FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

A Web Science View on Computer-Science Bibliography Data

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Stefan Honeder, BSc
Matrikelnummer 0625411

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuer: Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl

Wien, 18.11.2011

(Unterschrift Verfasser) (Unterschrift Betreuer)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

A Web Science View on Computer-Science Bibliography Data

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Stefan Honeder, BSc
Registration Number 0625411

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl

Vienna, 18.11.2011

(Signature of Author)

(Signature of Advisor)

Erklärung zur Verfassung der Arbeit

Stefan Honeder, BSc
Nr. 165, 3664 Martinsberg

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Martinsberg, 18.11.2011 _____

(Unterschrift Verfasser)

Acknowledgements

First of all I want to thank my advisor Dieter Merkl. His lectures about data mining and information retrieval sparked my interest in this subfield of computer science. The discussions with him formed the basis for this thesis. His engagement and encouragement is unparalleled and his constructive feedback has been of great value for me.

I also want to thank my parents Margarete and Leopold Honeder who made all this possible for me by supporting me all the time during my study.

Abstract

Classification: H.3.4 [Systems and Software]: World Wide Web (WWW)
H.3.1 [Content Analysis and Indexing]: Indexing methods
Keywords: Web Science - Bibliography analysis - Cultural differences

Web Science is the name of an initiative started in 2006 by a workgroup consisting of, among others, Tim Berners-Lee, Wendy Hall and Nigel Shadbolt. The initiative originated from the observation that understanding the Web is crucial for its further success and therefore a new research discipline, Web Science, has to be established with the interdisciplinarity of the Web as main research focus. The intention for the Web in the early 1990 was the interchange of scientific research papers, but it has undergone many changes, grown to a worldwide scale, influencing the society and the way we work, and although it is that powerful, research in this field, with the Web as main focus, is rare.

Consequently a part of this science, deals with the current shift from Web 2.0 user generated content, to the Semantic Web. Research in this field helps to understand crucial success factors for this shift, and also offers the opportunity to influence the development. This next generation Web, which was already foreseen 1996 by Tim Berners-Lee, is based on semantically rich data. Such Linked Data forms the basis for further development of the Web and gives the opportunity for a bunch of new applications.

We chose to analyze the power of connected datasources in the field of computer science bibliography data. Beside the reason that there are different sources (DBLP, ACM Digital Library, Microsoft Academic Search, Google Scholar) it offers the opportunity to answer questions like, 'Is there a cultural influence on research in computer science?'.

For building the datasource to answer questions like above, we started with the downloadable data from DBLP, interlinked this information with data from Microsoft Academic Search and connected it further with ACM Computing Classification System (CCS) information. Prior the analyses and visualization the data was processed by basic text indexing, tokenizing, stop word removal and stemming. For analyzing the collected information we used visualizations based on this data, regression models as well as clustering and classification methods.

We analyzed the influence of authors and their co-authors to the quality of an institution. It turned out that scientists with a scientific career between 5 and 10 years provide the most valuable contribution to the quality of an institution. Also a bias between the country of origin of the authors in the dataset and the amount of authors was found. A cultural influence on research was observed which shows a tendency of more Mathematical research in Eastern Europe, parts

of Asia as well as Middle America. Also a relationship between the location of a conference and the scientists joining this conference is noticed for non-top conferences.

So this master's thesis demonstrates a proper application of Web Science with Linked Data, and the generated dataset can be used as basis for further applications.

Kurzfassung

Klassifizierung: H.3.4 [Systems and Software]: World Wide Web (WWW)
H.3.1 [Content Analysis and Indexing]: Indexing methods
Schlüsselwörter: Web Science - Bibliographische Analyse - Kulturelle Unterschiede

Tim Berners-Lee, Wendy Hall und Nigel Shadbolt gründeten mit Kollegen 2006 eine Initiative mit dem Namen Web Science. Diese Initiative wurde ins Leben gerufen, da eine fehlendes Verständnis für das Web als Ganzes beobachtet wurde. Obwohl sich das Web seit seinen Anfängen 1990, als es hauptsächlich für den Austausch von wissenschaftlichen Materialien genutzt wurde, grundlegend verändert hat und die gesamte Gesellschaft beeinflusst, gibt es keine Forschungsdisziplin die sich mit dem Web ansich beschäftigt. Der Fokus von Web Science liegt darauf die Interdisziplinarität des Webs zu erforschen. Der Kerngedanke dahinter ist, dass es notwendig ist das Web zu verstehen um seine Weiterentwicklung und den zukünftigen Erfolg zu gewährleisten.

Ein Teil von Web Science beschäftigt sich nun mit der Evolution des Web 2.0 zum Semantic Web. Forschung in diesem Gebiet soll diesen Übergang unterstützen. Diese nächste Generation des Webs wurde bereits 1996 von Tim Berners-Lee definiert, und basiert auf Daten mit semantischen Informationen. Mit diesen semantischen Informationen ist es möglich verschiedene Datenquellen im Web zu verbinden und so die Weiterentwicklung des Webs voranzutreiben.

Um die Möglichkeiten von verbundenen Datenquellen zu demonstrieren, verwenden wir in dieser Arbeit bibliographische Angaben aus dem Bereich der Informatik. Diese Daten sind einerseits in einer Vielzahl von Quellen zugänglich (DBLP, ACM Digital Library, Microsoft Academic Search, Google Scholar) und des Weiteren können Fragen in der Art 'Kann man einen kulturellen Einflusse auf die Forschung in der Informatik feststellen?' beantwortet werden.

Unsere Datenquelle wurde aus verschiedenen Teilen zusammengesetzt. Den Anfang machte eine Momentaufnahme der Daten aus der DBLP, diese wurden mit Informationen der Microsoft Academic Search verbunden und mit ACM Computing Classification System (CCS) Information angereichert. Die Daten wurden vor der Visualisierung und Analyse grundlegender Text Indizierungen unterzogen. Dabei wurden häufig verwendete Wörter entfernt sowie Wordstammreduzierung durchgeführt. Die Analyse wurde mittels Visualisierungen, Regressionsmodellen, Clustering und Klassifizierung durchgeführt.

Wir analysiert wie sich Autoren und deren Koautoren auf die Qualität einer Institution auswirken. Es stellte sich heraus das Wissenschaftler mit einer Karriere zwischen 5 und 10 Jahren den größten Einfluss auf die Qualität einer Organisation besitzen. Eine ungleiche Verteilung der

Menge an gefundenen Autoren und dem Herkunftsland wurde ebenfalls in unserer Datenquelle festgestellt. Der kulturelle Einfluss auf die Forschung konnte nachgewiesen werden und ein Trend zu einer eher mathematischen Forschung in Osteuropa, Teilen von Asien und Mittelamerika ist evident. In einer weiteren Analyse wurde die Beziehung zwischen dem Austragungsort einer Konferenz und dem Herkunftsland der Autoren die auf dieser Konferenz publizieren für 'non-top' Konferenzen nachgewiesen.

Contents

CHAPTER 1

Introduction

This chapter gives an introduction to this master's thesis. It consists of the motivation for this work, the problem statement, aim of the work, the methodological approach and structure of the work.

1.1 Motivation

Looking at the Internet as it was proposed by Tim Berners-Lee 1990 [4] we can see that it has undergone many changes. First designed for scientific purpose, it nowadays contains information in an amount and dimension not expected in these early years. People use the Internet as source of information, as a place where they interact with each other through social networks, where governments provide access to their data through Open Data[1] initiatives, where companies do business and where people have the opportunity for a global audience. As we can see the Internet itself has changed and it has also changed the society A a whole, how we live and work together.

Looking at the main purpose of the Web in the early 1990's we can see that the exchange of scientific articles and materials was the primary goal, and hence the amount of web pages was manageable. People have managed lists and bookmarks of Universal Resource Locator's (URL) to access their desired information. This has changed, the Web has grown in an incredible speed, up to 11.5 billion pages in 2005 [27], and as the amount of web sites has grown the need for search engines evolved. It was simply not possible anymore to find every desired information as the number of places to search was not manageable. So the need for search engines evolved which also has undergone some developments, from applying offline information retrieval techniques to specialized techniques for the Web and finally a breakthrough with the

[1] `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData` (accessed 26-October-2011)

PageRank algorithm of Google [10]. This algorithm, which ranks the search results by calculating the importance of a website through the amount of sites linking to it as well as using the description of these links as description for the site, revolutionized searching on the Web as it took into account the characteristics of the Internet for the search process.

In the beginning of the Web the information provided on it, was just consumed by the users, this has changed with the evolution of the so called Web 2.0, where users got the possibility to generate content, and not only receive content, by writing a blog post, answering in a forum and so on. In this ongoing process of change the next challenge is the evolution of the Web of documents to the Web of data, also called Semantic Web. This next form of the Web is supposed to not only contain plain information but also semantic for this information and hence provide the opportunity to machine process this data, creating new valuable information out of this [7]. This is interesting because search engines still basically try to guess the users intention behind the small amount of search terms provided and return the best results found for this terms. By providing not just documents on the Web, but also data with semantic, it is possible to connect information from various sources, and then try to answer questions properly. An example how this approach can look like is demonstrated by Auer and Lehmann in [2], who answer questions like; 'Which films star an Oscar winner (as best actor) with a budget of more than 10 million US dollars?', and receive the proper answers by using the semantic information included in the articles of Wikipedia[2]. So the evolutionary step to enrich the huge amount of available data on the internet with semantic, targets the possibility of processing data automatically, interlink it and hence provide more accurate and cross referenced information for users [29].

A science which deals with this changes on the Web and tries to investigate new developments, is Web-Science, a new science discipline proposed by among others Tim Berners-Lee [6]. The key idea of this research area is, that understanding the Web is crucial for it's further success, and hence new developments like the shift to the Web of data has to be understood to find solutions to maybe upcoming problems, like privacy issues in the Web of data.

Hence this work demonstrates a possibility how this shift to the Web of data, by adding semantic information and interlinking datasources, provides new insights. This approach is demonstrated by analyzing bibliography data in the computer science area. This is an interesting field to be investigated, because it provides insights in different areas. For example questions like 'Is there a cultural influence on research in computer science?' or 'Is there a bias between the venue of a conference and the origin of authors joining the conference?' can be answered by examining bibliography data. This is possible because basic information about publications, like title, name of the authors, publication year, source of publication, is free accessible and a lot of datasources on the Web provide access to bibliography citation information. The reason for this public accessibility is, that these informations have to be freely accessible to reach a proper audience and hence get acknowledged accordingly. Papers not available online, or better said not findable online, have a lower chance of being cited and therefore the tendency to provide easy access to science publications is an important factor to gain proper recognition [42]. This trend therefore leads to a quite large amount of various datasources providing bibliography data available for the proposed investigation. So beside the fairly easy access to this data the reasons

[2] `http://www.wikipedia.org` (accessed 14-September-2011)

to investigate citation information are that it is possible to analyze the data concerning trends in the different computer science subfields, to question and investigate performance indicators of authors and institutes, to examine differences between regions and the list could be continued but the intention behind these scenarios is to give an overview of a research area and provide insights so that maybe political decisions regarding education or funding decision are faced from a different point of view.

We motivated our work in this section and argued why it is interesting to investigate the evolution of the Web. We described the current shift of the Internet to a Web of data and we want to provide an example which demonstrates the power of semantically enriched data. Therefore the analysis of bibliography citation information is selected where the data is not only easy to access, but also provides interesting properties worth a study, like the cultural influence on research.

1.2 Problem Statement

As stated in Section 1.1, data from one source nowadays often can't be connected to other sources because it lacks the semantic information which is necessary for machines processing the information to make correct relations. Consequently current analysis of bibliography data in the computer science field often concentrate on a single source which also produces results, but locks you into a predefined environment setting where additional information or new insights can't be unveiled that easily.

So we connect more than one source with each other. The typical bibliography citation datasources can be distinguished between two types, on the one hand automatically indexed libraries, which crawl the Web for publications and index them, and on the other hand manually maintained registers. The automatically indexed ones provide access to informations like the author of a publication, title, year, sometimes abstract and if freely available also the reference to the fulltext. For the manually maintained libraries normally the same information is available and depending on the provider, the access to fulltext is possible with a subscription.

The following list contains some common sources of bibliography information.

1. Association for Computing Machinery (ACM) Digital Library[3]
   This datasource is manually maintained by the ACM. It contains beside the citation information also the fulltext of their own produced journals and held conferences which are accessible through subscription.

2. DBLP[4]
   DBLP provides access to bibliographic information in computer science by containing citation information about major journals and conferences. The term DBLP originally

[3] `http://dl.acm.org` (accessed 03-April-2011)
[4] `http://dblp.uni-trier.de` (accessed 03-April-2011)

stands for 'Data Base systems and Logic Programming' as this was the focus in the beginning of DBLP. Nowadays it is expanded to cover a much broader range of computer science and the term more generally is interpreted as 'Digital Bibliography & Library Project'. The data is maintained and added manually and consists of information about publications, like the title, year, source and the name of the authors. Fulltext is in some cases accessible through a stored web reference, but essentially not targeted.

3. SpringerLink[5]
   SpringerLink is the online library of the publishing house Springer, and thus provides access to the publications of Springer. It is maintained manually and provides access to the basic citation information for free. The access to fulltext is possible through subscription.

4. Google Scholar[6]
   The index of Google Scholar is created automatically, by collecting scientific papers freely available, as well as following citation information in this papers. Hence basic data like the title, author, year are available and if found also the reference to the fulltext. Google Scholar does not limit the thematic area indexed and therefore not only computer science articles can be found.

5. CiteSeerX[7]
   Is similar to Google Scholar, as it is an automatically created digital scientific literature library and search engine. It is maintained by the Pennsylvania State University's College of Information Sciences and Technology. In contrast to Google Scholar this datasource concentrates only on the computer science domain.

6. Microsoft Academic Search [8]
   The Microsoft Academic Search is another example of an automatically created index. It currently covers 16 different domains (as of October 7th, 2011). Beside providing a search interface like Google Scholar and CiteSeerX, Microsoft Academic Search also offers other visualization options, for example a domain trend chart, showing proper development of trends. Additionally to the basic bibliography information, information about the author, like performance indicators, and the institution the author works for, like the geographic coordinates, are provided.

When we now look at current approaches of analyzing bibliography data in the computer science domain [8, 21, 51, 64], we see that the analysis uses either DBLP or CiteSeerX as their source. As we listed above there are various other sources for bibliography information. The ACM Digital Library, SpringerLink, Microsoft Academic Search or Google Scholar, also offer access to bibliography data in different quality and with different characteristics.

In Table 1.1 we see the different characteristics of the datasources. The characteristics are *Downloadable*, which means if the whole dataset is downloadable for free. *BibTeX* indicates

[5] `http://www.springerlink.com` (accessed 03-April-2011)
[6] `http://scholar.google.com/` (accessed 03-April-2011)
[7] `http://citeseerx.ist.psu.edu/` (accessed 03-April-2011)
[8] `http://academic.research.microsoft.com/` (accessed 09-October-2011)

	ACM	DBLP	Springer	CiteSeerX	Microsoft	Google Scholar
Downloadable	no	yes	no	yes	no	no
BibTeX	yes	yes	yes	yes	yes	yes
Publications	1,7mill	1,7mill	NA	NA	27,1mill	NA
Abstract	yes	no	yes	yes	yes	partly
Fulltext	no	no	no	partly	partly	partly
Automated	no	no	no	yes	yes	yes
Author Information	yes	no	no	no	yes	no
Author/Institute association	yes	-	-	-	yes	-
Institute Geoinformation	no	-	-	-	yes	-
Classification	yes	no	no	no	no	no

Table 1.1: Characteristic of Bibliography Datasources (as of September 9th, 2011)

if the BibTeX[9] of a publication can be downloaded. The amount of indexed publications is represented by *Publications*. The characteristic *Abstract* indicates if the abstract of a publication is accessible in the datasource. A similar characteristic as for the abstract is represented by *Fulltext*, which shows if the fulltext of an indexed publication is available. The kind of index, if it is created automatically or manually is represented by *Automated*. If more information about an author than his or her name is available it is represented by *Author Information*. Beside additional author information, an assignment to an institution or organization is indicated by the characteristic *Author/Institute association*. The geographical position of an institution is available in datasource where the property *Institute Geoinformation* is indicate with a 'yes'. Beside this additional information about institution, the availability of thematic information about publications is represented by *Classification*. So relying the study on a single source does not provide the option of analyzing all aspects of bibliography data, to do so interlinking and connecting of the available information is necessary. A question which can't be answered by just looking at one source is for example; 'Is there a bias between the venue of a conference and the origin of authors joining the conference?'. Data in geographic context is used to answer other questions, like 'Is there a cultural influence on research in computer science?'. With such information maybe answers to the performance and efficiency of the funding system and the proper influence on research can be found or at least new insights can be unveiled. Also popularity of research fields can be evaluated and reasons for maybe unbalanced research efforts in different disciplines can be discovered.

To overcome the shortcomings of current works, the datasources have to be connected, by using proper semantic information and then be analyzed regarding the new possibilities .

[9] `http://www.bibtex.org` (accessed 25-October-2011)

1.3 Aim of the Work

The goal of this work is now to analyze computer science bibliography data with Web Science approaches by interlinking different sources of information and so provide new valuable information.

If we look at the different sources from Table 1.1 we see that the amount of available data is enormous, although not all of this information is publicly accessible. First we have to identify the proper source(s) of information to answer questions in the form of 'Is there a cultural influence on research in computer science?' or 'Is there a bias between the venue of a conference and the origin of authors joining the conference?'. The identified sources then have to be analyzed on behalf of their available information and structure of this information. Transforming this information into correct semantic format to be linked to the other sources is another step, and finally the visualization of the results.

To be able to answer the questions from above, one of this semantic information is the categorization of articles into subfields of computer science. To do so either classification information from sources directly can be used, for example the ACM Computing Classification System (CCS). CCS is a taxonomy created by the ACM to provide authors a framework for categorizing their publication. On the other hand an own categorization can be established (e.g. tag-cloud created from titles). The geographic context is also important for answering our question and thus we have to enrich the bibliography information, which consists mainly of titles and author names, with the geographic information about authors. With this enriched dataset proper analyses are performed.

Investigating discovered differences is an additional step. Explanation may be broken down into different parts. The funding system of a country in combination with the performance indication, which is important to satisfy the public funding, influences research on long term and can therefore be an explanation. What also has to be considered is the digital divided world we live in. As not every country and institution has the same access to the World Wide Web as we are used to it, significant differences might be detected. What we can't influence is the language problem, there are for sure a lot of publications in different languages, e.g. Chinese but these can't be taken into consideration in this evaluation if they are not listed in the available data on the Internet. So we have to clarify, that crawling the Web for additional publication data was not seen as an objective of this Thesis. Instead, we combined and related already available data sources.

1.4 Methodological Approach

The first step is a literature research to gain insight into current analysis of bibliography data and their outcomings. Also quality aspects concerning the available data from the different sources can be figured out by looking at current works in this field.

The cultural influence on research is also one part of literature research. Hypothesis are formulated regarding the influence on research and the expected outcomes in the analysis of the data afterwards.

The final theoretical part should be a current state of Web Science, describing possibilities and expected future application for this science.

After analyzing the theoretical parts of current works an investigation of the necessary toolkit for performing the analysis is done.

Subsequent to this step, data is collected from different sources. The proper modification of the data to interlink it, the subsequent analysis of the data and the corresponding interpretation and representation are the final steps.

1.5 Structure of the Work

In Chapter 2 we show the current work in the different subfields included in this work. In Section 2.1 we formulate hypothesis regarding cultural influence on research based on findings from the humanities. In Section 2.2 we look at the current approaches in the analysis of bibliography data which also should answer the question of the quality of the data source. Finally in Section 2.3 we look at the current state of Web Science and the proper achievements there.

The next chapter, Chapter 3 describes the methodology used for this work. This consists of different parts, from document indexing, information search to clustering.

Chapter 4 consists of the realization of the work. The description of the data source in Section 4.1 provides the necessary background to understand with which data we deal in this work. Section 4.2 describes the different tools used for visualization of the analyses. Finally in Section 4.3 we present the analyses performed on the dataset.

The last chapter, Chapter 5 concludes the whole work and presents possible future directions of research, which can be based on the insights from this work, as well as which other approaches can be realized with the available data.

CHAPTER 2

State of the Art

The aim of this chapter is to review the State of the Art in our area of research. This is done in different subfields, first we look at cultural influence on research to motivate or hypotheses about possible outcomes of our analysis. In a next step we investigate current approaches of analyzing bibliography data for getting an overview of already found insights and also the different approaches used for investigation. In the last section we look at the current status of the new research discipline Web Science.

2.1 Cultural Influence on Research

In this chapter we give a brief introduction in current opinions on cultural influence on research. This is interesting, as it may support findings during our analyses concerning the research effort in the different subareas of computer science in different regions. The cultural influence has different viewpoints which have to be considered. On the one hand there are influences based on the country or region where research takes place, on the other side the society itself influences research.

In sociology the society was analyzed typically along national borders, but studies also point out that this restriction may not be accurate anymore, and that there are more functional dependencies. In his work Knelangen also states that for example in Europe the inter-national differences are reduced but intra-national differences, like the social inequality, increases [37]. This is also reflected by Greif, who traces this back on cultural beliefs, which are the expectation of people into the behavior of others [25]. His statements are justified by a comparison of the evolution in trading of the two premodern societies of the Maghribi and Genoese, whereas the experiences of each group in the past are the reason for their evolution. He concludes that the developed West is formed by an individualistic society which allows economic transaction through different cultures and contract enforcement through a independent organization. Hence this leads to a vertical social structure with an uneven welfare distribution [25]. This development is also observed by Gibbons who states that the collaboration between scientists not only

changed to a more international one because of the ease of communication, but also to a more transdisciplinarity one. All this leads to a distributed knowledge production [24].

These findings would lead to the result that cultures influenced by similar history behave also similar. Now looking at this findings we formulate the hypothesis that for developed countries from North America and Europe people from the same social hierarchy, like scientists, interact with each other and thus the difference between these countries in research is small.

Beside the discussion if cultural influence can be measured along national borders, another observations is the change how research is performed generally. Before the 90s research was taken by individuals or small groups which interacted a lot in their group but not with other groups. Funding was mainly provided by state owned institutions and the proper research areas were chosen by the scientists. The knowledge flow more or less from the research community to the society and the knowledge produced during research was accepted as profound [24]. The science influenced the culture, by creating new technologies and advances for example in the medical sector [55].

Nowadays this has changed dramatically in some cases the society creates the input for the research community by creating demands. Applied research is problem oriented and the universities must legitimate their expenses [47].

This shift in funding, which basically is an indirect control of universities, has undergone a process started from the post World-War-II time. The assumptions, that basic research in any way produces valuable knowledge which increases the welfare of society and that knowledge is a public good, opened the possibility for universities to act independently. During the economic crisis in the 1970s the budgets of governments were cut and as universities lost prestige in the public, fundings declined. Universities had the task to increase short-term efficiency and produce applied knowledge which can be directly handled by the industry. The funding system was also changed to steer into this direction by providing direct governmental funding, which is a goal oriented contract based funding for which institutions have to compete against each other. Not to forget the contextualization of research by the increase of private and industrial funding.

These changes can have unintended consequences, for example resources are concentrated on some institutions which are supposed to be top. This would lead to the problem that capabilities of lower ranked institutes can never fully be unveiled. Another problem could be the short-term research endeavor which do not allow long-term projects to be conducted. Additionally the self-reinforcement mechanism is a big problem, which means that scientists or groups who were successful are more likely to receive further fundings which increases the reputation again [23].

To sum it up research nowadays has to be responsive to social demands and therefore is influenced by it, which leads to the phenomena that research often is performed problem oriented [47]. So another hypothesis which is formulated out of this findings is that funding systems and the society contextualize research.

We looked already at the influence through geographical and social differences and now we take a look also at the influence through personal differences. Craig Rusbult proposes in his work a model which sums up this influence on research [57]. He states that a scientist is always

affected by cultural-personal factors. These factors are from various type, they can either be very practical ones, like the curiosity of a scientist and the simple joy of science and search for satisfaction and success which influence him. Also so called metaphysical worldviews are a factor, for example the empirical consistency which assumes to produce reproducible results. Ideological principles are another factor, which are basically subjective values how things should work, be it inspired religious, political, gender related and so on. Not to forget opinions of authorities which either influence a scientist by not publishing research work or not funding it [55]. So a scientist always works influenced by these different factors which are interactive [56].

As this would result in a last consequence that independent research can't be performed, Rusbult also mentions that the intensity of the influence of these factors vary widely for different fields and also individuals. He proposes so called thought styles, which are shared beliefs among groups or also individuals, about how things should be. They are delivered to scientists first during their education by thought styles of educators being transported to students. This is continued by scientific communities transporting it to the researchers [55]. For the science to be most effective a recognition of these influence and minimization is desired [56].

So another hypothesis which is formulated out of these findings is, that independent research is not possible because it always is influenced by personal factors.

We found out that there are various sources which influence research. Cultural influence which can be traced back on history, social influence through funding justification and personal factors. To be aware of this different areas is important to justify proper difference observed during the analyses.

2.2 Analysis of Bibliography Data

In this section we deal with current works analyzing bibliography citation information, present different approaches followed by the authors and interesting findings.

Many analyses are based on the difference between subareas in computer science or just concentrate on one of these subareas, for example Data Base Systems. Assigning publications to subareas is therefore a first step in many works. This assignment is either done manually by the authors, by simple adding categorization information to their publication as it is done with publications in the ACM Digital Library, where authors assign CCS classes to their publications. A different approach is to categorize publications automatically by using available semantic information. Typically for citation purpose the source where the publication was published is also available, be it a journal or a conference, and this information can be used to assign the publication source to topics. For example if a subarea as stated above is Data Base Systems, then the conference VLDB (Very Large Data Bases) can be assigned to this subarea, and hence all publications made their and the authors are categorized. This approach is for example followed by Biryukov [8], Elmacioglu [20] and Reitz [51]. We now take a look at different previous works on analyzing bibliography data and some conclusions of them.

As stated above Biryukov [8] used the approach of assigning conferences to subareas to create a dataset which is split into topical groups. In total fourteen subareas were identified and conferences assigned to them.

The aim of the work was to analyze computer science communities in DBLP, which are represented by the co-authorship in the 14 different subareas. Basically find out how scientific communities evolve and communicate with each other. The basis for the publications were build by 2626 distinct conferences which they called CS dataset. For analyzing the communities, for example the collaboration trends between subareas based on the co-authorships, they also wanted to investigate differences between relevant and not so relevant conferences. Therefore they had to create a dataset which contained top, in the sense of influence on the research area, and one with non-top, conferences. This was done by assigning to the fourteen subareas high quality conferences. To decide the quality of conferences, which is subjective, they chose commonly agreed high impact conferences and consulted different sources [1,2,3] to validate their choice. This dataset then was called TOP. On the other hand they also assigned conferences which are not of high impact to the subareas and created thus a dataset called $NONTOP$. So these three created datasets build the basis for their investigation. For analyzing population stability, which is the relation between new authors publishing at a conference and people who left the conference and the publication growth rate, which is the increase in publications at conferences per year, the co-authorship graphs of the three datasets were build and analyzed. This graphs where named G_{CS}, G_{Top} and G_{nonTop}.

