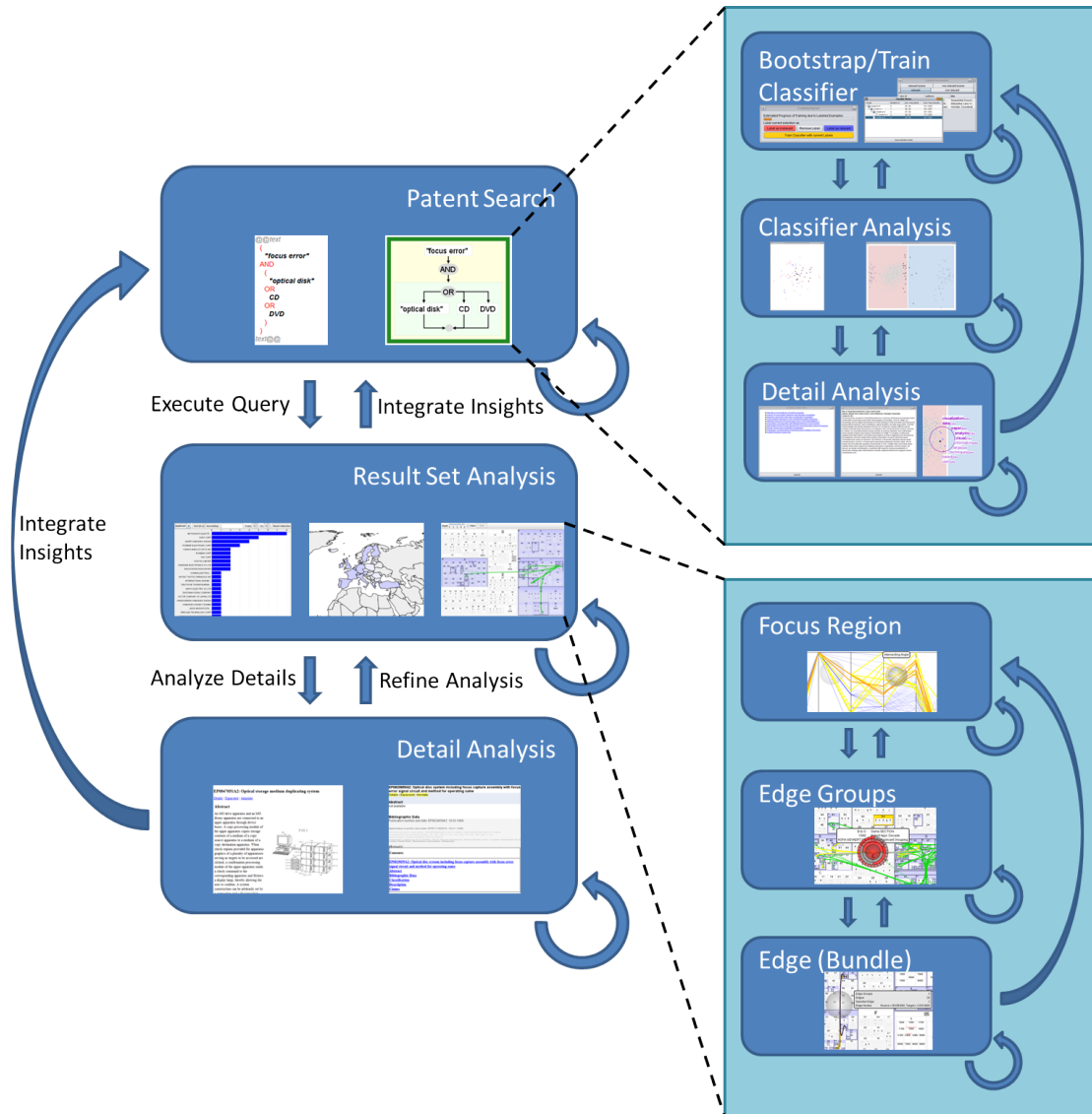


Figure 4.18 — Integration of the classifier creation process and the edge exploration process as realized with EdgeAnalyzer into the overall patent analysis process.

and their very specific requirements in terms of data characteristics and data size to be analyzed, e.g., streamed data from social networks or simulations as opposed to large, less rapidly changing data sources, it is very difficult to find a general scheme for such a holistic integration approach. On the level of subtasks, as with the methods presented in this chapter, it might be feasible.

The possibility to integrate solutions addressing subtasks into other visual analytics systems makes them scalable from the analysis perspective. Besides the mentioned issue many other scalability aspects play an important role in visual analytics and are discussed with details of the presented approaches' scalability characteristics in the next chapter.

Scalability, Provenance and Reporting

Scalability was identified by [Thomas and Cook \[2005\]](#) as a major challenge for developing visual analytics approaches. They identify the following scalability aspects to be addressed: *information scalability*, *visual scalability*, *display scalability*, *human scalability*, *software scalability*, as well as *security*, *privacy* and *globally distributed analysis tasks*. According to [Thomas and Cook](#), information scalability relates to extracting relevant information from large data sources and presenting it in an audience-tailored manner. Representations and interactions spanning various scales of information are subsumed under this category as well. Visual scalability refers to the depiction of large amounts of information through visualization, while display scalability requires accounting for different display sizes and qualities, e.g., from mobile devices and desktop computers up to large display walls. In order to achieve human scalability, visual analytics systems have to account for collaborative approaches. Software scalability relates to integrating aspects of automated data analysis and interactive visualization to support scenarios that are difficult or impossible to support without this integration. Many visual approaches have to deal with some, if not all of the listed scalability issues.

Different schemes have been proposed to characterize visual analytics challenges that are related to scalability, which offer different perspectives on the problem area. Characteristics such as *problem size*, *visualization richness*, *interaction pace*, *level of computational analysis*, and *comprehensiveness* can also be employed to

look critically at visual analytics solutions.¹ Those will be referred to as well during the following sections, since these characteristics are interesting in the context of patent analysis, too. During the development of the field in recent years, additional challenges and fields of application have been identified [Thomas and Kielman, 2009; Wong et al., 2012].

The approaches presented in this thesis were developed with the aim to explicitly address certain scalability aspects and related challenges that play an important role in the patent domain, as well as for document retrieval in general. The addressed scalability considerations roughly follow the categorization of Thomas and Cook [2005], but are adapted to the properties of document retrieval and analysis. In addition to general scalability considerations, ideas for deriving analytic provenance, history recording of analytic processes, and collaboration aspects are presented in the following. Some of the design facets described in the subsequent section have already been realized in the prototypes presented in Chapters 3 and 4. However, some ideas are introduced as prototypes indicating promising solutions, but have not yet been developed in depth.

This chapter is partly based on the following publications:

S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative Integration of Visual Insights during Patent Search and Analysis. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*, pages 203–210, 2009

S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative Integration of Visual Insights during Scalable Patent Search and Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):557–569, 2011

A. Panagiotidis, H. Bosch, S. Koch, and T. Ertl. EdgeAnalyzer: Exploratory Analysis through Advanced Edge Interaction. In *Hawaii International Conference on System Sciences (HICSS 2011)*, pages 1–10, 2011

H. Bosch, D. Thom, M. Wörner, S. Koch, E. Püttmann, D. Jäckle, and T. Ertl. ScatterBlogs: Geo-spatial document analysis. In *IEEE Conference on Visual Analytics Science and Technology (VAST 2011)*, pages 309–310, 2011

¹ These characteristics were proposed by Helwig Hauser in his keynote talk at the i-KNOW 2012 conference in the TAVA special track.

continued...

F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual Classifier Training for Text Document Retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012

5.1 Scalability Aspects of Patent Literature Analysis

As emphasized in Chapter 3, searching and analyzing patents are complex problems, while at the same time the risk of missing an important document is high and can have severe economic consequences. In this regard, information scalability, and specifically extraction of relevant documents is of particular importance. Therefore, all approaches presented in this thesis aim at improving retrieval scenarios, with respect to completeness, speed, and analytic coherence, meaning that whole analytic cycles and feedback loops are integrated seamlessly and can be carried out in an interactive visual manner. In relation to problem size, the amount of data that has to be considered during patent analysis can be seen as being large: databases of the EPO, for example, contain more than 70 million patent documents². However, taking into account a specific information need and the number of patent documents a professional has to deal with during a corresponding analysis, typically not more than a few thousand documents have to be handled after an initial search. These are further drilled down to a set that can finally be analyzed in detail, i.e. read. It requires experience and iterative approaches to retrieve a manageable number of documents to be analyzed further in subsequent steps. Therefore, automatic methods in the form of retrieval systems are needed. Besides Boolean retrieval, which is still the most commonly used retrieval method in patent search, other retrieval back-ends can be employed (see Chapter 3). Clearly, data scalability and task scalability are significant issues in patent literature analysis.

In order to support analysts with a coherent interface, all of the search back-ends are – again using Boolean combination – made accessible through a coherent visual interface in order to account for visual scalability. EdgeAnalyzer was developed to address the same problem on a much smaller scale and within a specific type of visualization. However, it also facilitates Boolean combination of findings and advanced filtering methods. The approach which allows analysts to create

² <http://www.epo.org/searching/free/espacenet.html>

task-tailored classifiers deviates from the retrieval back-ends of PatViz. Here, the users can determine the goals of an automatic technique according to their requirements. Since it can be integrated into Boolean retrieval approaches as well, it can be an interesting choice for cases where keyword search, for instance, does not suffice.

In the following subsections scalability aspects of the three presented approaches are discussed. Each of the sections focuses on one scalability issue and how it is addressed respectively.

5.1.1 Software and Data Scalability

A key concept of the PatExpert system is its distributed architecture with several data sources, search engines and services. The two central components in this architecture are the *broker* (Figure 3.6) for integrating different search engines as well as their data sources, and the PatViz front-end steering the search and analysis process. With the method for interactive visual classifier building, an approach has been developed that makes the analysts not only users of retrieval systems which rely on preprocessed documents but creators of new automatic methods. Its integration into larger retrieval systems has not been accomplished so far, but it is designed in way that accounts for this option. EdgeAnalyzer has been created to be extensible, and with future employment in a large variety of line-based visualizations in mind.

PatViz

The PatViz front-end is part of a larger approach developed in the PatExpert project and represents one module in a distributed system facilitating a number of services. Two principal types of communication interfaces connect PatViz to PatExpert's back-end system: a search request interface and the document request interface (see Figure 3.6). The difference between both request types can be described as follows: search requests define a constraint without available knowledge about concrete documents, thereby formalizing the users' information needs, while document requests are used to retrieve the (meta)data associated with given patent numbers. As a response to the search request interface only a list of patent document IDs are returned to the visualization front-end and according to the data needed for visualizing patent aspects, additional information is requested through the document request interface. This means, for example, that the search engine for metadata does not need to cope with the textual content of the patent document, and the semantic search engine can ignore the bibliographic data.

As a benefit, the lean data-exchange mechanism makes the system applicable to other domains requiring iterative analysis of unstructured data (other document types, images, etc.) and related metadata. For the visual components, however, further mechanisms have to be planned for to make the system adaptable (see Section 5.1.2 below).

The described design decision has not only been made to achieve a greater separation of concerns, but also to reduce the initial data transfer, since it is not known which data associated with a document is needed by the visualization module later on. Through this simplification, the network load is reduced and the latencies of request response cycles of both request types are low. In addition to that, combined queries can be distributed to the corresponding different search back-ends and run in parallel. Depending on the query structure, the system throughput can even be increased by letting the query broker send back partial results of those query parts that are connected disjunctively. A system can be created that does not have single points of failure, by making the back-end services (including the broker) redundantly available. Even parallel execution of subqueries would theoretically be possible, but neither feature has yet been implemented in the prototype system. Furthermore, the described architecture provides a large degree of freedom with respect to the system's extensibility. This applies to additional back-end services, e.g., new query facilities or analysis services and also to new visual methods that can be incorporated into the front-end as described in Section 5.1.2.

All back-end components are connected via XML³-based Web Services to allow their easy integration into a common web infrastructure. Within these components, the broker service takes a key role. If a query will be executed, an XML representation of the query is generated to be encapsulated in a Web Service request. As sketched in Figure 3.6, the broker is responsible for interpreting this XML request, decomposing it into the different subqueries, querying each search service, merging the results, eliminating duplicate results, and delivering them back to the visualization module. In addition to the Boolean requests mentioned earlier, the broker is also capable of handling fuzzy queries and has mechanisms to incorporate user feedback for adjusting the score of result set entries. For more details on the query broker service see the description by Codina et al. [2008].

The visualization component PatViz is modularized into different parts which can be mapped to the different stages of the visual analytics cycle in the patent analysis process. Search/query stage, results set analysis stage, and detail examination stage provide different visualizations to give users the opportunity to take different perspectives on the problem space within the corresponding stage. All views define interfaces that allow for their interoperability with each other and, at a coarser

³ <http://www.w3.org/XML/>

level, with the specific requirements of the back-end. The interfaces needed for the visual components will be discussed in the following paragraphs.

The visualizations that are available for analysis have been designed in a flexible way, because a bar chart, for example, can be used to display any kind of object as long as there is some nominal and some scalar data available that can be related to each other in a meaningful way. Each view in PatViz therefore provides capability information with respect to the type of data it can display. On the one hand, this allows for dynamic checks on the suitability of data routed to the view, given that the data is self-descriptive enough. On the other hand, extensibility with views for other scenarios taking the same type of data is guaranteed.

In PatViz this is solved by using an additional object-oriented model of the data available for visualization tasks in the front-end. Unfortunately, knowledge of the model is not enough to create abstractions for user-steered interactive selections in order to provide advanced brushing & linking mechanisms, because the semantics for the selection is not taken into account. Highlighting data in other views can be accomplished by considering only primary domain objects or their ID respectively, as it is done for search responses. Well-defined selection semantics are necessary for the creation of complex insights and for reusing them in subsequent analytic iterations though (see Section 3.3). The views for result set and detail analysis are easily exchangeable, as long as they can provide information about what kind of data they are capable of displaying, which kind of data they are displaying as concrete instances, and how the selection semantics are defined on them. Defining specific semantics is often difficult, because, depending on the selection operations available within a view, ambiguous interpretations of the constraint that should be applied are almost unavoidable. Visualization components must therefore implement three interfaces, including a capability interface, an interface that makes the displayed data accessible, and an interface that provides information on what kind of semantics is associated with selection gestures. To realize them, an object-oriented data model (see Aligned Data in Figure 2.3) is employed within the client and façades [Gamma et al., 1995] are provided for the single views to allow them to access the data in the way they need: in a set or table-oriented manner, for instance. With the object-oriented domain model and the visual components knowing which kind of data they need, lazy loading of model data can be achieved. A minimal object model solely consists of a set of primary object IDs. Because views know what data they are meant to display, activation of such a view can trigger the loading or computation of the corresponding additional data. Even more fine-grained loading strategies are possible, e.g., when zooming into the details of a view; although this carries the risk of unresponsiveness to user interactions, since possible latencies emerging from the back-ends web infrastructure are difficult to control. In some cases, such as loading of patent texts or images into separate

views as a response to user interaction, this disadvantage is acceptable. The palette of available visual analysis views can be adapted according to the size of the patent set under analysis, since the requirements regarding the amount of data needed for each view is known before. At least a user commitment can be requested before perspectives are activated which would have negative effects on the client's performance.

The visual perspectives used for query creation do not only need to fulfill the contract to the described internal interfaces, but must also obey external restrictions, which are dictated by the expressiveness of the available back-end search facilities. Extensibility of the query views requires an additional mechanism, which allows for the integration of further visual search perspectives. In PatViz this is technically realized through a hierarchical parser/generator module as depicted in Figure 3.7. The module is capable of parsing/accepting the textual as well as the visual representation of the combined query language. If expressions or visual constructs of a specific sublanguage are encountered, they are forwarded to the corresponding subparser. Both representations are updated automatically, if the changes applied by a user to one of them are syntactically correct, which is not guaranteed in the case of modifying the textual query. Due to the hierarchical parser/generator concept, the query system can also be adapted to other domains or extended by additional search facilities by adding new subparsers that can be created semi-automatically. Such an approach is very helpful for the automatic generation of the textual part of the query. However, the visual representation and the representation that is sent to the broker and finally to the corresponding search engine, still have to be developed manually. The development of solutions which also facilitate the creation of interactive visual metaphors using a descriptive formalism have not been fully developed in PatViz yet. First tests for Semantic Web data with techniques borrowed from the Fresnel display vocabulary⁴ indicate that this is possible.

User-steered Classifier Creation

With the approach for classifier creation (see Section 4.2), which has not yet been integrated into larger systems, possibilities have to be provided for employing it in distributed retrieval environments. The classifier training and its application are both scalable enough to run on a normal desktop client machine, but its integration would break with some of the the design decisions mentioned above. A solution running on the server could be created straightforwardly, since the approach for classifier training is designed in a scalable way, using a Lucene repository as its back-end, which is also employed for storing and searching patent documents in the PatViz back-end. However, this would require that analysts connect directly to the

⁴ <http://www.w3.org/2005/04/fresnel-info/>

server in order to build or update classifiers. In case a setup is desired that permits only certain specialists to build classifiers, this might be a viable solution.

If a broader group of analysts should be enabled to create classifiers, it must be available as a client-solution and would require that a document repository holding the documents to build the classifier is available on client machines. Since the first step in of the presented approach for classifier building is bootstrapping, the set of relevant documents decrease greatly in comparison with the whole document set. It would therefore be possible to transfer a collection of vectorized documents to a client system, if the connection to the document server(s) allows it. Since non-relevant documents are important classifier training as well, sampling strategies could be used to retrieve those documents, in order to transfer a big enough, representative number of documents. A subset of a document corpus is sufficient for classifier training with visualization-centric methods. If necessary, the sampling step could be repeated for each training iteration. An alternative strategy for handling large training sets is to restrict the classifier view to a subset of the whole corpus by displaying only the most uncertain documents (e.g. 5,000 of them) together with the training documents, and to hide all other documents.

Created classifiers must then be transferred back to the server-side retrieval facility, where they are available for subsequent use and can potentially be shared and used by other analysts too. Since the classifier itself is solely defined through the hyperplane, the model that has to be transferred is relatively small and no scalability issues arise from this step.

