Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart
Universitätsstraße 5b
D - 70569 Stuttgart

Bachelorarbeit Nr. 348

Improving Author Co-Citation Analysis in Scientific Literature by Using Citation Function Classification

Constantin Seibold

Studiengang:	Informatik
Prüfer:	Prof. Dr. Sebastian Padó
Betreuer:	Dr. Roman Klinger
begonnen am:	31.5.2016
beendet am:	29.9.2016
CR-Klassifikation:	H.3.3, I.2.7, I.5.3

Erklärung (Statement of Authorship)

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollstndig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift

I hereby declare that the work presented in this thesis is entirely my own. I did not use any other sources and references that the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

Place, Date, Signature

Abstract

The concept of co-citation analysis is a possible approach for the interpretation of the relations between scientific papers or authors. Most of the previous work regarding author co-citation analysis, however, doesn't take the citation context into account. In this thesis, I propose a method for letting citation functions, which are functions that represent the intention of an author assigned to the corresponding references, directly influence the author co-citation analysis procedure. This approach is based on a faceted citation classification scheme, which allows comparisons between references. This should allow an easier representation of author groups, as authors, which are working together, usually share the same view on science and, therefore, are likely to be cited similarly. As there is no real gold standard for author groups, the evaluation of this approach tests the textual coherence of clusters created by this procedure based on the authors' oeuvres and compares the nationality of authors within clusters. The results indicate a correlation between author groups and similar citation functions.

Inhaltsangabe

Die Thematik der Co-Zitation Analyse beschreibt einen möglichen generellen Ansatz zur Interpretation von Beziehungen zwischen einer Menge von wissenschaftlichen Arbeiten oder Autoren. Die meisten Methoden zur Erstellung von Autorenclustern, die bisher zu diesem Themengebiet vorgestellt wurden, nehmen keinen Bezug auf den Kontext, in welchem die jeweiligen Zitationen vorkommen. In dieser Arbeit stelle ich eine Möglichkeit vor, durch die Zitationsfunktionen direkten Einfluss auf Autorcozitationsanalyse nehmen kann. Eine Zitationsfunktion gibt die Intention des Autors wieder, welche zu einer Referenz geführt hat, und ist somit an diese Referenz gebunden. Der hier vorgestellte Vorgang bezieht sich auf eine facettiertes Zitationsklassifikationsschema, durch welches Referenzen verglichen werden können. Dies sollte eine einfachere Möglichkeit darbieten Autorengruppen darzustellen, da diese, dadurch dass sie ähnliche Ansichtspunkte auf ihre Fachgebiete haben, auf ähnliche Art und Weise zitiert werden. Da es keinen richtigen Goldstandard für Autorengruppen gibt, bezieht sich die Evaluation auf den textbezogenen Zusammenhang zwischen den Oeuvres von Autoren und die Repräsentation von Nationalitäten innerhalb von Clustern. Die Resultate, die in dieser Arbeit gefunden werden, deuten darauf hin, dass es tatsächlich einen Zusammenhang zwischen Authorengruppen und ähnlichen Zitierweisen dieser Autoren gibt.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

A scientific publication is a document, which displays a certain view on a specific topic and allows the author to offer a new view or even a new feature concerning this topic. Hence every publication contributes at least a bit to a certain scientific society and so is part of a bigger picture, which continues to expand with each new publication in each corresponding field. And the growth of annualy new literature continues increasing with every year. While in 2000 it was an amount of nearly 1,250 more publications per year, the overall growth increased by 500 more publications to the date 2010 to an amount of roughly 1,750 additional publications per year as seen in 1 (Bornmann and Mutz, 2015). Bornmann and Mutz report that at the time of 2012

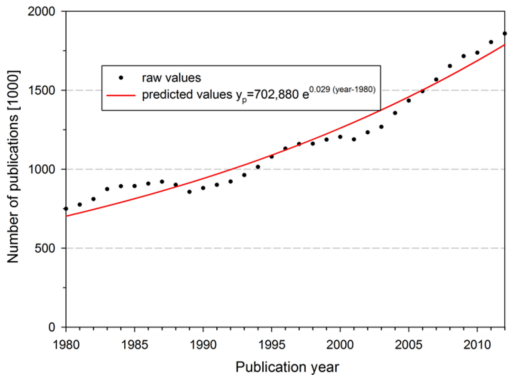

Figure 1: Statistics provided in Bornmann and Mutz (2015)

there was an approximately 3% growth of the global scientific publication output . However, while the amount of publication increases, the amount of references in total decreased since the beginning of the 21'st century. In this context an argument of Tabah (1999) gets mentioned:

„... the faster a literature or a given journal grows (...),
the more rapidly it ages“

The more rapidly a field ages, the stronger becomes the focus on recent publications, while older ones will rather be referenced by summarizing literature

and hence move into the background. With rapid growth, the fields will move away from fundamental work and towards more specific aspects. Fundamentals will be integrated into summarizing work or into more recent general publications and mostly work from those specific aspects will be investigated on, the amount of referenced work decreases. Hence for a publication to stay relatively relevant over time, it has to make some impact on its field. In the reverse, it is possible to conclude that writings, that had an above average influence in their regarding field of science, are also more likely to be cited more than the average document in that regard.
Emphases like this and other possible correlations between scientific publications based on its citations allow an analysis of different fields of science and grant the possibility to map its progression over time. This falls into the category of citation analysis.

Citation analysis, which gets described by Garfield et al. (1972), is a part of science, which specializes in the study of citations and references and how those reflect on different fields of science. It uses various methods to gain information from citations,which occurs when a document gets mentioned in another publication, references, which occur when a document mentiones another publication and related matters.
Many different approaches can be used to achieve a reasonable interpretation of such information. This work, however, focuses on a method, which is called co-citation analysis.
Co-citation analysis was first introduced in Small (1973) and used an approach published in Rosengren (1968). Co-citation counts for two different documents, how often those two documents are cited together. This can easily be achieved by comparing the list of documents, which reference those. The higher this co-citation count is, the more those two documents got cited together and if the co-citation count is above average, one can conclude, that those two documents have an above average correlation between one another. And documents part of document pairs achieving a high co-citation count will be cited therefore be cited relatively often, as documents are in most cases not referenced with the other document all the time.
This concept has not just restricted use on publications. White and Griffith (1981) presented a way to translate this method to analyse the relations between authors. There an author is defined by all of his writings, his oeuvre. And the concepts mentioned above still apply, as authors in this aspect are basically not considered as persons, but as collection of his publications. The first mapping of scientific authors by White and Griffith (1981) displayed

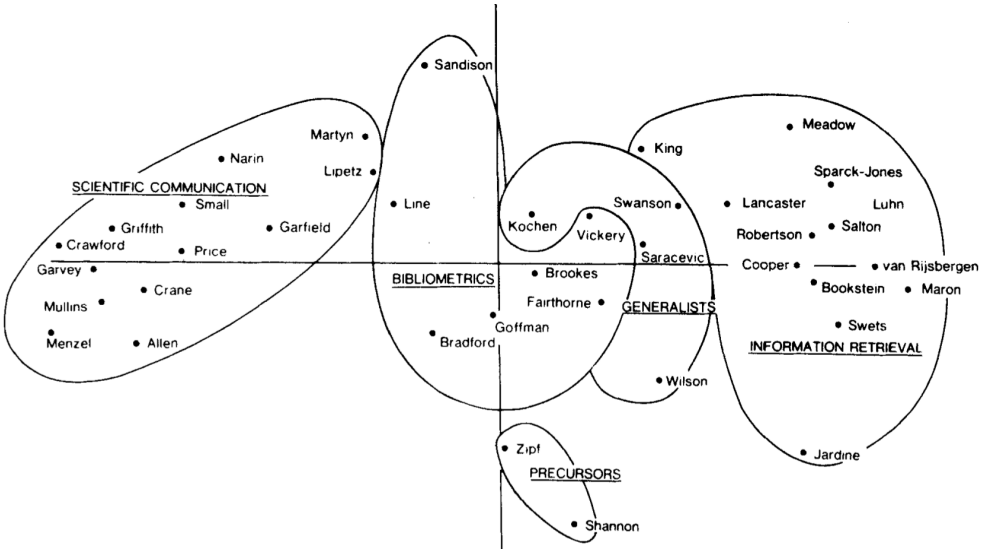

Figure 2: Author co-citation analysis in NLP literature from White and Griffith (1981).

in figure 2 shows a representation of 39 most cited authors in the field of NLP literature. Hereby authors in that focused on roughly the same field are shown closer to each other. By just using the sheer co-occurance of authors a relatively accurate representation was made, which shows the potential of this method.
But even while having the possibility to create maps of fields of science, citation analysis has clear problems as stated in MacRoberts and MacRoberts (1989). Problems like biased citing, citing based on self interest and similar are mentioned. Those also translate on its procedures, like co-citation analysis. Those problems exist, as the way of referencing of authors differs from the traditional concept of scientific work. Hereby an author is seen as a direct medium of science. Only portraying what is relevant, while being a passive observer. This, however, is not the case.
It can be seen that science is influenced by author's subjectivity, historical aspects and the social network, which surrounds each field of science.
As original citation analysis approaches only factored in if,how and where documents or authors are referenced, they do not take into consideration, in which context those citations are delivered, in which relation the authors are, if an author cites himself and many other factors.

Those problems raise questions like:

- What could be possible ways to adjust those methods to fix these problems?
- Would the influence of adjustments even produce different results?
- If they do, what will those results imply?

This thesis will investigate the influence of citation context in the field of co-citation analysis and evaluate the differences and possible advantages and disadvantages compared to a standardized co-citations analysis approach. The underlying idea is that documents and authors cited in a similar way have a stronger connection than those, that are cited in completely different ways. The main research question will be to find an appropriate way to include the citation context and how it influences the co-citation process.

1.2 Background

Every variation of co-citation analysis follows the same procedure.
At first an $author \times author$ matrix raw gets arranged by calculating the co-citation count, which is the amount of documents, which cite both authors of each author pair, for each author pair:

$$raw_{i,j} = \#documents\ that\ cite\ both\ i\ and\ j, \qquad i, j \in author$$

As the co-citation count between two documents is the same for $raw_{i,j}$ and $raw_{j,i}$, the resulting matrix is symmetrical. Because of this it is sufficient to only save upper triangular matrix, as it completely represents all co-citation counts between authors.

Based on this matrix a new matrix Cor also of size $author \times author$ containing the correlation c between all author pairs:

$$cor_{i,j} = c(i, j)$$

The correlation between the authors can be calculated using one of the various methods like Pearson's r, K50 or others. There has been some research on

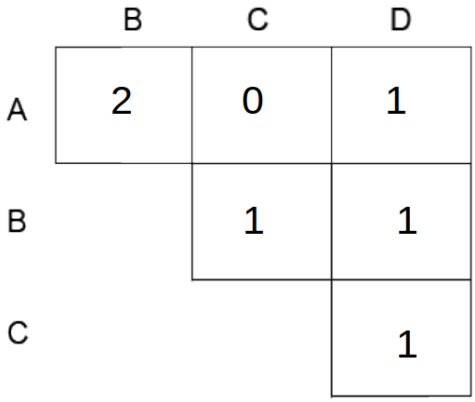

Figure 3: Example for a co-citation matrix *raw* consisting of four authors using the standard approach of author co-occurence in reference lists.

the improvement of the similarity measures needed for the co-citation clustering. In Boyack et al. (2005) several measures were compared, showing that similarity measures like Pearson's r have weaknesses as they are either not much scalable or do not provide a balanced clustering of documents. This work suggests a new normalized frequency measure based on co-citation. The so-called K50 measure is a method similar to a cosine measure. It is supposed to weight relationships between smaller journals stronger if they were above an expected value, to prevent larger journals from overshadowing. Leydesdorff compares Jaccard's similarity measure with Salton's cosine and Pearson's r in a web environment. There he notes that the Jaccard index is a proper basis for normalization as long as the *raw*-matrix is only based on co-occurrence since the Jaccard index does not take the complete rows and columns into account. The cosine should rather be used for visualization. Egghe and Leydesdorff show the relation between Pearson's r and Salton's cosine measure in the context of author co-citation analysis.
The *cor*-matrix allows the creation of a distance matrix *dis*, by using a distance measure d based on correlation:

$$dis_{i,j} = d(c_{i,j})$$

The distance matrix *dis* can then allow multiple clustering methods, i.e. k-means, single linkage, etc. The created clustering then provides a mapping of the fields of science.