Before Biryukov [8] analyzed the communities they focused on analyzing authors and interesting properties, like the career length and inderdisciplinarity, of them.

Biryukov [8] found out that looking at the length of a scientists career in the TOP dataset, these conferences are dominated by researchers with a career not longer than five years. Only Algorithm&Theory and Cryptography build an exception, were most of the scientists have a career between 10 and 15 years. In the whole dataset CS only around 1.4% have a career longer than 10 years.

In looking at these long time scientists, they most likely (71%) work in more than one subarea (avarage 2.2). In analyzing the performance of these scientists it turns out, that they produce most of their publication per year between the 6th and 10th year of their career, which could be interpreted as their time after the PhD degree, where their academic position depends mainly on their productivity.

As interesting as these findings are, it is also worthwhile to mention that only 30% to 60% of the publications from one scientiest are made on top ranked venues, the majority of researchers relies on a mixture of conferences where they publish their papers.

After investigating these properties of authors Biryukov concentrated on analyzing communities.

In a first step, the absolute and relative publication growth rate, which are dynamic measures, were examined. The absolute growth rate of a topic A_i in year y is defined as $AbsGr_{A_{i,y}} =$

[1] `http://www3.ntu.edu.sg/home/assourav/crank.htm` (accessed 12-October-2011)

[2] `http://webdocs.cs.ualberta.ca/~zaiane/htmldocs/ConfRanking.html` (accessed 12-October-2011)

[3] `http://www.cais.ntu.edu.sg/content/research/conference_list.jsp` (accessed 12-October-2011)

$\frac{Publ_{A_{i,y}}}{Publ_{A_{i,y-1}}}$ where $Publ_{A_{i,y}}$ stands for the amount of publications in topic A_i in year y. Performing the analysis on the whole dataset CS showed considerable differences. For example Computer Networks stabilized in the early 90's by 1 ± 0.1 and Natural Language Processing and Information Retrieval vary 3 times much, from year to year, up to now. As a bias between conferences may be present, the same test was done on the TOP and $NONTOP$ dataset. Performing this analysis showed a systematically higher growth rate in the $NONTOP$ dataset. The relative growth rate, defined as $RGr_{A_{i,y}} = \frac{AbsGr_{A_{i,y}}}{AbsGr_{CS_{i,y}}}$, based on the overall computer science activity, showed trendy subareas by forming out peeks in the proper years, for example a burst for the Data Mining, Data Engineering, Machine Learning field in the beginning of the 90s.

For answering the question about collaboration trends, the number of coauthors per paper and per author as well as the clustering coefficient [62], which express the relation between the connections of direct neighbors in a graph and a complete graph, were analyzed. An interesting finding was, that the relation between interdisciplinarity and connectivity of a field is weak. For example the two subareas, Graphics and Security, have similar clustering coefficient, but Graphics is the most homogeneous and Security the most heterogeneous area with respect to coauthors of an author from the same subfield. This means that in the Graphics area, the co-authors per author typically are from the same area, and in Security not.

Another analysis concentrated on the population stability, which was measured in newcomers and leavers of a conference. Additionally to newcomers and leavers a value named pure newcomers was introduced, which measures the amount of people publishing at a conference without knowing someone who has already published there. This measure was interpreted as a value which explains friendship driven publication, which means if the probability to publish at a venue where a co-author has already published is more likely then publishing at a conference where this relationship does not exist. It turned out that at top-ranked venues there is no general rule, some subareas are very stable with low newcomer and leaver rates while others, mainly young conferences show a more dynamic picture. On the other side, at non-top ranked conferences, the rate of newcomers was above 75% and the fraction of pure newcomers also about 75% which suggests that non-top ranked conferences are more often chosen as a starting point for a scientific career.

In contrast to the work above from Biryukov [8], Elmacioglu [20] concentrated only on one part, the Database community, of the DBLP data, but the approach of receiving data for this subarea was similar, by selecting venues (19 journals and 81 conferences) which represented this subfield.

The field was analyzed statistically to gain insight into the Database community, therefore properties like new authors rate per year, the active authors rate per year, average number of papers per author and so on were analyzed. An interesting finding was that more than half (63%) of the authors have only one paper published and just a few a large amount (only 18 authors with more than 100 publications), which shows that the law of power is present.

The amount of new authors per year, more than 1000 in 1991 and over 3000 in 2003, and the trend to collaborate, from around 2 collaborators per author to over 3 in 2003, grew with a steady increase. The increase of co-authorship can be seen as either the pressure of making

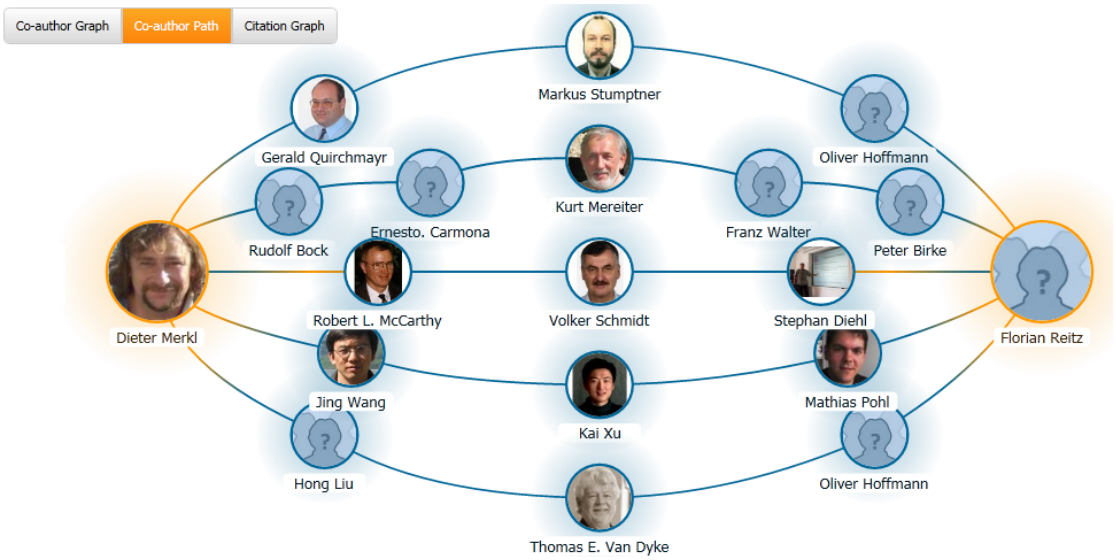


Figure 2.1: Example of the geodesic from author Dieter Merkl to author Florian Reitz based on data from Microsoft Academic Search (as of November 14th 2011) visualized by VisualExplorer [17]

publications to justify fundings or the easier way of communication through the WWW. Purely single authors, in the sense of authors who just publish on their own, seem to diminish as the fraction of active single authors to active authors in 2003 was below 0.1%. The activity of the community was meassured by looking at the number of papers and number of authors per year. It turned out that the avarage amount of new papers per year per author has stabilized by around 0.3 papers. A steady increase in published papers, is therefore explained by the also steady increase of authors. And as mentioned above a trend to collaboration was found during the analysis.

Beside the information gained about authors, also the Database community was analyzed, therefore the structure was investigated with different approaches. First the giant component was identified, which consists of the largest interconnected subset of nodes in the co-authorship graph. The size of this component was steadily growing to 57% (18542 authors) in 2003, and the second largest one only consisted of 51 authors. This finding shows, that the Database community has one center, although it was believed that this component would be even bigger. Additionaly the geodesic (the shortest path) between authors was analyzed, this value is calculated as the avarage of the pairwise pathlength from one author to another author by following the co-authorships, and it turned out that from avarage 8 in 1983 it stabilized around 6 from then on, which shows the 'small world' [62] effect of the Database community with this relative small avarage distance between authors. For an example we visualized the co-author path between Dieter Merkl from the Vienna University of Technology and Florian Reitz from the University of Trier with the VisualExplorer [17] based on the data from the 14th November 2011 in Microsoft Academic Search (see Figure 2.1).

Reitz [51] again concentrated on the whole bibliography data of DBLP, and used a very similar approach as Biryukov [8] for connecting publications with topics, by assigning conferences to thematic subcategories and hence classify the publications, the framework used in this work is based on Laender [40]. Laender created for his work a framework consisting of 30 theme groups. These groups are for example *Algorithms and Theory*, *Artificial Intelligence*, *Programming Languages* and so on. As Biryukov, Laender also assigned to every group conferences, the major difference is, that Laender used also conferences which are not listed in DBLP, and so an overview of the coverage of a subfield in DBLP is made.

The research question in the work of Reitz [51] was focused on analyzing the coverage of the different computer science subareas in the DBLP dataset, the work concentrated more on the structure of the datasource. As analyzing DBLP data always has a bias on the proper coverage of the subarea investigated, and in the very well beginning DBLP was mostly concentrated on database technology and logic programming, research about the evolution of this coverage is vital. To reconstruct the evolution, backups of the system from 1995 to 2009 were used. The historic analysis showed, that more than half of the papers were added a year after publication and also the dominant (over 95%) kind of publications are conference papers and journal articles.

As Elmacioglu [20] explains the steady increase of research papers in the Database community by the growing amount of authors, in this work the high increase in new conference papers and journal articles is justified by a growing average number of papers per author because the amount of new authors per year (90,000) has not changed since 2003.

For analyzing the coverage of the different computer science subareas in DBLP, in a first step conferences with proceedings in DBLP were compared to the conferences assigned to this subfields. It turned out that there are significant difference in the coverage. Database, Information Retrieval, Digital Libraries and Data Mining, representing one theme group out of the 30 from Laender, is covered with 95% of their conferences in contrast to Computer Education with just 37%.

Looking at the progress over time, the best covered themes, database and logic programming, which were the main focus in the beginning of DBLP, always had a better coverage then all the other themes. Subareas covered with not more than 10% before 2000, still remain underrepresented by not more than 60% of their conferences.

These analysis showed that beside the analysis of the existing data a critical view on the datasource itself is also important.

Zaiane [64] used a different approach for identifying thematic subareas. In contrast to assigning sources of publications to topics he used the information implicit available in the citation. He analyzed the titles and used the frequent items contained in the titles for representing thematic information. As this approach also found a lot of non topic relevant items, he concentrated for the topical representation on frequent item-sets with a length of two which reflected topics in a good manner, for example Relational Database, Neural Network and so on. This information was used to assign papers to topical subfields by matching occurence of frequent bi-grams with occurrence of them in the title. As the amount of all bi-grams was to huge, he limited this to the 1000 most frequent items. For analyzing community properties the co-authorship graph was build.

After reflecting the framework behind the work of Zaiane I introduce the system proposed in his work [64] which is called DBConnect[4]. This system not just analyzes the communities in the DBLP data through co-authorship but also proposes collaborators for research based on a random walk approach. Therefore the datasource described above was represented as graph, once as bipartite (with author-conference) and once as tripartite (author-conference-topic) graph.

To include the information about co-authorship, the random walk algorithm was extended. The modification took place in representing the conferences with additional virtual leaves in the graph, with directed edges to express the inherited information. As the resulting adjacency matrix was to big to fit in the main memory a graph partition was performed before analyzing an authors network. The computational problem and also the topic assignment, based on the relative short titles, is seen as future work and subject to be improved.

Dalibor Fiala [21], in contrast to the previous works used CiteSeer, the predecessor of CiteSeerX, as datasource and not DBLP. As he also took the co-authorship graph for his analysis he created it out of the publication citation graph. The graph consisted of 411 thousand authors and 4.8 million citations. Beside a different datasource the work is mainly oriented to find influential scientists and validate the quality of computer generated bibliographic data in CiteSeer by comparing differnt rankings to the ACM SIGMOD E. F. Codd Innovations Award winners.

The number of citations and in-degree of the author citation graph were used as basic metrics beside other different approaches, one of them was an extended approach of the PageRank concept, which weight citations from foreign research higher than from colleagues. On evaluating the ranking of the award winners with the calculated metrics, it turned out that in-degree rankings and citation have the largest overlap with the list of awardees and therefore provide a good basis for finding influential scientists.

This section gave an overview on current works in the field of citation analysis. Two approaches of how categorization information can be added to publication citation information were presented, either by classifying the source of publications or by analyzing the topical information from publication titles. Additionally, interesting findings are the small amount of authors with a career longer than 10 years and the fairly large amount of authors with just one paper published. The finding that scientists have an average of 2.2 different topics they investigate during their life and are most productive between their 6th and 10th year of publication. It also turned out that non-top conferences seem to be selected more often by scientists to publish their first paper. That the co-authorship in the Database community follow a small world effect and that citation seem to represent the value of an author pretty good are also interesting observations. Additionally structural findings of the DBLP citation dataset, for example that publications typically added a year after their publishing and that the various subfields of the computer science domain are differently covered, are important for further investigations.

[4] `http://webdocs.cs.ualberta.ca/~zaiane/pub/dbconnect.html` (accessed 09-October-2011)

2.3 Web Science

This section introduces Web Science, by providing information about the background behind this science, and is followed by current works in this field.

Web Science is the name of an initiative started in 2006 [6] by a workgroup consisting of, among others, Tim Berners-Lee, Wendy Hall and Nigel Shadbolt. The initiative arose through the observation that understanding the Web is crucial for its further success and therefore a new research discipline, Web Science, has to be established with the interdisciplinarity of the Web as main research focus. The intention for the Web in the early 1990 was the interchange of scientific research papers, but it has undergone many changes, grown to a worldwide scale, influencing the society and the way we work, and although it is that powerful, research in this field, with the Web as main focus, is rare.

A basic understanding already evolved in Web Science. That the Web is more than the sum of its pages, and emergent properties are transforming society, are accepted facts [59]. Some insights already exist, for example PageRank, the ranking of search results based on the importance of a site, calculated through the amount of links pointing to a site, is a finding based on understanding the structure of the Web. The scale-free characteristic of the link structure, which means that there is no average number of links to and from a site, the small world effect, which allows to get from one site to every other site by a certain amount of clicks are additional insights already gained.

In [6] it was stated that the pure extraction of semantic out of textual data is unrewarding, and that a trend to data sources with logic in its elements is there, but the challenge would be connecting independently built sources. As there is an enormous amount of documents already on the Web, this information also has to be filled with semantic, and the only way to achieve this is by the users themselves, therefore Shadbolt [59] argues that we have to analyze the society, in the aspect of what's the motivation for people to add semantic to available data, to gain semantically enriched data in bulk.

Hendler [31] also deals with the problem like Berners-Lee [6], that although the Web has such a big influence, and changed completely how scientific research is performed, most of research about the Web gets classified miscellaneous. Although on curricula you can find Web design and similar courses, it seems that the underlying principles a rarely covered. TCP/IP and fundamental networking is taught, but the Web is still considered just as an application running on top of it, for delivering content, rather than a phenomenon worth an investigation of its own.

Therefore Hendler [31] also states, that the new science of the Web should merge the analytical part of physics with the synthetic part of computer science to try to understand the Web on its own. Analyzing technical improvements and pieces of engineering at the micro scale if designated goals are reached build the basis, continued by the interaction of humans and followed by the analysis of emergent properties at the macro scale, which is proposed by [31, 59]. To do so a new understanding of the software development process on the Web is presented. The Web hasn't used a typical design, implement, test approach, it was more built on the frame-

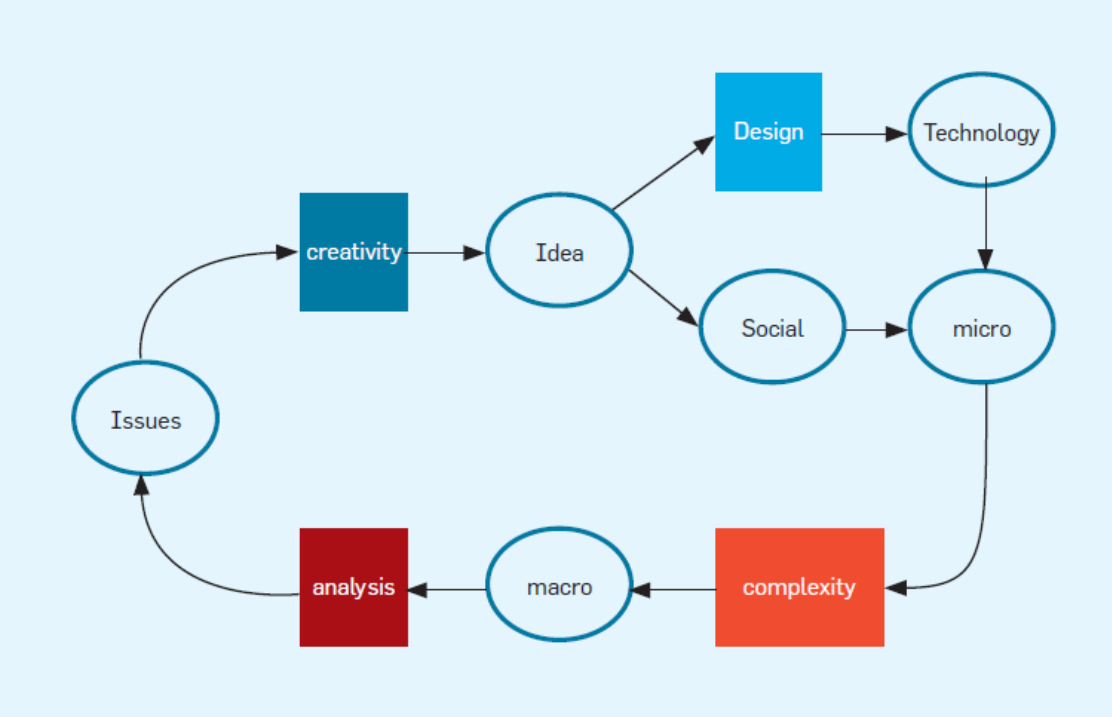


Figure 2.2: Development Process on the Web [31, S. 62]

work shown in Figure 2.2. An application on the Web is created out of an *Idea*. It is based on a *Design* which is realized by available *Technology*. This Web application is then used by a small group, where the *Social* effects at *micro* scale are tested. With an increasing amount of users, *complexity* increases and emergent properties at *macro* scale evolve. An *analysis* of these properties is necessary, as with the macro scale also new *Issues* appear. These issues build the basis for further development as they support *creativity*, which closes the loop by building new Web applications. Issues at macro scale are hardly predictable, to form the future of the Web, systems which produce the desired effect at macro scale are necessary.

The success of the PageRank algorithm was mentioned, it is based on the representation of the Web as a graph, which is still very common, but it is just one abstraction of the Web. Data which is represented depending on the type of the request, the 'deep' Web, is not accounted in current graph models. Dynamics which consist in the Web are therefore not well covered in the representation as a graph.

Hendler [31] picked an interesting example in his paper, by showing that the link structure of Wikipedia, although mainly a managed corpus, has the same scale-free properties as found on the Web generally. What is interesting is, that although many other sites build on the same software, MediaWiki[5], no other site has gained that influence and size as Wikipedia has. So the community behind such sites, contributing and supporting it, and therefore being responsible

[5] `http://www.mediawiki.org` (accessed 26-October-2011)

for the success of such initiatives must be better understood. Such social machines are in a very early phase, and today are based on trial, use, and refinement. To successfully engineer such machines some questions have to be answered, among others, 'How do cultural differences affect the development and use of social mechanisms on the Web?' [31, S. 67]

Consequently a part of Web Science, deals with the current shift from Web 2.0 user generated content, to the Semantic Web. This next generation Web, which was already foreseen 1996 by Tim Berners-Lee [5], is based on semantically rich data. Such Linked Data forms the basis for further development of the Web and gives the opportunity for a bunch of new applications.

Berners-Lee showed in [7] how our life may change with the Semantic Web. An agent, just parameterized with a few properties, can automatically find and combine information from the Web and generates the proper output for the user, in this example a medical treatment plan.

The basic technologies for the future Web already exist since 2001. Knowledge representation, the access to structured information and set of inference rules for automatic reasoning, are parts of this technologies. One extension, which in similar way was also crucial for the success of the WWW, that total consistency can't be guaranteed, has to be extended on the traditional systems. Unanswerable questions have to be possible in the usage of theses systems, this drawback has to be accepted to form a proper suitable approach for the Web.

The Ressource Description Framework (RDF) [36] also builds a key part in the possibility of adding semantic to data on the Web. This approach of subject, predicate and object is suitable for expressing most semantic. Providing ontologies, which consist of predefined predicates, inference rules for relations and the possibility of automatic reasoning allow us to create new valuable information by describing data with RDF.

Wendy Hall took a look at the evolution of the Web and provides some interesting observations [29]. That although the vision of the Semantic Web was already formulated in 2001, until 2009 it was still difficult to set up connected semantically rich information. Hall said that Web 2.0 technologies are a driving force for the Semantic Web, by generating a huge amount of user-generated content. For the success of Semantic Web, communities started to develop applications based on these principles, and so it is now important to analyze and investigate the properties of such developments and the proper consequences to better understand the future evolution.

Also interesting is, that eScience, a computational intensive science, nowadays also often provides the data in a semantic enriched form, like the COMBECHEM project [61]. Hendler [31] also states that nowadays it is possible for students to experiment with large-scale Web-programming based on the distributed computational power available and usable through Web technologies.

As Hall stated above, communities already created applications around this data. Auer shows in [1] an example for the next generation of the Web, the Semantic Web. He describes DBpedia[6],

[6] `http://dbpedia.org` (accessed 26-October-2011)

which provides structured access to information from Wikipedia as well as connections to various other sources. In his work he explains the process of extracting information and providing semantically rich interfaces for machines as well as humans.

To answer semantically rich questions a broad data base is necessary, and the typical top-down approach, building an ontology before gaining data breaks on the scale of the Web. Therefore a more grass-roots-style is desirable, which leads to the challenging problem of collaboratively edited data, contradictory data and inconsistent taxonomical conventions. As a first step, DBpedia concentrated on extracting Wikipedia information and build structured knowledge out of it. Extraction, providing access, interlinking to other datasets, and accessible Web services were the main goals.

For the extraction of the data, dumps from Wikipedia were used, and enriched with semantic information. Different access methods, Linked Data[7], the SPARQL[8] protocol and RDF dumps were provided for the data. Browsing the linked data, navigates through the URIs added in the extraction process, and returns a meaningful information for each URI. The SPARQL endpoint provides the opportunity to query information, although limited to protect service overloads. Dumps can be downloaded and further processed in any other application.

DBpedia not only adds semantic information, it also interlinks with other open datasets. Links pointing to other web sources and also links pointing to DBpedia provide the option of gaining cross referenced information about a topic. To enable all this, RDF, as described in [7], is the key enabler. Subjects and objects are identified with unique URIs, and so the possibility to link from anywhere to this information is provided.

Bizer [9] analyzed the status of DBpedia again in 2009 and explained in more detail the extraction process.

Beside the described method from [1] with loading Wikipedia dumps into the application, a live extraction was implemented. With this approach the time lag of DBpedia information to Wikipedia information is now between 1 to 2 minutes, which means basically steady up to date information in DBpedia.

One main problem still is the extraction of the user-generated content, as the information is heterogeneous, the templates for infoboxes for example vary from different types for the same information to different attribute names. If you look at the infobox for cities in Japan[9] in contrast to the infobox for a town in Switzerland [10]. Both of them describe a city, but beside they consist of different fields, they also name same properties different. What for Switzerland is the 'postal_code' is for Japan 'CityHallPostalCode'. Therefore a generic extraction approach as well as a mapping-based one is implemented. With the generic approach a wide coverage is achieved in contrary the mapping-based approach tends to increase the data quality. As it is typical at the scale of the Web, it is impossible for the DBpedia project to create mappings for all kind of infoboxes and so crowd sourcing is necessary to fulfill this task.

[7] `http://linkeddata.org` (accessed 26-October-2011)

[8] `http://www.w3.org/TR/rdf-sparql-query` (accessed 26-October-2011)

[9] `http://en.wikipedia.org/wiki/Template:Infobox_city_Japan` (accessed 09-October-2011)

[10] `http://en.wikipedia.org/wiki/Template:Infobox_Swiss_town` (accessed 09-October-2011)

The size of the knowledge base reported in this paper [9] was 274 million RDF triples and 2.6 million entities, which is enormous. All these entities are classified with four different systems, a DBpedia ontology, manually created through the most used template boxes, Wikipedia Categories, which are kept up-to-date but don't form a proper hierarchical system, YAGO[11] which encodes a lot of information into the class itself and UMBEL[12], which was designed for interlinking Web content and data. As it can be seen above on the numbers, each entity is described by a lot of RDF triples, which are extracted generically, mapping-based or are basic general information. The distribution of the properties over the data follows a power law which is typical for small world networks.

For accessing the information, beside the interfaces, Linked Data, SPARQL and dumps, a lookup index was created. This service supports data publishers by proposing proper DBpedia URIs for a given label. Therefore a search after a given label returns proper URIs describing this label and ranked through a similar approach as PageRank.

During the first years, DBpedia has grown to a hub for the Web of data. It consists of 3.64 million entities[13] and if you look at the outgoing and incoming links, more than 4.9 million outgoing RDF links, and 23 external [9] data sources pointing to DBpedia demonstrate the dimension of it. As the topic coverage is, like Wikipedia, almost for every subfield, annotating blog posts and other content of the Web with URIs from DBpedia increase, which hence provide another step towards the Web of Data.

Mobile applications like DBpedia Mobile already take advantage of the data available in DBpedia. It provides semantically enriched information about the current geo location of a user. Relationship Finder and Query Builder also provide access to the dataset by allowing to query for information. Content annotation is implemented in many projects to point to DBpedia URIs, for example Thomson Reuters and BBC have implemented such approaches to semantically enrich their information. BBC therefore implemented the so called MuddyBoots which aim was to identify people and companies in news stories and add proper DBpedia identifiers for identification [38]. Thomson Reuters also maintains a project for this, which is called Open Calais[14]. It is free of charge and works in following way, if an unstructured text is submitted it returns RDF enriched data with identified entities, persons, companies and so forth, for an example see Figure 2.3. There the sentence 'If I want to travel to the city of Salzburg to visit the birthplace of Wolfgang Amadeus Mozart, which train can I take from Vienna?' is submitted and the Open Calais returns it enriched with the semantic information of Salzburg and Vienna as cities and Wolfgang Amadeus Mozart as person.

All in all, DBpedia has grown that much in the last years, that it builds a great test area for evaluation data integration, reasoning and uncertainty management techniques.

