PhD Dissertation

International Doctorate School in Information and Communication Technologies

DIT - University of Trento

A Semi-supervised Approach for Improving Search, Navigation and Data Quality in Autonomous Digital Libraries

Mikalai Krapivin

Advisor:

Prof. Maurizio Marchese

Università degli Studi di Trento

February 2010

Abstract

The current rapid uptake of Autonomous Digital Libraries [56] (both in scholarly and generalist domains) has driven the need for automated procedures for extracting, processing and representing the digital information contained in these digital repositories. Concurrently, the development of Web 2.0 technologies and applications has provided new opportunities and challenges for web-based information system and user interactions: novel features of social interactions as well as new usability and visualization model are having an impact on how people search, navigate and rank digital objects (being multimedia content like music or movies or - as in the case of Autonomous Digital Libraries - digital publications): as an example a simple quick glance at a user-based "cloud of tags" for a given Journal may now provide us a lot of information about its content and/or the way that the specific Journal is perceived by the people reading it.

In this thesis we have tackled two open research dimension in state-of-the-art Autonomous Digital Libraries, namely:

1. Automated key-phrases/tags extraction from digital scientific contributions (papers): the ever increasing dimensionality of modern ADLs does not permit realistic manual documents processing and needs efficient and high quality methods for automatic or semi-automatic key-phrases extractions.

2. Exploration of current metrics and proposal of novel metrics for ranking of digital objects in order to improve the navigation within a large

number of objects present in modern ADL.

Moreover, we have implemented a prototype for a Digital Library interface capable to integrate the tools developed on the base of the results obtained in the above research directions. The prototype supports the user: (1) in the search of documents related to a topic - using the novel semi-automatic key-phrases extractions techniques proposed in this work; and (2) in the navigation and identification of "relevant" documents for a given topic based on a number of user-selectable relevance metrics.

The first challenge we met in our work has been the lack of large, high quality and publicly available document datasets containing both the full text and human (experts) assigned key-phrases to be used for analytical assessment. Thus we constructed one from available public content sources and curated metadata repositories. Our dataset (named in the following as Trento Computer Science (TCS) dataset) consists in a subset of 2000 scientific papers published within 2003 and 2006 in Computer Science domain in the ACM Digital Library. The TCS dataset consists of the full text of papers, curated metadata (authors, title, affiliations, references etc.) and human (both authors and curators) assigned key-phrases. The original paper type (typically PDF, PS or LaTeX) has been processed and transformed into a textual format with the support of commercial pdf-to-text transformations tool and refined with the support of maximum entropy machine learning in order to improve the final quality of the full-texts.

For the semi-supervised key-phrases/tags extraction task we have compared several Machine Learning techniques - namely, Random Forest, Support Vector Machines a novel Fast Local Kernel SVM and the Naive Bayes learning-based system KEA on the same TCS dataset. In particular, we have performed a number of experiments and explored in details the effect of including in the chosen feature sets linguistic and domain specific knowledge. In our experiments, Random Forest has been identified as the most precise method outperforming KEA (used as baseline for key-phrases extraction) by 36% when using a novel feature sets including linguistic and domain specific knowledge. Moreover, compared with the other Machine Learning techniques, Random Forest is the best trade-off between accuracy and computational speed.

The second task taken on by this thesis - navigation and ranking - relates to the large dimensionality of current ADL. In fact, in the presence of a huge quantity of documents connected to a specific topic, it is hard to navigate and find "interesting" contributions. The challenge here is to be able to identify the most important set of papers in a specific topic or for a particular author. A number of used metrics are available for this task, namely citation count for papers and Hirsch-index for authors. We have applied them as well as novel metrics based on the PageRank metric, named PaperRank, Focused PaperRank and PaperRank h-index, that captures - where data is available - more information present in complete citation graphs. As part of our analysis, we have developed methods and tools for qualitatively and quantitatively analyzing metrics that evaluate content and people. We have used them to explore the differences between various metrics as well as to understand in more details what do they measure. We also believe, that these methods and tools could successfully be used to compare rankings in different domains (search engine, review processes, etc.). We have carried out an extensive investigation of the various ranking metrics on the dataset of over 266,000 scientific papers, and related citation graphs. We discovered that the difference in ranking results is indeed very significant for the different metrics and investigated in details the reasons of this difference.

Although initially this research has started as an independent line of research, it has found a significant number of important interactions with the European FET-Open project LiquidPub [1]. A specific prototype of plug-in used for tagged search and ranking has been incorporated in the LiquidPub portal[2] that allows the tagged search and ranking over the whole collection of 266,000 papers.

Keywords:
Page Rank based Indexes, Automated Keyphrases Extraction, Autonomous Digital Libraries, Standard Dataset Preparation.

[1] Grant number 213360, acronym LiquidPublication under THEME 3: FP7-ICT-2007-C FET OPEN.
[2] http://demo.liquidpub.org:8081/ResevalGUI/

Conditio sine qua non

Acknowledgments

I would like to express my gratitude to people who was supporting me, helping me and making me so sure I will do it and defend my PhD. First of all I would like to thank **Professor Maurizio Marchese** for continuous help and support during all my thesis steps, starting from comprehensive exam and finishing with the structure of the final thesis. I want to thank Professor Marchese for giving me support in different life situations, for helping me to buy a car, translating from Italian and many other things, for his patience and for encouraging me.

My appreciate to Professor Fabio Casati, Professor Fausto Giunchiglia, Professor Enrico Blanzieri, Professor Raymond Ng, Mr. Andrei Yadrantsau, Mr. Nicola Segata, Mr. Aliaksander Autaeu for many useful discussions.

I want to kindly thank my good friend Viktor Pravdin for discussing many non work related situations, drinking wine and successful fishing together!

Huge thanks to my wife Tania and my brother Viktor who have been inspiring me all time since I left my home country. Also great thanks to all russian language speaking people in Trento, specially Uladzimir Kharkevich (my "twin brother"), Andrej Somov, Ivan Minakov, Ivan Tankoev, Max Khadkevich, Alexander Birukou, Volha Bryl and Aliaksandr Autaeu for fruitful and interesting work together.

Talking to Professors Lee Giles, Alexander Smola, Raymond Ng and Fausto Giunchiglia was amazing! I want to thank University of Trento for this opportunity.

Contributions and publications

This Thesis has been done in collaboration with various researchers, in particular all papers are written with participation of Prof. Maurizio Marchese, part about scientometric research impact indicators is done in tight collaboration with Ptof. Fabio Casati. The text mining chapters are complete with help of Andrei Yadrantsau, Aliaksander Autayeu, Prof. Enrico Blanzieri and Dr. Nicola Segata. And in general this work is inspired by various discussions with Prof. Lee Giles and Prof. Fausto Giunchiglia.

Below we list major finding and innovative aspects of the present Thesis. First, regarding to the data mining for Autonomous Digital Libraries:

1. Innovative method of combination of Machine Learning (in particular Random Forrest [8], Support Vector Machines [4], FaLK-SVM [80, 79]) with Natural Language Processing techniques: POS (Part-of-Speech) tagging [94] and parsing dependencies [67] is proposed.

2. Large Dataset for Keyphrases Extraction is prepared [47] and freely shared in my web-page[3]. This dataset consist of 2000 of scientific papers with fulltexts and according meta data. We believe it may set a ground for the future comparison of different techniques for keyphrases extraction. Since each paper in the dataset is converted from PDF, texts contained a lot of "garbage": pieces of LaTeXformulae, pieces of tables, figures etc. We have developed and implemented text refinement procedure based on Maximum Entropy [74, 73] Machine Learning method. It showed 97% of precision in cleaning of garbage.

3. 50% improvement of precision recall and F-measure [97] comparing

[3] http://disi.unitn.it/~krapivin

with KEA [102] baseline method.

For ranking of objects in Autonomous Digital Libraries part we can enumerate the following innovative aspects:

1. We have proposed adaptation of a wellknown PageRank algorithm to the problem of ranking of items in scientific citations graphs. Such adaptation called PaperRank is simpler for computation and easy for embedding into a Digital Library.

2. Tradeoff between PageRank and Citation Count named Focused Page Rank has been proposed for scientific citing.

3. We have proposed how to compute analogue of popular author related metrics – Hirsch Index based on PageRank, or PageRankHircsh.

4. New method of comparison between several ranking schemas called divergence has been proposed.

5. New experimental visualization techniques have been created specially to visualize divergence of several ranking schemas. We apply it to put hundreds thousands of ranked scientific papers in one plot.

6. The visual and theoretical explanations of differences between Citation Count and Paper Rank are done.

Part of the material of the Thesis is published as papers in some conferences, workshops and journals, namely:

Journal papers

1) [53] Mikalai Krapivin and Maurizio Marchese and Fabio Casati, *Exploring and Understanding Citation-based Scientific Metrics*, Advances in Complex Systems (ACS) Journal 2010, accepted for publication Dec 14, 2009.

Conference papers

1) [49] Mikalai Krapivin, Aliaksandr Autayeu, Maurizio Marchese, Enrico Blanzieri and Nicola Segata, *Keyphrases Extraction from Scientific Documents: Improving Machine Learning Approaches with Natural Language Processing*, ICADL 2010, Gold Coast, Australia.

2) [52] Mikalai Krapivin and Maurizio Marchese and Fabio Casati, *Exploring and Understanding Scientific Metrics in Citation Networks*, COMPLEX 2009, ps. 1550–1563, [4], Shanghai, China.

3) [50] Mikalai Krapivin and Maurizio Marchese, *Focused Page Rank in Scientific Papers Ranking*, International Conference on Asian Digital libraries 2008, ps. 144–153, [5], Bali, Indonesia.

4) [45] M. Krapivin and M. Marchese and A. Yadrantsau and Yanchun Liang", *Automated key-phrases extraction from scientific papers using domain and linguistic knowledge*, International Conference on Digital Information Management. ICDIM 2008, ps. 105-112, London, GB.

Workshop papers

1) [46] Mikalai Krapivin, *Focused Page Rank Application for Scientific Papers Ranking in Digital Libraries*[6], European Conference on Digital Libraries/Very Large Digital Libraries Workshop 2008, Aarhus, Denmark.

[4] http://dx.doi.org/10.1007/978-3-642-02469-6_35

[5] http://dx.doi.org/10.1007/978-3-540-89533-6_15

[6] http://www.nmis.isti.cnr.it/manghi/VLDL2008/Krapivin.pdf

Technical reports, DISI University of Trento

1) [51] Mikalai Krapivin, Maurizio Marchese, Fabio Casati, *Exploring and Understanding Citation-based Scientific Metrics*, Technical Report DISI-08-022, DISI, University of Trento, Italy, May 2008[7].

2) [48] Mikalai Krapivin and Aliaksandr Autayeu and Maurizio Marchese and Enrico Blanzieri and Nicola Segata, *Improving Machine Learning Approaches for Keyphrases Extraction from Scientific Documents with Natural Language Knowledge*, Technical Report DISI-10-003, DISI, University of Trento, Italy, 2008[8].

3) [47] Mikalai Krapivin and Aliaksandr Autayeu and Maurizio Marchese, *Large Dataset for Keyphrases Extraction*, Technical Report DISI-09-055, DISI, University of Trento, Italy, 2009[9].

Submitted papers

1) Mikalai Krapivin, Aliaksandr Autayeu, Maurizio Marchese, Enrico Blanzieri and Nicola Segata, *Improving Machine Learning Approaches for Keyphrases Extraction from Scientific Documents with Natural Language Knowledge*, TKDE "Knowledge and Data Engineering" `http://www.computer.org/portal/web/tkde/topics` Journal.

[7] `http://eprints.biblio.unitn.it/archive/00001426/01/022.pdf`

[8] `http://eprints.biblio.unitn.it/`

[9] `http://eprints.biblio.unitn.it/archive/00001671/01/disi09055-krapivin-autayeu-marchese.pdf`

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this thesis we have addressed two open issues in state-of-the-art Autonomous Digital Libraries (ADL), namely:

1. Semi-supervised key-phrases/tags extraction from digital scientific contributions (papers)

2. Exploration of current metrics and proposal of novel metrics for ranking of digital objects in order to improve the navigation within a large number of objects present in modern ADL.

In this Chapter, we first present some current examples of state-of-the-art ADLs. Then we provide a first brief presentation of the problem statement considered in our work and a brief description on the main related open issues.