While working on the similarity measures can improve the end result, it is not

the only way to improve clustering. Another way to advance the methods of co-citation analysis is to choose different heuristics to create the *raw*-matrix, which will be used as a basis for the correlation matrix. One of the first variations came from Small (1997). He proposes methods such as fractional citation counting or ordination by triangulation.

However, those types of improvements do not work on the conceptual problems of co-citation analysis like the absence of the use of context. A publication that makes use of document context is from Boyack et al. (2013). By partitioning a document into a set of blocks and analogously converting each citation's position in the text into the position according to each block a new kind of *raw*-matrix is built. Based on the distance in the text between citations the co-citation weights are measured. The idea hereby is that two documents or authors have a stronger connection if they are closer to each other in a document's text. The resulting clustering showed improvements to the standardized method in terms of textual coherence.
But a citation can come in different possible ways. With a reference an author might show approval or disapproval concerning an idea or method, might show the use of previous results to compare his work to or the usage of a tool. By just taking the distance between those references into account, it is not clear, if the referenced authors are coherent in any way, as it might be by using the authors intentions.
Small (2011) investigates another way to include more context into co-citation analysis. He uses the concept of Teufel (2010) and Teufel et al. (2006) of citation function classification methods and tries to figure out how in terms of co-citation analysis those may help in the understanding of maps of science and how citation sentiment is connected to such structure.
Citation classification is a relatively new field. It focuses on the automatic assignment of citations to specific classes, which are most of the time a certain sentiment. Herefore it uses information given by the sentence, in which the citation occurs (Teufel et al. (2006),Teufel (2010)), or its related segment Athar and Teufel (2012). Previous work has hereby focused on the assignment of references to a certain sentiment or a citation function.
A citation function is a function assigned to a reference, with the intent to represent the authors motives for citing this work.
Teufel et al. (2006) split citation functions into four major categories.
Direct weakness, if the reference sentence shows any disapproval towards cited work, comparison, if the referenced work gets compared to the authors work or work of a third party, praise, if the author builds up his own paper

on the referenced work, as according to Teufel et al. (2006) there is no higher praise than building up on another authors work and lastly a neutral element. The neutral element is selected, if a reference sentence does not show any or not enough evidence that the author had any specific motivations for this reference.
Another way to assign classification functions was shown by Jochim and Schütze (2012). He relied on classification based on a citation class scheme from Murugesan and Moravcsik (1978).
This scheme allows differentiation between four different facets with binary classes, which help to specify the context of the citation. By assigning a citation four classes instead of just one, the classification problem becomes less complex since the classifier has to only decide between two different classes for each facet instead of several.

1.3 Goals of this Work

An author group is a set of authors, which are connected to each other author in that group in a more direct way of interaction. This can be the case by working together, having build friendships or similar.
By working together or similar a certain bias towards those each group emerges, as authors within an author group share roughly the same ideas and work in similar ways. As a result of this each author within such an author group will also reference work of other authors within this group. And as those authors usually work on the same fields with the same basic ideas, they will also often be referenced together and in a similar way as the topical difference between them is rather small.

This bias was one of the problems named by MacRoberts and MacRoberts (1989). Previous co-citation procedures are not taking this kind of problem into account. In this work, we will investigate if the citation function shows any coherence to this kind of bias and if they can be used to represent author groups more accurately.
Hence this thesis will revolve around the construction of various weighting schemes for the creation of author co-citation matrices, by taking the citation functions into account. For this purpose, all the citations get assigned with specific citation function. The focus for that will be on the classification scheme of Murugesan and Moravcsik (1978).

Since it is a relatively reasonable conception that authors within author groups would be cited in a similar way, the assumption is made that by including the citation context provided by the citation scheme of Murugesan and Moravcsik, it should be possible to enable a more accurate representation of science with a stronger focus on author groups. By letting citation function directly influence co-citation analysis through a weighting function, it would based on this assumption increase the correlation between authors within each author group, as authors sharing the same idea are more likely to be cited in the same publication. Resulting to this, author groups should be represented more accurately.

1.4 Outline of this theses

This thesis is structured as follows:

In chapter 2 all fundamentals, which are necessary to understand for this thesis, are explained. Those revolve around more common techniques that appear in the fields of information retrieval and machine learning. It involves classification, clustering and the evaluation of such techniques.

In chapter 3 the approach of this thesis will be explained. It focuses on the citation function classification as well as weighting schemes for the clustering process and the clustering process itself. It shows the steps, which were used to classify citation context and let it influence the clustering process.

This approach will then be used in an experiment, which will be expanded on in chapter 4. This involves the corpora, their classification and lastly the co-citation clustering with multiple weighting schemes. The results will demonstrate the national and textual coherence between multiple clusterings as well as other details concerning the classification and clustering.

In chapter 5 the experiment results, as well as the overall approach, will be discussed and a potential outlook concerning co-citation analysis will be mentioned.

I will end this work with a short summary in chapter 6.

2 Fundamentals

This chapter will describe some commonly used methods, which will influence the course of this thesis, in the field of maschine learning and text analysis.

2.1 Classification - Maximum Entropy Classifier

In this context, the term „classification“describes the problem of assigning a set of data with a category depending on previous data, which was provided in a so-called training set.
Since the process is built on the analysis of previous data, to then use the given information to make new decisions, classification is generally interpreted as part of the field of supervised learning. In order to classify commonly a mathematical learning function, which will get directly influenced by the already classified data appearing in the training set, will be used. This function is called a „classifier“. Its purpose is the mapping of unassigned data to a class. This function can be implemented manually, through rules or by calculating the likelihood for all classes for a certain case and then assigning this case to the class with the highest likelihood.

This work uses the so-called *Maximum Entropy Classifier*. It refers to the publication of Manning and D.Klein (2003). The *Maximum Entropy Classifier* is a classifier based on conditional probability. The classifying function makes the use of features and their parameters, which are calculated through the use of the training data sets.
A feature f is for this classifier usually seen as a logical function, that returns the value 0, if the result is 'false', or 1 if it is 'true', with the input being a specific class and data, that is to classify. It describes a relation between a class $y \in Y$ and underlying data $\mathbf{x} \in X$. Each feature gets combined with a weight parameter λ, which represents the relation between the class and the data, hereby negative values point to an unlikeliness to appear together, positive values show a certain likelihood between those. The addition of each feature increases the maximum likelihood of the data, however, decreases the amount of the maximum entropy. It is the goal to create a balance, which maximizes the entropy without, while also maintaining a relatively high maximum conditional likelihood of data.

To calculate the probability of a class for given data every feature has to be considered.

$\lambda_1, \lambda_2, ... \in \lambda$, where λ is the set of all weight parameters and each λ_i is assigned to the corresponding f_i.

$f_1, f_2, ... \in f$, where f describes the set of all defined features for a classification process.

The data $\mathbf{x}$ is then assigned to the class y that maximizes this probability function

$$p_\lambda(y|\mathbf{x}) = \frac{e^{\sum_i \lambda_i f_i(y,x)}}{\sum_{s'} e^{\sum_i \lambda_i f_i(y,x)}}. \tag{1}$$

The more complex part of this classifier is to learn the optimal weights λ for each feature.
Those can be obtained by maximizing the logarithmic conditional likelihood

$$\log p(Y|X, \lambda) = \sum_{x,y \in X,Y} \log p_\lambda(y|\mathbf{x}) \tag{2}$$

and therefore maximizes the probability of each class with each already classified part of the data set.

2.2 Clustering - Complete linkage

„Clustering“ describes a task to create a partition p of a set of objects O, which is dependent on the distance d between those objects. The distances can usually, be gathered from a matrix consisting of $O \times O$. Each entry equates to the distance between two objects, that are represented by row and column. In this thesis we will focus on symmetrical matrices, therefore w.l.o.g the distance $d_{i,j} = d_{j,i}$. A symmetric matrix allows to save only the part of the matrix above the diagonal, as the entries below the diagonal are also represented in the upper part. The diagonal will be excluded since the distance of an object to itself is here considered 0. There are various different methods to create clusters based on the closest distances between objects.

We will focus on a hierarchical approach called *complete linkage clustering.*

This bottom-up variant of *complete linkage clustering* as it is a hierarchical approach begins by referring to each object $o \in O$ as a cluster $c \in P$ only containing o, while $\bigcup_{c \in P} c = O$ and $\forall c_i, c_j \in P$, $c_i \cap c_j = \varnothing$ are maintained.
In each following step, there will be a concatenation between clusters depending on the minimal maximal saved distance between the clusters. So for every cluster, the maximal distance to each other cluster will be calculated. The cluster pair that holds the minimal maximum distance is then to be combined.

$$\underset{\{c_i,c_j\},c_i,c_j \in P, i \neq j}{\arg\min} f(c_i, c_j) \tag{3}$$

$$f(c_i, c_j) = \max_{n \in c_i, m \in c_j} d_{n,m}$$

This step will be repeated until a certain threshold gets exceeded or there is only one cluster left.

2.3 Evaluation

In this section several methods for the evaluation or validation of clusterings or classifications, that will be used later on. will be presented.

2.3.1 Cross-validation

The x-fold cross-validation as portrayed in Kohavi et al. (1995) is an accuracy estimation method. It can, therefore, be used to evaluate how good a classifier with a certain feature set can be. The advantages of this validation method compared to others, is that it can be used with a relatively low amount of data, as it works on the training set itself and so does not compromise between choosing a partition of a data set into training set and validation set, which can result in the loss of testing or classifying capability in case the available amount of data is small.

It partitions a dataset D, which used as training set for prediction methods like a classifier, randomly into x parts. This partitioning $\mathrm{Pt} = \{d_1, ..., d_x\}$ contains parts of nearly equal size and it fulfills following requirements $\forall d_i, d_j \in \mathrm{Pt}$, $d_i \cap d_j = \varnothing$ and $\bigcup_{d_i \in Pt} d_i = D$.

For each round of the validation process the prediction method then gets trained by using $x-1$ of the subsets as training set and uses the one subset, that has not been chosen as training data, as a test set. The results will be validated and to minimize variability this procedure will be performed x times, with the used training set and validation set rotating with each round so that every subset will be used as validation set once. The resulting accuracy of each step will added up and get averaged.

2.3.2 Rand index

The *rand index* is an evaluation method normally used for cluster evaluation. It does this by comparing pairings of objects to see if they belong and if they were actually grouped together.
Every object pairing will be observed. By using a gold standard data set, which assigns each object to an optimal class based on a specific assumption, it is possible to calculate the number of times objects were clustered accurately and the number of times they were not. A pairing will be seen as a true positive (tp), if both objects of a pairing share the same class and were assigned to the same cluster. It will be seen as true negative (tn), if both objects neither have the class nor their cluster in common. The rand index is then calculated through

$$ri = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{tn} + \mathrm{fp} + \mathrm{fn}} = \frac{\mathrm{tp} + \mathrm{tn}}{\binom{n}{2}}. \tag{4}$$

False positives(fp) and false negatives (fn) are only defining part of the equation since the normalizing factor consists of the amount of all pairings $\binom{n}{2}, n = |O|$, which is equivalent to $tp+tn+fp+fn$. The closer the rand index value gets to 1, the closer the clustering is to a perfect mapping of objects to a cluster.

2.3.3 F_1 score

This method has the purpose of evaluating how good a binary classification is for a set of items. It finds uses in information retrieval to evaluate searches and classifications.

The F_1-*score* f_1 consists of both precision pr and recall re and is calculated as following:

$$f_1 = \frac{2 \cdot \mathrm{pr} \cdot \mathrm{re}}{\mathrm{pr} + \mathrm{re}} \tag{5}$$

Precision shows the relations between how many items have been assigned correctly to the class in focus and how many were assigned it, but however should not have been. The recall the relation between the amount of correctly assigned items and the amount of objects that should have been assigned to this class but were not.

$$pr = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}} \tag{6}$$

$$re = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}} \tag{7}$$

This method shows how accurate a classification is for one set. The better a classification is, the more true positives there are. Hence the closer $F1$ score gets to 1, the more accurate the classification is.
However when examining a greater number sets, that have to be evaluated, one has two different ways of averaging the values of each case.
The way of macro-averaging computes the f_1-scores of each case and averages these scores, while when micro-averaging computes for all cases the amount of true positives, false positives, and false negatives and computes the f_1-score based on these values.