As stated in [7], some drawbacks have to be accepted, and similar key success factors as for the Web, openness and extendability, are valid for Linked Data. Tom Health et. al. show in [30]

[11] `http://www.mpi-inf.mpg.de/yago-naga/yago/` (accessed 26-October-2011)

[12] `http://umbel.org/` (accessed 26-October-2011)

[13] `http://blog.dbpedia.org/2011/09/11/dbpedia-37-released-including-15-localized-editions` (accessed 26-October-2011)

[14] `http://www.opencalais.com/` (accessed 10-October-2011)

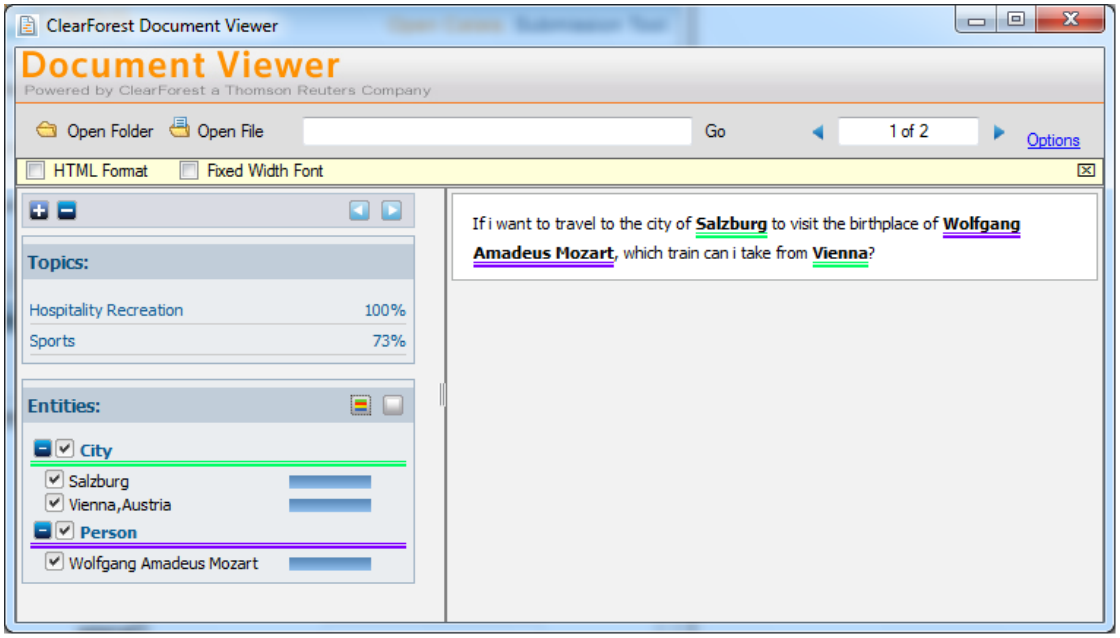


Figure 2.3: Demonstration of OpenCalais with the Calais Submission Tool [52]

the basic principles of Linked Data. As it is technically published data with machine-readable links it enables a bunch of new possibilities. As principles for data publishers four key points have emerged, use URIs, if possible HTTP to provide access to them, provide standardized interface for access, and link to other URIs to provide further information, are these principles in short.

As the other papers already showed, crowd sourcing, is a driving force involved in publishing Linked Data. Calculating the current size of the Web of Data is not possible, but estimations speak of about 4.7 billion RDF triples (as of May 2009) [15] [16]. The mentioned URIs which provide the possibility of interlinking, often are subject to the problem, that different URIs are used for the same physical things, but this also provides the opportunity of creating different views on the same things.

A main challenge is validating the accuracy of a dataset, therefore it is important to provide metadata, when the data was created, the method of creation and so forth, so that agents can choose which datasource they can trust. On the other side, publishing data in RDF manner is no problem anymore, a lot of frameworks and tools were developed to facilitate RDF creation. D2R Server[17] is one of these tools which provides the opportunity to publish relational data with semantic by providing a declarative mapping. Other tools for example are Virtuoso Universal

[15] `http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/LinkStatistics` (accessed 12-October-2011)

[16] `http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics` (accessed 12-October-2011)

[17] `http://www4.wiwiss.fu-berlin.de/bizer/d2r-server` (accessed 30-March-2011)

Server [18], Talis Platform[19] and Pubby[20].

A lot of applications also were created to fit the requirements, from Linked Data browsers which assist the users in navigating through the Web of Data, to search engines based on this datasets. Search engines can be divided into human-oriented ones, which provide mainly a similar interface than commonly known search engines and are keyword-based, and application-oriented ones, which purpose is to provide RDF links for an URI to interlink the proper data with it. Domain-specific applications, like already stated above also exist in numerous ways, from geo-applications, to data integration technology and so forth.

Future challenges are to overcome shortcomings in user interface, for example the possibility to add and remove data sources, the problem of the runtime link traversal, which must be overcome when the amount of data gets to huge. Data integration and schema mapping is also very crucial for the success as well as link maintenance. Although dead links are allowed, and build a key part of the Web, too many of them hinder the system growing. Trustworthiness and privacy issues are another field which has to be investigated for further development.

[18] `http://virtuoso.openlinksw.com` (accessed 12-October-2011)

[19] `http://www.talis.com/platform` (accessed 12-October-2011)

[20] `http://www4.wiwiss.fu-berlin.de/pubby` (accessed 12-October-2011)

CHAPTER 3

Methodology

This chapter contains the theoretical toolkit used during analyzing and processing the data in this work. We explain the methodologies, their background and the proper use cases in this work.

3.1 Document Indexing

In this section we introduce the basic concepts necessary for document indexing, this is important because we analyze a corpus of document titles and therefore have to build a proper representation of them. Document indexing is the process of building such a representation where the typical steps in such a process are parsing and tokenizing, stop word removal and stemming [46].

Parsing and Tokenizing

Digital documents are typically stored in a sequence of bytes. In document indexing the first step is parsing this byte sequence into a sequence of characters. To perform this step the correct encoding has to be determined, this can be done either manually by the user, or automatically by reading some metadata from the document.

Before we continue, some distinction of the used vocabulary for indexing is important, therefore we summarize it

- **token**: is an instance of a sequence of characters
- **type**: is the summary of tokens containing the same characters
- **term**: is a type used for indexing in the Information Retrieval system

Subsequent to converting a byte sequence to a character sequence, determining the document unit is important. The document unit describes which portion of the data should be treated as one document. For example, a file in a directory is one document, or a ZIP file should for example

be treated that each document in the ZIP file is one document for indexing. Also indexing a whole book or each sentence as one document is either way not appropriate and therefore the index granularity has to be chosen properly.

Succeeding to parsing the next step is tokenizing. If a character sequence is present and the document unit is defined, tokenizing is the process of splitting up the input into single parts, and eliminating punctuation. The following input is converted to the tokens presented by output.

Input : The question: 'Do you believe?'

Output : [The] [question] [Do] [you] [believe]

A main problem in tokenization is, 'What are the correct tokens a document should be split up?'. Typical approaches are, eliminate all punctuation and split on whitespaces and all non-alphanumeric characters. This approach is in some cases very helpful, but can also lead to major problems. Take the case of documents from the technical domain with very specific terms like 'C++', where the '++' should not be dropped or split because the meaning of just 'C' is completely different to 'C++'. Another example would be the term 'on-line' which can be written with a hyphen, a space 'on line' or as one word 'online'. The tokenization and index process should map these possible forms to the same term to provide the possibility to search for it. Hence a domain-specific tokenization is important. Beside the domain, the language is also a problem, as English is pretty simple, other languages like German have various problems like compound nouns (e.g. 'Postpaketzustelldienst'), to overcome such a problem approaches exist which split the token if parts of it can be found in a dictionary. Some Asian languages on the other side have the problem that there exists no whitespace at all. Hence beside domain specific indexing, the index process should be also language aware. As in this work, the text to be indexed is English no proper language aware indexer has to be used.

An important part of the indexing process is, that the tokenization process applied to the document corpus, also has to be applied to the query thereby consistency of the same representation of terms in the document collection and query is ensured [45, Ch. 2].

Stop Word Removal

Subsequent to parsing and tokenizing in a next step very common words which appear frequent in documents and are little to no help on searching for particular documents are eliminated. Such words are for example conjunctions like 'and', 'or' or 'then' and articles like 'the', which are called stop words. To be able to remove such words, in a first step a list of such frequent common terms has to be created. One approach for creating a stop word list is to determine the collection frequency (also known as inverse document frequency), which is a measure that expresses the value of a term for querying in regard to a collection of documents. The most frequent ones are then added to this list, which would subsequent be removed from the documents by the index process.

These lists always have to be hand-filtered in the context of not removing vital terms for the proper application. Beside the approach of creating stop word lists, there also exist predefined

stop word lists for example the stop word list from Onix Text Retrieval Toolkit[1] or the stop word list from the Apache Lucene Project[2], included in StopAnalyzer.[3]

Apache Lucene StopAnalyzer stop word list:

'a', 'an', 'and', 'are', 'as', 'at', 'be', 'but', 'by', 'for', 'if', 'in', 'into', 'is', 'it', 'no', 'not', 'of', 'on', 'or', 'such', 'that', 'the', 'their', 'then', 'there', 'these', 'they', 'this', 'to', 'was', 'will', 'with'

One major difference between these two lists is the size, as the list from Onix consists of 429 terms, the list from Lucene only consists of 33 which also shows the trend, that rather short lists are used nowadays because of the little impact on total costs in terms of query processing time [45, Ch. 2].

Stemming Algorithms

The next step in the indexing process is stemming. It is a basic process in document indexing, increasing recall (the ratio between relevant documents received for a query and total relevant documents in the collection) of queries on a document corpus, through replacing inflected and derived forms of words with a root representation and therefore normalizing the concepts. It is also very important in short documents [39].

The goal of stemming is to increase matches between queries and documents by normalizing tokens. For performing this, different approaches exist. Two examples are: applying a predefined set of rules or maintaining a list with relations between tokens, which are used to expand the query or the index [45, Ch.2]. The result of all of these methods should be the same, queries on a document corpus should also return results which contain the query terms in inflected or derived form.

As mentioned above different approaches exist for achieving this goal and performing proper stemming, some use a stem dictionary and other just rely on replacing suffixes. We introduce two of the most common used approaches, the Porter Stemmer and KStem.

Martin Porter proposed 1980 a suffix stemmer [48]. The motivation behind it is, that the removal of suffixes supports the reduction of the term vector which represents documents. This is achieved by conflating term groups into single terms, and hence reduces the size of an information retrieval (IR) system. The system is based on a set of rules containing a list of suffixes and the proper replacements as well as the criterion's under which the rules are applied.

The aim of this approach is not to create correct words, it targets the goal of creating representation of words created through applying the rules. These rules are applied during the index process and on queries. It would be an even better IR system if two distinct words are conflated to a single stem if it can be determined that the meaning is the same, but this can't be achieved by using Porter's stemmer.

[1] `http://www.lextek.com/manuals/onix/stopwords1.html` (accessed 15-October-2011)
[2] `http://lucene.apache.org/` (accessed 15-October-2011)
[3] org.apache.lucene.analysis.StopAnalyzer

Despite everything, Porter Stemmer is a fairly ease algorithm to implement which is also very fast and has therefore a reasonable diffusiveness. The algorithm is based on some definitions which I summarize here.

- **Consonant:** is a letter other than A, E, I, O and U and other than Y preceded by a consonant (TOY - Y is consonant, RELATIVITY - Y is vowel)

- **Vowel:** if a letter is no consonant, it is a vowel

A consonant is represented by a *c*, and a vowel by a *v*. Lists of consonants and vowels with a size greater than zero, are represented by a *C* respectively by a *V*. Optional parts are enclosed by [ ]. Consequently a word can be represented as $[C](VC)^m[V]$ where m is called *measure*. The set of rules applied to tokens is represented as:

$$(condition)S1 \rightarrow S2$$

The $condition$ mostly is of the form $(m > \mathbb{N})$. $S1 \rightarrow S2$ stands for, suffix *S1* is present and if the condition holds, it can be replaced by suffix *S2*. Additionally to the possibility to express the $condition$ as $(m > \mathbb{N})$ some special operators can be used in the condition, which are:

$(*S)$ the stem ends with the letter S (can be used with other letters as well)

$(*v*)$ the stem contains a vowel

$(*d)$ the stem ends in a double consonant (e.g. -LL)

$(*o)$ the stem ends *cvc*, where the second *c* is not W, X or Y (e.g. -DAS)

$(and/or/not)$ conditions can be connected with and/or or negated with not

With the definitions above, the set of rules can be defined. The rules are grouped into five steps, and if a stem matches more than one rule in one step, the rule with the longest matching *S1* is applied.

The five steps are performed during the stemming process, which conflates term groups to single terms. Step 1 is divided into three sections. We don't list the rules here in detail, see [48], but step 1 aims to stem plurals and past participles. Step 2 concentrates on replacing derived forms with their base form. Step 3 to 5 moreover try to remove unnecessary long stems, in total around 60 suffixes are checked for stemming. As stated before, this approach does not take care of linguistic basics, it just tries not to remove a suffix if the stem is too short by obeying the length of the stem through *m*.

Robert Krovetz in contrast to Porter, concentrates more on the morphology of words, as he described his stemming approach in 1993 [39]. This approach is nowadays widely known as KStem and is also used in Chapter 4 of this work.

Morphology is basically build out of two parts, the inflection of words and the derivation of them. As inflection more or less is rule based for building plural and tenses, the derivation of words can also change the meaning of it by transforming words into different types (e.g. 'gravity' in the sense of force, in contrast to 'graveness' in the sense of serious). As stated, Porter does not care about the meaning of words and does not try to find linguistic roots, consequently the rules applied to some words, hinder them from being removed as stop-word, like 'doing' which would be removed as stop word, but stemmed to 'doe' it will be kept in the index process.

Krovetz in a first step extended Porters approach by adding a dictionary, the Longman Dictionary of Contemporary English (LDOCE)[4], and checking the word against the dictionary before applying one of the 5 steps of Porter. The approach was tested on four different collections covering, computer science, law, physics and newspaper stories. As the performance of this algorithm was even worse in some cases, a entirely new algorithm was implemented.

KStem has 5 main aims:

1. produce words instead of stems
2. do not conflate words with different meanings
3. broad coverage (conflate as many word-forms as possible, limited to point 2)
4. provide a proper performance
5. be part of an algorithm for word-sense disambiguation

Started with the evolution of an inflectional stemmer, by looking at the inflectional endings in the collections, it turned out that around 50% of the endings are caused by plural and the rest is evenly divided between tenses and aspect. Prior any stemming the dictionary is inquired to check if the word is present there, and if it would be so, no stemming is performed and the word would be returned.

One of the first steps in KStem is to find the root form of words ending in 'ing'/'es'/'ed'. To achieve this endings are replaced by an 'e' or removed as a whole. First they are replaced by an 'e' and the dictionary is checked if the word is present, if not, the ending is removed completely and the dictionary checked again. For words appearing in the dictionary in both versions (with an 'e' and without) a exception list is created which is conducted for words ending in 'ing'/'ed' as the correct root for this words is the word without an ending. For example 'suited' will be reduced to 'suit', but 'suites' will be reduced to 'suite'. In total the inflectional stemmer uses a three step approach, by first converting plural to singular, second past tense to present tense and third removing trailing '-ing'.

The derivational stemmer, as extension of the inflectional, is conducted very conservative. If a word-form is found in the dictionary it is not stemmed, assuming that words appearing in

[4] `http://www.ldoceonline.com/` (accessed 15-October-2011)

a dictionary have a different meaning than derived forms of it, hence the derivation of words with different meanings is prevented. For improving the inflectional stemmer with derivational parts, a list of 106 derivational endings was created through LODCE and the collections were analyzed regarding the most frequent endings. It turned out that the most common endings are: 'er', 'or', 'ion', 'ly',' ity', 'al', 'ive', 'ize', 'ment', 'ble', 'ism', 'ic', 'ness', 'ncy' and 'nce'. The identified endings were included in the inflectional stemmer by an own procedure, replacing the endings, if the word is found in the dictionary without it, and therefore building a derivational form of the same meaning.

To give a better overview of how the stemming process changes a text, here an example of a text stemmed with Porter and KStem.

Original this example should show how the different stemming procedures create a stemmed text out of their input by performing modification on the original text

Porter thi exampl should show how the differ stem procedur creat a stem text out of their input by perform modif on the origin text

KStem this example should show how the different stem procedure create a stem text out of their input by perform modification on the original text

3.2 Information Retrieval

Looking at the history of Information Retrieval (IR) described by Lesk in [43] or by Singhal in [60], both state Vannevar Bush's article 'As we may think' from 1945 [11] as a ground breaking article for IR. The need of storing information and efficiently retrieve this stored information has been there, but Vannevar Bush was the first who described the automatic processing of a large amount of information and the retrieval of it. IR systems can be defined according to Lancester in the following way:'An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.' [41]

So an IR system returns results to queries and therefore the performance of such a system has to be evaluated. The typical measures for evaluation the performance of an IR system are recall and precision. These two terms got defined by Cyril Cleverdon in the 1960s [13, S. 34-36], where recall is defined as the ratio between relevant documents received and total number of relevant documents in the collection. Precision on the other side is the amount of relevant documents received divided by the number of received documents. In his work Cyril also states that an inverse relationship between these two factors exist, if returning more documents from the collection, the probability that more relevant ones are included, recall, increases, but also the probability of more irrelevant documents in the received documents, precision, drops.

$$Recall = \frac{\text{relevant Documents received}}{\text{total relevant Documents in the Collection}}$$

$$Precision = \frac{\text{relevant Documents received}}{\text{total Documents received}}$$

Two factors influencing recall and precision are the the level of exhaustivity of document descriptions and the level of specificity of the terms of the index language. Exhaustivity, as Sparck Jones describes [35], is a property concerning the description of a document. If all topics from a document can be described by the index terms then the document is exhaustively described, consequently a exhaustive described document, if the index vocabulary stays constant, more likely matches a query and therefore can be found in the document collection. Specificity on the other side is a semantic value describing an index term itself (e.g. iPad is more specific than Tablet PC). The problem here is, that although very specific terms can be assigned, and documents are described exhaustive, a term which is used very frequent for describing documents is not useful in finding proper documents, because out of the set of retrieved ones the relevant ones become less, precision drops. Hence some terms are more vital than others for the description of a document.

For receiving documents from an IR system by performing a search operation, a proper weighting of the query terms, and hence ranking of documents matching the query has to be implemented. Usually the weights assigned to terms of a query for a document are summed up to create a ranking of the retrieved documents. So the relevance of a document d according to a query q is mostly expressed as the sum of weights for query terms.

$$Relevance_{d,q} = \sum_{t \in q} w_{t,d}$$

For calculating the weights w for a term t in a document d different approaches exist. One approach is to just weight query terms based on the number of occurrence in a document. This approach is based on term frequency tf where the weight w a term t gets assigned in a document d is the sum of the term occurrence in this document [45, Ch.6].

$$w_{t,d} = tf_{t,d}$$

As this usually is a good approach when documents are of the same size, it is a problem if the document length differ widely. In a long document it is more likely that a term occurs more often than in a short one, therefore normalizing the term frequency is common by just dividing the term frequency by the total amount of terms in a document [46].

$$w_{t,d} = \frac{tf_{t,d}}{\sum_{k=1}^{n} tf_{k,d}}$$

A shortcoming of this weighting approach is, that each term of a query is seen equally important, although executing queries on collections should receive documents matching the query terms, if frequent and therefore less specific terms match, they should be weighted below non-frequent terms which are hence more specific. Redefining exhaustivity as the number of terms a document description contains and specificity the number of documents containing the term, a statistical interpretation of the two properties can be made [35]. Calculating the weight now based on the collection frequency, instead of the term frequency, tend to distinguish documents.

Therefore the inverse document frequency idf for a term t is defined as the logarithm of the quotient of the number of documents in the collection N divided by the document frequency df of term t, which is the number of documents containing the term t [54].

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

Now the inverse document frequency and the normalized term frequency defined, these two approaches combined are a very good representation for weighting terms, by weighting rarely used terms more than common terms through idf and also including the relevance in the document through tf. The tf-idf value is calculated by simple multiplying these two values [45, Ch. 6].

$$w_{t,d} = \frac{tf_{t,d}}{\sum_{k=1}^{n} tf_{k,d}} \cdot log \frac{N}{df_t}$$

Beside the weighting of search terms, the purpose of an IR systems is to retrieve relevant documents, and this can be achieved by various methods. Singhal [60] gives an overview of the most commonly used methods nowadays and we introduce them shortly.

The first and maybe simplest one is a boolean model. It uses the search terms connected by boolean operators to search on the indexed documents for matches and return the found documents accordingly. A more sophisticated approach is the vector space model, which was defined by Salton already in 1975 [58]. The idea behind this approach is, representing a document by a vector created out of the terms contained in the document. The dimension of the vector is of the size of the amount of distinct terms in the document collection. A query is also represented as a vector, and the best matches between the query vector and document vectors are returned as results. Another type of systems are based on probabilistic models defined by Robertson [53]. The key idea is to find probable relevant documents based on a query.

3.3 Clustering

Clustering is a form of unsupervised learning, with the aim to assign items, in this case documents, to groups based on their data. The groups created through clustering should have a high intra group association and a low inter group similarity [45, Ch.16]. In contrast to classification, a form of supervised learning, in clustering the classes are not available prior to the clustering process. Beside creating clusters and assigning items to them, it is also an approach to unveil hidden patterns in large data [50]. Hence clustering can be used in different ways in the context of IR:

Clustering in IR can be used in different ways [45, Ch.16]

1. Clustering documents on the terms they contain, to retrieve more efficiently the results for a query, or better present the results of a query

2. Clustering on the co-occuring citations of documents to give insights into the subarea and may unveil community patterns

3. Clustering terms based on the documents they co-occur, to create thesaurus out of it

While clustering methods differ, most commonly the researcher has the option to choose the measure of similarity on his or her own. The measures have in common that they try to express the degree of association between documents, but differ in their kind of calculation and field suitable for. Hence a clustering method can produce different clusters based on the chosen similarity measure [50].

Similarity Measures

We now introduce some common similarity measures, or also called distance measures. The distance D between a document x and y is calculated by comparing the different term vectors x_i respectively y_i of the documents to each other [50]. Before listening the different distance measures I sum up some basic properties of the measures [28, Ch. 4].

Some basic properties of distance measures

- The distance is always positive. ($D(x, y) > 0$)
- The distance between a document x and itself is zero. ($D(x, x) = 0$)
- The distance calculated between document x and document y results in the same value as calculating the distance between document y and document x. ($D(x, y) = D(y, x)$)
- The distance between document x and document z is always lower or equal the distance from document x to document y and document y to document z. ($D(x, z) \leq D(x, y) + D(y, z)$)

These basic properties of distance measures defined, we now list some of the typical approaches that are used.

Euclidean Distance

$$D(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

This measure is intuitive, if the documents are visualized as points in the vector space, then the euclidean distance calculates the shortest path between these two points.

Cosine coefficient

$$D(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

This measure can also be interpreted geometrically, as it is the angle between the two vectors in the vector space.

Clustering Methods

Now clustering introduced generally and also some similarity measures presented, we take a look at the different clustering methods. Basically clustering methods can be divided into two big groups, hierarchical and nonhierarchical, also called partitioning, methods [50]. The reason why these methods are called like this, is that the methods determine the produced cluster structure. Below each group is introduced.

Partitioning methods create k clusters in N documents with no overlap, this means every document can just be assigned to one cluster. These methods have in common that they are heuristic, which means some parameters, like the amount of clusters, has to be known in advance. The general process for nonhierarchical methods consists of 4 steps

1. select randomly the first representer of the k clusters
2. assign all documents to their closest cluster
3. recalculate the centroid of each cluster
4. repeat step 2 - 3 until no relocation of items is performed any more

On the other side, hierarchical methods create clusters, which are linked together this can either be done top-down (divisive) by starting with all data in one cluster and splitting it up, or from bottom-up (agglomerative), starting from unclustered data, joining one after another until all data is in the same cluster. The generic agglomerative method consists of 2 steps:

1. combine the two closest points (treating single items and clusters as points)
2. repeat step 1 until only one cluster remains

We already defined similarity measures for the distance between two documents and how they can be calculated, also the distance between clusters has to be computed. This value is often called linkage and can be calculated in various ways [28, Ch. 4].

Single link

The distance between two clusters is calculated on the closest points of unconnected clusters. Therefore no recalculation of centroids is required during the clustering process. (see Figure 3.1)

Complete link

This calculation is based on the distance of the least similar pair of the clusters, and the shortest one is selected for performing the next merge. (see Figure 3.2)

Group average link

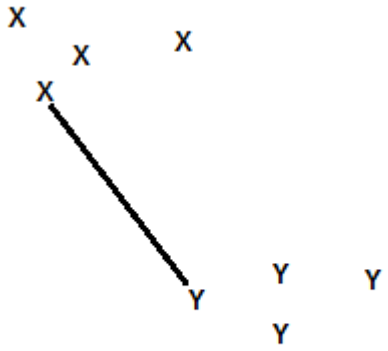


Figure 3.1: Single link distance between cluster X and Y [28, S. 179]

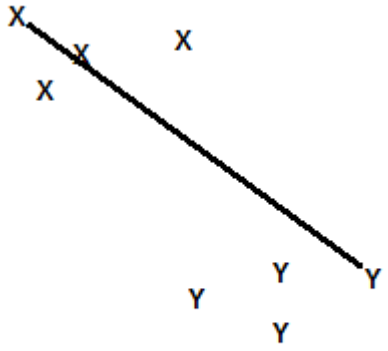


Figure 3.2: Complete link distance between cluster X and Y [28, S. 180]

This distance measure calculates the distance of all pairwise points in a cluster compared to another cluster and then builds the average of it, joining the clusters with the lowest distance. (see Figure 3.3)

Centroid

For the merge decision, the distance between the centroids of each cluster is calculated and used for decision. (see Figure 3.4)

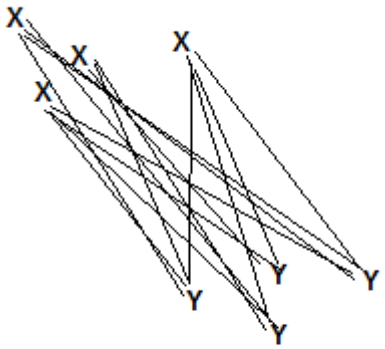


Figure 3.3: Average group link distance between cluster X and Y [28, S. 181]

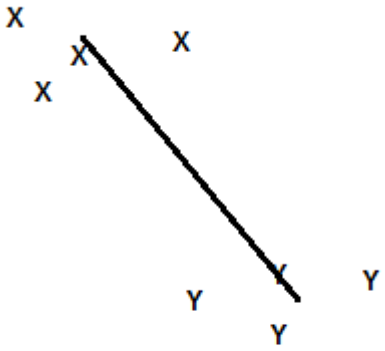


Figure 3.4: Centroid link distance between cluster X and Y [28, S. 181]

After giving an overview of the common methods, I would like to explain k-means, which is a representative for partitioning methods.

