Methods for Extracting Meta-Information from Bibliographic Databases

Maria Biryukov

19 October 2010

Acknowledgment

I am glad for an opportunity to express my gratitude to the people without whom this work would not be possible. First of all to the supervisor of my PhD thesis, Christoph Schommer, for his attention, insightful comments, advice and guidance. To my co-authors, Yafang Wang and Cailing Dong for their interest, enthusiasm and hard work. To Michael Ley for providing me with a DBLP parser and for the fruitful discussion of my first work related to DBLP. To Sascha Kaufmann for his help with the database setup and for fun to work together. To all members of MINE team for the very nice working atmosphere. To Pascal Bouvry, Pierre Kelsen, Alain Krief, Michael Ley and Christoph Schommer for having kindly agreed to read my work and for being on my PhD. committee. Finally I am most grateful to my family and especially to my husband Alex for his sharp and useful remarks and great support, and to my daughters, Alice and Sasha for providing constant but pleasant distraction from my work.

Contents

List of Tables

Abstract

Due to intensive growth of the electronically available publications in the last few decades, bibliographic databases have become widespread. They cover a large variety of knowledge fields and provide a fast access to the wide variety of data. At the same time they contain a wealth of hidden knowledge that requires steps of extra processing in order to infer it. In this work we focus on extraction of such implicit (or meta) knowledge from the research bibliographic databases by looking at them from sociolinguistic, text mining and bibliometric perspectives. We choose the Digital Library and Bibliographic Database — DBLP as a testbed for our experiments.

In the framework of the sociolinguistic analysis we build a statistical system for the language identification of personal names. We show also that extension of a purely statistical model with the co-authors network boosts the system's performance. There are several premises motivating our work. For example, it has been shown that the geographical proximity influences research. Moreover, research is constantly evaluated on the national and international basis. To make these and similar investigations less laborious in terms of human effort, ability to automatically assign personal names to the appropriate language seems to be useful.

In the text mining scenario, we perform a number of experiments that focus on topic identification and ranking. While our topic detection approach remains generic and can be used for any kind of textual data, the topic ranking metrics are built upon the information provided by the bibliographic databases. With respect to the topic ranking, our study aims at finding the ways of different topic ordering depending on the question that has

to be addressed: in some cases we have to know what are the "hot" research directions that attract a wide audience, while in some other cases, narrow and probably highly specialized topics have to be put forward. Our results show that one can achieve this goal by varying the type of bibliographic information used to compute the topic rank.

It is popular nowadays to bring technics from bibliometrics and scientometrics into the world of bibliographic databases to investigate the collaboration patterns and explore mechanisms underlying community development. The goal of our bibliometric study is to create a researcher's profile on DBLP and analyze some of the research communities defined by the different conferences, in terms of the publication activity, interdisciplinarity of research, collaboration trends and population stability. We also aim at exploring to what extent these aspects correlate with the conference rank.

Each of the above topics constitutes a method of meta information extraction from bibliographic databases. Moreover they can be used in combination in order to provide better results and in order to obtain a multidimentional view on research and research communities. Such techniques can also be applied to other similarly structured data sources.

Chapter 1

Introduction

In July 1945, Vannevar Bush [21] published an article *"As we may think"* where he described a *Memex* — "device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility." Memex is often mentioned as a kind of prototype of what we call nowadays *digital library* [27]. Informally digital libraries are very much like traditional libraries with the exception that their collections are digitally stored and electronically accessible [52]. These are the two crucial features that make digital libraries not only a useful service but also an object of multidisciplinary research that brings together scientists from many areas. Each area suggests its own view on the essence of digital libraries [105]:

- For library science, they are an ongoing trend toward library automation;
- For wide-area information service providers they are a Web application;
- For those who work on web technology development, they constitute a particular realization of hypertext methods;
- For information retrieval, digital libraries are a type of information retrieval system.

Although digital libraries are indeed a combination of all these features, for the purpose of the current study we will focus on the structural aspect that constitutes a core of any digital library, namely *bibliogrpaphic databases*.

These are defined as electronic collections of references to printed and non-printed materials, such as books, journal and newspaper articles, proceedings of the conferences and papers, technical reports, legal documents, patents, sound and video tracks. They consist of *bibliographic records* that give a uniform description of a document specifying *author(s) name(s), publication's title, year, and source* (adopted from [43, 118]).

Bibliographic databases span a large variety of knowledge domains, from humanities and social sciences to computer and life sciences, and differ in the amount of information covered by their records. Thus our description of a bibliographic record above refers to only the *basic elements* found in all well-formed records. Examples of more extended ones would include among others: *language*, *author's keywords*, *thesaurus descriptors*, *citations*, *abstract*, author's *affiliation* and even *full text*. Table 1.1 gives some examples of the bibliographic databases from different domains and of different coverage.

Note, that independent of the database topic and extent, the basic information about the document is always present. This is an important feature that ensures common ground between all the bibliographic databases and enables generic search strategies and analytical approaches. On top of it, the complexity of the investigation would obviously be conditioned on the amount of data offered by a particular database.

Bibliographic databases can be seen in two main ways: as source of information per se, and as research object in their own right. The first view corresponds to the definition of the bibliographic database that we gave earlier. For example, when we look up a paper in ACM digital library, ACM

[1] http://librarians.acm.org/digital-library

[2] http://www.csa.com/factsheets/assia-set-c.php

[3] http://www.rsc.org/Publishing/CurrentAwareness/aa/

[4] http://thomsonreuters.com/products_services/science/science_products/a-z/arts_humanities_citation_index

[5] http://www.bioone.org/

[6] http://citeseerx.ist.psu.edu/

[7] http://www.nova.edu/library/dils/lessons/compendex/

[8] Indexing, abstracts and full texts are available in approximately half of the cases.

[9] DBLP: http://www.informatik.uni-trier.de/~ley/db/

[10] http://ieeexplore.ieee.org/

[11] http://www.theiet.org/publishing/inspec/

[12] LNCS: http://www.springer.com/computer/lncs?SGWID=0-164-0-0-0

[13] http://www.apa.org/pubs/databases/psycinfo/index.aspx

[14] http://www.ncbi.nlm.nih.gov/pubmed

Table 1.1: Examples of bibliographic databases and their properties

Name	Area	Basic Info	Affiliation	Keywords	Descriptors	Citations	Abstract	Full text	Availability
ACM[1]	Computing	+	+	+	+	+	+	for subscribers	subscription for full access
ASSIA[2]	Soc. Sciences	+	-	-	+	-	+	-	by subscription
Analytical Abstracts[3]	Chemistry	+	+	-	+	-	+	-	by subscription
Arts and Humanities Citation Index[4]	Arts and Humanities	+	+	+	+	+	+	-	by subscription
BioOne[5]	Biology, Environment Sciences	+	+	+	-	+	+	+	free
CiteSeerX[6]	Computer and Information Science	+	-	-	-	+	+	+	free
Compendex[7]	Engineering and Technology	+	-	-	partially[8]	+	partially	partially	by subscription
DBLP[9]	Computer Science	+	-	-	-	-	-	-	free
IEEE Xplore[10]	CS, Electrical Engineering, Electronics	+	+	-	for subscribers	for subscribers	+	for subscribers	subscription for full access
Inspec[11]	Physics; Engineering; IT	+	+	+	+	-	+	-	by subscription
LNCS[12]	Computer Science	+	+	sometimes	-	sometimes	+	for subscribers	subscription for full access
PsychINFO[13]	Psychology	+	+	+	+	sample only	+	-	by subscription
PubMed[14]	Medicine, Life Sciences, Biomedicine	+	+	-	-	Related articles	+	sometimes	free

functions as a source of bibliographic information. On the other hand, when we want to know whether a certain author tends to publish alone or rather works in collaboration with other authors, we can use the information available from the ACM database to infer the answer, although it is not explicitly stated. In this case ACM becomes an object of extra processing and analysis. In this work we concentrate on the extraction of such hidden (or meta) knowledge from the bibliographic databases.

Analysis of bibliographic databases may help to address a wide range of questions. The question types depend on who asks them, and why. Let us give a few examples.

For **linguists**, a bibliographic database is a collection of texts written in multiple languages. The textual information comes from the authors' names, document titles, and also abstracts and full texts, if they are available. Language identification of these texts is one of the possibilities for linguistic analysis. In the context of collaborations between the authors (researchers) from around the world it might be useful to know where do the authors come from. The wealth of texts allows also for style analysis and in particular, can serve as data for the attitude mining in scientific texts — topic that has been attracting increasing attention in the last few years [85, 129, 145].

Text mining approach brings the analysis up to the semantic level and helps to discover meaningful connections between the textual elements. From this point of view the bibliographic database is a collection of topics and topic keywords in various domains. Hence, one of the tasks is to identify these topics, assign keywords to them and eventually incorporate other information in order to put the topics into a larger context. For instance, one may talk about the topic evolution and trends by taking temporal information into account. Association of the extracted topics to the authors' names will yield author – topic pairs. The venue names will add one more dimension to this analysis and allow for finding events related by both – topics and researchers.

From **graph-theoretic** perspective, elements of the bibliographic records can be connected into diverse graphs, such as a graph of co-authors, a graph of citations, a graph of authors and publications, and many others. Each graph type (or eventually a combination thereof) will help to answer a certain question. For instance, co-authorship graph is an informative source for studying *communities* — social groups (or *networks*) formed, in this particular case, by authors that work in the same or closely related area(s). It may shed light on how community membership evolves over time and whether there exists an overlap between the communities. Narrowing the view from

community to individual researcher's level, we might turn to a citation graph and evaluate the author's contribution to his research field, identify the most prominent ones and even compare citation practices across different scientific areas.

Bibliographic databases are targeted by the **data visualization** community. Here abundance of data allows for the development and fine-tuning of algorithms specially tailored for handling very large graphs [121]. At the same time visualization constitutes yet another approach to the data exploration aiming to reveal relationships between its various elements ranging from text entities [76] to citations and co-authorship [12].

In **bibliometrics and scientometrics** the wealth of bibliographic data is explored with the aim of research's output quantification, citation impact analysis, evaluation of scientists, venues and academic institutions, and understanding of the mechanisms underling the spread of scientific influence.

The final goal of all these approaches — taken separately or in combination, is to provide insight into the structure of research communities, identify processes underlying their development, measure the extent to which various research areas are similar to each other and what is unique to the specific fields. Within this large, far from being exhaustive, framework we have chosen to focus our attention on the area of computer science and explore the data available from the comprehensive computer science bibliography known as *Digital Bibliography and Library Project — DBLP* [82] (see also Table 1.1). We examine it in the contexts of: a) sociolinguistics; b) text mining, and c) research communities. In the first scenario we deal with the language identification of personal names. Although the cultural and geographical proximity seem to influence the researcher's communication [17, 72, 111], linguistic study of personal names remained beyond the scope of the bibliographic database investigation so far. In the second scenario we address the problem of topic detection and ranking. Various techniques have been proposed and some applications — real or prototype — have been built based on various topic extraction methods [34, 159]. It seems though that certain aspects of the topic treatment have not been given sufficient attention yet — such as for example differentiation between the topics, concepts and figures of speech typical for the technical writing. In the third scenario we combine an analysis of scientists and scientific communities with the elements of scientific venue evaluation. Both of these directions are actively explored nowadays albeit often considered independent of each other.

Our investigation is modular. On the one hand, each of the three perspectives constitutes a stand alone research topic as well as a method for extracting knowledge that is not explicitly present in the bibliographic databases — that is, meta-knowledge. On the other hand they all contribute to an extended multi-faceted view on the research and research community. Technically, the basic information provided by the bibliographic records is sufficient for gathering the test data. Although presence of abstracts or full text in a database constitutes an advantage for the topic treatment, our results show that the titles alone provide the necessary minimum of information that makes the task completion possible. It follows that the above methods are generic and can be ported to any bibliographic database independent of its thematic focus and record's depth. Moreover, they can also be applied to the databases that are not formally defined as bibliographic but have a similar structure, for example an International Movie Database, IMDB [65].

Let us now introduce DBLP in somewhat more details since its data is used in all our experiments. It has started in 1993 as a "test of Web technology" [82], and contained a server that gave access to the tables of content (TOCs) of important proceedings and journals in the fields of database systems and logic programming. Hence the first server name "Data Bases and Logic Programming". The service turned out to be useful for others and has evolved into a digital library that contains nowadays more than 1.4 million publications [83] coming mainly from the conference proceedings and journals, although other material types such as books, PhD and Master theses are also included. The thematic coverage has broadened too, and the actual database collection encompasses (in different proportions though) 27 subfields of computer science according to the thematic division introduced by Laender et al. and Martins et al. in 2008 and 2009 respectively [117].

In terms of structure, DBLP is a bibliography hypertext that consists of XML bibliographic records imported into the tables of content. An example of XML record is represented in Figure 1.1. The topmost field, "inproceeding" indicates that the record refers to a publication that appeared in the conference proceedings. Note the presence of the basic fields, namely "author", "title", "year" and "booktitle" — the last one stands for the source of the document. In the case of conference publications the source is the conference name, which in this particular example happens to be SPIRE. The record reflects the hypertext organization of DBLP: the field labeled "ee" is a link to the web page of the publication in the conference proceedings, and the field labeled "crossref" is an internal link to the web page of the

```
<inproceedings mdate="2002-09-19" key="conf/spire/Ley02">
<author>Michael Ley</author>
<title>The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives.</title>
<pages>1-10</pages>
<ee>http://link.springer.de/link/service/series/0558/bibs/2476/24760001.htm</ee>
<year>2002</year>
<crossref>conf/spire/2002</crossref>
<booktitle>SPIRE</booktitle>
<url>db/conf/spire/spire2002.html#Ley02</url>
</inproceedings>
```

Figure 1.1: Example of a DBLP XML bibliographic record

SPIRE's table of content of the given year in DBLP. The web interface of DBLP has a richer hypertext structure than an XML bibliographic record can capture. Thus, each author's name is hyperlinked to the page with all his / her publications and a complete hyperlinked list of his / her co-authors. The author's page in its turn allows to narrow ("refine") the view by either co-author name, publication venue or publication year. The effect of such organization of the data is that by following the hyperlinks a user navigates through the co-authorship, and author–title–venue graphs.

It is interesting to observe how the evolution of DBLP from a web application to its current state of an online hypertext bibliography instantiates a digital library from all the points of view that we have introduced in the beginning of this chapter. Our choice of DBLP is due to its open access, high accuracy and wide coverage of the computer science domain. As opposed to some other freely available sources like CiteSeerX (see Table 1.1) or Google Scholar[15] that automatically crawl Web for the information [20, 50], maintenance of DBLP is done with the substantial human effort which increases the consistency and quality of data [84]. Other options could include ACM Digital Library or Springer Lecture Notes in Computer Science (LNCS) series (see Table 1.1). Their strength is in presence of publications' abstracts and citation information (the last one is more relevant to the ACM Digital Library while citation lists on LNCS are irregular and not always freely avail-

[15] `http://scholar.google.com/` For the sake of precision note that Google Scholar is not a bibliographic database. It is a search engine that links to the sources holding the reference to the retrieved documents (and possibly the documents themselves). However the format Google Scholar uses to display the search results corresponds to the content of the basic bibliographic record. Therefore it can also be used for the bibliographic data exploration.

able). However in terms of coverage of the computer science materials they constitute only a subset of the information provided by DBLP.
The remaining part of this document is organized as follows: in Chapter 2 we present our approach to the language classification of the personal names in DBLP; Chapter 3 is devoted to the topic analysis; in Chapter 4 we discuss several properties of the computer science communities and conferences; Chapter 5 summarizes the work, presents the conclusions and offers a discussion on the future research options.

Chapter 2

Language Classification of Personal Names

In this chapter we consider bibliographic databases as an object of sociolinguistic research. We address the question of *"where do the authors come from?"* via language identification of the author's names. This is a two-steps process which involves primary classification based on the statistical models of languages, and classification refinement achieved with the analysis of the co-author network built from the bibliographic records. A system for automatic language identification presented here handles 14 different languages and requires no dictionary of names for training. The statistical models are built from the general purpose corpora for all Western European, Chinese, Japanese and Turkish languages. The system is fine tuned to achieve precision and recall above 90% for many languages, and provides better performance than some other systems aiming at the language identification of personal names. Tests on the DBLP data set have shown that the extension of the language model with the co-author network helps to improve classification results, especially in cases of closely related languages and mixed names. They have also demonstrated the usability of the system in applications such as data cleaning and trends detection.

2.1 Introduction and Related Research

With the constant growth of the volume of electronic publications, bibliographic databases and digital libraries become widespread. In Chapter 1 we

have pointed out their dual role of the re-searchable objects and introduced some of the research questions that have been or could be addressed based on their content. While authors are essential building blocks of the bibliographic records, the question *"where do the authors come from?"* does not seem to attract much attention so far. In [136], an attempt to capture the geographical background of the papers published in SIGIR[1] conferences is reported. The research is based on the rich data contained in the SIGIR's proceedings over the last three decades. However the name itself may shed light on the author's origin. In this study we propose a method for language attribution of personal names. Typically this information is not present in the bibliographic databases. However it might be useful in a variety of applications. Besides the assessment of the geographical scope of publications and spread of scientific productivity, it boosts efficiency of names transliteration and spell-check.

Detection of the language of a given sample of text is a well studied problem [22, 37, 40, 53, 92, 120]. The language detection systems achieve high accuracy for texts more than 100 bytes. However language identification of personal names remains a challenge because names are typically very short: from 2 and up to a few dozen characters, with only 13 characters on the average[2].

Previous work in the name language identification has been done in the area of speech synthesis, where knowledge of the name origin can help to generate correct pronunciation of that name [23, 81, 87]. In this work we study the problem in the context of digital bibliographies and libraries, but it is relevant for any databases which keep track of personal names. Application of our tool to the online computer science bibliography DBLP achieves a twofold goal:

- On the one hand it reveals a high number of names which cannot be unambiguously attributed to one language and thus affect the success rate of the tool. Consider for example the name "John Li": the first component suggests English, while the second one points to Chinese. In order for such names (*mixed names* thereafter) to be classified correctly the language model alone is not sufficient and additional knowledge is required. It could eventually be obtained from the external sources, for instance personal homepages or institute affiliations. Alternatively the

[1] SIGIR – Special Interest Group on Information Retrieval [137].

[2] This average was computed over more than 600,000 names in the DBLP database.

DBLP itself provides us with a kind of external knowledge coming in the form of co-author network which we examine to solve the problem of language assignement.

- On the other hand it serves as a testbed for two experiments which show potential usefulness of our system in the real life settings. In the first experiment it is applied to the data cleaning process, namely to the selection of the correct name spelling when multiple variations of the same name exist. The goal of the second experiment is to discover how the share of participation of different cultures in scientific publications has been evolving in the last 20 years.

This chapter is organized as follows: in Section 2.2 we describe the system and explain the language model used for the detection of the origin of a name. Evaluation of the results is demonstrated in Section 2.3. Section 2.4 introduces the name language classification approach enhanced with the co-author network analysis. Results of the evaluation of the refined method are presented in Section 2.5. In Section 2.6 we discuss the applications of our tool to the data cleaning and scientific trends discovery in DBLP. We conclude the chapter by a summary of the results in Section 2.7.

2.2 Finding origin of name: language based approach

2.2.1 System overview

The language detection system we have built, consists of a set of corpora and a set of metrics for the estimation of the probability that a character string A belongs to a language L. While the system is applied to the personal name language detection, the string A is not limited to represent a name. Rather it can be any valid string in some language L. The overlapping n-gram[3] model with $n = 4$ is chosen to represent both – the corpora and the names to be labeled. It is based on the assumption that the n-grams and their frequencies are specific to any given language and thus may serve as

[3]The term n-gram refers to the sequence of n characters, where $n \geq 1$. The word *overlapping* indicates that $n-1$ last characters of the $k_{th}-1$ n-gram are the first $n-1$ characters of the k_{th} n-gram.

a discriminating feature. The n-gram model of the language can be traced back to Shannon [130].

2.2.2 Corpus selection

Our system is trained to identify Chinese, Dutch, English, Finnish, French, German, Italian, Japanese, Portuguese, Spanish, a group of the three Scandinavian languages (Danish, Norwegian, Swedish), and Turkish. Except for the Chinese and Japanese, the texts of the training corpora come from the Wortschatz Corpora [156] and consist of sentences randomly selected either from newspapers or webpages. The Chinese corpus is constructed from a collection of various texts provided in the framework of the Gutenberg Project [116], and a cleaned segmented and romanized version of Chinese PH corpus – a collection of news-wire texts published by Xinhua News Agency in 1990-1991 [68]. A small Japanese corpus has been collected from the Internet and consists of literary works of Japanese authors converted into romaji romanization system.
Most of the languages listed above contain diacritic marks – signs added to letters in order to alter the pronunciation, mark the stress etc. Examples of such symbols are ê in French, {ä, ö, ü} in German, ñ in Spanish, or ø in Danish, etc. While the diacritics are sometimes replaced using certain conventions [22, 23], we do preserve the original character chart for any given language to avoid the loss of information which comes along with the substitution/replacement processes. The reason to keep the diacritics is that in bibliographic databases such as DBLP, or any Unicode based databases for which our language identification system could be applied, the diacritics would be preserved.

2.2.3 Calculation metric

For checking whether a string of characters $A = [a_0, a_1, \ldots, a_{l-1}]$ belongs to the language L we use the following formula:

$$P(A \in L) = p_L(a_0, a_1, a_2, a_3) \cdot \prod_{i=1}^{l-4} p_L(a_{i+3}|a_i, a_{i+1}, a_{i+2}).$$

Here, the probability p_L that the tetragram a_0, a_1, a_2, a_3 belongs to the language L is approximated by its frequency in the corpus of the language L,

divided by the size M of the corpus:

$$p_L(a_0, a_1, a_2, a_3) \approx fr_L(a_0, a_1, a_2, a_3)/M$$

and conditional tetragram probability is approximated as follows:

$$p_L(a_{i+3}|a_i, a_{i+1}, a_{i+2}) \approx \frac{fr(a_i, a_{i+1}, a_{i+2}, a_{i+3})}{fr(a_i, a_{i+1}, a_{i+2})}.$$

If we denote by $logFr$ the logarithms of the frequencies and normalize the result by the length of the string l, we get:

$$\log P(A \in L) = logFr(a_0, a_1, a_2, a_3) - \log M +$$

$$+ \sum_{i=1}^{l-4} logFr(a_i, a_{i+1}, a_{i+2}, a_{i+3}) - logFr(a_i, a_{i+1}, a_{i+2}).$$

$$CondTetrScore(A) = \frac{logP(A \in L)}{l}.$$

This metric estimates the probability that the string A belongs to the language L using the conditional tetragram approximation of the language. Tetragrams which occur < 3 times in the corpus are not considered, since their frequency may not be a close approximation of the real tetragram probability. For the n-grams that cannot be found or are infrequent in the language the default solution is to evaluate their weight to -1000 ("penalty"). It might be the case though that the corpus for that language is not sufficiently large to include all the possible n-grams that may occur in the names. To account for such cases, the n-gram is approximated by an $(n-1)$-gram (for example for the tetragrams):

$$\log P(a_{i+3}|a_i, a_{i+1}, a_{i+2}) \approx$$

$$\approx logFr(a_{i+1}, a_{i+2}, a_{i+3}) - logFr(a_{i+1}, a_{i+2}).$$

Building an n-gram model based on an $n-1$-gram model is called *backoff*, and is used for the n-gram smoothing [69][4]. At this stage if the $(n-1)$-gram

[4]An alternative could be a Laplace's sample size correction, as used in [37]. It has been pointed out however that this method tends to highly overestimate the probability of the unseen events at the cost of reducing the probability estimates of more frequent events [91], which reduces its attractiveness.

Table 2.1: Most frequent Italian tetragrams: names vs. text

Frequency rank	1	2	3	4	5	6	7	8	9	10
Names	*_mar*	*o_ma*	*ini_*	*i_ma*	*ano_*	bert	marc	erto	*rto_*	andr
Text	*_il_*	*del_*	*che_*	*_di_*	*_la_*	*_ato*	*con_*	*_lla*	*per_*	dell

is not found in the language the conditional tetragram is penalized. However before recurring to the $n-1$ approximation we check whether a tetragram in question exists with a certain minimal frequency in the other languages. If the n-gram is sufficiently frequent in at least one of the other languages, we give the penalty weight in the language which is currently being checked. This way we increase the discriminating power of the computational model.