2.3.4 Textual coherence

Textual coherence can be measured through many different ways, with each having its pros and cons.
Instead of examining the average cosine similarity between and within clusters or other methods, this work concentrates on a concept shown in Boyack and Klavans (2010). It uses the Jensen-Shannon divergence (Lin, 1991) as a basis to calculate the textual coherence. The coherence between objects in a clustering is hereby greater, the greater the coherence value becomes.

As a divergence measure, the value of the Jensen-Shannon divergence converges to 1, if the text related to the objects in a cluster are relatively different

and to 0, if the set of words doesn't differ a lot. It is based on the Shannon Entropy and Jensen's inequality.
Entropy H_n, $n \in \mathbb{N}$ of a discrete random variable on a limited set of values A as defined by Shannon portrays the expected information of words $a \in A$.

$$H_n = -\sum_{a \in A^n} p_a \cdot \log(p_a) \tag{8}$$

The entropy for the cases $\lim_{p \to 1} p \cdot \log p = 0$ and $\lim_{p \to 0} p \cdot \log p = 0$ shows that single words, which will appear in nearly every occasion or words that appear so rare that one appearance is nearly negligible, hold nearly no information.
Whereas Jensen's inequality states that for a real convex function f and non-negative weightings λ, with each assigned to another variable $x \in \mathbb{R}$, with $\sum \lambda = 1$ applies:

$$f(\sum \lambda \cdot x) \leq \sum \lambda \cdot f(x) \tag{9}$$

By combining those methods Lin (1991) provides the Jensen-Shannon divergence jsd, which in a textual environment concerning clusters can be seen as **?**:

Be Voc_c the vocabulary of all words build over a single cluster $c \in P$, and p_d, q_c probabilistic vectors. While p_d represents the probabilities $p_d(w)$ of all $w \in \text{Voc}$ for a single document $d \in c$, q inherits the probabilities $q_c(w)$ the words for the whole cluster.
The Jensen-Shannon divergence for a cluster will then calculated by taking the average of the divergence of the documents in the cluster.

(10)

$$\text{jsd}_{c\_d}(p_d, q_c) = \frac{\sum_{i=1}^{|\text{Voc}|} p_d(\text{Voc}_i) \cdot \log(\frac{2 \cdot p_d(\text{Voc}_i)}{p_d(\text{Voc}_i) + q_c(\text{Voc}_i)}) + q_c(\text{Voc}_i) \cdot \log(\frac{2 \cdot q_c(\text{Voc}_i)}{p_d(\text{Voc}_i) + q_c(\text{Voc}_i)})}{2}$$

$$jsd_c = \frac{\sum_{d \in c} \text{jsd}_{c\_d}(p_d, q_c)}{|c|}$$

The textual coherence for a single cluster as in Boyack and Klavans (2010) is then calculated in relation to the size of the cluster. Because clusters with a greater size usually consist of a higher variety of words, the coherence of a

cluster c will be seen as the difference between the jsd_c-value and the jsd-value of a cluster, which has been arranged completely randomly and has the same size as c. This value based on a random cluster will be referred to as jsd_{random}.

$$Coh = \frac{\sum_{c \in P} |c| \cdot \mathrm{Coh}_c}{\sum_{c \in P} |c|} \tag{11}$$

$$Coh_c = \mathrm{jsd}_{\mathrm{random\_}|c|} - \mathrm{jsd}_c$$

3 Approach

This chapter will describe the followed approach, that was used during this thesis.

This work describes a procedure, which will use information given in citation sentences, to create multiple different clusterings. This requires multiple steps. First of all the data has to be extracted and classified. Then the weighting schemes, which will be used to create the matrices, that are then used to calculate the correlations between authors, will then be shown and the thought process behind those will be explained. The next step would be the clustering approach itself. This includes the creation of the regular matrices as well as the correlation matrices and clustering method. The last step would be the evaluation. There the clustering results according to each weighting scheme will be compared to each other in term of contenxual connection as well as a regional connection.

3.1 Citation function classification

The used classification scheme equates to the faceted classification scheme provided by Murugesan and Moravcsik (1978). In comparison to many other classification schemes this scheme shows two major differences.
The first is the faceted structure, with each facet being independent from every other one. Due to this the classification is rather easy as it is only a decision between two classes for each facet. This facetted approach also allows for a relatively easy possibility for comparisons between citation functions, as citation functions that have more facets the same are bound hence cited in a more similar way. This type of comparison can become more difficult if the citation function is only based on a single assignment of a class, since the differences between those classes are not easy to identify and would, therefore, result into difficulties for creating a proper weighting scheme.
The other difference is the absence of a neutral class, which will be assigned if not enough evidence towards another certain class is shown. As most scientific publications stay try to stay politcally correct and therefore do not want to show any specific sentiment regarding a reference the huge majority of

references would be assigned to the neutral class as seen in Figure 4 in 4. With nearly two thirds of all references being assigned a neutral function a

Neut	PUse	CoCoGM	PSim	Weak	PMot	CoCoR0	PBas	CoCoXY	CoCo-	PModi	PSup
62.7%	15.8%	3.9%	3.8%	3.1%	2.2%	0.8%	1.5%	2.9%	1.0%	1.6%	1.1%

Figure 4: Distribution of classes of Teufel's classification scheme(Teufel et al., 2006).

comparison can become relatively pointless, as no new information will be won through it. In addition to that it is not known to which other class a reference assigned with neutral element would even be close to.
As Murugesan and Moravcsik do not fall back on such a neutral class some information can always be gained by the comparison of citation functions.

3.1.1 Annotation scheme

The annotation scheme of Murugesan and Moravcsik (1978) consists of four facets, which are representing binary classes, that were defined as followed:

- **Conceptual** or **Operational**
  The first class differentiates between „**Conceptual**"and „**Operational**", abbreviated „**Conc**"and „**Op**". A reference can be seen as conceptual, if an idea or concept was used in a way to develop the authors thought process or state an alternative understanding. However if a citation describes a used tool or dataset, which helps the author to demonstrate his finding, it is rather looked at as operational citation.

- **Organic** or **Perfunctory**
  Hereby a reference is considered as „**Organic**", short „**Org**", if the cited work build a basis for the citing work and as „**Perfunctory**", short „**Perf**", if the work cites alternative approaches or other types of references, which will not really be fleshed out textually in the citing work and therefore do not seem as necessary to the author as „organic"citations.

- **Evolutionary** or **Juxtapositional**
  A reference assigned with „**Evolutionary**", abbreviated „**Evol**"generally

indicates, that the author builds builds up on the cited work and in that way helps him to develop his thought or process.
If it is assigned with „**Juxtapositional**“, abbreviated „**Jux**“, it suggests the cited work to not contribute a lot to the citing work at the given position in the text.

- **Confirmative** or **Negational**
  The last class describes if the author agrees or disagrees with cited work. Hence a reference gets the class „**Confirmative**“, abbreviated „**Conf**“, if the author agrees with the cited work and „**Negational**“, abbreviated „**Neg**“, if he disagrees,isn't sure about its correctness or even claims it to be incorrect.

It is undeniable, that a certain connection between those classes exists and that those facets are not evenly distributed in the scientific literature. There will be fewer cases, in which an author will base his work on a reference, he disagrees with and more, where he agrees with cited work. This describes a connection between „**Neg**“/„**Conf**“and „**Evol**“This is just one example of such connections. But even though there are relations between classes, it doesn't mean, that one class appears exclusively with one another.
Some of those classes will also be harder for a computer to classify than for a person since some require more insight either in the field of science, the citing work is part of, or in sentence construction in scientific papers. The classes affected the most by such problems are „**Org**“-„**Perf**“and „**Conf**“-„**Neg**“.
The reason this annotation scheme got chosen over the likes of Teufel et al. (2006), etc. was that, while sentiment can help to indicate author network, sentiment would normally not be easy to filter out, because scientific literature has the tendency to stay overall neutral and project everything in a neutral way. Therefore the sentiment would most of the time end up being neutral, which can be seen in Teufel et al. (2006), where more than half of all citations appear in a neutral sentiment, while Murugesan and Moravcsik's annotation scheme always has a rather descriptive way to annotate citations, without relying on the author need to express any sentiment.
This, as opposed to sentimental annotations, would therefore rather result in a clustering, which will emphasize more on groups, that are more likely to have a contextual relation, than a sentimental one.

3.1.2 Feature selection

A comparison of various different classification features was made by Jochim (2014) and he evaluates how useful those can be in an environment based on Murugesan and Moravcsik's classification scheme. He examines features based on frequency, sentiment, linguistic structure, textual location, word-level linguistic and lexical fundamentals on one dataset, which consists of the ACL Proceedings from the ACL Anthology Regerence Corpus of the year 2004, which has been taken from a corpus consisting of all documents from the ACL Anthology Reference Corpus to the date of February 2007, collected in Bird et al. (2008). His results show, that each facet has different emphasis on a feature type. The highest accuracy for „**Conc**"-„**Op**"was provided by lexical features, for „**Evol**"-„**Jux**"by frequency type features, for „**Org**"-„**Perf**"by location type features and for „**Conf**"-„**Neg**"by structural ones.
He presents that for the classes „**Conc**"-„**Op**"and „**Conf**"-„**Neg**"a mixture of 19 different features from all categories, whereas the usage of all, which are 42 different features, described features wielded the best result for „**Perf**"-„**Org**"and „**Evol**"-„**Jux**".
Both of these approaches need a lot of effort since all those features must be evaluated before classifying the reference. When comparing his model against a standard bag-of-words model, which was used as a baseline for his model, it is notable that the bag-of-words model isn't significantly worse, even while showing a bit poorer results.
Therefore the bag-of-words feature model will be used by a *Maximum Entropy Classifier* following the implementation of Manning and D.Klein (2003).

3.2 Weighting scheme

This section focuses on different weighting schemes used, which were used for calculating the similarity between authors.

The standard variant used in co-citation analysis for the creation of a co-citation matrix, consisting of $Author \times Author$, which shows the co-citation count of all author pairs, which is calculated by the amount of documents, in which two authors are cited together, over all used documents. This allows for the observation on how authors are cited together, as well as on how an

author is cited overall. One can then estimate roughly in which category the cited authors fall and how those authors might stay in a work relation with one another.
But since an usual author cites more people than just the ones, he stays in contact with, and generally has to resort to fundamental documents or has to integrate new discoveries, the mapping based on such co-citation count table, without any particular weighting, will more likely result in just a representation fields of science and less in a representation of author workgroups.
With the assumption being that author research groups will have more likely similar idea and approach on their research subjects, the usage of a citation function in form of Murugesan and Moravcsik's citation scheme should lead to results, which are more likely to represent research groups.
The base idea, which will serve as basis for the different weigting schemes, would be, that author, which stay in a connection with each other, will usually be cited by the citing author in a similar way or context, as authors of one author group usually discuss related subjects. The citing author, if he wants to go into detail in one part of his work, will usually cite multiple sources to elaborate his thought process. This will show a connection between multiple cited authors concerning the elaborated subject. The difference between those will, therefore, become visible by their annotated citation function.
So the first weighting scheme will weight the co-citation count between two cited authors in one document more the more the citation function between those matches. An approach will be looked at in various ways. There will be a multiplicative and additive way with both high and low values to compare the impact the weighting will have. Each approach will use the average over all citations pairings occurring in a document between an author pair. The reason for this is that every citation in a document has to be valued. But just going by citation function alone might lead to a skewed result. Because of this, there will be two base ideas, albeit the focus will stay on the first:

The following fundamental definitions apply for the whole section 3.2.

A be a set consisting of all considered authors of a dataset and D a set of all considered documents. Then be $i, j \in A$ with $i \neq j$ and $d \in D$. r be a function, which returns the reference list for a given document d. cit be a function, which returns a set of citation functions Cit_{i_d} with $cf_{i\_d\_1}, cf_{i\_d\_2}, ... \in Cit_{i\_d}$ being a single citation function, which describes the four facets described above in the form of $cf_{i\_d\_n} = \{cf^1_{i\_d\_n}, cf^2_{i\_d\_n}, cf^3_{i\_d\_n}, cf^4_{i\_d\_n}\}$ with

$$cf^1_{i\_d\_n} = \begin{cases} Conc & \text{,if the citation was conceptual} \\ Op & \text{,if the citation was operational} \end{cases},$$
$$cf^2_{i\_d\_n} = \begin{cases} Evol & \text{,if the citation was evolutionary} \\ Jux & \text{,if the citation was Juxtapositional} \end{cases},$$
$$cf^3_{i\_d\_n} = \begin{cases} Org & \text{,if the citation was organic} \\ Perf & \text{,if the citation was perfunctory} \end{cases},$$
$$cf^4_{i\_d\_n} = \begin{cases} Conf & \text{,if the citation was confirmative} \\ Neg & \text{,if the citation was negational} \end{cases}$$

, of author i for a document d.
m_{simple} be a function that compares one of four facets for two citations, with the input being two citation functions of two different authors and returns 1 if this facet is equal for both citations and 0 if it isn't. Whereas m_{com} be a function, which returns 1, if the two citations match for either `Conc`,`Evol`,`Perf` or `Conf`, which are the more common parts of the facets and 0 else and m_{ncom} returns 1, if they match for either `Op`,`Jux`,`Org` or `Neg`,which are the lesser common parts of the facets as will be shown later and 0 else. m_{foc} extends the input of the other m-methods by the class, which will be focused, and returns 1, if the both citations contain the same focused class and m_{focN} returns 1 on all other matching occasions and 0, if the facet between the two functions doesn't match or matches in the focused class.
The Co-Citation weight C of the pair i and j for d be calculated as described in the following part:

The co-citation weight will be using in addition to the standard co-citation occurrence an additional value, which will represent the similarity between the citation functions.