All client-side solutions entail the benefit of moving computational effort for classifier training and the creation of the visual perspectives from servers to client-machines. For distributed setups, however, enabling analysts to create their own tools for document retrieval clearly requires a tighter integration of services and necessarily increases the amount of data to be transferred. Fortunately, the active learning principle or the proposed adaptation for user-steered classifier building diminishes this effect greatly, since it is specifically designed to train classifiers more quickly, and reduces the number of documents an analyst has to label. If a good set of examples can be determined, e.g. through sampling, the number of documents to be transferred is reduced accordingly.

Also hybrid solutions are conceivable but depend on the size of document repositories and data characteristics. Users who are interested in building their own classifiers could be provided with a mechanism for bulk download of the document repository, and transfer back only created classifiers. Such a solution would be feasible only for medium-sized repositories fitting in client machine's storage, and holding documents that do not change dynamically over time, since the latter might require adaptation of classifiers and consequently frequent updates of the repository.

Classification training as well as applying the classifier to large document sets is also scalable through the use of linear support vector machines (LSVMs). The proposed solution is based on the LibLinear [Fan et al., 2008] library, which is suitable for solving LSVMs for high-dimensional data vectors and large training sets very fast. Hence, it is perfect for dealing with text documents that are represented as large vectors in a so-called ‘bag-of-words’ model. The integration of Lucene and LibLinear ensures high scalability of the presented approach with respect to corpus size. Well-built classifiers also adhere to the ‘create once, apply often’-idea and can be reused. Classification also has to take into account highly skewed classes. If, for example, the number of ‘relevant’ documents in a large corpus is very low compared to the number of ‘non-relevant’ documents, even active learning strategies might take a long time to converge and are not guaranteed to succeed. Furthermore, sampling strategies are likely to miss relevant documents. With the bootstrapping approach these problems are addressed, since good keyword queries guarantee the inclusion of relevant documents in the training sample.

EdgeAnalyzer

The EdgeAnalyzer approach can be employed within any edge or link-based views different from the patent co-classification scenario. It is scalable in this regard from a software engineering perspective. However, it relies on the interfaces for geometric edge properties/intersection tests and access to edge metadata if corresponding grouping mechanisms are to be used (see 4.3). It has been specifically designed to work on large numbers of edges of densely connected graphs. Furthermore, it can be easily extended with respect to employed grouping and visualization mechanisms, since the respective interfaces are provided. However, implementors of these interfaces have to acknowledge the the amount of edges to be inspected within the intended tasks and take care to provide grouping and visualization methods that are fast enough from a computational perspective. Otherwise, fluid interaction is not possible and will restrain analysts’ explorative tasks. Figure 4.7 shows its application within a multiple coordinated view demonstrating this flexibility.

Regarding the abovementioned alternative characterization of visual analytics problems and approaches, problem size and comprehensiveness are most closely related to the scalability issue described above. The problems tackled with the approaches described in this thesis can be categorized as medium to large sized (hundreds to tens of thousands ‘data objects’ must be processed), from a visualization and data perspective. If the overall size of available information in patent repositories is considered, the high-dimensionality and diversity of the data, as well as the complexity of the documents that have to be interpreted by human specialists, the problem size could also be described as huge. With respect to comprehensiveness, in particular the EdgeAnalyzer technique and the method for classifier building can be

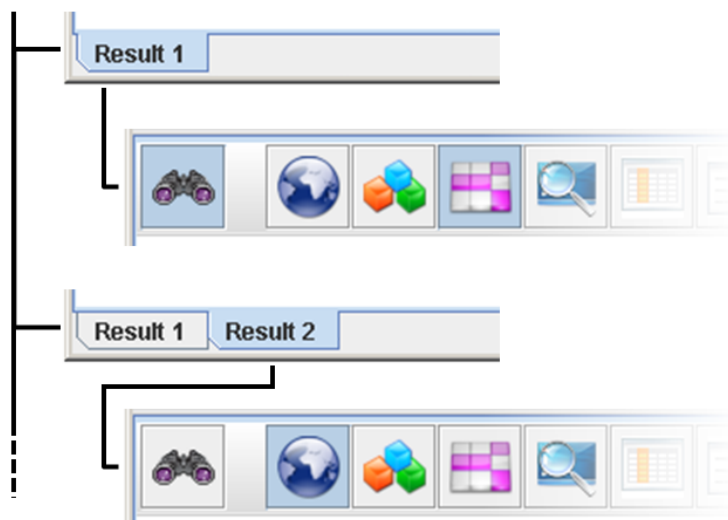


Figure 5.1 — For each query iteration a new tab is generated depicting the state of the corresponding query and result set views.

considered as being flexible. While the overall approach of PatViz is restricted to a specific domain, some of its components, e.g., the selection management technique, have successfully been exploited in the context of other developments, too.

5.1.2 Visual and Interaction Scalability

The efficient usage of available display space is an important aspect of visual scalability, which can be improved through different visualization approaches and interaction methods. As many other techniques, the approaches presented here address this aspect on a per-view level by providing common interaction techniques, such as zooming, panning, and scrolling for most of the single views available in the user interface. Visual aggregation and abstraction of patent information, are uses as well for increasing scalability. This section highlights the measures taken to improve usage of display space for the multiple coordinated view approaches in PatViz, user-steered classifier creation interface, and the impact of EdgeAnalyzer on scalability.

PatViz

PatViz provides different tabs to hold the multiple coordinated views for each iteration of the analytic cycle, namely the query view, result views, and detail views. On (re)submission of a modified query, a new tab containing the set of

views and a copy of the submitted query is generated (cf. Figure 5.1). Former analytic results remain available and can be accessed at a later point in time. All views can be detached from the main PatViz window to allow for the concurrent use of different views that do not reside in the same tab, This is also beneficial if several monitors are available, helping to avoid situations where views occupy different displays at the same time with distracting monitor frames in between. Because drag gestures are used either for moving insights from one view to another or for highlighting a selection, overlapping windows pose a problem. While most windowing systems provide techniques to perform dragging between overlapping windows, visual inspection of all views used at the same time is particularly beneficial for complex analytic tasks, as long as enough display space is available. However, good alignment of a variety of perspectives within a multiple coordinated views environment, in order to prevent occlusion, is not easy to achieve. Especially if the view's constraints, such as minimum perceivable size and aspect ratio, are taken into account. PatViz's solution is quite simple but works well for arranging multiple views in a delimited rectangular area, i.e., a tab. It applies an ordered⁵ and squarified treemap layout [Kandogan and Shneiderman, 1996; Shneiderman and Wattenberg, 2001; Bruls et al., 2000] to place the views within one tab. The hierarchy for creating the treemap is shallow and simple; it just consists of the tabbed parent window and just one child level, containing all currently activated views. Users have two options for ordering the views: either in sequence of their activation or in the same order as the corresponding buttons in the tool bar. The idea behind this kind of placement is to help users to keep their mental maps, thereby allowing them to track the available views more easily if new ones are added or old ones removed. Furthermore, the squarification of the views is an approximation of a good aspect ratio for most visualizations, and users can still use interactive zooming and panning to adjust them to their needs. One other option tested was to change the views' orientation automatically, reducing structural zoom, and even recomputing the layout of view content wherever possible. However, it quickly became obvious that such a massive change in multiple views created too much confusion for the users.

User-steered Classifier Creation

Visual classifier creation also uses a multiple coordinated view system in order to offer analysts different perspectives on the classifier's state during each training iteration, where views are handled similar to the PatViz approach. The employed views either use point-based, scatter-plot-like techniques, aggregated views, or detail perspectives. Scalability is not an issue for the latter two. For the scatter-plot-based

⁵ At the moment the pivot-by-middle method for the ordered treemap layout is employed.

views zooming and panning interaction is available. In the case of classifier view, about 50,000 documents can be represented without experiencing performance issues. Of course, this is already more documents than an analyst could reasonably inspect in detail for classifier creation. With the employed interaction mechanism in the form of the term lens this problem can be decreased, but the effect with the highest impact on scalability clearly comes from active-learning-inspired interaction approaches and task-tailored layout. The term lens summarizes documents under the lens by showing the 10 most frequent terms with respect to document frequency. The layout of the cluster view directly depicts the classifier's decision border giving users a good idea where best to label documents. Furthermore, the layout of document nodes around the decision border reduces overlap and aims at simplifying interacting with them. This lets users explore densely populated regions quickly, while still providing feedback on the underlying documents' content. In the cluster view, the number of depicted document items is restricted to the 100 most uncertain, in order to guarantee good labeling impact and less training iterations.

As described in Section 4.2, employing user-defined classifiers is not intended as a stand-alone retrieval method, but should complement other retrieval approaches. Such an integration, e.g. into PatViz or other visual interactive retrieval environments would be possible by extending the Boolean search interface to include user-created classifiers as well. With the introduction of a new classification node, either existing classifiers could be made available for selection or an option to create a new one, thus calling the interface for classifier creation.

EdgeAnalyzer

The EdgeAnalyzer approach is intended to increase scalability of edge and line-based visualizations. This can be achieved in scenarios where edges have been bundled, as well as in densely-connected line-based views. However, it is particularly useful if analysts are required to work with edges or their corresponding metadata properties directly. In the case of bundled edges, details are reduced for the sake of a better overview. This can be an improvement, since visual clutter, as caused by many edge-crossings, not only restrains overview, but also prevents analysts from perceiving details. EdgeAnalyzer equips analysts with the possibility to explore details in visualizations that have been improved through edge bundling, thereby adding an additional benefit to such views. In case of link-based visualizations without bundling, it facilitates detail inspection as well, and can be used to create 'local overview', meaning that focused links or edges can be bundled in order to increase clarity in a region.

If seen from the point of view of visualization richness, all categories from simple to rich are covered with the presented approaches. While rather primitive views are

employed in the PatViz approach and for user-steered classifier creation (charts, scatter plots and maps), their interplay and linkage within the coordinated views environment adds a certain expressiveness to the overall visual approach. Some of the presented single views can be regarded as advanced or even complex: the EdgeAnalyzer's approach utilized on top of the co-classification treemap would be such an example.

5.1.3 Platform Scalability

In companies, patent search is often carried out in research and development, strategic planning, and in other contexts involving different departments. Platform scalability therefore plays an important part, if heterogeneous IT-infrastructures and different working environments like workstations dedicated to patent analysis, common office computers, or laptops should all be usable for analytic tasks. Open communities, such as researchers and other groups interested in patent analysis, also rely on easy access to such a system. Visualization systems and applications that are usable via standard web browser technologies or which can be easily deployed via the Internet greatly affect human scalability as has been shown, e.g., with ManyEyes [Viégas et al., 2007]. Different systems with different capabilities in terms of processing and graphic performance can be addressed if flexibility for the visualization techniques is planned for. The PatExpert back-end allows for flexible solutions concerning the location of data storage and execution of search requests. To achieve a similarly flexible solution for deployment of and access to the PatViz system, a variety of prototypes have been created in order to evaluate different web-based deployment approaches.

The rich client PatViz front-end, as described in this thesis, can be easily deployed to any computer with a current Java Runtime Environment using the Java WebStart technology.⁶ This was especially useful to demonstrate the prototype to remote partners in the consortium. Even without an available installation of Java, large parts of the PatViz system can still be used. In this case the application runs completely on the server component. To interact with the system, the user accesses the website with a query interface and dynamically updated images of the result set views. Whenever the user clicks on an image, the event is registered and transmitted to the server together with the location of the click, using asynchronous JavaScript. The server reacts to the event and renders the result into an image to be sent to the user. While this works well for click events and keystrokes, drag operations suffer from the transmission delay. Currently, this solution comes at the cost of reduced interaction support.

⁶ <http://www.java.com/>

In between the two extremes, search requests and result set views are also executable in Java Applets. This allows for their integration into existing web environments like forums and Wikis. An approach for collaborative patent search using visualization applets in a Wiki is described in [Giereth and Ertl \[2008\]](#).

EdgeAnalyzer as well as the approach for user-steered classifier creation are also based on Java Technology, which makes them applicable on various platforms. In the case of the approach for classifier building, the dependencies on other client-side views are only marginal. As long as a possibility for starting the interface and means for integrating trained classifier models into existing systems are available, the method can be used independently.

5.1.4 User, Task and Process Scalability

Scalability issues also arise with respect to those who use analysis systems and their specific tasks. On the one hand, users with different levels of knowledge and experience want to use the analytics system to accomplish a variety of tasks. On the other hand, very complex tasks, such as patent analysis, can be sped up, made more reliable, or even simplified through providing visual analytics approaches. This, however, is often also reflected in increased visualization and interaction complexity that has to be learned before it can be exploited successfully. This section will give a brief overview of what has been done to support both aspects in the discussed approaches, with exception to collaboration aspects discussed separately in Section 5.2.

PatViz

In PatViz it is possible to define presets for result set visualizations that are well-suited for specific tasks including prior art search, freedom-to-operate search, etc. However, users can select more than one or alternative views if they wish, and are not prevented from using arbitrary views either. This feature has been explicitly requested by patent specialists of the PatExpert consortium. The combination of view and layout presets with parameterized forms allows for a task- and user-oriented customization of the PatViz system.

Varying user experience levels are addressed at different stages of the analysis cycle as well. This is valid for the query formulation as shown in Section 3.1 and for the result set analysis, which provides both very common views, e.g., charts, and more complex ones for experienced users. This makes it possible to use the parts of the system with lower complexity from the very beginning, while having the option of switching to complex views and workflows at a point in time when basic features are mastered.

The technique for selection management further increases task scalability, by externalizing subtasks, such as selection operations in single views or their boolean combination. As a result, analysts are freed of keeping in mind what they found out in previous steps, in the form of interesting sets, or (in)validated hypotheses. This is an important feature for approaches that support analytic reasoning processes:

“Reflective thought requires the ability to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward, sometimes backtracking when a promising line of thought proves to be unfruitful. The process takes time.” – Norman [1993]

The visual selection management technique is also scalable with respect to the possibility of creating different selection and filter paths in parallel. Users can always switch back, choose and change one of the (intermediate) results that have been created, and check the effects of their refinements by re-displaying them in linked views.

User-steered Classifier Creation

In case of classifier building, the evaluation (see Section 6.1.5) showed that users who are not well trained, or have not understood certain aspects of the approach, are at risk of creating bad classifiers with the user-steered method. In some cases much simpler methods, employed during the comparative user evaluation, turn out to deliver more robust results, even if this comes at the cost of decreased trust in the created classification models. This shows again that offering methods acknowledging users' expertise in carrying out a task is of great importance.

Visual classifier creation can also take more time than with the simple method. Here, the test subjects spent large amounts of time on exploring the documents. While techniques facilitating interactive, visual exploration⁷ are often described as being superior to less interactive approaches, they carry the risk of requiring more time for such a task. In order to increase exploration speed, the term lens was introduced to diminish the effect. Again, spending too much time on exploration is likely to decrease during training. Developing meaningful and successful task workflows is therefore a key issue for all complex analytics tasks. Ignoring the workflows and principles of classic analysis (if available in a field) tends to reduce the acceptance of visual analytics approaches and leads to a decrease in analytic performance, too.

⁷ Here only approaches are referred to that facilitate explorative approaches that are meaningful for the task at hand, i.e. when an analyst's information need is unclear and has therefore to be shaped first through explorative tasks.

However, if patent experts are well-trained in user-driven classifier creation, it has the potential to scale well with their requirement of building trust in search solutions, which is the main argument for employing Boolean search. Here, visual perspectives support users by letting them interactively refine and build classifiers, without the need to understand all technical details of the machine learning technique. By visualizing the classifier's state and its mode of operation adequately, users can build more trust in such automatic solutions and integrate them into their existing search strategies, as opposed to out-of-the-box solutions, whose behavior is hardly understood and difficult to assess.

Facilitating interactive feedback loops tailored to analytic workflows can greatly help to increase efficiency in carrying out analytic tasks. Feedback loops for speeding up patent literature retrieval have been described in detail in the Chapters 3 and 4 and are therefore not discussed again in this chapter.

5.1.5 Scalability Conflicts

Some scalability aspects are inherently conflicting. While software scalability can be addressed by using single generic visualizations in order to reuse them in different scenarios, specialization of views with respect to the problem domain can improve analytic performance. Flexible approaches, allowing for multiple ways of performing an analysis can increase a technique's power greatly, but are at the same time more difficult to learn and understand for analysts.

Another conflict exists between the creation of scalable distributed multi-tier software systems and the aim to interact with and influence automatic back-end systems in order to realize task-driven and user-adapted approaches. The presented technique for classifier building is such an example. While retrieval quality can be increased, software scalability is reduced at the same time, since the decoupling of systems is at least reduced. As can be seen from the modified information visualization model (see Figure 2.3), potentially all steps in the processing pipeline can take part in feedback loops, and data flow between these step is bidirectional as well. An increase in dependencies between components can lead to monolithic, tightly coupled approaches. Therefore, particular care must be taken to keep visual analytics approaches scalable with respect to clear separation of concerns and modularization. The development of back-end systems that acknowledge back-flow of information and other properties important in the context of visual analytics is still at its beginning. Chapter 6 in [Keim et al. \[2010\]](#) offers an extensive discussion of infrastructure issues in visual analytics.

Even if specialized visualizations are designed to be compact, precise, and scalable, they might require more user knowledge to be interpreted correctly, hence they initially decrease user scalability in terms of the time needed to learn using them.

For visual scalability the goals may be contradictory as well. On the one hand, the system should organize views in a space-efficient manner, while retaining ordering, aspect ratio, and zoom levels; on the other hand, single views should be placeable on auxiliary displays. Additionally, visual and interaction scalability are closely tied to task scalability. If tasks are sufficiently complex, interaction and/or visualization complexity will increase also in the number of employed views, their complexity, or their need for advanced interaction techniques.

Finding appropriate solutions for these conflicts is possible in some situations, but typically comes at the cost of suboptimal analytic efficiency, steeper learning curves, tightly coupled systems, and incoherent workflow support. In the end, visual analytics approaches must always find a compromise regarding conflicting scalability requirements, depending on the focus of the tasks to be achieved and the processes to be supported.