2.2.4 Choosing a baseline

To assess the effectiveness of the proposed technique we have to choose a model that performs well in the similar task. One of the well known language identification systems has been suggested in [22]. It uses an n-gram language representation, where the n-grams to be considered are chosen based on their frequency distribution in the training corpora. For the classification task only N most frequent ones are taken into account and compared to the n-grams of strings to be assigned a language. The system has been applied to the language identification of strings from $300-1700$ bytes and achieved about 98% accuracy. One of the shortcomings of this approach is that it requires a correlation between the training and test corpora. Table 2.1 illustrates the idea: ten of the most frequent tetragrams obtained from the list of about 2000 Italian names is contrasted to the ten most frequent tetragrams generated from a general italian text of approximately the same size ($\approx$ 70KB). We see no overlap between the two lists. The n-grams in the right column come from auxiliary words (determiners, prepositions, etc.)— the ones which typically have the highest frequencies in the general texts. This distribution is not at all representative for the names and hence cannot be used to score them given that we train the system on the general corpora.
A more appropriate way for us would be to consider all possible n-grams of a certain length obtained from the corpus and assign each one a probability based on its relative frequency. These counters would serve as ground for name (word) scoring. Since this method has shown about 90% accuracy in the language classification of proper nouns [81] we adopt it as a baseline.

Thus we apply the formula:

$$\log P(A \in L) \approx \sum_{i=0}^{l-3} logFr(a_i, a_{i+1}, a_{i+2}) - \log M.$$

$$BaseScore(A) = \frac{\log P(A \in L)}{l}.$$

to calculate the probability that a string A belongs to the language L.

2.2.5 Workflow and testing

Our system is built in a way which allows an easy addition of new languages and new string evaluation metrics. The program takes as input a list of personal names for which their language origin has to be identified, and the parameter, which indicates the choice of the string evaluation metric. The system outputs separate files with the names attributed to each language, ranked by the metric of choice. For each name the second best choice is given, as well as the values of the metric across all the languages.
The system has been tested on 100 names for each of the 14 languages[5] as well as on the joint list of 1400 names, all collected from the Wikipedia people lists [154]. The first setting allows us to accurately assess the recall and precision achieved by the system when given a monolingual set of names. The second setting approximates the "real life" conditions of a database with a multilingual set of names.

2.3 Experiments and Evaluation

Table 2.2 shows the results of the test runs obtained for each of the 14 languages. In this table each row corresponds to a monolingual test with 100 names, each column corresponds to the number of names from different lists attributed to the target language by our system. For example, the test against 100 Italian names assigns 94 names to Italian, 3 to Spanish, 1 to Turkish, and 2 to Portuguese. The values on the diagonal show the recall for the respective language labeling the row. We notice that recall for

[5]For the Scandinavian languages a combined list of 300 names, with 100 names for each of the three languages, has been used.

Table 2.2: Recall/precision for 100 name lists of 14 languages using tetragram metric.

100 names	Italian	English	Japanese	Dutch	French	Spanish	Scandinavian	Turkish	German	Portuguese	Finnish	Chinese
Italian	**94**	0	0	0	0	3	0	1	0	2	0	0
English	0	**67**	0	8	4	1	12	0	7	1	0	0
Japanese	0	2	**90**	0	0	0	2	1	3	0	2	0
Dutch	1	0	0	**85**	0	0	9	1	4	0	0	0
French	1	1	0	3	**90**	0	3	0	1	1	0	0
Spanish	3	2	0	0	0	**85**	2	0	0	10	0	0
Scandinavian	0	0	0	1	0	0	**95**	0	3	0	1	0
Turkish	0	1	0	0	0	0	0	**92**	7	0	0	0
German	0	0	0	1	0	0	4	0	**94**	0	1	0
Portugese	4	0	0	0	0	2	1	0	0	**93**	0	0
Finnish	1	0	1	1	0	0	2	0	0	1	**94**	0
Chinese	0	3	0	2	0	0	0	1	2	0	0	**92**
Total names(recall %)	**104(.94)**	**76(.67)**	**91(.90)**	**101(.85)**	**94(.90)**	**91(.85)**	**130(.95)**	**96(.92)**	**121(.94)**	**108(.93)**	**98(.94)**	**92(.92)**
Baseline (recall %)	105(.91)	90(.72)	98(.96)	116(.83)	97(.84)	105(.85)	86(.71)	90(.86)	134(.90)	108(.82)	96(.88)	100(.98)
Total precision %	**0.90**	**0.88**	**0.99**	**0.84**	**0.96**	**0.93**	**0.73**	**0.96**	**0.77**	**0.86**	**0.95**	**1.00**
Baseline precision %	0.87	0.80	0.98	0.72	0.87	0.81	0.82	0.96	0.67	0.76	0.92	0.98
F_1-measure	**0.92**	**0.77**	**0.94**	**0.84**	**0.93**	**0.89**	**0.83**	**0.94**	**0.85**	**0.89**	**0.94**	**0.96**
Baseline F_1 measure	0.89	0.76	0.97	0.77	0.85	0.83	0.76	0.91	0.77	0.79	0.90	0.98

Chinese, Finnish, French, German, Italian, Japanese, Portuguese, Scandinavian languages, and Turkish is $\geq$ 90%. It is above 80% for Dutch and Spanish. English language demonstrates only 67%, with many names being misclassified to Dutch, French, German and Scandinavian languages. With exception of French, all the other languages belong to the family of Germanic languages. Indeed, decision between related languages can be challlenging even for a human expert. As of French language, there are many words in English that have French origin, such as "art", "machine", "table" to mention a very few. Moreover, the survey of $10,000$ words taken from several thousand business letters has found that English language has 41% of words of French origin [155].

Language similarity constitutes the reason why we collapse Danish, Norwegian and Swedish into a single group of Scandinavian languages. Treating them on the individual basis affects both – recall and precision, that can be seen from Table 2.3. We observe also that for any of these three languages,

Table 2.3: Individual performance of Scandinavian languages

100 Names	Danish	Norwegian	Swedish	Other
Danish	70	24	1	5
Norwegian	14	77	6	3
Swedish	1	8	85	6

the second best score has almost always been a Scandinavian language. The difficulty in discrimination between closely related languages is caused by the overlap in n-grams and similarity of their respective frequencies. Develop-

ment of efficient algorithms for the classification of names from highly similar languages remains an interesting open problem.
We notice that the matrix is not symmetric. For instance, no English name was misclassified as Chinese, while 3 Chinese names were erroneously considered English. Similarly, no German name fell into the Japanese group, while 3 Japanese names were decided German. These figures may shed light on the nature of the corpus. We have seen that there are Chinese words (or names) in the English corpus, and Japanese in the German one, but not vice versa.
The bottom lines of Table 2.2 show the total recall and precision obtained on multilingual list of 1400 names from 14 languages. Figures for the conditional tetragram metric are given in bold. We compare them to the baseline performance in the following subsection. With regard to the overall performance, the recall typically does not change because the language-wise decision does not depend on the number of languages in the test set, and the number of names per category in the complete test set remains the same. Precision rate depends on the number of names erroneously assigned a given category in addition to the correctly classified names. Scandinavian languages attract the highest number of foreign names (mostly from the languages of the Germanic group), and thus have the lowest precision. Combined recall-precision F_1 measure[6] sums up the overall system performance. It weights evenly recall and precision, and is calculated by the formula:

$$F_1 = 2 \cdot (\mathrm{P} \cdot \mathrm{R})/(\mathrm{P} + \mathrm{R}).$$

where P and R stand for precision and recall respectively.

2.3.1 Comparison to the baseline and other systems

Recall, precision, and F_1 measure obtained with the baseline metric of unconditional trigrams are quoted at the bottom of Table 2.2 under the corresponding figures for the conditional tetragrams technique. We notice that the baseline has better recall for the Japanese and Chinese languages. This can be explained by the tendency of the conditional metric to overestimate the importance of rare events. Our frequency-based baseline technique has

[6] F_1 measure is a measure of test's accuracy. It can be seen as a weighted average of the recall and precision, where the recall and precision are considered equally important and the weight coefficient is 1 [91].

no such drawback. With regard to all the other languages it performs worse. Pairs of similar languages such as (Spanish, Portuguese) or (German, Dutch) especially suffer. The reason of the lower accuracy of the unconditional metric is that it considers the n-grams as independent units which does not model correctly the real situation in the language.

A work that has inspired the choice of the baseline technique reports $\approx 90\%$ accuracy when classifying three languages: English, Russian, and Arabic. One of the possible explanations of the high success might be that these three languages are very different and hence are easier for discrimination. Another interesting observation is that the overall system performance depends on the total number of languages to be considered. Results obtained by running our system with the varying number of languages, starting fromEnglish, Chinese, Japanese, and adding one language at a time, have shown that system's accuracy was gradually dropping from 0.98 to 0.87[7]. It shows that the task of classification into a large number of languages is quite challenging.

Comparing the techniques and systems, it might be of interest to compare some results demonstrated by our system with those described in [23]. It aims at the language classification of personal names, considers 4 languages (English, German, French and Portuguese), uses conditional trigrams obtained from the most frequent syllable units, and computes the probability of a name belonging to a language with the Bayesian decision rule. The results are given in terms of the confusion ratio which for every pair of languages represents the average percentage of mutually misclassified names. Results obtained from running of our system with exactly the same set of languages have shown that both systems suffer from confusion of English and German names. However, the percentage of confusion yielded by our system is slightly lower (6.5% vs. 7.6%). Several reasons might lead to this result. Our system is based on tetragrams which carry more information than the trigrams. In [23] the n-grams are chosen from a subset of the most frequent syllables which might not necessarily be the most representative for a language. In our system, n-grams are generated from the entire texts and thus we do not loose the information. Another interesting observation is that in [23], $< French, Portuguese >$ is the most confused pair (11.4%). In our system the percentage of confusion between these two is extremely low (0.5%). This result points to the importance of preserving the diacritic marks which turns to be a powerful discriminating feature.

[7]The values are expressed in terms of F_1 measure averaged across the languages.

Closing the evaluation section we would like to stress that in the previous works on the topic the systems had been trained on the corpora compiled from personal names [23] or containing substantial portion of personal names [87, 81]. Our system however uses the general purpose corpora. The results obtained by our system suggest that the general purpose corpora suits well to the task. This observation may have a practical advantage since the general purpose corpora might be easier to obtain.

2.4 Language detection using co-author network

So far, we have been analyzing the system's performance on the specially compiled test corpus. In this section we turn to the discussion of a number of issues which have arisen when running the system on the list of the author names extracted from the DBLP.

2.4.1 Assigning language to the authors: DBLP case

As we described in detail in Chapter 1, DBLP is a publicly available database which provides bibliographic information on major computer science journals and proceedings. The records typically list the co-author name(s), publication title, date and venue. For the first attempt of the language identification only personal names have been considered and processed in isolation from other information contained in the records. We run our experiments on the DBLP release from February 2008[8] which has listed 609411 personal names. To increase the accuracy of classification the system only deals with the names whose complete length is ≥ 4, which has amounted to 608350 names. While the system has shown promising results during the test runs, applying it to the DBLP brings out a number of differences between the settings:

- Language scope. Presumably DBLP contains names from much more languages than our system in its current state can handle (all Eastern-European, Indian, Korean, Arabic, etc.). To detect such names and avoid them from being randomly assigned to one of the existing categories, we adopt the following method:

[8]The up-to-date versions of DBLP are available for download from `http://dblp.uni-trier.de/xml/` in XML format.

Recall that the weight of a name in the language is determined by the frequency of its n-grams in that language. Hence, names from unknown languages are especially prone to penalties according to the "penalization policy" described in subsection 2.2.3. Should the name receive at least one penalty in all 14 languages, it is labeled "other" and is sent to the file which collects names from languages not covered by the system.

- Uncertain names. Even for the 14 languages the system deals with, the decision is not always unambiguous. In Section 2.3 we stressed one reason for the uncertainty, namely the language similarity. Names whose components are typical for more than one language constitute another reason. For instance "Robert" or "Charles" occur (and are written in the same way) in both, English and French, and assignment of the name "Charles Robert" to English or French is almost equally likely. In terms of the name scores, such cases would have a very small difference between the 1st and 2nd best choices, and thus the classification cannot be accepted with confidence[9]. Such names are assigned the language where they have gained the highest score, but labeled "uncertain".

- Mixed names. Mixed names are the ones, whose components belong to the different languages. For instance, in the name "Thomas Xavier Vintimilla" the first given name is English (Welsh origin), the second one – Spanish (Basque origin, written as Javier in modern Spanish, also popular in France, US), and the family name is probably Spanish. Mixed names do not necessarily have close 1st and 2nd best ranks, and hence are not always recognized as "uncertain". They are often misclassified.

To increase the system's performance in the real life conditions we enhance the model with the co-author network. The idea is that research is a communicative process and in order for the researchers to communicate they have to share some common ground. Are there factors besides the common research interests that would facilitate the collaboration? *Geographical distance*, for example, has been classified as a negative factor despite the high number of

[9]In the experiments described here decision is confirmed if the difference between the two highest scores ≥ 0.5.

remote research communications [72, 111]. A certain alikeness between the researchers — let it be the same *gender* or the same *research unit* people belong to, has proved beneficial in establishing collaborations [16]. Our assumption is that the same *language* is one of the commonalities that makes collaboration easier and that monolingual collaborations are more widespread than the multilingual ones. Thus, if a person whose name is labeled "uncertain" with the highest rank in Italian, has mainly Italian co-authors (as classified by the system), it can be identified as Italian with increased certainty. In the same spirit, misclassified names can be reassigned the most appropriate language category. Of course, this method is not a substitute for the languages that are not covered by our system. However it may help to correct the initial classification by transferring names erroneously labeled "other" to one of the languages known to the system based on the co-author list assignment. On the other hand, co-author classification serves as support for the author name classification, in case they agree.
Bellow we describe the application of co-author network to the personal name language classification in more details.

2.4.2 DBLP as a co-author network

To conduct the experiments we transform the DBLP into a network of co-authors represented by a graph $G = (V, E)$, where V is the set of vertices which correspond to personal names, and E is the set of edges which are defined by the co-authorship: there is an edge between two authors $\langle a, b \rangle$ if they have at least one common publication. Based on the DBLP data from February 2008, the network graph consists of 609411 vertices, and 3634114 edges. In average, there are 2.51 authors per publication, and 4.1 publications per author, out of which 3.69 are made in collaboration with the other authors. For every co-author b of an author a we calculate the relative strength of their co-authorship via the formula:

$$w_b(a) = \sum_{i=1}^{n} 1/(A_i - 1),$$

where A_i is the number of co-authors in the ith common publication of a and b, and n is the number of the common publications. There are on average 5.96 co-authors per author, and the co-authorship strength across the database is 0.63.

2.4.3 Enhancing language model with the co-author network

In this altered approach the language classification consists of three steps:

- Personal name language detection for every vertex in V. This step is done according to the procedure described in the subsection 2.2.3. The result is partitioning of the DBLP personal names into language categories, as described in the subsection 2.4.1.
- Verification of the initial classification. The objectif is to determine for every $a \in V$ the dominating language category of his/her co-authors.
- Refine the classification by merging the results of the two independent classifications (via linguistic structure of the name and via co-author network).

We have implemented three different methods of computing the language category of the co-authors.

2.4.4 Classification using probabilistic voting approach

This method represents a kind of "voting system", where each co-author votes for the language to which his personal name has been attributed by the first round of the classification process.[10] We will also describe other more refined models later in this section. Consider the following example: Suppose that out of 30 co-authors the highest vote for a single language (say, Italian) is 10. Is this a chance event or a strong bias towards Italian? In order to determine the threshold we propose the probabilistic method described below.
This method determines how much the probability of selecting one of the 14 languages by co-author voting is higher than a chance selection. We iterate over the co-authors b_i of $a \in V$, count for each language the number of co-authors that have been assigned to it, and determine the language with the largest counter c_{max}. We assume that the language counters are binomially distributed $B(n, p)$ with $p = 1/14$ (independent choice of one of the 14 languages) and n – being the number of co-authors of a. For some

[10] We consider all $a \in V$ that have ≥ 5 co-authors, and ≥ 3 works produced in collaboration, i.e. sufficient co-authorship strength. In total 131989 DBLP authors pass this criteria.

language the probability that the number X of co-authors assigned to it is $< c_{max}$ is expressed by:

$$P(X < c_{max}) = F(c_{max}; n, p),$$

where F is the cumulative function of the binomial distribution. This cumulative function can be evaluated using the regularized incomplete beta function, as follows:

$$F(c_{max}; n, p) = P(X < c_{max}) = I_{1-p}(n - c_{max}, c_{max} + 1),$$

provided that $0 < c_{max} \leq n$. Thus taking into account the 14 languages treated by the system we can compute the probability P that in some language the number of co-authors is higher than c_{max}, applying the formula:

$$P = 1 - I_{1-p}(n - c_{max}, c_{max} + 1)^{14}.$$

If $P < p_{min}$, we accept that having c_{max} co-authors voting for the same language is not a chance event. In our experiments p_{min} is set to 0.01. (We have checked other possibilities for p_{min}, from 0.02 to 0.05, and kept 0.01 as producing the most accurate results).
This model can be further refined due to the following observations:

- By using only a single vote per co-author we loose the possibly relevant infomation that is contained in the second best, third best, etc. languages proposed by our linguistic model. We can still accomodate this information by giving points to the top five languages for each person, with some decay factor. For example: a vote of 1 for the first language, a vote of 0.5 for the second, 0.25 for the third, etc. (decay factor 1/2). The reason why we work with points rather than with linguistic weights is that the later depend on the corpus size, frequency and the total number of the unique n-grams in the language. They are also influenced by the corpus frequency of names. Thus it makes no sense to compare absolute weight values across the languages.

- The second observation is that a co-authorship strength $w_b(a)$ varies between the co-authors and thus giving all the co-authors the same voting power may not be optimal. We may thus weight the vote of each co-author by his/her co-authorship strength with the target author a.

In all these methods we do not consider co-author names labeled "other" because they mainly belong to the languages not covered by the system. If all the co-authors of a given author are "others", the author is skipped.
Finally, we check whether or the language category suggested by the co-authors corresponds to the one obtained by the author in the first classification step. Results produced by this method are discussed in the following section.

2.5 Evaluation of the Results and Discussion

We apply the three methods described in Subsection 2.4.4 to the 131989 DBLP authors who satisfy the co-authorship strength criteria. From that list 100 names have randomly been chosen to assess the quality of the classification. Table 2.4 summarizes the results.

Category	True	False	Chance
Methods agree	36	0	–
Methods differ	37	5	–
Chance	–	–	22
Total	73	5	22

Table 2.4: Evaluation of the name language classification using co-author network.

We notice that 22 names out of 100 have not been classified because the language selection made by the co-author voting has been considered a chance selection by all the three methods. In the other 36 cases the language selected by the co-authors corresponds to the one initially attributed to the name by the linguistic method, and in 42 cases – the two classifications disagree. To check the correctness of these results we have searched for the information concerning the author's current or past affiliation. As the evaluation table suggests, the co-author based classification is true in most of the cases (only 5 errors out of 78 cases, i.e. above 90% success rate). The match between the linguistic and the co-author based classifications speaks for the hypothesis that people tend to collaborate within monolingual communities. The disagreement between the two usually occurs in one of the following scenarios:

- The name is classified with uncertainty or misclassified. For example in our test set there are 27 such names out of 42, and 7 among them

Table 2.5: Example of closely spelled Chinese names from DBLP.

Name	Score	Name	Score	Human Expert Evaluation
funping bu	(sw,-3.53)	fanping bu	(ch,-2.67)	1st is wrongly written
fugang li	(ch,-3.05)	fufang li	(ch,-2.97)	both
fouxiang shen	(ch,-2.35)	fuxiang shen	(ch,-2.27)	both, but 2nd is often used
fuyun ling	(ch,-2.85)	fuyung ling	(ch,-3.39)	2nd is not chinese
ge gao	(ch,-2.00)	ge guo	(ch,-1.76)	both
geng-dian huang	(ch,-1.46)	geng-dian hwang	(ch,-2.78)	2nd is not chinese
guang-sha qui	(fr,-2.91)	guang-sha qiu	(ch,-1.67)	1st is wrongly written

are initially labeled "other" while they actually fall into the scope of languages processed by the system. Due to the co-author based classification we could correct the initial assignment.

- Person works outside of his/her native linguistic environment (for example, in another country). We have encountered 10 such names out of 42 in our test set. In that case co-authors attribute the name to the language of community to which he/she contributes.

The technique-wise comparison shows that all the three methods usually produce the same language selection for a single author. However the method which takes into account the co-authorship strength $w_b(a)$ may select the language of the strongest co-author, if there is one. This feature makes it useful for discoverig special patterns in co-authorship, for example: $\langle professor, PhD-student \rangle$.

2.6 Personal name language detection at work: two experiments

In this section we show how a digital bibliographic database such as DBLP may benefit from the tool we have been describing in the previous sections.

2.6.1 Application to data cleaning

The widely known problem which affects the quality of the services provided by the bibliographic databases and digital libraries is the maintenance of

personal names. Personal names constitute the core of the bibliographic information. However due to the variety of sources from which the information is gathered for these databases the name spelling in the databases is not always consistent. There might be several variations of the same name in a database which results in multiple records that in fact belong to the same person. A lot of research that aimed at detection of the misspelled names had been done in the recent years [45, 57, 80, 119], and various string processing techniques had been suggested and tested [7, 25, 44] to list a few. However even when the set of strings that could be attributed to the same name has been detected, there remains a question: which one of these strings corresponds to the correct spelling of the name. An automatic database cleaning tool should be able to resolve this problem by itself since the end-user would unlikely be an expert in the foreign name spelling. We believe that our tool could be applied in this scenario. The underlying idea is that erroneously written names will contain n-gram(s) which either do not exist in a given language or are at least more rare than typical n-gram(s) of the target language. This will result in the lower overall score of the misspelled name. To test this hypothesis we have performed the following experiment. We first implemented an algorithm which allows us to identify all names in DBLP that are close in terms of the Levenshtein distance [55]. Then we combined this list with the results of the name language classification as described above. As a result, a subset of 100 name pairs with at least one Chinese name was selected and given to a native speaker for the evaluation. As can be seen from Table 2.5 names marked as "not Chinese" or "wrongly written", have been systematically receiving lower weight than the correctly spelled ones[11]. When both spellings are possible, the names have close weights and the gap between them is considerably smaller than the weight difference between a correct and misspelled names. Based on this score the correct spelling could automatically be suggested to the user. Such a tool can also be used to learn correct spelling of various names or find the most correct transliteration of the name.

[11]The two letters code in the parenthesis represents the language label assigned by the system. It is followed by the name weight.

2.6.2 Application to trend discovery

Automatic detection of trends in various aspects of scientific activities has recently become a popular research task. The language classification of personal names may bring a new dimension to the trend analysis. For example, one might be interested in popularity of various research topics among the nations. Or yet in combination with the time aspect, one might want to trace the evolution of topics in some scientific domain within certain time period and national group(s). Here we present a simple example of trends discovery in combination with the name language identification.

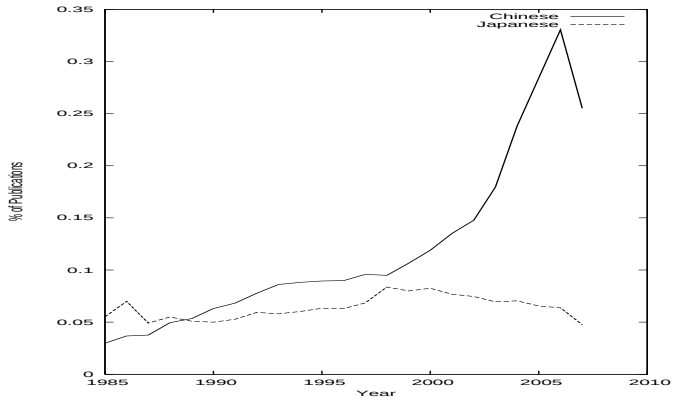

Figure 2.1: Discovering of national trends in computer science based on DBLP data from 1985-2007.

Suppose we are interested in learning how the share of participation of different cultures in scientific publications has been evolving with time. Figure 2.1 demonstrates this idea for the area of computer science based on the publications data from the DBLP. For the purpose of this example, we use publications produced by the Chinese and Japanese researchers since these languages are identified with the highest accuracy[12]. The curves are calculated from the number of documents attributed to a nation normalized by the total number of papers in the given year. We notice the constant growth in publication activity of Chinese, while Japanese demonstrate quite stable behavior with some small picks of activity around 1986 and $1997 - 1998$.

[12]It might be useful to combine the language and time data with the demographic information such as population counts when producing such general trends plots.

It is worth mentioning that rapid development of science in China over the past few decades have been reported in [75] whose investigation was based on the analysis of research papers and reviews along with their citations, available from the journals indexed by Thomson ISI[13]. One may also notice that the curves go down around 2006. This can be explained by the fact that the publication data of the last years has not been entered completely into DBLP yet.
Note that trends' discovery can be viewed as a particular assignment within the global task of the research evaluation at both national and international levels. Results of these evaluations have been shown to influence various aspects of the research management such as fonds allocation and scientific programs development [14, 18, 54, 96]. Similar to the trends' discovery, we believe that a system capable of the automatic personal name language identification could be applied in these scenarios.