`eqAddMore`

$$C_{eqAddMore(i,j)}(d) = \begin{cases} 1 + p_{eqAddMore(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{12}$$

$$p_{eqAddMore(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|} \sum_{o=1}^{|Cit_{j\_d}|} \sum_{k=1}^{4} m_{simple}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m})}{|Cit_{i\_d}| \cdot |Cit_{i\_d}|}$$

`eqAddLess`

$$C_{eqAddLess(i,j)}(d) = \begin{cases} 1 + p_{eqAddLess(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{13}$$

$$p_{eqAddLess(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|} \sum_{o=1}^{|Cit_{j\_d}|} \sum_{k=1}^{4} 0.5 \cdot m_{simple}(cf_{i\_d\_n}^{k}, cf_{j\_d\_m}^{k})}{|Cit_{i\_d}| \cdot |Cit_{i\_d}|}$$

`eqMultMore`

$$C_{eqMultMore(i,j)}(d) = \begin{cases} 2^{p_{eqMult(i,j)}(d)} & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{14}$$

$$p_{eqMult(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|} \sum_{o=1}^{|Cit_{j\_d}|} \sum_{k=1}^{4} m_{simple}(cf_{i\_d\_n}^{k}, cf_{j\_d\_m}^{k})}{|Cit_{i\_d}| \cdot |Cit_{i\_d}|}$$

`eqMultLess`

$$C_{eqMultLess(i,j)}(d) = \begin{cases} 1.1^{p_{eqMult(i,j)}(d)} & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{15}$$

The co-citation weight will only be influenced by the similarity between the citation functions of the author's citations.

`eqNone`

$$C_{eqNone(i,j)}(d) = \begin{cases} p_{eqNone(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{16}$$

$$p_{eqNone(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|} \sum_{o=1}^{|Cit_{j\_d}|} \sum_{k=1}^{4} m_{simple}(cf_{i\_d\_n}^{k}, cf_{j\_d\_m}^{k})}{|Cit_{i\_d}| \cdot |Cit_{i\_d}|}$$

Jochim (2014) showed the distribution between the different classes on the ACL Proceedings from the ARC of the year 2004, which is displayed in table 1. It is noticeable, that all four facets have a distribution of 1:9 between the two classes of each facet. If a rarer way of citing occurs for both cited authors, it might be a stronger implication that two authors might have some correlation between one another. Therefore the second base approach will be

Conceptual (`Conc`)	1792	Evolutionary (`Evol`)	1804
Operational (`Op`)	216	Juxtapositional (`Jux`)	204
Organic (`Org`)	203	Confirmative (`Conf`)	1836
Perfunctory (`Perf`)	1805	Negational (`Neg`)	172

Table 1: Distribution between the four different facets as shown in Jochim (2014) from a total amount of 2008 citations from the Proceedings of ACL ARC of the year 2004.

to compare each facet and weight each match between rarer facets stronger than a match between more common facets. Here will also be differentiated between having just the co-occurrence as baseline and not having it. focop-more

`difAddMore`

$$C_{difAddMore(i,j)}(d) = \begin{cases} 1 + p_{difAddMore(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{17}$$

$$p_{difAddMore(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|}\sum_{o=1}^{|Cit_{j\_d}|}\sum_{k=1}^{4}(1\cdot m_{com}(cf^k_{i\_d\_n},cf^k_{j\_d\_m})+3\cdot m_{ncom}(cf^k_{i\_d\_n},cf^k_{j\_d\_m}))}{|Cit_{i\_d}|\cdot|Cit_{i\_d}|}$$

`difAddLess`

$$C_{difAddLess(i,j)}(d) = \begin{cases} 1 + p_{difAddLess(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{18}$$

$$p_{difAddLess(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|}\sum_{o=1}^{|Cit_{i\_d}|}\sum_{k=1}^{4}(0.1\cdot m_{com}(cf^k_{j\_d\_n},cf^k_{j\_d\_m})+0.5\cdot m_{ncom}(cf^k_{i\_d\_n},cf^k_{j\_d\_m}))}{|Cit_{i\_d}|\cdot|Cit_{i\_d}|}$$

`difMultMore`

(19)

$$C_{difMultMore(i,j)}(d) = \begin{cases} 2^{pC_{difMultMore(i,j)}(d)} \cdot 4^{pUC_{difMultMore(i,j)}(d)} & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases}$$

$$pC_{difMultMore(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|}\sum_{o=1}^{|Cit_{i\_d}|}\sum_{k=1}^{4} m_{com}(cf^k_{j\_d\_n},cf^k_{j\_d\_m})}{|Cit_{i\_d}|\cdot|Cit_{i\_d}|}$$

$$pUC_{difMultMore(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|}\sum_{o=1}^{|Cit_{i\_d}|}\sum_{k=1}^{4} m_{ncom}(cf^k_{j\_d\_n},cf^k_{j\_d\_m})}{|Cit_{i\_d}|\cdot|Cit_{i\_d}|}$$

`difMultLess`

(20)
$$C_{difMultLess(i,j)}(d) = \begin{cases} 1.1^{pC_{difMultLess(i,j)}(d)} \cdot 1.5^{pUC_{difMultless(i,j)}(d)} & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases}$$

$$pC_{difMultLess(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|} \sum_{o=1}^{|Cit_{i\_d}|} \sum_{k=1}^{4} m_{com}(cf^k_{j\_d\_n}, cf^k_{j\_d\_m})}{|Cit_{i\_d}| \cdot |Cit_{i\_d}|}$$

$$pUC_{difMultLess(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|} \sum_{o=1}^{|Cit_{i\_d}|} \sum_{k=1}^{4} m_{ncom}(cf^k_{j\_d\_n}, cf^k_{j\_d\_m})}{|Cit_{i\_d}| \cdot |Cit_{i\_d}|}$$

`difNone`

(21)
$$C_{difNone(i,j)}(d) = \begin{cases} p_{difAddMore(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases}$$

$$p_{difNone(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|} \sum_{o=1}^{|Cit_{j\_d}|} \sum_{k=1}^{4} (1 \cdot m_{com}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}) + 3 \cdot m_{ncom}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}))}{|Cit_{i\_d}| \cdot |Cit_{i\_d}|}$$

In a third approach the specific facets get focused to observe the individual influence of the co-occurrence of each rare facet. Therefore a additive weighting scheme will be used. It works similar to the shown `dif`-functions, the difference being, that the highlight will be on only one of the uncommon classes instead of all uncommen classes. As with all other methods they viewed viewed on two different ways, once with high and once with low values.

`focOpMore`

(22)
$$C_{focOpLess(i,j)}(d) = \begin{cases} 1 + p_{focOpLess(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases}$$

$$p_{focOpLess(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|} \sum_{o=1}^{|Cit_{j\_d}|} \sum_{k=1}^{4} (2 \cdot m_{foc}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}, \texttt{Op}) + 0.5 \cdot m_{focN}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}, \texttt{Op}))}{|Cit_{i\_d}| \cdot |Cit_{i\_d}|}$$

`focOpLess`

(23)
$$C_{focOpMore(i,j)}(d) = \begin{cases} 1 + p_{focOpMore(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases}$$

$$p_{focOpMore(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|} \sum_{o=1}^{|Cit_{i\_d}|} \sum_{k=1}^{4} (0.5 \cdot m_{foc}(cf^k_{j\_d\_n}, cf^k_{j\_d\_m}, \texttt{Op}) + 0.1 \cdot m_{focN}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}, \texttt{Op}))}{|Cit_{i\_d}| \cdot |Cit_{i\_d}|}$$

`focJuxMore`

$$C_{focJuxMore(i,j)}(d) = \begin{cases} 1 + p_{focJuxMore(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{24}$$

$$p_{focJuxMore(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|}\sum_{o=1}^{|Cit_{i\_d}|}\sum_{k=1}^{4}(2\cdot m_{foc}(cf^k_{j\_d\_n}, cf^k_{j\_d\_m}, \texttt{Jux}) + 0.5\cdot m_{focN}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}, \texttt{Jux}))}{|Cit_{i\_d}|\cdot|Cit_{i\_d}|}$$

`focJuxLess`

$$C_{focJuxLess(i,j)}(d) = \begin{cases} 1 + p_{focJuxLess(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{25}$$

$$p_{focJuxLess(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|}\sum_{o=1}^{|Cit_{i\_d}|}\sum_{k=1}^{4}(0.5\cdot m_{foc}(cf^k_{j\_d\_n}, cf^k_{j\_d\_m}, \texttt{Jux}) + 0.1\cdot m_{focN}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}, \texttt{Jux}))}{|Cit_{i\_d}|\cdot|Cit_{i\_d}|}$$

`focOrgMore`

$$C_{focOrgMore(i,j)}(d) = \begin{cases} 1 + p_{focOrgMore(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{26}$$

$$p_{focOrgMore(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|}\sum_{o=1}^{|Cit_{i\_d}|}\sum_{k=1}^{4}(2\cdot m_{foc}(cf^k_{j\_d\_n}, cf^k_{j\_d\_m}, \texttt{Org}) + 0.5\cdot m_{focN}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}, \texttt{Org}))}{|Cit_{i\_d}|\cdot|Cit_{i\_d}|}$$

`focOrgLess`

$$C_{focOrgLess(i,j)}(d) = \begin{cases} 1 + p_{focOrgLess(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{27}$$

$$p_{focOrgLess(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|}\sum_{o=1}^{|Cit_{i\_d}|}\sum_{k=1}^{4}(0.5\cdot m_{foc}(cf^k_{j\_d\_n}, cf^k_{j\_d\_m}, \texttt{Org}) + 0.1\cdot m_{focN}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}, \texttt{Org}))}{|Cit_{i\_d}|\cdot|Cit_{i\_d}|}$$

`focNegMore`

$$C_{focNegMore(i,j)}(d) = \begin{cases} 1 + p_{focNegMore(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{28}$$

$$p_{focNegMore(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|}\sum_{o=1}^{|Cit_{i\_d}|}\sum_{k=1}^{4}(2\cdot m_{foc}(cf^k_{j\_d\_n}, cf^k_{j\_d\_m}, \texttt{Neg}) + 0.5\cdot m_{focN}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}, \texttt{Neg}))}{|Cit_{i\_d}|\cdot|Cit_{i\_d}|}$$

`focNegLess`

$$C_{focNegLess(i,j)}(d) = \begin{cases} 1 + p_{focnegLess(i,j)}(d) & \text{if } i,j \in r(d) \\ 0 & \text{else} \end{cases} \tag{29}$$

$$p_{focNegLess(i,j)} = \frac{\sum_{n=1}^{|Cit_{i\_d}|}\sum_{o=1}^{|Cit_{i\_d}|}\sum_{k=1}^{4}(0.5\cdot m_{foc}(cf^k_{j\_d\_n}, cf^k_{j\_d\_m}, \texttt{Neg}) + 0.1\cdot m_{focN}(cf^k_{i\_d\_n}, cf^k_{j\_d\_m}, \texttt{Neg}))}{|Cit_{i\_d}|\cdot|Cit_{i\_d}|}$$

Those weighting schemes will be used to create weighted co-citation count matrices $raw_{weighted\_s}$, which use the scheme $s \in F$, whereas F represents the set of all just presented weighting schemes.