The k-means method typically represents its clusters by their centroids, which is the weighted average of all points of a cluster. The method for calculating similarity between elements is the Euclidean distance measure. The steps performed in the clustering process are:

The k-means basic steps [28, S. 172]

1. k randomly selected seeds, which form the initial clusters and centroids
2. each document is assigned to the cluster where the similarity measure, in this case the Euclidean distance, is minimal
3. the centroids of each cluster are calculated based on the new cluster assignments
4. the process is repeated until a termination condition is reached

As stated in Step 4, a termination condition must be reached to stop the k-Means algorithm. For this condition different possibilities exists.

Typical termination conditions for k-means [45, S. 361]

- the process runs a predefined amount of iterations I
- there is no change in the cluster assignments of the objects anymore
- the centroids do not change anymore
- the average Euclidean distance is below a threshold, and hence convergence is reached

K-means has some drawbacks, which are important to be considered. One major drawback is that the created clusters depend strongly on the initial selected seeds, to overcome this problem a iterative approach can be performed, by first selecting random seeds, and repeating the process with different seeds to compare the results. Also being careful on selecting no outliers as starting seeds can reduce this problem. Additionally to this, k-means produces a local optimum which could be far from the global. The size of the clusters is not checked with k-means which can result in a fairly uneven distribution of elements. And as already stated above the k has to be selected by the user, which is not trivial. If possible the selection should be founded on some prior knowledge about the data to create a proper representation of the data [3] [28, S. 177].

For performing clustering on heterogeneous data in the mean of data ranges, it is important that the attribute values are converted to a similar scale. This is vital because otherwise the calculation would be much more influenced by higher values than by lower ones. On the other side this gives the opportunity to weight some properties different, if they are more vital for the cluster representation they can be scaled. Typically the values are scaled to a range between zero and one. One approach of scaling is to divide each attribute by the mean value of the attribute, or dividing by the difference between the largest and smallest value or standardizing can also be performed. Standardizing is achieved by subtracting the mean of the attribute and dividing by the standard deviation. This transforms each numerical attribute to a value between -1 and +1 [28].

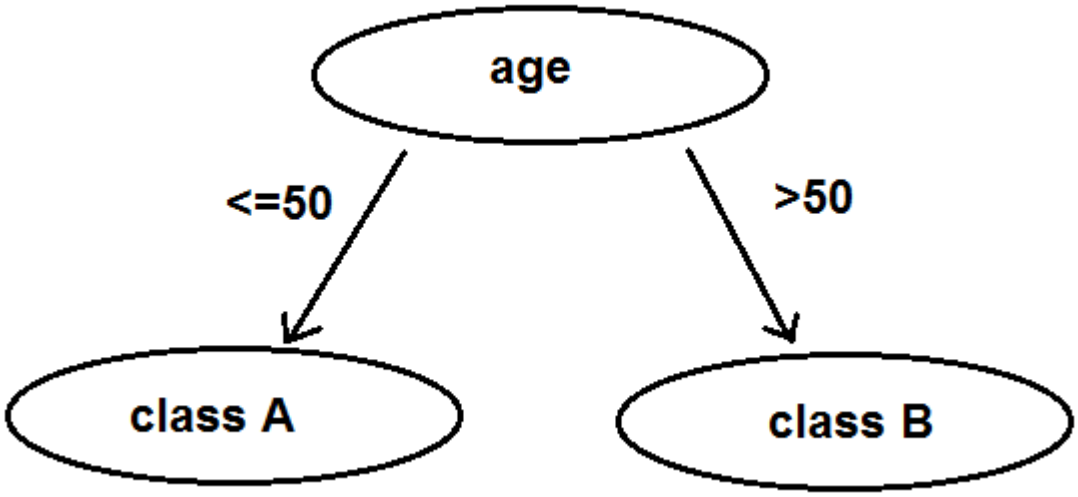


Figure 3.5: Example of a decision in a decision tree with an internal node and two leaves

3.4 Classification

As classification is also used in a small part of Chapter 4 we give a short introduction here. Classification is a form of supervised learning. A classification system is trained with data which already is classified, and out of this data a model is created which then can assign newly added objects to proper classes based on their property values. The basic concept of classification is, to be able to classify objects based on their property values, and hence be able to make predictions. The properties which are used for deciding to which class an instance belongs are the independent properties or attributes, the class itself is the dependent attribute. It is important to be aware of that a model can never be 100% accurate but this is also not the aim to achieve [28, Ch. 3].

Decision Tree

A typical classification method is a decision tree. The tree consists of internal nodes and leaves, whereas the internal nodes represent a decision based on a property value and the leaves represent the class an instance gets assigned (see Figure 3.5). The tree is created during the induction phase with the training data, and then the model can be applied to any data instances for classification.

We now explain the tree induction process a little bit more in detail. The induction process is basically a top-down greedy algorithm trying to create leaves as homogeneous as possible [28].

Basic steps of the induction process [28, S. 120]

1. The training data T is used to represent a single tree node.

2. If all instances of the training data are classified with the same class, stop the process.
3. Split the node by selecting an independent attribute that best divides the remaining training set, and create a decision node.
4. Split the training data according to the selected attribute from 3.
5. Stop the induction process if one of the criteria is met, otherwise continue with 3.
    a) If the divided training data is split, so that the subsets belong only to one class.
    b) If no attribute is left for further dividing.

The major step is Step 3, selecting the next attribute for splitting. The basic concept behind this selection is to select the attribute which best divides the remaining data into homogeneous groups of elements from the same class.

CHAPTER 4

Realization

So far we have given an introduction to the topic in Chapter 1. In Chapter 2 current works in the different fields were presented, to provide background information as well as sources for assumptions made during this work. Chapter 3 contains the toolkit used during the realization process.

This chapter contains the practical realization of this work. It provides a Web Science view on computer science bibliography data, by first introducing the used datasources, then explaining the techniques used for visualization and afterwards presenting the performed analyses.

4.1 Data Sources

This section deals with the acquisition of the data used during the analysis process. We first describe each datasource on its own, with its properties, advantages and disadvantages, subsequently we describe how the data was gained and in which way it was processed and used during the analyses.

The ACM Digital Library

As already mentioned in Section 1.2 the ACM Digital Library[1] (ACM DL) provides access to scientific papers and material from the Association for Computing Machinery (ACM). As the name suggests, the main focus is on the scientific computing society, which is also the focus of this work. ACM, which was established 1947, sponsors over 150 conferences in fields like, management of data (SIGMOD); data communications (SIGCOMM); knowledge discovery and data mining (KDD), to name just a few, which are all recognized as high impact conferences [8]. Beside conferences, also over 40 different publications are produced by ACM, e.g. Communications of the ACM. Additionally ACM also bestows different awards among others the A. M. Turing Award.

[1] `http://dl.acm.org` (accessed 06-October-2011)

Code	Title
A.	General Literature
B.	Hardware
C.	Computer Systems Organization
D.	Software
E.	Data
F.	Theory of Computation
G.	Mathematics of Computing
H.	Information Systems
I.	Computing Methodologies
J.	Computer Applications
K.	Computing Milieux

Table 4.1: ACM Computing Classification System Top Level [33]

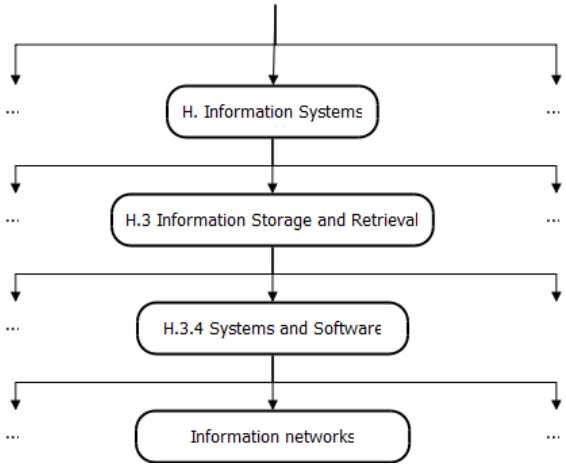


Figure 4.1: Example encoding for ACM CCS classification 'H.3.4.' with subject 'Information networks'

The ACM also proposes an own taxonomy of the computing field, the ACM Computing Classification System (ACM CCS), for categorization of publications. This classification was first published 1964, which was completely revised 1982, since then modifications were published 1983, 1987, 1991 and the last one 1998. The ACM CCS is organized as a four-level tree, where three levels are coded and the last level is made of subjects. The top level consists of the letters A to K (see Table 4.1 for details) an example of such an encoding is presented in Figure 4.1. The purpose of this system is, that each author publishing at ACM has to categorize his or her work, and hence provide colleagues an easier way to find related work and also give some basic information about the work by classification. For example the work from Gulli [27] is indexed with H.3.3, where H. means Information Systems, H.3 Information Storage and Retrieval and H.3.3 Information Search and Retrieval.

The classification of publications in the ACM DL are an advantage, because thus the articles and papers are thematically categorized which is basically nothing more than inherent semantic information. This information is also used for providing different possibilities to browse the publications. The information can either be searched with a search input field, where the best matches for a query are returned. Other options to browse the data are by journals, conferences, publishers or the previously mentioned CCS structure. Hence publications classified with a specific term can easily be found.

For accessing all information in the DL it is necessary to have a subscription. The search function as well as citation information, in various formats for example BibTeX, is public accessible. If available following additional information about publications can be accessed:

- title: the title of the publication
- author(s): information about the author(s)
  - name: the name of the author
  - affiliation history: a list of institutions he or she worked for
  - publication years: the span of years when publication have been made by this author (e.g. 1980-1997)
  - citation count: the amount of citations in the datasource
  - list of publications: all the publication he or she authored or co-authored in the datasource
  - publication count: the number of publications
  - institution: information about the institution he or she works for
    - name: the name of the institution
    - authors: a list of authors from this institution
    - citation count: the amount of citations of the publications made from scientists at this institution
    - publication count: the number of publications made under this institution
    - list of publications: a list of publications assigned to this institution
    - tag cloud: over the publication's assigned to this institute
- type: the type of the publication (e.g. article, inproceedings)
- source: the conference/journal where the work was published
- year: the year of the publication
- abstract: the abstract of the work
- references: the list of references included in the work

- citations: a list of publication which cite this one
- index terms: the ACM CCS classification
- table of contents: the table of contents of this work

Looking at the amount of data, the DL contains bibliographic citation information from 1,720,329 and fulltext of 317,740 (as of October 8th 2011) publications [34].

The DBLP Computer Science Bibliography

DBLP[2] started in the beginning, 1993, as a test of Web technologies, by hosting a HTTP server and some content at the University of Trier.[3] Michael Ley who set this system up added as content some tables of contents of proceedings and journals from the database system and logic programming subarea and hence this was the reason for calling it DBLP 'Data Base systems and Logic Programming'. For demonstration purpose and to improve the searchability hyperlinks to author pages were created, which contained the publications of this person, and the names of the co-authors as well as the source of the publication. This very simple setup seemed to have value for other scientists, in sense of providing a source where the publications of an author can be found, and was the beginning of DBLP [44]. In this early phase the resources were limited, 1997 some funding was provided and a project from the ACM SIGMOD was started to add the historical publications of the database subarea to the DBLP. From this time on recognition of DBLP grew and more conference organizers wanted their publications listed in DBLP.

As DBLP is a service run by a university they provide deeper insights into the structure of the data and the system behind than a private institution. The data of DBLP was in the beginning just stored in the HTML documents but then was shifted to a XML document. Up to now, there is no relational database system behind DBLP, all the bibliographic records are stored in a simple XML file which can be downloaded for free.[4] The XML file is build on publication basis, each publication is an entry in the XML file, identified by a unique key. Out of these publication records the author pages are generated, which contain the publications of an author ordered by publication year. Here you can see the structure of the XML file.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE dblp SYSTEM "dblp.dtd">
<dblp>
        <article>
                ...
</dblp>
```

As you can see the XML file is based on the DTD 'dblp.dtd', this file defines the structure. Basically after the root element 'dblp', different types of publication can follow. These types are

[2] http://dblp.uni-trier.de/ (accessed 06-October-2011)
[3] http://www.uni-trier.de/ (accessed 06-October-2011)
[4] http://dblp.uni-trier.de/xml/ (accessed 06-October-2011)

'article', 'inproceedings', 'proceedings', 'book', 'incollection', 'phdthesis', 'mastersthesis' and 'www'. The types itself are defined very simple, they can contain elements of the entity type 'field'. The entity 'field' hence can contain the information starting from 'author', 'year', 'title' up to 'url', 'isbn' and so forth. The only mandatory part of each element is the attribute 'key'. See below the DTD for details.

```
<!ELEMENT dblp
        (article
        |inproceedings
        |proceedings
        |book
        |incollection
        |phdthesis
        |mastersthesis
        |www)*>

<!ENTITY
        % field
                "author|editor|title|booktitle
                |pages|year|address|journal
                |volume|number|month|url|ee
                |cdrom|cite|publisher|note
                |crossref|isbn|series|school|chapter">