5.2 Collaboration, History Recording, and Analytic Provenance

Collaboration is seen as a very important aspect of achieving user scalability. In this context, collaboration does not refer to the integration of automatic methods, human reasoning, and their interplay, but to collaboration between users, who are involved in the same analytic effort. This section discusses collaboration along with analytic provenance and history recording. There is an inherent connection between the latter two, since provenance information has to be attached to search/analysis history. Collaboration approaches, reporting, and also evaluation, as discussed in the next chapter, can benefit from this provenance information.

The need for capturing analytic provenance for visual analytics approaches was already formulated by [Thomas and Cook](#). With VisTrails [[Callahan et al., 2006](#)], an extension to VTK⁸ was proposed that allows for recording ‘visual provenance’ information on the visualization process and data manipulation steps. Logging of users’ interactions was employed to evaluate how users explore information in InfoVis systems and perform reasoning when using visual analytics systems [[Pohl et al., 2010](#); [Dou et al., 2009](#)]. Other suggestions for dealing with analytic provenance were formulated by [Gotz and Zhou \[2009\]](#). They propose HARVEST, a visual analytics prototype system that combines manual and automatic provenance recording. Furthermore, they define a taxonomy for user interaction on different abstraction levels: tasks, subtasks, actions, and events. If compared to the HARVEST approach, PatViz captures analytic steps on a coarser level and focuses on

⁸ <http://www.vtk.org/>

the action and subtask levels that [Gotz and Zhou](#) identify as the ones critical for capturing analytic provenance. [Heer et al. \[2008\]](#) present a design space analysis of interactive, visual history tools. With the ‘CzSaw’ visual analytics system, [[Kadivar et al., 2009](#)] introduce editable and replayable history recording aimed at making analyses more comprehensible and reusable within other analyses.

PatViz’s visual selection management as described in Section 3.2.9 enables users to save and combine selections in order to actively steer the search and analysis process. The resulting graph is stored for later (re)use. By applying this selection management technique, analysts automatically document steps of their analytic sessions, thereby generating an abstract form of analytic provenance data that reflects mainly high-level analytic tasks. Furthermore, saved graphs can be applied to new result sets, serving as ready-to-use analysis steps. Saving the various stages of a performed analysis may also be of value with respect to accountability. However, some patent researchers have expressed concerns with regard to such a functionality. The reason for this is the legal practice in some countries to increase the fine for infringing patents if there is evidence that someone in the sued organization had prior knowledge of the patent in question.

In addition, every analysis cycle, i.e. analysis of result sets, query modification and subsequent requesting of a new result set, is recorded in order to enable analysts to access previous findings and to perform undo steps if a hypothesis turns out to be unsuccessful regarding the aims of an analytic task. This feature can be exploited in collaborative scenarios, since discrete improvement steps of the query are easily traceable for other analysts. This combination of global query management and local selection management has the potential to serve as documentation helping analysts to continue work commenced by others, thereby enabling collaboration. Further, it can also be used to teach new examiners which steps to take during a patent search.

For user-driven classifier creation, a similar approach has been taken, even if it is less comprehensive and does not reflect single sets of documents. Here the approach could be seen as an even more abstract form of history recording. Its main purpose is to serve as a reminder for the classifier-creating analyst, which training steps were applied to develop a classifier and how training steps influenced subsequent versions of classifiers. The amount of labeled documents for each class, as well as the balance of the classification outcome are recorded and made available. As part of the classifier history view, it also facilitates undo and redo operations, enabling analysts to start over from ‘older’ versions of the classifier. Besides, the classification history can be exploited in collaborative scenarios. Analysts can retrain available classifiers to adapt them to new situations. Classifiers can theoretically be shared easily, and others can exploit the effort previously invested by a colleague.

To ensure that all necessary steps have been taken, analysts often employ templates that can be followed for common tasks. In PatViz, these templates come in the form of parameterizable queries. A parameterized query consists of a template query and a list of variables. These variables have fixed positions once the query is saved. Upon loading the query, the user will be requested to provide a value for each parameter. This mechanism allows for easy creation of forms that can start an initial query for a special patent search task or at least provide templates and basic strategies for common search objectives. Parameterized queries can be seen as a very short and condensed script on how to address a certain search problem.

The means created for representing and recording analytic steps so far are still rough prototypes. In order to use them effectively in larger systems, a more fine-grained recording of actions has to be developed. However, the advantages the shallow recording already offers can be demonstrated with the approach for reporting described in the following section.

Automatic Report Generation

Because visual analytics tasks and systems increase in complexity, reporting the gist of findings becomes an important issue. As one of their top 10 observations for visual analytics systems [Thomas and Kielman \[2009\]](#) consider reporting as being “*Critical to analytical assessment [...]*.” Analysts themselves might not be authorized to make decisions. Consequently, the findings must be presented adequately to the person or group of persons in charge – the decision makers. A similar need occurs if there is a team of analysts working on larger problems, and intermediate results must be distributed among them effectively to leverage collaboration⁹ as discussed above. Furthermore, reports of analytic tasks and findings can be a valuable resource during evaluation, archival, and future reuse of analysis results. Interesting thoughts on analytic reporting and possible directions were discussed in [Chinchor and Pike \[2009\]](#) and [Lipford et al. \[2010\]](#) describe an approach for letting analysts track and store their analytic steps in order to help them remember and reconstruct their reasoning processes. This could be seen as a means for ‘self-reporting’, as is the case with many of the related approaches that were mentioned above in context of provenance recording.

Creating analytic reports manually, without (semi)automatic support can be a tedious task. Firstly, visual analytics methods often lack an export option for data extracted during analyses, which is at least true for most of the visual analytics approaches presented in academia, despite the fact that such options are highly relevant in practice. Accordingly, a variety of commercial visual analytics systems

⁹ Presentation of analytic results can also be seen as a special form of collaboration.

already integrate reporting facilities, e.g., IBM's I2 Analyst's Notebook¹⁰ or Oculus's Sandbox [Wright et al., 2006].

Manual creation, however, forces users to perform additional actions to transfer the 'result data' into the reporting method, or, in the worst case, to transfer them manually by reproducing analytic insights and through indicating the important parts of the data. Additionally, it is difficult to judge a finding's importance before it was (cross-)validated, and other analysis tasks have been carried out. Thinking about reporting when carrying out the analytic task may hinder the analysis itself, because it interrupts an analyst's workflow. All methods for recording analytics tasks that require active, manual user interaction to store an intermediate result, to make a screen shot, etc., suffer from the mentioned issues, independent of whether they are integrated into an analytics system or not.

Secondly, it is quite natural that reporting comes into focus *after* an analysis has been performed. If no measures are taken to record analytic artifacts, e.g., interesting data (subsets), automatically during the analytic process, analysts have to reproduce and retrace the process with the reporting aspect in mind, thereby taking notes and screen shots to document their work. The more traces that are followed and the more hypotheses that are checked during an analysis, the harder this gets. The more complicated the analytic task is with respect to the number of automatic and visual methods involved and the required human reasoning, the longer it takes.

The creation of automatic reporting mechanisms for visual analytics approaches is closely related to, or even depends on, a variety of different aspects, such as history recording, provenance, and collaboration, discussed in the previous section. Solutions for semi-automatic support of recording reasoning processes that could also be used as a basis for reporting have been suggested by Shrinivasan and van Wijk [2008]; Chen et al. [2009, 2010]. Active construction of analysis as proposed with facilities such as the selection management technique can be directly exploited for reporting the results of an analysis. Here, the analyst externalizes findings and combines them to test hypotheses, thereby creating a visually represented solution for an analytic task. In order to make findings identifiable at a later point in time, analysts can attach meaningful labels to them. Every time a node is created in the selection management technique (see Section 3.2.9) or existing ones are combined, the analyst externalizes a potentially valuable finding. Node creation and manipulation actions can therefore be seen as indicators for the creation of new insight and the fulfillment of an analytic subtask. After an analysis is finished, a graph of findings – ideally also holding the solution to the analyzed problem – is available. The idea is to exploit this graph for report generation just by identifying

¹⁰<http://www.i2group.com/>

the node(s) of interest. Since all sets are defined in a formal way, the analytic story of how they were created can be told automatically, by traversing the graph backwards to its root, starting from the node of interest. In this way all nodes and operations that were part of the refinement of the interesting finding are considered. The backward traversal of the graph acknowledges that the analytic result decision makers might base their actions on is presented first, but the whole process of its construction can be tracked and followed. As mentioned before, analytic processes are rarely linear. It is therefore no surprise that analysis graphs created with the selection management technique are typically non-linear as well (except for very simple cases). In order to tell a non-linear story, a medium had to be chosen that acknowledges this aspect, facilitates navigational exploration for following the construction of findings, and allows for combining textual and visual descriptions. HTML¹¹ is an obvious solution to fulfill these requirements: it facilitates non-linear browsing through hyperlinks, it can be viewed on a considerable number of devices making it rather platform independent, it can be shared easily through common world wide web infrastructure, and non-analysts should be able to work with it through common web-browsers as well.

During the VAST 2011 mini challenge ¹², which dealt with the analysis of geo-located microblogs, the selection management technique was employed, and the first prototype of an automatic reporting component has been added [Bosch et al., 2011]. By making geo-temporal selections of sets of microblog messages filtered through keywords, it was possible to create insight graphs to answer the questions posed in this challenge (Figure 5.2 depicts an overview of the approach). Of course, this is a simplified scenario on a test data set, but it should suffice to illustrate the idea of exploiting provenance information for automatic reporting.

One subtask in the mini challenge was to answer questions about the spreading of ‘disease’s symptoms’ and the consequences for the affected people. The data set to be analyzed consisted of about 1 million microblog messages. From these messages’ content and their geo-temporal distribution, the answers had to be derived. The approach aims at finding spatio-temporal anomalies of term usage in messages, and placing important terms as labels on the map through an automatic extraction procedure. The latter can be seen in Figure 5.3.

The approach turned out to be scalable enough for applying it in real-world scenarios, such as the analysis of Twitter ¹³ messages [Thom et al., 2012]. In both approaches it is possible to apply spatial, temporal, and content-based filtering for carrying out analyses and gain insights from microblog messages that can be externalized and combined using said selection management technique.

¹¹<http://www.w3.org/TR/html-markup/>

¹²<http://hcil.cs.umd.edu/localphp/hcil/vast11/>

¹³<https://twitter.com/>

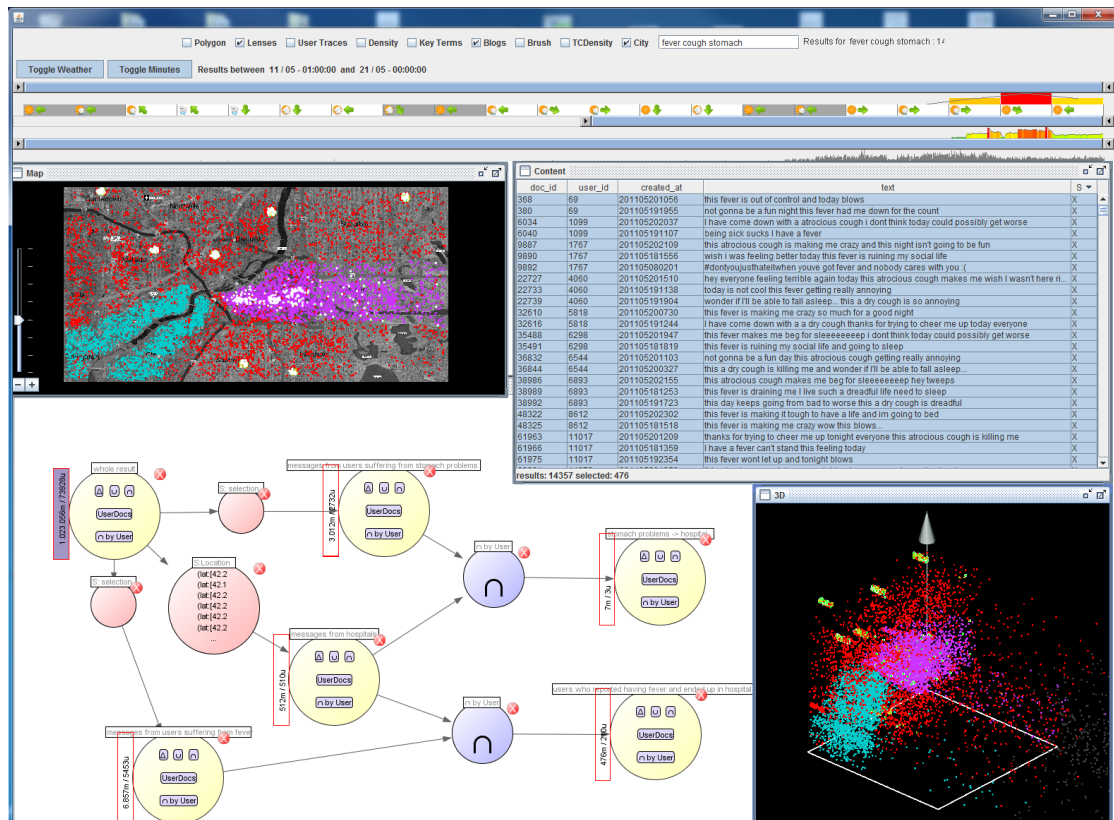


Figure 5.2 — Desktop of the approach for analyzing geo-located microblog messages showing a hierarchical time slider at the top, the map view (middle, left), the microblog list (middle, right), the selection management tool (bottom, left), and a spatio-temporal 3D view (bottom, right). Details on the system can be found in Bosch et al. [2011].

The usage of the reporting technique is exemplified with the following analytic subtask. Two different groups of symptoms could be identified during the analysis: persons reporting stomach problems, and persons suffering from fever symptoms. In order to analyze the consequences of these two ‘outbreaks’, the selection management facility was employed. First, the messages of all persons talking about stomach and fever problems are selected and added as selection nodes (see nodes i) and ii) in Figure 5.4). As a second step, messages from the vicinity of hospitals are selected talking about ‘stomach’ or ‘fever’ as can be seen with node iii) from Figure 5.4. Afterwards, messages of sets i) and ii) are joined by the microblog’s user’s id with iii) resulting in the nodes iv) (hospitalized persons, who had fever) and v) (hospitalized persons who suffered from stomach issues). From the numbers

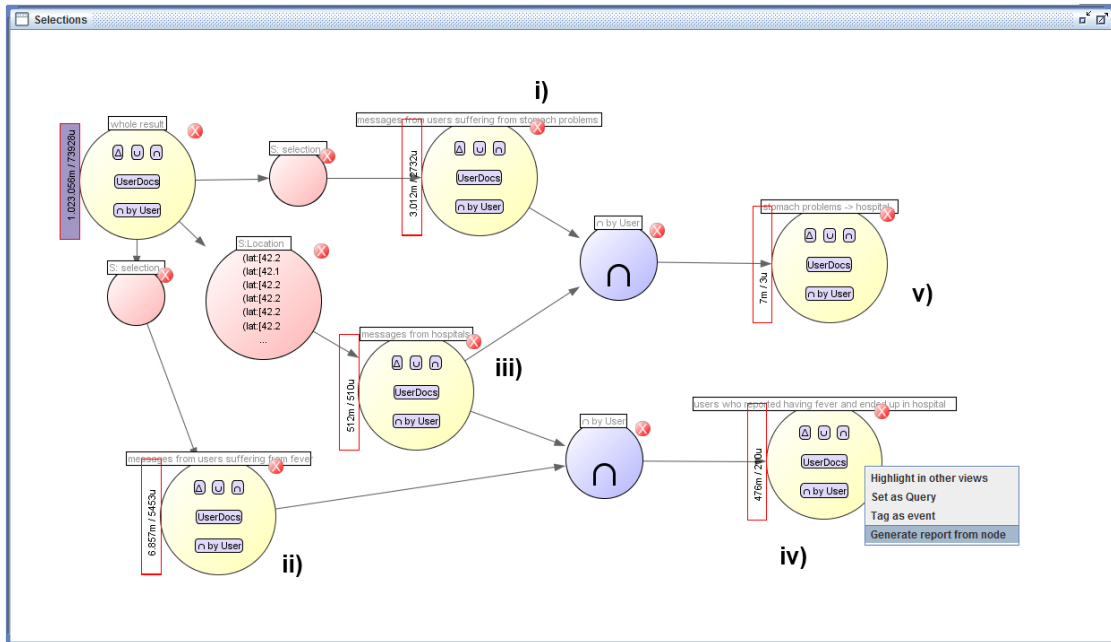


Figure 5.4 — Selection graph that has been constructed by creating a node for people talking about stomach problems i) and persons reporting fever issues ii). Node iii) represent persons writing messages from the location of hospitals on fever and stomach issues. Subsequently, both ‘symptom’-selections i) and ii) are intersected by user with the message set from the hospitals iii). Inspecting nodes iv) and v) it becomes obvious that people reporting fever symptoms end up in hospital much more often than those suffering from stomach issues. In the view shown, the context menu of node iv) has been activated in order to generate a report for the corresponding finding.

only rudimentary statistical information is provided with the report. Right now, generating reports is only possible for one selected node or the the whole graph. This will be extended to select several nodes of interest for report generation. Future enhancements of the proposed method will integrate these missing aspects.

Evaluation, Results and Discussion

This chapter presents results gained through user studies carried out with the approaches introduced in Chapters 3 and 4. Additionally, benefits as well as identified problems are discussed in the context of related work.