3.3 Co-citation clustering process

By using the just described weighting schemes, the *raw*-citation matrices will be created. The entries $raw_{\text{function}(i,j)}$ with $i, j \in A$ and $i \neq j$ are created by using the weighting scheme corresponding the function in question by summing C_{function}, whereas the function refers to whatever weighting scheme in question, over all documents in the dataset:

$$\text{raw}_{\text{function}(i,j)} = \sum_{d}^{D} C_{\text{function}(i,j)}(d) \tag{30}$$

and $raw_{i,i}$ will be irrelevant comparable to Boyack et al. (2005), where the diagonal was considered missing since those entries will have no real value for the ongoing process.

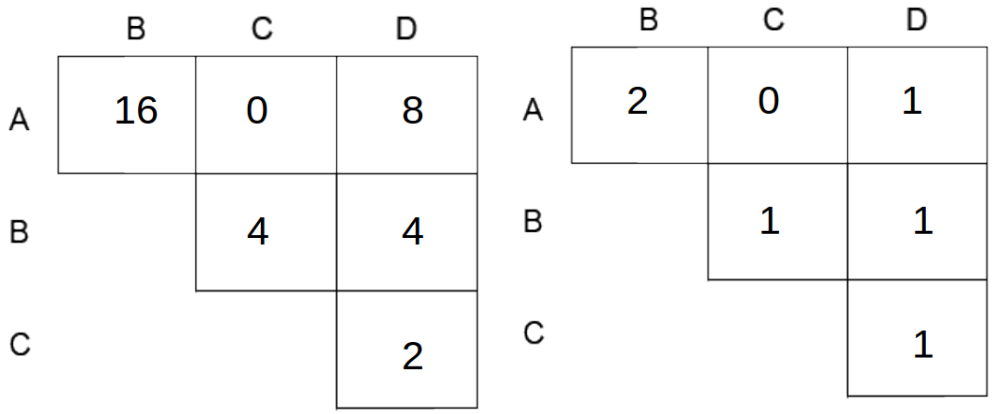

Figure 5: Example of a co-citation matrix *raw* using the `eqMultMore`-weighting scheme (left) in comparison to the standard co-citation procedure (right) with the same basis as in Fig. 3.

From the *raw*-citation matrix the *cor*-co-citation matrix will be derived. To calculate the correlation between two authors, the correlation coefficient known as Pearson's r, established in Pearson (1909), will be used. The formala equates to the one provided in Boyack et al. (2005).

$$(31) \qquad r_{i,j} = \frac{\sum_{n=1}^{|A|}(C_{i,n} - avgROW(i)) \cdot (C_{j,n} - avgROW(j))}{\sqrt{(\sum_{n=1}^{|A|} C_{i,n} - avgROW(i)^2} \cdot \cdot \sqrt{(\sum_{n=1}^{|A|} C_{j,n} - avgROW(j)^2}}$$

$$avgROW(i) = \frac{\sum_{n=1}^{|A|} raw_{i,n}}{|A|},\ n \neq i$$

To calculate the specific correlation between two authors, all citation counts between one of the observed authors and every other author influence the calculation. This ensures that even authors with an overall low amount of co-occurrences get assigned with an appropriate value.
Over the years there has been critique on this correlation measure concerning the usefullness in the field of author co-citation analysis, i.e. in Ahlgren et al. (2003). It argues that the Pearson correlation coefficient won't accurately represent author groups in case new authors, which do not co-occur with many already existing authors, are added and therefore add a number of entries to the matrix with the value 0 since they were cited with many of the already existing ones, which might skew the result. White (2003) breaks down the problem shown in Ahlgren et al. (2003). There White has shown, that even while the fundamentals, declarated inAhlgren et al. (2003), were failed to achieve by Pearson's r, the produced mapping is still acceptable, as the focus will shift from differences within author groups to the differences that reign between author groups. He also claims, that it is more important to note that the reason as to why authors are co-cited together are more important than the strict obedience to the mentioned fundamentals.
Later Boyack et al. (2005) showed their own similarity measure, which they called K50, and that it showed superiority in the fields of scalability and visual cluster representation. However the mutual information of the clusterings based on the two different similarity measures stood relatively even.
As the mutual information between the two clusterings is nearly the same, the lack of visualization in this work and the usage of Pearson's r allows for comparison to broader variety of previous work, Pearson's r was chosen as similarity measure.
The entries $dis_{i,j}$ of the *dis*-co-citation matrix will be created as follows:

$$(32) \qquad dis_{i,j} = \frac{1 - r_{i,j}}{2}$$

After the *dis*-co-citation matrix is created the clustering process begins. A

complete linkage clustering approach on the *dis*-co-citation matrix will be used, as it prevents chaining. The actual mapping of the author co-citation analysis will not be elaborated and the focus lies more on the resulting values rather than a visualization.

4 Experiment

In this chapter, the approach will be implemented.
Hereby we evaluate the practicability of all shown weighting measures in section 3.2 with the intentions of finding accurate author groups or alternate connections between authors. As there was no gold standard for author groups, which could have been used for evaluation, the clustering results will be analyzed in the contrast of nationality and textual coherence concerning the authors' oeuvres of each author cluster. Since a lot of previous studies working on citation classification, this work will use documents taken from NLP-literature as corpora as well. This should allow haivng a similar basis to also examine the similarities as well as dissimilarities of comparable weighting schemes as the ones described in the previous chapter on other citation classification procedures in future work.

4.1 Corpus

As mentioned above the used corpus will consist of documents from the field of natural language processing. The corpus shown in Bird et al. (2008) is a collection of conference and journal papers concerning natural langue and computational linguistics[1]. It consists of 10,921 articles in both XML- and PDF-format. A problem that was stated for this corpus was the text extraction of the source PDF-files, which results in an inaccurate representation of some sentences or references.
This corpus will be referred to as the main corpus.

Total Articles	10,921
Total References	152,546
References to articles inside ACL ARC	38,767 (25.4%)
References to articles outside ACL ARC	113,779 (74.6%)

Table 2: Statistics of the main corpus provided by Bird et al..

[1] Available http://acl-arc.comp.nus.edu.sg to the time of the release of this thesis.

Conc-Op	Evol-Jux	Org-Perf	Conf-Neg
0.6416	0.4963	0.5401	0.4891

Table 3: Macro f_1-scores for a 10-fold cross validation for the classification using the Stanford MaxEnt classifier with a bag-of-words feature on the „IMS Citation Corpus“.

Every citation of each document will be associated with the corresponding author and a citation function, which will be assigned as mentioned in section 3.1. In order to classify a training set is necessary, which provides a certain amount of annotated citations in a similar field of science. The thought behind the necessity of having a training set in a similar environment as the data, which has to be classified, is that certain words might have varying interpretations applied in another setting as well as such fields might provide whole different sets of used words or ways to cite other authors.
For this reason the „IMS Citation Corpus“of Jochim and Schütze (2012) will be used as training corpus and referred as such. It consists of 2008 annotated citations over 84 documents with the distribution as shown in table 1. It uses the same classification scheme as Murugesan and Moravcsik and the used documents are part of the main corpus, hence the citation sentences have the same background of NLP as the remaining documents of the main corpus.

4.2 Classification

The classification of all citations follows the procedure described in section 3.1. For the implementation of the maximum entropy classifier the „Stanford Classifier“[2] was chosen.
To show the accuracy of the classification a 10-fold cross validation on the training set was done and the macro f_1-score was calculated. The results for this can be seen in table 3.

[2] http://nlp.stanford.edu/software/classifier.shtml

4.3 Clustering

The clustering approach described in 3.2 and 3.3 was applied on two trials, T_1 and T_2, which only differ in size. This process includes the n most cited authors, which will then be partitioned into clusters of different sizes. T_1 has $n = 500$, while T_2 has $n = 1000$. However for the sake of simplicity multiple authors with the same last name will be seen as one author.
The reasoning for a clustering on such a small number of authors was influenced by the size of the main corpus as well as the scalability of the used similarity measure Pearson's r. However using an overall improved similarity measure should only lead to improved results.
This experiment set up allows to see in which way author groups generated in each clustering process are coherent textual and nationality wise and how the inclusion more lesser cited authors affects the result.

4.3.1 Results

To analyze those methods those methods will be compared to a baseline consisting of a standard co-citation clustering, which is just based on the co-occurrence of authors, and a random clustering. This leads to 17 clusterings for each trial. We evaluate the cluster solutions by using the rand index for author nationalities, which are accessed by location of the facilities an author was attached to, and the textual coherence method provided by Boyack and Klavans (2010). However since the nation, in which an author is researching, might change over time, it changes the way the rand index will be applied. As a result, the used method differs from a standard way of applying the rand index. A pair of authors will be seen as true positive, if they share a nation in their residence list and as true negative, if this isn't the case. This could be changed by considering and comparing the residencies for a certain time period of the authors of an author pair.
The textual coherence will be evaluated over titles of each work in an author's oeuvre. The title of a publication reveals the topic of it to a certain degree. Hence the coherence will provide a certain insight to what extent a cluster of a cluster solution has a topical agreement between its authors.
As shown in table 2 the majority of all references point at documents, which are not part of the main corpus. This caused the decision use the complete

oeuvre of each author, which was crawled [3] [4] and not restricting it to the main corpus.
This method could also be improved by using a whole abstract of an author's work. The main reason to use only the titles of an author's publications was directly influenced by the accessibility. While the titles of all publications of an author could easily be obtained, it has been more complicated to gain access to all abstracts of an author.

First, the textual coherence of the clusters will be observed, as this fairly protrays the relatedness of clusters and therefore allows seeing how this method can be used as a general method to map fields of science and authors. Then the second assumption will be reviewed with the help of the rand index described above. An increase of the rand index of the standard co-citation method, will show that there might be some sort of bias, which results to citing authors of the same nation similarly.
Each review covers four parts, which cover the values for some weighting schemes for both trials.
The first part covers the methods based on weighting rarer classes of facets stronger.

Figure 7 displays that for an increasing cluster amount and therefore decreasing cluster size the methods behave relatively the same as the standard method and show superiority over a random clustering method. Greater differences are rather be seen with small cluster amounts. The standard method shows to have quite a weak textual coherence for cluster amounts towards 1, while `dif`-methods seem to have a higher textual for bigger clusters. However the real difference between the methods becomes noticeable for cluster amounts from around 30 to 120 with an average cluster size from 8 to 33. All methods seem to be weaker than a random clustering, with the standard method being the most coherent of all shown co-citation clustering of this part, while `difAddMore` and `difMultMore` appear to be the weakest methods in this range.
Trial 1 in figure 6 seemingly shows nearly equal results with smaller average cluster sizes. Both trials imply that the multiplicative methods are slightly better than their additive counter parts in the range between 30 and 120 as seen in table 4 and table 5. This however could be reduced to variance.

[3] Author profiles on scholar.google.com
[4] Author profiles on ACL Anthology Network http://clair.eecs.umich.edu/aan/index.php

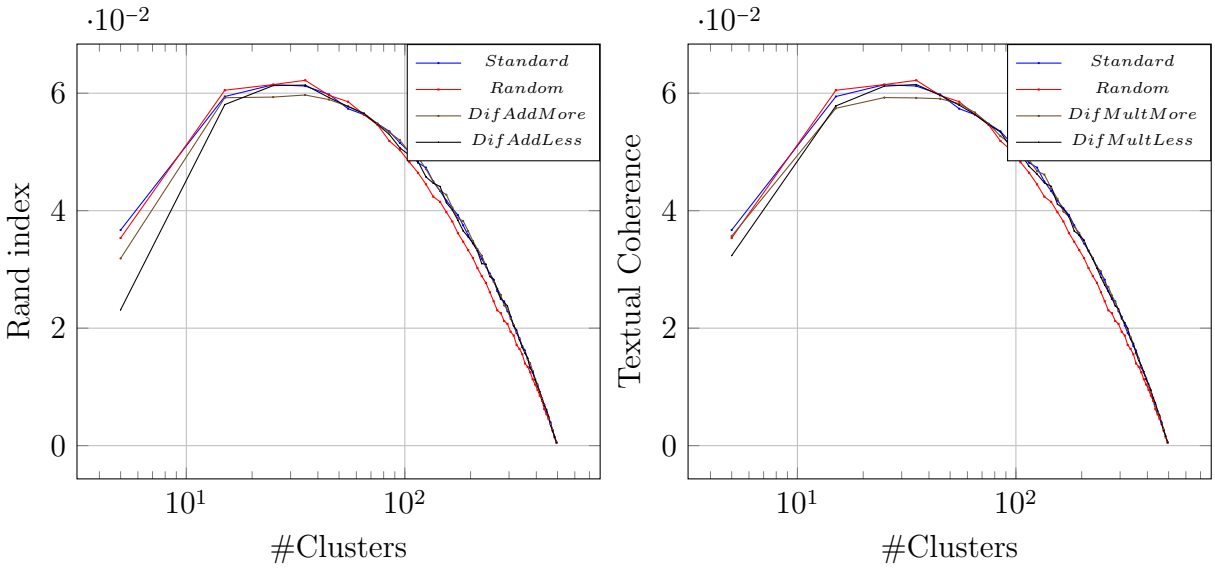

Figure 6: Textual Coherence over the number of clusters for differing weighting for classes of facets for T_1.