1.1 Open Issues in Modern Digital Libraries

In the present time, as information storages getting cheaper and bigger, it seems that all humankind knowledge can be stored and preserved in digital format. A large number of various kind of so-called Digital Libraries are appearing, from well-established, reputed and commercial publishers like Elsevier and Springer to more web-based and community-driven endeavors like Google Scholar [17] Citeseer [30], CiteSeerX [31], Rexa [61].

This PhD Thesis is focused to some open problems present in the second type of Digital Libraries – also named Autonomous Digital Libraries and at present mainly devoted to scientific publishing. With the continuous growth of Autonomous Digital Libraries, the problem of management of autonomous knowledge accumulation, warehousing and dissemination is getting more and more actual. One of the most popular knowledge accumulation methods is web crawling [29, 20, 34]. In fact very popular public digital libraries like CiteseerX, GoogleScholar and Rexa have been created exactly in this way. Crawling is a fully automated process that may be enhanced with heuristics. Crawler or "Spider" system starts from a number of "seeds" web sites and then it browses the "adjacent" sites following the links inside web-pages. When crawling millions of journal articles, book chapters, proceedings papers, Master and PhD thesis manuscripts, it is impossible to handle all the information processing (i.e. metadata extraction) manually. That is one of the reasons, people started referring to these type of crawling-based Digital Libraries as "Autonomous Digital Libraries". "Autonomous" stands for absence (or limited presence) of human supervision.

More recently, a number of initiatives (LiquidPub [12], CiteULike [84], Menedeley [90]) are tackling the problem of curating a very large number of digital documents by exploiting the power of the large number of users in social networks, where the whole community of users is involved in the process. Although this kind of community-based improved documents can be partly handled manually, it also can use appropriate automatic procedures of information extraction and retrieval in order to facilitate search and navigation.

1.1.1 Examples of Autonomous Digital Libraries

Let us briefly present few examples of real-world public scientific Autonomous Digital Libraries currently available in the Web.

The most comprehensive initiative at present is Google Scholar [17]: it possesses the largest collection of available internet papers making use of the outstanding performance and coverage of the Google crawler. After being collected by the crawler, papers are analyzed and indexed by the Google search engine. Google Scholar proposes an "user interface" where the user can perform free-text search, or use more advanced search criteria like by author, venue, date and others. The user can also choose to filter by subject area within the following major categories:

1. Biology, Life Sciences, and Environmental Science
2. Business, Administration, Finance, and Economics
3. Chemistry and Materials Science
4. Engineering, Computer Science, and Mathematics
5. Medicine, Pharmacology, and Veterinary Science
6. Physics, Astronomy, and Planetary Science
7. Social Sciences, Arts, and Humanities

Scholar provides some more useful features, for instance each found references may have a link to the source if it is publicly available, or may have a link to the site of publisher, where a referenced paper may be acquired for money. Despite Google Scholar is the biggest ADL in the world it has weak points caused Scholar's automatic way of creation [77]. Result of querying Google Scholar contains unrelated information like projects deliverables, reports, communication notes. To the other end, incomplete information occurrence is possible, for instance Google Scholar cuts authors names, venues, descriptions or title of a paper when they are too long. Names and surnames of authors are not disambiguated in Scholar, so simple query like "John Smith" causes 2,560,000 of results in responce.

A second interesting initiative is CiteseerX [30, 31]: a public Scientific Digital Library initially developed at NEC Research Institute and later at Pennsylvania State University. Initially (under the name Citeseer) the site was focused on Computer Science domain. At present, CiteseerX is rapidly growing in other domains and includes contributions from Chemistry (ChemSeer) and Physics. The coverage of CiteseerX is not so broad as GoogleScholar in terms of quantity of terabytes of information it owes, but it has a wider spectrum of services for users, like presence of bibtex files, relative papers, disambiguated search, extraction of information from tables and figures.

Another interesting ADL example is Rexa [61], a library which is smaller than Google Scholar or CiteseerX in coverage, but has a simple (and thus attractive) but powerful queries mechanism, including search by generic topics (not just few fixed broad categories like Google Scholar proposes). in this regard Rexa is similar to the idea proposed in this thesis.

Finally, there is a number of important commercial Digital Libraries that possess large amount of proprietary content (typically submitted from numerous Conferences, Proceedings, Workshops and Journals). Among many others, IEEE [40], ACM [26], Springer [85] and ScienceDirect (from Elsevier [23]).

1.2 Problem Statement

1.2.1 Automatic and Semi-Automatic Information Extraction in ADL

How can we cope with the exponential increase in the number of digital documents and artifacts? How do we find the relevant and related documents? These issues are central to current information age and have a central role in the Digital Libraries domain. Modern Digital Libraries need automated procedures of texts refinements, text meta information (like authorship, affiliations etc.) extraction, indexing and disambiguation.

Classifications have been used for centuries with the goal of cataloguing and searching large sets of objects. Before some document can be classified, a set of keywords/keyphrases must be defined or retrieved. There are a set of real-world digital libraries that use keyphrases to give a quick impression about a content, for instance free digital library Google Books `http://books.google.com/` has automatically extracted keyphrases in the quick book description. Commercial DL, Amazon `http://www.amazon.com/`, also provides keyphrases when searching a book. However the definition of natural language keyphrases is a time-consuming task for human experts and show its limitations when one tries to scale the process to the very large number of current digital objects.

Machine learning methods are commonly and successfully used to support unsupervised or supervised such information mining tasks. The large majority of research work in data mining domain is dedicated to the extraction of information from web pages, mails, news and typically short and unstructured type of digital content (see for instance [93]). A specific challenge lies in the domain of scholarly papers [37] and is related to the current development of Autonomous Digital Libraries in academia domain [20]. To achieve automated (unsupervised or semi-supervised) information extraction, classification and categorization processes, machine learning techniques are often used [80, 101, 70, 69].

More specifically, the domain of information extraction from scholarly papers contains two broad classes of tasks:

- recognition of structural information which is present inside the paper body (like authors, venues, title, abstract, text parts like sections, tables, figures, author assigned keyphrases (the ones that follow after word "Keywords:" from a new line in a header);

- extraction of information which is only implicitly present, such as generic keyphrases or tags, which are not explicitly assigned by the authors or editors.

First task is well-investigated and state-of-the-art extraction precision is very high (up to 97% obtained by two groups: Giles [37] with help of Support Vector Machines and McCallum [62, 70] with help of Hidden Markov Models or Conditional Random Fields, on a benchmarking dataset taken from Rexa [61] scientific crawler). The second type of extraction (implicit information) is a very different problem and most of the approaches applied to the extraction of explicit information are not suited for it.

In this thesis work, we have focused on the second, more challenging task of extraction of implicit information. Specifically, we analyze the effect of the use of Natural Language Processing (NLP) techniques and the use of specific NLP-based heuristics on the improvement of current Machine Learning (ML) approaches.

Machine learning methods are successfully used to support automatic information extraction tasks for short news, mails, web pages [93], and also for the problem of keyphrases extraction [99, 38, 39]. Key phrases are short phrases representing the main concepts from a document. They are useful for search, navigation and classification of digital content.

The present work extends and improves on a current state-of-the-art approaches and methods, namely:

1. We extend the use of state-of-the-art NLP tools [65] for the extraction, definition and use of linguistic-based features such as part of speech and syntactic relations extracted by dependency parsers [67].

2. We apply the proposed NLP-based approach to a number of different ML methods namely traditional SVM, innovative Local SVM and Random Forests.

3. We analyze in details the effect of different NLP features and dataset size on the overall quality of extracted keyphrases.

4. We perform a comparative analysis of the computed quality measures and obtained keyphrases with the various ML techniques (enhanced with NLP) and the popular Bayesian learning system KEA.

1.2.2 Ranking metrics in ADL

The "curse of dimensionality" causes another interesting problem. There are a lot of scientific domains, which hold many thousands of documents (for instance Google Scholar proposes 21,500 of papers for the "PageRank computation" query). How to recognize the most important papers? How to cut the long tail in this "crowd" of papers? This question leads to the problem of applying different ranking schemas to score research papers or other scientific contributions according to their impact.

The area of scientific metrics (metrics that assess the quality of scientific productions) is an emerging area of research aiming at the following two objectives:

1. measuring scientific papers, so that "interesting" papers can be identified and so that researchers can quickly find useful contributions when

studying a given field, as opposed to browsing a sea of papers; and

2. measuring individual contributions, to determine the impact of a scientist and to help screen and identify candidates for hiring or promotions in both industry Research and Development labs and academia.

Until only 20 years ago, the number of researchers and of conferences was relatively small, and it was relatively easy to assess papers and people by looking at papers published in international journals. With small numbers, the evaluation was essentially based on looking at the paper themselves. In terms of quantitative and measurable indexes, the number of publication was the key metric (if used at all). With the explosion of the number of researchers, journals, and conferences, the "number of publications" metric progressively lost meaning. On the other hand, this same explosion increased the need for quantitative metrics at least to "filter the noise". For example, a detailed, individual, qualitative analysis of hundreds of applications typically received today for any job postings becomes hard without quantitative measures for at least a significant preliminary filtering. Recently, the availability of online databases and Web crawling made it possible to introduce and compute indexes based on the number of citations of papers (citation count and its variations or aggregations, such as the impact factor and the h and g indexes [36]) to understand the impact of papers and scientists on the scientific community. More and more, Universities (including ours) are using these indexes as a way to filter or even decide how to fill positions by "plotting" candidates on charts based on several such indexes.

In this thesis, we have performed an experimental study of scientific metrics (and, in particular, citation-based metrics) with the goal of

- assessing the extent of differences and variations on the evaluation results when choosing a certain metric over another, and

- understanding the reasons behind these differences. Besides "traditional" metrics, we also present and discuss metrics for papers and authors inspired at how the significance of Web pages is computed (essentially by considering papers as web pages, citations as links, and applying a variation of PageRank). PageRank-based metrics are emerging as important complement to citation counts as they incorporate the "weight" (the reputation or authority) of the citing paper and its density of citations (how many other papers it references) in the metric. In addition, the fact that they have been working very well for the Web suggests that they may be insightful for papers as well.

Besides the introduction of the PageRank-based index and its computation algorithm, the main contributions of this thesis lie:

- in the experimental analysis of metrics, so that people and developers in "ranking" papers and people are aware of how much choosing different indexes results in different versions of the truth, and why this is the case, and
- in the identification of a generally applicable analysis method and of a set of indicators to assess the difference between ranking algorithms for papers and people.

We performed the analysis on a dataset consisting of over 266K ACM publications. The analysis was conducted by

1. computing the various citation-based indexes;
2. analyzing the extent of the differences in ranking of papers and people depending on the metric,
3. developing "meta-indexes" whose purpose is to help explore the reasons for these differences, and

4. using these exploration indexes to derive conclusions of when and why page rank and citation measures differ and what to make of this difference.

1.2.3 Availability of a large and high quality document dataset

Although state-of-the-art approaches for keyphrases extraction is quite extensive, it is very hard to compare different methods because:

- there are no publicly available standard data sets of proven quality;
- the majority of reported experiments refer to manually collected papers/news with manually-assigned keyphrases by the same persons conducting the experiments (see Chapter 2);
- the size of a current data sets is limited and usually varies from 10 to 500 documents.

Since supervised or unsupervised Machine Learning is very set-dependent task, it would be very useful to have in this domain a publicly available dataset (like the TREC [16] dataset in the Information Retrieval community) in order to test the various approaches and compare their performances on the same data. In order to perform our experiments - both on key phrases extraction and ranking - we prepared a large and high quality dataset of scientific documents - all in computer science domain - including independent expert-assigned key phrases. We named this dataset Trento Computer Science (TCS) dataset and have publish it [47] for the use of the community.

1.3 Structure of the Thesis

This Thesis is organized in the following six chapters.

1. Chapter 1 presents short introduction to the context and the problem statement of the thesis work.

2. Chapter 2 is devoted to the description of the definition and preparation of a large and high quality data set of scientific documents to be used as the standard set for the subsequent experiments.

3. Chapter 3 introduces the state-of-the-art in keyphrases extraction and presents the proposed methodology - based on the use of Natural Language Processing - to enhance state-of-the-art machine learning approaches. For the evaluation we have uses the data set defined in the previous chapter. Evaluation shows interesting results that outperform state-of-the-art Bayesian learning system KEA improving the average F-Measure from 22% (KEA) to 30% (Random Forest) on the same dataset without the use of controlled vocabularies. Finally, we report a detailed analysis of the effect of the individual NLP features and data set size on the overall quality of extracted keyphrases.