6.1 Evaluation

There is an ongoing discussion about how visual analytics approaches can and should be evaluated. The biggest problem is that even for smaller information visualization approaches formal and thorough evaluations are difficult and costly to undertake, at least if starting on the cognition and perception level. Pohl et al. [2012] discuss the relevance of important theories from psychology and HCI in the context of visual analytics approaches,¹ but cannot offer concrete evaluation procedures. If multiple connected views are used, which are in combination more powerful than their single components, evaluation gets even more complicated. Accordingly, pragmatic ways are followed in order to close the gap between theoretical foundations and evaluation strategies that can be used. While promising suggestions have been made for evaluating specific aspects of smaller tasks and corresponding methods in the context of information visualization [Plaisant, 2004], and visual analytics [Plaisant et al., 2008], holistic solutions for evaluating approaches addressing more complex analytic scenarios in detail do not seem to be on the horizon. The complexity of an

¹ They also see Pirolli and Card's [2005] sensemaking model as a probable way to describe visual analytics approaches as is done in this thesis (see 2.7).

approach is also dictated by the complexity of the scenario which makes bottom-up evaluation often too expensive. To some extent this is an inherent problem of visual analytics scenarios that can potentially find broad application in various fields, where large amounts of data have to be studied by human analysts. The problems arising from evaluating such scenarios are discussed in the next section.

This thesis does not claim to present a satisfying solution for evaluating complex analysis methods like PatViz. However, it adds another instrument to the VA evaluation toolset with an approach adapted from the field of information retrieval, which is extended to involve users in the test procedure, while keeping the effort manageable. The said approach has been applied during the evaluation of visual classifier creation and is described in detail in Section 4.2. Additional ideas about exploiting suggested features of visual analytics approaches offering new directions for evaluation are presented as well.

This chapter is partly based on the following publications:

S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative Integration of Visual Insights during Patent Search and Analysis. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*, pages 203–210, 2009

A. Panagiotidis, H. Bosch, S. Koch, and T. Ertl. EdgeAnalyzer: Exploratory Analysis through Advanced Edge Interaction. In *Hawaii International Conference on System Sciences (HICSS 2011)*, pages 1–10, 2011

F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual Classifier Training for Text Document Retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012

6.1.1 The Difficulty of Evaluating Visual Analytics Approaches

Undertaking extensive user studies for evaluating visual analytics systems and approaches is hindered by a variety of problems. These include, but are not limited to:

- task complexity
- tool/method complexity

- duration of tasks
- diversity of domains
- availability of domain experts
- lack of suitable (large enough) test data
- lack of ground truth data and gold standard data
- lack of comparable VA approaches
- lack of VA approaches' maturity
- lack of suitable evaluation criteria

Many of these problems are not independent, but rather intertwined. One of them is the *complexity of visual analytics tasks*. If the analysis scenario is complex and difficult enough, methods for addressing the problem are likely to increase in complexity as well, e.g., with respect to the employment of different sophisticated views, complex automated methods, and advanced interaction (*tool/method complexity*). Visual analytics aims to provide solutions for complex problems of this particular kind, meaning that this is an intrinsic problem. Of course, a primary goal of visual analytics systems is to reduce the complexity, ideally making analytics problems solvable by lay users or semi-experts as well, but some problem domains still require expertise. With an increase in analytic quality or a speed-up of an analytic process, complex approaches, unsuitable to be used without previous training, can constitute significant progress over the state of the art, but come at the cost of being difficult to evaluate. Closely linked to task complexity is a potentially long duration of analyses that hampers a broad evaluation of such approaches with many test subjects. The *duration of tasks* can therefore be another problem, as is the case with patent search, where analysis sessions can easily take a whole working day [Joho et al., 2010].

Analytic tasks which can potentially benefit from visual analytics approaches are highly diverse, since large amounts of data that cannot be analyzed solely automatically are nowadays generated in almost every sector of modern society. Due to this *diversity of tasks*, visual analytics approaches are problematic to compare during evaluation. Quite often, however, the people interested in analyzing this data, and accordingly also the analysis tasks and questions, are domain experts. This poses additional problems, because addressing the needs of domain experts naturally requires the same experts to be the test subjects for evaluating corresponding visual analytics approaches (*availability of domain experts*).

Moreover, suitable data sets must be found for evaluation tasks. While there should be plenty of them available as a result of the often-quoted data explosion, especially in those fields where the pressure to analyze data is high (business and economy), the willingness to provide real data sets is marginal, because most often broad dissemination of internal information does not comply with companies' interest (*lack of test data*). This is even the case if the data has been successfully analyzed before, which would be a prerequisite for having a ground truth of insights that can be drawn from this data. But even if the data is freely available, the sheer amount of data makes the finding or definition of a ground truth difficult (*lack of ground truth data and gold standards*). The broad variety of domains to be covered adds another obstacle, making data sets and tools, and as a result their evaluation, difficult to compare. Other research disciplines, such as information retrieval and NLP in general, have undertaken enormous efforts to create ground truth data sets, e.g., available from evaluation forums such as CLEF² and TREC³. Corresponding efforts in the field of visual analytics are still in their beginnings, even if artificial ground truth data sets were created as part of the VAST challenges and Infovis contests⁴.

Unlike in other fields, automatic benchmarking and testing strategies are difficult to use for evaluation of visual analytics approaches, since by definition human analysts' skills of reasoning and sensemaking are an integral part of any analytics process in this field. This has consequences for the comparability of competing approaches and the required quality of the tools to be evaluated. Even if competing approaches are available, common evaluation procedures must be found in order to make them comparable (*lack of comparable approaches*). This still does not take into account that different test subjects, whether they are experts in their field or not, have different expertise and will most likely choose different analytic strategies of a given task. It further raises the question of how much training or introduction has to be provided to account for similar starting situations. Additionally, provided tools must have near production maturity, since omitting certain convenience functions, or approaches exhibiting low robustness, and other issues are likely to affect analytic efficiency greatly (*lack of approaches' maturity*).

There is still a *lack of suitable evaluation criteria* for visual analytics approaches. However, the problems were recognized in the visual analytics community and a variety of suggestions and methods for carrying out user evaluations under the described circumstances were made. The assessment of insights gained by test subjects during a user evaluation is one common method to rate the suitability of a visual analytics approach. Most often the methods for collecting the gained insights

² <http://www.clef-initiative.eu>

³ <http://trec.nist.gov/>

⁴ <http://hcil.cs.umd.edu/localphp/hcil/vast/archive/provenance.php>

are think-aloud procedures. Methods for counting insights were proposed, but it is questionable whether this is a meaningful measure on its own [Saraiya et al., 2005; North, 2006]. Designing visual analytics approaches in a participatory way has the potential to increase usability, but cannot serve as a qualitative measure. Relatively recent work suggests the combination of these methods and considers relations of insights and test subjects' prior knowledge [Smuc et al., 2009].

Often the abovementioned problems are addressed by reducing the number of participants for evaluation, simplifying analytic tasks, finding similar ones in different fields that can be evaluated more easily, employing small test sets with known ground truth, generating artificial test data sets, etc. To some extent this was also necessary during the evaluation of the approaches presented here.

6.1.2 Exploiting Analytic Provenance Data for Evaluation

Since visual analytics tasks can be extremely complex, why not apply the same visual analytics methods intended to facilitate those tasks to the evaluation itself as well? In principle, some of the features requested for visual analytics approaches, such as provenance recording and report generation, can be exploited for evaluation directly, which increases their importance even more. Clearly, the goals of gathering analytic provenance information while using a visual analytics tool and at the same time evaluating visual analytics tools overlap. A discussion of these points is provided in the context of scalability issues in Chapter 5. However, exploiting provenance recordings for evaluation also brings along a variety of additional requirements. First of all, the visual analytics tools to be tested would require the implementation of provenance and reporting mechanisms, which are often omitted in research prototypes in order to reduce the implementation effort. Furthermore, the recording of provenance data must comply with the level of insights study organizers are interested in and have time stamps included if analytic efficiency is to be evaluated as well. Analytic provenance capturing must also consider data recording beyond positive analytic insight. Otherwise, it would be difficult to derive potential mistakes and dead-ends of a user's analysis from this data. Another requirement would be the description of findings by users, including invalidated hypotheses, in order to capture and reproduce the analyst's chosen path of reasoning. 'Tool insights' as described by Smuc et al. [2009] that are valuable for improving a visual analytics tool's quality are difficult to be attained with this method alone.

Nevertheless, such an approach has some intriguing aspects, and, at least if insights are considered on a task or process level, it could be realized with some of the visual analytics methods discussed in this thesis. Given a certain degree of tool quality, such an approach has the potential to decrease the evaluation effort. The reporting

facility shown during the VAST Challenge 2011 [Bosch et al., 2011] comes closest to achieving this. With some limitations, such a strategy could be seen as reusing analytic provenance information for evaluation and as a substitution or addition to think-aloud procedures. Additionally, the analytic provenance data recorded during longitudinal studies could be evaluated using, again, visual analytic approaches as proposed by Smuc et al. with the RIO approach.

6.1.3 Evaluation of the PatViz Approach

According to Trippe and Ruthven [2011], measuring the performance of patent retrieval systems is questionable if one relies solely on retrieval performance indicators like recall and precision as determined through evaluation setups following the Cranfield paradigm [Voorhees, 2002]. Typical information retrieval evaluations consider a predefined set of documents, where all relevant documents according to a specific query or information need to be known in advance and automatic test procedures are carried out to assess a system's quality without considering the user of a retrieval system. To some extent Trippe and Ruthven are right; at least if the process of searching and analyzing patent information is seen, as within this thesis, as always involving human reasoning and sensemaking. However, performance of the automatic parts of patent retrieval techniques can be improved with traditional evaluations, even if this means that patent experts cannot immediately use such techniques, because it would require to change their search strategies and to learn and deeply understand alternative search back-ends.

For different domains the variance in search effort can be very high. Quite a large number of analysis tasks, techniques, and systems require user expertise in order to judge the quality of a search task's results. This is especially true for difficult tasks and those where the cost of missing relevant documents is high as well – patent retrieval is a good example for this. Domain experts might be able to coarsely judge whether the number of returned relevant documents is reasonable or not and base their decision to continue or cancel a search subtask on this experience.

Trippe and Ruthven [2011], certainly taking the perspective of patent professionals, suggest to “develop [...] evaluation approaches that help estimate the confidence [...] in different system components” and to estimate confidence by the level of ‘trust’ that can be established by users for (parts) of the retrieval process. While they aim specifically at retrieval aspects and do not explicitly take into account visualization, their general idea to develop process-based measures unsurprisingly matches at least partly the ideas for evaluating visual analytics processes. However, they do not offer practical solutions regarding concrete evaluation methods and how to impose trust or confidence measures. As a consequence to the problems discussed at the beginning of the section, the evaluation procedures had to be simplified. For

the approaches taken in the PatViz system interface, two evaluation tasks with two different groups of participants were conducted. The viability of using a visual query representation with respect to its understandability was evaluated through a questionnaire sent out to persons knowledgeable in Boolean search, including patent searchers, via email and resulted in 15 replies. The evaluation of central approaches, such as the interactive reintegration of visually detected insight, was much more challenging to carry out due to problems described above. Especially finding experts in the specific field of ‘optical recording’ and ‘machine tools’ was difficult, since the prototype system was restricted to these patent domains. The length of typical searches also limited the evaluation procedure, because the patent professionals could not afford to spend a whole day testing the system. The most important results of this evaluation are provided in the next sections.

Visual Query Building

As described in Chapter 3, the visual query system consists of two coordinated views - a text-based and a visual one. The tools were developed in close cooperation with patent professionals, but this did not warrant the suitability of the coordinated views for a broader user spectrum. To guarantee that the chosen visual metaphors can be interpreted correctly by users, a questionnaire was drawn up for which test subjects had to interpret single and combined visual metaphors, correlate textual query representations with visual ones, and translate visual into textual queries. All evaluators were asked to answer questions regarding the following aspects:

- *Suitability of the chosen visual metaphors*
- *Comprehensibility of visual metaphors*
- *Recognition of the scopes of Boolean operators*
- *Helpfulness of interactive exploration for query understanding*
- *Creation of Boolean queries*
- and *Composition of complex queries including different search facilities.*

To cross-check the results, most of the aspects were addressed in two different questions, whereby some of the questions incorporated two or more of the aspects above. If required, the evaluators could also include comments and questions as part of their email reply containing the results.

The test subjects were asked to decide whether the provided visual metaphor for Boolean AND and OR operation within the PatViz query approach was appropriate. The evaluators disagreed on whether the Boolean AND operator should

be represented by a sequential or a branching metaphor (analogous to the OR operator). Nevertheless, none of them had difficulties to interpret combinations of the metaphors correctly. There is a strong indication that the visual metaphors are *suitable*. In order to prevent misinterpretation of the visually represented metaphors, additional labels, placed on the links representing operators, were introduced. The *comprehensibility* of the provided visual query example (without labels) was high. All except one of the testers interpreted the visual example queries correctly. The same holds for the testers' ability to *recognize operator scopes* accurately. Thirteen of the testers deemed scope highlighting a useful feature for the exploration of queries. With respect to the *creation* of Boolean queries, three participants mentioned that they would prefer a purely textual query interface over a visual one. All others preferred the combined approach which has been applied in PatViz. Twelve of the test persons expressed the opinion that the approach is suitable for the composition of complex queries including the integration of multiple search facilities. Three were undecided. The result of the questionnaire's evaluation suggests that, even without using the query tool for direct insight integration, the approach already offers an advantage over a purely textual approach.

Iterative Insight Integration

The viability of the concept for insight integration into subsequent search and analysis cycles is much more demanding to test. As already discussed, correct interpretation of patent documents requires at least some experience with the technical field under analysis. For this task, the employment of patent specialists as test subjects was a must, in order to be able to judge the suitability of the developed tools. Since it was difficult to find patent specialists knowledgeable in the field of 'optical recording' or 'machine tools', three patent practitioners from the consortium were asked to take part in a think-aloud evaluation. The actions of the participants as well as their 'loudly spoken thoughts' were recorded. Naturally, the validity of such a test is limited by the relatively small sample for this evaluation. The fact that not enough patent experts knowledgeable in the field of optical recording could be recruited, even within the consortium, exacerbated the problem.

One frequently expressed comment indicated that most of the patent experts had never worked with a system providing linked and interactive visual interfaces. While this was also one of the system's properties most appreciated by the users, it became clear that such features are very difficult to use without previous training. In order to carry out the 'think-aloud' evaluation, the test persons were given access to an online version of the system prior to inviting them for the test itself. Additionally, the evaluators were introduced to brushing and linking within the multiple coordinated views interface and to the meaning and usage of the available

views. Subsequently, they were asked to carry out the same analysis tasks they are performing in their daily work.

All patent practitioners agreed that the visual interface provides a valuable means for creating and editing complex queries for different search engines, but some of them were puzzled when they had to use it for the first time. In subsequent discussions it became clear that this was related to the fact that conventional, mostly form-based, interfaces for patent search are designed in the same way patent documents are structured. Of course, this is not reflected within an interface that allows for arbitrary combinations of different constraints for search facilities; however, it might be a good starting-point for future enhancement of the query visualization tool providing a third view taking this issue into account. Practitioners who were used to employ formal Boolean languages instead appreciated the visual representation from the beginning.

Another observation was that most of the patent experts used views like the tag cloud, the legal entity charts, and the world map more frequently than the more sophisticated ones. A probable explanation for this behavior is that users may tend to perform their tasks with tools they are accustomed to. Nevertheless, after a quick introduction, the testers were able to integrate the other views successfully into their analysis. The most significant benefit identified by the test users was the support for iterative refinement of queries and patent sets. Also the synergetic effects of using different views of the same set in parallel were appreciated by the users and the linking and brushing facilities were used extensively after a short period of familiarizing themselves with the system. The testers commented positively on the flexibility and power of the system resulting from the degrees of freedom in moving back and forth between the stages of the analysis process and between different perspectives within one stage of the process.

6.1.4 Advanced Focus+Context

In order to evaluate the fundamental idea of the EdgeAnalyzer approach only, again the think-aloud protocol was employed to evaluate an early version of the prototype. Due to the large number of combinations of parameters and methods for the proposed technique, it was not feasible to perform a comparative evaluation of all features. Instead a subset of the implemented grouping mechanisms and available views was accessible during the evaluation. The arc wheel and the employment of multiple lenses were also not available during evaluation, since these features were developed later. Participants could use a single lens (moving/resizing), geometric grouping, and parameters were fixed. The evaluation should therefore be seen as a rationale for introducing features such as the arc wheel and to illustrate users' general acceptance of the approach.

Eleven students took part in the study that involved two tasks embedded into the example scenario about patent co-classification analysis. A data set consisting of 1000 patent documents and 169 co-classification relations was used for both tasks. The test subjects received a brief introduction to the IPC schema, the corresponding tree map view, and the basic idea of patent co-classification. Afterwards they were given a few minutes to familiarize themselves with the lens-based focus+context technique and to ask questions regarding its usage until they felt confident enough to start with the tasks. All participants stated they were able to use the technique in less than two minutes of training time. During the study, the participants were encouraged to think aloud, ask questions if they got stuck, and make general comments on the tool even if they were not related to the tasks at hand.

The first task was to explore a dense edge cluster that was partially collapsed to a single bundle due to edge bundling, which constitutes a common situation in dense node-link diagrams (Figure 4.3 depicts a similar situation). Edge groups were visualized with the local de-bundling method, that splits edges in the focused region. For grouping three methods were available: grouping by intersection points with the lens, grouping by angle between lens-intersecting line segments, and ‘no grouping at all’ – meaning that each edge was allocated to a separate group, whereas the first two methods are variations of geometric grouping. The test subjects could switch between the different grouping mechanisms.

Three participants found that ‘no grouping’ is superior in this particular task, while most test subjects commented, that the geometric grouping methods seemed to yield unreliable and incomprehensible results. Nevertheless, two participants preferred them as they produced fewer interaction elements. One reason why participants favored ‘no grouping’ was the fact that the number of focused edges was instantly perceptible. Test subjects also mentioned that multiple lenses could have been useful for this task.

The second task asked participants to find the IPC group with the highest number of connections using the row representation of edge groups (see Figure 4.4, right image). While the saturation of edge bundles already gave a hint to the solution, still many nodes appeared as potential candidates in this scenario. All participants found the correct answer, and two were surprised how deceptive the opacity of bundles was, since no visual difference could be perceived when the amount of edges exceeded a certain number. Asked about their approach, eight participants answered that they first used the opacity of edges as a starting point and then the lens for further inspection.