<!ELEMENT article (%field;)*>
<!ATTLIST article
        key CDATA #REQUIRED
        reviewid CDATA #IMPLIED
        rating CDATA #IMPLIED
        mdate CDATA #IMPLIED
>
...
```

Out of this data, HTML pages are created and represented to users. As stated above the data is free for non commercial use. The information can either be downloaded as a whole, or also just the citation information in form of BibTeX for individual publications.

The amount of data contained in DBLP as of October 8th 2011 was citation information about 1,765,547 publications and 1,015,013 distinct names of authors.

Microsoft Academic Search

The Microsoft Academic Search[5] (MAS) is a search engine for academic material created by Microsoft Research[6]. Beside offering the search service it is used for testing data mining techniques and different forms of data visualization. As the indexing process is automated, the publications are not limited to the Computer Science domain, as of October 8th 2011 the domains Agriculture Science, Arts & Humanities, Biology, Chemistry, Computer Science, Multidisciplinary, Economics & Business, Engineering, Environmental Sciences, Geosciences, Material Science, Mathematics, Medicine, Physics, Social Science and Space Science are covered.

As it is a product created by a company not a lot of information about the techniques behind the data acquisition process are provided nor the structure of the data. Basically it can be assumed that the typical approach of a search service is followed, by crawling the Web and indexing scientifc publications an index for scholar purpose like Google Scholar[7] is created. In contrast to Google Scholar, MAS does not just provide title information about publications, it also creates profiles of scientists and institutions. Beside the possibility to search for publications, the intention is also to provide an interface where users can explore conferences, authors, organizations and so on [15].

As already mentioned above, MAS provides the possibility to explore the data in different ways. The most simple way is the well known search input field, where you can either enter a name of an author, an institution or search for a publication. Proper results are then presented and ordered according to a ranking, which is based on relevance for the query and global importance of the results in the MAS [15]. Beside this search option you can also browse the structure of the data on domain basis. For the Computer Science domain for example you can either navigate through selecting authors, publications, conferences, journals, keywords, organizations or subdomains and get publications for the selection presented (see Figure 4.2).

For visualizing the data five additional options are available for users. The Academic Map[8], the Call for Papers Calendar[9], the Domain Trend visualizer[10], the Organization Comparison chart[11], and the Visual Explorer[12] for analyzing co-author and citation graphs.

The purpose of the Academic Map is to display institutions filtered on specific domain basis on a geographical map. Users then can select institutions and get an overview of the authors assigned to this institution displayed.

Call for Papers Calendar displays conferences for different domains in a timeline or also on a geographical map with the information about the deadline for the paper submission and duration of the conference as well as the venue. The list can be filtered on domain and region basis

[5] `http://academic.research.microsoft.com/` (accessed 06-October-2011)
[6] `http://research.microsoft.com/` (accessed 06-October-2011)
[7] `http://scholar.google.com` (accessed 06-October-2011)
[8] `http://academic.research.microsoft.com/AcademicMap` (accessed 06-October-2011)
[9] `http://academic.research.microsoft.com/CFP` (accessed 06-October-2011)
[10] `http://academic.research.microsoft.com/DomainTrend` (accessed 06-October-2011)
[11] `http://academic.research.microsoft.com/Comparison` (accessed 06-October-2011)
[12] `http://academic.research.microsoft.com/VisualExplorer` (accessed 06-October-2011)

Figure 4.2: Navigation through Computer Science domain on MAS [16]

and should provide scientists an overview of locations and conferences were they may want to submit their papers.
The Domain Trend is limited to the Computer Science subfield. It provides an overview of the evolution over time in the Computer Science subfields, based on publication categorization, and displays top authors, based on the calculations and observations of MAS, in the proper fields. The time span to be evaluated for selecting top-authors can be modified by the user.
The Organization Comparison chart allows you to search for two institutions and compare their data available in MAS. The evolution of publications or citations can be viewed along with a tag-cloud about keywords and a list of authors.
The Visual Explorer is a tool for dynamically search through the co-author graph. Beside the co-author graph also the geodesic between two authors and the citation graph can be visualized with this tool.

Additionally to the typical research performance indicators, publication count and citation count, MAS also calculates the H-Index and the G-Index for authors, as well as the H-Index for institutions. The H-Index was proposed by Jorge Hirsch, and should measure the impact of publications of an author or institution by expressing following. 'A scientist has index h if h of his or her N_p papers have at least h citations each and the other $N_p - h$ papers have $\leq h$ citations each.' [32] As improvement of the H-Index the G-Index was proposed by Leo Egghe which measures the scientific performance of a set of articles, basically the articles of an author. It is defined as: 'If a set of articles is ranked in decreasing order of the number of citations that they received, the g-index is the (unique) largest number such that the top g articles received (together) at least g^2 citations' [19].

Now we take a closer look at which information is accessible, if you search for a publication. Beside the common exportation option to BibTeX following information is available if found:

- title: the title of the publication
- author(s): the name of the author(s)
- abstract: the abstract of the publication
- conference: the name of the conference if the publication was published at one
- fulltext: links to fulltext if freely accessible
- source: the conference/journal where the work was published
- year: the year of the publication
- references: the list of references included in the work
- citations: a list of publications which cite this one

As we can see this information is very similar to that of the ACM DL, but looking at the information about authors which is provided we can see the difference. About authors following information is provided:

- name: the name of the author
- institution: the name of the institution he or she is assigned to
- publication years: the span of years when publication were made (e.g. 1980-1997)
- citation count: the amount of citations in the datasource
- list of publications: all the publication he or she authored or co-authored in the datasource
- publication count: the number of publications
- G-Index: the calculated G-Index
- H-Index: the calculated H-Index
- interests: a list of 3 topics he or she is interested based on the classification of his or her work
- co-authors: list of co-authors
- conferences: a list of conferences he or she has published
- keywords: a tag cloud of keywords found in his or her works

Comparing this information about an author to the information available in the ACM DL we can see that it is more complete. Additionally the information about institutions consists of:

- name: the name of the institution

- authors: a list of authors from this institution
- citation count: the amount of citations of the publications made from scientists at this institution
- publication count: the number of publications made under this institution
- H-Index: the calculated H-index of the institution
- top-areas: a list of 5 areas the institutions is famous for based on the level of authors publishing for the institution
- geoinformation: the latitude and longitude coordinates of the institution
- keywords: a tag cloud of the keywords from this institution

In addition to the more complete information about the author also the information about research institutions is more complete.

Looking at the amount of data MAS covers we can see also the difference between an automatically created datasource and a manually maintained one. As of October 8th 2011 MAS consists of 36,672,635 publications and 18,868,730 authors indexed ranging over all domains. For the Computer Science domain, in which we are interested 2,914,432 publications, 1,467,364 authors and 9,007 institutions were indexed.

Additional Datasets for Analyses

In addition to the bibliography information from the various sources stated above, for some of the analyses data from other sources was used. For example, we needed the assignment of coordinates expressed by latitude and longitude, to countries, as well as country assignments to continents. For these purpose other sources were connected with the bibliography data. They are listed below with a short description of their characteristics and use case in this work.

- GeoNames [63]

  GeoNames is a geographical database licensed under creative commons attributions licence, which allows a free of charge usage if a reference to GeoNames is provided when the data is used. It was founded by Marc Wick, and contains geographical names as well as their positions. For example searching for 'Vienna University of Technology' returns the dataset of the university with the proper latitude and longitude values 48.1989 and 16.37. In addition for searching positions through a web interface, Webservices[13] as well as a daily database dump[14], are provided to access the data and hence the process of searching for institution location could be automated or built into an own application.

[13] `http://www.geonames.org/export/#ws` (accessed 14-September-2011)
[14] `http://www.geonames.org/export/#dump` (accessed 14-September-2011)

We used for our purpose once the CountryCode[15] Webservice which takes as arguments the latitude and longitude position of a place, and returns the ISO-3166 alpha 2 countrycode [22] of this position. This was necessary to assign institutions to countries, as this information was not provided on the MAS website. Additionaly to the Webservice ContryCode we also used the information of the countries provided[16]. This information contained the ISO countrycode as well as the fullname, population, area, capital and the continent of the country. This data was used for further analyses on relations between data from the different continents.

- Webometrics Ranking of World Universities [18]

  The Webometrics Ranking of World Universities is provided by the Cybermetrics Lab of the Consejo Superior de Investigaciones Cientificas[17] which is a public research organization in Spain. In a first approach the aim of this ranking was to encourage Web publication, as the ranking is based on the visibility of an institution on the Web. Hence the open access to scientific material should be provided by institutions. The ranking is based on the Web domain of an institution analyzing the visibility, as inlinks from other sites, the size of the Web presence calculated through the amount of web pages, the amount of rich files which are public accessible scientific material and scholar information about publications found in bibliography databases. Since 2004 twice a year rankings for tertiary educational institutions are provided, and the data of the top 500 universities can be downloaded[18] and freely used if the source is cited.

  For our analyses we used the provided data of the top 500 universities as of July 2011 to verify our results of evaluated top universities. The data is provided for download [12]. The data contains beside the list of the top 500 universities with name, country and region a summation of the institutions per country and region in the top 100, 200, 500 and total data.

Creating the Dataset for our Analyses

In Figure 4.3 we can see a graphic of the datasources with the amount of information in each of them. To give a better overview, the characteristics of the datasources are summed up in Table 4.2.

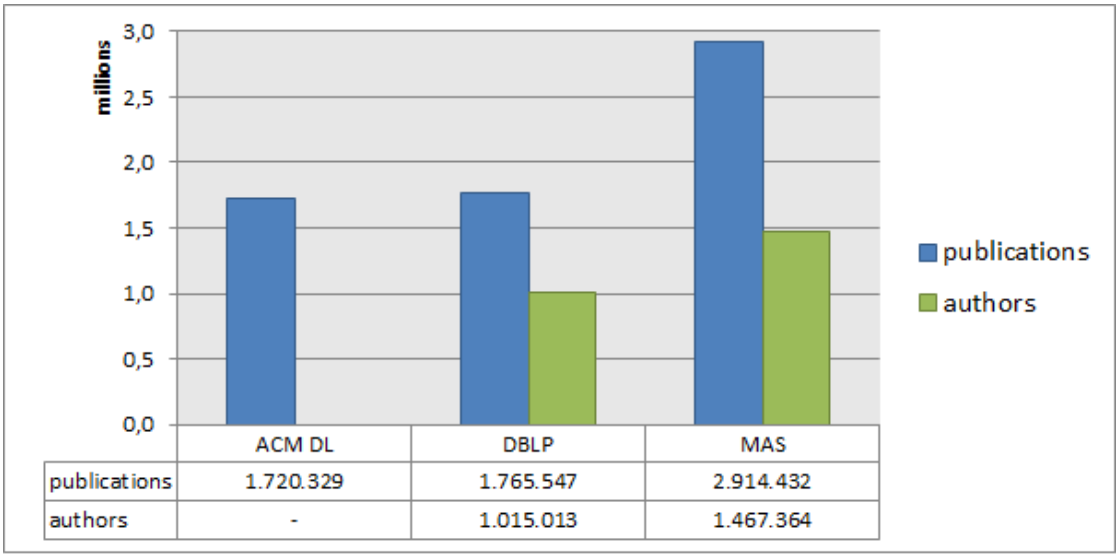
As stated in Section 1.3 we need classification of the publications to answer our questions like 'Is there a cultural influence on research in computer science?' because we have to distinguish between the different subdomains of the computer science field to answer this question. Hence the first approach was to gain a dataset from the ACM DL, as it already provides classification information, author information and institutions assignment and enrich this data with

[15] `http://www.geonames.org/export/web-services.html#countrycode` (accessed 14-September-2011)

[16] `http://www.geonames.org/countries/` (accessed 04-October-2011)

[17] `http://www.csic.es` (accessed 08-October-2011)

[18] `http://www.webometrics.info/premierleague.html` (accessed 08-October-2011)



	ACM DL	DBLP	MAS
publications	1.720.329	1.765.547	2.914.432
authors	-	1.015.013	1.467.364

Figure 4.3: Amount of bibliographic information in the Computer Science domain for different datasources (as of October 8th, 2011)

	ACM DL	DBLP	MAS
Downloadable	no	yes	no
BibTeX	yes	yes	yes
Publications	1,7mill	1,8mill	36,6mill*
Automated	no	no	yes
Author Information	yes	no	yes
Author/Institute association	yes	-	yes
Institute Geoinformation	no	-	yes
Classification	yes	no	no

* Amount of publications in all domains

Table 4.2: Characteristics of datasources used for this work (as of October 9th, 2011)

geoinformation about institutions. This enrichment would have been achieved by taking institution names and locating their geographical location by searching on GeoNames.

Unfortunately the data of the ACM DL is not downloadable and it was not possible for us to receive a snapshot of the data. As most of the information is browseable for free, crawling the site and fetching the data as well as extracting the information (scraping) was planned, but this approach was not followed as scraping the dataset of 1,7 million publications with around 10 seconds per publication would last for more than 190 days. Therefore we switched to the next possible option, using the data from DBLP, which is downloadable for free, as our primary source.

So as starting point for creating our dataset we chose to use the data from DBLP. As mentioned above the data from DBLP is downloadable in XML format for free, for our purpose we needed the data in a relational database model which is offered in Hannover by the L3S Research Center[19]. They offer a MySql dump of the data as they also host a faceted search interface as well as a RDF view through the use of a D2R server. The dump contains the same data as the XML file but already converted in a relational database structure which can be processed further. The snapshot from April 23rd 2011[20] was taken which contains 942,346 distinct authors and 1,638,158 publications. We refer to this dataset as $DBLP_{all}$.

With this dataset we got the titles and information about the source, year and type of publications as well as the name of the authors for a publication. In a first step, stop word removal and KStem was performed on the titles of the publications. To do this a Java program was written, using Apache Lucene Core[21]. Apache Lucene Core is a Java library implementing document indexing as well as search functionality. As in the Apache version 3.0.1 of Lucene Core there is an implementation of the Porter stemming algorithm[22] but no implementation for the KStem, we used the certified distribution from Lucid Imagination, LucidWorks 3.0.1[23] which contains beside the Apache Lucene Core version 3.0.1 also an implementation of KStem[24]. The titles were first processed by Lucene StandardAnalyzer[25] and then by the KStem. The StandardAnalyzer tokenized the titles based on words, removing punctuation and hyphens and removed english stop words after this the KStem was applied resulting in normalized titles for the publications. This operation was performed on all of the 1,638,158 publications.

In a next step the goal was to add information about the authors to the database. As we can see in Table 4.2 information about 1.47 million authors can be found on MAS. Beside this fact, we can dedicated search in MAS for authors in contrast to the ACM DL where this can only be done through the keyword search. As mentioned above the number of authors in the DBLP dataset is 942,346 which is practically impossible to be searched hence we took a look at the structure of the authors. As you can see in Figure 4.4 a lot of authors (493,510) just have one publication

[19] `http://dblp.l3s.de` (accessed 04-October-2011)
[20] `http://dblp.l3s.de/dblp++.php` (accessed 05-May-2011)
[21] `http://lucene.apache.org` (accessed 22-May-2011)
[22] `org.apache.lucene.analysis.PorterStemFilter`
[23] `http://www.lucidimagination.com/products/certified/lucene` (accessed 17-May-2011)
[24] `com.lucidimagination.luceneworks.analysis.LucidKStemFilter`
[25] `org.apache.lucene.analysis.standard.StandardAnalyzer`

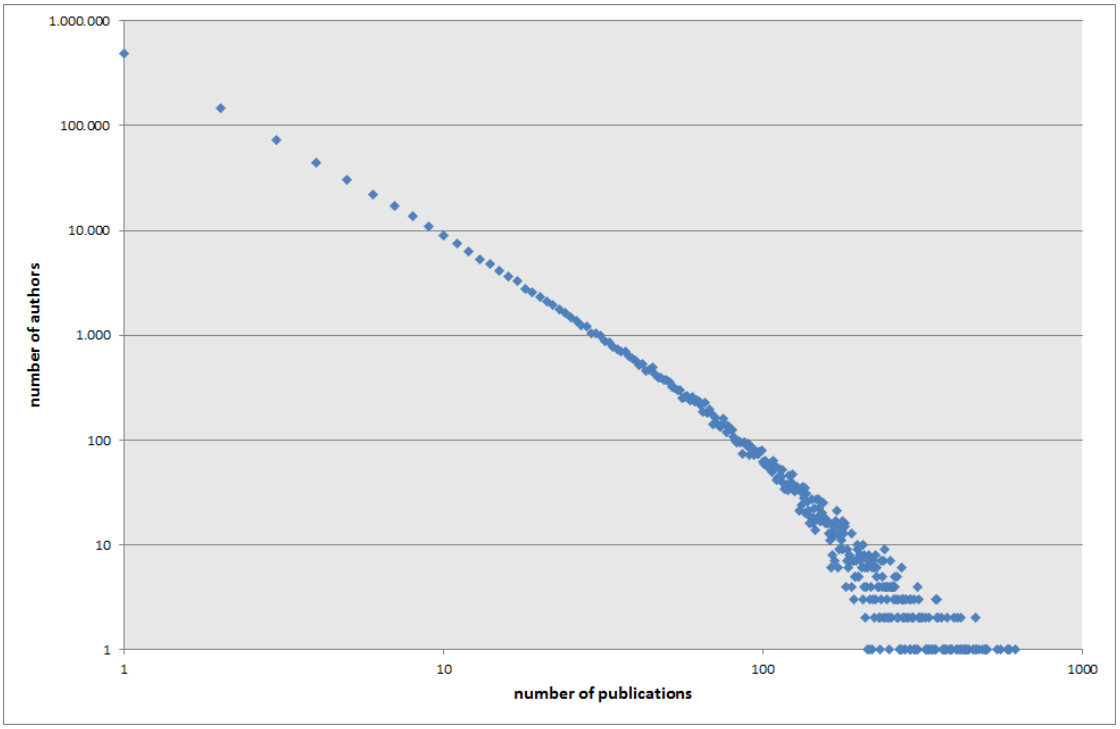


Figure 4.4: Relation between number of authors and number of papers in DBLP (as of April 2011)

and just 2,872 have more than 100 publications. Looking at some statistical values, the median of this distribution is still 1 publication, which means more than half of the authors have just one publication, and just 24.16% have more than 3 publications. For this reason we decided to just search the top-authors on MAS, by interpreting top-authors as authors with more or equal 10 publications. Looking at the DBLP data 88,852 authors have more or equal 10 publications. We refer to this dataset as $DBLP_{top}$.

For fetching the author information from MAS a tool was written in Java, with the help of HtmlUnit version 2.8[26]. This tool searched on MAS for the authors name, and if a match was found the author information was downloaded. The information consisted of the name of the author, an internal id, the URL of the personal website of the author if found, the amount of citations and publications as well as the G-Index and H-Index calculated on base of the data in the MAS and the assigned institute if available (see Figure 4.5).

If an organization was assigned to an author, which is based on the observations of the data at MAS and hence can be error prone, the information about this organization was also fetched and saved. Beside the name of the institution, the amount of publications and citations as well as the calculated H-Index, an internal id and the location information in the form of latitude and longitude values was fetched (see Figure 4.6).

[26] `http://htmlunit.sourceforge.net/` (accessed 02-May-2011)

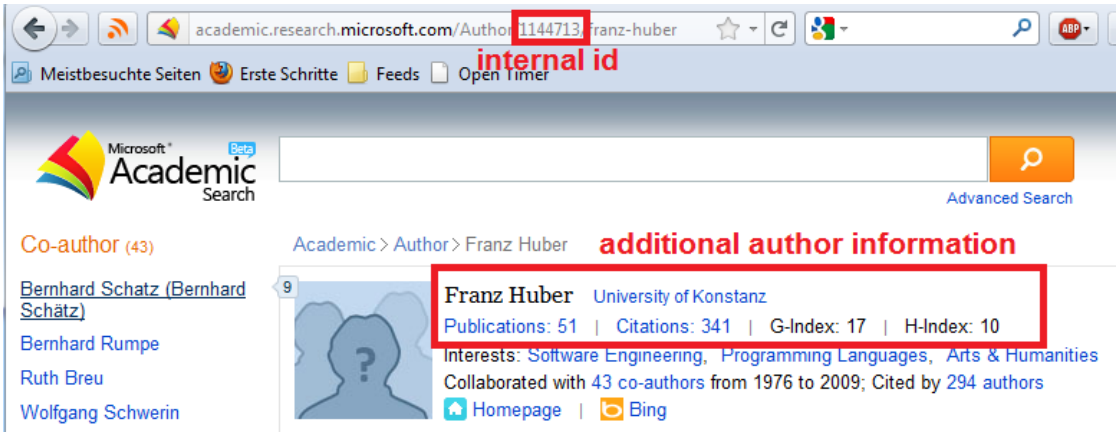


Figure 4.5: Example of author information page on MAS with fetched information highlighted (as of October 10th 2011)

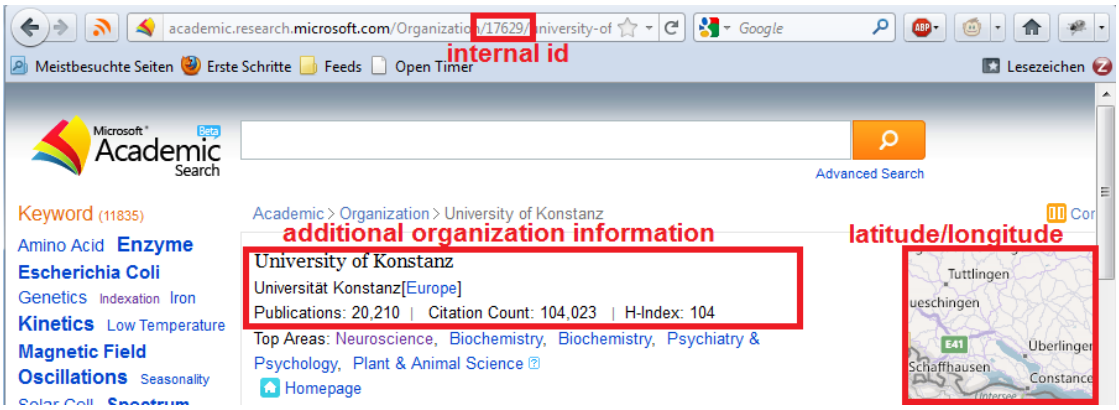


Figure 4.6: Example of organization information page on MAS with fetched information highlighted (as of October 10th 2011)

The searching of authors from the dataset $DBLP_{top}$ resulted in 76,974 authors and 3,427 institutes which were found on MAS. We refer to this dataset as MAS_{all}. Not all institutions found had geolocation assigned, hence also not all authors have a geolocation assigned. In total 3,313 organizations had a latitude and longitude property as well as 62,728 authors were assigned to them. We refer to this dataset as MAS_{geo}.

Looking at our question 'Is there a cultural influence on research in computer science?' we want to answer, we need beside the geolocation of an author, the thematic classification into a subfield of computer science for him or her. This is necessary, because this combined information can be analyzed concerning a relationship between geoinformation and thematic information and hence cultural influences regarding research subfields can be unveiled. Therefore we first analyzed the indexed publication titles and created tag-clouds for some authors to get an

Figure 4.7: Tag cloud of the indexed titles of Tim Berners-Lee's publications in DBLP (as of April 23rd 2011) created with IBM Word Cloud Generator [14]

overview of the data and information contained in the titles. As you can see by looking at the title tag cloud of Tim Berners-Lee in Figure 4.7, some semantic is included, as the terms 'world wide web' seem to appear more frequent in his publication titles than anything else.

This observation was used as basis to create a framework which enriches the dataset with topic information. As described in Subsection 'The ACM Digital Library' it is possible to browse the ACM DL based on the CCS. In addition downloading just the title information of a publication is much faster than downloading the whole information about a publication. So one part of the framework to enrich the data with topic information is, that properly classified title information about publications can be download in an appropriate quantity. Subsequent to this step, the extraction of the thematic information from the publication titles of the authors is done. The final task was to match the topic information of the authors to the downloaded classified publications titles, so that the publications of an author are properly classified through the titles of the ACM DL.

Therefore in a first step we downloaded for the 318 ACM CCS classes, which are leaves in the classification tree, titles of publications. To receive a broader overview of the titles from the ACM DL different sortings of the search results were used. Sorting after relevance, publication date, citation count, download count overall, download count last 6 weeks and download

count last 12 months were used. The titles were fetched, and also indexed with the same method (Apache Lucene Core) as the titles of the DBLP publications. For the 318 CCS classes, 311,801 titles were downloaded. These titles then were stored in a Apache Lucene Directory, this directory is an index created from the Lucene library for providing fulltext search. It can be seen as a datastructure on which searching can be performed, and the Lucene library takes care of creating this directory and performing the search. The user just adds documents to the directory, in this case titles of publications, and then can search for results, getting an ordered list returned, based on the query matching.

In a next step the topic information from publication titles of authors was extracted. For this approach the frequent terms of titles for authors were calculated and stored. As basis for authors which were investigated the dataset MAS_{all} with 76,974 authors was used. As a finding from Section 2.2 a scientist typically has more than one research area, in average 2.2 [8], two approaches were used for calculating the frequent terms. In a first approach all titles of an author were treated as they represent one class, hence the frequent terms over all publications were calculated. For 54,103 authors frequent items were found for a support of 0.2 in their title information, which means that the terms were present in at least 20% of the publications of an author. As an example, the frequent terms in the publication titles of author Anca-Andreea Ivan are *matching semantic service web*. In a second approach the publications of each author were first clustered with k-means and a k of 3. With this preprocessing a distinction of the authors titles regarding his or her research fields were possible. After this the frequent terms of each cluster of publications for each author were calculated. This resulted in 73,707 authors having at least in one of their three clusters frequent terms present. To demonstrate the difference, the frequent terms for author Anca-Andreea Ivan in cluster 0 are *combine semantic matching web service*, in cluster 1 they are *data metric* and in cluster 2 *network value model*. The terms constructed during this two approaches were then used as search terms for the previously constructed Lucene directory. Searching the terms and storing the best match resulted in classification of the author. For the first approach, assuming the author just has one topic over his or her whole life, we could find for 54,098 authors an ACM CCS class, calling this ACM_{one} dataset. To continue our example, Anca-Andreea Ivan's frequent terms over all her publications got classified as, *D.2.12 [Software] [Software Engineering] Interoperability*. The second approach, treating the life of a scientist as three classes, we could assign 73,692 authors to an ACM CCS class, calling this $ACM_{cluster}$ dataset. Anca-Andreea Ivan's frequent terms in cluster 0 got classified also as *D.2.12 [Software] [Software Engineering] Interoperability*, the terms from cluster 1 as *E.m [Data] Miscellaneous*, and cluster 2 as *I.6.5 [Computing Methodologies] [Simulation and Modeling] Model Development*. As this author was just randomly selected, we also want to provide an example which demonstrates the improvement of the results by the clustering approach. For this purpose we chose Karen Sparck Jones, who defined in her paper [35] the inverse document frequency. The calculation for the frequent terms over all of her publications delivered no result. This means it was not possible to find frequent items and hence no classification was possible. Looking at the results of the frequent terms calculation of the clustered publications we can observe the improvement. For cluster 0 we still were not able to locate frequent terms, for cluster 1 they are *document retrieval spoken* and for cluster 2 *information retrieval*. The frequent items for cluster 1 result in a classification as *I.2.7 [Computing Methodologies] [Artificial Intelligence] Natural*