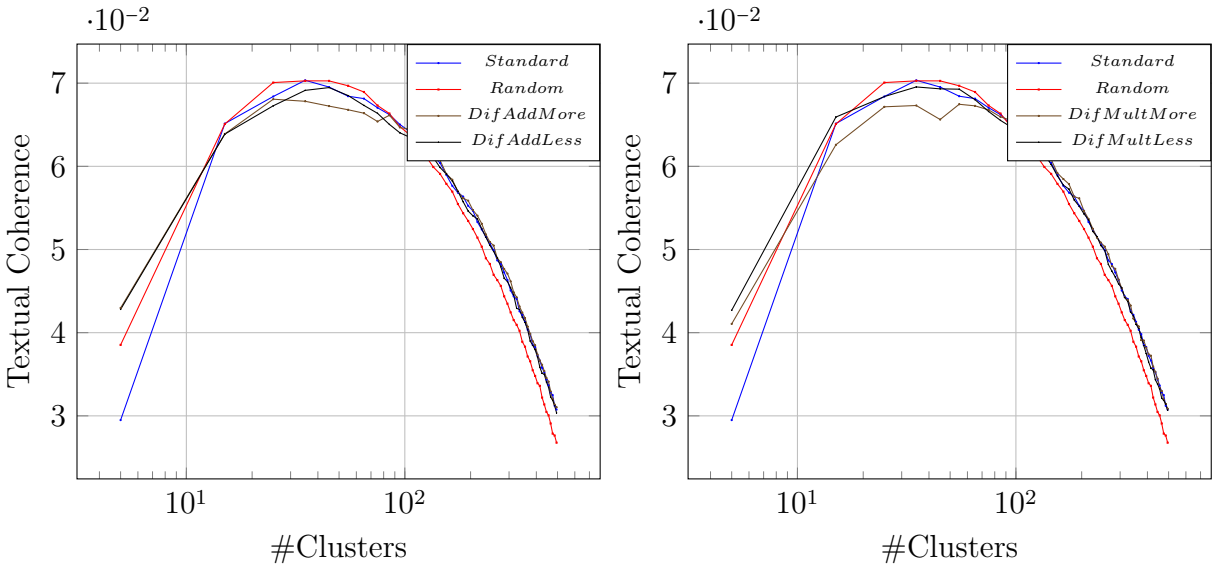

Figure 7: Textual Coherence over the number of clusters for differing weighting for classes of facets T_2.

`difMultLess`	0.06313	`difAddLess`	0.06299
`difMultMore`	0.06262	`difAddMore`	0.06279

Table 4: Textual coherence of `dif`-type methods in the range from 30 to 100 for trial T_1.

`difMultLess`	0.07563	`difAddLess`	0.07508
`difMultMore`	0.07467	`difAddMore`	0.07462

Table 5: Textual coherence of `dif`-type methods in the range from 30 to 100 for trial T_2.

Next the methods based on equal weighting `(eq)` of all classes of facets. Table 8 represents hereby the results for trial T_1 and table 9 the results of the trial T_2.

The results are similar to the previous ones, as towards a lower average cluster size the coherences is nearly equal to the standard co-citation clustering approach and the methods with lower factors perform better than the ones with higher factors in the range from 30 - 120. The outcome is somewhat the same, even though there are slight inferiorities in the „$AddMore$“-method.

`eqMultLess`	0.06311	`eqAddLess`	0.06309
`eqMultMore`	0.06266	`eqAddMore`	0.06300

Table 6: Textual coherence of `eq`-type methods in the range from 30 to 100 for trial T_1.

`eqMultLess`	0.07572	`eqAddLess`	0.07465
`eqMultMore`	0.07458	`eqAddMore`	0.07368

Table 7: Textual coherence of `eq`-type methods in the range from 30 to 100 for trial T_2.

The methods, based on only counting the similarity of citation context and ignoring general co-occurrence are displayed in figure 10.

These yield similar results for T_1 and T_2 to previous methods. Both methods seem to perform equally in the range from 30 - 120 in T_2, but also performing worse than the standard method and better than the random method for

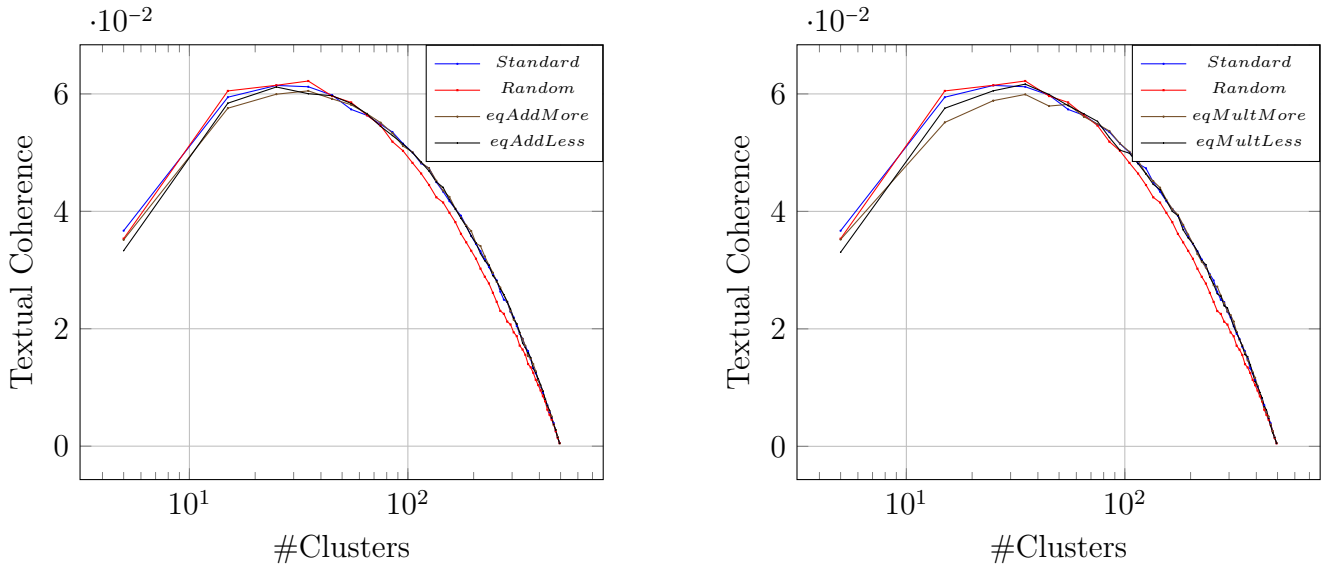

Figure 8: Textual Coherence over the number of clusters for equal weighting for classes of facets for T_1.

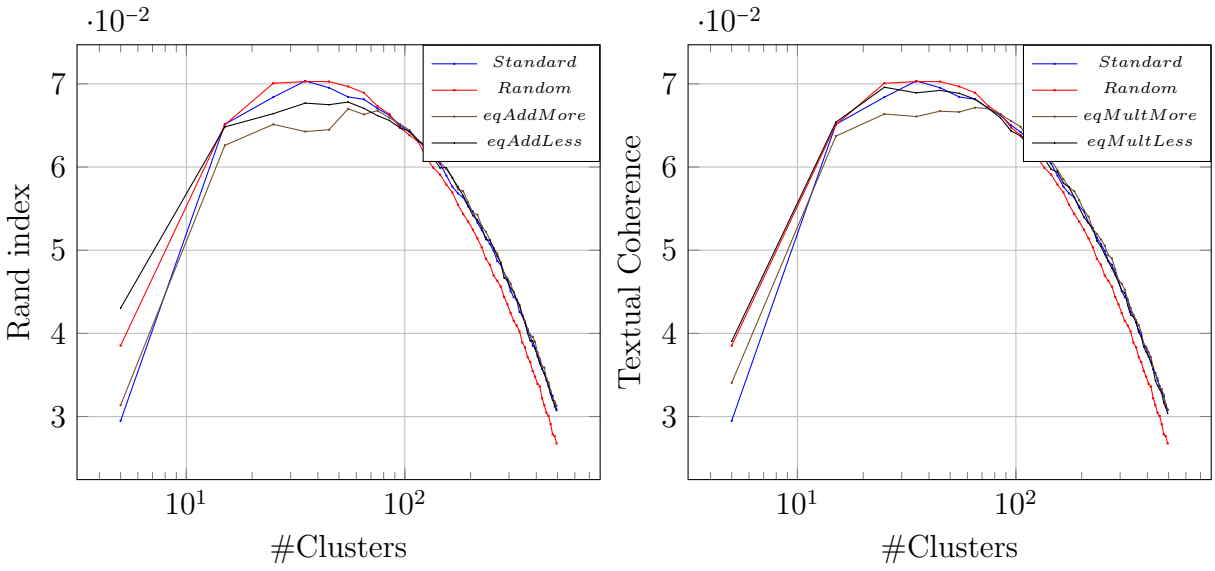

Figure 9: Textual Coherence over the number of clusters for equal weighting for classes of facets for T_2.

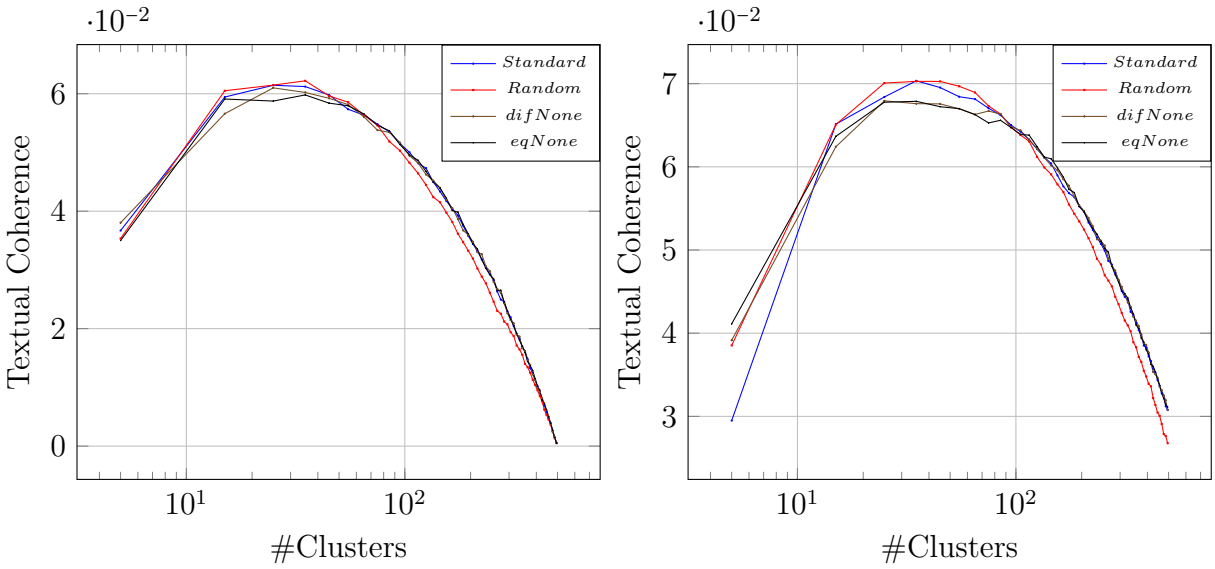

Figure 10: Textual Coherence over the number of clusters for weighting, which ignores simple co-occurrence, for classes of facets for T_1 (left) and T_2 (right).

smaller average cluster sizes.

Lastly methods with special focus on a certain rarer class of a facet (`foc`), which are represented in figure 11, will be looked at.
While all the methods of this category share the normal behavior of all other methods towards smaller average cluster sizes, there are differences to be seen in the usual range of cluster sizes between 30 to 120. Especially „$focJuxLess$“appears to be equal to the standard co-citation clustering method in T_1, however, provides higher values in the highest varying range of 30-120 clusters in T_2 .