After completing the two tasks, the test subjects were asked to state their opinion regarding the usefulness of the technique for edge exploration. All participants appreciated the proposed interaction mechanism for exploring large graphs. Five

participants could also imagine to select edges by stretching a user-defined shape over the edges, i.e. employing traditional brushing interaction, instead of moving the lens-shaped metaphor. The option to change the size of the lens was used frequently. Seven participants appreciated that the lens stays visible after a focus region has been selected. Two other participants disliked the overdraw of the focus region by the lens and the visual representation of edge groups. The tooltip was highly valued by all participants. Three participants wanted to move the tooltip, which was not possible in this scenario. Seven participants stated that this interaction technique was confusing at first and took some time to get used to.

As a consequence of this study, metadata based grouping mechanisms were implemented, since users seem to expect a comprehensible, predictable, and repeatable outcome when grouping edges, which is not guaranteed for geometric grouping criteria that can also depend on the lens' relative location to intersecting edges. In addition, EdgeAnalyzer was extended by the arc wheel and with support for using multiple lenses. Furthermore, focused edges are always highlighted entirely by default, meaning that not only the focused part is affected. Non-focused edges can be hidden on demand to further reduce clutter.

During this study, clearly more insights regarding the tool than the underlying data set were derived, as was intended for this early stage study, aiming more at the approach's usefulness [Greenberg and Buxton, 2008] than at usability or for determining how much insight could be gained from data analysis. Accordingly, the tasks were predefined and limited to specific exploratory usage of the tool. As a result, of course, usability aspects could be addressed afterwards by adding additional interaction mechanisms as described above.

6.1.5 Classifier Creation

For evaluating the approach of visual interactive classifier creation, it was possible to choose a more comprehensive method. The evaluation procedure could be designed in a comprehensive manner, since the tasks to be carried out were much less demanding and complex as opposed to the analytic possibilities offered by PatViz. Furthermore, the method for user-steered classifier creation as described in Section 4.2 is applicable to a broader, almost domain-independent spectrum of text retrieval scenarios, aiming at improving high-recall search tasks. Another important aspect was the availability of test data sets with given gold labels, or data where gold labels could be derived from. Additionally, two extra tools were developed for carrying out the same analytic tasks in order to compare the visual interactive approach with two other, more basic techniques. These additional tools and their functionality are briefly described along with the evaluation setup in the

next section. Primarily, the user study aimed at finding out whether a classifier can be trained visually and interactively. Further goals included a comparison of the tool's effectiveness with other methods, gaining feedback regarding the employed tool's usefulness, and assessing the insights the test subjects could derive for the tasks. To achieve these results, a task-based user evaluation including two different tasks and all three alternative methods for classifier creation was carried out, complemented by think-aloud procedures, questionnaires, and open discussions with the participants, as described below.

Evaluation Setup and Procedure

For the qualitative evaluation three different data sets were employed, two of which are well-known and widely-used benchmark corpora with gold labels for text document classification. In the context of the classifier building task one problem was to create an artificial information need that fits the existing gold labels of the corpus. As a consequence, information needs were derived from the gold labels in the corpus and communicated to the participants to ensure that participants' labels, applied during the classification procedure, fit the original gold labels.

Through introducing an artificial information need, the comparison of different participants' results as well as comparing the approaches' effectiveness was possible. This also introduces a problem: When using standard test corpora or tasks and information needs that do not reflect the participants' interest in the data, test subjects are less motivated in carrying out the task – an effect described by [Saraiya et al. \[2006\]](#). It can therefore be expected that the reported results would be better for all methods if text corpora of an analyst's interest were used in combination with a real information need.

Three text corpora were used: *20 newsgroups*⁵, *Reuters RCV1*, and one corpus that has been assembled from the abstracts of *VisWeek*⁶ publications. The 20 newsgroups corpus (short: 20ng) consists of usenet postings from 20 different newsgroups. The corpus was assembled by the author of the Newsweeder system [Lang \[1995\]](#) and has often been used as a benchmark corpus since then. The version applied during the evaluation was one where duplicates were removed resulting in about 19,000 remaining postings. As the example task for this corpus all computer-related posts had to be labeled as relevant for classifier training and all others as non-relevant. The gold labels were created by defining all posts from the comp.* newsgroups as relevant, and others as non-relevant.

⁵ The corpus itself can be downloaded from the web, e.g. at <http://people.csail.mit.edu/jrennie/20Newsgroups/>

⁶ IEEE VIS (formerly known as VisWeek) is a yearly held major forum for conferences in the fields of scientific visualization, information visualization, and visual analytics <http://ieevis.org/>

The second corpus is a subset of *Reuters RCV1* [Rose et al., 2002] (RCV1). While the original corpus consists of over 800,000 Reuters newswire articles, a subset of 12,000 documents was chosen for the evaluation to keep the document pool that each user had to handle at a manageable size. A discussion of scalability aspects, including the employment of meaningful amounts of documents for classification training, can be found in Section 5.1.1. All articles in the RCV1 corpus are labeled according to their topics. For the RCV1 corpus, the task of separating all sports news articles (relevant) from the other news (irrelevant) was chosen.

The third corpus contains about 1,200 VisWeek abstracts. There are no gold labels for this corpus, since it was only used for familiarization with the interface. The exercise task consisted of the identification and labeling of all those abstracts as relevant that talked about a natural language processing or text visualization component. As all of the participants had a visualization background, this was expected to be a rather straightforward task for them.

The documents of each corpus were tokenized and represented as normalized *tf-idf* vectors (see Section 2.5.2). Stemming of tokens was omitted to avoid confusing the participants by displaying terms that are difficult to recognize as regular words.

In order to carry out a comparative user study two other tools were employed. At first, a rather basic method was created resembling active-learning-based procedures usually carried out for NLP labeling tasks (see Figure 6.1). Typically this procedure is used to produce labeled data sets as a basis for creating classifiers afterwards, without involving the annotators in the training process or in the assessment of a trained classifier's quality. After the initial bootstrapping step, the basic method presents the most uncertainly classified document to the annotator who must label it as relevant or non-relevant to the given task. In case annotators are undecided regarding the document's class, they can also reject the document without labeling it; they are then presented with the second most uncertain document, etc. Each labeling action automatically entails a subsequent training step. Feedback on the current classifier's choice is provided by showing the classifier's guess on class membership of the document. Labeling actions continue until test subjects have the impression that the classifier works adequately enough.

The second method already provides the interface as described in Section 4.2 for the visual, interactive, user-steered method. It enables participants to explore the current classifier's state visually, but enforces the active learning procedure as described for the basic method. This means that all selection interactions available in the user-steered method are disabled, the labeling panel is replaced with the same view shown for the basic method, but without stating the classifier's guess on the displayed document's classification. The classifier history is not available either.

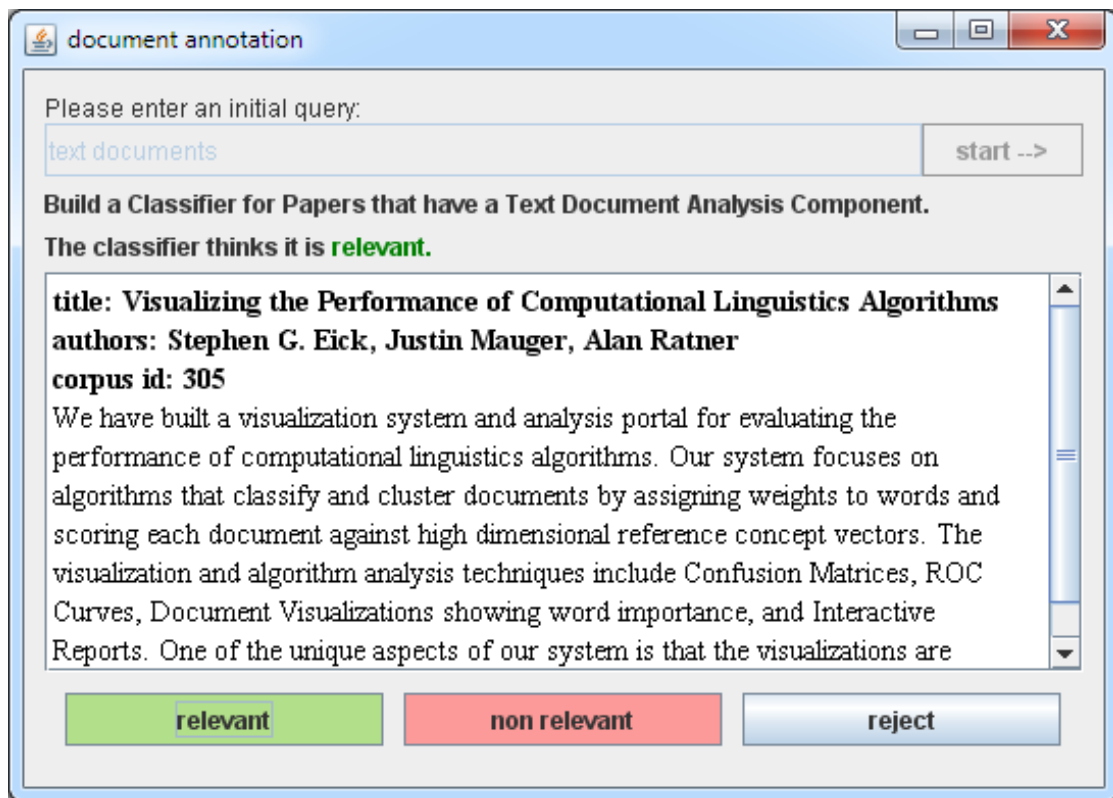


Figure 6.1 — The interface for the basic method showing just one document at a time. Users can either label it as relevant, non-relevant, or reject it if they are unsure to which class it belongs.

The idea of introducing this method was to detect whether the visual feedback would help users in assessing the quality of a classifier, e.g., to stop the training if sufficient classification quality was achieved.

The participants for this evaluation were twelve PhD students from the visualization department. They were asked to create classifiers with two of the three methods. In order to prevent users from becoming familiar with the data sets, they were presented different corpora for their first and second task. Additionally, the order of the methods was permuted to diminish the effect that test subjects become familiar with the tools, since parts of the different methods are similar. Each combination of method and labeling task has thus been executed exactly four times by four different participants in different order.

During the evaluation the participants were encouraged to think aloud about all aspects of the tool or the task that came to mind. All comments during the evaluation sessions were recorded on paper. In addition, the participants' actions

were observed and interesting behavior, including mistakes and employed strategies of usage, were logged.

The evaluation procedure for each participant comprised all of the following ten steps:

i) initial instruction: Participants were instructed about the evaluation process and informed that they could stop the evaluation at any time and without giving any reasons for stopping. They were also informed about which data was recorded during the evaluation sessions and that all of the recordings were fully anonymized.

ii) colorblindness test: Each participant was tested for colorblindness with the Ishihara color plates. This was necessary because the methods contained red-green color differences by default, but could be switched to another color mapping if necessary.

iii) tutorial for the first task: Participants were introduced to the tool for the first method and received a short tutorial. Afterwards, they were able to use the tool with the exercise task until they felt confident enough to start the real labeling task with the next corpus. Questions about the tool or the labeling tasks were answered at any time.

iv) tool evaluation for the first task: The first task was started. The participants had a maximum time of 15 minutes to accomplish this task, but they were also allowed to stop at any time for any reason (e.g. because they were satisfied with the classifier's performance or the training did not make any progress).

v) questionnaire for the first task: Participants were asked to complete a questionnaire about the first task.

vi) - viii): Tutorial, tool evaluation, and questionnaire for the second task.

ix) final questionnaire: Participants were asked to complete a questionnaire with questions about their age and previous knowledge.

x) comments and discussion: Finally, the two methods and tasks were discussed with the participants. Each evaluation session took about one to one and a half hours.

Quantitative Evaluation

In order to compare the classifiers' performances, predefined queries for each task were used. The queries were kept constant for each participant and corpus, in order to guarantee the same starting situation and to explicitly rule out suboptimal starting configurations caused by problematic initial queries. The artificial information need was presented to the participants in terms of the initial query and in form of a short task description displayed in each of the three tools.

Corpus	Initial Query
20ng:	computers network motherboard graphics
RCV1:	sports baseball basketball tennis game
VisWeek (tutorial):	text

Table 6.1 — The fixed initial queries used for classifier bootstrapping with the three corpora in the evaluation

Accordingly, all participants started with the exactly same configuration. The preselected initial queries were as follows:

For measuring the performance of the classifiers trained on the selections of the participants, each of the corpora was split randomly into an 80% training set and a 20% test set. The evaluation task was performed entirely on the training set. The smaller test set was used to evaluate the performance of the classifiers with respect to the gold labels.

Figures 6.2 and 6.3 show the results of the quantitative evaluation for the two corpora with the three methods. Each of the diagrams depicts the classifier’s evolution curve⁷. Hence, each diagram contains four classifier evolution curves of four different users. The classifier evolution curves generated by the participants are compared to a random sampling baseline (in blue) and simulated active learning (in black). The random sampling curve depicts the average evolution (ten simulation runs) of a learner that randomly picks a document from the training set and assigns its gold label. The simulated active learning curve has been generated with the basic method as proposed for interactive systems by [Tong and Koller \[2000\]](#). The resulting classifiers’ performance was then measured on the test set of the respective corpus. Simulated learning was realized through a perfect labeler using the gold labels, in order to have a common baseline which all curves generated by the participants could be compared to. In contrast to the classifiers’ evolution curves produced by the participants of the user study, which assigned labels to documents according to the labeling task assigned to them, some reduction in performance due to different labeling decisions can be expected. For the simulation of AL (the black curve in Figures 6.2 and 6.3), the same initial training documents as for the user evaluations were used to keep the results comparable. All of the classifier evolution curves are identical up until 100 instances due to using the same preliminary bootstrapping step in each case, and are therefore cut off below this border. The dashed black line depicts classification performance when the classifier is trained with all gold-labeled examples. The dashed line in each of the diagrams

⁷ In the literature such diagrams are known as *learning curves*. To avoid confusion with users’ learning progress, the term *evolution curves* is used here in the context of classifiers instead.

indicates the performance of the classifier when trained on the complete training set with the gold labels.

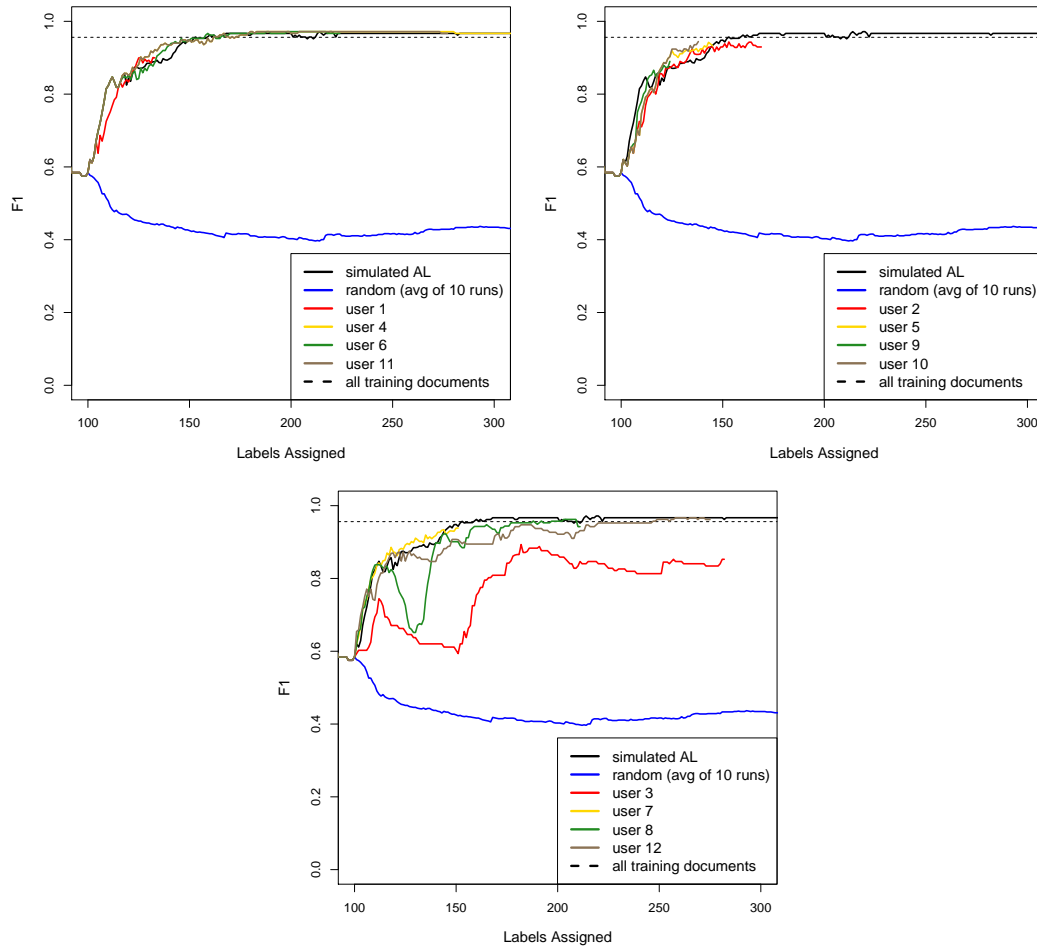


Figure 6.2 — Evaluation results as F_1 -scores over labeled instances for the *RCV1* dataset. The upper left diagram shows results for the basic method, the upper right for the visual method, and the bottom one for the user-steered method. The dashed, black lines represent the training of the classifier with all available training instances. The blue lines in each case show 10 averaged classification runs with random sampling (perfect labeling). The continuous black lines depict the classifiers’ evolution with a perfect labeler using simulated active learning. All other curves show results achieved by the test users.