2.7 Summary

In this chapter we have described a statistical system which performs personal name language classification using a general corpora for training. We have tested the system on a set of 14 languages which have included all the Western-European languages, Chinese, Japanese and Turkish. Our system has demonstrated high accuracy in terms of recall and precision for most of the languages when tested on the list of 1400 names. We have also tested our system on the collection of more than $600,000$ names taken from computer science bibliographical database, DBLP. Using the DBLP as a test bed we have shown how the initial, based on the language model only, classification can be improved via extending it with the co-author network built from the bibliographical records. Experiments with the data cleaning and trends discovery have demonstrated potential usefulness of our tool for the real life bibliographic databases such as DBLP.

[13]Thomson ISI, formerly known as *Institute of Scientific Information*, provides bibliographic services, citation indexing and analysis. It covers more than $14,000$ academic journals in dozens of languages, that represent a large variety of scientific fields ranging from exact science and engineering to social sciences and arts [146].

Chapter 3

Topic detection and ranking in bibliographic databases

This chapter is devoted to the investigation of the bibliographic content from the text mining point of view. Recall from Chapter 1 that topic detection constitutes one of the tasks and research questions being addressed by the text analysis of the bibliographic data. Here we do not only focus on the identification of topics but also motivate the relevance of various task-dependent topic ranking mechanisms. We demonstrate and discuss the results of some possible rankings.

3.1 Introduction and Related Research

"How to identify topics in the large amounts of textual data?" is one of the research questions that has been attracting serious attention since the very beginning of the automatic text processing. This interest is easy to explain since the detection of topic(s) is a precondition for many text-based operations such as summarization, abstracting, classification and clustering. As the research itself becomes an object of investigation, the new topic-centered questions arrise: a) what is in focus of scientific interest? b) What conference to choose to submit a paper? c) Whom to choose to work with? d) What are the prospective research areas for the investment? This list is not exhaustive.

To answer these and related questions various approaches have been tried. Kleinberg [77] proposes to identify topics as bursts of activity corresponding

to the appearance and disappearance of terms in a given research area. He uses sate transition in a Markov chain to implement the model and applies it to the DBLP document collection. Zaïne, Chen and Goebel [159] detect research topics based on the word frequency distribution in publication titles and apply the extracted topics to the analysis of the research communities on DBLP. Diedrich and Balke [34, 35] use the author-provided keywords as main building blocks for the automatically created hierarchical topic facets, and demonstrate how to apply them to the topic-based user modeling.

Tracking of topic dynamics and trends constitutes another active branch of the topic-related studies. Wang and McCallum [151] combine the word co-occurrences and document's time-stamp to identify topics and model topic behavior over time. Kanagasabi and Tan [71] pursue the similar goal via using self-organizing neural networks. Ke, Börn and Viswanath [74] study major research topics and trends as represented in ACM digital library from the data visualization perspective.

Yet another approach to the topic analysis couples textual information with the analysis of the co-author networks built from the bibliographic data. Rosen-Zvi et. al. [122] describe a generative probabilistic process of the topic modeling and extend the model so that the topic weights are determined by the authors of the documents. Steyvers et. al. [143] apply the same author-topic model to the collection of documents recorded in CiteSeer and generate author-by-topic and topic-by-author rankings as well as discover scientific trends in the period from 1990 to 2002. Mei et. al. [95] combine a text-based statistical technics for topic terms detection with the refinement of the initial topic assignments by the co-author network analysis. Zhou et. al. [160] generate a topic–author social network represented by a Markov chain, and use it in order to discover authors that have influenced emergence of various research topics.

In line with these works, our methods for topic identification and ranking rely on the text and bibliographic data analyses. We combine these two types of information in various ways to construct ranking schemes that promote topics with different breadth and scope. We extend the synergy between the text mining and social network-based approaches to the topic treatment by exploring how some of the topic properties correlate with the formal properties and quantitive characteristics of the co-author networks.

The chapter is organized as follows: Section 3.2 describes the process of topic generation. In Section 3.3 the various ways of topic ranking are introduced. Section 3.4 presents the experiments and discusses the results. A summary

of our study and a synopsis of the future work are given in Section 3.5.

3.2 Topic generation

The goal of this study is to extract topics from the bibliographic data and distinguish between the broad and narrow topics via the combination of three sources of information: text, co-authorship graph, and time. We start from extracting topic using conference publication titles which constitute the textual component for the purpose of this work.

3.2.1 Extracting Topics

In this work a *topic* is defined as a *collocation* composed of n consecutive words, where $2 \leq n \leq 3$. Requiring the topic components to be a collocation implies that they are semantically related, together convey a certain meaning which is different from the meaning of individual words, and the probability of their co-occurrence is higher than it would be expected if the words were independent [91]. In this context, expressions like "data mining" or "disjunctive logic programming" are examples of topics. One possibility to extract such topics would be to identify all tuples of the most often co-occurring words (direct neighbors). This technique incremented by some post-processing fine-tuning has been successfully applied in [34, 159]. However it has been shown by [70] that frequency alone does not always function as a discriminative indicator of not-by-chance word co-occurrence. We therefore choose another procedure and apply *likelihood ratio test for binomial distribution* [36] in order to decide whether or not a sequence forms a collocation. This technique performed well in the similar task when applied to the large general purpose text collections such as news, as well as highly specialized ones like textbooks and full texts of scientific articles [86, 100]. Inspired by its success we apply it to a new type of corpora derived from the bibliographic data.
The likelihood ratio test belongs to the class of *hypothesis* tests where one formulates two hypotheses: a) *null hypothesis* which expresses the word independence, and b) *not-null hypothesis* under which the words are semantically related and their co-occurrence is not a chance event. The equations 3.1 and 3.2 formalize these hypotheses for the case of testing two words but can be

extended for longer expressions.

$$H_0 : P(w^1 w^2) = p = P(w^2|\neg w^1) \tag{3.1}$$

$$H_1 : P(w^1 w^2) = p_1 \neq p_2 = P(w^2|\neg w^1) \tag{3.2}$$

By taking the ratio of the likelihoods of the two hypotheses λ one may say how much more likely one hypothesis is than the other. The null hypothesis H_0 is rejected if $p_1 \gg p_2$. It has been shown in [36] that the quantity $-2log\lambda$ is asymptotically χ^2 distributed. Hence we can use the χ^2 distribution table to determine for each word sequence the confidence level of its $-2log\lambda$ value, and compare it to the treshold value required for a collocation which is set to 10.83 with confidence level $p = 0.001$. All candidates which satisfy the treshold are considered valid collocations and make up the resulting list of preliminary topics.
We discuss the topic lists in Section 3.4.

3.2.2 Topic terms Refinement

As mentioned above we allow topic terms composed of two and three words (bi- and tri-grams further in this text). Any trigram can be seen as an extension of some bigram by one word. Presumably there are cases when $-2log\lambda$ values are sufficiently high to retain both - a bigram and its corresponding trigram(s) as topic terms. Thus we obtain terms like *"generative model"* as well as *"discriminative generative model"* and *"probabilistic generative model"*. However in some other cases selecting a trigram along with its bigrams may yield false positives. For example in *"world wide web"* only the trigram itself makes sense but neither *world wide* nor *wide web* are valid by themselves. If we think of any topic-based application, for instance, a search engine that visualizes topics, we have to minimize such cases. We therefore complete the process of topic generation by applying *subsumption approach* proposed in [125] for the deriving of concept hierarchies from text. The original idea is the following: given two terms x and y, x **subsumes** y if the documents which y occurs in are a subset of the documents which x occurs in. Since x subsumes y and because it is more frequent, x is the parent of y. We adopt this idea and modify it in such a way that it serves in two different scenarios.

- **Cleaning topic list from meaningless collocations**. Given a bigram x and its extension, trigram y, we **eliminate** x as having no stand alone meaning if it occurs in 80% of the documents (i.e. publication titles) which y occurs in. In other words, x is removed from the list of topics if it occurs as part of y in at least 80% of the cases. Note that we do not require a complete overlap between the occurrences of x and y. Doing so would lead to preserving a high number of meaningless bigrams just because of a few cases in which x did occur without y.

- **Defining clusters of lexically related terms**. Given a bigram x and its multiple extensions $Y = \{y_1, y_2, ..., y_n\}$, **the cluster is formed** with the central term being x, and the member terms $\{y_1, y_2, ..., y_n\}, y_i \in Y$.

After the refinement we can proceed with studying some of the topic properties.

3.3 Reranking of the Topics

Since collocations are semantically meaningful units, the ranked list obtained in the way described above could already serve as a final ranked list of topics. However we consider the re-ranking due to the following observations. First of all, the two and three word collocations are generated separately, which results in two independent topic lists. Because bi- and tri-grams have different ranges of weights there is no straightforward way to compile them into one ranked list of topics without recurring to any external information. Second remark addresses the meaning of the collocation weight in general. The $-2log\lambda$ value of a topic reflects its relevance to the corpus as a whole. However it fails to capture the information about topic generality or specificity although one often needs to classify topics in this way. To overcome the lack of such information we define additional metrics for topic ranking.

3.3.1 Ranking of topics by citation

The mechanisms of the topic detection and refinement described so far can be considered generic as they apply to any type of textual data — let it be a collection of free texts or a corpus constructed of some elements of scientific writing, such as titles, abstracts or full texts. Alternatively our approaches to the topic re-ranking discussed in this and the following subsections take

advantage of the information that is either explicitly present in the bibliographic databases or can be mined from them.
It is common to interpret citations as a recognition of importance of an object or event [13, 49]. Citations are also widely used for the topic related tasks. For example co-citation information along with the word-based title similarity is analyzed by [66] in order to identify thematically related scientific papers. Citations are employed by [158] in order to detect topics in large-scale linked document collections. In [102] textual information and citations are coupled for the purpose of topic modeling of CiteSeer data. A strategy of leveraging topic analysis by various bibliometric measures including citation counts is described by [90]. Unfortunately citations are far from being always available in the bibliographic databases (see for instance Table 1.1 in Chapter 1). This is the reason why we propose a metric that substitutes the literal citation although does capture its semantic meaning. To decide on salience of a topic we define two types of citations: *citation by title* and *citation by conference*. The idea behind it is to consider every apparition of the given topic after its first occurrence as a reference to or citation of the original topic. Note that at this point we pass from the global corpus-wise represenation of topics to a structure that associates each topic to the publication titles and venues. Moreover we incorporate time dimension into the analysis.
To compute the new weight $weight_{t_i}$ of a topic $t_i \in T$ where T denotes the list of topics produced via the collocation extraction as described in subsection 3.2.1, we define:

- Citation by title $cite_{t,i}$ as a number of titles which topic t_i occurs in after the first apparition.

- Citation by conference $cite_{c,i}$ as a number of different conferences which topic t_i occurs in after its first apparition.

Then the resulting topic weight is given by the product of the two types of citations:

$$weight_{t_i} = cite_{t,i} \times cite_{c,i} \tag{3.3}$$

This metric favors topics which have high counters for both, titles and conferences. Consequently we expect topics that reflect broad trends to outrank the more locally focused ones.

3.3.2 Ranking topics by co-authorship

Until now we have been exploiting textual, temporal and some aspects of bibliographic information in order to create, refine, and re-rank the topics. Similar to the previous subsection, a ranking metric described here aims at distinguishing between broad and focused topics but it uses co-author graph properties to do so. Using graphs for text mining and information retrieval is a long time tradition. For instance, in [98] *TextRank* – a graph-based ranking model for keywords and sentence extraction is introduced. Graph-based metrics are used by [41] to compute sentence salience for the purpose of text summarization. In [48] novelty is detected in texts via their graph-based representation. Here the idea of employing co-author graph for the topic ranking is inspired by viewing scientific writing as a communicative process that we have introduced in Chapter 2. Intuitively more general topics will be spread among many not necessarily related to each other authors. On the contrary, more specific topics will be likely to link individuals that have been working on them into "socio-epistemic" networks [112], revealing tight co-author clusters behind themselves. This is the reason why we suppose that the metrics employed in the social network analysis are appropriate for the given task. According to Watts and Strogatz [153] social networks are characterized by the presence of local communities (or clusters) in which the number of people who know each other exceeds the average. Watts and Strogatz have also introduced a method that tests the actual network data for having such clusters. They defined a *clustering coefficient* that quantifies how close the direct neighbors of a vertex in a graph are to form a complete graph.

To compute the topic weight in this co-authorship graph-based metric we build a co-authorship graph G_t for each topic $t_i \in T$, with vertices $\{V'\}$ being the authors of all the papers which t_i occurs in, and edges $\{E'\}$ defined by the co-authorship relation between the authors in G_t. The topic weight $weight_{t_i}$ is given by the clustering coefficient of G_t, cc_{G_T}, and is computed as follows:

$$weight_{t_i} = cc_{G_T} = \frac{|E'|}{(|V'| \times (|V'| - 1))/2} \tag{3.4}$$

where the nominator is the number of edges in G_t, and the denominator is the maximal number of edges that would have been in G_t if it was fully connected.

We observe that such graphs are sparse: they represent a set of typically

unrelated cliques. That is, the edges in G_t are mainly the ones which connect the authors of every given paper, but there are almost no edges between the authors of the different papers. However one may assume that some $v'_i, v'_j \in V'$ are connected to each other but not necessarily via particular t_i. It follows that G_t might not fully reflect the co-authorship relations between the authors related to t_i. To remedy the situation we complete the G_t with information from the global graph $G = \{V, E\}$, where $\{V\}$ are the authors of all publications listed in the bibliographical database, and there is an edge $e_{i,j} \in E$ between some v_i and $v_j \in V$ if they co-authored at least one paper. The process of building G_t is now modified in the following way: after the authors of all papers containing t_i are introduced and appropriately connected in G_t, every pair of unconnected vertices v_i, v_j is checked for having an edge in the global graph G. Should there be one, an edge $e_{i,j}$ is added to the G_t. After all the vertices $\{V'\} \in G_t$ have been checked a new clustering coefficient cc'_{G_T} is computed with the updated number of edges $\{E''\} \in G_t$. In terms of expected quantitive indicators, we suppose that the higher is the cc'_{G_T} value the more focused is the topic.

3.3.3 Ranking of topics by $tf.idf$ value

Term frequency - inverse document frequency ($tf.idf$) is another way of separating terms into general and specific. Introduced in [141] it has been widely used in the field of information retrieval. We use it here as a benchmark for the two other metrics introduced in subsections 3.3.1 and 3.3.2. The metric combines the term *salience* for the collection of documents (tf) with its *informativeness* (idf) presuming that the more focused terms will be concentrated in a fewer number of documents than more general ones which would be spread throughout the collection. We apply this metric as follows:

- term t_i = topic $t_i \in T$;
- document $d_j = c_j$, where c_j is a conference from the list of all conferences C in the database;
- $tf_{i,j}$ is the number of titles which t_i occurs in;
- cf_j is the number of different conferences which t_i occurs in.

The weight of each topic $t_i, t \in T$ is given by (3.5):

$$weight(i,j) = \begin{cases} (1+\log(tf_{i,j}))\log\frac{C}{cf_i} & if \quad tf_{i,j} \geq 1 \\ 0 & if \quad tf_{i,j} = 0 \end{cases} \tag{3.5}$$

where $f(tf) = (1+\log(tf_{i,j})), tf > 0$ is the dampening function. (See page 542 of [91] for a detailed explanation). We expect that more general topics will be featured not only by the high number of hosting titles but also by the high number of conferences which they occur in, as opposed to the more specific ones, grouped in relatively small number of venues.
In Section 3.4 we compare the results of all the three different metrics.

3.4 Experiments and evaluation

In this section we discuss experiments that have been performed to test the methods described above. We focus on conference publications[1] and use computer science bibliographic database DBLP as a test bed. Our experiments are run on the DBLP release from February 2008.

3.4.1 Data collection and preparation

The XML file is parsed and the data is stored in a database. Then it is organized into two independent sets. One is intended for the collocation extraction and contains titles of conference papers. The initial list consisting of 610895 items is further preprocessed by converting to the low case, removing stop words (we use a list provided by the Lingua package [114]), punctuation, and titles which contain non-ASCII symbols. These constitute $\sim 2\%$ of the total number, and are mostly French and German ones with a few occurrences of the mathematical notation. The resulting list contains 599456 titles. In the second set we store complete information about the publications, including author names, title, year, and venue. It counts 610895 titles, 609053 authors, and 3996 conferences in the range of 49 years, from 1959 to 2008.

[1]Conferences have different roles in different scientific fields. It has been argued that in Computer Science conferences play a more important role than journals do [79, 99]. This is the reason why we choose conference publications for our experiments.

3.4.2 Evaluation of topics on DBLP

The preprocessed list of titles serves as the input to the program which generates topics. (We use the NLP package for collocation extraction [4], with loglikehood ratio test λ as a statistic metric, and 10.83 as a cutoff weight for the $-2log\lambda$ value.) The process yields 392994 bi- and 3150332 trigrams. Since the titles were modified during the preparation stage, not all the collocations are valid. We then conduct a post-processing which amounts to:

- matching collocations to the original titles. Collocations that contain punctuation marks and/or stopwords, or which components fail to represent a sequence, are eliminated.
- merging singular and plural cases into one entry;
- subsumption, as described in subsection 3.2.2.

At the end of the post-processing we obtain a structure known in information retrieval as *inverted file* where for each entry the number of occurrences and an array of hosting titles are stored. The number of retained topics is reduced to 124480.
Table 3.1 shows some examples of the subsumption process. The first row illustrates elimination of a meaningless bigram "adaptable user". The second row is an example of a cluster which is formed around the bigram "ada programming". It is covered by the corresponding trigrams but is not eliminated. Analysis of the list of such clusters shows that many bi-grams while covered by some set of trigrams have a meaning of their own and could potentially serve for topic labeling. The last row is an example of a cluster built around the bigram "application software". The topic designated by the bigram is broad enough and is not covered by the cluster members.

3.4.3 Experiments with topic re-ranking

In 2008, the data stored in DBLP has spanned 49 years. However it can be seen from the Figure 3.1, that the number of publications increases considerably toward mid eighties. That is the reason why we restrict our experiments to topics that appeared no earlier than 1988. (The sharp fall of the curve toward the end of 2010 is explained by the fact that the data from $2007-2008$

Table 3.1: Examples of subsumption procedure.

Bigram	Frequency	Trigram	Frequency	Covered
adaptable user	9	adaptable user interface	8	Yes
ada programming	9	ada programming environment	2	
		ada programming language	2	
		ada programming support	3	
		advanced ada programming	2	Yes
application software	39	application software development	3	
		application software systems	2	
		embedded application software	2	
		mobile application software	2	
		generic application software	2	No

had not been completely introduced into the database by the time we downloaded the file.). Additionally we restrict the minimal topic frequency to 5 for the bi-grams, and 2 for the tri-grams.

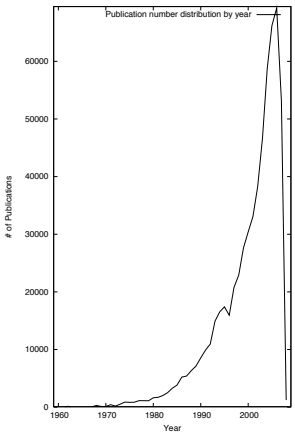

Figure 3.1: Paper distribution in DBLP from 1959 to 2008.

Results of the ranking by citation

Table 3.2 lists 20 top ranked topics according to the citation ranking computed using the equation (3.3).
We observe that the ranking results agree with our expectations, as almost all twenty topics designate broad areas of computer science. They are featured by high numbers of both - conferences and papers, and reflect "trendy" research directions of the last 15 years. The metric captures a high interest in relatively new topic - "semantic web": despite its shortest span (8 years),

Table 3.2: The 20 top ranked topics by the citation metric.

topic	weight	# of conferences	# of titles	year	span
web service	2039826	654	3119	1994	13
sensor network	1777047	501	3547	1993	12
data mining	1045044	572	1827	1993	16
ad hoc network	1004598	441	2278	1995	13
wireless sensor network	648999	351	1849	1999	10
mobile agent	622362	474	1313	1994	15
wireless network	563178	371	1518	1992	17
semantic web	495624	386	1284	2001	8
multi agent system	492063	403	1221	1991	18
support vector machine	379874	341	1114	1996	13
mobile ad hoc	363025	325	1117	1998	11
virtual environment	359755	341	1055	1990	18
digital library	293112	236	1242	1991	17
association rule	261318	291	898	1993	16
face recognition	256522	251	1022	1990	18
context aware	241696	332	728	1996	12
web application	238924	322	742	1996	13
reinforcement learning	218240	248	880	1988	20
evolutionary algorithm	195487	233	839	1993	15
virtual reality	185472	288	644	1990	18

Table 3.3: Topics on the $500th_s$ rank.

topic	weight	# of conferences	# of titles	year	span
handwriting recognition	6850	50	137	1993	15
distance measure	6688	76	88	1990	15
heterogeneous computing	6649	61	109	1989	17
online game	6608	59	112	2001	7
authenticated key	6771	61	111	1993	12
soc design	6630	51	130	2000	8
aspect oriented programming	6528	64	102	1997	11
predictive control	6510	62	105	1995	11
protein folding	6435	65	99	1992	16
image denoising	6292	52	121	1997	11

and relatively recent emergence (2001) it scores seventh on the total list of topics.

As we descend toward the lower ranked topics we notice that they gradually become more focused. Table 3.3 shows more specific topics such as "handwriting recognition"; concept terms, names of techniques and processes like "authenticated key", "image denoising" or "protein folding"; or yet multidisciplinary technical terms like "distance measure". The conference – paper relation suggests that these are concentrated at a smaller number of venues than the trendy topics from Table 3.2.

Table 3.4: Top versus Bottom ranked topics ordered by the clustering coefficient.

topic	vertices	edges (local)	edges (global)	cc'_{G_T}
spiral architecture	19	40	43	0.25146
blue gene	209	3059	3523	0.16208
proof planning	39	53	114	0.15385
proof carrying code	21	30	32	0.15238
related key	44	89	135	0.14271
parameterized complexity	50	121	165	0.13469
defeasible logic	62	202	211	0.11158
functional logic program	42	65	90	0.10453
secure computation	72	104	251	0.09820
american sign language	77	264	283	0.09672
...	...	...	...	...
multi agent system	2259	3491	4595	0.00180
virtual environment	2330	3987	4725	0.00174
mobile agent	2245	3452	4029	0.00160
support vector machine	2459	3427	4469	0.00148
wireless sensor network	3785	7067	8964	0.00125
wireless network	3311	4945	6737	0.00123
data mining	3641	5779	7563	0.00114
ad hoc network	4254	6183	8482	0.00094
web service	5732	10561	14698	0.00089
sensor network	6475	12883	16730	0.00080

Results of the ranking by the clustering coefficient

Let us now look at the topic list ranked according to the clustering coefficient cc'_{G_T} described in subsection 3.3.2. Table 3.4 shows 10 out of the top 20 topics, and 10 out of the last 20 topics on the list. The top ranked topics represent quite specific research fields such as theorem proving, cryptography, branches of logic or linguistics. The metric also puts forward certain product names like "blue gene" – a very well known yet narrow focused concept. On the contrary topics with the lowest rank represent the broad areas of computer science. Moreover they almost exactly mimic the top ranked topics according to the citation metric. Comparison of the resulting lists of the two rankings shows that the close inverse correspondence between the $< topic, rank >$ pairs holds for at least one thousand topmost or lowest ranks depending on the ranking scheme. This experiment proves our expectations that the clustering coefficient may serve to distinguish between broad and focused topics and gives priority to the more specific ones.

Table 3.5: 10 top most ranked topics by the $tf.idf$.

topic	weight by $tf.idf$	# of conferences	# of papers	rank by citation	rank by clustering coefficient
research note	40.05	4	128	4289	4680
interactive presentation	34.97	4	61	7293	8121
co chair	33.92	12	135	1745	1251
output analysis	33.75	4	51	8344	2000
parallel manipulator	33.16	10	99	2581	8759
poster abstract	32.80	7	68	4536	9119
workshop chair	32.74	4	44	9229	1579
simulation optimization	32.70	7	67	4557	7423
digital government	32.16	9	76	3431	5765
low voltage	31.68	36	337	288	5568

Results of the ranking by $tf.idf$

Table 3.5 presents the 10 top entries from the topic list ranked according to the $tf.idf$. Since this metric gives the maximal weight to items which occur in 1 document we set the minimal number of documents (i.e. conferences in our case) to 3. We do so after the manual check of the results on an unrestricted set, which put forward dozens of terms like "session chair", "extended abstract", etc. Despite this measure, we immediately notice that among the selected items there is a high number of non-topic terms such as "research note" or "interactive presentation". The mixture of topic and non-topics terms happens everywhere throughout the list. Note also that the figures in the last two columns which correspond to the **topic rank** assigned by the citation and clustering coefficient metrics respectively, do not allow to establish dependency between this and the two other metrics. We explain such a behavior by the fact that $tf.idf$ is the less informed of all and clearly prefers items with the high paper-to-conference ratio which does not model the topic properties correctly.