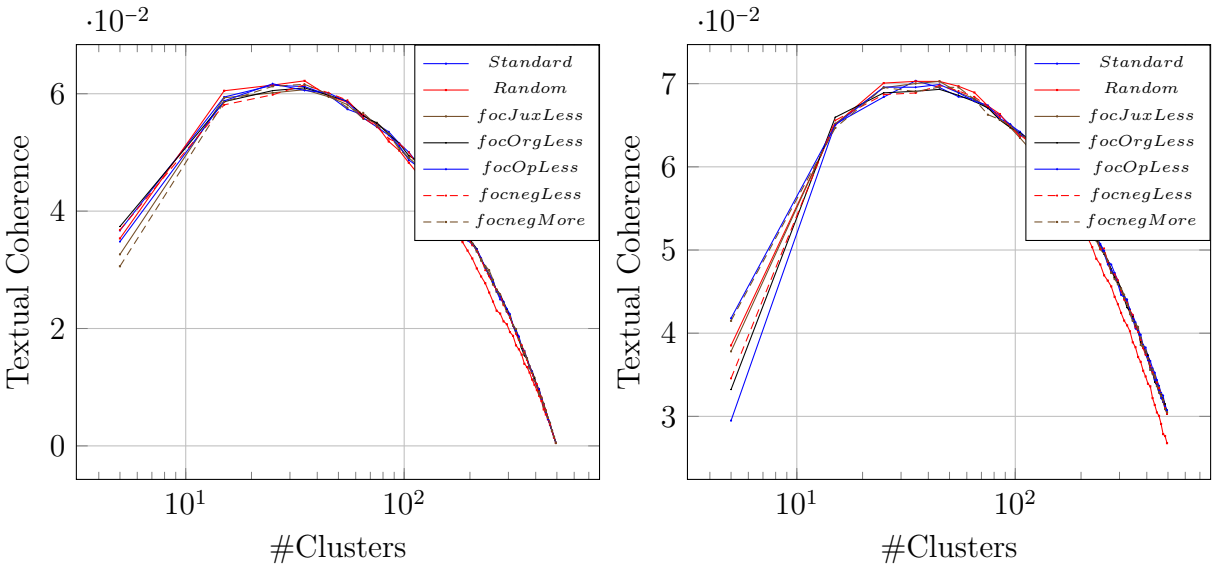

Figure 11: Textual Coherence over the number of clusters for weighting, which focuses on certain aspects of the facets, for classes of facets for T_1(left) and T_2(right).

The rand index for `dif`-type methods is portrayed in figure 12 and figure 13. For T_1 all methods seem to perform nearly the same and there don't appear to huge differences. Therefore the focus will be more on T_2 as variances become more apparent for more authors.

The `dif`-type measures as well as the standard co-citation clustering converge towards the same value, while having relatively similar values from a cluster amount of 100 in T_1 and from around 200 in T_2. It is to note that there is a huge difference in rand index between the two trials and a random cluster is way inferior compared to other methods.
This time the stronger weighted methods provide better results than the standard co-citation clustering, whereas the lighter weighted methods perform worse in regards to the rand index. Especially „$difMultMore$"shows vast improvements to its counterpart „$difMultLess$"and shows a higher rand index than „$difAddMore$"in both trials.

This differs from `eq`-type measures in some way. While „$eqMultMore$"is the strongest measure as well and shows a huge gap to „$eqMultLess$", there

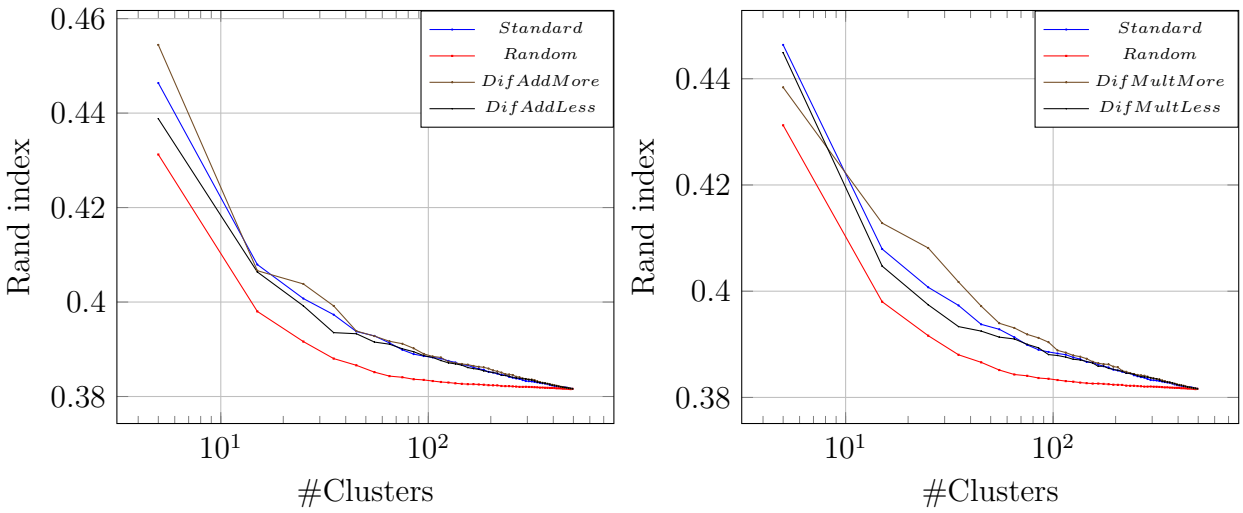

Figure 12: Rand index over the number of clusters for differing weighting for classes of facets for T_1.

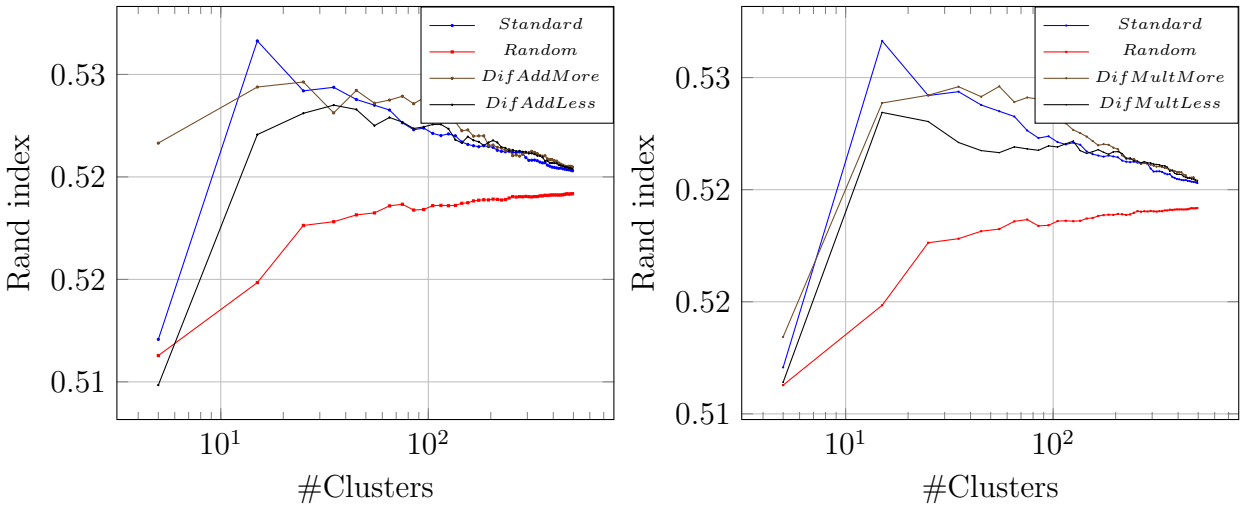

Figure 13: Rand index over the number of clusters for differing weighting for classes of facets for T_2.

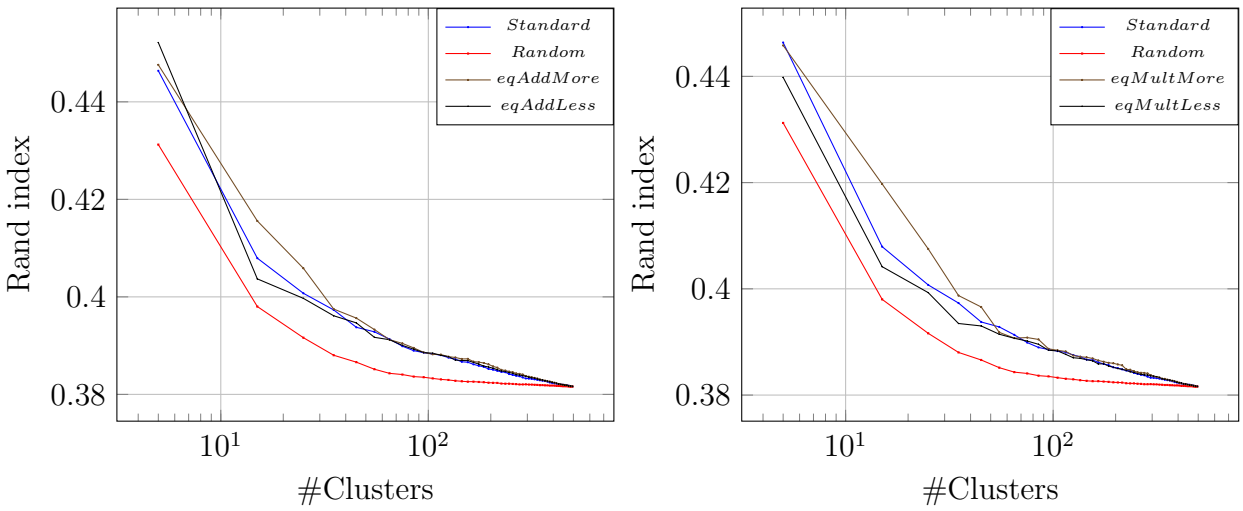

Figure 14: Rand index over the number of clusters for equal weighting for classes of facets for T_1.

is a huge variation in the additive measures compared to `dif`-based methods. „$eqAddLess$“is in T_2 strongly superior to „$eqAddMore$“in the range of 30 to around 120, while being inferior to it in T_1. It also has a spike in value at 20 to 30 clusters in T_2, which falls of towards greater average cluster sizes. This is similar to the standard co-citation clustering, which also spikes at an amount of nearly 20 clusters. The other methods, however, seem to be more consistent, while having its highest values at around 40 to 100 clusters with an average cluster size of 10 to 25 authors.

`None`-methods perform better than the standard co-citation clustering in regards to the rand index from a cluster amount of 70, while performing less previously. „$difNone$“performs overall better than „$eqNone$“in T_2 , which, however, proved slightly better results in T_1 between cluster amounts of 1 to 80. Just as the standard method „$difNone$„seems to spike for bigger average cluster sizes around 25 - 33 in T_2 and hence performs similar to „$eqAddLess$“, whereas slightly worse as seen in table 8.

In both T_1 and T_2 the methods focusing on a special class (`foc`) seem to perform generally worse than the standard method in figure 17. The focus on `neg` delivers the best methods in this regard, with the stronger weights

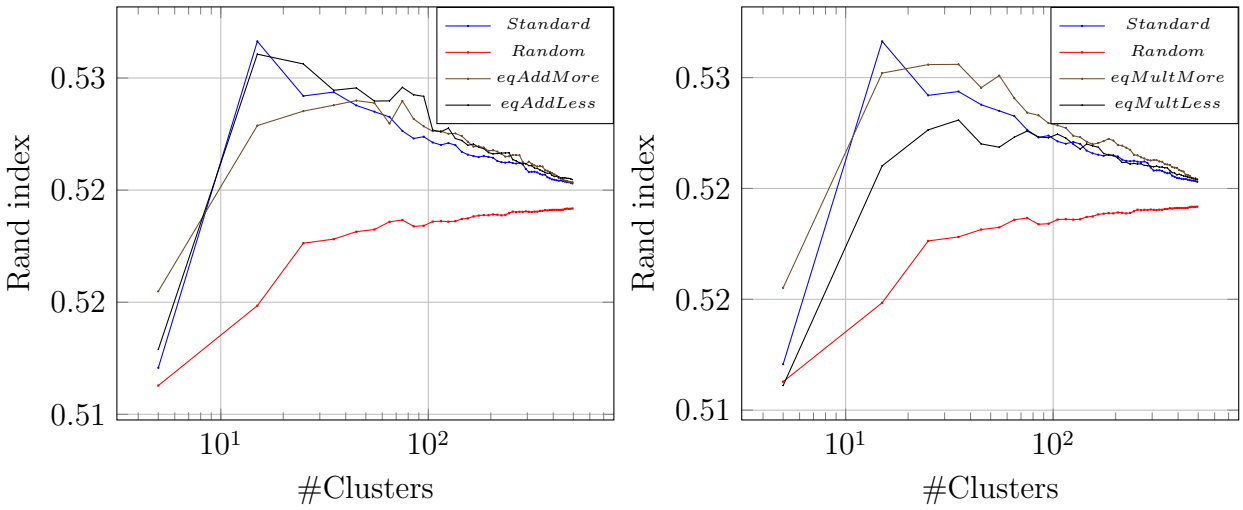

Figure 15: Rand index over the number of clusters for equal weighting for classes of facets for T_2.