The positive effect in speeding up the creation of a high-quality classifier by applying active learning can be clearly perceived from all diagrams with the exception of

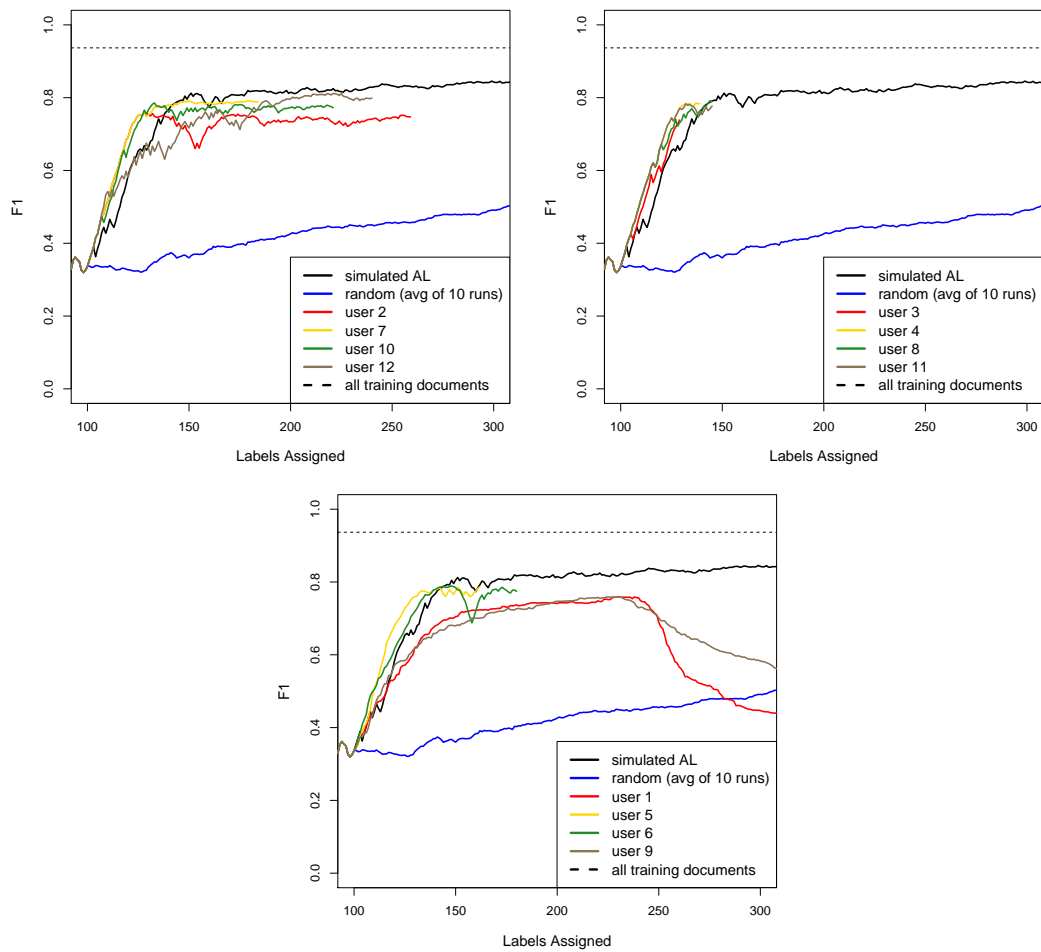


Figure 6.3 — Evaluation results as F_1 -scores over labeled instances for the *20ng* dataset. The curves in the diagrams are to be interpreted analogously to Figure 6.2. Upper left: basic method; upper right: visual method; bottom: user-steered method.

the diagram for the *20ng* task with the user-steered Method in Fig. 6.3. It is thus possible to effectively train an LSVM with all three methods producing comparable results. Taking a look at the best results achieved during the evaluation, the research question, whether it is possible to train high quality classifiers with an visual analytics approach, can be answered positively.

Classic active learning and the user-steered method learn slower on the *20ng* task compared to RCV1, which is due to the greater diversity of the relevant class in the

task	basic method	visual method	user-steered method	all docs
RCV1	0.97	0.94	0.96	0.96
20ng	0.80	0.79	0.80	0.94

Table 6.2 — The shown performance values are measured in F_1 , the *all docs* column specifies the performance of a classifier trained on all gold-labeled documents of the training set.

20ng corpus. The evaluation indicates that methods with lower degrees of freedom in labeling are more robust against labeling actions that have a severe impact on the trained classifier’s performance.

Examples for such disadvantageous labeling actions can be seen in Fig. 6.3 (user-steered). Here, user 1, depicted in red, labeled vast amounts of the negatively classified documents as positive in one training iteration. This resulted in a strongly skewed classifier. Unfortunately, the undo functionality was not used afterwards, but other mass labeling actions were performed in order to ‘repair’ the classifier. This led to over 2,000 labeling actions and a slight improvement ending with an F_1 score of 0.42. User 9 made selections mainly using the term chart view without much further refinement. This resulted in many incorrectly labeled instances and thus in a bad classifier performance. Since the number of labeling actions in these two cases was very high, most of them are cropped in favor of preserving details of the other participants’ actions. The F_1 values come back up during subsequent training rounds, but do not come close to the levels achieved by other users.

What can be clearly learned from these results is that the additional degrees of freedom the visual analytics approach provides are accompanied by risks of choosing wrong labeling candidates. This is at least the case for rather good-natured tasks used in the described evaluation setup. Here, uncritical queries for the bootstrapping steps were chosen, and an active learning procedure with uncertainty sampling works well. If many documents are incorrectly labeled due to the initial query, it can be difficult to create a good classifier without relabeling these documents. Relabeling documents, however, is not intended in ‘classic’ AL approaches as applied in the basic and the visual method. In the context of dynamic data sources, such as blogs and forums, visual verification of the classifier performance is important. Otherwise, new relevant subtopics that were not encountered during classifier training are not detected and the need for retraining the classifier cannot be recognized anyway.

The basic method outperforms the visual and the user-steered method with respect to efficiency and higher robustness. However, this comes at the cost that users

cannot judge the quality of the produced classifier very well, which was one aspect participants complained about most when using the basic method.

Qualitative Evaluation

Qualitative feedback was requested from the users through a questionnaire on an optional basis after they carried out each evaluation task. The first part of the questionnaire contained questions of the standard NASA-TLX test [Hart and Staveland, 1988] for measuring the task load on the participants. Other questions asked about the participants' trust in the classifier, why participants stopped the training, and the usefulness of the methods' views. The final questionnaire requested information on the participants' age, gender, and their expertise in using web search engines, machine learning, using interactive visualizations, and carrying out information finding tasks.

No correlation could be found between the performance of the trained classifier and the level of participants' experience in machine learning. This could be seen as a first indication that all three tested methods can be used successfully by machine learning non-experts. The 20ng corpus labeling task turned out to be more difficult than the RCV1 task, which was reflected in the test subjects' answers. Furthermore, the participants had higher trust in the classifiers built for RCV1, independent of the applied method, and they reported lower workload as well. Task complexity obviously plays a bigger role than the employed method.

Most of the participants stopped training because the time limit was reached. The classification view was rated as being most beneficial for the task, followed by the detail view. Some users found the term chart view helpful, while others hardly ever used it. The interaction mechanism provided with the term lens was regarded as equally helpful by the test subjects.

The affiliation of the test persons as well as their expertise in computer science and visualization may raise questions regarding their representativeness. In addition to their experience in the field of visualization, they matched the scheme of being specialists in their field, and should have some familiarity with recall-biased retrieval problems such as related work search. However, it became obvious through the participants' questions and loudly spoken thoughts that not all of them were able to develop a meaningful interaction workflow for training a classifier with the user-steered method. It could be observed that some participants were still learning and exploring certain aspects of the tools while carrying out the tasks for the visual and user-steered method. This might be a possible reason for the two quite unsuccessful attempts in training the LSVM with the user-steered method (see Figure 6.3). These findings indicate that the learning curve to successfully exploit this method is high.

Further comments indicate that inspecting a considerable number of documents sequentially, as required with the basic method, is tiring, and that the participants are getting inattentive quickly. Two of them even reported physical load in the TLX questions and mentioned eyestrain and weariness in the later discussions. Getting visual feedback is preferred by most of the participants.

In summary, most participants achieved good classification performance. Since almost all of them used the whole 15 minutes for their labeling tasks, the number of labeled instances in the diagrams shows that participants were faster using the basic method. The influence of a task's difficulty on the amount of labeled instances can be seen in Figures 6.2 and 6.3, when comparing the results of the RCV1 task and the 20ng task. It is likely that the time planned for the tutorial was insufficient for mastering the complexity of the user-steered method. This has an additional negative influence on the number of labeling actions. Better user training would most probably diminish this influence.

6.2 Discussion

The different evaluations of the presented approaches revealed benefits regarding analytic effectiveness but also flaws. Both are summarized and discussed in the following.

6.2.1 PatViz

On the one hand, the flexibility and implicit functionality provided by the developed prototype is difficult to comprehend if users start working with the system without previous instruction. To some extent, this problem can be reduced by providing appropriate context-sensitive help systems. On the other hand, a powerful and flexible demonstration prototype that facilitates diverse patent analysis tasks was created. This fact has been recognized by the test users and was positively emphasized by them during the test sessions. The main contributions of PatViz are the visual representation of queries to address different retrieval back-ends, the result set views that allow for an advanced visual form of faceted browsing, while still preserving context, and the selection management and filtering approach. In combination, these techniques support a variety of interactive feedback loops required for a coherent and powerful analysis environment and suitable to leverage human information discourse. In particular the possibility that insights gained during analysis can be directly exploited in subsequent query iterations has the potential to be useful for other text retrieval scenarios. The selection management approach has already been successfully exploited in different visual analytics

prototypes and always led to an increase of analytic power, besides offering a coarse form of provenance recording.

6.2.2 EdgeAnalyzer

The approach presented with EdgeAnalyzer is specifically interesting to explore complex visualizations containing a large number of visual elements. Its application within PatViz, however, has to be seen critically, since it deviates from other views regarding increased interaction complexity. One aspect is the break in conformity with respect to integrating two almost opposite approaches with overview+detail plus brushing&linking in multiple coordinated views and a focus+context technique within one of these views. No evaluation has been carried out to detect corresponding negative effects, but from the feedback received from patent specialists for the PatViz interface, it can be deduced that patent analysts need quite some training to use interactive visual tools efficiently. Offering sophisticated interaction facilities, such as with EdgeAnalyzer, adds further complexity and would most probably require an extended training phase.

However, the approach itself provides a powerful interaction mechanism and can be applied to many different scenarios where relational connection of data attributes is expressed with visual links or edges. It further replicates some aspects of the selection management a filter approach on an intra-visualization level through the employment of advanced and stacked filters, and the support of multiple lenses for the combination of findings and selections. In scenarios where an explorative approach for rather dense, link-based views is needed, it can be a useful add-on.

6.2.3 Classifier creation

The evaluation as carried out above showed that the basic method tends to be more robust regarding the quality of the trained classifiers in comparison to the suggested visual interactive procedure. However, user-steered classifier creation offers a variety of benefits that could not be addressed adequately in the user study presented above. In the evaluation, users only had to deal with well-selected, unproblematic initial classifiers. Depending on the task and predefined query this cannot be guaranteed in general. While the bootstrapping step through a user-defined key term search is suitable to leverage the creation of an initial classifier, it also poses risks. The initial query should be carefully created to achieve high precision. Otherwise there is a chance that a considerable number of non-relevant documents will be wrongly labeled as relevant. This results in displaying them as regular training data, possibly even far away from the decision border. If many

documents are incorrectly labeled due to the initial query, it can be difficult to create a good classifier without relabeling these documents. Relabeling documents, however, is not possible in ‘classic’ AL approaches as applied in the basic and visual methods that were used during the evaluation.

Similarly, a query with low recall can lead to false negative training examples displayed in the same way as false positives on the other side. Depending on the document set to be classified and the concrete query formulation, this however gets more unlikely if the selectivity of the query is low and/or the document collection is huge.

Furthermore, visual control and corrective actions are important in the context of dynamic data sources, such as (micro)blogs and forums. Otherwise, new relevant subtopics, that were not encountered during classifier training, are not detected and the need for retraining the classifier cannot be recognized. Situations might arise where a set of documents cannot be classified according to an analyst’s needs at all. Such situations are hard to detect without visual feedback as well.

Until now the approach for visual classifier creation is realized as a separate software prototype. If it were to be used as part of larger systems, e.g., such as PatViz, a variety of architectural considerations would have to be taken into account. As described in Section 5.1, it is reasonable to organize large information retrieval systems into several distributed layers. The communication channels between visual front-end and the system’s back-end interfaces are used to send queries and retrieve the corresponding results. However, such systems are usually not designed to integrate user-defined retrieval methods in the back-end.

Even if the overall complexity is considerably lower than with PatViz, the created prototype showed similar problems during evaluation as encountered there. While it could be handled effectively by some users during the evaluation, they had to understand the tool and develop useful workflows. In this case, there is no break with conventions, because analysts are not confronted with classifier training and assessment for ad hoc tasks at the moment (see next section). Nevertheless, it is worth pursuing this further, since high-recall searching requires that analysts build trust in their search methods, and visual feedback as well as interactive methods for fixing/improving a search method can be a suitable means for achieving this.

6.2.4 General Considerations

A key issue of visual analytics is providing appropriate interactive feedback loops at different levels of abstraction, leveraging analysis approaches by facilitating hypotheses testing, and better integration of automatic methods – these are the most important protagonists of human information discourse. In particular non-

linear workflows can be supported with such feedback loops as is often the case with tasks that require human reasoning. The approaches presented in this work all facilitate such feedback loops, as part of domain-tailored systems like PatViz, but also within more generic approaches addressing analytic sub-tasks, as shown for visual classifier creation and the sophisticated interaction mechanisms depicted in EdgeAnalyzer. As can be seen from the analytic process models in the Figures 2.9, 4.2, and 4.17, the presented approaches follow the aforementioned Shneiderman mantra “*Overview first, zoom and filter, then details-on-demand*” [Shneiderman, 1996], or Keim’s adaptation thereof “*Analyse First – Show the Important – Zoom, Filter and Analyse Further – Details on Demand*” [Keim et al., 2006] on different levels of abstraction. The better feedback loops can be integrated with the help of interaction, the fewer breaks occur within analytic workflows, and the better analysts are supported in fulfilling their tasks.

However, feedback loops also raise new questions regarding system design and suitable software architectures. Automatic methods, often employed in earlier stages of visual analytics pipelines (i.e., on the data source and the aligned data level as can be found in Figure 2.3), should be made steerable and adaptable for situations where analytic tasks require such an adaptation. As a consequence, the integration of automatic approaches and visualization has to be designed tighter, which can cause conflicts with software engineering principles, such as separation of concerns, thereby raising new scalability issues (see Section 5.1.1). The larger the feedback loops are with respect to the stages and steps they are spanning (even if this would be very advantageous for the analytic workflow), the larger these problems get. This issue is also related to back-end systems, such as databases and retrieval systems, that exist in rather standardized form and often do not provide means to change their settings or working behaviors during analytic tasks. Both issues are severely hindering scalable visual analytics solutions.

With respect to comprehensiveness of the presented approaches, PatViz is targeted at the patent domain, while EdgeAnalyzer and user-steered classifier creation are more flexible, even if they were also presented in the context of patent literature analysis. Nonetheless, PatViz can be adapted to related domains like searching for scientific literature and other document retrieval tasks requiring high recall. In particular the query approach could be applied in many other scenarios. The other two approaches have fewer constraints regarding their field of deployment. EdgeAnalyzer can be used for every dense, line-based visualization, as long as a need for inspecting relational details represented by those lines exists. User-steered classifier creation can be employed in text retrieval scenarios, where more effort spent on the retrieval task can be justified by higher quality of the results - either because the task is worth it, or because it can be reused many times thereafter.

Both can, again, be integrated into larger analytics systems such as PatViz as has been described in Chapter 4.

Explicitly depicting found insights and the testing of hypotheses not only in the model but also visually are key aspects of visual analytics approaches. Including means enabling analysts to actively construct their analysis is therefore beneficial for a broad variety of aspects considered to be important for visual analytics approaches. Such an approach has been presented in this thesis with the selection management tool as proposed in Section 3.2.9. Constructive analysis instead allows for externalizing single analytic subtasks in arbitrary sequence, which acknowledges human reasoning processes, typically not being linear. Letting users connect these insights through operations results in an analysis graph where combined and higher level insights can be represented as well. Apart from helping analysts during the analysis by permitting them to externalize insights to work on a specific subtask that can later be integrated into the analytic graph, this construct can be used to support collaboration, history recording, analytic reporting, and probably even user evaluations of analytic approaches, by comparing the created analytic graphs.

Crouser and Chang [2012] suggest a framework for describing different facets of human computer collaboration and for assessing the complexity of visual analytics approaches which facilitate this cooperation, employing an ‘affordance-based’ perspective. The term *affordance* was originally coined by Gibson [1977] and was transferred to the field of human-computer interaction by Norman [1988] to describe opportunities for the action an object offers to a human (inter)actor. Crouser and Chang generalize the term to a set of ‘human affordances’ and ‘machine affordances’ in order to be able to describe the abovementioned collaboration. In a case study Crouser and Chang come to the conclusion that PatViz provides too many of these affordances. They see them as the source of PatViz’s problems as identified in Section 6.1. However, Crouser and Chang mix up affordance with perceived affordance, constraints, and conventions, at least if they derive their notion of affordance from Norman’s understanding of it – a misinterpretation that has been made before by others [Norman, 1999]. Referring to the addition of graphical interaction elements to a user interface Norman states:

“Usually they [graphic designers] mean that some graphical depiction suggests to the user that a certain action is possible. This is not affordance, either real or perceived. [...] It is a symbolic communication, one that works only if it follows a convention understood by the user.”

What Crouser and Chang identify as a problem resulting from too many affordances in the case of PatViz is rather the fact that it breaks with conventions, to phrase it in accordance with Norman’s terminology. And it not only breaks with one of them,

but violates a variety of rules suggested for designing user interfaces in general, including *conformity with user expectations*, and *self descriptiveness* and *providing user guidance* (for new and complex views and interaction techniques). These issues are discussed subsequently in the context of visual analytics approaches.

Conformity with user expectations: To some extent the idea of visual analytics is to find *new ways* of leveraging human reasoning capabilities by providing visual and automatic methods that are combined through interactive means [Thomas and Cook, 2005]. For situations, tasks, or processes that could not be addressed before, visual analytics approaches can define the target course and come up with completely new ways of addressing problems visually, as well as from an interaction perspective. For already established, hard-to-solve analytic tasks, such as patent analysis, new ways have to be developed to speed them up, make them more reliable, or raise the quality of analytic outcomes. What is important in both cases, is that visual analytics approaches are designed in a task- and process-tailored manner. The approaches presented in this thesis were developed with task-appropriateness in mind.