4. Chapter 4 introduces the state-of-the-art in the problem of ranking of scientific documents in citation networks and presents a detailed description and analysis of novel approaches.

5. Chapter 5 describes a prototype system that provides an interface for querying scientific publications using the extracted key phrases and the investigated different ranking metrics.

6. In Chapter 6 we summarize major results and related discussion.

Chapter 2

Trento Computer Science (TCS) Dataset Preparation

This chapter is devoted to a large dataset construction with the aim to use it for the machine learning-based automatic keyphrase extraction task. The constructed dataset has a high quality and contains 2,000 scientific papers from Computer Science domain, all published by ACM [26]. Each paper has several keyphrases assigned by the authors and verified by the reviewers. Different parts of papers, such as title and abstract, are separated, enabling extraction based on a part of text. The content of each paper is converted from PDF to plain text. The pieces of formulae, tables, figures and LaTeX mark up were removed automatically. For removal we have used Maximum Entropy Model based machine learning and achieved 97.04% precision. Preliminary investigation with help of the state-of-the-art keyphrase extraction system KEA shows keyphrases recognition accuracy improvement for refined texts.

We hope it will establish a ground for fair evaluation and comparison of different keyphrase extraction systems.

2.1 Introduction

Modern Digital Libraries like CiteSeerX[31] or Google Scholar[17] contain millions of documents. Typically, the crawler downloads a document, converts it to a plain text format and then extracts all necessary information. For instance automatic extraction of keyphrases is one of the challenges in information extraction task. The state-of-the-art contains complaints about absence of standard benchmarking sets for keyphrase extraction validation and methodology proof [66].

We claim that our dataset is characterized by the following features, necessary for any good *automatically crawled* dataset [32]:

- Correctness, that is the dataset should be correct and of high quality;
- Complexity or "hardness", which addresses the fact that state-of-the-art mining systems mine differently.

Below we will argue in favor of these points regarding to our dataset.

2.2 Existing State-of-the-art Datasets

Let us briefly mention some previous works about keyphrases extraction from the point of view of benchmarking set usage. Chronologically the pioneer in successful keyphrase extraction was Peter Tourney [95]. He proposed very detailed investigation of decision trees based algorithms and several links to freely available datasets. For instance, NEXOR[1], FIPS[2] and others [95]. However, that was more than decade ago and *all* those links are no longer available and we have failed to find any of the proposed datasets

[1] http://www.nexor.com/public/aliweb/search/doc/form.html
[2] http://www.itl.nist.gov/div897/pubs/

in internet. Later work which is one of the most valuable in the domain is KEA[3] [102]. Its algorithm is based on Naive Bayes classifier. KEA is a free software and can be downloaded through KEA website, but there are no standard datasets in the download package. KEA inventors mention that they obtained Tourney dataset directly from the author. Nguen *et al* [66] directly pointed out to the impossibility to find any proper datasets and used their own dataset constructed from **250** crawled documents.

We emphasize that most of the datasets used in state of the art belong to the area of scientific papers. For instance, Tourney [95] used **75** scientific papers from different domains: Neuro Science, Behavorial and Brain Science and Chemistry on one hand. He also used **311** email messages and up to **140** of web pages from different domains. Dataset of a similar size was mentioned in [24]. In the more recent work Tourney proposed **500** of scientific papers from Physics domain taken from arXiv.org e-Print archive[4]. Papers were taken in PostScript (PS) format and author did not mention neither how they converted them to text nor what is the conversion quality. In the [99] authors proposed a dataset consisting of **160** scientific papers without mentioning particular domains. Annette Hulth [39] took **198** pieces of short Swedish texts related to social activities. In previous work she proposed commercial dataset from Inspec[5] [38].

We have recently proposed a novel method combining state of the art Support Vector Machines learning in combination with Stanford NLP Parser [45] upon **400** of scientific papers in Computer science domain published in ACM.

[3] http://www.nzdl.org/Kea/
[4] http://arxiv.org/
[5] http://www.theiet.org/publishing/inspec/

2.2.1 Problems with existing datasets

Let us summarize major dataset related problems in the state-of-the-art. Apart from already mentioned complexity and correctness, we point out the following:

- **Dataset size** is one of the biggest problems of any keyphrase extraction research papers. To the best of our knowledge no one used dataset containing more than 500 scientific papers. We think this is caused by the difficulties in dataset construction and further results evaluation. Most of the trials were done *manually*, which is extremely time consuming. However, increasing the dataset size may lead to significant improvements in precision and recall of tasks based on supervised machine learning methods.

- **Availability and sustainability** are the main problems for most of the datasets considered in the state of the art. It is really hard to compare new algorithms and methodologies with previous work because the results may vary from dataset to dataset drastically. Even taking papers from the same storage and nearly same domain may change the results depending on machine learning method.

In the present chapter we address all these issues:

- we propose the largest dataset in the state-of-the-art (about 2000 documents with full texts);

- the dataset has high quality;

- the dataset is freely available through internet for further competition.

2.2.2 Related work

Creation of benchmarking sets is not a new field. There are some datasets well-known in Information Retrieval. For example, Reuters Dataset[6], prepared by David Lewis. This dataset carries thousands of short news texts with labels and helps to evaluate classification algorithms. Another example is a large dataset called TREC [16] [7]. TREC collection is dedicated to web mining, indexing and query answering. It fits well to semantic search community tasks and has been used in different semantics and Natural Language Processing-based evaluations. There is the dataset constructed by Giunchiglia *et. al.* [32] by crawling (as we do here) for ontology matching. And there are many more samples of datasets available in web.

2.3 TCS Dataset Description

The dataset we present contains papers from Computer Science domain published by ACM [26] in the period from 2003 to 2005. All these papers are written in English and stored in UTF-8 text encoding. Each text has clearly indicated:

- Title.
- Abstract.
- Body.
- References (recognized by our method [41]).
- References crawled from ACM portal.

[6] http://www.daviddlewis.com/resources/testcollections/reuters21578/
[7] http://trec.nist.gov/

- References to citing papers (also taken from ACM).

The separation of the parts enables to use them as an additional training material for training text part recognition. Moreover, they can be used to restrict search for a keyphrase to a part of the text. For example, search can be restricted to abstract and references only [45, 99]. This is convenient for computationally expensive methods like SVM [99].

Each file holds full text of a paper and has the name like "[id].txt" where "[id]" is a valid ACM[8] document id, for instance "1005858.txt" corresponds to a real paper with id "1005858". One may find this paper at http://portal.acm.org and make sure it is a paper "A framework for architecting peer-to-peer receiver-driven overlays" with attached keyphrases "congestion control, peer-to-peer streaming". Keyphrases for particular file are located in file "[id].key". This format is common for machine

learning community and used in KEA [102]. Dataset contains 2304 papers freely available in internet[9]. It is not separated into a training set and a test set, so we presume applying of cross-validating procedure (see for example [68]). The papers full texts were downloaded from CiteSeerX Autonomous Digital Library.

2.4 TCS Dataset Preparation

We took the papers in PDF format from CiteseerX, skipping all corrupted or "unconvertible" PDFs (such as PDF stored as image). Meta-information like titles, references and abstracts was taken from ACM portal. We have mapped ACM meta-information to Citeseer texts on the bases of crawled id mappings and information kindly shared with us by professor Lee Giles,

[8]http://www.acm.org/
[9]http://disi.unitn.it/ krapivin/

creator of Citeseer and CiteseerX digital libraries. Then we converted PDF to plain text using a commercial system, then information was processed step by step as described in [41]. While doing this we have found some "garbage", or lexically meaningless pieces of information, which trapped into texts as a result of double conversion: from LaTeX to PDF and then from PDF to text. While using Natural Language Processing tools may improve keyphrase recognition rate [45, 38] this "garbage" decreases the precision of Natural Language Processing tools. We have used Maximum Entropy Model based tool to eliminate the "garbage".

2.4.1 Garbage cleaning

PostScript and PDF formats are current standard of presenting scientific papers. While they have many advantages of allowing rich formatting, complex formulas and figures to be used, for many tasks requiring natural language processing this presents an additional challenge of extracting plain text out of a PDF document.

Many tools address the issue of PDF to plain text conversion. However, the resulting plain text document often contains remains of LaTeX markup, various extra punctuation symbols, clusters of brackets. For example, Figure 2.1 shows the example of a "garbage" remaining in plain text after the conversion from PDF.

Figure 2.1: PDF converted to plain text.

```
a linear system Ax = b, in which
satisfy k = (M \Gamma1 N), so the iteration
```

These markup and punctuation pieces restrain modern NLP tools from achieving maximum performance and even cause failures in less robust tools. Therefore, it is desirable to clean up this "garbage" from the text.

Figure 2.2 shows how cleaned text looks like. Cleaned in this way text eliminates failures in NLP tools and allows them to achieve better results.

Figure 2.2: Cleaned plain text.

```
a linear system b, in which
satisfy, so the iteration
```

Due to the large size of the dataset, manual cleaning will take a lot of time and is unfeasible. Our approach is to use supervised machine learning. The task of identifying the garbage in a text could be seen as deciding for each token its category, which could be either “text” or “garbage”.

We annotate a small sample of the dataset, consisting of 6 documents containing together about 53,000 tokens. To each token we attach a tag identifying whether it is a “text” token or a “garbage”. The task of classifying text tokens into different categories is well-known in NLP as part-of-speech tagging.

We train and evaluate two state-of-the art part-of-speech taggers, Stanford POS tagger [94] and OpenNLP tools [65] POS tagger on our annotated dataset. Both of them are based on Maximum Entropy Models [74]. We tried several combinations of options available in taggers, however the best performance was achieved using default settings.

For tagging we use approach described in [73]. We use our own very small tag set of 2 tags, namely `T` for text and `G` for garbage. We extract and use the following features to make tagging decisions:

- up to 4 prefixes made of first 4 characters
- up to 4 suffixes made of last 4 characters
- presence of punctuation characters inside a token

- presence of initial capital letter in a token
- presence of digits inside a token
- 2 previous words and their tags
- 2 successive words and their tags

We evaluate both taggers using 10-fold cross-validation on our annotated sample. Table 2.1 summarizes taggers performance.

Table 2.1: POS taggers performance, precision per token, %.

	Overall	Garbage
Stanford	95.55	83.19
OpenNLP	97.04	87.21

For the better performing OpenNLP POS tagger Figure 2.3 shows precision improvement during incremental training.

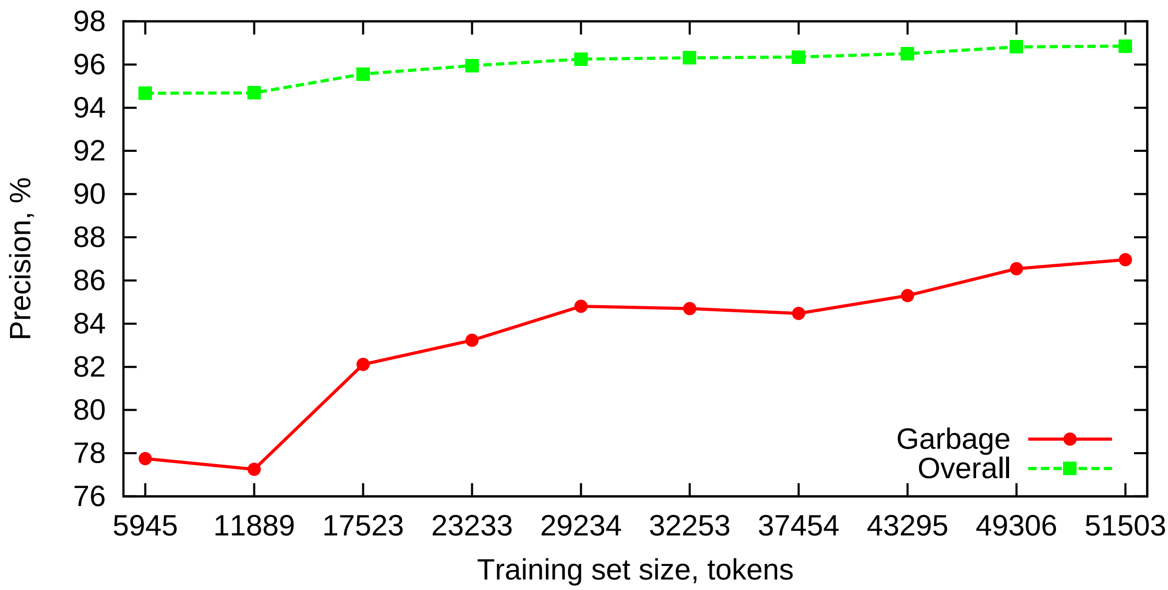

Figure 2.3: Garbage and overall detection precision for incremental training.