3.5 Summary and Future work

In this chapter we have described the way of research topic extraction based on the titles of scientific publications. We have introduced and compared three different methods of topic ranking aiming at distinguishing between general and specific topics. The rankings by citation and clustering coefficient have yielded topic lists which corresponded to our expectations: the first metric put forward the broader topics, while the second favored the more focused ones. Ability of both metrics to differentiate between the topic scope suggests flexibility of their application — one can choose either of these two

ranking schemes depending on the task at hand. On the contrary, the $tf.idf$ weighting has failed to generate a coherent list, mixing up topic and non-topic terms. Such an outcome shows that the paper-to-conference relationship alone does not provide sufficient ground for the topic ranking.

Our approach to the topic generation and refinement is generic and can be used with any kind of textual data. As of the re-ranking mechanisms, they rely on the information that is either explicitly present in the bibliographic databases — for example, venue titles or time, or can be inferred from the bibliographic data — for example, co-authorship network. The citation and co-authorship graph based ranking methods possess practically useful features. The first one is interesting because it captures the meaning and functionality of citations without explicitly requiring them to be present. The second method may shed light on the community-wise collaborative practices.

Topics have been extracted in this study based on the publication titles only. Being relatively short, titles contain a reduced amount of textual information which does not allow to capture semantic relations between the topics and they are processed as atomic. Extending textual data with the publication abstracts would alleviate this problem and permit ranking of semantically related topic clusters rather than individual topic terms.

Another point that requires further investigation refers to the ability of our methods to not only separate between the broad and narrow topics but also distinguish between the topics, concepts and technical terminology even though all the three categories are often labeled "topics" in the context of information retrieval and text mining. To address this question we will have to investigate the topics from the point of view of their informativeness and revise the distribution among the authors and venues. For the first part, an idea of extending the h-index[2] application from individual scientists to a collection of scientific papers proposed by [24], is worth checking. For the second part it might be useful to split the entire text collection into a number of broad thematic categories (using for example Latent Dirichlet Allocation [9] and Topical N-grams [152] techniques for this purpose) rather than working with semantically unstructured corpus.

[2]h-index is a measure of the individual scientist's productivity defined as the number of papers with citation number h. [59]

Chapter 4

Analysis of computer science communities and conferences

Chapters 2 and 3 dealt with the authors and topics from the sociolinguistic and text mining points of view. Authors and topics are considered in this chapter as well, but from the perspective of social communities and conferences. Here we use the DBLP data to investigate the author's scientific career and provide an exploration of some of the computer science communities. We compare them in terms of productivity, population stability and collaboration trends. Besides we use these features to compare the sets of top-ranked conferences with their lower ranked counterparts.

4.1 Introduction

Being broad and constantly growing field, computer science comprises various subareas each of which has its own specialization and characteristic features. At the same time there exist multiple connections between the areas. Thus for example *Information Retrieval* combines computer science, linguistics, cognitive psychology, and mathematics. Yet another example, from the area of the *World Wide Web*: its rapid growth requires efficient techniques for management of the large volumes of data — a task that has traditionally been associated with the field of *Databases*. The interdisciplinary nature of research is reflected by the conferences' content. Take for instance the *Conference on Information and Knowledge Management* (CIKM): besides the topic spelled out in the conference title, it has two other, equally important,

streams: *information retrieval* and *databases*. While different in size and granularity, research areas and conferences can be thought of as scientific communities that bring together specialists sharing similar interests. What is specific about conferences is that in addition to scope, participating scientists and regularity, they are also characterized by level. In each area there is a certain number of commonly agreed upon top ranked venues, and many others – with the lower rank or unranked. In this study we aim at finding out how the communities represented by different research fields and conferences are evolving and communicating to each other. To answer this question we survey the development of the author career, compare various research areas to each other, and finally, try to identify features that would, along with the already existing ones, allow to distinguish between venues of different rank. We believe that such an insight might be of interest for advanced students who are about to choose their specialization; young researchers looking for an appropriate conference to submit their work; authorities who decide on funding of diverse research areas.
This chapter is organized as follows: in Section 4.2 we give an overview of the related work. Section 4.3 elaborates on the data collection. In Section 4.4 we discuss the author profiling. Section 4.5 focuses on the comparison between various communities and venues. Section 3.5 concludes the chapter.

4.2 Related Work

Communities are nowadays actively analyzed in the context of social networks and mechanisms responsible for their life-cycle. While mentioned several times in the preceding chapters, social networks have not been formally introduced so far. According to Newman [107], *a social network* is a set of people each of whom is acquainted to some extent with some or all the others in this set. Thus, virtual blogs, school or university teams, colleagues in a company, conference attendees would all be an example of a social network. Social networks have been investigated from both theoretical and empirical perspectives. Watts and Strogatz [153] contributed to the networks analysis by elaborated discussions on topology, clustering patterns and comparison of random and regular networks. Newman [26, 109, 110] has been studying a wide variety of social networks and investigating their essential properties, such as degree distribution, centrality, betweenness, and assortativity, to name a few. The theoretical insight into the principles of social networks

yielded a great deal of interest in studying research communities and their properties based on the coauthorship networks. Nascimento [104] has studied network properties of the SIGMOD co-authorship graph. Hiemstra et.al. [58] suggested a topological analysis of the Information Retrieval community extracted from the SIGIR records. Backstrom, Huttenlocher and Kleinberg [3] have studied mechanisms underlying the membership, growth, and change of the user-defined communities in LiveJournal and DBLP. An extensive bibliometric study has been performed by Elmacioglu and Dongwoon Lee [38]. Using DBLP to build a co-authorship network they have investigated various properties of the Data Base community and came to the conclusion that DB is a "small-world" community. Using CiteSeer as a source of bibliograhic records, Huang et. al. [61] applied bibilometric techniques to the analysis of a number of computer science fields in order to study dynamic properties of the underlying netwoks. Based on the top ranked venues recorded in DBLP, Bird et. al. [8] identified 14 computer science communities and studied collaboration patterns and interdisciplinary research at the individual, within-area, and network levels.

An important product of scientific activity is research. Research assessment constitutes one of the preoccupations in bibliometrics and scientometrics and has recently become an active research topic in the computer science environment. Several evaluation strategies have been developed throughout the years.

In 1972 Garfield [49] proposed an *Impact Factor (IF)* — a metric targeted to reflect the journal salience based on the average number of citations received by that journal's publications within the two years window. While yearly calculated for hundreds of scientific journals, IF has a number of weaknesses: it is inflated by the survey and long articles that attract high number of citations and thus alter the IF of a journal; it is not representative for the quality of the individual articles; it does not correct for self-citations [127]; neglects authority of citing source [10]. The limitations of IF gave raise to the other citation-based metrics. For example, Bollen [10] suggested a *weighted PageRank* that like the original PageRank algorithm [19] accounts for not only the number of citations but also for the authority of a citing source. Rowland [123] suggested *Journal diffusion factor* as a complimentary metric that quantifies the transdisciplinary influence of research. Analyzing the scope of *Angewandte Chemie International Edition* (AC-IE), Bornmann and Daniel [15] proposed to augment the research evaluation criteria by *citation's spread speed* as they have shown that the papers accepted for the publication

in AC-IE have a higher chance to be cited than the papers rejected by AC-IE but published elsewhere. Citation-based *h-index* [59] is used to rank an individual's research. Other entities evaluated on the citation basis include digital libraries [144], system of prize awarding [134], conferences [133, 148, 149], authors and documents in the heterogeneous networks [161].
Let us take a closer look at the ranking of conferences since conferences constitute one of the main concerns of our study. Along with the citation-based metrics other approaches to the conference assessment have been tried. Elmacioglu et. al. [39, 162] suggest to decide on the conference quality based on the quality of its program committee members. Yan and Lee [157] propose a way of ranking venues based on the scientific contribution of the individual scholars. Souto et. al. [140] compose an evaluation list of factors like sponsorship (ACM, IEEE, SIAM, IFIP, ...), length of the accepted papers, venue status (main conference versus co-located workshop), proceedings' publisher(ACM, IEEE, SIAM, ...), and scope (International, National, Regional). Waister et. al. [149] augment the above list with the submission and acceptance rates as well as the conference's life-time.
Despite the variety of attempts to rank conferences there remain a number of observations to make. First of all, it stems from the brief overview above that there is no unique, commonly agreed upon evaluation criteria. Second, many of the listed experiments require data which is not explicitly present in the bibliographic databases - let it be citations, program committee members or submission / acceptance rates.
Our work bears on the previous research in that it focuses on a statistical investigation of scientists and scientific communities. Its contribution consists in:

- extension of a framework for the author's analysis in order to build a comprehensive profile of the researchers on DBLP;
- an attempt to bring together the community analysis and some aspects of the research evaluation relying on the data directly available from the bibliographic databases.

4.3 Data Collection

Like in the two preceding chapters, we use DBLP to conduct our investigation. We downloaded the XML file in August 2009 and used conference

Table 4.1: Example of Conference Name Integration

Resulting Name	Individual Names	Time span
AAAI	Agent Modeling	1
	Deep Blue Vs kasparov: the Significance for Artificial Intelligence	1
	AAAI Workshop on Intelligent Multimedia Interfaces	1
	AAAI/IAAAI, Vol.1	1
	AAAI/IAAAI, Vol.2	1
	AAAI/IAAI	5
	AAAI	17

publications for the corpus construction. Although DBLP covers 50 years of publications, the data before 1970 is rather irregular. This is the reason why we consider publications from 1970 on.

The complete list of conferences accounts for 4449 distinct conference names. Manual examination of the conference pages in DBLP has shown that some venues have changed their names one or more times since they had been established. It follows that we cannot treat conference names as unique because there is no guarantee of capturing the entire history of a venue. Fortunately all instances of the same conference can be automatically identified with the XML tags in the original file. We use this feature and integrate all events of a venue with multiple names under the name of a component with the longest history. Table 4.1 illustrates the idea. *National Conference on Artificial Intelligence (AAAI)* had been recorded in DBLP under the names like "Agent Modeling", "AAAI/IAAI", "AAAI/IAAAI Vol.1", etc., accumulating seven different name variations in total. However the name with the highest number of occurrences was "AAAI" and it has therefore been chosen to label this conference. Due to the name unification, the number of conferences is brought down to 2626. Publications from these conferences constitute the most general data set we use for our experiments. It is denoted *DBLP dataset*.

As we are interested in a comparative analysis of different scientific communities and venues we have to split the entire set of publications into topical subareas. One of the ways to do so is to specify sets of conferences that correspond to every subarea we want to analyze. Thus we select 14 subareas each of which is represented by a set of relevant top ranked conferences with at least 10 years time span for the sake of data stability[1].

[1]We have had to relax the "min 10 years time span" requirement when dealing with conferences in Computational Biology and World Wide Web because these are young areas

Table 4.2: Research Communities and Corresponding Top Conferences

Abbreviation	Area	Conferences
ARCH	Hardware&Architecture	ASPLOS, DAC, FCCM, HPCA, ICCAD, ISCA, MICRO
AT	Algorithm&Theory	COLT, FOCS, ISSAC, LICS, SCG, SODA, STOC
CBIO	Computational Biology	BIBE, CSB, ISMB, RECOMB, WABI
CRYPTO	Cryptography	ASIACRYPT, CHES, CRYPTO, EUROCRYPT, FSE, PKC, TCC
DB	Data Bases & Conceptual Modeling	DEXA, EDBT, ER, ICDT, PODS, SIGMOD, VLDB
DMML	Data Mining, Data Engineering, Machine Learning	CIKM, ECML, ICDE, ICDM, ICML, KDD, PAKDD
DP	Distributed&Parallel Computing	Euro-par, ICDCS, ICPP, IPDPS, PACT, PODC, PPoPP
GV	Graphics&Computer Vision	CGI, CVPR, ECCV, ICCV, SI3D, SIGGRAPH
NET	Networks	ICNP, INFOCOM, LCN, MOBICOM, MOBIHOC, SIGCOMM
NLIR	Computational Linguistics, Natural Language Processing, Information Retrieval	ACL, EACL, ECIR, NAACL, SIGIR, SPIRE, TREC
PL	Programming Languages	APLAS, CP, ICFP, ICLP, OOPSLA, PLDI, POPL
SE	Software Engineering	ASE, CAV, FM/FME, Soft FSE, ICSE, PEPM, TACAS
SEC	Security	CCS, CSFW, ESORICS, NDSS, S&P
WWW	World Wide Web	EC-web, ICWE, IEEE/WIC, ISWC, WISE, WWW

The idea of relying on the top ranked conferences is inspired by works of [8, 61, 157, 162], and is grounded on the assumption that high quality conferences are clearly defined in terms of topics they cover. While every area has a modest number of commonly agreed upon top ranked venues, the assignment remains subjective. This is the reason why we validate the choice of venues by consulting several hand-made conference ranking sources [29, 30, 31] and considered the estimated venue impact provided by [42]. To enable a fair comparison we represent each subarea by the same or nearly the same number of conferences[2]. Table 4.2 shows the resulting data set which is denoted *TOP dataset*. The data in the table is represented in the following format: the first column contains the area abbreviation followed by the abbreviation's meaning given in the second column, and the third column lists the conference names that we have chosen to represent each area.

One might notice that "Artificial Intelligence" (AI) does not make part of the chosen areas. While AI is a large, dynamic and widely researched domain, it is extremely interdisciplinary and includes fields that can be considered in their own right, for example *Natural Language Processing, Machine Learning, Computer Vision, Artificial Neural Networks*, to name a few. AI is therefore difficult to delimit and, as opposed to other research works that do treat it as a separate area, we refer to a subset of its smaller components.

As one of our goals is to identify a set of features that would help to distinguish between the top and non-top conferences, we need a selection of conferences that do not belong to the set of top ranked venues. Using the same human-made sources we select 6 areas with 5 representative conferences each. They are given in table 4.3, and constitute the *NONTOP dataset*. (Table 4.3 has the same format as table 4.2.)

Note that there are some differences between the two sets in terms of top-

that have started off at the end of 90s.

[2]In a few cases renowned conferences with less than 10 years history have been chosen

Table 4.3: Research Communities and Corresponding Non-Top Conferences

Abbreviation	Area	Conferences
AT	Algorithms &Theory	APPROX, ICCS, SOFSEM, TLCA, DLT
CB	Computational Biology & Medicine	APBC, ICB, ISBRA, CBMS, DILS
DB	Data bases	IDEAS, ABDIS, ADC, WebDB, DOLAP
DM	Data Mining	MLDM, IndCDM, ADMA, KES, IDEAL
SeC	Security & Cryptography	SCN, ISC/ISW, ISPEC, ACISP, WISA
WWW	World Wide Web	WEBIST, SAINT, WECWIS, ESWC, ICWE

ical partitioning and number of covered subareas. This is explained by the fact that the data about the lower ranked conferences is less consistent and agreeable, and we have preferred to construct smaller though more reliable sets.

In these three sets above we exclude all publications that have incomplete bibliographic data such as missing authors, title or year. These constitute 0.052% of the records. The remaining publications are used to build co-authorship graphs G_{DBLP}, G_{Top}, and G_{nonTop}, where $G_{Top}, G_{nonTop} \in G_{DBLP}$. These are undirected graphs where the authors constitute the set of vertices $\{V\}$, and two vertices $v_i, v_k \in \{V\}$ are connected by an edge $e^{'} \in \{E\}$ iff v_i and v_k have coauthored at least one paper. Our experiments are based on these graphs along with other bibliographic data such as number of records, venue, year.

4.4 General Researcher Profiling

The authors in a co-author network are typically investigated from the point of view of their contribution to the research. Thus particular attention is paid to the members of program committees [162], "fathers" of the influential research directions [160], authors with high citation index [135] or yet those researchers who get often acknowledged [51] or invited to give a talk [67]. Such a "celebrity focused" view on scientists is due to *cumulative advantage process* [115], also known as *Matthew effect* [97] which postulates that fame propagates fame, and that influential authors will gain even more in influence. Validity of this phenomenon has been proved statistically [115, 128] and studied in the contexts of publishing activity and recognition [73], influence of academic institutions [78, 94], and specific features of various research fields [32, 126, 147]. The natural consequence of Matthew effect is that it

to maintain consistency of the sets' size.

yields only a partial image of the researchers' community. In this section we aim at providing a broader view on the authors in entire DBLP and the areas described above by looking at their typical career[3] length, interdisciplinarity of interests, individual performance pattern and publication distribution with respect to the top and non-top venues. Since our NONTOP dataset covers only a small part of the lower ranked venues listed in DBLP, we do not compare the TOP and NONTOP datasets to each other in this setting. Rather we contrast the data in TOP dataset to the global author statistics in DBLP.

4.4.1 Author career length

DBLP contains hundreds of thousands of distinct authors. But how many of them pursue a long scientific career?

Figures 4.1 and 4.2 give a full account on the authors career length distribution among the various research areas in the TOP and DBLP datasets. The first chart represents percentage of authors with ≤ 5 career length, while the second one covers periods from 6 to 20 years. It turns out that top-ranked venues are dominated by authors with ≤ 5 years experience, and only $\approx 2\%$ stay publishing at top ranked conferences for more than 10 years. This is consistent with the figures obtained on the whole DBLP set: $\approx 3\%$ of authors have a longer than 10 years career. Based on these figures we assume that the main component of DBLP authors is represented by PhD students who, after having finished their studies, leave the active scientific career. This hypothesis is in line with the results provided by a number of surveys targeted to explore the amount of doctoral student contribution to research [5, 56, 150]. Although abundance of PhD students is not surprising in itself, in the background it correlates with the elitist representation of the research community discussed above. These are mostly the long-term researchers that one would think about in relation with some discipline rather than a PhD student, yet these lifelong researchers constitute a minority of the entire scientific population.

With respect to the research subareas, AT and CRYPTO have the lowest percentage of researchers with a short career and the highest percentage of people whose career length ranges between 10 and 15 years. The explanation

[3]For the purpose of this study, *career* is defined in terms of the author's publication record track on DBLP measured in years.

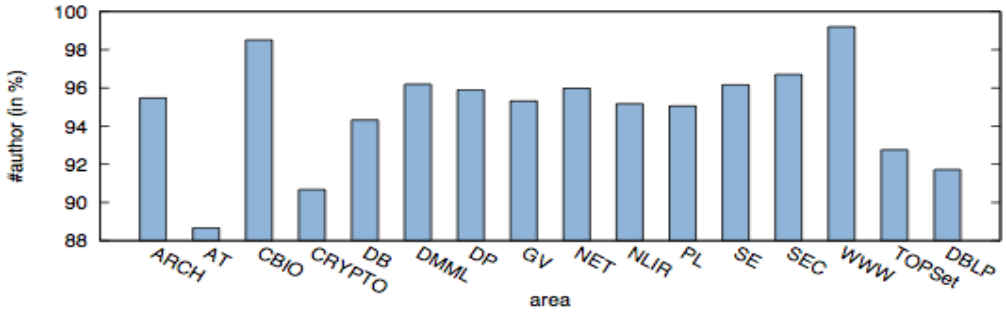

Figure 4.1: Percentage of authors with $\leq$ 5 years career in TOP set and entire DBLP.

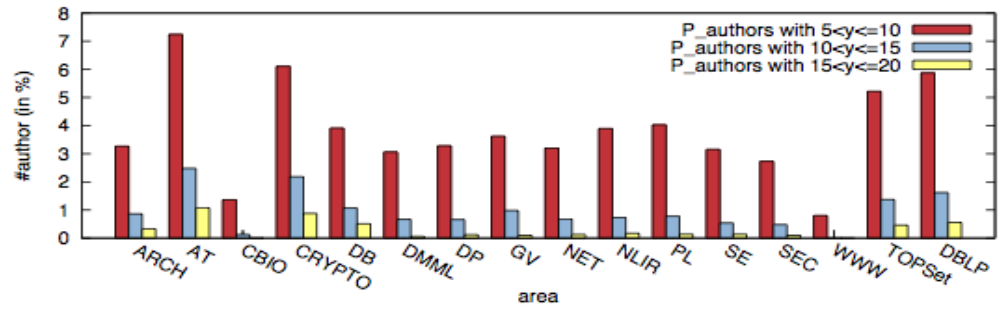

Figure 4.2: Percentage of authors with $6 \leq career \leq 20$ years in TOP set and entire DBLP.

lays probably in that fact that these domains require substantial mathematical background and thus time to obtain it which makes them harder to get in for the short time scientists, and more difficult for switching for those who have spent so much time on it.

4.4.2 Interdisciplinarity of Interests

We have discussed earlier in this chapter the complex nature of computer science which is multi- and- interdisciplinary at the same time. In this section we analyze the interdisciplinarity from the researchers' perspective. Indeed, scientists do not necessarily stay in one and the same field throughout the whole career. But how many areas and at what time of their career do they typically join? What is the probability for a researcher to join one more area

given that he is already publishing in some field?
There are 102928 authors in our TOP set. Out of them only $\approx 22\%$ works in one area only. The remaining 78% join multiple areas with the average value of ≈ 2.2. We have analyzed the data distribution and found that they typically publish in more than one area from the very beginning of the career with a small spike between the 5th and tenth years.
It is logical to assume that the interdisciplinarity of the researcher interests serves as an indicator of the area relatedness that can be calculated. Moreover since the sets of authors are formed based on the conference data, results of the calculations may shed light on the topical orientation of various conferences and connections between them. To quantify the inderdisciplinarity, let A_{start} be an area in which the author a_i started to publish[4]. Next, build a transition matrix P_{A_i} with probabilities $P_{transition} = P_{A_j}|P_{start}$ such that $1 \leq j \leq 14$, and $j \neq start$. Note that there exist two basic scenarios:

- a_i publishes in more than one area in one year, and
- a_i publishes in one area in a given year while overall he is active in multiple areas.

We treat these two cases equally when computing P.
The diagram in Figure 4.3 shows the most probable transitions between the areas. Each circle represents an area, and its size is defined by the number of people working in it. The thickest arrows connect the most related areas, the thinner but solid arrows correspond to the second choice and the dotted ones (when present) to the third. The diagram shows clearly that the area relatedness is asymmetric. For example, Data Mining and Machine Learning (DMML) is primarily related to the Data Bases (DB). At the same time information retrieval (NLIR), computational biology (CBIO), graphics (GV), and WWW have their closest relationship to the DMML, indicating that the authors from these domains publish actively at DMML conferences. One plausible explanation is that these more practical areas constitute a field of application for the data mining and machine learning algorithms.
It is also interesting to note that our rather global results that capture the state of interdisciplinarity in computer science in the last 40 years, are comparable to the yearly snapshots of the area overlap, found in [8]. For example,

[4]When calculating the most related areas we assume that an author is publishing in some area iff he has ≥ 2 publications in it.

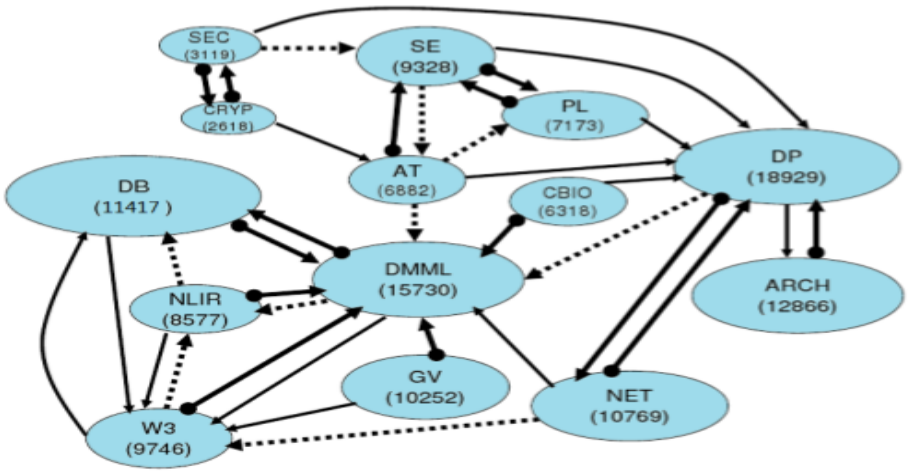

Figure 4.3: Area relatedness based on the researchers' multidisciplinary interests.

both claim that there is a considerable authors' overlap between CRYPTO, Security (SEC), and theory (AT); Programming Languages (PL), Software Engineering (SE), and Distributed Computing (DP); Networks (NET) and DP. The similarity of findings that results from static and dynamic computations might point to the long-term relatedness between the areas.

4.4.3 Some characteristics of "experienced" scientists

We now turn our attention to the authors with ≥ 10 years experience since they are more probable to influence scientific community than "short time" researchers. There are 16192 ($\approx 3\%$) such authors in the whole DBLP set, and 2623 researchers have ≥ 10 years publication record in the TOP set. We characterize this latter group in terms of productivity distribution and focus on the author publication distribution over time and venues.
Analysis of the scientific productivity is an important factor in researcher's evaluation and has been studied in litterature. As early as 1926, Alfred Lotka explored "the part which men of different caliber contribute to the progress of science." [28]. He assumed that the scientific publications obey Pareto distribution according to which the relative portion of scientists with n publications is proportional to $\frac{1}{n^2}$ [124]. This relation, known as *Lotka's Law*,

states that there are only a few authors who have a high number of publications while the majority of authors have only a few publications. It has been tested during the decades of bibliometric studies and proved applicable to a large variety of disciplines of the academic and industrial communities [60, 89, 103, 106].
In our work we are interested in the productivity distribution in the context of the career periods and quality of venues. For the temporal distribution analysis we distinguish between the following three groups of authors:

- Authors with $\geq$ 10 years experience of publishing in TOPset conferences and focusing on one area only;
- Authors with $\geq$ 10 years experience of publishing in TOPset conferences and focusing on multiple areas;
- Authors $\in$ the TOPset with $\geq$ 10 years experience of publishing in the DBLP dataset, irrespective of the number of areas and conference rank.