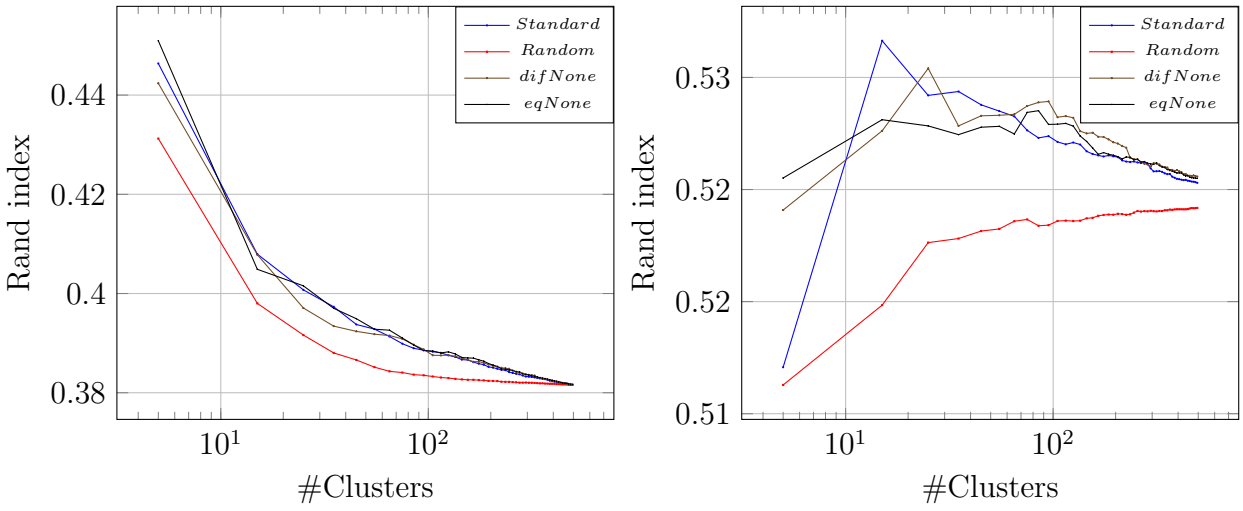

Figure 16: Rand index over the number of clusters for weighting, which ignores simple co-occurrence, for classes of facets for T_1(left) and T_2(right).

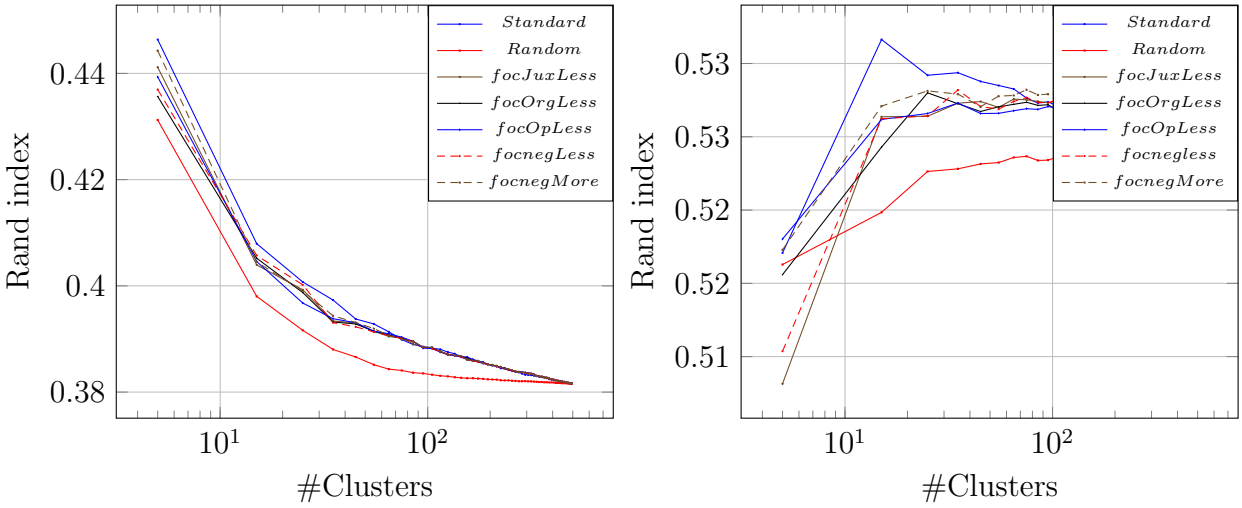

Figure 17: Rand index over the number of clusters for weighting, which focuses on certain aspects of the facets, for classes of facets for T_1(left) and T_2(right).

being superior to the smaller weights. For average cluster sizes towards 1 the methods behave the same as it was to expected barring previous results.

`eqNone`	0.52733	`eqAddLess`	0.52771	`eqMultMore`	0.52828
`difNone`	0.52764	`difAddMore`	0.52860	`difMultMore`	0.52705
`standard`	0.52721	`random`	0.52094	`focnegMore`	0.52549

Table 8: Average rand index comparison of several best methods for each type of measure in the range from 5 to 50 for trial T_2.

In table 8 are the values for the better methods of each part together with the standard and the random clustering enlisted for average cluster sizes from 20 to 200 in T_2. According to this table the best methods for a relatively high cluster size are „$difAddMore$“followed by „$eqMultMore$“.

5 Conclusion

The result of this experiment provides some evidence that the use of citation context do not increase the accuracy of co-citation clustering as the clusters created by those methods do not increase the textual coherence. While there are some methods that come close to the coherence of the standard co-citation clustering method there are none, which show any improvements over it. Therefore the closer the method was to the standard method, the closer the coherence to it was. This is seen as the methods which focus only on a specific class provide nearly identical results, whereas methods that have heavy weighting by equal context like „$difAddMore$“or „$eqAddMore$“show a significant downside usually at an amount of 30 to 80 clusters.
Although these methods result in a lesser textual coherence in clusters, the results they bring in terms of rand index show quite a huge improvement. While the standard method shows the best values for a cluster amount of about 20, it gets outclassed by at least one method of each category excluding `foc`-type methods for higher numbers of clusters. The methods, which are outperforming the standard method in terms of rand index, are the ones that use the highest valued parameters like „$difAddMore$“or „$eqAddMore$“. It can be seen that the higher the results in terms of rand index become, the lower the coherence in between clusters becomes.
So it appears that the stronger the influence of the citation function becomes in this procedure the higher is the likelihood of authors within the same cluster to be in the same nation, but are wider spread in terms of topology.
Hence people that are cited in the same way might appear to have a connection indicates that the citation function does in fact influence the co-citation process and that there is a certain bias for authors concerning references.
So authors that are cited together in the same way show some connection, which was in this work evaluated through nationality. Which might be correlated to author groups as those are usually in the same country.

However these might indicate a better representation of research groups since these, while also having the possibility to be international, are often in the same country. Hence context based co-citation clustering may help to map research groups in future work.

Although this thesis presents a way to include context into co-citation clustering and shows its behavior in terms of topic coherence and nationality

based rand index over the ACL ARC , there is still a lot left to examine. The ACL ARC corpus is a relatively small corpus and so might not represent the behavior of context based co-citation analysis on bigger corpora as for example a „PubMed“-corpus, which comprises several million documents. Also in addition to the evaluation techniques used in this thesis an attempt could be made to detect author cliques within clusters. An author clique of size n is hereby defined as a group of n authors, in which every author is shown to have worked with every other author of the clique at least once. If a higher value results from a context driven co-citation clustering approach than from a standard co-citation clustering, it will be a strong evidence that context based co-citation clustering is advantageous for the mapping of research groups.

Direct advancements of the technique could be made by combining context with Boyack et al.'s position based method. By weighting citations stronger that appear in the same sentence or bracket it could show a direct improvement over the here presented methods in terms of representing author groups. However as seen here the topic of the here proposed methods is relatively wide spread according to the textual coherence of the author ouvres. This might lead to poorer results in terms of actual textual coherence compared to Boyack et al. (2013).

6 Summary

In this thesis, we expanded on the theories in the field of author co-citation analysis. The fundamentals of co-citation analysis and citation classification get explained. A new concept was then presented to include citation context directly into the clustering process. Authors, which were cited in similar ways, are interpreted as having a higher correlation and therefore closer and if they are cited in not similar ways the relative distance between those authors increases compared to usual co-occurrence based distance. A lot of different weighting schemes were created to focus on various features. These were then used in an experiment, which tests their potential to find author group relations. This experiment was executed in an NLP environment and it gets shown that, even though the clusters don't seem to center as much around one topic as in a standard co-citation clustering, the author clusters have a higher nationality share.
We concluded that this might be due to author research groups being more likely to be cited together, as they share ideas and concepts regarding their topics.

7 Acknowledgements

I would like to express my gratitude to my supervisor Dr. Roman Klinger for the useful comments, remarks, and engagement through the process of this bachelor thesis. Furthermore, I would like to thank Sebastian Padó as well for examining this thesis. Lastly, i would like to thank anyone, who has supported me during the time of this thesis.

References

Per Ahlgren, Bo Jarneving, and Ronald Rousseau. Requirements for a cocitation similarity measure, with special reference to pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54(6):550–560, 2003.

Awais Athar and Simone Teufel. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 597–601. Association for Computational Linguistics, 2012.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan R Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*, 2008.

Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015.

Kevin W Boyack and Richard Klavans. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12):2389–2404, 2010.

Kevin W Boyack, Richard Klavans, and Katy Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005.

Kevin W Boyack, Henry Small, and Richard Klavans. Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, 64(9):1759–1767, 2013.

Leo Egghe and Loet Leydesdorff. The relation between pearson's correlation coefficient r and salton's cosine measure. *Journal of the American Society for information Science and Technology*, 60(5):1027–1036, 2009.

Eugene Garfield et al. Citation analysis as a tool in journal evaluation. American Association for the Advancement of Science, 1972.

Charles Jochim. *Natural language processing and information retrieval methods for intellectual property analysis*. PhD thesis, University of Stuttgart, IMS, 2014.

Charles Jochim and Hinrich Schütze. Towards a generic and flexible citation classifier based on a faceted classification scheme. 2012.

Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

Loet Leydesdorff. On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85, 2008.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Michael H MacRoberts and Barbara R MacRoberts. Problems of citation analysis: A critical review. *Journal of the American Society for information Science*, 40(5):342, 1989.

6C. Manning and D.Klein. Optimization, maxent models, and conditional estimation without magic. *Tutorial at HLT-NAACL 2003 and ACL 2003*, 2003.

Poovanalingam Murugesan and Michael J Moravcsik. Variation of the nature of citation measures with journals and scientific specialties. *Journal of the American Society for Information Science*, 29(3):141–147, 1978.

Karl Pearson. Determination of the coefficient of correlation. *Science*, pages 23–25, 1909.

Karl Erik Rosengren. *Sociological aspects of the literary system*, volume 4. Natur och kultur, 1968.

Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269, 1973.

Henry Small. Update on science mapping: Creating large document spaces. *Scientometrics*, 38(2):275–293, 1997.

Henry Small. Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, 87(2):373–388, 2011.

Albert N Tabah. Literature dynamics: Studies on growth, diffusion, and epidemics. *Annual review of information science and technology (ARIST)*, 34:249–86, 1999.

Simone Teufel. The structure of scientific articles: Applications to citation indexing and summarization (center for the study of language and information-lecture notes). 2010.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110. Association for Computational Linguistics, 2006.

Howard D. White. Author cocitation analysis and pearson's r. *Journal of the American Society for Information Science and Technology*, 54(13): 1250–1259, 2003. ISSN 1532-2890. doi: 10.1002/asi.10325. URL `http://dx.doi.org/10.1002/asi.10325`.

Howard D White and Belver C Griffith. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for information Science*, 32(3):163–171, 1981.