Note the stabilization of the overall precision curve of the POS tagger around 97%. While the overall precision stabilizes, we note that solid line showing precision per tag `G`, which indicates "garbage" to remove, is not stable yet. This precision might be improved further by increasing the size of our manually annotated training set.

2.4.2 Correctness and completeness, preliminary evaluation

We evaluated proposed set for KEA [102], and Machine Learning + Natural Language Processing method, recently proposed in [45, 48]. Evaluation shows that both methods, very different by their nature, have the overlap of true positive extracted keyphrases of about 55%. This indicates that different systems mine different keyphrases, so the dataset is "hard" or complete [32].

From other point of view the dataset is "naturally" correct, because it has editor assigned keyphrases of proven quality, and at least one keyphrase appears in each text at least once.

We perform preliminary evaluation with KEA, carrying out the experiments for keyphrases recognition using refined full texts and not refined ones. We see small but stable (*c.a.* 4%) improvement in keyphrases recognition using refined full texts. This is not major aim of the present work, to evaluate the effect of cleaning, moreover, since KEA does not use syntactic knowledge, it cannot improve too much with refined texts. But, since we have removed all "garbage" KEA has less chance to extract linguistically senseless information like piece of tag or formula.

2.5 Discussion

We have prepared and presented a large dataset for keyphrase extraction. The novelties of the dataset are:

- It is at least 5 times bigger than any of previously used datasets.
- It is a set of full texts of scientific papers, which is typical for the keyphrase extraction domain.
- It has author assigned and editor corrected keyphrases.
- It is verifiable and reproducible, because all presented information may be found through CiteseerX and ACM portal.
- It is public and available for use by researchers.
- It is refined for better NLP processing to get more syntactical and semantical knowledge.
- The dataset is hard because it presents different challenges for different state-of-the-art machine learning systems.

The proposed dataset may also be used for classification tasks, because all presented documents have classification labels which may be found on ACM portal. Another possible use of the dataset is the text parts detection.

Chapter 3

Automatic Keyphrases or Tags Extracting in ADL

In this chapter we study the use of Natural Language Processing (NLP) techniques to improve different machine learning approaches (Support Vector Machines (SVM), Local SVM, Random Forests), to tackle the problem of automatic keyphrases extraction from scientific papers. For the assessment we used a large high quality dataset: 2000 ACM papers from the Computer Science domain (see Chapter2). Evaluation shows promising results that outperform state-of-the-art Bayesian learning system KEA improving the average F-Measure from **22%** (KEA) to **30%** (Random Forest) on the same dataset without the use of controlled vocabularies. The assessment is performed by comparison with expert assigned keyphrases. Finally, we report a detailed analysis of the effect of the individual NLP features and data sets size on the overall quality of extracted keyphrases.

We organize the Chapter as follows: In Section 3.1 we present a brief review of the state-of-the-art in the domain and a discussion of relevant related work. Section 3.2 provides a detailed description of the dataset used in our experiments. Section 3.3 presents the details of the proposed extraction methodology (that we name hereafter ML+NLP), specific feature set, text processing tasks and result assessment methodology. In Section 3.4 we present the results of all four approaches, Bayesian Learning (KEA), Support Vector Machine, Fast Local Kernel Machine and Random Forrest training. In Section 3.5 we present a quantitative comparative analysis of the obtained quality measures as well as a qualitative analysis of the extracted keyphrases. Section 3.6 is devoted to the conclusions.

3.1 State-of-the-art

Exponential growth of information in web era made autonomous digital libraries very popular. Autonomy means automatic information harvesting, processing, classification and representation and it brings several challenges.

A specific challenge lies in the domain of scholarly papers [37], accumulated in autonomous digital libraries [20, 29, 56] like CiteSeerX[1][31], Google Scholar[2][17] or Rexa[3][61]. Library spider crawls the web for scientific papers. Having retrieving the paper, the crawler must convert it into a text format. Then it extracts relevant metadata (like title, authors, citations) and finally documents are properly analyzed (classified, identified, ranked, stored). Metadata helps to categorize or classify papers, simplifying and enhancing users' searches, but often metadata is not available explicitly.

Information extraction from scholarly papers contains two broad classes of tasks:

- recognition of structural information which is present inside the paper body (like authors, venues, title, abstract, text parts like sections,

[1] http://citeseer.ittc.ku.edu
[2] `http://scholar.google.com`
[3] http://rexa.info

tables figures, author assigned keyphrases (the ones that follow after word "Keywords:" from a new line in a header) and extraction of it;

- mining of information which is only implicitly present, such as generic keyphrases or tags, which are not explicitly assigned by the authors (which happens quite frequently, specially for the earlier publications).

First task is well-investigated and accomplished with extremely high (up to 97% of F-Measure) performance by two groups in parallel: Lee Giles [88] did it with help of Support Vector Machines and Andrew McCallum performed the same task with help of Hidden Markov Models usage [62] first, and with Conational Random Fields [70] later (the winner method), on a benchmarking dataset taken from Rexa scientific spider system. We emphasize that extraction of implicit information is very different problem, where methods like Conditional Random Fields are unapplicable. But SVM is very universal and fits most of known learning tasks, so it might be adopted to the keyphrases extraction problem.

In this Chapter we focus on the task of extraction of implicit information, which we believe is more challenging. We analyze the effect of the use of Natural Language Processing (NLP) techniques and the use of specific NLP-based heuristics on the improvement of current Machine Learning (ML) approaches.

In the state-of-the-art in keyphrases extraction people usually adopt short news, mails, web pages [93] as datasets, and Support Vector Machines (SVM) [45] (variations of SVM like SVM with unbalanced cardinalities [48] or SVM known as Least Square SVM [99]), decision trees-based [95] methods [38, 39] or probabilistic Bayesian method KEA [102]. Let us define a Keyphrase as a short phrase (from 1 to 5 tokens like "Web services composition") representing a concept of a document. Keyphrases are usually used for search, navigation and classification of various texts.

Let us briefly outline text classification problem and show how keyphrases may help in it. The most common and primitive method of classification is bag-of-word plus vector space representation [76], where a document is converted into a vector of very large dimensionality (typically 5-10 thousands) where component equal to 1 if word presents in document and 0 if not. So distinct quantity of words in a set of documents we want to classify is the dimensionality of a vector space (so-called features space) we work with. Assuming that basis of the feature space is orthogonal, the scalar product of two vectors (documents) indicates how similar documents are. So categorization is simply a similarity between category vector and document vector. In this primitive approach we cannot deal with keywords or keyphrases. But if we will consider phrases as instead of simple tokens, and weight each phase more if it is a keyphrase we will get more accurate classification [33, 78]. So as we see, keyphrases are indeed useful for text categorization task.

The most popular state-of-the-art system for keyphrases extraction is KEA [102]. It uses Naïve Bayes classificator and few heuristics, namely filtering out keyphrases that are subphrases of more narrow keyphrases, using particular stemmer, creative text tokenization algorithm and taking into account frequent keyphrases only. The best results reported by KEA team show about 18% of F-Measure [102] in the extraction of keyphrases from generic web pages. Usage of domain specific vocabularies may improve the result up to 28.3% for Recall and 26.1% for Precision [63].

Another approach is suggested by Tourney and uses GenEx algorithm [95]. GenEx is based on a combination of parameterized heuristic rules and genetic algorithms. The approach provides nearly the same precision and recall as KEA. In a more recent work [93], the author applies web-querying techniques to get additional information from the Web as background knowledge to improve the results. This method has a disadvantage: mining the Web for information and parsing the responses is a time and resource consuming operation. This is inconvenient for Digital Libraries with millions of documents. In this approach the author measures the results by the average number of correctly found phrases vs. total number of extracted phrases ratio.

Recent works by A. Hulth *et al.* took into account domain [38] and linguistic [39] knowledge to search relevant keyphrases. In particular, contribution [39] used thesauri trying to get domain knowledge. Recall reported in this work is very low, namely 4-6%. The approach proposed in [38] introduced a heuristic related to part-of-speech usage, and proposed training based on the three standard KEA features plus one linguistic feature. Authors reported relatively good results (F-Measure up to 33.9%). However, it is hard to compare their results with the others due to the strong specificity of the used data set: short abstract with on average 120 tokens where around 10% of all words in the proposed set were keyphrases.

A recent interesting work with regard to the application of linguistic knowledge to the specific problem is reported in [24]. The authors used WordNet[25] and "lexical chains" structures based on synonyms and antonyms. Then they applied decision trees as a ML part on about 50 journal articles as the training set and 25 documents as the testing set. They reported high precision, up to 45%, but did not mention recall. This makes difficult any comparison with other techniques, and, as it will be investigated below, we think that such dataset is too small and biased for comparison.

Other ML technique, Least Square SVM [99] shows 21.0% Precision and 23.7% Recall in the analysis of web-mined scientific papers. Also in this case the described experiments are limited to a very small testing dataset of 40 manually collected papers, which is again very small set.

Let us briefly consider some other methods all summarized in the recent paper [58], they are:

1. Word clustering: here the basic idea is to cluster words using unsupervised clustering method, for example [86]. Than, after the phrases are clustered, we induce an additional weight to a phrase's TF, making it proportional to summarized TF's of all phrases in a cluster.

2. Using the sentence salience score [72] as a feature. This score is a weight of sentence in which a phase is located. It based on the notion of vector space, where we represent each sentence in a document or cluster of documents as a vector, and than compute the inner product between each pair of vectors. Having linear combination of scholar vector product, position of a sentence in a document and centroid score (the measure of centrality of a sentence in a single document or cluster of documents) we denote a notion of salience score as it is defined in [72]. Salience score is also used for keyphrases mining in [57].

3. Graph-based score is also may give an additional weighting, here we exploit the idea that important phrases are connected into a graph, such graphs may be constructed with help of iterative reinforcement algorithm [98]. Same statement is valid for sentences which are also bound in a graph. Reinforcement algorithm is the iterative method that recomputes the weight of a node (single word or sentence) in a graph until convergence. So it may be employed as an additional weighting.

Sticking all mentioned methods together we may assign each single phrase with a complex score, than, getting threshold we may take n most "important" phrases and treat them as keyphrases. Evaluation of this method in the recent work [98] shows an improvement of F-Measure from 25% (TFxIDF baseline approach) to 29%. This improvement is quite similar (a bit less) with what we have obtained, but it is done on the smaller manually tagged by volunteers annotators set of meeting notes.

There is another way of machine learning: unsupervised learning or learning without a trainer. Such kind of training does not require any training set (while, for sometime it may use some "seeds" which are not necessary). For supervised keyphrases extraction problems there is just one publicly available system that may be used without any additional preparations: KEA [102]. For unsupervised learning (without necessity to have training set) there is also just one publicly available [4] *keywords* extraction system: Leximancer [35]. To the best of our knowledge there are no any other systems which may be used as a "blackbox", working by the following scheme: 1) input text, 2) push the button and 3) get keyphrases. Leximancer system [83, 100] designed and developed for completely unsupervised keyword extraction. We say keyword instead keyphrases since Leximancer is able to extract merely one-token phrases. This limitation is caused by the nature of a method. For information extraction Leximancer team uses so-called context based analysis, i.e. each word considered within context, the piece of text around a word. This methodology came from earlier works of David Yarowsky [105, 106] dedicated to the problem of statistical word-sense disambiguation. We believe that context-based analysis with combination of proposed int the present Thesis methodology may improve keyphrases extraction and make a first step towards keyphrases assignment, where keyphrase is not presented in text, but assigned to it due to analysis of set of "similar" texts.

This chapter extends and improves on a preliminary work describing our initial concepts and presenting initial results [45] obtained using standard SVM approaches, namely:

1. We extend the use of state-of-the-art NLP tools [65] for the extraction,

[4] It is free of charge for research purposes, see licence at `https://www.leximancer.com/`

definition and use of linguistic-based features such as part of speech and syntactic relations extracted by dependency parsers [67].