Language Processing and for cluster 2 they yield in *H.3.3. [Information Systems] [Information Storage and Retrieval] Information Search and Retrieval*. We can observe, that these results represent the research area of Karen Sparck Jones very well, demonstrating also the improvements through the clustering approach.

As we can see not all data from MAS (MAS_{all}) contained geographic information (MAS_{geo}), and not all could be assigned to ACM CCS classes (ACM_{one}/$ACM_{cluster}$). The overlap of these datasets, with authors assigned to institutes and furthermore geolocation information, and authors classified results in the base dataset applicable for performing the desired analyses. Looking at the overlap of MAS_{geo} and ACM_{one} 43,465 are thematically classified and have a geolocation assigned. We call this dataset $Base_{one}$. MAS_{geo} and $ACM_{cluster}$ have an overlap resulting in 59,540 authors classified and geographically located, we call this dataset $Base_{cluster}$.

This final datasets then basically contained information about publications, like the publication year, the source, be it a conference or journal, the author's name, some statistics about the author's institution he or she might work for as well as the information about the location of this institution.

For a better overview of the available data, see the entity relationship model in Figure 4.8 which lists the attributes and relationships between the data. Additionally in Figure 4.9 you can see a visualization of the data acquiring process.

4.2 Visualization

In this chapter we explain in short which techniques were used for creating the visualizations later on.

IBM Word Cloud Generator

The IBM Word Cloud Generator [14] was used for creating word clouds. This was mainly a preprocessing work, visualizing the titles of DBLP publications for better overview. IBM Word Cloud Generator is also basis of the famous online word cloud generator Wordle[27] by Jonathan Feinberg. The process of creating a word cloud consisted of three steps:

1. Deciding which word cloud to create
   a) word cloud of an author over all his or her publications or a particular year
   b) word cloud of a conference over all years or a particular year
2. Load the indexed (stop words removed, stemmed) titles of the selected publications
3. Execute the IBM Word Cloug Generator by handing over the proper parameters

[27] `http://www.wordle.net/` (accessed 10-October-2011)

Figure 4.8: ER diagram of created dataset

Google Fusion Tables

The service of Google Fusion Tables[28] is to provide storage for datasets online which hence can be accessed from the Web. The basic layout is similar to a relational database system, where datatables can be created and filled with data. The data can be uploaded via a comma separated file. The tables can either be held private, which means no one can see the data, public, which means the data can be found by search and accessed by anyone, or unlisted which means the tables can't be found by search but accessed by anyone who has the data source id/table id. This id gets uniquely assigned to every fusion table created and identifies exactly one table. The purpose and advantage beside the sharing of the data is that it can be, if geopositions are included in the data, visualized on a Google Map by just one click.

In the visualizations provided, Google Fusion tables was used to store the information online, and hence can be accessed from everywhere.

[28] `http://www.google.com/fusiontables` (accessed 10-September-2011)

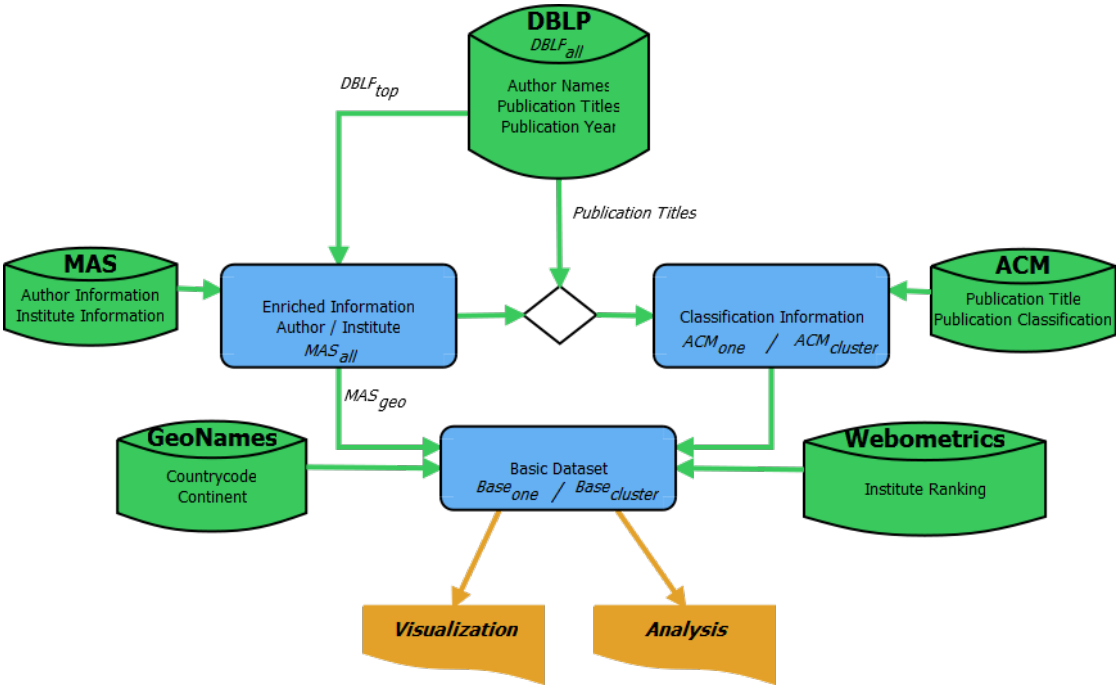


Figure 4.9: Visualization how the basic dataset is created and which sources are included

Google Maps JavaScript API

For some of the visualization the Google Maps JavaScript API V3[29] was used. This was basically done through including the JavaScript into a local HTML file. Through some drop down boxes selections of the datasource could be made and then the data is visualized on the map, by positioning markers on the map. For better overview, a clustering of markers is possible through the usage of MarkerClusterer[30]. This library creates automatically clusters, based on the added markers for better overview, calculating the size and position of them.

Google Chart Tools

For another type of visualizations I used Google Chart Tools[31], especially the GeoChart[32], which offers the opportunity of displaying the values for countries as well as color coding the countries according to their value.

[29] `http://code.google.com/apis/maps/documentation/javascript/` (accessed 01-September-2011)

[30] `http://google-maps-utility-library-v3.googlecode.com/svn/trunk/markerclusterer/` (accessed 01-September-2011)

[31] `http://code.google.com/apis/chart/` (accessed 03-September-2011)

[32] `http://code.google.com/apis/chart/interactive/docs/gallery/geochart.html` (accessed 03-September-2011)

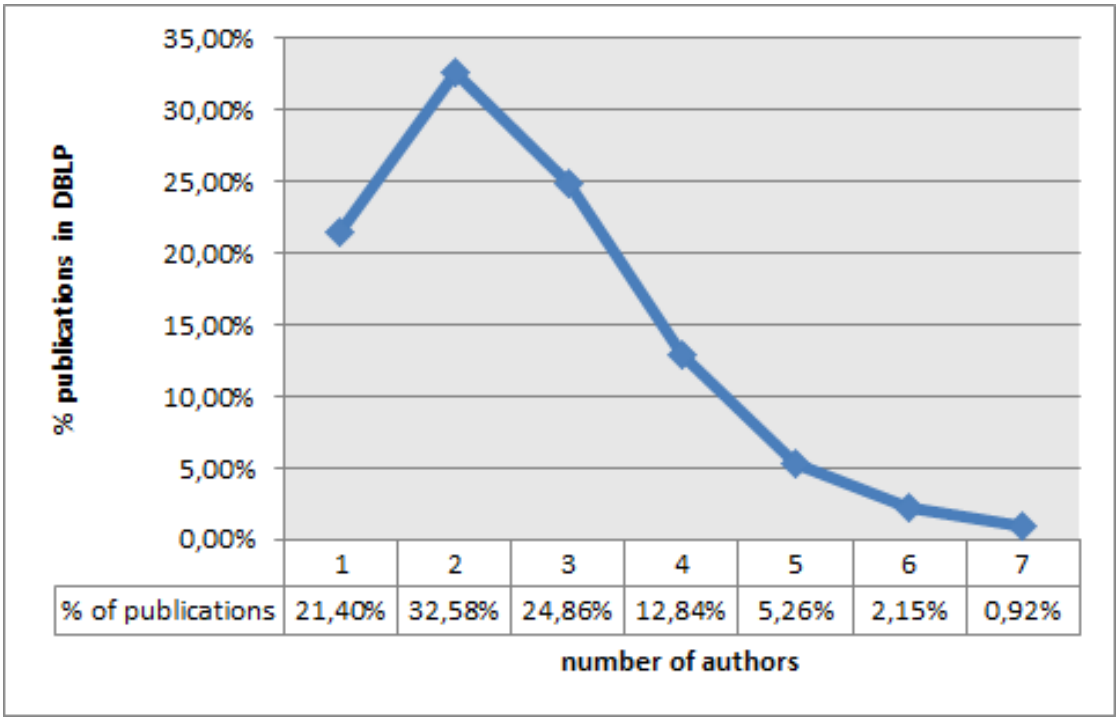


Figure 4.10: Distribution of publications with number of authors in $DBLP_{all}$

4.3 Analyses

In this section we describe the performed analyses on the created dataset. We give first a basic overview of the dataset in 'Analyses of the DBLP data' and then answer the questions 'Can we observe a relationship between the ranking of an institute and the authors of an institute?', 'Are there regional differences concerning the amount of scientists?', 'Is there a cultural influence on research in computer science?' and 'Is there a bias between the venue of a conference and the origin of authors joining the conference?'.

Analyses of the DBLP data

Some basic analyses were performed on the $DBLP_{all}$ dataset. First we analyzed the distribution of the amount of authors per publication in the DBLP. In the dataset $DBLP_{all}$ 21.40% of the publications are written by a single author, 57.43% by 2 or 3 authors and 42.95% by 4 to 6 authors. Figure 4.10 is a visualization of this distribution. Looking at this numbers we can see, that the amount of publications written by a single author is not even half the size as the publications written by two or three authors.

The look at the distribution over the whole timespan in $DBLP_{all}$ gave a first impression of the structure. In a second approach we analyzed also the amount of authors per publication, but we looked at the variations over the course of time. This is done because it shows trends of

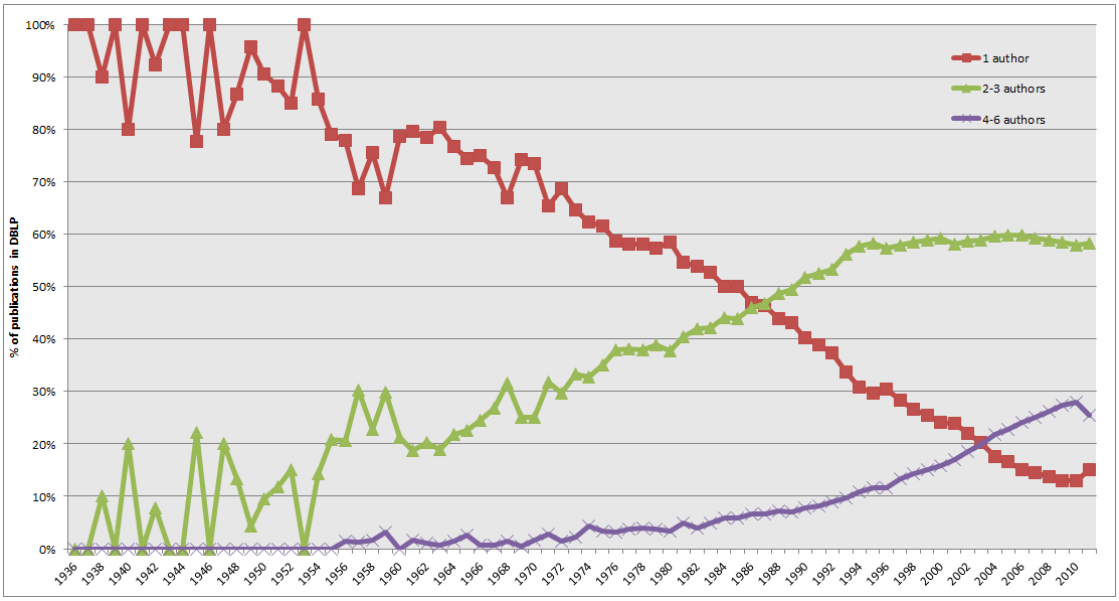


Figure 4.11: Distribution of publications with number of authors per year in $DBLP_{all}$

collaboration. In the first years up to 1970 a heavy fluctuation can be seen, this is caused by the very small amount of just 7,081 (0.4%) publications in $DBLP_{all}$ before 1970. Starting from 1970 on, the amount of publications available increases from year to year. If we look at Figure 4.11 we can see that the graph gets continuous from the 1970s on. This continuous evolution shows a clear trend of a decreasing amount of publications authored by just one person, and a trend to collaboration in authoring publications. The amount of papers published by two or three authors rose until 1993, which then stabilized between 56%-59%. Papers published by more than 3 authors started to rise from 1982 on up to now. The year 2011, which shows a decrease in papers published by more than 3 authors has to be considered with care, as the snapshot was from April 23rd, thus, the data from 2011 is highly incomplete and not quite representative. The numbers and the investigation of the graph support the finding of Elmacioglu [20], who observed for the Data Base community in DBLP an increase in the co-authorship count as well as interdisciplinarity, and shows that the field of computer science nowadays hardly can be investigated by a single person.

In this first subsection we gave an overview over the data. The aim was to provide some basic insights so that we can continue in the next subsections with answering the questions.

Can we observe a relationship between the ranking of an institute and the authors of an institute?

This analysis investigates the relationship between the attributes of authors from an institution and their influence on the performance measure of this institution. The key question is: 'Is there a correlation between the authors of an institution and the performance of this institution?'. To

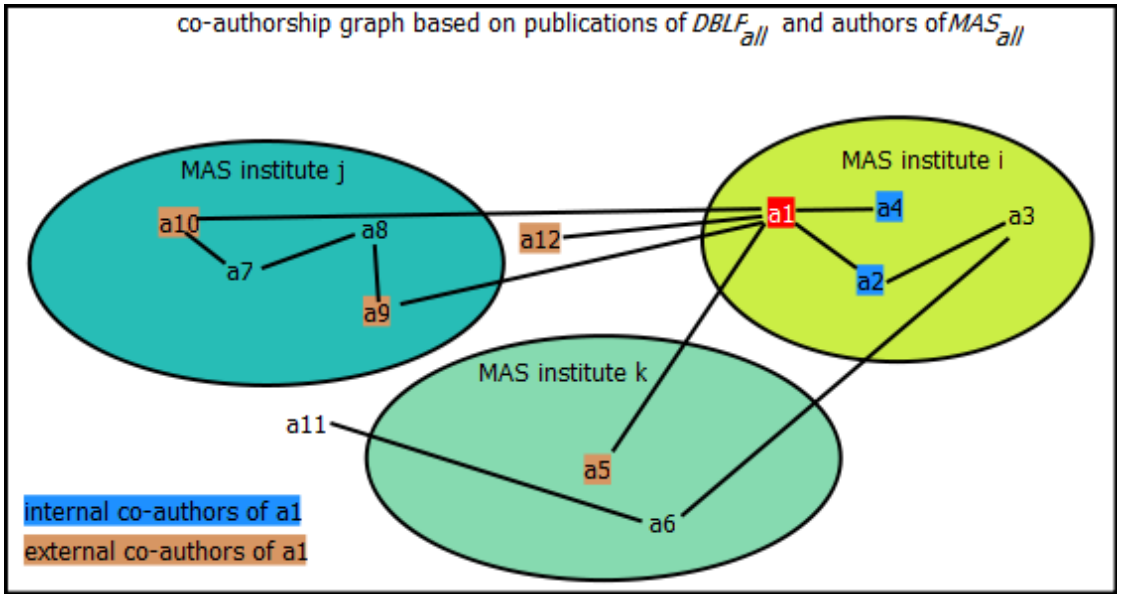


Figure 4.12: Visualization of the concept $internal$ and $external$ co-authors

be able to answer this question, we first have to define how the performance of an institution is expressed. We assume that the H-index of an institution is a proper representation of its performance. Additionally to the institute performance we have to represent the authors of an institute with some values. To express the quality of an author we chose to use his or her co-authors H-index. This means that we assume that the co-authors of an author a can represent the quality of the author's institution i. This assumption is based on the finding in Section 2.2, which states citation information as a proper source for representing the quality of a scientist and hence the H-index, which is calculated through citations, is also a valid measure. The representation of an author by his co-authors is a little bit more sophisticated. Instead of using all co-authors of an author as one group, we split this up into two disjoint groups, an $internal$ co-author group and an $external$ co-author group. These two groups are determined through the co-authorship graph represented by the publications in the DBLP. Co-authors of an author who are found on MAS and assigned to the same organization are part of the $internal$ group. On the other side, co-authors who are not assigned to the same institution are part of the $external$ group. The concept of this principle is visualized in Figure 4.12.

We will now define the parameters which are used for our analysis. Beside the already mentioned H-index of co-authors for different groups the first year of publication of author a is also included. The assumption is that this value adjusts the H-index values by taking the length of a scientists career into account.

- $Hindex_{a,i}$

  is the H-index of organization i to which author a is assigned, observed by MAS and fetched.

- $Hindex_{intern_a}$

  is the average H-index calculated over the co-authors of author a of his or her publications from $DBLP_{all}$ where the co-authors are assigned to the same organization i as author a

- $Hindex_{extern_a}$

  is the average H-index calculated over the co-authors of author a of his or her publications from $DBLP_{all}$ where the co-authors are assigned to a organization j which is different to the organization i from author a

- $firstyear_a$

  the year of the first publication of author a in $DBLP_{all}$

- $firstyear_{normalized_a}$

  is $firstyear$ reduced by 1970 to normalize it

For answering the question we built a regression model based on this defined variables. The model is based on the dependent attribute $Hindex_{a,i}$, which is the H-index of institute i of author a, representing the performance of this institution, and the independent attributes of the author, which are the average H-index of the internal and external co-authors as well as the first year of publication of the author.

$$Hindex_{a,i} = \beta_0 + \beta_1 Hindex_{intern_a} + \beta_2 Hindex_{extern_a} + \beta_3 firstyear_{normalized_a}$$

For testing the model, observations have been created by calculating for all authors from MAS_{all} these values. Out of the 76,974 authors from MAS_{all}, 63,440 values were calculated. This 63,440 observation were used as basis for the regression analysis. In a first attempt we tested the regression model over all observations. The result was a model with a standard deviation for $Hindex_{a,i}$ of more than 67, which is for a value range of $Hindex_{a,i}$ between 2 and 366 too much. In a second approach 250 observations were randomly selected to perform the same test again. This lead to a similar result as above with a too high standard deviation concerning the value range and hence no meaningful model.

Findings from Section 2.2, that each subfield of Computer Science has its own properties, and that the coverage in DBLP for the different areas also differs, suggest that looking at just a particular subarea of computer science provide better results. So limiting the observations on a specific subfield exclude these varying conditions. To perform this test we narrowed the data to 'H.3.3 Information Search and Retrieval' which lead to 1175 observations. We tested the regression model on this narrowed dataset and the outcome was, that the standard deviation even got worse (69). To exclude not only the influence of other subfields, but also the historical evolution and trends within a subfield, a restriction on the publication years was applied. We limited the narrowed dataset again, by looking only at the years 1990 to 1995 in subarea 'H.3.3'. The model tested on this 92 observations was not significant anymore. As these test were performed only

on one subfield, the same procedure was run through with 'F.2.1 Numerical Algorithms and Problems', which lead to the same observations.

An additional finding from section 2.1, that each culture can also have different properties, suggests that a geographical restriction would exclude this influence and hence a significant outcome can be produced. The observations were limited to the data from Austria, leading to 568 values. The model was tested, but the assumption of a meaningful result was not proved. A standard deviation of 22.8 at first glance seemed good, but a reduced value range for $Hindex_{a,i}$ from 1 to 87 relativates this finding. The test was also performed on a dataset for Austria which was restricted by time, with 59 observations from 1990-1995 the model was not significant anymore. To perform a cross check the same procedure was applied to the data from Sweden. A meaningless model for the 619 observations from Sweden and 67 observations from Sweden for 1990-1995 was the result.

Although the two previous modifications lead on their own to no significant outcome, it is suggested that a combination of both of them can achieve this. So we restricted the data in three dimensions, geographically, thematically and in the time dimension and performed a regression model test. The data was for a first test restricted to the United States and the H.3.3 class. These 455 observations resulted in a not significant model. Subsequently the data was restricted to China and the H.3.3 class, 56 observations resulted also in a not significant model anymore. To check if something is explained in Austria for a thematic restriction to 'H.3 Information Storage and Retrieval', 49 observations were tested. The model was significant, but suffers the same drawback as the other significant models, a standard deviation of 20 for a value range of $Hindex_{a,i}$ from 8 to 87 is too high. For Sweden this approach was also tested on H.3, but with only 22 observations the model was not significant anymore.

The outcome of the regression tests on this data is, either without a restriction nor with a restriction be it geographically, thematically or for the time a significant pattern can be observed and thus no answer to the question can be made. One minor finding was that the parameter in the regression model which most likely lost its influence was $Hindex_{extern_a}$. This clearly can be linked to the definition of the H-index, as the H-index x for an institute i is represented by together x publications of all authors a of institution i are cited at least x times. So no external influence on this value.

As with the previous approach, regression analysis, no answer to the question could be found, a different method was applied. The observations were aggregated and calculated for each organization, and not for each author. The dataset was created out of all institutes of MAS_{all}. Beside changing the way how the values are calculated, also the method was changed. Instead of performing a regression analysis, the approach was to cluster the data, and limit the cluster size to three, so that the result of the clustering process are three distinct clusters where one consists of top institutes, one of middle institutes and one of low institutes concerning their values. Subsequent to this step, a classification model is trained on the gained information from the clustering process, so that for institutes not evaluated in this dataset, classification can be carried out.

As already stated above, the dataset was created on organizational level. To take into account the different authors assigned to one institution, the data was not only split into an $internal$ and

$external$ group, but also into age groups of authors within these two sets. The age of authors was determined based on the publication age. Publication age is the year of the last publication in $DBLP_{all}$ subtracted by the year of the first publication in $DBLP_{all}$ of an author a. To create a proper datasource the authors were first divided into $internal$ authors and $external$ co-authors and then grouped into young (publication age ≤ 5), middle (< 5 publication age ≤ 10) and old (publication age > 10) authors. The $internal$ authors are all authors a from MAS_{all} assigned to institute i. This means each author assigned to institution i is an internal author of this institution i. The $external$ authors are determined in the same way as in the regression analysis with just the difference that the set is created through union of all $external$ authors of all $internal$ authors a of institute i. The values have been calculated for 3,422 institutes out of 3,427 institutes from MAS_{all}. The names of the attributes as well as a description and example instance can be seen in Table 4.3.

The defined attributes were in a first step analyzed regarding their correlation to $hindex$, to find out if an influence can be observed. It turned out that the amount of authors found for an institute ($cntInternHindexAll$) correlates with the H-index from the institute ($hindex$) with 0.78. This would mean that the amount of authors is more vital than their individual performance. The best correlation between the calculated attributes and $hindex$ is observed for $cntExternHindexOld$ which is 0.82. This would suggest that a lot of external relationships also have a positive influence on the institution's performance. The average H-index values do not correlate that much (between 0.16 and 0.48) but we can observe that the correlation of the average values for the old authors are higher than the correlation for young and middle aged ones and that the internal correlation values are all higher than the corresponding external ones ($avgExternHindexOld = 0.29 > avgExternHindexMiddle = 0.24$, $avgInternHindexOld = 0.48 > avgInternHindexMiddle = 0.36$).

This basic correlation analysis gave an overview of the dataset. The next step was the clustering process. K-means as method, with a k of 3 and the Euclidean distance as distance measure, was used. The clustering was performed based on several combinations of attributes. For example performing the clustering on $cntInternHindexMiddle$, $cntInterHindexOld$, $cntExternHindexMiddle$ and $cntExternHindexOld$ resulted in only 2 clusters. In k-means a standardization of the value range for different attributes is important so that there is no significant influence on the clustering process by outliers. To overcome this problem, the data was normalized for all numerical values by subtracting the mean and dividing through standard deviation of the proper attribute to limit the value range to $(-1, 1)$. These values have the same name as the numerical attributes in Table 4.3, just a leading std was added representing the standardized value of the same attribute.

The clustering process was performed on the standardized values of $stdAvgInternHindexYoung$, $stdAvgInternHindexMiddle$, $stdAvgInternHindexOld$, $stdAvgExternHindexYoung$, $stdAvgExternHindexMiddle$ and $stdAvgExternHindexOld$ resulting in 3 clusters with a distribution of 40/23/37% of insitutes per cluster. Another combination of attributes, with the standardized values $stdAvgExternHindexYoung$, $stdAvgExternHindexMiddle$ and $stdAvgExternHindexOld$ produced also 3 clusters with an fairly even cluster distribution of 30/47/27%. Another good result, in the sense of an even distribution of the cluster size, was produced through clustering over the $stdAvgInternHindexAll$, $stdAvgInternHindex$-

Attribute	Description	Value
$institute$	Name of the institution	Vienna University of Technology
$countrycode$	Countrycode of the country the institution belongs to	AT
$continent$	Continent where the country belongs to	EU
$hindex$	H-index fetched from MAS	81
$cntInternHindexAll$	Amount of all internal authors	209
$avgInternHindexAll$	Average of the H-index of all internal authors	5.73
$cntInternHindexYoung$	Amount of young internal authors	46
$avgInternHindexYoung$	Average of the H-index of young internal authors	3.41
$cntInternHindexMiddle$	Amount of middle internal authors	74
$avgInternHindexMiddle$	Average of the H-index of middle internal authors	4.01
$cntInterHindexOld$	Amount of old internal authors	85
$avgInternHindexOld$	Average of the H-index of old internal authors	8.66
$cntExternHindexAll$	Amount of all external authors	1134
$avgExternHindexAll$	Average of the H-index of all external authors	8.85
$cntExternHindexYoung$	Amount of young external authors	75
$avgExternHindexYoung$	Average of the H-index of young external authors	3.15
$cntExternHindexMiddle$	Amount of middle external authors	291
$avgExternHindexMiddle$	Average of the H-index of middle external authors	5.90
$cntExternHindexOld$	Amount of old external authors	768
$avgExternHindexOld$	Average of the H-index of old external authors	10.52

Table 4.3: The name, description and example instance of the attributes calculated for clustering for each institute

$Young$, $stdAvgInternHindexMiddle$, $stdAvgInternHindexOld$, $stdAvgExternHindex$-All, $stdAvgExternHindexYoung$, $stdAvgExternHindexMiddle$ and $stdAvgExternHindex$-Old attribute. The result consisted of 3 clusters with a distribution of 35/24/41% institutes per cluster.

As it is not practicable to test clustering on every combination of this 16 numerical attributes, we had to evaluate the quality of our results against an external source. We checked the cluster assignments against the top 500 institutes according to Webometrics Ranking of World Universities[33]. To check the assignments we first included into our created dataset of the institutes from MAS_{all}, the cluster assignments from the several performed clustering runs. Subsequent the institutes from MAS_{all} were matched to the corresponding institute of the top 500 from Webometrics, were out of the 500, 416 could be assigned. The cluster assignments of those 416 institutes were analyzed regarding their ranking in Webometrics. To be able to evaluate the results, the top 500 institutes were split up into groups of 50 institutes each. Then the amount of institutes per cluster for each of this fifties group was calculated by considering the 416 assigned institutes. These values were then visualized and analyzed and so the most accurate clustering for the top 500 was determined. The assumption is, that the clustering representing the top 500 institutes according to Webometrics properly, also represents the whole dataset best, in the sense of splitting the data into top, middle and low institutes.