The average number of publications produced by each category of authors per 5-years periods are plotted at Figure 4.4. The data reveals an inter-

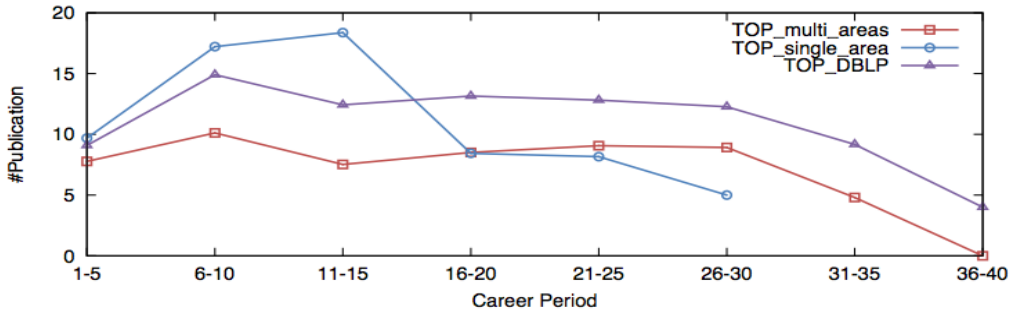

Figure 4.4: Author productivity within the different periods of career.

esting pattern: researchers in all three categories are much more active in the second period of their career, and the single-area authors are even more active in the $3rd$ period. After that the productivity drops in the fourth period and remains stable with some minor fluctuations. Since career periods could be translated into age categories it seems interesting to compare our findings to the earlier research that investigated the impact of age on the publication activity. Lehman and later Pelz and Andrews [46] demonstrated

that scientists are the most productive in their late 30s and early 40s as far as their major contributions are concerned. Although DBLP does not provide information on the researcher's age, this period corresponds intuitively to the second and third phases of the career in our experiment which are also marked by growth of the research activity. The same and some later studies [11] demonstrated that the scientific productivity declined with the age. Note however that none of them confirmed regression in intellectual capacities after the early 40s. Rather influential factors might be: a) relax in motivation and b) increasing specialization which affects the acute viewpoint required for breakthroughs. While increasing age is often associated with the transition from mostly research oriented positions toward the administrative ones which entrain more teaching load, no evidence had been found that scientist's productivity suffers from taking on supervisory duties [33, 46, 78, 131].

In our turn we propose to interpret the plot on Figure 4.4 in terms of the principle milestones in the scientists' life: the first 5 years correspond roughly to the PhD studies during which one typically produces a certain (not necessarily high) number of publications. The next $5 - 10$ years (2nd period) are of great importance to those who stay in research. In that time authors are evaluated on the international scale and their academic position depends heavily on their productivity. Recall also from the Subsection 4.4.2 that the small raise in the number of areas joined by researchers occurs in the same time and is in line with the overall bust of scientific activity characterizing this period. The later stages correspond to the scientific maturity when scientific output stabilizes on average.

With respect to the publication rate values, they are much higher for the single-area authors during the spike periods. There is no additional evidence that would help us to explain this phenomenon. We might hypothesize that by working in one field only it is easier to get more papers published, since the author knows better the research criteria of his community. We might also assume that there exists a correlation between the level of interdisciplinarity and researcher's productivity rate.

To analyze the author - publication distribution over venues we calculate for each author $a_i \in$ TOP dataset the percentage of his publications in the top-ranked conferences relative to all his publications recorded in DBLP. Next we combine the results into the 10%-intervals and match them against the corresponding percentage of authors.

The results are shown at Figure 4.5. It turns out that only about 1.5% of

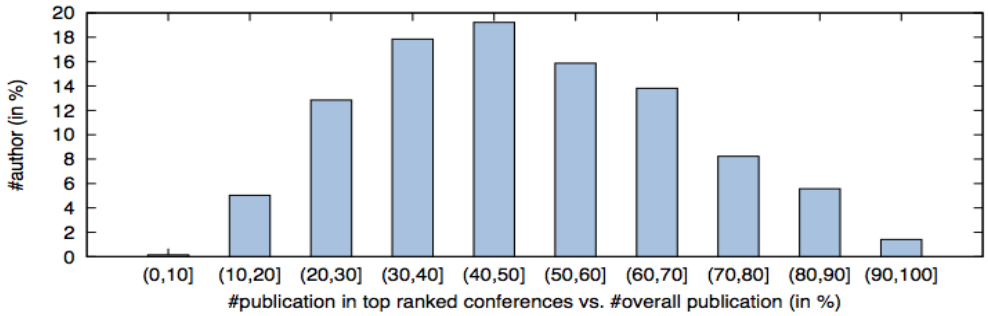

Figure 4.5: Author - venue distribution: percentage of publications at top ranked conferences compared to the overall author production.

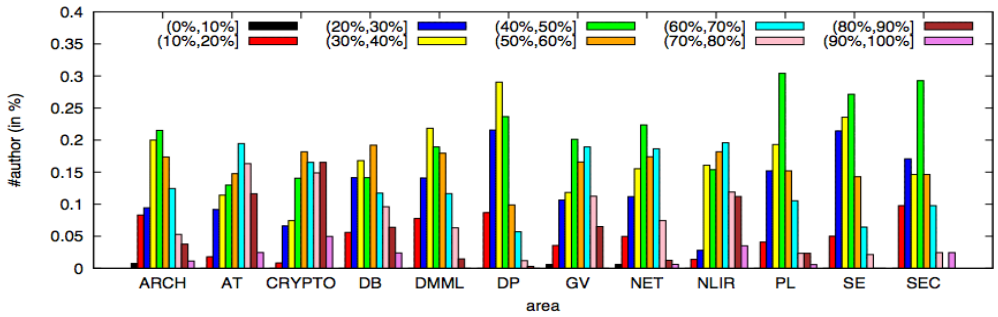

Figure 4.6: Author - venue distribution in various areas.

authors in the TOP dataset publish exclusively or mostly at the top-ranked venues. Typically the top-ranked conference publications constitute from 30% to 60% of the author's conference production. It suggests that the majority of researchers appears in the mixed set of venues.

To look closer at the publication distribution over venues in the topical sets we first assign each author $a_i \in$ TOP dataset to the area he contributes at most (frequency based majority voting), and perform the same computation as before[5].

Figure 4.6 presents the results. Notice that majority of areas are dominated by people who publish between $40-50\%$ of their publications in the top ranked conferences, and in DP and DMML the prevailing range is $30-40\%$. These values confirm the general tendency of publishing in the mixed set of

[5]CBIO and WWW are not considered as the resulting sets of authors are too small to produce consistent results.

venues. On the contrary, authors from DB, CRYPTO, AT and NLIR show more adherence to the top-ranked venues as proportion of researchers who publish $50 - 70\%$ of papers at top-ranked conferences outranks the other categories. These observations suggest that publication habits differ among scientific fields.

4.5 Scientific Community Analysis

The previous section dealt with the author characteristic with respect to DBLP and the research areas defined in Section 4.3. In this section we take a closer look at the areas themselves and investigate them in terms of the *publication growth rate*, *collaboration trends*, and *population stability*. Selection of the evaluation criteria is not random. We believe that it may help to highlight the peculiarities of the individual domains and compare them to each other. We perform the same tests with top and non-top ranked conferences and eventually find differences between these two categories of venues.

4.5.1 Publication Growth Rate

Publication growth rate is commonly perceived as an important indicator of scientific output [2, 79]. It provides an evidence for the area "well-being"and sheds light on how much interest there is in it at the given moment. It is a dynamic measure that traces yearly changes in the area productivity. We distinguish between the *relative* and *absolute* growth rates.
The *absolute growth rate* $AbsGr_{A_{i,y}}$ of an area A_i in year y is a ratio of publications in A_i within two consecutive years y_i and y_{i-1} such that $AbsGr_{A_{i,y}} = \frac{Publ_{A_{i,y}}}{Publ_{A_{i,y-1}}}$. We have calculated the values for all areas and found that except for the fluctuations corresponding typically to the beginning years, the fields differ considerably from each other. For example, Computer Architecture (ARCH) and Computer Networks (NET) have stabilized at early 90s, their absolute growths rate values oscillate around 1 ± 0.1. On the contrary, Natural Language Processing and Information Retrieval (NLIR) productivity may vary three times as much from year to year, up to nowadays. Difference in growth rate indices with regard to the various fields have been stated in multiple bibliometric studies throughout the years [46, 79, 131]. However the reasons of such a diversity do not seem to be explained. In the particular case

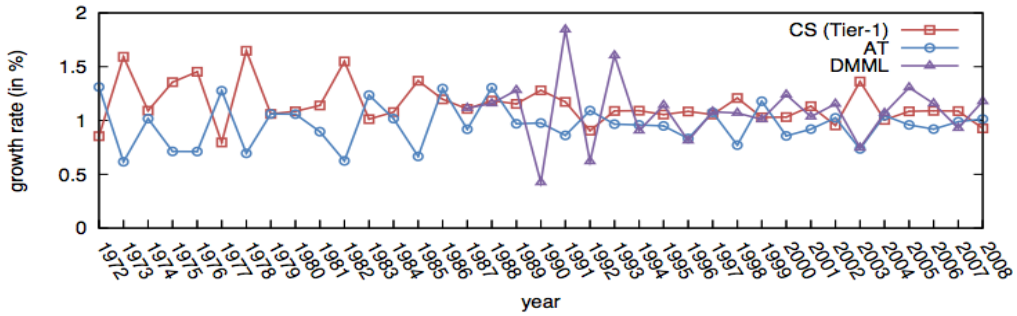

Figure 4.7: Relative growth rates of AT and DMML vs absolute growth rate of CS

of our conference based collection it could probably result from within-venue conventions that define the number of yearly accepted papers. We therefore compare the conferences in our TOP and NONTOP data sets with regard to the absolute publication growth rate. It turns out to be systematically higher in the non-top conferences. We can translate this result in terms of *publication acceptance rates* (information that is typically not present in the bibliographic databases though it is one of the important parameters for conference evaluation [162, 157, 149]), and conclude that they are lower for the top venues.

The *relative growth rate* of an area A_i in year y, $RGr_{A_i,y}$ is a measure of its activity compared to the overall activity in Computer Science (CS)[6]. It is calculated as a ratio between the area absolute growth rate and the computer science absolute growth rate in the given year: $RGr_{A_i,y} = \frac{AbsGr_{A_i,y}}{AbsGr_{CS_i,y}}$ Thus $RGr_{A_i,y} > 1$, indicates a raise of interest to the area A_i in some year y.

Figure 4.7 illustrates the idea. As of CS, we observe considerable fluctuations in its growth rate with the overall tendency to raise in the 70s - 1st half of 80s. One possible explanation is that many areas had started off in that period. At the same time the diapason in conference productivity is large in the beginning, and this is the reason why the curve goes up and down rather than increasing steadily. An additional explanation of the unstable behavior of the curve is the incompleteness of the DBLP data for the corresponding period. On the contrary, influx of the new disciplines becomes much smaller

[6]Here, CS is formally represented by either TOP or NONTOP set. However due to the relatively small size of the NONTOP set and the limited number of areas it contains, we focus rather on the TOP set when discussing this metric.

from the second half of the 80s on, and we notice only two modest spikes - at the end of 90s and in the first years of 2000 which reflect most probably the contribution of the new-born Computational Biology, and World Wide Web.
We chose DMML and AT to visualize the concept of the relative growth rate. On the background of the global development of CS, the bursts of activity in DMML can be seen in the beginning of 90s, and several times in the 2000s, though on a smaller rate. It corresponds well to the evolution of the area which has become very popular in the late 80s - beginning of 90s and attracts a great deal of attention nowadays. On the contrary, relative growth rate in AT remains most of the time bellow one. We suppose that the same considerations that we have mentioned in Subsection 4.4.1 prevent the area becoming "trendy".

4.5.2 Collaboration trends

In Chapters 2 and 3 we studied the effects of collaborative work on the language and topic distributions among scientists. In this chapter we compare collaboration patterns over the various computer science subfields and communities described by the TOP and NONTOP datasets.
Collaborative research has been actively investigated by those who are interested in evaluation and quantification of the research outcomes. Their findings suggest that there is a strong tendency to work by groups in all branches of science [12, 61, 101, 113, 139, 162]. Moreover they show that collaborative works have higher probability to be accepted and cited, attract bigger number of citations and have longer citation history than the papers produced by a single author [6, 12, 138, 142]. At the university and industry levels, international collaborations have been found more valuable than the national ones [47, 64, 131]. We can summarize these findings saying that there seems to be a strong correlation between the fact of working in teams and the quality of the research results. From the perspective of our investigation two questions naturally arise: a) whether various subfields of computer science exhibit similar collaborative behavior, and b) whether or not do the collaborative research and conference rank correlate?
Analysis of collaborations shows how much community is connected. We would expect that a highly interdisciplinary area such as *Data Mining* will exhibit lower connectivity than *Information Retrieval* which is focused on a much smaller number of topics and thus facilitates the collaboration. To

quantify the connectivity we use the average number of coauthors per author, the average number of authors per paper [107], and clustering coefficient [153] introduced in Chapter 3. We use co-authorship graphs G_{DBLP}, G_{Top}, and G_{nonTop} defined in Section 4.3 along with publication statistics to perform these computations.

Previous analysis of the co-author network in ACM data set has shown that the number of collaborators per author increases steadily over the years [162]. It has been confirmed by [61] who used CiteSeer as the experimental testbed. Our results obtained from the DBLP show that the increasing average number of co-authors per authors as well as the average number of authors per paper characterize all the subareas we deal with. Tables 4.4 and 4.5 summarizes our findings.

In the TOP set data, CBIO and WWW have the highest average number of authors per paper along with the highest clustering coefficient which suggest intensive collaborations throughout the entire community. With regard to CBIO, it is interesting to notice that our result is just the opposite from that obtained by Persson et al. [113] on the dataset provided by Medline. The reason lies probably in the network differences caused by the data sources used in these two studies.

Moving forward to the remaining disciplines, we observe that AT, CRYPTO and PL (Programming Languages) have the smallest number of authors per paper, highest percentage of singleton authors (i.e. authors working alone) and the lowest clustering coefficient among all 14 disciplines. It follows that the authors in these three areas have a strong preference for working in small groups when collaborating. Moreover these groups turn to be weakly connected which results in a network composed of rather isolated cliques. It is worth mentioning that [8] found that among other CS areas, CRYPTO has also the highest collaborative *assortativity* [108] which means that the authors tend to collaborate with those authors who have similar number of coauthors.

Surprisingly, the figures in the table do not confirm our assumption about the connectivity of DMML and NLIR. The higher percentage of coauthors per author coming from the same area (63%) in NLIR proves its lower interdisciplinarity compared to DMML where $\approx$ 51% coauthors per author belong to other disciplines. However it does not seem to have an impact on the connectivity pattern, and the clustering coefficient of NLIR is a little smaller than that of DMML. Alternatively it can be explained by the fraction of singletons which is almost twice as much in NLIR as in DMML and

Table 4.4: Collaboration trends in TOP set

Area	Vertexes	# of Authors per paper[a]	# of Coauthors per Author in the same area 1^{st} year	# of Coauthors per Author in the same area average over the entire period	# of Coauthors per Author in TOP set	# of Coauthors per Author in DBLP	% of Singletons	CC
ARCH	12866	1.68-3.05	1.05	5.6	7.9	16.62	3.2	0.71
AT	6882	1.22-2.22	0.48	4.8	9.1	18.3	8.3	0.5
CBIO	6318	2.80-3.36	2.43	4.9	7.8	15.6	2.3	0.79
CRYPTO	2618	1.50-2.27	1.07	4.61	8.09	15.4	7.3	0.57
DB	114173	1.36-2.70	0.6	4.95	8.87	19.1	5.7	0.68
DMML	15730	2.13-2.80	2.9	4.13	8.34	19.35	3.2	0.67
DP	18929	2.05-2.80	1.5	4.36	7.77	19.04	3.1	0.66
GV	10252	2.64-3.02	3.85	4.26	5.8	16.9	2.5	0.67
NET	10769	1.94-2.84	1.25	3.98	6.93	17.61	2.4	0.66
NLIR	8577	1.56-2.63	1.52	4.71	7.45	16.74	7.0	0.66
PL	7173	1.77-2.35	1.55	3.74	8.01	18.3	8.7	0.61
SE	9328	1.90-2.54	1.5	3.83	7.29	18.64	7.3	0.64
SEC	3119	1.52-2.62	0.72	3.7	9.36	21.21	6.1	0.68
WWW	9746	2.79-3.12	2.71	3.75	8.01	21.58	2.4	0.74

Table 4.5: Collaboration trends in NONTOP set

Area	Vertexes	# of Authors per paper	# of Coauthors per Author in the same area 1^{st} year	# of Coauthors per Author in the same area average over the entire period	# of Coauthors per Author in NONTOP set	# of Coauthors per Author in DBLP	% of Singletons	CC
DB	2893	2.03-2.39	2.0	2.84	3.7	7.9	5.2	0.62
AT	2761	1.66-1.97	0.95	2.17	3.5	8.09	14.9	0.55
CBIO	4866	2.90-3.26	3.07	4.3	4.7	8.87	2.0	0.83
DM	9434	2.53-2.87	2.43	3.22	3.57	8.09	3.2	0.71
SEC	1727	2.07-3.01	1.58	3.34	3.78	8.34	3.4	0.69
WWW	6205	2.38-3.04	2.17	3.73	4.2	9.13	3.8	0.75

[a]In Tables 4.4, 4.5, the average number of authors per paper is given by the tuple $\langle 1^{st}$ year of an area, 2009$\rangle$.

naturally lows down the connectivity rate of the former. The weak relation between the interdisciplinarity of a field and its connectivity is best seen with {GV (Graphics), SEC (Security)} pair. The clustering coefficient of both is slightly above average (0.67 and 0.68 vs 0.65). At the same time GV is the most homogeneous area out of all 14 (73% of coauthors per authors belong to GV), while SEC is the most heterogeneous one: only 40% of coauthors per authors come from the same discipline.
The data in Table 4.4 reveals that on average only 43% of coauthors per author belong to the set of authors publishing at top ranked conferences. It is in line with the author/venue distribution discussed in Subsection 4.4.3, and confirms that the same researchers publish at top and non-top ranked venues. In general, the NONTOP set (Table 4.5) is featured by the slightly higher number of authors per paper and higher clustering coefficient (DB is an exception), although the values are close in both sets. Note also that if we were to sort the areas by the clustering coefficient, the order would be the same as in the TOP set (DB and DMML switched around). We therefore conclude that in the given collaborative setting our results do not reveal whether or not the non-top ranked conferences exhibit distinctive behavior compared to the top-ranked ones.

4.5.3 Population Stability

In Section 4.4 we discussed area interdisciplinarity as suggested by author transitions between the fields. In this section we concentrate on the mechanisms that influence researchers' dynamics. For this we analyze changes in conference populations in terms of new members that join a venue (*newcomers*), and those who leave it, *leavers*. In the context of this section, the large *communities* corresponding to the research areas are decomposed into the conferences each of which is understood as an individual community.
The information flow in a community is controlled by so-called *hubs of collaboration* [162] — that is, people known to many other people. For example, Persson [113] found that "two scientists are much more likely to have collaborated if they have a third common collaborator, than are two scientists chosen at random from the community". According to Backstrom et. al. [3], the membership in a community is driven by the similar social process under which the fact of having "friends" may influence the decision to join that community. Thus some researchers are more likely to submit their paper to a conference if they have previously coauthored with someone who had already

published over there. This theory has been tested on the LiveJournal and DBLP (set of 84 conferences with at least 15 years history) communities, and proved valid. We take on this approach and investigate whether this property holds equally in different areas and venues. We therefore define:

- Newcomer $New_{c_{k,y}}$: an author who had no publications at conference c_k before year y. We define a fraction of newcomers in a conference c_k in the year y as $NewComers_{c_{k,y}} = \frac{\sum New_{c_{k,y}}}{TotalAauthors_{c_{k,y}}}$;
- Pure newcomer $Pnew_{c_{k,y}}$: an author who had neither publications nor has he coauthored with an author already member of c_k before year y. The pure newcomers are calculated as $PnewComers = \frac{\sum Pnew_{c_{k,y}}}{NewComers_{c_{k,y}}}$;
- Leaver $Leaver_{c_{k,y}}$: an author who has no more publications in c_k after year y. The fraction of leavers in $c_{k,y}$ is formalized as $\frac{\sum Leaver_{c_{k,y}}}{TotalAauthors_{c_{k,y}}}$.

The most interesting results of the computations are given in Tables 4.6, 4.7.

Let us discuss some of the TOP set conferences. All venues in AT and CRYPTO prove stable and moreover are the most stable venues in the whole TOP set. They are characterized by low percentage of Newcomers, Pure newcomers, and Leavers, compared to the average values across the whole TOP set. Note that fraction of Pure newcomers is an important parameter as it sheds light on how "friendship" phenomenon affects the inflow of the new authors: the higher the fraction is, the smaller is the friendship influence. We have found that AT and CRYPTO are friendship driven as about 50% of new authors joining venues have co-authored with authors who had already published over there.
Contrarily to the two fields above, WWW conferences are the most dynamic ones, featured by the high values for the Newcomers, Pure newcomers, and Leavers' fractions. Friendship does not seem to alter the influx of new authors as the Pure newcomers typically count for $\approx 60 - 80\%$ of all the Newcomers. Note that the member conferences are young - except of ISWC that has started off in 1997 all other venues have appeared in 2000s. It is natural to postulate that the population stability of a venue is directly related to its age. In the given set of conferences, our assumption is immediately confirmed by the ISWC which has the lowest values for all three aspects. Note however that the above relation holds in many but not all the cases. Thus for example

Table 4.6: Population stability in TOP set

Area	Conference	1^{st} year	Average NewComers	Average PnewComers	Average Leavers
ARCH	FCCM	1995	0.72	0.53	0.70
	HPCA	1995	0.65	0.44	0.63
	ICCAD	1990	0.56	0.31	0.54
	ISCA	1973	0.64	0.45	0.59
	MICRO	1987	0.63	0.44	0.59
	ASPLOS	1982	0.78	0.56	0.74
	DAC	1985	0.61	0.38	0.57
AT	FOCS	1970	0.48	0.44	0.41
	ISSAC	1988	0.49	0.57	0.48
	LICS	1986	0.53	0.54	0.51
	SODA	1990	0.51	0.39	0.42
	STOC	1970	0.44	0.43	0.38
	COLT	1988	0.44	0.48	0.40
	SCG	1986	0.45	0.32	0.41
CRYPTO	EUROCRYPT	1982	0.48	0.45	0.46
	FSE	1993	0.50	0.47	0.46
	ASIACRYPT	1990	0.60	0.56	0.58
	CHES	1990	0.64	0.63	0.64
	CRYPTO	1981	0.47	0.45	0.46
	PKC	1998	0.63	0.53	0.61
	TCC	2004	0.52	0.29	0.49
DMML	ECML	1987	0.74	0.72	0.64
	ICDE	1984	0.63	0.44	0.55
	ICML	1988	0.60	0.51	0.52
	KDD	1994	0.67	0.53	0.59
	PAKDD	1998	0.74	0.67	0.68
	CIKM	1992	0.76	0.65	0.68
	ICDM	2001	0.75	0.66	0.69
NLIR	EACL	1983	0.82	0.8	0.76
	ECIR	1997	0.76	0.7	0.65
	ACL	1979	0.66	0.64	0.52
	SIGIR	1971	0.64	0.63	0.55
	SPIRE	1998	0.67	0.66	0.65
	TREC	1992	0.49	0.40	0.43
	NAACL	2001	0.74	0.59	0.61
SEC	ESORICS	1990	0.77	0.69	0.72
	NDSS	1997	0.78	0.64	0.75
	CCS	1993	0.73	0.61	0.58
	CSFW	1988	0.55	0.59	0.50
	S&P	1980	0.75	0.65	0.70
WWW	ISWC	1997	0.70	0.57	0.68
	EC-web	2000	0.82	0.80	0.85
	ICWE	2003	0.71	0.73	0.76
	IEEEWIC	2001	0.82	0.79	0.79
	WWW	2001	0.73	0.58	0.70
	WISE	2000	0.83	0.75	0.83

in Security, CSFW (1988) is less dynamic than S&P (1980), and ICCAD (1990), the most stable community in Architecture, is much younger than ISCA (1973) which scores second in terms of stability. The interpretation of these observations is that while population stability does depend to the certain extent on the conference age, it is also influenced by other, conference specific factors.