2. We apply the proposed NLP-based approach to a number of different ML methods namely traditional SVM, innovative Local SVM and Random Forests.

3. We define and publish a large high quality dataset of 2000 documents [47], available through internet, with experts assigned keyphrases in Computer Science field.

4. We analyze in details the effect of different NLP features and dataset size on the overall quality of extracted keyphrases.

5. We perform a comparative analysis of the computed quality measures and obtained keyphrases with the various ML techniques (enhanced with NLP) and the popular Bayesian learning system KEA.

3.2 TCS Dataset Characterization and Linguistic Processing

3.2.1 TCS Dataset Characterization

The dataset presented [47] contains a set of papers published by the ACM in the Computer Science domain in 2003-2005 years. The documents are included in the ACM portal[5] and their full texts were crawled by CiteSeerX digital library as PDFs, but we place them to internet as the text files. These text files names are unique ACM ids, so dataset may be easily verified online through ACM portal. In our pre-processing tasks, we separated different parts of papers, such as title and abstract, thus enabling

[5] `http://portal.acm.org`; available also at `http://dit.unitn.it/~krapivin/`

extraction based on a part of an article text. Formulae, tables, figures and eventual LaTeX mark up were removed automatically. We share this dataset and welcome interested communities to use it as a benchmarks set for information extraction approaches.

For our investigations we separate existing keyphrases into two categories:

- author assigned: located inside each document in the header sections after the prefix "Keywords:", we removed them;
- editor assigned: manually assigned by human experts in a particular domain.

Our experimental dataset consists of 2000 documents with keyphrases assigned by ACM editors. It is important to note that in our preparation of the above dataset, we have selected only papers that contain at least one expert assigned keyphrase in the full text of a document. So we are not in the more challenging case of *completely* implicit extraction. In our dataset, each document has on average about 3 unique human assigned keyphrases (see Figure 3.1).

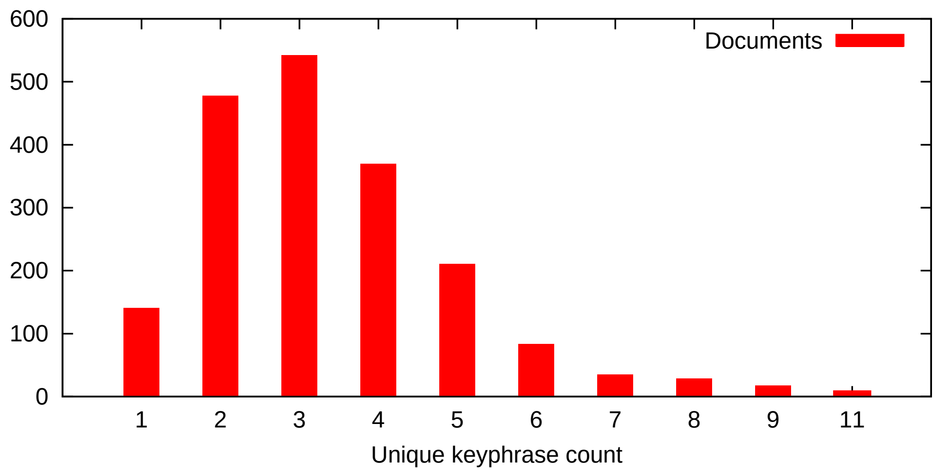

Figure 3.1: Distributions of unique assigned keyphrases per document.

3.2.2 Linguistic Analysis of Keyphrases

We performed a NLP analysis of the keyphrases to study their linguistic properties. This is the base for the proposal of heuristics and the definition of the features used in the machine learning step. This improves the quality of generated keyphrase candidates while simultaneously reducing their quantity.

We analyzed a sample of 100 random documents using OpenNLP tools. We applied tokenizer, Part of Speech (POS) tagger and chunker to explore differences between POS tags and chunk types for normal text documents and the corresponding keyphrases set. Fig. 3.2 shows POS tag distributions for the most common POS tags, such as nouns: NN, NNP, NNS; prepositions: IN, adjectives: JJ and verbs: VBN, VBP, VBG, VBD. Fig. 3.3 shows distributions for chunk types, such as noun phrases: B-NP, I-NP; prepositional phrases: B-PP; verbal phrases: B-VP, I-VP. To improve readability we have omitted values close to zero.

One can note from the figures that the distributions differ significantly between normal text and keyphrases sets. The major differences in POS tags distribution confirm that the majority of keyphrases consist of nouns, singular as well as plural, and adjectives. The difference in chunk types distribution also confirms and reinforces this hypothesis, adding to it that the overwhelming majority of keyphrases are noun phrases.

We did another analysis using MaltParser[67] to explore differences between dependencies of keyphrases and the ones of normal text. Fig. 3.4 compares keyphrases and normal text dependency distributions. We use the results of this analysis to compose features set.

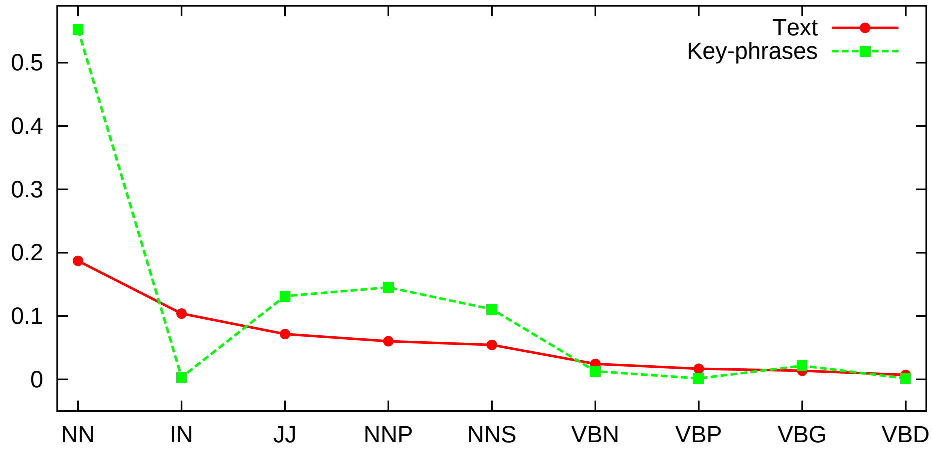

Figure 3.2: POS tags distributions for normal text and keyphrases.

3.2.3 Text processing

Before any extraction task, text needs to be pre-processed to assure a reasonable quality of extraction [103]. Modern scientific papers are mostly available in PDF format, thus first we need to convert them to plain text. Further preprocessing includes sentence boundary detection, tokenization, POS tagging, chunking, parsing, stemming and recognizing separate blocks inside the article, such as Title, Abstract, Section Headers, Reference Section, Body.

We used OpenNLP suite [65] to do standard steps of text processing. Namely, we apply sentence boundary detector, tokenizer, part of speech tagger and chunker consequently. Then we apply a heuristic, inspired by the previous linguistic analysis of keyphrases.

The heuristic consists of two steps. First we filter by chunk type, leaving only NP chunks for further processing. Then we filter the remaining chunks by POS. We leave only chunks with tokens belonging to the parts of speech from the top of the distribution in Fig. 3.2, such as NN, NNP, JJ, NNS, VBG and VBN. Table 3.1 shows an example sentence and ex-

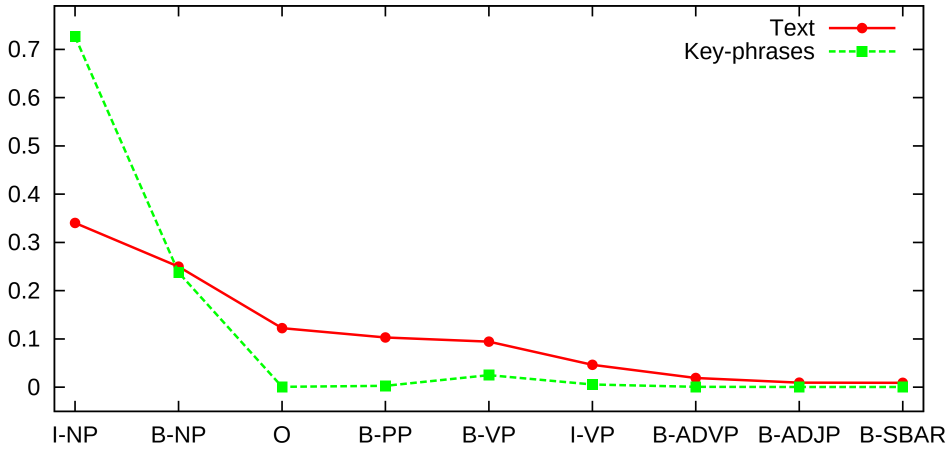

Figure 3.3: Chunk types distributions for normal text and keyphrases.

Table 3.1: Keyphrase candidates extracted by the heuristic.

Sentence	Therefore, the seat reservation problem is an on-line problem, and a competitive analysis is appropriate.
Candidates	seat, seat reservation, seat reservation problem, reservation, reservation problem, problem, on-line problem, analysis, competitive analysis

tracted keyphrase candidates. This heuristic extracts for further analysis only linguistically meaningful keyphrase candidates.

We use S-removal stemmer, embedded into KEA[102], to avoid problems with the same words written in different forms. This is the stemmer specially adopted by WEKA[103] group for the problem of keyphrases extraction, it is different from the other WEKA-embedded stemmers like snowball stemmer[92], Porter's[71] or Lovins' stemmer[59]. One of the core differences is that it does not remove "ing" form since such form changes the sense of a word significantly in rare cases. It called "s-removal" just because it distinguishes plural and single or in other words removes "s" in

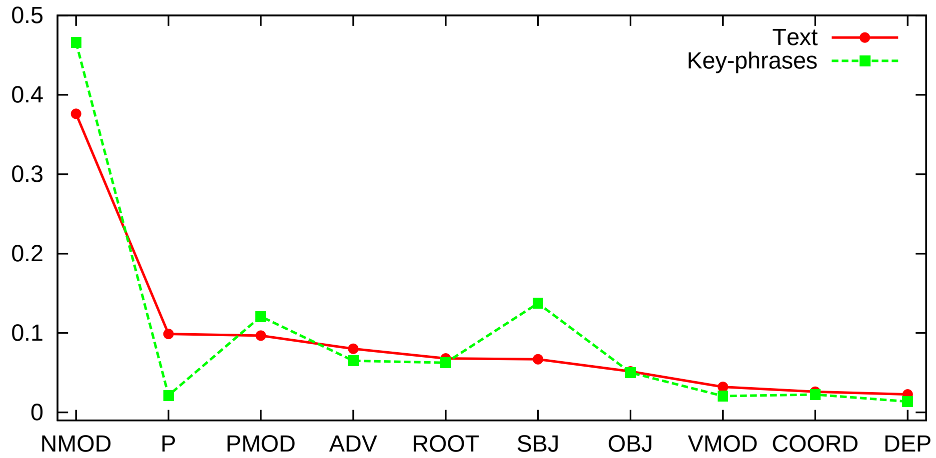

Figure 3.4: Dependencies distribution for normal text and keyphrases.

Table 3.2: Original and stemmed forms

Original	Stemmed
Multiple selections	multipl selection
text editing	text editing
SPFD-based global rewiring	spfd base global rewiring
Glauber dynamics	glauber dynamic
networking	networking
network	network

the end of a word. Table 3.2 shows the examples of original and stemmed forms.

In addition, we apply MaltParser to extract dependencies which we use as additional features for machine learning.

3.3 Enhancing Machine Learning with Natural Language Processing

3.3.1 Features selection

Proper feature space selection is a crucial step in information extraction. Many features may be used for accurate information extraction and their characteristics are strongly domain dependent.

The feature set we propose is detailed in Table 3.3. Features 1, 2 and 3 are common and widely used in most information extraction systems [93]. Less traditional features are: feature 4 quantity of tokens in a phrase, used in [88], feature 5 – the part of a text, successfully used in [99]. The features numbered in Table 3.3 from 6 to 20 are based on linguistic knowledge. We consider keyphrase containing a maximum of 3 tokens, with indices 1, 2 and 3 in the feature names referring to the first, second and third token of the candidate, respectively. Features 6 to 8 contain part of speech tags of the tokens of a keyphrase candidate.