The clustering, which represented the top 500 best, was created out of the standardized attributes $stdAvgInternHindexAll$, $stdAvgInternHindexYoung$, $stdAvgInternHindex$-$Middle$, $stdAvgInternHindexOld$, $stdAvgExternHindexAll$, $stdAvgExternHindexYoung$, $stdAvgExternHindexMiddle$ and $stdAvgExternHindexOld$, where 6 institutes got assigned to cluster0 (low institutes), 297 to cluster 1 (top institutes) and 113 to cluster 2 (middle institutes). As you can see in Figure 4.13, this clustering represents the top 500 very well, by putting most of the institutes in cluster 1 and least in cluster 0.

The clustering information from above was in a next step used to train a classification model. To create a dataset on which the decision tree is trained, all of the 3,422 institutes from MAS_{all} and the added cluster assignment as class information were used. The aim was to create a decision tree which can be used to assign institutes to top/middle/low class based on their attributes. A trade off between the amount of attributes used for training the model, the size of the decision tree and the accuracy of the model in regard of the cluster assignment, had to be taken. Different combinations of attributes were tested. The classification model based on the attributes used for the clustering process (standardized average of young/middle/old/all in intern/extern) resulted in the most accurate model by classifying 98% of the instances correctly, but with a tree size of 134 leaves it is not applicable. The combination of the attributes $avgInternHindexAll$, $avgInternHindexYoung$, $avgInternHindexMiddle$, $avgInternHindexOld$, $avgExtern$-$HindexAll$, $avgExternHindexYoung$, $avgExternHindexMiddle$, $avgExternHindex$-Old, $countrycode$ and $continent$ resulted in a fairly good classification model, in respect of size and accuracy. 78% correctly classified instances and a tree size with 24 leaves is acceptable.

To check the accuracy of the result, the same approach as for clustering was followed, the class assignment from the decision tree was checked against the top 500 institutes of Webomet-

[33] `http://www.webometrics.info` (accessed 14-November-2011)

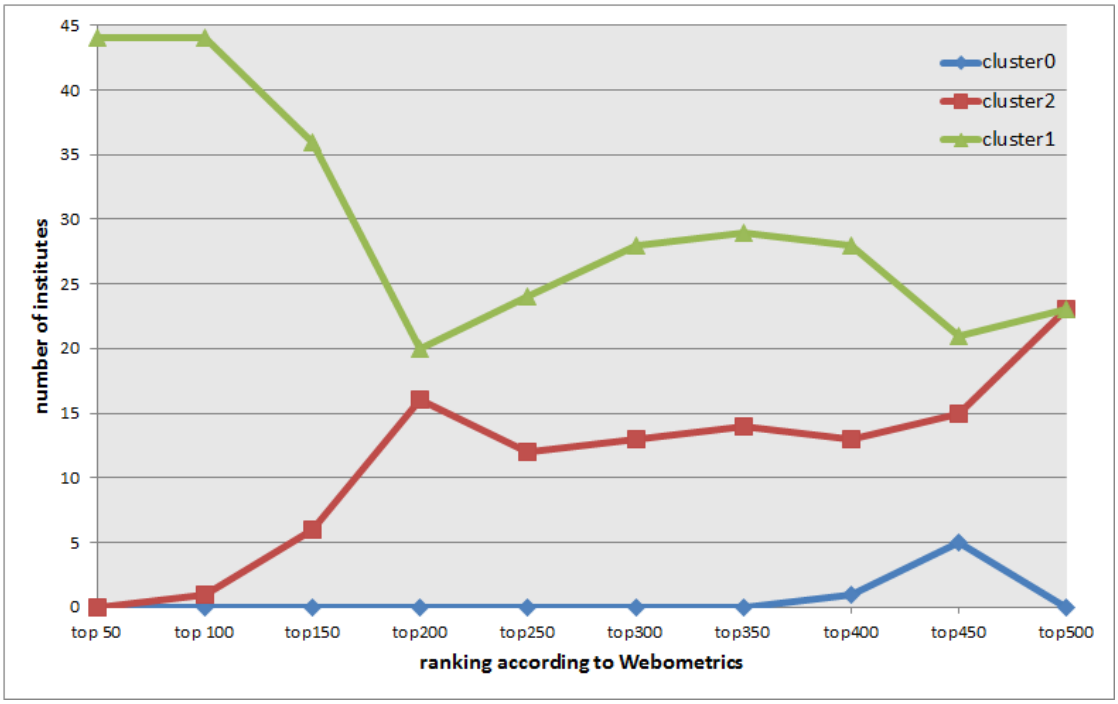


Figure 4.13: Visualization of the cluster assignments compared to the ranking of Webometrics

rics. Out of the 416 institutes available in the dataset, 15 got classified as low, 113 as middle and 288 as top institute in the top 500 ranking. For an overview of the distribution see Figure 4.14. We can observe a very similar picture as in Figure 4.13 and hence this classification model is assumed as accurate to represent the influence of authors of an institute to the ranking of the institute.

The decision tree created through the classification model is visualized in Figure 4.15. If we take a closer look at the tree, and analyze some of the paths interesting findings are unveiled. For example, it seems that the total average of H-index external and internal are used for the big decisions. That means if you are not having at least an average H-index for all external authors above 6 and for all authors internal above 4, the chance for getting classified as a top institution is just 12.5% (2 out of 16 leaves). What also can be observed is that all 5 decision resulting in a top classification are dependend on the average H-index of the middle aged authors, which leads to the assumption that scientists with a publication age between 5 and 10 years are a very vital group for the performance of an institution. This can also be traced back to a finding from Biryukov [8], that the most productive time in a scientific career is between the 6th and 10th year.

We can see that the classification model answers the question 'Can we observe a relationship between the ranking of an institute and the authors of an institute?', by providing the proper insights in the decision tree in Figure 4.15.

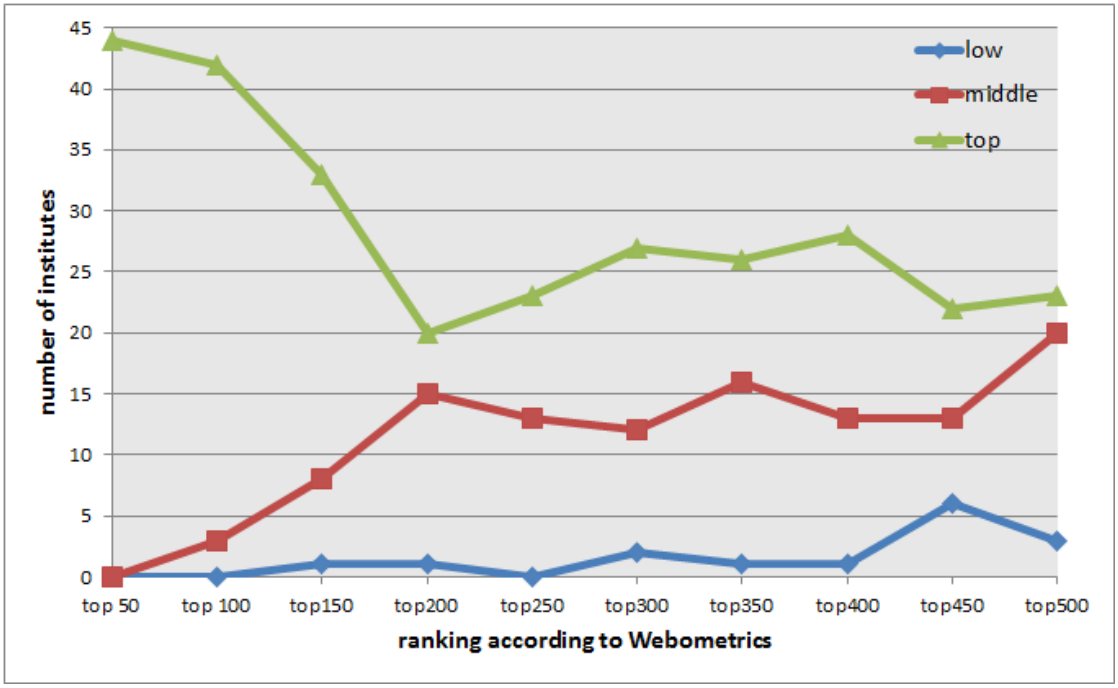


Figure 4.14: Visualization of the classification compared to the ranking of Webometrics

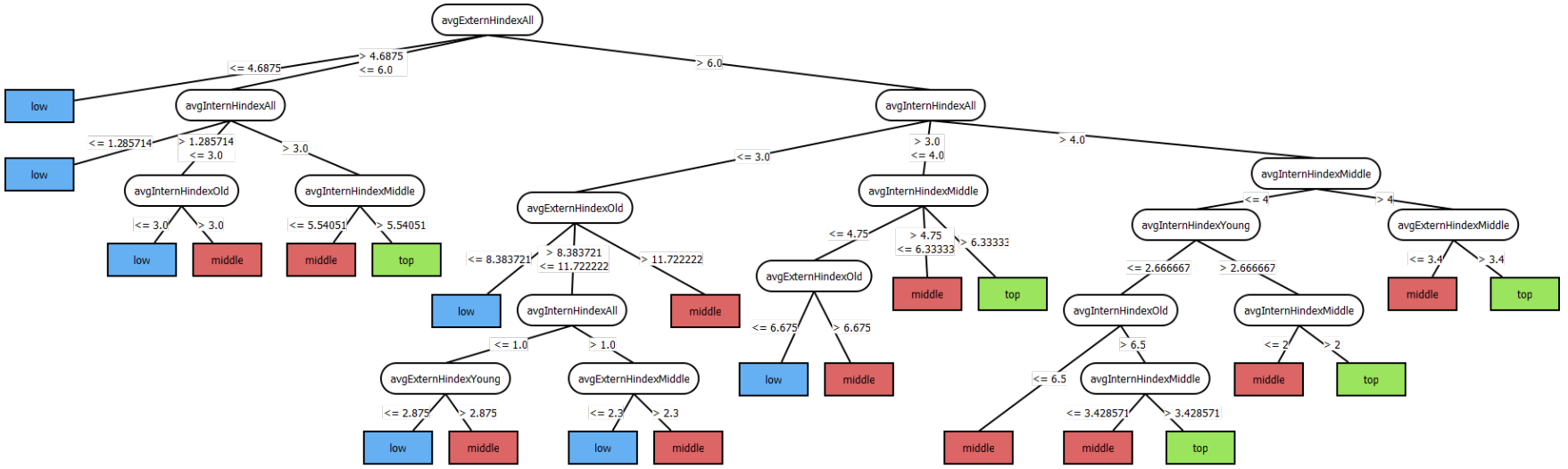


Figure 4.15: Decision tree created through the classification model

Continent	Amount of Authors	Percentage of all Authors
Africa	337	0.5%
Asia	12,215	19.3%
Europe	23,718	37.4%
North America	24,516	38.6%
Oceania	1,642	2.6%
South America	1,017	1.6%
Antarctica	0	0.0%

Table 4.4: Amount of authors assigned to continents from MAS_{geo}

As a finding in Section 2.1 is that the funding system can influence research, we investigated this relationship in our data. The dataset of the 3,422 institutes was aggregated to countries and combined with additional information. From the OECD Education at a Glance report 2011 [49] the data about expenditures on educational institutions as a percentage of GDP by source of fund for the tertiary education was used (Table B.2.3[34]). This data was related to the percentage of institutes classified as top per country. In Figure 4.16 you can see the result of this comparison. You can see the high amount of private funding in the United States, Canada and Japan and the high amount of public funding in the Scandinavian countries like Norway, Finland and Denmark. But either way there seems to be no systematic patterns and a test on significant influence of public and private funding on the amount of top institutes resulted in a rejection of this hypothesis.

Are there regional differences concerning the amount of scientists?

The aim of this analysis is to determine if a bias between the amount of authors and countries can be observed. This is interesting as it provides a conclusion back on the structure of the dataset. To perform the analysis we took the dataset MAS_{geo} and visualized all found authors on a Google Map. In a first approach each single marker was visualized (see Figure 4.17) for a better overview the markers were clustered in a second approach (see Figure 4.18).

The distribution of the amount of authors in Figure 4.17 and Figure 4.18 shows a clear picture. Europe and parts of the US are fairly good covered with scientists in the dataset. In Table 4.4 you can see that Africa, South America and Oceania together not even have 5% of authors in the MAS_{geo} dataset.

For answering our question if there are regional differences a χ^2 test on uniform distribution for this data resulted, as it was expected, in a rejection of the hypothesis, which means no uniform distribution is present.

Subsequent in the next step we investigated the relationship between a continent and amount of authors. This was done by taking the dataset MAS_{geo} and performing a regression analysis. The observations used for testing the regression model were created for each country c. The dependent attribute was the amount of authors $authors_c$ for this country c, and the independent attributes were a binary encoding for the continent of country c. This binary encoding

[34] `http://dx.doi.org/10.1787/888932463802` (accessed 17-October-2011)

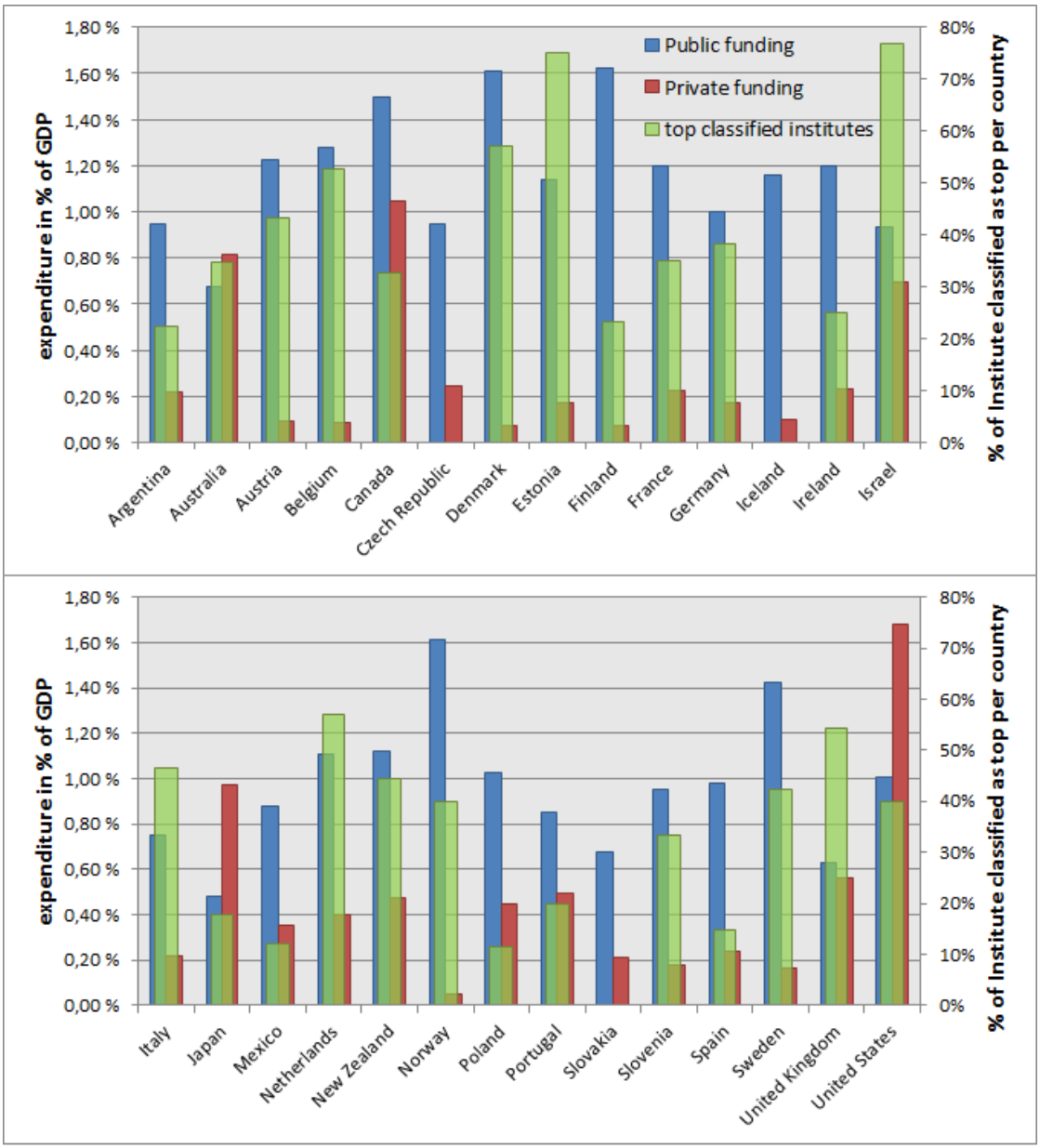


Figure 4.16: Comparison between percentage of top classified institutes per country to public and private funding in percentage of GDP

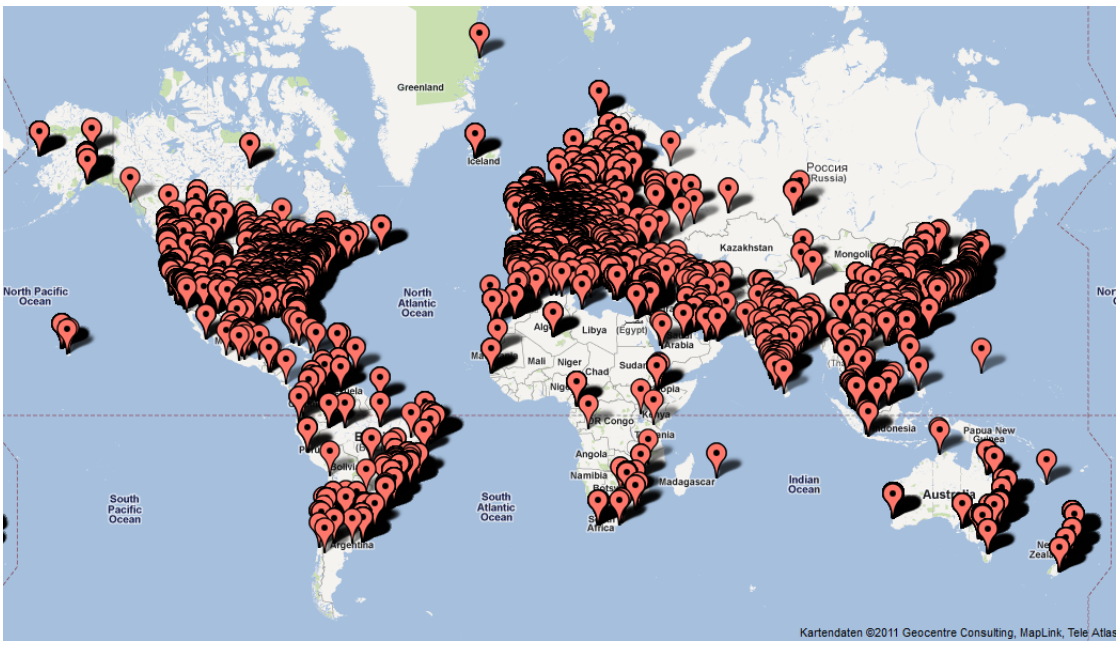


Figure 4.17: Google Maps visualization of authors from MAS_{geo} as markers

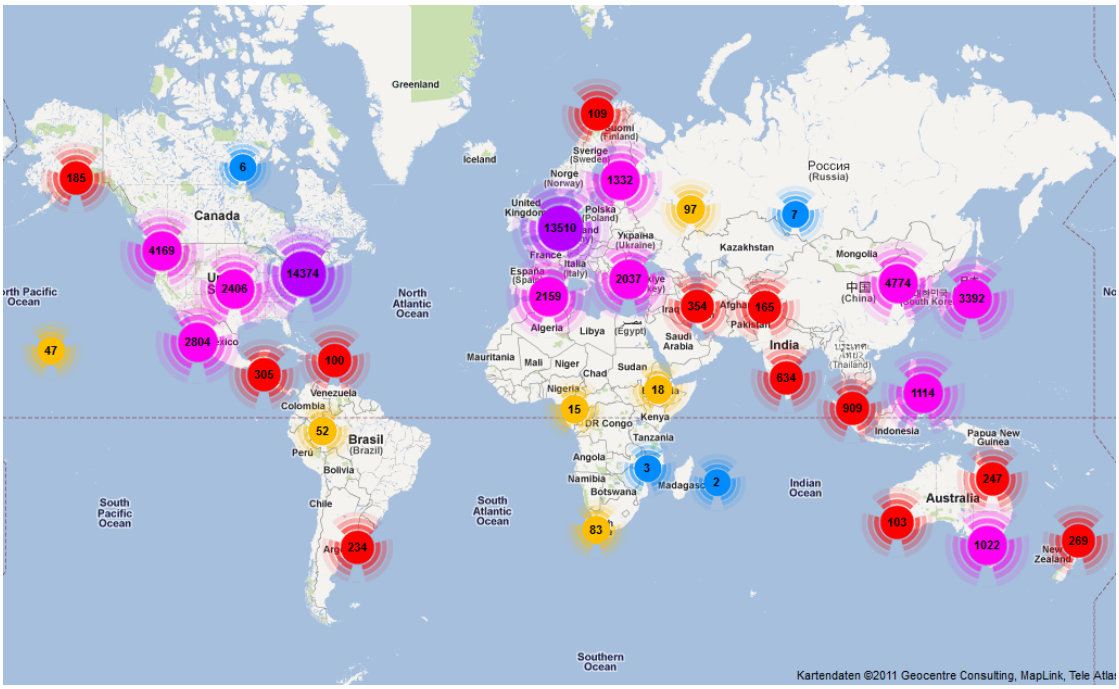


Figure 4.18: Google Maps visualization of authors from MAS_{geo} as clustered markers

for the continent is built such that the continent the country origins is represented by a 1 and the other continents by 0. For example the country Austria ($c = Austria$) would have a 1 for $Europe$ and 0 for $Africa$, $NorthAmerica$, $SouthAmerica$, $Asia$ and $Oceania$ as well as $author_{Austria} = 573$. The dataset created for each country in MAS_{geo} with this attributes was then tested on the following model.

$$authors_c = \beta_0 + \beta_1 Africa_c + \beta_2 Asia_c + \beta_3 Europe_c + \beta_4 NorthAmerica_c \\ + \beta_5 Oceania_c + \beta_6 SouthAmerica_c$$

The first test with a constant element in the model resulted in a not significant model. Holding the constant $\beta_0 = 0$ resulted in a significant model (test value of $F = 2.74 > F^*(0.95, 5, 109) = 2.30$). Looking at the parameters showed the expected result, that the only parameter significantly different from zero at a confidence level of 0.95% was b_4 for North America, b_3 for Europe is significant for a confidence niveau of 0.9%.

With the result from the regression analysis we can answer the question about regional differences. A significant influence between the continent of an author and the amount of authors from this continent is present. This can either be justified by the reason that no research is done in these areas or it is not represented accordingly in citation databases, and hence in such an analysis. The underrepresentation also conforms with findings from Chapter 2, as one problem is the quality of the citation database. Another problem are possible cultural differences resulting in this significant results.

Is there a cultural influence on research in computer science?

To answer this question we have to investigate our datasets $Base_{one}$ and $Base_{cluster}$, to find out if cultural influence can be determined. To get an overview over the data, visualizations are created in a first approach. Two different forms of visualization have been selected, in one of them a Google Map with markers is presented. The markers represent authors and can be filtered by year of publication and subfield of computer science. With this approach it is possible to either look at the evolution of a proper subfield of computer science over time, or compare the status of two different subareas in the same year. The second visualization is based on Google GeoChart, providing an overview of the world with color encoding of countries. The color codes are determined by the numerical values of the absolute or relative amount of authors for each country filtered by topical subareas. Such a visualization provides the opportunity to compare countries and their performance in various subfields. Both visualizations are based on the data of $Base_{one}$ and $Base_{cluster}$.

To demonstrate the visualization we take a closer look at the data from $Base_{cluster}$ in the subfields 'H. Information Systems' and 'G. Mathematics of Computing' for the Russian Federation in contrast to Sweden. The total amount of authors assigned to Sweden is 619 and 106 to Russia. These 725 authors represent slightly above 1% of all authors in the dataset with geoinformation assigned. Now if we take a look at amount of authors assigned to the subfield G., 37.7% (40) authors of Russia and 13.4% (83) authors of Sweden are assigned to this class. In contrast to the subfield H., where 18.9% (20) of the Russian authors and 18.9% (117) of the

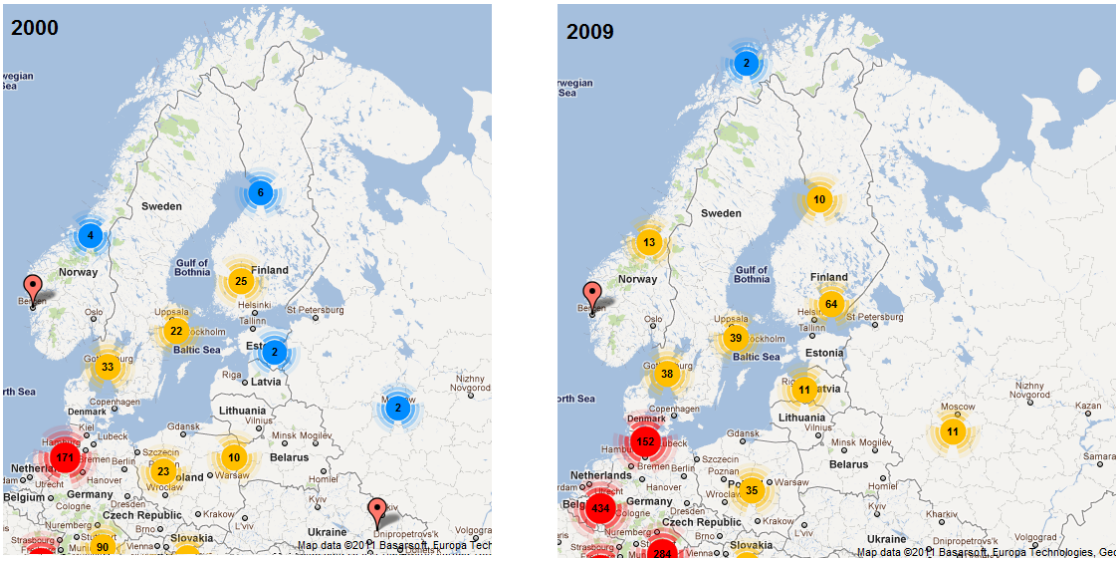


Figure 4.19: Development of publications made in class 'H. Information Systems' in $Base_{cluster}$ dataset

Swedish authors are assigned to this subfield. This observation shows a tendency of a more mathematical research in Russia than in Sweden.

To investigate if this relationship can also be observed by looking at a particular time span we compared the publication amount for the years 2000 and 2009. Looking at Russia we can see a stable amount of publications in Mathematics of Computing of 20% in 2000 (14 publications) and 2009 (41 publications). In contrast to Sweden where the amount of publications in the G. class shrinked from 5% (27 publications) in 2000 to 3.7% (44 publications) in 2009. For the H. class an increase of publications for Russia between 2000 and 2009 of 3.8%, from 4.3% (3 publications) to 8.1% (17 publications) can be observed. Sweden also increased the amount of publications in Information Systems from 2000 to 2009 by 0.4%, from 8.1% (43 publications) to 8.5% (101 publications). This evolution is visualized in Figure 4.19 and Figure 4.20. So we can see that the finding about authors and the regional influence on the publications can also be observed in the evolution over time.

The insights gained through the first visualizations were further analyzed with the visualization of the relative amount of authors of a country in a particular subfield. Looking at the CCS class 'F.4 Mathematical Logic and Formal Languages' in Figure 4.21, we can see that in countries like Russia (12%), Poland (10.8%), Ukraine (12.5%), Romania (9.3%) and Serbia (8.5%) the relative amount of authors in this field is fairly high for each country compared to United States (2.4%), China (2.2%), Austria (2.8%), Germany (3.3%), Australia (2.1%) and Spain (2.1%) for example.

To support our visual findings a regression analysis was performed. We created a dataset which consisted of the fraction of authors from a country c in a particular CCS class ccs to the total amount of authors from this country c on basis of the data in $Base_{cluster}$. This fraction, the

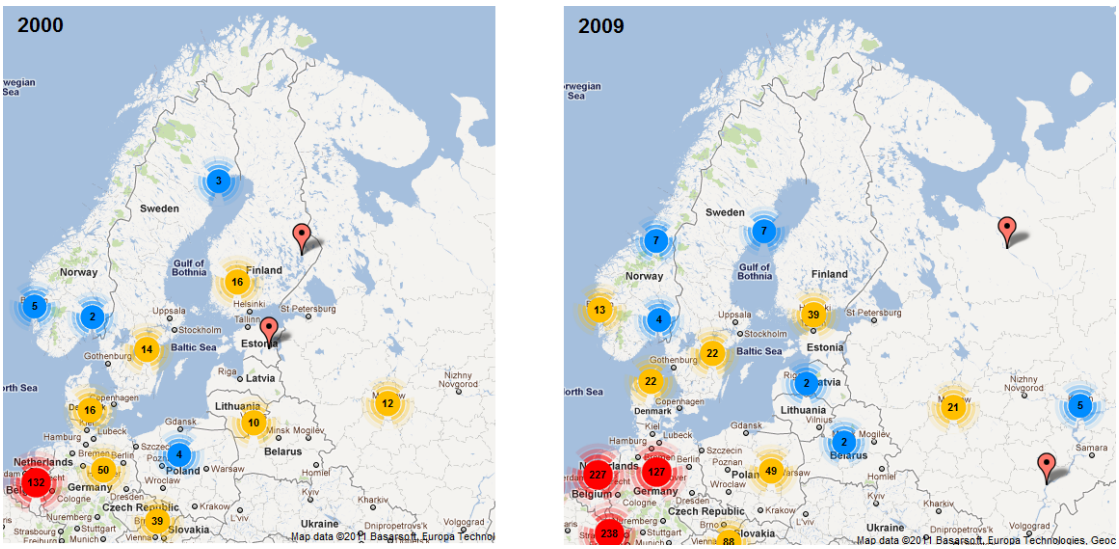


Figure 4.20: Development of publications made in class 'G. Mathematics of Computing' in $Base_{cluster}$ dataset

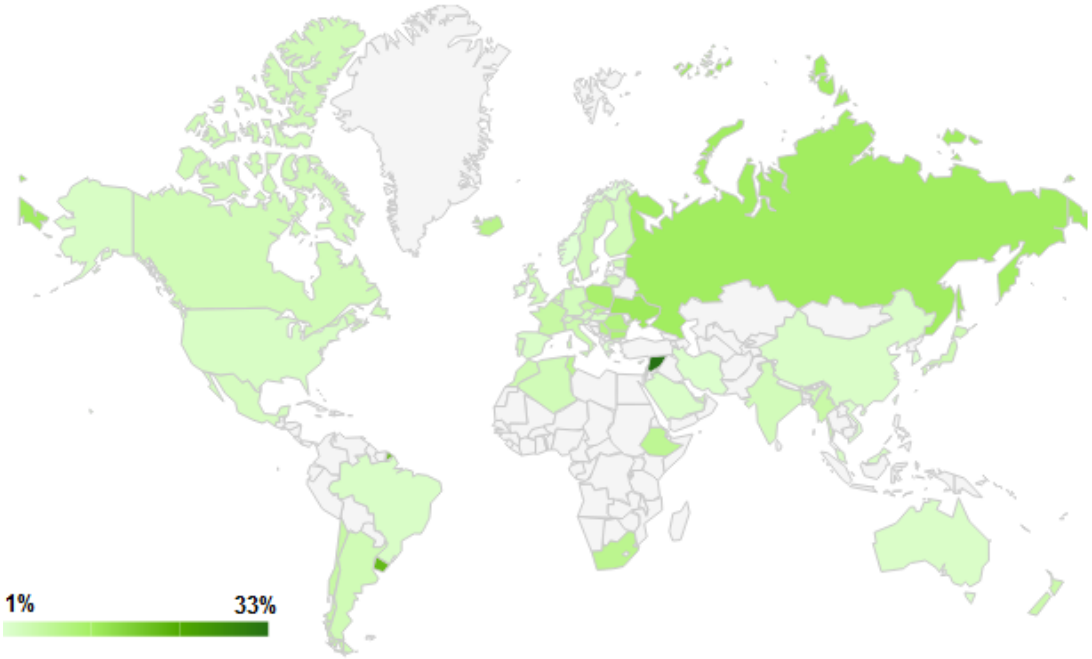


Figure 4.21: Visualization of the relative amount of authors per country in class 'F.4 Mathematical Logic and Formal Languages' from $Base_{cluster}$

relative amount of authors from a country c in a CCS class ccs, is represented by $authors_{ccs,c}$. The top CCS classes from A. to K. were used. To be able to evaluate cultural differences, the countries are grouped into regions according to the United Nations Statistic Division standard M.49[35]. Normally the regions for Africa would be, Northern Africa, Western Africa, Middle Africa, Eastern Africa and Southern Africa. America is split into, North America, South America, Central America and the Caribbeans. Asia into, Western Asia, Central Asia, Southern Asia, Eastern Asia and Southeastern Asia. Europe into, Western Europe, Northern Europe, Southern Europe and Eastern Europe. And Oceania into, Melanesia, Micronesia, Polynesia, Australia and New Zealand. For our purpose we merged the regions Western Africa, Middle Africa and Eastern Africa to 'Africa Middle'. Also Central America and the Caribbeans were merged to 'Americas Middle' and Oceania is used as continent. To create observations on which the regression model can be tested a similar structure as the model in the previous question was used. The relative amount of authors of a country c for a CCS class ccs as dependent attribute $author_{ccs,c}$, is expressed by a binary encoding of the regions. The encoding for the regions is created such that the region r to which country c belongs is encoded with 1 and the other regions with 0. For example the relative amount of authors from Austria in class H. Information Systems is $author_{H.,Austria} = 0.26$. The regions Northern Africa, Southern Africa, Africa Middle, North America, South America, Americas Middle, Northern Europe, Southern Europe, Eastern Europe, Central Asia, Southern Asia, Southeastern Asia, Western Asia, Eastern Asia and Oceania get assigned a 0 and Western Europe 1. For each of the CCS top classes from A. to K. observations were created by expressing $authors_{ccs,c}$ for every country c in the dataset with following model.

$$\begin{aligned} authors_{ccs,c} = \beta_0 &+ \beta_1 NorthernAfrica_c + \beta_2 AfricaMiddle_c + \beta_3 SouthernAfrica_c \\ &+ \beta_4 NorthAmerica_c + \beta_5 SouthAmerica_c + \beta_6 AmericasMiddle_c \\ &+ \beta_7 WesternEurope_c + \beta_8 NorthernEurope_c + \beta_9 SouthernEurope_c \\ &+ \beta_{10} EasternEurope_c + \beta_{11} SoutheasternAsia_c + \beta_{12} WesternAsia_c \\ &+ \beta_{13} CentralAsia_c + \beta_{14} SouthernAsia_c + \beta_{15} EasternAsia_c \\ &+ \beta_{16} Oceania_c \end{aligned}$$

This model was tested on the different CCS classes, which resulted in a confirmation of the visual findings. If we look at the structure of the model from above, the outcome of the analysis are values for the parameters β_0 to β_{16} expressing the relative amount of authors per country for a subfield. This is caused by the binary encoding of the regions. If the model is significant with a $\beta_0 \neq 0$ an average relative amount of authors for this subfield is present and expressed by β_0, and the influence of a region β_1 to β_{16} can be positive or negative. If the model has to be restricted such that $\beta_0 = 0$ to be significant, the parameters β_1 to β_{16} can only be positive, expressing the relative amount of authors per country for this region in a proper subfield.

The regression model for the CCS class H. is significant. It shows an average amount of relative authors per country of 35.1% (β_0), this average value is for a confidence level of 0.9

[35] `http://unstats.un.org/unsd/methods/m49/m49regin.htm` (accessed 27-October-2011)

significantly reduced for Eastern Europe by $\beta_{10} = -16.9\%$ to a relative amount of authors in this field of 18.2%. Also Western Asia has a significant negative influence on this average value with $\beta_{12} = -17.4\%$ resulting in an relative amount of 17.7% authors from this region working in the subfield Information Systems. The only region positively influencing this average amount, is Africa Middle with $\beta_2 = 23.4\%$ resulting in 58.5%. For the other regions no significant influence on the average amount of relative authors per country of 35.1% in subfield H. was observed.

G. Mathematics of Computing was also analyzed with the regression model, but to obtain a significant model, the constant value β_0 was restricted to 0. Within this model, countries from Northern Europe ($\beta_8 = 14.3\%$) have beside countries from Western Europe ($\beta_7 = 13.0\%$) the lowest relative amount of authors in this field. In contrast countries from Eastern Europe with $\beta_{10} = 28.4\%$ are part of the top 4 in this subarea.

The cultural differences, and in this case cultural is based on geographic regions as defined above, for the various subareas in the computer science field based on the top classes from ACM CCS A. to K. are summed up here.

A. General Literature

   For this subfield we could not find any significant model or influence because of the little amount of scientists (677) assigned to this area.

B. Hardware

   To get a meaningful result we had to restrict the constant element β_0 to 0. With this restriction a significant model can be observed. The parameters for Africa Middle, North America and America Middle were not significant. Countries belonging to all the other regions have a relative amount of authors in this subfield between 11.9% and 23.1%. The only two regions having a greater relative amount of authors in this subfield are Southern Africa with $\beta_3 = 26.9\%$ and Southeastern Asia with $\beta_{11} = 30.3\%$.

C. Computer Systems Organization

   The analysis of the regression model for this class was significant with a constant element unequal 0. β_0 expressing the average relative amount of scientists in this subfield per country has a value of 36.4%. This influence is just enforced by the region Africa Middle with $\beta_2 = 27.9\%$ resulting in an relative amount of authors for each country in this region of 64.3%.

D. Software

   A meaningful result was observed in this analysis. The parameter for the average amount of relative authors in this field β_0 is 23.7%. Similar as with the class C., Africa Middle is the only region with a significant influence on the amount of researchers in this field with $\beta_2 = 31.1\%$ resulting in a relative amount of 54.8% in this field for Africa Middle.

E. Data

The regression analysis returned a significant model and interesting insights in this subfield. The parameter β_0 expressing the average relative amount per country, is 30.7%. Taking a closer look turns out that North America and Africa Middle are the only regions holding this average value, because their parameters β_4 and β_2 are zero and so do not influence the average value. The other regions all significantly reduce this average value which leads to a relative amount of authors per country in this subfield for the other regions between 3.9% and 14.1%.

F. Theory of Computation

A statement for this subfield is difficult. To get a significant result we had to set $\beta_0 = 0$, this lead to a model with parameters for each region. Southern Africa with $\beta_3 = 27.8\%$, North America with $\beta_4 = 25.2\%$, Americas Middle with $\beta_6 = 35.2\%$ and Western Asia with $\beta_{12} = 20.1\%$ are the top four regions concerning the relative amount of authors. The remaining regions vary between 10% and 20%.

G. Mathematics of Computing

This subarea can be expressed by a model with $\beta_0 = 0$. The model shows a highly positive influence with more than 28% relative amount of authors per country for the regions Eastern Europe ($\beta_{10} = 28.4\%$), Southern Asia ($\beta_{14} = 28.6\%$), Western Asia ($\beta_{12} = 30.5\%$) and Americas Middle ($\beta_6 = 34.9\%$). Other regions like Southern Africa, North America and Oceania have no significant parameter. The remaining regions vary between 13.0% and 23.7%.

H. Information Systems

Information Systems are fairly covered in all regions, which is expressed by an average relative amount of $\beta_0 = 35.1\%$ scientists per country in this field. Africa Middle even enforces this by $\beta_2 = 23.4\%$ to 58.5%. A significantly decline of the mean 35.1% can be observed for Eastern Europe and Western Asia resulting in an average of 18.2% respectively 17.7% relative amount of authors per country in this region.

I. Computing Methodologies

This subfield with most of the scientists assigned to (31,716) shows a picture, that only Southern Africa (β_3) has no significant influence. The constant value β_0 for the model is 0 and all the regions except Southern Africa have an average of 45.7% to 70.0% for the relative amount of authors per country in this field.

J. Computer Applications

An average relative amount can't be observed in this subfield ($\beta_0 = 0$). The regions of Southern Africa, Africa Middle and Southern Europe have an significant influence of $\beta_3 = 28.8\%$, $\beta_2 = 42.9\%$ and $\beta_9 = 16.8\%$ for the relative amount of authors. The remaining regions have no significant influence on this subarea.

K. Computing Milieux

The Computing Milieux is an interesting area. The average value for the relative amount of authors per country in this field is significant with $\beta_0 = 35.8\%$. Beside Southeastern Asia, North America, Middle Africa and Oceania which have no impact on this average value, the other regions significantly reduce this value. The value range for this remaining regions is between 9.3% and 15.6%.

We answered in this subsection the question about cultural influence on research by first visually analyzing the data and then performing a regression analysis. It turned out that a cultural bias is present and can be proved either visually or analytically.

Is there a bias between the venue of a conference and the origin of authors joining the conference?

With this question we find out if a relation between the conference venue and the origin of the authors publishing at a conference is present. This is interesting as it can show how important a conference location is and how carefully it should be selected. For this reason we visualized authors publishing papers at a conference on a map and looked at the venue of the conference itself. To find out the venue, some manual research on the Web was performed. To figure out if a bias exist we chose the conference on 'Very Large Data Bases VLDB', which is according to Biryukov [8] a top conference.

In a first approach a visual check was performed for the conference VLDB of the years 1978 and 1979. In 1978 the conference was held in Berlin, and in $DBLP_{all}$ 117 distinct authors published at this conference, out of them 35 (29.9%) are available in MAS_{geo}. In 1979 it was held in Rio de Janeiro, with 101 authors in $DBLP_{all}$ and 38 (37.6%) in MAS_{geo}. Looking at Figure 4.22 and Figure 4.23 we can observe a clear bias between the conference location and the authors publishing there. 1978 no authors from Brazil nor South America published, which changed 1979. Also an increase of authors from North America in contrast to a decrease of authors from Europe can be observed from 1978 to 1979.