The key observation concerning the NONTOP set of venues, is that all of them irrespective of time span (which ranges from 17 to 3 years) and domain, are very dynamic. (The only exceptions are ICCS and DLT (AT) whose behavior is closer to AT venues from the TOP set). Typically the Newcomers constitute about $75 - 85\%$ of all authors, and the average value of the Pure

Table 4.7: Population stability in NONTOP set

Area	Conference	1^{st} year	Average NewComers	Average PnewComers	Average Leavers
AT	APPROX	1998	0.75	0.72	0.64
	ICCS	1992	0.53	0.59	0.48
	SOFSEM	1995	0.82	0.83	0.79
	TLCA	1993	0.66	0.74	0.65
	DLT	1993	0.56	0.66	0.54
DM	MLDM	1999	0.85	0.86	0.75
	IndCDM	2001	0.86	0.84	0.75
	ADMA	2005	0.85	0.75	0.87
	KES	1997	0.79	0.75	0.75
	IDEAL	2000	0.81	0.75	0.81
SEC	SCN	2002	0.75	0.71	0.74
	ISCISW	1997	0.83	0.75	0.83
	ISPEC	2005	0.65	0.58	0.84
	ACISP	1996	0.86	0.76	0.62
	WISA	2003	0.84	0.79	0.87
WWW	WEBIST	2005	0.89	0.90	0.88
	SAINT	2001	0.81	0.72	0.78
	WECWIS	1999	0.84	0.81	0.81
	ESWC	2004	0.75	0.62	0.69
	ICWE	2003	0.71	0.73	0.76

newcomers is about 75% which suggests that the friendship influence on the decision to join a venue is rather negligible. The turnover of authors is also remarkable since the fraction of Leavers is often comparable to that of Newcomers and constitutes up to 88% of all the authors. As such, population stability might be considered as a candidate feature that helps to distinguish between the top and non-top venues.

4.6 Conclusions and Future Work

In this chapter we have analyzed computer science communities in different settings. We have performed statistical analysis of authors, and found that the DBLP community is dominated by the short-time researchers whose career does not exceed 5 years. We have also discovered that scientists from the top-ranked venues tend to join multiple research communities and that the long-term researchers produce the highest number of publications between the fifth and fifteenth years of their career. Typically they publish in a mixture of top and non-top ranked venues.

We have also compared communities from 14 research areas of computer science in terms of publication growth rate, collaboration trends and population stability, and have shown that the disciplines are not always alike. In addition, we applied the same criteria to the comparison between the top and non-top ranked conferences and discovered that the publication growth

rate and population stability could count among the features that help to separate the two sets.

An important characteristic of the evaluation criteria considered in this study, is that its individual parameters are either directly available or can be inferred from the bibliographic databases without recurring to the external sources. It has a twofold implication: a) in the local context of this work, it meets our goal of setting up an evaluation criteria based solely on the data available from the bibliographic databases (see Section 4.2); b) given that our experiments require only a basic bibliographic information that is typically present in all databases irrespective of their thematics and scope (see Chapter 1), our investigation could be conducted on a wider range of scientific disciplines also beyond the computer science area. An apparent generality of the evaluation model does not imply the generality of the results and their interpretation, though. Our examples show clearly that they do not only vary from discipline to discipline but are also influenced by the source of data chosen for the experiments.

In this work we have manually divided the broad area of computer science into 14 topics. A better alternative would be to substitute this rather ad hoc approach by an automatic partitioning of the entire dataset into topics. In DBLP this could be done based on the conference and publication titles by applying for example, *Latent Dirichlet Allocation(LDA)* [9] technique. Compared to the other classification mechanisms, LDA has a number of advantages: it can be used for both - classification and learning the best number of topics into which the given data can be divided; it can efficiently deal with the synonymy and polysemy — two natural properties of language that are typically difficult to be dealt with for the majority of classification algortithms. An automatic classification will help to avoid the subjectivity caused by the manual assignment of conferences to topics and ranks. It will also allow for a deeper insight into the multi- and interdisciplinarity of the scientific domains.

Another objective of the future research has to do with the very nature of our investigation. Currently it has rather explorative character. An extension of the evaluation criteria and their incorporation into a learning algorithm would provide our approach with a predictive capacity.

Chapter 5

Summary and future work

In this work we have proposed a multi-faceted analysis of the bibliographic data using the *Digital Bibliography and Library Project — DBLP* as a testbed. We have examined the data from the linguistic, text mining and bibliometric perspectives and put our investigation in the social context of the research collaboration patterns.

In the framework of the linguistic analysis we have first developed a system that classified personal names originating from 14 languages with high accuracy measured in terms of recall and precision. Then we applied it to the classification of more than $600,000$ names recorded in DBLP. In that scenario along with the correctly classified names there were cases of misclassification due to the high number of names that belong to the languages outside of the system's scope and those names that have a mixed nature represented by the components coming from different languages. We have shown that the last problem can be alleviated by taking advantage of the bibliographic information and its structure, namely by extending the pure language approach underlying our classification system with the co-author network built from the bibliographic records. We have also demonstrated how our system can be used in the real-life settings such as data cleaning and trends discovery.

One of the crucial steps toward the improvement of the system's performance is to increase the number of languages it operates on. Working with such multilingual collection as DBLP or any other international databases that contain personal names, requires an ability to identify all Eastern European languages, Korean, Indian, Persian and many more. The work in this direc-

tion will not only include construction of the appropriate corpora but also finding the transliterating tools which is often the most difficult part.

From the text mining perspective we conducted series of experiments aiming at topic discovery and ranking in DBLP. The goal of this study was to not only identify the research topics but also to find the ways of organizing them in response to the given practical needs: in some cases one may want to know what are the most popular broad research directions, while in some other situations the more specific, treated by rather small researchers' groups, topics will be in focus of interest.

Like in the previous scenario we performed a statistical analysis of the pure textual data coming in this case from the publication titles, along with the analysis of information derived from the co-author networks. This mixed approach made our methods suitable for a variety of document collections. In particular, our methods for topic generation and refinement can be applied to any kind of textual data, stretching beyond the scientific publications organized in the form of bibliographic records. On the contrary our topic re-ranking mechanisms benefit from the explicit or implicit information available from the bibliographic databases.

An interesting and useful future investigation can be done in order to achieve a more fine-grained topic classification that would allow to distinguish between real topics, concepts and technical terms specific to a particular area or more general ones which are used across the fields. This step will require new methods as well as extension of the textual information in order to boost semantic analysis that has rather limited power when applied to short texts like publication titles.

Our bibliometric investigation was aimed at the analysis of researchers and research communities in DBLP within the framework of the top and non-top ranked conferences. On the individual author's level it appeared that the young researchers with ≤ 5 years scientific career constitute the majority of the DBLP population. With regard to the researchers from the set of top-ranked conferences, they have multidisciplinary interests and work in more than one area throughout their entire career irrespective of its length. Those following the academic path are most active between the $5th$ and $15th$ year of their career, and tend to publish at both – top and non-top ranked conferences.

We analyzed the research communities in terms of the publication growth rate, collaboration trends and population stability. The results have been interpreted with two main purposes in mind: a) to compare the communities and b) to check whether or not there is a relationship between the above parameters and conference quality. According to our findings, publication growth rate and population stability seem to correlate with the conference level and as such could make part of the conference evaluation criteria.

Our work leaves room for further improvement. On the methodological level, the topic selection and conference assignment to topics have been done manually. In the future it would be more appropriate to use one of the machine learning algorithms for an automatic selection of data. Latent Dirichlet Allocation — a recently developed classification technique seems to be suitable for this task. On the content level, an interesting further research could be done toward the extension of the evaluation criteria and building a fully automated stand alone tool for the conference ranking.

In this work we have concentrated on the extraction of meta-information from the bibliographic databases using various approaches that can be used in combination strengthening each other and producing better results.

Appendix A

Language identification of personal names

In this appendix we present some examples of the language classification of personal names performed by our system that we have described in Chapter 2. The names are originating from the DBLP authors' list that we have used for our system's application to DBLP.
Figure A.1 represents the first fifty names with the highest rank with regard to French, while the Figure A.2 represents the first sixty top-most ranked names with regard to German. The figures should be read is follows: the left-most column of numbers indicates the score obtained by the name in the target language – French and German respectively. The number between the parentheses shows the second best choice. Thus for example for *Jean-Jacques le Jeune* from Figure A.1 it is -2.27 in Dutch, and for *M. Deutscher* from Figure A.2 it is -2.65, also in Dutch. The third column quotes the author's name in its original form, as it appears in DBLP, followed by the form used by the system while processing the names. As we have explained, a certain minimal name's length has to be satisfied in order for the name, or its component, to be retained. That is the reason why *R. Champagne* (Figure A.1, fourth row) is transformed into *champagne*. Note also, all the names are lowercased. The figures in the remaining 13 columns demonstrate the name's scores with respect to the 13 out of 14 languages. Notice, that the language a name is attributed to is given in the left-most column once. The names are sorted in descending order with respect to the name's best score.

Figure A.1: The first fifty top-ranked names attributed to French.

-1.47 (-2.27 du)	Jean-Jacques le Jeune (| jean jacques le jeune)	-808.4	-52.6	-2.3	-102.7	-152.5	-3.2	-152.4	-52.8	-152.3	-751.2	-52.8	-252.4	-103.0
-1.47 (-2.27 du)	Jean-Jacques Le Jeune (| jean jacques le jeune)	-808.4	-52.6	-2.3	-102.7	-152.5	-3.2	-152.4	-52.8	-152.3	-751.2	-52.8	-252.4	-103.0
-1.58 (-2.27 du)	Jean-Pierre de la Croix (| jean pierre de la croix)	-418.5	-3.0	-2.3	-2.5	-139.1	-3.1	-93.4	-93.9	-2.8	-637.6	-2.9	-139.0	-139.3
-1.60 (-2.19 du)	R. Champagne (| champagne)	-751.6	-3.3	-2.2	-3.5	-252.8	-3.2	-2.5	-2.8	-2.6	-751.7	-3.2	-3.5	-252.8
-1.60 (-2.19 du)	F. Champagne (| champagne)	-751.6	-3.3	-2.2	-3.5	-252.8	-3.2	-2.5	-2.8	-2.6	-751.7	-3.2	-3.5	-252.8
-1.60 (-2.19 du)	D. Champagne (| champagne)	-751.6	-3.3	-2.2	-3.5	-252.0	-3.2	-2.5	-2.0	-2.6	-751.7	-3.2	-3.5	-252.0
-1.64 (-2.25 sp)	Camille Constant (| camille constant)	-667.7	-2.4	-2.3	-2.3	-2.9	-2.3	-2.4	-2.6	-2.4	-933.7	-2.4	-3.0	-3.3
-1.65 (-2.59 en)	T. Champion (| champion)	-716.2	-2.6	-2.8	-3.6	-146.0	-3.8	-3.0	-2.8	-2.9	-573.7	-3.5	-3.2	-3.6
-1.65 (-2.59 en)	R. E. M. Champion (| champion)	-716.2	-2.6	-2.8	-3.6	-146.0	-3.8	-3.0	-2.8	-2.9	-573.7	-3.5	-3.2	-3.6
-1.65 (-2.59 en)	D. Champion (| champion)	-716.2	-2.6	-2.8	-3.6	-146.0	-3.8	-3.0	-2.8	-2.9	-573.7	-3.5	-3.2	-3.6
-1.67 (-2.41 sp)	T. de la Rue (| de la rue)	-252.4	-4.1	-3.0	-2.4	-128.1	-3.4	-3.4	-4.1	-3.7	-581.7	-3.6	-4.0	-4.6
-1.68 (-2.41 du)	Dominique Jean (| dominique jean)	-463.8	-2.6	-2.4	-2.7	-232.6	-2.7	-3.0	-2.6	-79.4	-617.0	-79.3	-79.1	-79.3
-1.69 (-2.65 it)	Jean-Pierre Le Cadre (| jean pierre le cadre)	-422.4	-3.5	-55.2	-55.0	-55.6	-3.4	-2.7	-56.0	-55.6	-790.7	-55.7	-3.0	-55.8
-1.69 (-2.26 po)	Dominique Mille (| dominique mille)	-438.1	-2.5	-2.4	-2.3	-216.1	-2.3	-2.5	-2.3	-2.4	-573.2	-73.4	-73.8	-73.4
-1.71 (-2.25 du)	Jacques Mairesse (| jacques mairesse)	-667.8	-3.0	-2.2	-69.6	-69.2	-2.7	-3.2	-2.4	-69.0	-469.1	-69.1	-282.2	-3.3
-1.71 (-2.28 ge)	Jean-Charles Tournier (| jean charles tournier)	-601.2	-2.5	-2.3	-3.0	-53.1	-3.3	-3.1	-2.3	-52.7	-552.0	-3.1	-2.8	-3.3
-1.71 (-2.27 du)	Jean-Jacques Michel (| jean jacques michel)	-722.9	-2.7	-2.3	-3.4	-2.9	-3.4	-2.9	-2.4	-2.5	-668.1	-2.8	-113.6	-3.0
-1.73 (-2.73 en)	Jean Demaison (| jean demaison)	-417.9	-2.7	-2.8	-3.3	-86.6	-3.5	-4.1	-2.9	-169.2	-86.7	-3.4	-86.1	-3.3
-1.73 (-55.03 du)	Jean-Louis Le Maigne (| jean louis le maigne)	-685.0	-55.4	-55.0	-160.6	-100.1	-100.1	-100.2	-100.1	-100.1	-685.9	-100.2	-160.5	-100.5
-1.75 (-2.74 du)	Jean-Christophe Devaux (| jean christophe devaux)	-809.9	-2.7	-2.7	-145.7	-50.7	-50.5	-98.3	-50.2	-50.6	-763.0	-98.1	-98.0	-193.0
-1.75 (-2.32 du)	Jean-Christophe Sainte (| jean christophe sainte)	-667.6	-2.6	-2.3	-97.0	-3.2	-2.7	-50.5	-2.5	-2.7	-573.0	-50.1	-50.1	-145.4
-1.75 (-2.69 du)	Jean-Pierre Toutant (| jean pierre toutant)	-279.7	-3.0	-2.7	-58.5	-114.1	-3.6	-3.3	-58.7	-58.4	-391.3	-58.4	-3.0	-3.3
-1.75 (-2.69 du)	Jean Pierre Toutant (| jean pierre toutant)	-279.7	-3.0	-2.7	-58.5	-114.1	-3.6	-3.3	-58.7	-58.4	-391.3	-58.4	-3.0	-3.3
-1.76 (-2.60 ge)	Dominique Boucher (| dominique boucher)	-314.5	-2.7	-2.7	-65.0	-252.0	-2.8	-2.9	-2.6	-64.9	-581.7	-127.1	-127.0	-127.3
-1.77 (-2.52 du)	Jean-Jacques Vieillot (| jean jacques vieillot)	-908.2	-102.3	-2.5	-3.1	-102.6	-3.2	-102.7	-52.8	-102.5	-801.1	-103.0	-302.3	-102.3
-1.77 (-2.30 du)	Jean-Jacques Charlot (| jean jacques charlot)	-737.4	-2.4	-2.3	-3.1	-2.9	-3.0	-3.0	-2.4	-2.5	-685.8	-2.9	-107.8	-2.8
-1.77 (-2.56 du)	Pierre Maison-Blanche (| pierre maison blanche)	-302.2	-3.1	-2.6	-2.6	-53.4	-2.7	-2.9	-2.8	-102.3	-551.2	-2.8	-52.7	-3.1
-1.79 (-2.65 po)	Philippe Devienne (| philippe devienne)	-688.4	-3.0	-3.0	-65.6	-65.5	-2.7	-2.9	-3.0	-2.9	-688.2	-3.1	-3.2	-65.5
-1.79 (-2.62 du)	Franck Beaucoup (| franck beaucoup)	-787.2	-3.1	-2.6	-217.2	-359.7	-75.3	-145.7	-216.3	-145.5	-1000.0	-216.9	-216.9	-359.4
-1.79 (-2.63 du)	Jean-Marc Boucher (| jean marc boucher)	-626.4	-2.8	-2.6	-190.4	-65.7	-4.3	-3.5	-2.7	-65.1	-581.9	-65.1	-65.1	-65.7
-1.79 (-2.52 du)	Jean-Charles Sanchez (| jean charles sanchez)	-527.4	-2.5	-2.5	-2.6	-188.5	-2.8	-3.1	-2.9	-187.9	-528.2	-188.1	-160.4	-160.4
-1.79 (-2.39 du)	Philippe Fortemps (| philippe fortemps)	-813.1	-65.0	-2.4	-65.5	-65.9	-2.5	-65.4	-2.7	-64.9	-937.6	-64.8	-65.1	-65.7
-1.80 (-2.88 du)	A. de la Cuvellerie (| de la cuvellerie)	-468.7	-4.7	-2.9	-70.3	-70.8	-3.8	-282.0	-3.4	-3.0	-880.7	-135.9	-3.8	-136.9
-1.80 (-2.42 da)	Dominique Millet (| dominique millet)	-468.1	-2.5	-2.5	-2.5	-281.7	-2.7	-2.9	-2.7	-2.4	-681.6	-68.7	-68.9	-68.9
-1.80 (-2.43 du)	Jean-Charles Quillet (| jean charles quillet)	-798.8	-2.7	-2.4	-2.7	-187.9	-2.9	-3.2	-3.0	-55.8	-685.8	-55.2	-55.2	-187.7
-1.81 (-2.49 da)	Nicolas Champion (| nicolas champion)	-667.7	-2.8	-2.7	-2.9	-69.8	-3.8	-2.7	-2.6	-2.5	-734.6	-3.1	-2.9	-3.1
-1.81 (-2.44 du)	Jean Christophe Bour (| jean christophe bour)	-798.8	-2.6	-2.4	-188.8	-3.3	-3.8	-55.7	-2.4	-2.7	-633.2	-55.2	-55.2	-160.4
-1.81 (-2.43 du)	Pierre France (| pierre france)	-419.2	-2.7	-2.4	-2.6	-3.3	-2.9	-2.9	-2.9	-2.7	-917.2	-2.9	-3.0	-3.1
-1.82 (-2.43 du)	Christophe Saint-Jean (| christophe saint jean)	-808.6	-2.6	-2.4	-152.6	-3.2	-3.8	-53.1	-2.6	-2.8	-880.5	-52.7	-52.8	-152.6
-1.82 (-2.47 du)	G. D. de Villiers (| de villiers)	-701.1	-2.8	-2.5	-3.5	-3.4	-3.2	-3.4	-3.3	-2.7	-980.6	-2.7	-2.9	-3.9
-1.82 (-2.47 du)	P. J. A. de Villiers (| de villiers)	-701.1	-2.8	-2.5	-3.5	-3.4	-3.2	-3.4	-3.3	-2.7	-980.6	-2.7	-2.9	-3.9
-1.82 (-2.47 du)	J. H. de Villiers (| de villiers)	-701.1	-2.8	-2.5	-3.5	-3.4	-3.2	-3.4	-3.3	-2.7	-980.6	-2.7	-2.9	-3.9
-1.82 (-2.47 du)	C. de Villiers (| de villiers)	-701.1	-2.8	-2.5	-3.5	-3.4	-3.2	-3.4	-3.3	-2.7	-980.6	-2.7	-2.9	-3.9
-1.82 (-2.47 du)	M. R. de Villiers (| de villiers)	-701.1	-2.8	-2.5	-3.5	-3.4	-3.2	-3.4	-3.3	-2.7	-980.6	-2.7	-2.9	-3.9
-1.82 (-2.47 du)	C. De Villiers (| de villiers)	-701.1	-2.8	-2.5	-3.5	-3.4	-3.2	-3.4	-3.3	-2.7	-980.6	-2.7	-2.9	-3.9
-1.83 (-2.41 ge)	Dominique Fournier (| dominique fournier)	-296.4	-2.6	-2.4	-2.7	-237.2	-2.9	-2.9	-2.4	-61.4	-648.7	-61.3	-61.0	-120.1
-1.83 (-2.59 du)	Jean Mairesse (| jean mairesse)	-501.7	-3.3	-2.6	-86.6	-86.2	-3.4	-3.5	-2.8	-86.8	-86.9	-86.1	-86.1	-3.6
-1.83 (-2.34 du)	Jean-Charles Boulanger (| jean charles boulanger)	-525.1	-2.8	-2.3	-2.9	-98.6	-2.9	-3.1	-2.5	-2.7	-620.7	-2.9	-2.9	-3.3
-1.83 (-2.37 du)	Jean-Marc Pierson (| jean marc pierson)	-501.6	-2.8	-2.4	-127.6	-3.7	-4.2	-3.1	-3.4	-3.0	-626.8	-3.2	-3.1	-3.2
-1.83 (-2.44 ge)	Christophe Boucher (| christophe boucher)	-706.8	-2.7	-2.8	-179.1	-61.8	-3.8	-61.8	-2.4	-61.2	-765.2	-119.8	-119.9	-237.6
-1.84 (-2.55 du)	Charles L. Bernier (| charles bernier)	-644.5	-2.9	-2.6	-2.9	-74.8	-2.9	-3.2	-2.6	-3.1	-715.7	-3.3	-3.2	-3.6
-1.84 (-2.54 du)	Jean Champagnat (| jean champagnat)	-644.8	-3.0	-2.5	-3.7	-145.5	-3.7	-73.9	-2.6	-3.0	-573.5	-3.3	-2.9	-145.6
-1.84 (-2.60 en)	Charles A. Boucher (| charles boucher)	-644.4	-2.6	-2.9	-74.2	-74.7	-3.2	-3.3	-2.7	-74.1	-644.4	-74.3	-74.2	-74.4
-1.84 (-2.60 en)	Charles Boucher (| charles boucher)	-644.4	-2.6	-2.9	-74.2	-74.7	-3.2	-3.3	-2.7	-74.1	-644.4	-74.3	-74.2	-74.4
-1.84 (-2.47 du)	Jean Pierre Fournier (| jean pierre fournier)	-378.4	-2.9	-2.5	-3.1	-55.4	-3.7	-3.1	-2.8	-55.4	-686.0	-3.8	-2.8	-3.2
-1.84 (-2.45 po)	Bruno Marques (| bruno marques)	-667.5	-2.6	-2.5	-2.7	-86.2	-2.5	-2.7	-2.9	-3.8	-584.2	-169.1	-2.9	-85.8
-1.84 (-2.45 po)	Bruno F. Marques (| bruno marques)	-667.5	-2.6	-2.5	-2.7	-86.2	-2.5	-2.7	-2.9	-3.8	-584.2	-169.1	-2.9	-85.8
-1.84 (-2.53 ge)	Jean-Christophe Bouteiller (| jean christophe bouteiller)	-808.6	-82.2	-2.6	-82.9	-82.7	-42.8	-122.9	-2.5	-42.4	-641.3	-82.5	-162.2	-122.6
-1.84 (-2.81 du)	S. de Maisonneuve (| de maisonneuve)	-463.1	-79.8	-2.8	-156.5	-306.6	-156.8	-79.9	-3.8	-306.6	-309.9	-88.8	-310.8	-88.4
-1.84 (-2.41 po)	Enrique Dans (| enrique dans)	-183.5	-2.9	-93.3	-2.9	-275.4	-2.4	-184.1	-3.3	-93.2	-456.7	-184.5	-2.9	-185.2

Figure A.2: The first fifty top-ranked names attributed to German.