The next set of features uses dependencies given by the MaltParser. Each dependency contains a head, a dependent and a labeled arc joining them. Dependencies help us to capture the relations between tokens, the position and the role of the keyphrase in the sentence. Specially, features 9-11 contain part of speech tag of a head of the token for each token of the candidate. Features 12-20 refer to the relations within the keyphrase and relations attaching a keyphrase to the sentence. They consist of three groups, one group for each token of the keyphrase candidate, and have similar meaning. Let us consider in detail the first group, features 12-14. Feature 12 refers to the label of the arc from the first token of the candidate to its head. It grasps the relation between the keyphrase and the sentence

Table 3.3: The adopted Feature Set, $i \in [1..3]$

#	Feature	#	Feature
1	term frequency	6-8	i-th token POS tag
2	inverse document frequency	9-11	i-th token head POS tag
3	position in text	12,15,18	i-th token dependency label
4	quantity of tokens	13,16,19	distance for i-th incoming arc
5	part of text	14,17,20	distance for i-th outgoing arc

or between the tokens of keyphrase. Features 13 and 14 grasp the cohesion of keyphrase and its relative position in the sentence. Feature 13 refers to the distance between the first keyphrase token and its dependent if it exists. As a distance we take the difference between token indexes. Feature 14 refers to the distance between the first token and its head.

3.3.2 Machine Learning Methods used for comparison

Random Forest.

Nowadays ensemble learning is getting more popular, one may be divided into two main branches:

- bagging [7] and
- boosting [75].

The core idea of ensemble learning is in the construction of many decision trees (or other) classifiers and voting for the best result. Bagging and boosting main difference in a way of constructing decision trees. Random Forest is an extension of earlier "bagging" technique proposed by Leo Breiman in 2001 [8, 14]. The RF algorithm randomly takes piece of a training set to grow each tree, and does not need costly cross-validation procedure. Being scalable and relatively fast RF improves state-of-the-art in a different machine learning-based classification tasks.

Support Vector Machines (SVMs)

[19] are classifiers with sound foundations in statistical learning theory [68] which are now considered the state-of-the-art classification method for a wide range of computational tasks. The reasons of their success are related to their ability to find the optimal solution, the possibility of highly non-linear mappings of the input space, the handling of noisy data with a soft-margin approach and their robustness to the curse of dimensionality. Differently from many text classification approaches based on the "bag of words" representation (where each text is encoded in a binary vector denoting which words are present) that causes a very high dimensionality of the data, we are working here with only 20 features. When the dimensionality is high a linear classifier is frequently the best choice, while with a reduced number of features a non-linear approach is needed. For this reason we adopt SVM with the Gaussian radial basis function (RBF) kernel in the form:

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{\sigma}\right) \tag{3.1}$$

where σ is a non-negative constant, called kernel width, which regulates the level of locality of the kernel and needs to be tuned tuning model selection.

FaLK-SVM

[80] is a kernel method based on Local SVM [6] which is scalable for large datasets. There are theoretical and empirical arguments supporting the fact that local learning with kernel machines can be more accurate than SVM [81]. In FaLK-SVM[6] a set of local SVMs is trained on redundant neighborhoods in the training set selecting at testing time the most ap-

[6]FaLKM-lib [79] source is available at `http://disi.unitn.it/~segata/FaLKM-lib`

propriate model for each query point. The global separation function is sub-divided in solutions of local optimization problems that can be handled very efficiently. This way, all points in the local neighborhoods can be considered without any computational limitation on the total number of SVs which is the major problem for the application of SVM on large and very large datasets.

KEA

[102] represents the state-of-the-art for keyphrase extraction tasks and it is based on the bag-of-words concept. The Bayes theorem is used to compute a probability of a phrase to be a keyphrase using frequencies gathered from a text. So in the end each phrase in the text has a probability to be a keyphrase. After that KEA takes top q of phrases and calls them keyphrases. Naïve Bayes learning is widely used for other text-oriented tasks like spam filtering or text classification.

There are some more accurate machine learning used for classification, for instance so called "Bayesian learning", most promising techniques are RVM, or Relevance Vector Machine [91] or Gaussian Processes [101] that may potentially improve the accuracy of prediction, but they are too slow comparing even with SVM.

3.3.3 Training with unbalanced class cardinalities

In keywords and keyphrases extraction tasks it is natural that the number of key elements (and of candidate key elements) is dramatically lower than the number of non-key elements. In our dataset the keyphrases represent about the 1.8% of the total number of phrases taken into account, meaning that we have more than 55 times fewer positive examples than negative ones. Classification with unbalanced datasets is a challenging task which is very often handled assigning different misclassification costs to the classes.

In SVM classification this is possible associating different soft-margin regularization parameters (C) to the classes as discussed for the first time in [68]. Obviously the assignment of different regularization parameters to each class enlarges the set of parameters that needs to be tuned during model selection phase.

The same approach can be used for FaLK-SVM with the only difference being that the soft-margin regularization parameters (C) are set locally.

Random Forest also can handle unbalanced data using "weight" parameter in implementation [14].

3.3.4 Result assessment methodology

Usual IR performance measures are Precision, Recall and F-Measure. Defining with A the set of true keyphrases that have not been recognized as keyphrases, with B the correctly recognized keyphrases and with C the set of phrases incorrectly recognized as keyphrases, the formal definition of precision (P), recall (R) and F-measure (FM) are:

$$P = 100\% \cdot \frac{B}{B + C} \tag{3.2}$$

$$R = 100\% \cdot \frac{B}{B + A} \tag{3.3}$$

$$FM = \frac{2P \cdot R}{P + R} \tag{3.4}$$

Here it is important to underline, that we consider each phrase as occurrence individually. This means that one phrase may be included into the paper text several times as a keyphrase, and several times as not keyphrase. It is difficult to judge whether a phrase is a keyphrase or not in such approach. The heuristic behind our judging about keyphrase is as follows: if the phrase has been recognized as a keyphrase at least once, we treat one as a keyphrase. We will call P, R and F-measure based on it as *document-based* measures in the future which is also the methodology adopted by KEA [102]. Another approach is to take into account each phrase occurrence separately, obtaining *occurrence-based* precision, recall and F-Measure. This family of measures are not suitable for the final evaluation of a keyphrases extraction method since one phrase can be considered as a keyphrase several times. Even worse, it can be considered few times as a keyphrase, and few times as a non keyphrase. However, occurrence-based F-Measure is effective as measure to be maximized during SVM model selection because it follows the formal representation of numerical SVM data.

3.4 Experimental evaluation

In this section we give the details of the experiments carried out for the analysis of the discussed keyphrase approaches. To assess the results we used standard IR performance measures: Precision, Recall and F-Measure [68, 88].

3.4.1 Dataset splitting

We divided the whole dataset of 2000 documents into 3 sets: training set (TR), validation set (VS) and testing set (TS) respectively with 1400, 200 and 400 documents each. To investigate the optimal dataset size we further divided the training set into 7 subsets of 200 documents each. All sets are selected randomly assuring however the balancing with respect to the year of publication, namely assuring that all the sets have the proportional quantity of papers published in a given year.

3.4.2 Experiment 1. Comparison of ML Methods Enhanced by NLP

Random Forest

: four parameters to tune are the number of trees in the ensemble I, the splitting parameter K, the balancing parameter w and the depth of a tree d. Experimentally we discovered the following tricks to reduce training tries:

- take 3 different K parameters: default, half and double of default;
- stop the algorithm as soon as an increase in the number of trees does not improve significantly the solution;
- the depth of tree usually should not overcome the quantity of selected features, and should not be much smaller than them.

We found experimentally the best tuning ranges. We used the fast open source implementation[7] compatible with WEKA [103].

SVM

: the hyper-parameters we tune are the regularization parameters of the positive and negative classes (C^+ and C^- respectively) and the width σ of the RBF kernel. These parameters are selected using 10 fold cross-validation with a three-dimensional grid search in the space of the parameters. The model selection is performed maximizing in this parameter space the occurrence-bases F-Measure. For SVM training and prediction we use LibSVM [13].

[7] http://code.google.com/p/fast-random-forest/

Table 3.4: The results for SVM, FaLK-SVM, RF and KEA. Best values in bold

	Precision	Recall	F-Measure
FaLK-SVM	24.59%	35.88%	29.18%
SVM	22.78%	**38.28**%	28.64%
Random Forest	**26.40**%	34.15%	**29.78**%
KEA (best q)	18.61%	26.96%	22.02%

Table 3.5: KEA results for different threshold q values. The best precision, recall and F-Measure among all q values are in bold.

q	P, %	R, %	F-Measure, %
6	17.31	**30.10**	21.98
5 (default)	18.61	26.96	**22.02**
3	21.47	18.41	19.98
2	**24.87**	14.41	18.25

FaLK-SVM

: in addition to the SVM parameters (C^-, C^+ and σ) we have to set the neighborhood size k used for the local learning approach. Model selection is thus performed as described for SVM but using a four-dimensional grid-search.

KEA

: KEA has one tuning parameter q which is threshold, experimentally we have found that $q = 5$ produces the best F-Measure (see Table 3.4.2).

Table 3.4 summarizes the results. We see that the best result in F-Measure is achieved with the Random Forest using all 20 proposed NLP features. FaLKM and SVM follow very closely while KEA is much lower. The difference between three best methods is not very big (RF outperforms SVM by about 4%), therefore it is important to understand what are the most important factors: particular features or peculiarities of the dataset.

These has led us to investigation of both mentioned features in the next set of experiments.

3.4.3 Experiment 2. Training set size analysis

An increase in training set size may bring an improvement of prediction quality. However, training on a large amount of data is computationally expensive. Thus it is relevant to estimate which dataset size is enough to obtain the best prediction performance. To study this, we carried out experiments at increasing training set sizes as summarized in Figure 3.5. One can see that i) F-Measure improves as the training set size increases; ii) the improvement levels off after ca. 400 documents. We can conclude that for the task of keyphrases extraction it is important to have rather large training sets, but training sets with more than 400 documents are very likely to experience computational difficulties without a relevant increase in prediction ability.

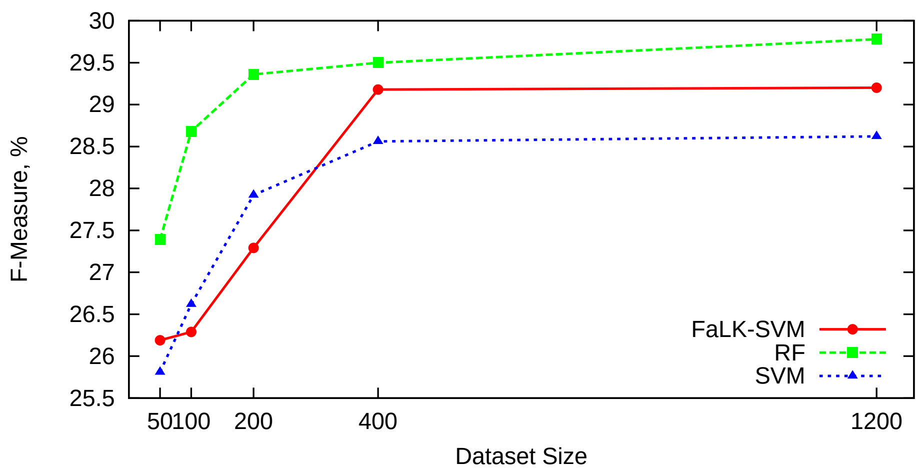

Figure 3.5: F-Measure behavior with dataset size growth.

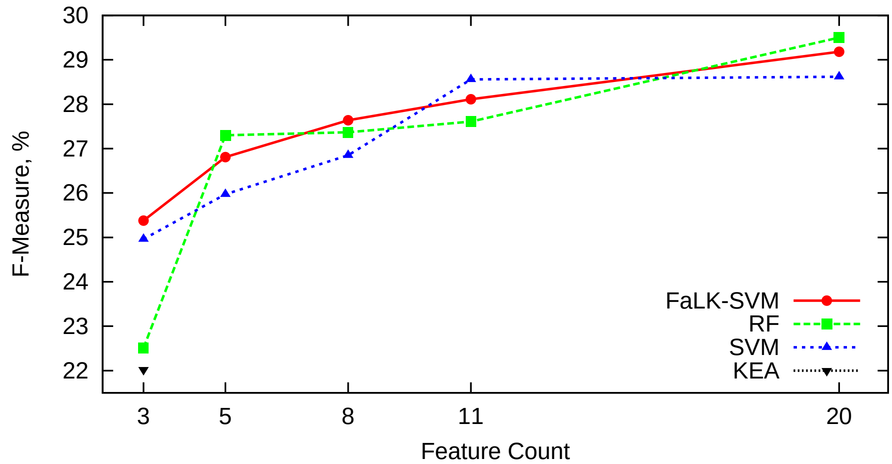

Figure 3.6: F-Measure behavior with feature count growth.