Following the insight from above, that a clear bias between the location of the conference and the authors publishing there was present in the 1970s, we checked this influence on current VLDB conferences again. Therefore we looked at the VLDB conference of 2006 held in Seoul and 2007 in Vienna. The amount of authors for example from Asia stayed constant at 41, an increase of authors from North America and Europe can also be observed, but this can be justified by the total increase of authors. Looking at Figure 4.24 justifies this finding, that for current VLDB conferences no bias between the location and the authors joining the conference is present anymore.

Subsequent to this investigation of a top conference, we chose nontop conferences according to Biryukov [8] in the same subfield, Databases. The 'International Workshop on the Web and Databases' short 'WebDB' and 'International Database Engineering and Applications Symposium' short 'IDEAS' are two candidates for this investigation. Looking at Figure 4.25 we can see that there is a change in the authors landscape publishing at IDEAS 2006 in Delhi, India and 2007 in Banff, Canada. Whereas 2006 authors from India and Australia were at the conference,

Figure 4.22: Visualization of authors published at VLDB conference 1978 in Berlin from MAS_{geo}

Figure 4.23: Visualization of authors published at VLDB conference 1979 in Rio de Janeiro from MAS_{geo}

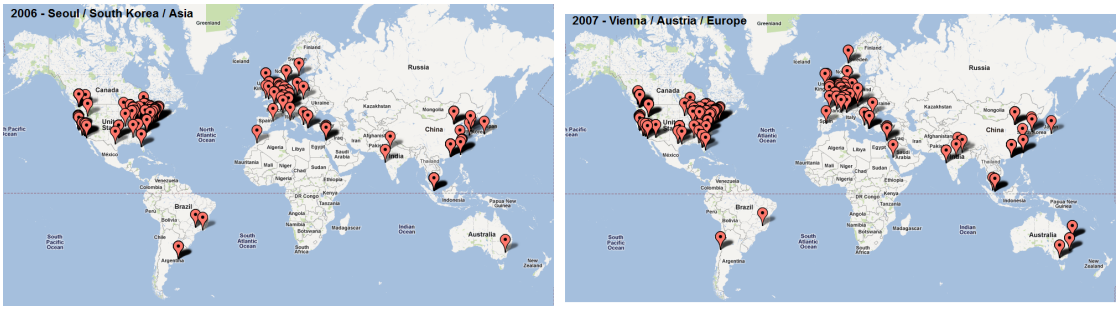


Figure 4.24: Comparison of authors published at VLDB conference 2006 and 2007 (MAS_{geo})

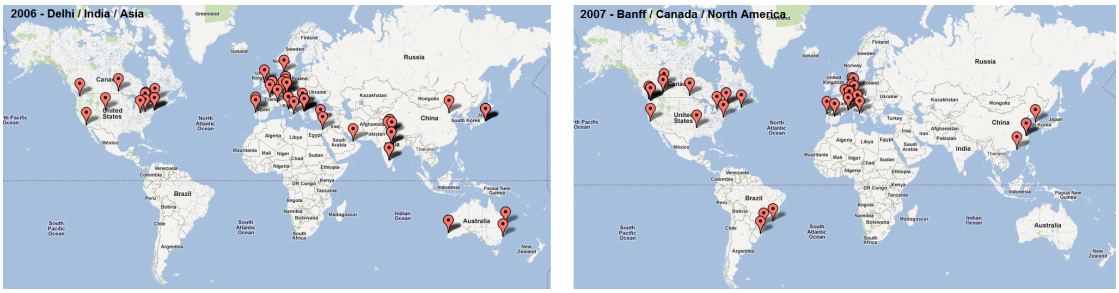


Figure 4.25: Comparison of authors published at IDEAS conference 2006 and 2007 (MAS_{geo})

2007 no authors from this countries joined the conference. This shows that a bias between the venue of a conference and the authors joining can be observed.

To verify the visual findings a regression analysis was performed on the data from VLDB, WebDB and IDEAS. Therefore in a first step the model was defined. The amount of authors from a continent per year $author_{c_i}(year)$ is the dependent attribute, which is expressed by a binary encoding of the continent the conference was held. The variables $AF(year)$, $AS(year)$, $EU(year)$, $NA(year)$, $OC(year)$ and $SA(year)$ are this binary encoded variables, which means $AF(year)$ is 1 if the conference was held in the year $year$ on the African continent, and 0 else. According to this the same was used for AS representing Asia, EU Europe, NA North America, OC Oceania and SA South America.

$$c_i = \{Africa, Asia, Europe, NorthAmerica, Oceania, SouthAmerica\}$$

$$\begin{aligned} authors_{c_i}(year) = \beta_0 + \beta_1 AF(year) + \beta_2 AS(year) + \beta_3 EU(year) \\ +\beta_4 NA(year) + \beta_5 OC(year) + \beta_6 SA(year) \end{aligned}$$

This regression model was tested on the observations of the VLDB conference from 1990 to 2010, for the WebDB from 1998 to 2010 and for IDEAS from 1997 to 2010 for each continent, which resulted in following observations.

- VLDB
  Performing the regression analysis on the VLDB data from 1990-2010, no significant model could be verified for any continent. This means that it was not possible to find a bias between the continent the VLDB conference is held and an influence on the amount of authors from a particular continent publishing there.

- WebDB
  The regression analysis was performed on the data from the first WebDB 1998 until 2010. For the observations of Europe a significant model is present. Meaning that, the amount of authors publishing at WebDB is related to the continent it takes place. If the conference is held in Europe this positively influences the amount of authors from Europe publishing there.

- IDEAS
  Looking at the other nontop conference, the IDEAS, which was analyzed with data from 1997-2010, a significant model was found for the amount of authors from Asia, which tend to be higher if the conference is held in an Asian country.

So we can see, that the visual observations can be analytically verified with the data. As an insight in Section 2.1 was that young scientists tend to first publish at nontop conferences a conclusion could be, that proper distribution of the venues can help to get young researchers elaborated, as they tend to join a conference if it is held locally. If we reason that publishing at a conference increases the personal H-index, because the paper gets more often cited, the scientist also contributes to the ranking of the institution he or she works for. This would hence imply also that locally held conferences, even if they are ranked nontop, would increase the status of an organization/country or region.

CHAPTER 5

Conclusion and Future Work

After analyzing the data, this chapter summarizes the work and suggests areas for future investigation.

5.1 Conclusion

We demonstrated in Chapter 4 how we can connect several datasources and create a dataset which then can be analyzed. The created dataset which was used for answering several questions has also some restrictions which is important to name.

First of all, the whole dataset is created upon the information in DBLP and as Reitz [51] shows that there are differences in the coverage of the subareas of Computer Science this bias is also present in our dataset. As a lack of other sources providing that amount of information in that quality and a different approach for thematic classification of publications in contrast to the work of Reitz it is a minor drawback. Beside the thematic bias, the selection of top-authors based on publication count can also be seen critical, as this is seen as a not very good representation for the quality of a scientist's work [32]. Due the fact that the amount of authors has to be limited, to be able to investigate the dataset, this approach seems applicable and out of the 88,852 top-authors 76,974 (86.6%) were found on MAS. The way how thematic information is added to the dataset is different from common approaches. The results of the analysis show that it is actually a good approach and furthermore it can be automated. The visualizations, especially the Google Tools (Fusion Tables, Maps and Chart Tools) demonstrate the variety of options available on the Web to represent data.

Now looking at the analyses of this work we can see that the structure of the dataset supports findings of previous works [20], for example the decreasing amount of single authors which is a major shift in Computer Science, expressing that a topic nowadays most certainly is investigated by a group of people. This evolution also can be linked to the fact that nowadays it is easier to communicate over the Web which enforces the interdisciplinary work.

The question 'Can we observe a relationship between the ranking of an institute and the authors of an institute?' gives an interesting insight into the performance of an organization and

the correlation to the scientists from this organization. As we were able to verify that clusters are related to the ranking of institutes by checking the results against an external ranking (Webometrics), we could find a classification model which represents this performance ranking. The decision tree gives interesting insights, for example that the average H-index of all external co-authors of an institution is important in the way, that if it is below 6, the organization gets assigned as low institute with a chance of 50% (3 out of 6 leaves). The performance of the old authors (publication age > 10 years) has no big influence. The internal old authors are included in 11.76% of the decissions (2 out of 17 nodes), one between low and middle classification and one between middle and top. The external old authors are also included in two decissions, but both distinquishing between low and middle classficiation. Very interesting is, that the path through the tree to get classified as top institute always passes the average of all authors internal and external and average internal for the middle aged scientists (publication age between 5 and 10 years). If the value of average H-index for all external co-authors is above 6, for all internal co-authors above 3 and for the internal middle aged scientists above 4.75 a classification as top or middle institute is sure. These findings would imply that for an institution the best way to improve recognition is to hire middle aged scientists ($5 <$ publication age ≤ 10) who have a good external network of co-authors, and a proper performance on their own which would lead to an increase of the internal performance and also the average of the middle aged scientists at this institution, and thus a better recognition of the organization.

The next finding of this work based on the question 'Are there regional differences concerning the amount of scientists?' is that there are systematic differences in the amount of authors in the database concerning regions. This can be traced back to the problem stated above, the uneven representation of subfields in the datasource, or also to the fact of cultural differences in research and an uneven knowledge distribution over the world.

Cultural influence has also been examined by the question 'Is there a cultural influence on research in computer science?'. We have found significant influences for the different subareas of Computer Science for regions. The subfield 'H. Information Systems' is significantly less researched in Eastern Europe (by just 18.2%) and Western Asia (by just 17.7%). 'G. Mathematics of Computing' on the other side is positively biased by 28.4% for Eastern Europe. Some other interesting observations are, that 'E. Data' is popular for North America and that the region Middle Africa often appears with a significant influence. This can be traced back, if we look at the question 'Are there regional differences concerning the amount of scientists?' on the uneven distribution of authors per continent. As Africa has only 0.5% of all authors assigned, the few authors influence their research area a lot, pushing the relative amount of their subfields to a high value.

In the last question 'Is there a bias between the venue of a conference and the origin of authors joining the conference?' we could show that high impact conferences are not biased with the location they are held. So to say an elaborated conference will be joined by scientists even if they have to travel far. But we showed also that nontop conference have a relation between the venue and the authors country of origin. This finding with the assumption that conferences are a place to network and build connections to other institutions and help to elaborate the work of local scientist by getting recognized and cited, can increase the local authors H-index. We close the loop by the finding that top institutes are represented by a high external as well internal

H-index, which can be achieved by the assumptions and findings from above by nontop conferences, and so implying that nontop conferences have their justification and purpose to boost the performance of an institution from the region the conference is held.

5.2 Future Work

As this master's thesis showed some applications of using Web Science to investigate the area of Computer Science bibliography information, with the gained data further analyses can be performed.

- **Investigation of the co-authorship per country:** We already have seen that the co-authorship tends to grow in the sense of an increasing amount of publications written by more than one author. Now this analysis could be further extend to analyze this information in a geographical context of countries or regions. With this approach relationships between the co-author amount and regions can be determined, which inherently could reflect the funding system of such regions or countries by imposing a proper structure.

- **Analysis of the loyalty of authors for conferences:** A relationship between the location of a conference and the authors publishing there can be observed in the data. Subsequently, we can analyze the loyalty of authors for conferences. This means is it more likely that an author publishes at a conference he or she published before by additionally considering the classification of the conference as top or nontop.

- **Investigation of the thematic impact on authors career length:** The created dataset $Base_{cluster}$ also provides the opportunity to investigate if a bias between the first subfield a scientist is working on and the length of his or her career can be observed. The data can be analyzed with respect to career length, or another to be defined measure for a successful career, and classification information. Another approach could be to look at popular topics per country and the career length of authors in this country, as it can be assumed that popular topics are longer investigated and also passed from professors to students.

- **Trends in topics:** The dataset also offers the opportunity to create a time series regarding the amount of publications in a subfield per year, and hence a trend of popular topics can be expressed.

Beside further analyzing the dataset already created, modifications of the approach used in this work could help to improve the quality and accuracy of this data.

- **Extending dataset:** Beside the used bibliography citation datasources, the data can be extended with additional ones like CiteSeerX. This would lead to a better coverage of the different areas of Computer Science and hence reduce the drawback of differently covered subfields in DBLP.

- **Automatization:** The approach in this work was done by a lot of manual steps, at least some of which could be automated. For example the MAS offers since July 12, 2011[1] an API. So it was released after this dataset was created. This API has the possibility to query for authors, organizations, domains, publications, conferences, journals and keywords and therefore an automated version of the approach could be realized, to analyze the most current data.
- **Semantic access to data:** The created dataset can also be published by an semantic endpoint, like a D2R server, to provide a source for people who want to visualize or perform further analyses on the data.

[1] `http://social.microsoft.com/Forums/en-US/mas/thread/9a23b2d6-6599-4853-acf5-c1692a64365e` (accessed 28-October-2011)

Bibliography

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: a nucleus for a Web of open data. The Semantic Web (4825):722–735, 2007. [Cited on pages 19 and 20.]

[2] S. Auer and J. Lehmann. What have Innsbruck and Leipzig in common? extracting semantics from wiki content. In *Proceedings of the 4th European conference on The Semantic Web: Research and Applications*, page 503–517, 2007. [Cited on page 2.]

[3] P. Berkhin. Survey of clustering data mining techniques. page 25–71. Springer-Verlag, Berlin/Heidelberg, 2006. [Cited on page 37.]

[4] T. Berners-Lee. Information management: a proposal. http://cdsweb.cern.ch/record/369245, 1989. (accessed 7-September-2011). [Cited on page 1.]

[5] T. Berners-Lee. WWW: past, present, and future. *Computer*, 29(10):69–77, October 1996. [Cited on page 19.]

[6] T. Berners-Lee, W. Hall, J. Hendler, and D. J. Weitzner. Creating a science of the Web. http://journal.webscience.org/2/, August 2006. (accessed 22-March-2011). [Cited on pages 2 and 17.]

[7] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 2001. [Cited on pages 2, 19, 20, and 21.]

[8] M. Biryukov and C. Dong. Analysis of computer science communities based on DBLP. In *Proceedings of the 14th European conference on Research and advanced technology for digital libraries*, ECDL'10, page 228–235, Berlin, Heidelberg, 2010. Springer-Verlag. ACM ID: 1887792. [Cited on pages 4, 11, 12, 13, 15, 41, 56, 68, and 80.]

[9] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - a crystallization point for the Web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7:154–165, September 2009. ACM ID: 1640848. [Cited on pages 20 and 21.]

[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998. [Cited on page 2.]

[11] Vannevar Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, 1945. [Cited on page 30.]

[12] CCHS-CSIC. Top 500 Webometrics ranking of world universities july 2011. http://www.webometrics.info/Webometrics library/Top 500 Webometrics Ranking of World Universities July 2011.xls, July 2011. (accessed 14-October-2011). [Cited on page 50.]

[13] C. W. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. *ASLIB Cranfield project, Cranfield*, 1966. [Cited on page 30.]

[14] IBM Corporation. IBM Word Cloud Generator. http://www.alphaworks.ibm.com/tech/wordcloud, 2011. (accessed 18-May-2011). [Cited on pages 55 and 57.]

[15] Microsoft Corporation. Help center - Microsoft Academic Search. http://academic.research.microsoft.com/About/Help.htm, 2011. (accessed 8-October-2011). [Cited on page 46.]

[16] Microsoft Corporation. Microsoft Academic Search. http://academic.research.microsoft.com, 2011. (accessed 9-October-2011). [Cited on page 47.]

[17] Microsoft Corporation. VisualExplorer - Microsoft Academic Search. http://academic.research.microsoft.com/VisualExplorer, 2011. (accessed 8-October-2011). [Cited on page 14.]

[18] Cybermetrics Lab CSIC. Ranking Web of world universities: Home. http://www.webometrics.info/, 2011. (accessed 21-October-2011). [Cited on page 50.]

[19] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69:131–152, April 2006. [Cited on page 47.]

[20] E. Elmacioglu and D. Lee. On six degrees of separation in DBLP-DB and more. *ACM SIGMOD Record*, 34:33–40, June 2005. ACM ID: 1083791. [Cited on pages 11, 13, 15, 61, and 85.]

[21] D. Fiala. Mining citation information from CiteSeer data. *Scientometrics*, 86:553–562, March 2011. ACM ID: 1938397. [Cited on pages 4 and 16.]

[22] International Organization for Standardization. ISO - maintenance agency for ISO 3166 country codes. http://www.iso.org/iso/country_codes.htm, 2011. (accessed 21-October-2011). [Cited on page 50.]

[23] A. Geuna. The changing rationale for European university research funding: Are there negative unintended consequences? *Journal of Economic Issues*, 35(3):607–632, 2001. ArticleType: research-article / Full publication date: Sep., 2001 / Copyright © 2001 Association for Evolutionary Economics. [Cited on page 10.]

[24] M. Gibbons, C. Limoges, and H. Nowotny. *The new production of knowledge: the dynamics of science and research in contemporary societies*. SAGE, 1997. [Cited on page 10.]

[25] Avner Greif. Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy*, 102(5), 1994. [Cited on page 9.]

[26] V. N. Gudivada, V. V. Raghavan, W. I. Grosky, and R. Kasanagottu. Information retrieval on the World Wide Web. *Internet Computing, IEEE*, 1(5):58–68, 1997. [Cited on page 32.]

[27] A. Gulli and A. Signorini. The indexable Web is more than 11.5 billion pages. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, WWW '05, page 902–903, New York, NY, USA, 2005. ACM. ACM ID: 1062789. [Cited on pages 1 and 42.]

[28] G. K. Gupta. *Introduction to data mining with case studies*. PHI Learning Pvt. Ltd., 2006. [Cited on pages 33, 34, 35, 36, 37, and 38.]

[29] W. Hall, D. De Roure, and N. Shadbolt. The evolution of the Web and implications for eResearch. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890):991 –1001, March 2009. [Cited on pages 2 and 19.]

[30] T. Heath, T. Berners-Lee, and C. Bizer. Linked Data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009. [Cited on page 21.]

[31] J. Hendler, N. Shadbolt, W. Hall, T. Berners-Lee, and D. J. Weitzner. Web Science: an interdisciplinary approach to understanding the Web. *Communications of the ACM*, 51:60–69, July 2008. ACM ID: 1364798. [Cited on pages 17, 18, and 19.]

[32] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, November 2005. PMID: 16275915 PMCID: 1283832. [Cited on pages 47 and 85.]

[33] ACM Inc. The 1998 ACM Computing Classification System - Association for Computing Machinery. http://www.acm.org/about/class/ccs98-html, 2010. (accessed 4-October-2011). [Cited on page 42.]

[34] ACM Inc. Results. http://dl.acm.org/results.cfm, 2011. (accessed 8-October-2011). [Cited on page 44.]

[35] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972. [Cited on pages 31 and 56.]

[36] G. Klyne and J. J. Carroll. Resource Description Framework (RDF): concepts and abstract syntax. http://www.w3.org/TR/rdf-concepts/, 2004. (accessed 4-October-2011). [Cited on page 19.]

[37] W. Knelangen. Soziologische und kulturwissenschaftliche Beiträge zur Integrationstheorie. Technical report, Mimeo, 2001. [Cited on page 9.]

[38] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media meets Semantic Web - how the BBC uses DBpedia and Linked Data to make connections. In *The Semantic Web: Research and Applications*, volume 5554. Springer-Verlag, Berlin, Heidelberg. [Cited on page 21.]

[39] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 191–202. ACM Press, 1993. [Cited on pages 27 and 29.]

[40] A.H.F. Laender, C. J. P. de Lucena, J. C. Maldonado, E. de Souza e Silva, and N. Ziviani. Assessing the research and education quality of the top Brazilian computer science graduate programs. *ACM SIGCSE Bulletin*, 40:135, June 2008. [Cited on page 15.]

[41] F. W. Lancaster. *Information Retrieval Systems: Characteristics, Testing, and Evaluation*. John Wiley & Sons Inc, January 1969. [Cited on page 30.]

[42] S. Lawrence. Free online availability substantially increases a paper's impact. *Nature*, 411(6837):521, May 2001. [Cited on page 2.]

[43] M. Lesk. The seven ages of information retrieval. In *Proceedings of the Conference for the 50th anniversary of As We May Think*, page 12–14, 1995. [Cited on page 30.]

[44] M. Ley. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *Proceedings of the 9th International Symposium on String Processing and Information Retrieval*, SPIRE 2002, page 1–10, London, UK, 2002. Springer-Verlag. [Cited on page 44.]

[45] C. Manning, P. Raghavan, and H. Schütze. *An introduction to information retrieval*. Cambridge University Press, Cambridge, 2008. [Cited on pages 25, 26, 27, 31, 32, and 37.]

[46] D. Merkl. Information search on the Internet, 2010. http://www.ec.tuwien.ac.at/~dieter/ISI-complete-slides-2010.pdf (accessed 16-October-2011). [Cited on pages 25, 31, and 32.]

[47] H. Nowotny, P. Scott, and M. Gibbons. *Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty*. Blackwell Publishers, February 2001. [Cited on page 10.]

[48] M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14:130–137, 1980. [Cited on pages 27 and 28.]

[49] OECD Publishing. *Education at a Glance 2011*. Organisation for Economic Co-operation and Development,, Paris :, 2011. [Cited on page 71.]

[50] E. Rasmussen. Clustering algorithms. *Information Retrieval*, page 419–42, 1992. [Cited on pages 32, 33, and 34.]

[51] F. Reitz and O. Hoffmann. An analysis of the evolving coverage of computer science subfields in the DBLP digital library. In *Proceedings of the 14th European conference on Research and advanced technology for digital libraries*, ECDL'10, page 216–227, Berlin, Heidelberg, 2010. Springer-Verlag. ACM ID: 1887791. [Cited on pages 4, 11, 15, and 85.]

[52] Thomson Reuters. Calais submission tool | OpenCalais. http://www.opencalais.com/documentation/calais-submission-tool, 2011. (accessed 8-October-2011). [Cited on page 22.]

[53] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33:294–304, 1977. [Cited on page 32.]

[54] S. E. Robertson and K. S. Jones. Simple, proven approaches to text retrieval. *University of Cambridge Computer Laboratory Technical Report*, 356, 1994. [Cited on page 32.]

[55] C.F. Rusbult. Cultural influence in science - causes & effects. http://www.asa3.org/ASA/education/science/cp2.htm, 1997. (accessed 29-March-2011). [Cited on pages 10 and 11.]

[56] C.F. Rusbult. Culture and science - cultural influences & effects. http://www.asa3.org/ASA/education/science/cp.htm, 1997. (accessed 22-October-2011). [Cited on page 11.]

[57] C.F. Rusbult. *A Model of Integrated Scientific Method.* PhD thesis, University of Wisconsin, 1997. [Cited on page 10.]

[58] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, November 1975. [Cited on page 32.]

[59] N. Shadbolt and T. Berners-Lee. Web Science emerges. *Scientific American*, October 2008. [Cited on page 17.]

[60] A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43, 2001. [Cited on pages 30 and 32.]

[61] K. Taylor, R. Gledhill, J.W. Essex, J.G. Frey, S.W. Harris, and D. De Roure. A semantic datagrid for combinatorial chemistry. page 8 pp. IEEE, 2005. [Cited on page 19.]

[62] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, June 1998. [Cited on pages 13 and 14.]

[63] M. Wick. GeoNames. http://www.geonames.org/, 2011. (accessed 24-March-2011). [Cited on page 49.]

[64] O. R. Zaiane, J. Chen, and R. Goebel. DBconnect: mining research community on DBLP data. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, page 74–81, New York, NY, USA, 2007. ACM. ACM ID: 1348558. [Cited on pages 4, 15, and 16.]