Self descriptiveness: Following Norman's idea, interactive information visualization has to be interpreted as means for *symbolic communication* as well. Especially advanced information visualization is not understood by a large number of 'readers' and not 'spoken' (by interacting with it) either. It does not always *'follow a convention understood by the user'* and it is therefore not comprehended automatically by users - it has to be learned. This is also one of the reasons why new developments in information visualization take time to reach practical application. The problem is intrinsic to visualization. Yet visualization also offers the chance to introduce new ideas, that will be learned by specialists drawing large benefit from them, and which might become conventional for a broader audience, if they turn out to be useful to them over time.

Of course, analysts should not be left alone in coping with a new analytic system. And in the context of visual analytics, it is also important to ensure a careful trade-off between expert users and casual users. While the effort to learn new visualizations might pay off for expert users, this might not be the case either for casual users. Ad-hoc usage of visual analytics systems should offer more conventional views, that represent 'common grounds' and which are more easy to learn.

PatViz does account for self-descriptiveness by providing tooltips for visual items and employs descriptive labels where possible to reduce this gap, but does, for instance, not provide online help as would be necessary for a regular product. This brings up the topic of user guidance.

User guidance: Whether the sensemaking loop as described by [Pirolli and Card, 2005], Keim’s model for depicting visual analytics approaches, or the extension of Card’s information reference model as suggested in this thesis (see Section 2.4) is used to describe visual analytics processes and approaches, all have one aspect in common: they are highly flexible with respect to the analytic paths that can be followed, and they rely on exchanging data in between all modules involved in architectural models. It would be a great mistake to reduce analytic flexibility of such systems in general, since this would force users into analytic paths they might not want to follow.

In PatViz, user guidance was addressed by introducing the measures described in Section 5.1.4, in order to account for certain analytic setups or to support lay users, but it also provides the flexibility to address all feedback loops that are necessary for effective patent search and analysis. Visual analytics approaches will always have to be developed under the conflicting priorities discussed above.

There is also no doubt that an approach such as PatViz can be improved into many directions. However, the approach provides advances over the state of the art, since otherwise, patent experts would not have appreciated the feedback loops it provides. For evaluating new approaches and addressing existing analytic problems this poses a problem, since experts from the field will always be confronted with ‘unfamiliar territory’. As a consequence, learning curves for such systems can be steep - a fact that should not be omitted if presenting visual analytics approaches for complex scenarios⁸.

If aiming at commercializing new ideas for problem domains with already existing solutions, it is therefore wise to follow the guideline ‘evolution instead of revolution’ and try to keep old workflow patterns while also offering new ones. To some extent this has been realized with PatViz, that offers some well-known views and options, such as textual Boolean query definition, list-based result set views, and classical detail views, but also provides a variety of alternatives, regarding visual methods, interaction techniques, and alternative workflows. In the case of PatViz, traditional views were integrated to explain alternatives to patent experts and in order to kick-start evaluation tasks. Through the linkage of traditional and new views, additional synergetic effects are achieved, that were positively acknowledged by domain experts (see Section 6.1.3).

⁸ Patent examiners, e.g. from the EPO, need to have a university degree in a technical field, speak at least the three European languages English, French, and German, and undergo intensive training during their first years of employment. This indicates that the learning curves for traditional patent analysis are high as well.

CHAPTER



Outlook

This thesis presented and discussed visual analytics approaches for patent search and analysis and related tasks. The described approaches take effect at different levels of granularity and abstraction.

The PatViz system spans the whole patent analysis process, focusing on iterative refinement of search queries and analytic feedback loops, which facilitate the transfer of insights made during the analysis back into earlier stages of the process. The approach aims, in particular, at increasing patent analysts' trust in their created search statements, which is one of their primary requirements to counteract the economic risks they are faced with. The high risk levels inherent to this industry and the associated desire to trust the analysis are more than ever visible in the ongoing 'Patent War' in the mobile industry¹. The trust building in PatViz is based on the interactive visual exploration of result sets, the derivation of insights from these sets, the combination of these insights, and their re-integration for improving search and analysis. However, still much effort is needed to improve patent research further. While PatViz supports search and analysis on the process level, taking into account patent metadata and content as well, the extraction of patent content must be a main concern for future research within this field. Still, understanding and assessing the complex patent content is an extremely time consuming task. Progress in this direction will also leverage better and more precise retrieval approaches, and has the biggest potential to speed up patent analysis and to make it more reliable. With user-steered classifier creation, one

¹ http://en.wikipedia.org/wiki/Patent_war

such promising, scalable approach was presented in this thesis. Future directions in intellectual property analysis also include the detection of emerging trends and finding promising areas where new developments and, consequently, patent applications could be directed. Again, visual analytics approaches seem to offer the ingredients needed to find such uncharted waters.

Research in the field of information retrieval and its tight, seamless integration into analytic tasks are one of today's big challenges taking into account the enormous amounts of digital documents produced. In this broader context, PatViz and the presented user-steered classifier creation technique contribute visual, interactive workflows and the integration of automated methods with human reasoning capabilities, thereby forming new visual analytics approaches to address this goal.

Currently employed (web) search approaches aim at increasing automation and personalization to guess what (casual) users are looking for based on their search behavior. A discussion is starting [Pariser, 2011] whether this restricts users to 'personal ecosystems of information' and how severe such effects are. As a consequence of this discussion, demands for more transparency on search results spring up. Visual analytics approaches are potentially suitable to provide and increase this transparency. In particular, this requires methods for capturing provenance and analytic reporting. Promising ideas on how such an expansion can be achieved with minimal additional effort required on the part of the analyst have been proposed in this thesis. However, there are obstacles to be overcome if such approaches are to find broad application. These include technical ones, such as scalability problems arising from close integration of visual and automatic techniques, but also factors, such as steadily increasing economic interests, availability of regulatory restrictions, etc.

Transparency of electronic transactions is generally becoming more important nowadays. Government institutions, companies, journalists, and scientists base their conclusions and decisions on the analysis of electronically available information. At some point, these decisions are likely to be the objective of assessment – be it in the context of law suits, as is the case with patents, ensuring compliance with the law and constitution, or compliance with scientific or institutional rules. At the same time large amounts of private information is passed to governments and companies, or is deliberately disseminated, and often becomes accessible through the Internet, which raises privacy concerns. Again, the data *and* the process of analysis on which a decision was based on will shift the balance here. If visual analytics approaches are going to be employed as the basis for important decisions of this kind, analytic provenance must be accessible as a basis for critical review and inspection. However, provenance alone, as presented in this work, will not be sufficient in such situations. Uncertainty of information regarding the automatic procedures employed to derive it from underlying data and facts, correctness of

visual representation, as well as the validity of human analytic reasoning based on their interpretation, will have to be assessed as well.

Taking this train of thought further, the aim of these efforts should be to achieve accountability [Weitzner et al., 2008]. In order to accomplish accountable analyses, considering only the abovementioned measures might not be sufficient. The visual analytics technique itself, in the version it was used to carry out the analysis, plus the data the analysis was based on, both also have to be preserved for later confirmability. Given the contemporary availability of cheap digital storage space, storing the visual analysis tool, the involved data, and analytic provenance information together for later examination might come into reach.

Since the analytic processing of big data becomes available, visual analytics will provide insights into (public) data in unprecedented detail and with negligible effort. It is to be hoped that the visual analytics community also takes into account the responsibilities that arise from the ability to make big data analytically available, and that it considers its consequences.

Bibliography

- World Intellectual Property Indicators - 2011 Edition. Technical report, World Intellectual Property Organization (WIPO), 2011. [page 32]
- C. Ahlberg and B. Shneiderman. Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 1994)*, pages 313–317, 1994. [pages 24 and 25]
- W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer Publishing Company, Incorporated, 1st edition, 2011. [page 12]
- D. Alberts, C. B. Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, and D. DeMarco. Introduction to Patent Searching. In M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*, pages 3–43. Springer Berlin Heidelberg, 2011. [pages 34, 51, 52, 54, and 80]
- J. Alsakran, Y. Chen, D. Luo, Y. Zhao, J. Yang, W. Dou, and S. Liu. Real-Time Visualization of Streaming Text with a Force-Based Dynamic System. *IEEE Computer Graphics and Applications*, 32(1):34–45, 2012. [page 41]
- R. Amar and J. Stasko. BEST PAPER: A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations. In *IEEE Symposium on Information Visualization (INFOVIS 2004)*, pages 143–150, 2004. [page 11]
- K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann. The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities. *Information Visualization*, 1(3-4):166–181, 2002. [page 20]
- K. H. Atkinson. Toward a more rational patent search paradigm. In *ACM Workshop on Patent Information Retrieval (PaIR 2008)*, pages 37–40, 2008. [page 33]
- R. Baeza-Yates. Visualization of large answers in text databases. In *ACM Workshop on Advanced Visual Interfaces (AVI 1996)*, pages 101–107, 1996. [page 25]
- R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York., 1999. [page 18]

- J. Barnes and P. Hut. A hierarchical $O(N \log N)$ force-calculation algorithm. *nature*, 324:446–449, 1986. [page 68]
- M. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424, 1989. [page 80]
- T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5):28–37, 2001. [page 55]
- J. Bertin. *Semiologie Graphique*. Gauthier-Villars, Paris, 1967. [page 13]
- E. Bertini and D. Lalanne. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery (VAKD 2009)*, pages 12–20, 2009. [page 94]
- E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose. Toolglass and magic lenses: the see-through interface. In *ACM SIGGRAPH Computer Graphics and Interactive Techniques*, pages 73–80, 1993. [page 12]
- R. Blanch and E. Lecolinet. Browsing Zoomable Treemaps: Structure-Aware Multi-Scale Navigation Techniques. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1248–1253, 2007. [page 64]
- K. Börner and C. Chen. Visual interfaces to digital libraries. In *ACM/IEEE-CS joint Conference on Digital Libraries*, pages 425–425, 2002. [page 41]
- K. Börner, C. Chen, and K. W. Boyack. Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1):179–255, 2003. [pages 21, 25, and 41]
- H. Bosch, J. Heinrich, C. Müller, B. Höferlin, G. Reina, M. Höferlin, M. Wörner, and S. Koch. Innovative filtering techniques and customized analytics tools. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*, pages 269–270, 2009. [page 77]
- H. Bosch, D. Thom, M. Wörner, S. Koch, E. Püttmann, D. Jäckle, and T. Ertl. ScatterBlogs: Geo-spatial document analysis. In *IEEE Conference on Visual Analytics Science and Technology (VAST 2011)*, pages 309–310, 2011. [pages 77, 78, 131, 132, and 142]
- M. Bruls, K. Huizing, and J. van Wijk. Squarified Treemaps. In *Joint Eurographics – IEEE TCVG Symposium on Visualization (VisSym 2000)*, pages 33–42, 2000. [page 121]

- C. J. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998. [page 21]
- S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo. VisTrails: visualization meets data management. In *ACM International Conference on Management of Data (SIGMOD 2006)*, pages 745–747, 2006. [page 127]
- C. Campbell, N. Cristianini, and A. Smola. Query Learning with Large Margin Classifiers. In *International Conference on Machine Learning*, pages 111–118, 2000. [page 23]
- S. Card, J. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999. [pages vi, 12, 13, 14, 15, 16, and 82]
- Y. Chen, J. Yang, and W. Ribarsky. Toward effective insight management in visual analytics systems. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 49–56, 2009. [page 130]
- Y. Chen, S. Barlowe, and J. Yang. Click2Annotate: Automated Insight Externalization with rich semantics. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2010)*, pages 155–162, 2010. [page 130]
- E. H. Chi. A Taxonomy of Visualization Techniques Using the Data State Reference Model. In *IEEE Symposium on Information Visualization (InfoVis 2000)*, pages 69–75, 2000. [page 13]
- N. Chinchor and W. A. Pike. The Science of Analytic Reporting. *Information Visualization*, 8(4):286–293, 2009. [page 129]
- A. Cockburn, A. Karlson, and B. B. Bederson. A review of overview+detail, zooming, and focus+context interfaces. *ACM Computing Surveys*, 41(1):2:1–2:31, 2009. [page 12]
- E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970. [page 23]
- J. Codina, E. Pianta, S. Vrochidis, and S. Papadopoulos. Integration of Semantic, Metadata and Image Search Engines with a Text Search Engine for Patent Retrieval. In S. Bloehdorn, M. Grobelnik, P. Mika, and D. T. Tran, editors, *SemSearch*, pages 14–28. 2008. [pages 44, 48, and 115]

- C. Collins, S. Carpendale, and G. Penn. DocuBurst: Visualizing Document Content using Language Structure. *Computer Graphics Forum*, 28(3):1039–1046, 2009. [page 25]
- M. Correll, M. Witmore, and M. Gleicher. Exploring Collections of Tagged Text for Literary Scholarship. *Computer Graphics Forum*, 30(3):731–740, 2011. [page 41]
- T. Cox and M. Cox. *Multidimensional scaling*, volume 88. Chapman & Hall/CRC, 2000. [page 20]
- N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000. [page 21]
- R. J. Crouser and R. Chang. An Affordance-Based Framework for Human Computation and Human-Computer Collaboration. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2859–2868, 2012. [page 161]
- W. Cui, H. Zhou, H. Qu, P. C. Wong, and X. Li. Geometry-Based Edge Clustering for Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1277–1284, 2008. [page 83]
- H. Cunningham, V. Tablan, I. Roberts, M. A. Greenwood, and N. Aswani. Information Extraction and Semantic Annotation for Multi-Paradigm Information Management. In M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*, pages 307–327. Springer Berlin Heidelberg, 2011. [page 55]
- G. Di Battista. *Graph drawing: algorithms for the visualization of graphs*. An Alan R. Apt book. Prentice Hall, 1999. [page 83]
- S. dos Santos and K. Brodlié. Gaining understanding of multivariate and multidimensional data through visualization. *Computers & Graphics*, 28(3):311–325, 2004. [pages 13 and 16]
- H. Dou, V. Leveillé, S. Manullang, and J. D. Jr. Patent analysis for competitive technical intelligence and innovative thinking. *Data Science Journal*, 4:209–236, 2005. [page 40]
- W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. Lipford, and R. Chang. Recovering Reasoning Processes from User Interactions. *IEEE Computer Graphics and Applications*, 29(3):52–61, 2009. [page 127]

- C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr. Rapid Understanding of Scientific Paper Collections : Integrating Statistics , Text Analytics , and Visualization. *Journal of the American Society for Information Systems and Technology*, pages 1–28, 2011. [page 41]
- S. G. Eick. Graphically Displaying Text. *Journal of Computational and Graphical Statistics*, 3(2):pp. 127–142, 1994. [page 25]
- G. Ellis and A. Dix. A Taxonomy of Clutter Reduction for Information Visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1216–1223, 2007. [page 83]
- R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems*. Addison-Wesley, Reading, Massachusetts, 2003. [page 23]
- N. Elmqvist, J. Stasko, and P. Tsigas. DataMeadow: A Visual Canvas for Analysis of Large-Scale Multivariate Data. *Information Visualization*, 7(1):18–33, 2008. [page 75]
- A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. Observation-level interaction with statistical models for visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST 2011)*, pages 121–130, 2011. [page 94]
- A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2012)*, pages 473–482, 2012. [page 41]
- R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008. [pages 96 and 119]
- J.-D. Fekete, J. J. Wijk, J. T. Stasko, and C. North. Information Visualization. chapter The Value of Information Visualization, pages 1–18. Springer-Verlag, Berlin, Heidelberg, 2008. [page 11]
- N. Fuhr and T. Rölleke. A probabilistic relational algebra for the integration of information retrieval and database systems. *ACM Transactions on Information Systems*, 15(1):32–66, 1997. [page 48]
- G. W. Furnas. Generalized fisheye views. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 1986)*, pages 16–23, 1986. [page 12]
- E. Gamma, R. Helm, R. Johnson, and J. Vlissides. Design patterns: Elements of reusable object-oriented design, 1995. [page 116]

- E. Gansner, Y. Hu, S. North, and C. Scheidegger. Multilevel agglomerative edge bundling for visualizing large graphs. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 187–194, 2011. [page 83]
- J. Gantz and D. Reinsel. Extracting value from chaos. *White Paper, IDC*, 2011. [page 1]
- J. Gibson. The concept of affordances. *Perceiving, acting, and knowing*, pages 67–82, 1977. [page 161]
- M. Giereth. *An Architecture for Visual Patent Analysis*. PhD thesis, Universität Stuttgart, 2012. [pages 55 and 62]
- M. Giereth and T. Ertl. Visualization Enhanced Semantic Wikis for Patent Information. In *International Conference on Information Visualisation (IV 2008)*, pages 185–190, 2008. [page 124]
- M. Giereth, S. Koch, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, and L. Wanner. *A Modular Framework for Ontology-Based Representation of Patent Information*, pages 49–59. IOS Press, 2007a.
- M. Giereth, S. Koch, M. Rotard, and T. Ertl. Web Based Visual Exploration of Patent Information. In *International Conference on Information Visualization (IV 2007)*, pages 150–155, 2007b. [page 62]
- M. Giereth, H. Bosch, and T. Ertl. A 3D treemap approach for analyzing the classificatory distribution in patent portfolios. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2008)*, pages 189–190, 2008a. [page 65]
- M. Giereth, M. Wörner, H. Bosch, P. Baier, and T. Ertl. Utilization of Semantic Annotations in Interactive User Interfaces for Large Documents. In H.-G. Hegering, A. Lehmann, H. J. Ohlbach, and C. Scheideler, editors, *GI Jahrestagung (2)*, pages 706–711. 2008b. [page 74]
- C. Görg, Z. Liu, N. Parekh, K. Singhal, and J. Stasko. Visual Analytics with Jigsaw. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2007)*, pages 201–202, 2007. [page 41]
- D. Gotz and M. X. Zhou. Characterizing Users’ Visual Analytic Activity for Insight Provenance. *Information Visualization*, 8(1):42–55, 2009. [pages 127 and 128]
- S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2008)*, pages 111–120, 2008. [page 147]