3.4.4 Experiment 3. NLP features analysis

In this experiments we analyze the individual effect of the features on the prediction ability. We performed experiments omitting features one by one and monitoring the effect on the overall quality. Assuming that our features are logically grouped we decided to exclude the following groups of features sequentially:

- Arcs (11 features left)
- Head POS Tags (8 features left)
- POS tags (5 features left)
- TFxIDF and relative position (3 features left).

Figure 3.6 summarizes the results. We see that in case of Random Forest using only first three features decreases F-Measure essentially to KEA results. This is very interesting, because Bayesian learning of KEA is only possible considering just 3 features (an increase of features quantity will break KEA). In our comparison of four methods we have two "statistical learning" methods (SVM and FaLK-SVM), and two "probabilistic" methods which give close results having the same three basic features that regard simple count of tokens. Figure 3.6 shows that the various methods capture different features in different ways: arcs are important for Random Forest and POS tags a most important for SVM. Moreover, while there is a tendency for the overall quality to level off, the experiments do not indicate clearly to have reached a "plateau" behavior (other relevant features may be found).

3.5 Comparative analysis of extracted keyphrases

Aside from the detailed quantitative analysis of the standard quality measures used to compare the different approaches performance, some important insights on the specific characteristics of each approach, can be gained by a direct analysis of the extracted keyphrases. To this end we have focused our attention on the best results obtained respectively by KEA and by our proposed approach (RF+NLP).

In Table 3.6 we have collected statistics from our experiments related to correctly and incorrectly recognized keyphrases. Figure 3.7 and Figure 3.8 show the distribution of the numbers of correctly and incorrectly extracted keyphrases per document by all four approaches, respectively.

From these data, apart from the generic improvements in the recognition performance of the proposed SVM+NLP approach already presented in Table 3.4 and discussed in the previous section, we can identify some other interesting characteristics and differences in the two approaches, namely:

- There is an overlap between the extracted true keyphrases in the two approaches: approximately 57% of RF+NLP correct keyphrases are also extracted by the KEA system. However, there exists a signifi-

Table 3.6: Comparison between Found/Not found keyphrases counts for the best results.

Keyphrases type	Keyphrases count
KEA Correct	360
RF Correct	463
KEA Incorrect	1575
RF Incorrect	1280
ACM Total	1332
Correct keyphrases overlapped for KEA *and* RF	264
Correct keyphrases uniquely for KEA *or* RF	559

cant number of distinct keyphrases extracted only by KEA and only by RF+NLP. This can also be seen in last 2 rows of the Table 3.6. Let us call varieties of correctly extracted by KEA keyphrases as (KEA-c) and by RF+NLP approach as (RF+NLP-c). The intersection of (KEA-c) and (RF+NLP-c) (see next-to-last row) contains 264 keyphrases, while the union of (KEA-c) and (RF+NLP-c) (the last row) is nearly two times bigger. That allows us hypothesize that a combination of the two approaches may improve keyphrases extraction performance.

- As already noticed in Table 3.4 RF+NLP recognizes less incorrect keyphrases than KEA or other methods. In addition we can notice from Figure 3.8 that ML+NLP approaches have a lot more documents with few or zero incorrectly recognized keyphrases, while KEA often makes 4 or 5 errors;

- Both approaches provide a similar "coverage" of correctly extracted keyphrases per document. Here "coverage" means the property of a given extraction approach to identify "at least" one correct keyphrase per document. In fact the total number of documents with correctly extracted keyphrases increases from 66% in KEA to 73% in RF+NLP.

From these observations we can conclude that KEA looses to ML+NLP approaches in precision and in recall both by 44% and 30% respectively. From the recall viewpoint KEA is more competitive with ML+NLP due to the higher number of extracted keyphrases it proposes, thus causing a much lower precision. Thus, even if there is a rather relevant number of correct keyphrases extracted only by KEA than can potentially enhance the ML+NLP recall, combining the two approaches might not be a good idea, because incorrectly recognized keyphrases must be merged too and thus the final precision (and the F-Measure) sensibly decreases.

Table 3.7: Examples of *top* results for RF+NLP approach

#	**ACM Keyphrases stems**	**RF+NLP Keyphrases stems**
1	data clustering, constructive induction, bayesian network, em algorithm	*data clustering, constructive induction, bayesian network, em algorithm*, bayesian multinet
2	creg, register allocation, graph coloring, register	*creg, register allocation, graph coloring*
3	software prefetching, software pipelining, vliw machine, locality analysi	*software prefetching, software pipelining, vliw machine,* modulo scheduling
4	two-variable fragment, controlled language, natural language, logic	*two-variable fragment, controlled language, natural language*
5	order-sorted logic, knowledge representation, terminological knowledge, resolution system	*order-sorted logic, knowledge representation, terminological knowledge*, label-based formula, hierarchical representation

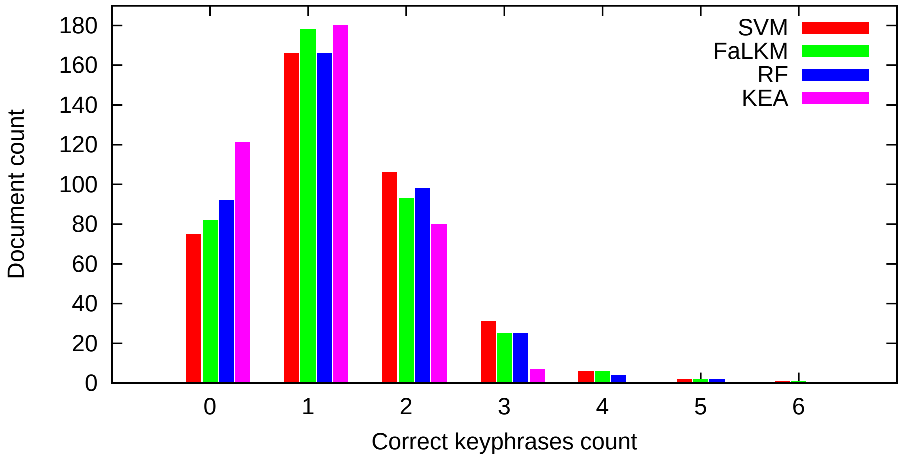

Figure 3.7: Comparison of KEA and ML+NLP distributions of correctly extracted keyphrases per document.

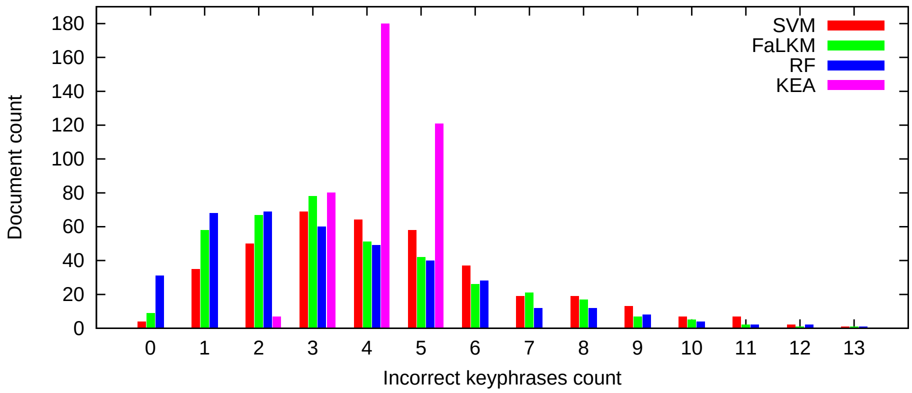

Figure 3.8: Comparison of KEA and ML+NLP distributions of incorrectly extracted keyphrases per document.

As a last analysis, we have explored qualitatively the keyphrases extracted by RF+NLP method. First, we have looked at the best results: as examples of the type of correct matches between original ACM keyphrases stems and correctly extracted keyphrases stems, we collected in Table 3.7 the top 5 results[8]: 2 documents have 100% match, and 3 documents with almost all keyphrases recognized. We see that unrecognized keyphrases

[8]Keyphrases in the table are stemmed with S-Removal stemmer

Table 3.8: Examples of *bad* results for RF+NLP approach

#	**ACM Keyphrases stems**	**RF+NLP Keyphrases stems**
1	distributed environment, persistent object, distributed system, modularity, object-oriented approach, distributed programming system, replication, migration	*distributed environment, persistent object, distributed system*, database system
2	flexible transaction, concurrency control, transaction management, serializability	*flexible transaction, concurrency control*, heterogeneou distributed database, distributed database, distributed database environment, database environment, database system, multidatabase system, multidatabase
3	knowledge-based query processing	knowledge-based approach
4	security, partial equivalence relation, semantic, powerdomain, noninterference	secure information, secure information flow, information flow, sequential program, security properti
5	reflection, program transformation	generic reification technique, reflective language

are synonyms or related to the recognized ones, for instance instead of "bayesian network" we have "bayesian multinet", or "hierarchical representation" instead of "knowledge representation".

More interesting is to look at the *bad* results of our approach, that is at the results at the bottom of the distribution where no correct keyphrases (the ones included in the ACM list) has been extracted. Table 3.8 provides 5 examples of such results. In this case we cannot say "the bottom 5" because we have a tail of equally *bad* results, so we have randomly selected 5 examples among all documents with no correctly extracted keyphrases. For completeness we report that we had 92 such documents out of 400 total. A preliminary qualitative analysis of the data in the table confirms the lack of correct keyphrases. However, the extracted keyphrases seem related to the human assigned keyphrases that we use for assessing the quality of our approach. Specifically, it seems that the human-assigned keyphrases may be more generic and tend to have a higher level of abstraction while the keyphrases extracted from the actual text via RF+NLP tend to be more specific. For instance: "security" → "secure information", "reflection" → "reflective language".

This behavior, if confirmed by further and more comprehensive data, would enhance the trust in the use of automatic keyphrases extraction. Automatically extracted keyphrases cannot really be used *instead* of assigned ones, but they potentially could enhance them. We know that it is not possible to draw out general trends from such a limited number of instances. However, exploration and analysis of the intrinsic qualities of automatically extracted keyphrases seems an interesting direction for future work.

3.6 Discussion

In this Chapter we presented the application of Natural Language Processing-based knowledge to a number of different Machine Learning methods: traditional SVM, a local variant of SVM and Random Forest for automatic keyphrases extraction from scientific documents. The proposed NLP-based approach shows promising results. We have performed a detailed evaluation of the performance of all ML methods by comparing the extracted keyphrases with human assigned keyphrases on a subset of 2000 ACM papers in the Computer Science domain.

Evaluation shows that adding the syntactic knowledge to the problem of keyphrases extraction improves the quality of extraction. Talking about machine learning part we can conclude that the best tradeoff between extraction quality and computation speed is Random Forrest. Baseline method KEA is extremely fast and simple and despite it uses just two features i) TFxIDF and ii) the relational position of the first occurrence of a phrase in a document, KEA shows relatively good performance, namely 22% of F-Measure. For comparison, the best result of Random Forrest is F-Measure 30%. We see 36% improvement comparing with KEA, which seems to be interesting impact without a usage of controlled vocabularies.

The proposed hybrid ML+NLP approach may also be valid with different data like news, emails, abstracts or web pages. One limitation of the present work (and of all the works based on the instance learning) is in the assumption of the presence of the searched keyphrases inside the documents (assumption that has been used in the construction of our dataset). Indeed, our learning method cannot find (without additional supporting knowledge) a specific keyphrase in a document when the document does not contain at least one instance of the keyphrase. To tackle such a challenging keyphrase *assignment* task one needs to take into account documents or keyphrases similarities. For example, one may forecast that documents with similar topics may have similar keyphrases. Alternatively, we have to move from syntactic to semantic relations between words in order to access (implicitly) related keyphrases. Possible future work may also focus on deeper analysis of a quality of extracted keyphrases, because "incorrect" ones are not necessarily the "bad" ones. There is a chance they may be better than author or editor assigned keyphrases.