-1.52 (-2.65 du)	M. Deutscher (| deutscher)	-875.9	-2.7	-2.6	-377.4	-3.0	-3.5	-3.1	-3.2	-2.9	-626.7	-3.5	-3.3	-4.3
-1.52 (-2.65 du)	R. F. Deutscher (| deutscher)	-875.9	-2.7	-2.6	-377.4	-3.0	-3.5	-3.1	-3.2	-2.9	-626.7	-3.5	-3.3	-4.3
-1.52 (-2.65 du)	J. Deutscher (| deutscher)	-875.9	-2.7	-2.6	-377.4	-3.0	-3.5	-3.1	-3.2	-2.9	-626.7	-3.5	-3.3	-4.3
-1.54 (-2.54 du)	Wolfgang von der Saal (| wolfgang von der saal)	-750.4	-52.8	-2.5	-282.9	-282.4	-253.0	-3.5	-52.8	-102.3	-750.7	-52.6	-102.8	-152.8
-1.56 (-2.41 du)	Wolfgang von der Linden (| wolfgang von der linden)	-591.7	-2.9	-2.4	-139.2	-138.6	-184.9	-3.4	-3.0	-92.9	-682.7	-2.5	-2.7	-93.9
-1.57 (-2.53 du)	Wolfgang Nieder (| wolfgang nieder)	-643.6	-3.1	-2.5	-200.0	-200.4	-431.1	-216.9	-145.0	-216.3	-706.4	-74.0	-74.4	-217.1
-1.63 (-2.14 en)	Mark Williard (| mark williard)	-917.5	-2.1	-2.6	-252.8	-86.2	-169.7	-168.9	-168.3	-2.3	-917.4	-2.4	-86.0	-252.5
-1.65 (-2.36 du)	Mark A. Auslander (| mark auslander)	-769.9	-155.9	-2.4	-233.2	-80.0	-233.1	-156.6	-156.2	-2.4	-770.3	-2.0	-79.2	-3.1
-1.65 (-2.28 du)	Peter Schlichtiger (| peter schlichtiger)	-706.8	-128.3	-2.2	-296.6	-237.7	-238.3	-179.5	-128.3	-119.8	-766.0	-178.7	-128.2	-61.4
-1.72 (-2.68 du)	Rainer A. Deutschmann (| rainer deutschmann)	-589.5	-2.6	-2.6	-296.5	-3.2	-62.0	-3.2	-2.9	-2.6	-589.7	-2.8	-2.9	-3.5
-1.73 (-2.29 du)	Wolfgang Heiden (| wolfgang heiden)	-429.8	-3.0	-2.3	-360.1	-145.6	-360.0	-74.3	-3.0	-145.0	-572.7	-2.5	-2.9	-145.1
-1.73 (-2.58 du)	C. Hamburger (| hamburger)	-751.3	-2.6	-2.6	-3.3	-3.8	-4.1	-3.4	-3.2	-2.6	-876.4	-3.0	-2.9	-3.2
-1.73 (-2.61 no)	Wolfgang Schmidt (| wolfgang schmidt)	-888.3	-2.8	-2.6	-681.8	-335.7	-668.6	-282.8	-2.9	-135.3	-933.7	-2.6	-70.8	-282.7
-1.74 (-2.77 no)	Wolfgang Freund (| wolfgang freund)	-786.1	-2.8	-2.9	-288.7	-431.0	-360.2	-74.6	-145.7	-145.2	-857.6	-2.8	-3.1	-288.0
-1.75 (-2.46 du)	Thomas In der Rieden (| thomas in der rieden)	-632.4	-2.9	-2.5	-3.3	-55.5	-3.3	-3.2	-3.4	-2.7	-475.3	-2.9	-2.7	-108.2
-1.75 (-2.68 du)	David D. Wirtschafter (| david wirtschafter)	-941.5	-61.2	-2.6	-354.8	-237.4	-296.4	-296.3	-128.1	-128.0	-1000.0	-62.0	-2.8	-128.2
-1.75 (-2.58 du)	Walter M. Neuenschwander (| walter neuenschwander)	-708.9	-52.9	-2.5	-482.1	-252.6	-352.2	-282.6	-252.1	-182.4	-781.0	-182.6	-152.2	-282.6
-1.76 (-2.64 da)	K. von der Heide (| von der heide)	-418.6	-3.3	-2.7	-86.9	-3.0	-3.2	-4.0	-3.1	-2.6	-335.8	-2.8	-3.1	-3.2
-1.76 (-2.48 du)	Torsten Schlieder (| torsten schlieder)	-758.6	-65.5	-2.4	-315.4	-190.2	-315.4	-66.1	-66.2	-65.8	-688.6	-65.0	-65.2	-2.9
-1.76 (-2.32 du)	H. Schlichter (| schlichter)	-888.9	-114.2	-2.3	-336.5	-225.2	-225.6	-114.6	-114.2	-113.8	-1000.0	-113.8	-114.3	-3.9
-1.77 (-2.32 du)	Frank Schlosser (| frank schlosser)	-928.9	-2.5	-2.3	-359.2	-74.6	-217.1	-217.1	-145.1	-73.8	-857.7	-74.2	-74.1	-146.3
-1.77 (-2.58 da)	William J. Schreller (| william schreller)	-938.1	-64.6	-2.7	-128.2	-2.7	-190.2	-65.7	-64.9	-2.5	-938.5	-64.9	-65.0	-190.3
-1.78 (-2.55 du)	Thomas Schlichter (| thomas schlichter)	-875.2	-65.2	-2.5	-190.6	-128.0	-128.0	-65.9	-65.4	-64.9	-937.7	-65.0	-65.2	-3.5
-1.78 (-2.53 da)	Peter Geneiner (| peter geneiner)	-386.2	-3.1	-2.6	-3.4	-79.5	-3.9	-79.8	-79.7	-2.5	-463.9	-3.1	-79.8	-3.6
-1.78 (-2.38 du)	Johann H. Schlichter (| johann schlichter)	-813.8	-65.3	-2.3	-253.2	-128.0	-128.4	-65.7	-65.6	-65.8	-813.8	-64.9	-65.2	-3.4
-1.78 (-2.75 du)	Wolfgang Sichermann (| wolfgang sichermann)	-398.8	-3.1	-2.7	-169.9	-169.7	-225.8	-3.2	-3.2	-169.2	-889.3	-58.3	-3.3	-169.6
-1.78 (-2.58 no)	Wolfgang Schreider (| wolfgang schneider)	-647.8	-2.8	-2.8	-355.1	-178.6	-472.7	-62.8	-2.7	-119.7	-824.1	-2.6	-128.3	-296.4
-1.79 (-2.36 du)	Wolfgang von Hansen (| wolfgang von hansen)	-556.5	-3.1	-2.4	-169.8	-169.5	-226.8	-3.5	-3.4	-113.2	-556.7	-2.4	-2.7	-114.1
-1.79 (-2.43 du)	Wolfgang Fischer (| wolfgang fischer)	-734.2	-2.6	-2.4	-269.9	-136.1	-269.7	-3.1	-2.7	-135.5	-933.7	-2.6	-2.8	-136.2
-1.79 (-2.43 du)	Wolfgang B. Fischer (| wolfgang fischer)	-734.2	-2.6	-2.4	-269.9	-136.1	-269.7	-3.1	-2.7	-135.5	-933.7	-2.6	-2.8	-136.2
-1.79 (-61.42 du)	Wolfgang Schlecker (| wolfgang schlecker)	-765.2	-61.5	-61.4	-472.4	-296.2	-473.8	-179.8	-120.0	-237.2	-941.5	-119.9	-61.6	-355.1
-1.79 (-2.51 du)	Wolfgang Lautenschlager (| wolfgang lautenschlager)	-637.3	-40.5	-2.5	-365.9	-184.7	-366.2	-139.8	-93.6	-138.7	-773.5	-40.0	-40.4	-184.2
-1.80 (-2.37 no)	Wolfgang Berger (| wolfgang berger)	-581.4	-2.8	-2.4	-288.7	-146.0	-289.0	-3.0	-2.7	-145.8	-857.7	-2.4	-2.8	-216.9
-1.80 (-73.76 du)	Volker Schlecht (| volker schlecht)	-786.1	-74.3	-73.8	-438.7	-216.8	-288.1	-217.0	-74.4	-74.2	-1000.0	-74.3	-74.5	-217.5
-1.81 (-2.41 du)	Mark Schlager (| mark schlager)	-751.8	-85.6	-2.4	-335.5	-86.2	-336.1	-252.6	-168.9	-85.5	-834.2	-85.6	-85.8	-169.8
-1.81 (-2.41 du)	Mark S. Schlager (| mark schlager)	-751.8	-85.6	-2.4	-335.5	-86.2	-336.1	-252.6	-168.9	-85.5	-834.2	-85.6	-85.8	-169.8
-1.81 (-2.65 du)	Wolfgang Hoffmann (| wolfgang hoffmann)	-688.3	-2.9	-2.6	-378.4	-128.7	-253.6	-66.5	-3.3	-127.4	-875.4	-2.7	-3.2	-127.9
-1.82 (-2.49 fr)	Wolfgang T. Donner (| wolfgang donner)	-572.5	-2.6	-2.6	-217.3	-145.9	-288.9	-3.0	-2.5	-145.2	-644.1	-2.8	-3.1	-217.2
-1.82 (-2.54 du)	Wolfgang Staffen (| wolfgang staffen)	-733.7	-2.7	-2.5	-482.9	-282.5	-336.1	-69.5	-3.1	-135.6	-933.7	-2.6	-2.9	-136.2
-1.82 (-2.67 du)	Wolfgang Bayer (| wolfgang bayer)	-463.4	-2.8	-2.7	-233.6	-157.0	-310.7	-3.2	-2.7	-156.1	-846.9	-2.8	-3.1	-156.5
-1.82 (-61.85 en)	Wolfgang Stolzmann (| wolfgang stolzmann)	-765.3	-61.9	-121.0	-473.1	-238.3	-414.5	-179.3	-179.0	-296.0	-882.8	-237.3	-62.3	-355.4
-1.82 (-2.71 du)	H. Neuenschwander (| neuenschwander)	-693.2	-88.4	-2.7	-617.2	-387.1	-540.6	-318.3	-386.7	-156.3	-464.1	-156.6	-233.8	-309.9
-1.82 (-2.47 du)	Wolfgang Schreiner (| wolfgang schreiner)	-647.9	-3.3	-2.5	-355.1	-296.5	-356.0	-3.2	-61.5	-178.7	-765.4	-120.0	-2.9	-179.1
-1.82 (-2.41 du)	Wolfgang Hannerschmidt (| wolfgang hannerschmidt)	-889.9	-2.7	-2.4	-478.5	-193.2	-431.1	-145.8	-58.3	-97.3	-985.2	-2.5	-58.7	-98.3
-1.82 (-2.59 du)	L. Auslander (| auslander)	-626.8	-252.3	-2.6	-252.7	-128.5	-252.5	-128.4	-252.3	-2.9	-627.5	-3.3	-127.5	-3.4
-1.83 (-2.36 du)	Rudolf Schlangen (| rudolf schlangen)	-667.5	-3.8	-2.4	-269.1	-136.7	-469.8	-69.9	-69.8	-2.5	-668.2	-2.5	-2.7	-136.2
-1.83 (-2.85 du)	Wolfgang Kohner (| wolfgang vohner)	-643.9	-145.5	-2.9	-582.1	-582.1	-645.8	-288.2	-287.9	-288.8	-928.9	-145.5	-145.8	-359.6
-1.83 (-2.58 du)	Wolfgang Walter (| wolfgang walter)	-714.8	-2.6	-2.5	-288.4	-217.1	-438.7	-74.3	-74.8	-145.1	-928.9	-2.7	-2.9	-217.2
-1.83 (-2.39 du)	L. Peter Deutsch (| peter deutsch)	-833.7	-2.7	-2.4	-252.4	-3.1	-3.2	-3.2	-3.8	-2.7	-585.2	-3.3	-3.8	-3.9
-1.83 (-2.65 no)	Matthias an der Heiden (| matthias an der heiden)	-524.9	-58.5	-2.7	-193.2	-145.2	-58.7	-51.8	-2.9	-58.1	-438.4	-2.7	-2.8	-2.7
-1.83 (-3.25 du)	Wissenschaftsrat (| wissenschaftsrat)	-867.4	-137.5	-3.3	-535.6	-269.2	-469.1	-535.9	-278.5	-3.6	-734.6	-3.5	-3.4	-137.4
-1.83 (-2.51 no)	Wolfgang Mann (| wolfgang mann)	-581.2	-3.8	-2.7	-253.2	-169.9	-336.9	-3.6	-3.2	-168.9	-758.9	-2.5	-3.2	-178.8
-1.83 (-2.52 du)	Steffen Schlager (| steffen schlager)	-667.7	-69.6	-2.5	-482.4	-78.3	-269.9	-282.9	-69.9	-69.1	-735.4	-69.1	-69.2	-136.9
-1.83 (-2.76 da)	Alexander F. Auch (| alexander auch)	-693.4	-2.9	-2.9	-88.4	-88.8	-2.9	-88.3	-3.8	-2.8	-769.7	-79.7	-156.3	-79.7
-1.83 (-2.49 da)	Klaus von der Heide (| klaus von der heide)	-612.4	-3.3	-2.6	-114.7	-58.2	-58.7	-3.9	-3.3	-2.5	-612.8	-2.7	-2.9	-2.9

Appendix B

Topic modeling with Latent Dirichlet Allocation

In this appendix we present some preliminary results of our current research. Its primary goal is to obtain an automatic conference classification and to develop an automatic conference ranking mechanism based on the data available from the bibliographic databases.

Latent Dirichlet Allocation (LDA) [9] has been suggested in Chapters 3 and 4 for the automatic classification of topics and conferences. As a first step toward this goal, we have constructed a corpus composed of the titles and abstracts of the conference publications available from the "Lecture Notes in Computer Science" (LNCS) [88] and "ACM Digital Library" (ACM) [1]. The resulting document collection contains $353,883$ items and amounts to $\approx 1,347$ MB. We have preprocessed the corpus in a standard way, removing punctuation marks, functional words, etc. Then we have classified the resulting collection of documents into 40 topics using Latent Dirichlet Allocation algorithm[1]. Selected topics are presented bellow in the form of $< topic\_id, keywords >$ sequences, where keywords are sorted by their relevance to the topic in descending order.

Topic 0 `data database query queries xml databases relational processing schema model spatial management object sources integration objects`

[1]We have used the LDA implementation available from the MALLET package for the topic classification [93].

views language access temporal join storage sql querying complex documents support structure set base schemas operations

Topic 1 data mining clustering patterns algorithm rules set cluster clusters sets algorithms analysis pattern association rule discovery time frequent attributes real knowledge rough series attribute high number temporal databases event streams detection

Topic 2 key scheme protocol schemes protocols signature authentication public encryption secret cryptographic signatures group model keys message random rsa party attacks cryptography private communication attack identity number exchange knowledge proof privacy computation card sharing

Topic 3 web search retrieval user documents content users document query pages digital text page semantic queries ranking relevance relevant data engine topic library collection metadata collections filtering engines evaluation similarity browsing searching sites feedback site

Topic 4 design hardware architecture level high processor implementation software embedded processing simulation fpga architectures time chip synthesis reconfigurable processors instruction speed data programmable core parallel memory low designs logic multi platform set power cost

Topic 6 learning classification neural data feature network training algorithm classifier networks features machine class vector selection classifiers set accuracy model support recognition decision algorithms svm kernel fuzzy function supervised experimental prediction multi classes

Topic 7 model models analysis distribution probability probabilistic estimation statistical random time data simulation parameters measure modeling number prediction evaluation markov measures accuracy sampling stochastic distributions error estimate values metrics process behavior

Topic 12 language programming object oriented languages java objects implementation code programs program level ada design features support type class apl environment compiler machine classes library interface data written abstract functional inheritance dynamic model software modules

Topic 14 algorithm optimization algorithms search genetic evolutionary optimal solution solutions programming local objective heuristic selection function multi solving space set heuristics strategy evolution solve global fitness number population ga functions experimental

constraints
Topic 18 gene protein biological model sequences sequence dna cell cellular molecular structure genes expression networks computational cells data alignment network biology genome proteins dynamics analysis neural evolution structures chemical models species immune interactions evolution
Topic 21 type calculus semantics proof order abstract types theory logic language programs rewriting functional rules theorem languages program algebraic higher proofs formal programming functions algebra transformation recursive termination notion proving prove set correctness specification
Topic 25 language text speech recognition word words natural translation automatic corpus grammar processing english document documents extraction languages retrieval linguistic texts analysis features semantic chinese sentences parsing
Topic 32 medical data brain registration images patient clinical patients health analysis diagnosis image imaging care model segmentation mr ct tissue mri disease healthcare heart treatment cardiac cancer surgery developed volume breast hospital surgical ultrasound magnetic detection

This example demonstrates one of the main strength of the LDA: it produces topics that can easily be interpreted and labeled. For example, Topic 0 refers to Data Bases, while Topic 1 represents Data Mining etc. (Notice that the task of topic labeling is left to the user). Moreover, it allows to focus on a set of topics according to the task at hand rather than treating the entire collection as one unit.

Note that LDA is a "bag of words" model — that is it does not take care of the semantic connections between the words and represent the topics by isolated words. This is of course a simplified way to represent the language. However this problem can be alleviated by using Topical N-grams [151].

B.0.1 Detection of related topics

One of the assumptions underlying the Latent Dirichlet Allocation approach to topic modeling is that a document is a mixture of topics, whereas each topic is present in the document to some extent. Therefore each document is

assigned to a number of topics, and each topic is characterized by some weight ranging from zero to one. Bellow is an example of a preprocessed document[2], followed by its topical assignment given by the $< topic\_id, weight >$ pairs, ordered by weight:

```
Network Analysis the Kinetics Amino Acid Metabolism Liver Cell
Bioreactor The correlation the kinetics amino acids ammonia and
urea liver cell bioreactor runs was analyzed and described network
structures Three kinds networks were investigated correlation networks
Bayesian networks and iii dynamic networks that obtain their structure
from systems differential equations Three groups liver cell bioreactor
runs with low medium and high performance respectively were investigated
The aim this study was identify patterns and structures the amino
acid metabolism that can characterize different performance levels
the bioreactor
```

$< 18, 0.83 >;\ < 7, 0.1 >;\ <32, 0.07 >$

Based on the topic keys from the previous example and our labeling, this document is assigned to "computational biology and bioinformatics", "statistics and probability", and finally, "medicine and health care", where the "computational biology and bioinformatics" (topic 18) is the most important one.

Topic-document distribution allows to decompose the entire document collection into the focused-by-topic subsets. Thus, for the purpose of this example, we have selected all the documents dominated by the topic 18 and constructed a subset of documents focused on the "computational biology and bioinformatics" (2488 documents). Next, we have exploited the topic-document distribution to automatically detect and rank topics in terms of their relatedness to the topic in question. Bellow is an example of five most related to the "computational biology" topics, represented by the $<$ *topic_id (topic label)*, *relative_weight* $>$ pairs, ordered by relative weight:

[2]The document in the example is an abstract of a paper published in the *Proceedings of the 5th International Symposium on Biological and Medical Data Analysis, ISBMDA*, 2004. Note that not all the functional words have been removed from the text. During the preprocessing we have removed the words composed of two and fewer characters, punctuation marks, diacritics, numbers, and made sure that the words are separated by exctly one blank space. The remaining functional and very frequent words ("stop-words") are eliminated during the topic generation process.

6 (Machine learning), 0.16
7 (Probabilistic methods), 0.15
1 (Data mining), 0.15
14 (Evolutionary algorithms), 0.14
21 (Formal methods), 0.14

After combining the topic ids to the topic keys above we can conclude that the three top most topics belong to the fields of Data Mining and Machine Learning along with the probabilistic methods of data analysis. Note that it is in line with the area relatedness results obtained independently based on the author's transition probability between the areas (see Figure 4.3 in Chapter 4). The close relatedness of the computational biology to the Data Mining and Machine Learning that outranks other logical choices, such as for instance "medicine and health care", can be explained by the high number of data mining conferences in our document collection, many of which have a track devoted to the computational biology and bioinformatics.

So far we have dealt with the individual documents and topic-wise document classification. Given that the documents in our corpus represent conferences we can naturally proceed for an automatic conference-wise topic classification. This will be the next step of our investigation.

Appendix C

List of publications

Peer-reviwed publications

1. Biryukov Maria, Cailing Dong (2010). "Analysis of Computer Science Communities Based on DBLP". In *Proceedings of* 14*th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2010*. Lecture Notes in Computer Science, pp 228 - 235.

2. Biryukov Maria (2009). "Topic Detection in Bibliographic Databases". In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, KDIR 2009*, pp. 236 - 242.

3. Biryukov Maria (2009). "Where do the Authors Come from?" *Journal of Digital Information Management*, Vol. 7, No. 4, pp. 211 - 218.

4. Biryukov Maria, Yafang Wang (2008). "Classification of Personal Names with Application to DBLP". In *Proceedings of the Third IEEE International Conference on Digital Information Management, ICDIM 2008.*

5. Biryukov Maria (2008). "Co-author Network Analysis in DBLP: Classifying Personal Names". In *Proceedings of the Second International Conference on Modeling, Computation and Optimization in Information Systems and Management Sciences, MCO 2008.* p. 399 - 408.

Bibliography

[1] The ACM Digital Library. URL: `http://librarians.acm.org/digital-library`

[2] Amsden Alice H., Mona Mourshed (1997). "Scientific Publications, Patents and Technological Capabilities in Late-Industrializing Countries" *Technology Analysis & Strategic Management*, Vol. 9, No. 3, pp. 343 - 359.

[3] Backstrom Lars, Daniel P. Huttenlocher, Jon M. Kleinberg and Xiangyang Lan (2006). "Group Formation in Large Social Networks: Membership, Growth, and Evolution". In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 44 - 54.

[4] Banerjee Satanjeev, Ted Pedersen (2003). "The Design, Implementation, and Use of the Ngram Statistics Package". In *Proceedings of 4th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2003*, pp. 370 - 381.

[5] Baird Leonard L. (1991). Publication Productivity in Doctoral Research Departments: Interdisciplinary and Interdisciplinary Factors". *Research in Higher Education*, Vol. 32, No. 3, pp. 303 - 318.

[6] Beaver Donald DeB. (2004). "Does Collaborative Research Have Greater Epistemic Authority¿'. *Scientometrics*, Vol. 60, No. 3, pp. 399 - 408.

[7] Bilenko Mikhail et. al. (2003). "Adaptive Name Matching in Information Integration". *Intelligent Systems, IEEE*, Vol. 18, No. 5, pp. 16 - 23.

[8] Bird Christian et. al. (2009). "Structure and Dynamics of Research Collaboration in Computer Science". In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 826-837.

[9] Blei David M., Andrew Y. Ng and Michael Jordan (2003). "Latent Dirichlet Allocation". *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022.

[10] Bollen Johan, Marko A. Rodriguez and Herbert Van de Sompet (2006). "Journal Status". *Scientometrics*, Vol. 69, No 3, pp. 669 - 687.

[11] Bonaccorsi Andrea, Cinzia Daraio (2003). "Age Effects in Scientific Productivity: The case of Italian National Research Council (CNR)". *Scientometrics*, Vol. 58, No. 1, pp. 49-90.

[12] Börner Katy et. al. (2005). "Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams". *Complexity*, Vol. 10, No. 4, pp. 57 - 67.

[13] Bornmann Lutz, Hans-Dieter Daniel (2008). "What Do Citation Counts Measure? A Review of Studies on Citing Bahvior.". *Journal of Documentation*, Vol. 64, No. 1, pp. 45 - 80.

[14] Bornmann Lutz (2010). "Mimicry in Science?" *Scientometrics*, Vol. 74, No. 1, pp. 153161

[15] Bornmann Lutz, Hans-Dieter Daniel (2010). "Citation Speed as a Measure to Predict the Attention an Article Receives: An Investigation of the Validity of Editorial Decisions at Angewandte Chemie International Edition". *Journal of Informetrics*, Vol. 4, pp. 83 - 88.

[16] Bozeman Barry, Elizabeth Corley (2004). "Scientists' Collaboration Startegies: Implications for Scientific and Technical Human Capital". *Research Policy*, Vol. 33, pp. 599 - 616.

[17] Bozeman Barry, Sooho Lee (2005). "The Impact of Research Collaboration on Scientific Productivity". In *Social Studies of Science*, Vol. 35 No. 5, pp. 673 - 702.

[18] Braun, T., W. Glänzel and A. Schubert (1989). "Assessing Assessments of British Science. Some Facts and Figures to Accept or Decline". *Sceintometrics*, Vol. 15, No. 3 - 4, pp. 165 - 170.

[19] Brin Sergey, Lawrence Page (1998). "The Anatomy of a Large-scale Hypertextual Web Search Engine". *Computer Networks and ISDN Systems*, Vo. 30, No. 1 - 7, pp. 107 - 117.

[20] Burright Marian (2006). "Database Reviews and Reports: Google Scholar – Science & Technology". In *Issues in Science and Technology Librarianship*, No. 45, Winter 2006. URL: `http://www.istl.org/06-winter/databases2.html`

[21] Bush, Vannevar (1945). "As We May Think". *Atlantic Monthly*, July 1945, pp. 101 - 108. URL: `http://www.theatlantic.com/past/docs/unbound/flashbks/computer/bushf.htm`

[22] Cavnar W. B., and J. M. Trenkle (1994). "N-gram-based Based Text Categorization'. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161 - 175.

[23] Chen Y. et. al. (2006). "Identifying Language Origin of Person Names with N-grams of Different Unit". In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006.*

[24] Chen Chaomei et. al. (2007). "Delineating the Citation Impact of Scientific Discoveries". In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries, JCDL 2007*, pp. 19 - 28.

[25] Christen Peter (2006). "A Comparison of Personal Name Matching: Techniques and Practical Issues". In *Proceedings of Workshops of the 6th IEEE International Conference on Data Mining (ICDM 2006)*, pp. 290 - 294.

[26] Clauset Aaron, Mark E. J. Newman and Cristopher Moore (2004). "Finding Community Structure in Very Large Networks". *Phys. Rev*, Vol. E, pp. 1 - 6.

[27] Cleveland, Gary (1998). "Digital Libraries: Definitions, Issues and Challenges". *Universal Dataflow and Telecommunications Core Program*, Occasional paper No. 8, March 1998, pp. 1 - 8. International Federation of Library Associations and Institutions (IFLA).