- R. Haber and D. A. McNabb. Visualization idioms: A conceptual model for scientific visualization systems. *Visualization in Scientific Computing*, pages 74–93, 1990. [page 13]
- S. G. Hart and L. E. Stavenland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, editors, *Human Mental Workload*, chapter 7. Elsevier, 1988. [page 156]
- M. A. Hearst. TileBars: visualization of term distribution information in full text information access. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 1995)*, pages 59–66, 1995. [page 25]
- M. A. Hearst. *Search User Interfaces*. Cambridge University Press, New York, NY, USA, 1st edition, 2009. [pages 18, 24, and 80]
- J. Heer, S. K. Card, and J. A. Landay. prefuse: a toolkit for interactive information visualization. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2005)*, pages 421–430, 2005. [pages 13, 42, and 68]
- J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala. Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1189–1196, 2008. [page 128]
- F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual Classifier Training for Text Document Retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012.
- N. Henry, J.-D. Fekete, and M. McGuffin. NodeTrix: a Hybrid Visualization of Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, 2007. [pages 83 and 84]
- D. Holten. Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, 2006. [pages 64 and 83]
- D. Holten and J. J. van Wijk. Force-Directed Edge Bundling for Graph Visualization. *Computer Graphics Forum*, 28(3):983–990, 2009. [page 83]
- D. Hunt, L. Nguyen, and M. Rodgers. *Patent searching: tools & techniques*. Wiley, 2007. [page 40]
- C. Hurter, B. Tissoires, and S. Conversy. FromDaDy: Spreading Aircraft Trajectories Across Views to Support Iterative Queries. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1017–1024, 2009. [page 83]

- C. Hurter, O. Ersoy, and A. Telea. MoleView: An Attribute and Structure-Based Semantic Lens for Large Element-Based Plots. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2600–2609, 2011. [page 90]
- A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *IEEE Visualization 1990*, pages 361–378, 1990. [page 91]
- T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning (ECML 1998)*, pages 137–142. 1998. [pages 21 and 94]
- T. Joachims. Transductive Inference for Text Classification using Support Vector Machines. In *International Conference on Machine Learning*, pages 200–209, 1999. [page 94]
- C. Jochim, C. Lioma, H. Schütze, S. Koch, and T. Ertl. Preliminary study into query translation for patent retrieval. In *International Workshop on Patent Information Retrieval (PaIR 2010)*, pages 57–66, 2010. [page 49]
- C. Jochim, C. Lioma, and H. Schütze. Expanding Queries with Term and Phrase Translations in Patent Retrieval. In A. Hanbury, A. Rauber, and A. Vries, editors, *Multidisciplinary Information Retrieval*, volume 6653 of *Lecture Notes in Computer Science*, pages 16–29. Springer Berlin Heidelberg, 2011. [page 49]
- H. Joho, L. A. Azzopardi, and W. Vanderbauwhede. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *Symposium on Information Interaction in Context*, pages 13–24, 2010. [pages 32, 35, and 139]
- P. Joia, F. Paulovich, D. Coimbra, J. Cuminato, and L. Nonato. Local Affine Multidimensional Projection. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2563–2571, 2011. [page 21]
- I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005. [page 20]
- S. Jones. Graphical query specification and dynamic result previews for a digital library. In *ACM Symposium on User Interface Software and Technology*, pages 143–151, 1998. [page 24]
- D. Jonker, W. Wright, D. Schroh, P. Proulx, and B. Cort. Information triage with TRIST. In *International Conference on Intelligence Analysis*, 2005. [page 25]
- N. Kadivar, V. Chen, D. Dunsmuir, E. Lee, C. Qian, J. Dill, C. Shaw, and R. Woodbury. Capturing and supporting the analysis process. In *IEEE Symposium on*

- Visual Analytics Science and Technology (VAST 2009)*, pages 131–138, 2009. [page 128]
- E. Kandogan and B. Shneiderman. Elastic windows: improved spatial layout and rapid multiple window operations. In *ACM Workshop on Advanced Visual Interfaces (AVI 1996)*, pages 29–38, 1996. [page 121]
- H. Kang, C. Plaisant, B. Lee, and B. Bederson. Netlens: Iterative exploration of content-actor network data. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2006)*, pages 91–98, 2006. [page 41]
- M. Kaufmann and D. Wagner. Drawing graphs: Methods and models (Lecture notes in computer science). 2001. [page 83]
- D. Keim and H.-P. Kriegel. VisDB: database exploration using multidimensional visualization. *IEEE Computer Graphics and Applications*, 14(5):40–49, 1994. [page 16]
- D. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in Visual Data Analysis. In *International Conference on Information Visualization (IV 2006)*, pages 9–16, 2006. [pages 15 and 160]
- D. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual Analytics: Scope and Challenges. In S. Simoff, M. Böhlen, and A. Mazeika, editors, *Visual Data Mining*, volume 4404 of *Lecture Notes in Computer Science*, pages 76–90. Springer Berlin / Heidelberg, 2008. [pages 14 and 15]
- D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering the Information Age: Solving Problems with Visual Analytics*. Eurographics Association, 2010. [pages 15, 23, and 126]
- S. Koch and H. Bosch. From Static Textual Display of Patents to Graphical Interactions. In M. Lupu, K. Mayer, J. Tait, A. J. Trippe, and W. B. Croft, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Kluwer International Series on Information Retrieval*, pages 217–235. Springer Berlin Heidelberg, 2011.
- S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative Integration of Visual Insights during Patent Search and Analysis. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*, pages 203–210, 2009. [page 79]
- S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative Integration of Visual Insights during Scalable Patent Search and Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):557–569, 2011.

- T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585, 2000. [page 21]
- R. Krüger, H. Bosch, S. Koch, C. Müller, G. Reina, D. Thom, and T. Ertl. HIVEBEAT - A Highly Interactive Visualization Environment for Broad-Scale Exploratory Analysis and Tracing. In *IEEE Conference on Visual Analytics Science and Technology (VAST 2012)*, pages 177–178, 2012. [pages 77 and 78]
- A. Lambert, R. Bourqui, and D. Auber. Winding Roads: Routing edges into bundles. *Computer Graphics Forum*, 29(3):853–862, 2010. [page 83]
- K. Lang. NewsWeeder: Learning to Filter Netnews. In *International Conference on Machine Learning*, pages 331–339, 1995. [page 148]
- J. H. Larkin and H. A. Simon. Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, 11(1):65–100, 1987. [page 82]
- D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 3–12, 1994. [page 22]
- H. Lipford, F. Stukes, W. Dou, M. Hawkins, and R. Chang. Helping users recall their reasoning process. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2010)*, pages 187–194, 2010. [page 129]
- S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration. In T. Gross, J. Gulliksen, P. Kotzé, L. Oestreicher, P. Palanque, R. Prates, and M. Winckler, editors, *Human-Computer Interaction (INTERACT 2009)*, volume 5726 of *Lecture Notes in Computer Science*, pages 392–404. Springer Berlin / Heidelberg, 2009. [page 72]
- C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. [pages 17, 18, 49, and 93]
- B. H. McCormick. Visualization in scientific computing. *SIGBIO Newsl.*, 10(1):15–21, 1988. [page 10]
- M. G. Moehrle, L. Walter, I. Bergmann, S. Bobe, and S. Skrzypale. Patinformatics as a business process: A guideline through patent research tasks and tools. *World Patent Information*, 32(4):291–299, 2010. [page 41]
- D. A. Norman. *The psychology of everyday things*. Basic Books, 1988. [page 161]

- D. A. Norman. *Things that make us smart: defending human attributes in the age of the machine*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1993. [page 125]
- D. A. Norman. Affordance, conventions, and design. *interactions*, 6(3):38–43, 1999. [pages 161 and 162]
- C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006. [page 141]
- C. North, A. Endert, C. Andrews, and G. Fink. The Visualization Pipeline is Broken. Technical report, Virginia Tech and Pacific Northwest Laboratories (in the context of the FODAVA initiative), 2010. [page 16]
- D. Oelke, D. Spretke, A. Stoffel, and D. Keim. Visual readability analysis: How to make your writings easier to read. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2010)*, pages 123–130, 2010. [page 41]
- F. Olsson. A literature survey of active machine learning in the context of natural language processing. Technical Report 06, Swedish Institute of Computer Science, 2009. [page 22]
- A. Paivio. *Mental representations: a dual coding approach*. Number 9 in Oxford Psychology Series. Oxford University Press, USA, 1986. [page 24]
- A. Panagiotidis, H. Bosch, S. Koch, and T. Ertl. EdgeAnalyzer: Exploratory Analysis through Advanced Edge Interaction. In *Hawaii International Conference on System Sciences (HICSS 2011)*, pages 1–10, 2011. [page 82]
- E. Pariser. *The filter bubble : what the Internet is hiding from you*. Viking, London, 2011. [page 166]
- F. Paulovich, L. Nonato, R. Minghim, and H. Levkowitz. Least Square Projection: A Fast High-Precision Multidimensional Projection Technique and Its Application to Document Mapping. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):564–575, 2008. [pages 21 and 100]
- F. V. Paulovich, M. C. F. Oliveira, and R. Minghim. The Projection Explorer: A Flexible Tool for Projection-based Multidimensional Visualization. In *Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*, pages 27–36, 2007. [page 100]
- P. Pirolli and S. Card. Information foraging in information access environments. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 1995)*, pages 51–58, 1995. [page 26]

- P. Pirolli and S. Card. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. *The Analyst*, pages 2–4, 2005. [pages vi, 4, 17, 26, 27, 28, 37, 77, 137, and 163]
- C. Plaisant. The challenge of information visualization evaluation. In *International Working Conference on Advanced Visual Interfaces (AVI 2004)*, pages 109–116, 2004. [page 137]
- C. Plaisant, J.-D. Fekete, and G. Grinstein. Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):120–134, 2008. [page 137]
- M. Pohl, S. Wiltner, and S. Miksch. Exploring information visualization: describing different interaction patterns. In *BEyond time and errors: novel evaluation methods for Information Visualization (BELIV 2010)*, pages 16–23, 2010. [page 127]
- M. Pohl, M. Smuc, and E. Mayr. The User Puzzle—Explaining the Interaction with Visual Analytics Systems. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2908–2916, 2012. [page 137]
- A. Potrich and E. Pianta. L-ISA: Learning Domain Specific Isa-Relations from the Web. In N. C. C. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, editors, *International Conference on Language Resources and Evaluation (LREC 2008)*, 2008. <http://www.lrec-conf.org/proceedings/lrec2008/>. [page 55]
- P. Proulx, S. Tandon, A. Bodnar, D. Schroh, R. Harper, and W. Wright. Avian Flu Case Study with nSpace and GeoTime. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2006)*, pages 27–34, 2006. [page 25]
- R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 1994)*, pages 318–322, 1994. [page 12]
- J. Roberts. State of the Art: Coordinated Multiple Views in Exploratory Visualization. In *International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, pages 61–71, 2007. [page 12]
- C. Rohrdantz, S. Koch, C. Jochim, G. Heyer, G. Scheuermann, T. Ertl, H. Schütze, and D. A. Keim. Visuelle Textanalyse. *Informatik-Spektrum*, 33:601–611, 2010.
- T. Rose, M. Stevenson, and M. Whitehead. The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *International Conference on Language Resources and Evaluation (LREC 2002)*, pages 29–31, 2002. [page 149]

- P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456, 2005. [page 141]
- P. Saraiya, C. North, V. Lam, and K. Duca. An Insight-Based Longitudinal Study of Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1511–1522, 2006. [page 148]
- C. Seifert and M. Granitzer. User-Based Active Learning. In *International Conference on Data Mining, Workshops*, pages 418–425, 2010. [page 94]
- D. Selassie, B. Heller, and J. Heer. Divided Edge Bundling for Directional Network Data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2354–2363, 2011. [page 83]
- B. Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. [page 22]
- B. Shneiderman. Dynamic Queries for Visual Information Seeking. *IEEE Software*, 11(6):70–77, 1994. [pages 25 and 75]
- B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, 1996. [pages 11, 15, and 160]
- B. Shneiderman and M. Wattenberg. Ordered treemap layouts. In *IEEE Symposium on Information Visualization (INFOVIS 2001)*, pages 73–78, 2001. [pages 62 and 121]
- Y. B. Shrinivasan and J. J. van Wijk. Supporting the analytical reasoning process in information visualization. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2008)*, pages 1237–1246, 2008. [page 130]
- T. E. R. Singer and J. F. Smith. Patentese: A dialect of English? *Journal of Chemical Education*, 44:111, 1967. [page 33]
- M. Smuc, E. Mayr, T. Lammarsch, W. Aigner, S. Miksch, and J. Gartner. To Score or Not to Score? Tripling Insights for Participatory Design. *IEEE Computer Graphics and Applications*, 29(3):29–38, 2009. [pages 141 and 142]
- A. Spoerri. InfoCrystal: a visual tool for information retrieval & management. In *International Conference on Information and Knowledge Management (CIKM 1993)*, pages 11–20, 1993. [page 24]

- J. Stasko and E. Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Symposium on Information Visualization (InfoVis 2000)*, pages 57–65, 2000. [page 89]
- C. Stolte and P. Hanrahan. Polaris: a system for query, analysis and visualization of multi-dimensional relational databases. In *IEEE Symposium on Information Visualization (InfoVis 2000)*, pages 5–14, 2000. [pages 13, 16, and 41]
- C. Stolte, D. Tang, and P. Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, 2002. [page 25]
- D. Tang, C. Stolte, and R. Bosch. Design choices when architecting visualizations. *Information Visualization*, 3:65–79, 2004. [page 13]
- D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 41–48, 2012. [page 131]
- J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005. [pages 9, 10, 14, 111, 112, 127, and 162]
- J. J. Thomas and J. Kielman. Challenges for visual analytics. *Information Visualization*, 8(4):309–314, 2009. [pages 10, 112, and 129]
- M. Tobiasz, P. Isenberg, and S. Carpendale. Lark: Coordinating Co-located Collaboration with Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1065–1072, 2009. [page 14]
- S. Tong and D. Koller. Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*, 2:45–66, 2000. [page 152]
- M. Tory and T. Möller. Rethinking Visualization: A High-Level Taxonomy. In *IEEE Symposium on Information Visualization (INFOVIS 2004)*, pages 151–158, 2004. [page 11]
- A. Trippe and I. Ruthven. Evaluating Real Patent Retrieval Effectiveness. In M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*, pages 125–143. Springer Berlin Heidelberg, 2011, 10.1007/978-3-642-19231-9_6. [pages 36 and 142]

- A. J. Trippe. Patinformatics: Tasks to tools. *World Patent Information*, 25(3): 211–221, 2003. [page 41]
- E. R. Tufte. *The visual display of quantitative information*. Graphics Press, Cheshire, CT, USA, 1986. [page 12]
- V. Vapnik. *Statistical learning theory*. Wiley, 1998. [pages 21, 22, and 94]
- F. Viégas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. ManyEyes: a Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007. [page 123]
- F. B. Viégas and M. Wattenberg. TIMELINES: Tag clouds and the case for vernacular visualization. *interactions*, 15:49–52, 2008. [page 70]
- T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. van Wijk, J.-D. Fekete, and D. Fellner. Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges. *Computer Graphics Forum*, 30(6):1719–1749, 2011. [page 83]
- E. M. Voorhees. The philosophy of information retrieval evaluation. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer Berlin Heidelberg, 2002. [page 142]
- S. Vrochidis, S. Papadopoulos, A. Moutzidou, P. Sidiropoulos, E. Pianta, and I. Kompatsiaris. Towards content-based patent image retrieval: A framework perspective. *World Patent Information*, 32(2):94–106, 2010. [page 54]
- L. Wanner, R. Baeza-Yates, S. Brüggmann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, G. Piella, I. Puhlmann, G. Rao, M. Rotard, P. Schoester, L. Serafini, and V. Zervaki. Towards content-oriented patent document processing. *World Patent Information*, 30(1):21–33, 2008. [page 39]
- C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004. [pages 12 and 24]
- C. Weaver. Building Highly-Coordinated Visualizations in Improvise. In *IEEE Symposium on Information Visualization (INFOVIS 2004)*, pages 159–166, 2004. [page 13]
- D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman. Information accountability. *Communications of the ACM*, 51(6):82–87, 2008. [page 167]

- N. Wirth. Systematic Programming: An Introduction. 1973. [page 45]
- J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Information Visualization*, pages 51–58, 1995. [pages 20 and 41]
- N. Wong and S. Carpendale. Supporting Interactive Graph Exploration Using Edge Plucking. In *IS&T/SPIE Symposium on Electronic Imaging: Visualization and Data Analysis*, 2007. [page 83]
- N. Wong, S. Carpendale, and S. Greenberg. Edgelens: an interactive method for managing edge congestion in graphs. In *IEEE Symposium on Information Visualization (INFOVIS 2003)*, pages 51–58, 2003. [page 83]
- P. C. Wong and J. Thomas. Visual Analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 2004. [page 9]
- P. C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner, and J. Thomas. IN-SPIRE InfoVis 2004 Contest Entry. In *IEEE Symposium on Information Visualization*, 2004. [page 20]
- P. C. Wong, H.-W. Shen, C. R. Johnson, C. Chen, and R. B. Ross. The Top 10 Challenges in Extreme-Scale Visual Analytics. *IEEE Computer Graphics and Applications*, 32:63–67, 2012. [page 112]
- W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort. The Sandbox for analysis: concepts and methods. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2006)*, pages 801–810, 2006. [page 130]
- W. Xu, M. Esteva, S. D. Jain, and V. Jain. Analysis of large digital collections with interactive visualization. In *IEEE Conference on Visual Analytics Science and Technology (VAST 2011)*, pages 241–250, 2011. [page 41]
- Y. Yang, L. Akers, T. Klose, and C. B. Yang. Text mining and visualization tools – Impressions of emerging capabilities. *World Patent Information*, 30(4):280–293, 2008. [page 41]