[28] Coile, Russel C (1978). "Lotka's Frequency Distribution of Scientific Productivity". *Professional Paper*, No. 216. Center for Naval Analyses, Arlington, Virginia.

[29] "Computer Science Conference Ranking". URL: `http://www3.ntu.edu.sg/home/assourav/crank.htm`.

[30] "Computer Science Conference Rankings". URL: `http://www.cs.ualberta.ca/~zaiane/htmldocs/ConfRanking.html`.

[31] "Conference listing". URL: `http://www.cais.ntu.edu.sg/content/research/conference_list.jsp`.

[32] Costas Rodrigo (2009). "Scaling Rules in the Science System: Influence of Field-Specific Citation Characteristics on the Impact of Individual Researchers". *Journal of the American Society for Information Science and Technology (JASIST)*, Vol. 60, No. 4, pp. 740 - 753.

[33] Crane Diana (1965). "Scientists at Major and Minor Universities: a Study of Productivity and Recognition". *American Sociological Review*, Vol. 30, No. 5, pp. 699 - 714.

[34] Diedrich Jörg, Wolf-Tilo Balke (2007). "The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems'. In *Proceedings of the* 11*th European Conference on Research and Advances Technology for Digital Libraries, ECDL 2007*, pp. 1 - 13.

[35] Diedrich Jörg, Wolf-Tilo Balke (2007). "Topic-Based User Models: Design & Comparison". In *Proceedings of the* 10*th DELOS Thematc Workshop on Personalized Access, Profile Management, and Content Awareness in Digital Libraries.*

[36] Dunning Ted (1993). "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics*, Vol. 19, No. 1, pp. 61 - 74.

[37] Dunning, Ted (1994). "Statistical identification of language". *Computer Research Laboratory Technical Report MCCS*, New Mexico State University, pp. 94–273.

[38] Elmacioglu Ergin, Dongwon Lee (2005). "On Six Degrees of Separation in DBLP-DB and More". *SIGMOD Record*, Vol. 43, No. 2, pp. 33 - 40.

[39] Elmacioglu Ergin, Dongwon Lee (2009). "Oracle, Where Shall I submit My Papers¿'. *Communications of the ACM*, Vol. 52, No. 2, pp. 115 - 118.

[40] Elworthy David (1998). "Language Identification With Confidence Limits". In *Proceedings of the Sixth Workshop on Very Large Corpora COLING-ACL'98*, pp. 94 - 102.

[41] Erkan Güneş, Dragomir R. Radev (2004). "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization". *Journal of Artificial Intelligence Research*, Vol. 22, pp. 457-479.

[42] "Estimated impact of publication venues in Computer Science". URL: `http://citeseer.ist.psu.edu/impact.html`.

[43] Feather John, Paul Sturges, eds (2003). *International Encyclopedia of Information and Library Science*. London: Routledge. p 127.

[44] Feitelson Dror G. (2004). "On Identifying Name Equivalences in Digital Libraries". In *Information Research*, Vol. 9, No. 4. Available online: URL: `http://informationr.net/ir/9-4/paper192.html`

[45] Folino Francesco, Giuseppe Manco and Luigi Pontieri (2006). "Effective Incremental Clustering for Duplicate Detection in Large Databases". In *Proceedings of the 10th International Database Engineering and Applications Symposium (IDEAS 2006)*, pp. 45-52.

[46] Fox Mary Frank (1983). "Publication Productivity Among Scientists: a Critical Review". *Social studies of science*, Vol. 13, pp. 285-305.

[47] Frenken Koen, Roderik Ponds and Frank van Oort (2010). "The Citation Impact of Research Collaboration in Science-Based Industries: A Spacial-Institutional Analysis". *Papers in Regional Science*, Vol. 89, No. 2, pp. 351 - 371.

[48] Gamon Michael (2006). "Graph-Based Text Representation for Novelty Detection". In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing ACL Workshops*, pp. 17 - 24.

[49] Garfield Eugene (1972). "Citation Analysis as a Tool in Journal Evaluation". *Essays of an Information Scientist*, Vol. 1, pp. 527 - 544. 1962 - 1963. Reprinted from *Science* (1972), No. 178, pp. 471 - 479.

[50] Giles Lee C., Kurt D. Bollacker, and Steve Lawrence (1998). "CiteSeer: An Automatic Citation Indexing System". In *Proceedings of the Third Conference on Digital Libraries, DL 98*, pp. 89 - 98.

[51] Giles Lee C., Isaac G. Council (2004). "Who Gets Acknowledged: Measuring Scientific Contributions Through Automatic Acknowledgment Indexing". In *PNAS*, Vol. 101, No. 51.

[52] Greenstein Daniel and Suzanne E. Thorin (2002). "The digital library: A Biography. Strategies and Tools for the Digital Library". *Digital Library Federation, Council on Library and Information Resources*, 2nd edition, 2002, pp. 1 - 78. URL: `http://www.clir.org/pubs/reports/pub109/contents.html`

[53] Grefenstette, G. "Comparing Two Language Identification Schemes". In *Proceedings of 3rd International Conference on Statistical Analysis of Textual Data (JADT)*, 1995.

[54] Gupta B. M., Avinash Kshitij and Charu Verma (2010). "Mapping of Indian Computer Science Research Output, $1999 - 2008$". *Scientometrics*, Vol. 85, No. 1, pp. 361 - 376.

[55] Gusfield Dan (1997). "Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology". Cambridge University Press.

[56] Hagen Nils T. (2010). "Deconstructing Doctoral Dissertations: How Many Papers Does it Take to Make a PhD¿' *Scientometrics*, Vol. 85, No. 2, pp. 567 -596

[57] Han Hui, Hongyuan Zha and C. Lee Giles (2005). "Name Disambiguation in Author Citations using a K-way Spectral Clustering Method". In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 334-343.

[58] Hiemstra Djoerd et. al. (2007). "SIGIR's 30*th* Anniversary an Analysis of Trends in IR Research and the Topology of its Community ". *SIGIR Forum*, Vol. 41, No. 2, pp. 18 - 24.

[59] Hirsh J. E. (2005). "An Index to Quantify an Individual's Scientific Research Output". *PNAS*, Vol. 102, No. 46, pp. 16569 - 16572.

[60] Holger Enst, Christopher Leptien, and Jan Vitt (2000). "Inventors Are Not Alike: the Distribution of Patenting Output among Industrial R&D Personnel". *IEEE Transactions on Engineering Management*, Vol. 47, No. 2, pp. 184 - 199.

[61] Huang Jian et. al. (2008). "Collaboration Over time: Characterizing and Modeling Netweork Evolution". In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008*, pp. 107 - 116.

[62] Ichise Ryutaro, Hideaki Takeda and Kosuke Ueyama (2005). "Community Mining Tool using Bibliographic Data". In *Proceedings of the* 9*th International Conference on Information Visualisation, IV 2005*, pp. 953 - 958.

[63] Ichise Ryutaro, Hideaki Takeda and Taichi Muraki (2006). "Research Community Mining with Topic Identification". In *Proceedings of the 10th International Conference on Information Visualisation, IV 2006*, pp. 276 - 281.

[64] Inzelt Annamária, András Schubert and Mihály Schubert (2009). "Incremental Citation Impact due to International Co-authorship in Hungarian Higher Education Institutions". *Scientometrics*, Vol. 78, No. 1 pp. 37 - 43.

[65] International Movie Database, IMDB. URL: `http://www.imdb.com/`.

[66] Jeh Glen, Jennifer Widom (2002). "SimRank: A Measure of Structural-Context Similarity". In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'02*, pp. 538 - 543.

[67] Jeong Senator, Sungin Lee and Hong-Gee Kim (2009). "Are You Invited Speaker? A Bibliometric Analysis of ELite Groups for Scholary Events

in Bioinformatics“. *Journal of the American Society for Information Science and Technology (JASIST)*, Vol. 60, No. 6, pp. 1118 - 1131.

[68] Jin G. ”The PH corpus of Mandarin Chinese“. URL: `http://bowland-files.lancs.ac.uk/corplang/phcorpus/phcorpus.htm`.

[69] Jurafsky Daniel and James H. Martin (2000). ”Speech and Language Processing“. *Prentice Hall*, 2000.

[70] Justeson John S., Slava M. Katz (1995). ”Technical Terminoligy: Some Linguistic Properties and an Algorithm for Identification in Text‘. *Natural Language Engineering*, Vol. 1, pp. 9 - 27.

[71] Kanagasabi Rajaraman, Ah-HweeTan (2001). ‘Topic Detection, Tracking, and Trend Analysis Using Self-Organizing Neural Networks”. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD'01*, pp. 102 - 107.

[72] Katz Sylvan J. (1994). ”Geographical Proximity and Scientific Collaboration“. In *Scientometrics*, Vol. 31, No. 1, pp. 31 - 43.

[73] Katz Sylvan J. (1999). ”The Self-Similar Science System“. *Research Policy*, Vo. 28, pp. 501 - 517.

[74] Ke Weimao, Katy Börner and Lalitha Viswanath (2004). “Major Information Visualization of Authors, Papers and Topics in the ACM Digital Library”. In *Proceedings of the* 10*th IEEE Symposium on Information Visualization (InfoVis 2004)*.

[75] King David A. (2004). “The Scientific Impact of Nations: What Different Countries Get for their Research Spending”. In *Nature*, Vol. 430, pp. 311 - 316.

[76] Anne Brüggemann-Klein, Rolf Klein and Britta Landgraf (2000). “BibRelEx: Exploring Bibliographic Databases by Visualization of Annotated Contents-Based Relations”. *IV*, pp. 19 - 24.

[77] Kleinberg Jon M. (2002). “Bursty and Hierarchical Structure in Streams”. *Data Min. Knowl. Discov.*, Vol. 7, No. 4, pp. 373-397.

[78] Knorr K. D., R. Mittermeir (1980). "Publication Productivity and Professional Position: Cross-National Evidence on the Role of Organizations". *Scientometrics*, Vol. 2, No. 2, pp. 95-120.

[79] Larsen Peder Olesen, Markus von Ins (2010). "The Rate of Growth in Scientific Publication and the Decline in Coverage Provided by Science Citation Index". *Scientometrics*, Vol. 83, No. 3, pp. 575 - 603.

[80] Lee Dongwon et. al. (2005). "Effective and Scalable Solutions for Mixed and Split Citation Problems in Digital Libraries". In *Proceedings of International Workshop on Information Quality in Information Systems (IQIS 2005)*, pp. 69 - 76.

[81] Lewis S., K. McGrath, and J. Reuppel (2004). "Language Identification and Language Specific Letter-to-Sound Rules". *Colorado Research in Linguistics*, Vol. 17, No. 1, pp. 1 - 8.

[82] Ley Michael (2002). "The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives". In*Proceedings of the 9th International Symposium on String Processing and Information Retrieval, SPIRE 2002*. Springer, Lecture Notes in Computer Science, pp. 1 - 10.

[83] Ley Michael. "The DBLP Computer Science Bibliography: Homepage".`http://www.informatik.uni-trier.de/~ley/db/`

[84] Ley Michael, Patrick Reuther (2006). "Maintaining an Online Bibliographic Database: The Problem of Data Quality". In *Proceedings of Extraction et Gestion des Connaissances (EGC'2006), Actes des Sixièmes Journées Extraction et Gestion des Connaissances, Lille, France, 17-20 janvier 2006, 2 Volumes*, pp. 5 - 10.

[85] Liakata, Maria et. al (2010). "Corpora for the Conceptualisation and Zoning of Scientific Papers". In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*.

[86] Lin Chin-Yew (1997). *Robust Automated Topic Identification: PhD Dissertation*. University of Southern California, 1997.

[87] Llitjos A. and A. Black (2001). "Knowledge of Language Origin Improves Pronunciation of Proper Names". In *Proceedings of EuroSpeech-2001*, pp. 1919–1922.

[88] Lecture Notes in Computer Sciences. URL: `http://www.springer.com/computer/lncs?SGWID=0-164-0-0-0`

[89] MacRoberts Michael H, Barbara R. MacRoberts (1982). "A Re-Evaluation of Lotka's Law of Scientific". *Social Studies of Science*, Vol. 12, pp. 443-450.

[90] Mann Gideon S., David M. Mimno and Andrew McCallum (2006). "Bibliometric Impact Measures Leveraging Topic Analysis". In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries, JCDL 2006*, pp. 65 - 74.

[91] Manning Christopher D. and Hinrich Schütze (2002). "Foundations of Statistical Natural Language Processing". *The MIT Press*, 2002.

[92] Martins, B. and M. - J. Silva (2005). "Language identification in web pages". In *Proceedings of SAC-05, the 2005 ACM symposium on Applied Computing*, pp. 764 - 768.

[93] McCallum Andrew Kachites (2002). "MALLET: A Machine Learning for Language Toolkit". URL: `http://mallet.cs.umass.edu`.

[94] Medoff Marshall H. (2006). "Evidence of a Harvard and Chicago Matthew Effect". *Journal of Economic Methodology*, Vol. 13, No. 4, pp. 485 - 506.

[95] Mei Qiaozhu et. al. (2008). "Topic Modeling with Network Regularization". In *Proceedings of the 17th International World Wide Web Conference (WWW'08)*, pp. 101 - 110.

[96] Menezes Guilherme Vale et. al. (2009). "A Geographical Analysis of Knowledge Production in Computer Science". In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009*, pp. 1041-1050.

[97] Merton Robert K. "The Matthew Effect in Science"(1968). *Science*, Vol. 159, No. 3810, pp. 56 - 63.

[98] Mihalcea Rada, Paul Tarau (2004). "TextRank: Bringing Order into Texts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004*, pp. 404 - 411.

[99] Meyer Bertrand et. al. (2009). "Research Evaluation for Computer Science". In *Communications of the ACM*, Vol. 52, No. 4, pp. 31 - 34.

[100] Moens Marie-Francine, Roxana Angheluta and Jos Dumortier (2005). "Generic Technologies for Single and Multi-Document Summarization'. *Information Processing and Management*, Vol. 41, pp. 569 - 586.

[101] Moody James (2004). "The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999". *American Sociological Review*, Vol. 69, pp. 213 - 238.

[102] Nallapati Ramesh et. al. (2008). "Joint Latent Topic Models for Text and Citations". In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'08*, pp. 542 - 550.

[103] Narin F. (1994). "Patent Bibliometrics". *Scientometrics*, Vol. 30, No. 1, pp. 147 - 155.

[104] Nascimento Mario, Jorg Sander and Jeffrey Pound (2004). "Analysis of SIGMOD's Co-authorship Graph". *SIGMOD Record*, Vol. 32, No. 3, pp. 8 - 10.

[105] Nürenberg, Peter J. et. al (1995). "Digital libraries: issues and architectures". *Digital Libraries*, June 11 - 13, pp. 147 - 153.

[106] Newby Georgy B., Jane Greenberg, and Paul Jones (2002). "Open Source Software Development and Lotka's law: Bibliometric Patterns in Programming". *Journal of American Society for Information Science and Technology*, Vol. 54 (1), pp. 1-11.

[107] Newman M. E. J (2001). "Scientific Collaboration Networks. I. Network Construction and Fundamental Results". *Physical review E*, Vol. 64, No. 016131, pp. 2-8.

[108] Newman Mark, E. J. (2003). "Mixing Patterns in networks". *Phys. Revue* E, No. 67.

[109] Newman Mark E. J. (2004). "Who is the Best Connected Scientist". In Ben-Naim Eli, Hans Frauenfelder and Zoltán Toroczkai eds. (2004), *Complex Networks*. Springer-Verlag. pp. 337 - 370.

[110] Newman Mark E. J., Albert-László Barabási and Dunkan J. Watts (2006). "The Structure and Dynamics of Networks". *Addison-Wesley Publishing Company*.

[111] Olson Gary M. and Judith S. Olson (2000). "Distance Matters". *Human-Computer Interaction*, Vol. 15, No. 2, pp. 139 - 178.

[112] Pepe Alberto (2008). "A Socio-Epistemic Approach to Identify Communities of Scientific Collaboration". *tripleC - Cognition, Communication, Co-operation*, Vol. 6, No.2, pp. 134 - 145.

[113] Persson Olle, Wolfgang Glänzel and Rickard Danell (2004). "Inflationary Bibliometric Values: The Role of Scientific Collaboration and the Need for Relative Indicators in Evaluative Studies". *Scientometrics*, Vol. 60, No. 3, pp. 421 - 432.

[114] Potencier Fabien, Marvin Humphrey (2004 - 2008). "Lingua: Stop words for several languages". URL: `http://search.cpan.org/~creamyg/Lingua-StopWords-0.09/lib/Lingua/StopWords.pm`

[115] Price Derek de Scolla (1976). "A General Theory of Bibliometric and Other Cumulative Advantage Processes". *Journal of the American Society for Information Science* , Vol. 27 (5-6), pp. 292-306.

[116] Project Gutenberg. "Online book catalog". URL: `http://www.gutenberg.org/browse/languages/zh`.

[117] Reitz Florian and Oliver Hoffmann (2010). "An Analysis of the Evolving Coverage of Computer Science Sub-fields in the DBLP Digital Library". In *Proceedings of the* 14*th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2010*, pp. 216 - 227.

[118] Reitz Joan (2004). *Dictionary for Library and Information Science*. Westport: Libraries Unlimited. p. 70.

[119] Reuther Patrick et. al (2006). "Managing the Quality of Person Names in DBLP". In *Proceedings of European Conference on Digital Libraries (ECDL)*, pp. 508 - 511.

[120] Reynolds D. and M. A. Zissman (2003). "Automatic Speaker and Language Recognition". In Proceedings of HLT-NAACL 2003, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.

[121] Rodrigues José F. Jr. et. al. (2006). "GMine: A System for Scalable, Interactive Graph Visualization and Mining". In *Proceedings of of the 32nd International Conference on Very Large Data Bases VLDB'06*, pp. 1195-1198.

[122] Rosen-Zvi Michael et. al. (2004). "The Author-Topic Model for Authors and Documents". In *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence UAI '04*, pp. 487 - 494.

[123] Rowlands Ian (2002). "Journal Diffusion Factiors: A New Approach to Measuring Research Influence". *Aslib Proceedings*, Vol. 54, No. 2, pp. 77 - 84.

[124] Saam Nicole J, L. Riter (1999). "Lotka's Law Reconsidered: the Evolution of Publication and Citation Distribution in Scientific Fields". *Scientometrics*, Vol. 44, No. 2, pp. 135-155.

[125] Sanderson Mark, W. Bruce Croft (1999). "Deriving Concept Hierarchies from Text". In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 206 - 213.

[126] Schwartz Charles (1999). "The Rise and Fall of Uncitedness". *College & Research Libraries*, Vol. 58, No. 1, pp. 19 - 29.

[127] Seglen Per O. (1997). "Why the Impact Factor of Journals Should Not Be Used for Evaluating Research". *BMJ*, Vol. 314, No. 7079.

[128] Seglen Per O. (1999). "The Skewness of Science. *Journal of the American Society for Information Science*, Vol. 43, No. 9, pp. 628 - 638.

[129] Shanahan James G., Yan Qu, and Janyce Wiebe (2010). *Computing Attitude and Affect in Text: Theory and Applications*. Springer, 2010.

[130] Shannon Claude Elwood (1948). "The mathematical theory of communication". *Bell System Technical Journal*, Vol. 27, pp. 379-423, 623-656.

[131] Shin Jung Cheil, William K. Cummings (2010). "Multilevel Analysis of Academic Publishing Across Disciplines: Research Preference, Collaboration, and Time on Research". *Scientometrics*, Vol. 85.

[132] Shinyama Yusuke. "Corpus Tools: Chinese Pinyin to ASCII converter". URL: `http://cs.nyu.edu/~yusuke/tools/`

[133] Sidiropoulos Antonis, Yannis Manolopoulos (2005). "Automatic Ranking Scientific Conferences using Digital Libraries". *Information Processing & Management*, Vol. 41, No. 2, pp. 22 - 29.

[134] Sidiropoulos Antonis, Yannis Manolopoulos (2005). "A Citation-Based System to Assist Prize Awarding". *SIGMOD Record*, Vol. 34, No. 4, pp. 54 - 60.

[135] Sidiropoulos Antonis, Dimitrios Katsaros, Yannis Manolopoulos (2007). "Generalized Hirsh h-Index for Disclosing Latent Facts in Citation Networks". In *Scientometrics*, Vol. 72, No. 2, pp. 253 - 280.

[136] Smeaton Alan F. et. al. (2003). "Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century¿'. In *SIGIR Forum*, Vol. 37, No. 1, pp. 49 - 53.

[137] SIGIR – Special Interest Group on Information Retrieval. URL: `http://www.sigir.org/`.

[138] Smart J. C., A. E. Bayer (1986). "Author Collaboration and Impact: A Note on Citation Rates of Single and Multiple Authored Articles". *Scientometrics*, Vol. 10, No. 5 - 6, pp. 297 - 305.

[139] Sonnenwald Diane H. (2007). "Scientific Collaboration: A Synthesis of Challenges and Strategies". *Annual Review of Information Science and Technology*, Vol. 41, No. 1, pp. 643681.

[140] Souto Maria Aparecida M., Mariusa Warpechowski and Joseé Palazzo M. de Oliveira (2007). "An Ontological Approach for the Quality Assessment of Computer Science Conferences". In *Porceedings of ER Workshops*, pp. 202 - 212.

[141] Spärk Jones, K. (1972). "A Statistical Interpretation of Term Specificity and its Application in Retrieval ". *Journal of Documentation*, Vol. 60, No 5, pp. 493-502, 2004.

[142] Stern Richard E. (1990). "Uncitedness in the Biomedical Literature". *Journal of the American Society for Information Science (JASIS)*, Vol. 41, No. 3, pp. 193 - 196.

[143] Steyvers Mark et. al. (2004). "Probabilistic Author-Topic Models for Information Discovery". In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'04*, pp. 306 - 315.

[144] Sun Yang, C. Lee Giles (2007). "Popularity Weighted Ranking for Academic Digital Libraries". In *Proceedings of 29th European Conference on IR Research, ECIR 2007*, pp. 605 - 612.

[145] Teufel Simone (1999). *Argumentative Zoning: Information Extraction from Scientic Text*. PhD thesis, University of Edinburgh, Edinburgh, UK.

[146] Thomson ISI. URL: `http://thomsonreuters.com/`

[147] Tol Richard S. J. (2009). "The Matthew Effect Defined and Tested for the 100 Most Prolific Economists". *Journal of the American Society for Information Science and Technology (JASIST)*, Vol. 60, No. 2, pp. 420 - 426.

[148] Waister Silva Martins et. al. (2009). "Assessing the Quality of Scientific Conferences Based on Bibliographic Citations".*Scientometrics*, Vol. 83, pp. 133 - 155.

[149] Waister Silava Martins et. al. (2009). "Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science'. In *Proceedings of the 2009 Joint International Conference on Digital Libraries, JCDL 2009*, pp. 183 - 202.

[150] Walford Geoffrey (1983). "Postgraduate Education and the Student's Contribution to Research". *British Journal of Sociology of Education*, Vol. 4, No. 3, pp. 241 - 254.

[151] Wang Xuerui, Andrew McCallum (2006). "Topics Over Time: A Non-Markov Continuous-Time Model of Topical Trends". In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'06*, pp. 424 - 433.

[152] Wang Xuerui, Andrew McCallum and Xing Wei (2007). "Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval". In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007)*, pp. 697 - 702.

[153] Watts Dunkan J., S. H. Strogatz (1998). "Collective Dynamics of "Small-world" Networks". *Nature*, Vol. 393, pp. 440 - 442.

[154] "Wikipedia, the Free Encyclopedia, Lists of People". URL: `http://en.wikipedia.org/wiki/Lists_of_people`

[155] Williams Joseph M. (1986). "Origins of the English Language". ,*Free Press*, 1986.

[156] University of Leipzig. "Leipzig Corpora Collection". URL: `http://corpora.informatik.uni-leipzig.de/download.html`.

[157] Yan Su, Dongwon Lee (2007). "Toward Alternative Measures for Ranking Venues: a Case of Database Research Community". In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, JCDL 2007*, pp. 235 - 244.

[158] Yookyung Jo, Carl Lagoze and C. Lee Giles (2007). "Detecting Research Topics via the Correlation Between Graphs and Texts". In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD'07)*, pp. 370 - 379.

[159] Zaïane Osmar R., Jiyang Chen, and Randy Goebel (2007). "DBconnect: Mining Research Community on DBLP Data". In *Proceedings of the 9th Joint WEBKDD and 1st SNA-KDD Workshop WebKDD/SNA-KDD'07*, pp. 59 - 76.

[160] Zhou Ding et. al (2006). "Topic Evolution and Social Interactions: How Authors Effect Research". In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, pp. 248 - 257.

[161] Zhou Ding et. al. (2007). "Co-Ranking Authors and Documents in a Heterogeneous Network". In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007)*, pp. 739 - 744.

[162] Zhuang Ziming et. al. (2007). "Measuring Conference Quality by Mining Program Committee Characteristics". In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 225-234.