Chapter 4

Ranking of Objects in ADL

In this Chapter we exploit the idea of ranking items in the graph-based citation networks. To perform ranking we propose few modifications of Google PageRank. If Autonomous Digital Library documents refer to each other, like web pages do, they are connected into a graph. This is true for scientific publications, where a paper usually has references. We apply PageRank and PageRank-based ranking schemas to the set of scientific papers which conform the graph of 266,000 nodes. We describe here also innovative visualization techniques to plot the difference between various ranks, and, in the end, we adopt graph-based ranking techniques to evaluate an impact of an individual researchers.

4.1 State-of-the-art

4.1.1 Scientometrics: state-of-the-art ranking metircs

After the Second World War, with the increase in funding of Science and Technology (S&T) initiatives (especially by public institutions), the need for supervising and measuring the productivity of research projects, institutions, and researcher themselves became apparent [27, 28]. Scientometrics

[96] was then born as a science for measuring and analysing quantitatively science itself [21]. Nowadays, the quantitative study of S&T is a rapidly developing field, also thanks to a greater availability of information about publications in a manner that is easy to process (query, analyze). The easiest measure to show any individual scientist's output is the total number of publications. However, this index does not express the quality or impact of the work, as the high number of conferences and journals make it easy to publish even low quality papers. To take quality and impact into account, the citations that a paper receives emerged, in various forms, as a leading indicator. The citation concept for academic journals was proposed in the fifties by Eugene Garfield, but received the deserved attention in 1963 with the birth of the Science Citation Index (SCI) [27]. SCI was published by the Institute for Scientific Information (ISI) founded by Garfield himself in 1960 and currently known as Thomson Scientific that provides the Web of Science on-line commercial database. The most studied and commonly used indexes (related to SCI) are, among others [64]:

1. P-index: or just number of articles of author.

2. CC-index: number of citations excluding self-citations.

3. CPP: or average number of citations per article.

4. Top 10% index: the number of papers of a person that are in the top 10% most frequently cited papers in the domain during the past 4 years.

5. Self-citation percentage.

6. Career length in years.

7. Productivity: quantity of papers per time-unit.

Although most of the indexes are related mainly to authors, they can also be applied to measuring communities, institutions or journal, using various forms of aggregation. In the last decade new indexes have been proposed. These indexes are rapidly gaining popularity over the more traditional citation metrics described above:

1. H-index, proposed by Hirsch in [36]. The H-index for an author is the maximum number h such that the author has at least h articles with h citations each. This index is widely used (including in our University), and comes in different flavors (e.g., normalized based on average number of authors of papers, on the average citations in a community, *etc*).

2. The G-index for an author is the maximum number g such that the most cited g papers of an author collectively received g^2 citations. The g index takes into account papers with very high citations, which is something that is smoothed out by the h-index.

In addition, we mention below some algorithm for ranking Web pages. They are relevant as many of them have been very successful for ranking web content, and papers share some similarities with Web sites, as they can be seen as a sort of hypertext structure is papers are seen as web pages and citations are seen as links.

1. Hypertext-Induced Topic Selection (HITS) [44]: based on graph linkage investigation, it operates with two notions: "authority" and "hub", where authority represents relevance of the page (graph node) to query and hub estimates the value of the node's links to other pages.

2. PageRank (described in more detailed in the following): a well-known and successful ranking algorithm for Web pages [9], based on net random walking probabilistic model. When modified for ranking scientific papers, it has been shown to give interesting results [15].

3. Hilltop [3]. This algorithm is based on the detection of "expert pages", i.e., pages that have many outgoing links (citations) and are relevant to a topic. Pages that are linked by expert ones have better rank.

In our work we adopt a variation of PageRank as one of the main indexes used for the analysis of differences among indexes. The intuition behind PageRank is that a web page is important if several other important web pages point to it. Correspondingly, PageRank is based on a mutual reinforcement between pages: the importance of a certain page influences and is being influenced by the importance of some other pages. From a computational point of view, PageRank is a statistical algorithm: it uses a relatively simple model of "Random Surfer" [9] to determine the probability to visit a particular web page. Since random browsing through a graph is a stochastic Markov process, the model is fully described by Markov chain stochastic matrix. The most intriguing question about PageRank is how to compute one for a dataset as huge as the web. The inventors of PageRank, Brin and Page, proposed a quite effective polynomial convergence method [9] (see please more in Sec. 4.2.3). Since then, a significant amount of research has been done in the exploration of the meaning of PageRank and proposals for different computation procedures [5, 60, 15]. When the attention is shifted from web pages to scientific citations, the properties of the citation graph - mainly its sparseness - has been used to simplify the computational problem [87]. In our work, we have based our computations on a variation of Page Rank (called Paper Rank) for ranking scholarly documents explained in detail in Section 4.2. From a computational perspective, the difference is that the algorithm we propose exploits the fact that in citations, unlike in web links, "cycles" or cross-citations (when paper A cites paper B and visa versa) are very rare, and should be considered as "unfair" citing [21]. Paper must be published with all it's references *before* being cited. In terms of comparison among scientific metrics for determining the difference in the ranking results they generate (and methods for evaluating such differences), there is quite small state-of-the-art to the best of our knowledge, we can mention here Kendall τ-distance [43], or the quantity of steps to arrange both ranks in the same order, usually it used for the top k ranked items. In the Statistical Theory there are some correlation measures [82], mostly suitable for measuring similarities between probabilistic processes-created values, (like throwing a dice results *etc*), but we think that they are not applicable to the problem of comparison of top k elements with respect to this elements *ranks.*

4.1.2 PageRank metric: definition, computation and evolution

Ten years ago Google[18] corporation with great success applied PageRank (PR) algorithm[9] to the problem of web-pages ranking. PR algorithm is purely statistical, and there is no need to analyze the content of each page lexically. It uses “Random Surfer” model, in which the process of browsing through the web pages links is modeled by the stochastic Markov process, fully described by Markov chain matrix. Recently Page Rank has been studied from several points of view including computational feasibility, modifications and adaptations to the different types of graphs and network models, probabilistic model, mathematical background[22]. Its popularity for ranking web-pages makes it popular in other domains, like ranking of scholarly publications. So let us return to the problem of PageRank computation it for the whole web. Complete internet contains terabytes of information, and being represented as a graph it exceeds modern computers memory. It is a creative engineering task to design fast access storage to compute PR. Let us briefly outline major methods for PR computation.

- The simplest one is the cyclic PR computation for all nodes in the graph one by one, using recursive formula 4.1 until convergence [15].

This method takes unit vector as initial rank approximation.

- PR authors, Sergey Brin and Larry Page proposed polynomial convergence method[9] with dumping factor usage, formula 4.3.
- This method was improved by Kamvar *et al.*, 1999 [42] using "block-based strategy", similar to implementations in relational database products.
- In 2004 Langville [55] invented the procedure with reduction of the iterations number with lucky initial approximation.
- In 2005 Haveliwala *et al.* [60], proposed quadratic extrapolation method to accelerate PR convergence and evaluated their methodology under roughly 81 millions of pages.

Most of mentioned above works are related to the WEB links ranking problem which usually deals with much larger graphs than scientific citing problem. So, the computation problem has been studied well enough and looks feasible.

4.2 Proposed Approach

In this section we consider dataset we used, different ranks applicable for scientific citing as a measure of research impact of a single paper or author.

4.2.1 Data set description and data preprocessing

The starting point for our analysis is a dataset of 266788 scientific papers published in ACM conferences or journals, and authored by 244782 different authors. The dataset was available as XML documents that for each paper describes information such as authors, title, year of publication, journal, classification and keywords (for some of the papers), journal volume and pages, and citations. A sample of the dataset format is available at the companion web page mentioned earlier. The set is biased in terms of citation information. For any given paper in the set, we have all its references (outgoing citations), but we only have citations to it (incoming citations) from other papers in the dataset, and hence from ACM papers. To remove the bias (to the possible extent), we disregard references to non-ACM papers. In other words, we assume that the world, for our citation analysis, only consists of ACM papers. Although we have no measurable evidence, given that we are comparing citation-based metrics we believe that the restriction to an "ACM world" does not change the qualitative results of the analysis. Including references to non-ACM papers would instead unfairly lower the measure for Paper Rank since, as we will show, Paper Rank is based on both incoming and outgoing citations. This being said, we also observe that the quality of the chosen dataset is very high. The majority of papers have been processed manually during the publishing process and all author's names have been disambiguated by humans. This is crucial since systems like Google Scholar or Citeseer contain errors in the disambiguation of authors names and citations. In fact, both Goodle Scholar or other autonomous digital libraries like Citeseer or Rexa use machine learning-based unsupervised techniques to disambiguate the information and are prone to introduce mistakes. A preliminary study of these errors in Google Scholar is presented in [77]. Besides disambiguation errors, crawled information may include spurious types of documents like deliverables, reports, white papers, etc. Indeed, Scholar includes in its statistics the citations coming from project deliverables or even curricula vitae, which are not commonly considered to be academically meaningful citations. Thus, although incomplete, the ACM dataset has a high level of quality in particular in respect to authors and citations. The full citation graph of the ACM dataset has 951961 citations, with an average of 3.6 outgoing citations per paper (references to other ACM papers). Figure 4.1 shows instead how many papers have a given (incoming) citation count (hereafter called CC). As expected, there is a very large number of papers with low, near-zero citations and a few papers with a high number of citations.

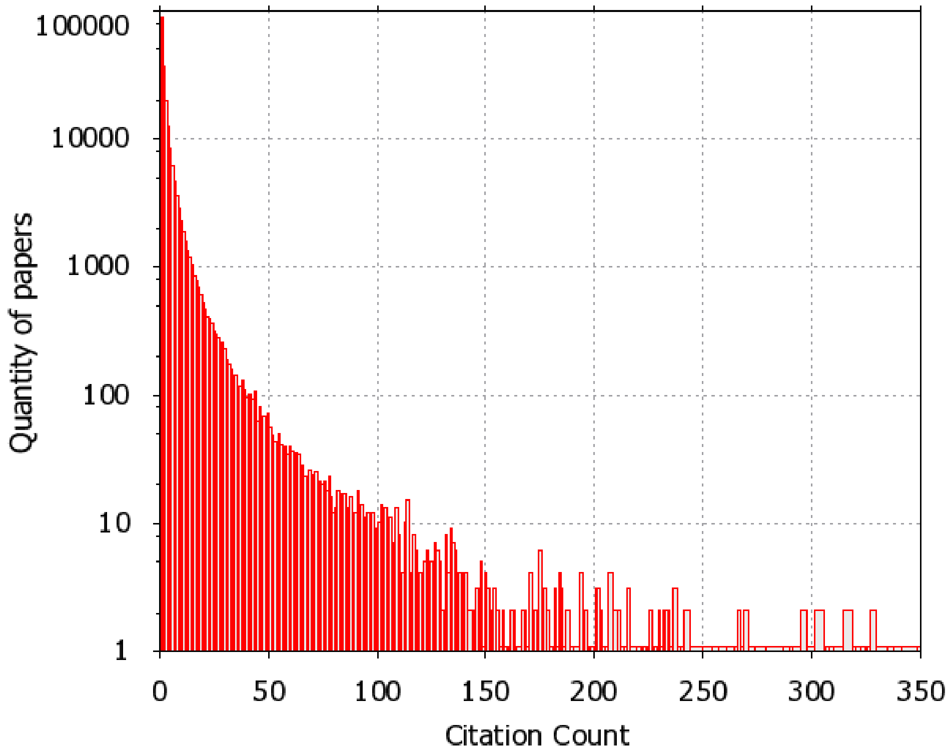

Figure 4.1: Distribution of papers by Citation Count.

The years of publication of the papers in the dataset vary from 1950 to 2005 with most emphasis on the recent two decades due to the increase in the number of publications.

4.2.2 Page Rank description

The original Page Rank algorithm [9] ranks the nodes of a directed graph with N vertices. Considering edges between nodes we may come to the following formalization: let us numerate all nodes in graph from 1 to N, so that each node has it's own unique sequence index. Each node has to be ranked by PageRank algorithm so that node number i has PageRank value P_i. Mapping this consideration to the problem of scientific citing we may conclude that a paper is a node, and citation is an edge of a graph. Rank of the node represents it's weight assuming that the more node is "referred" (cited in case of papers), the better should be the rank. The trick of a PageRank is that it considers not only the *quantity* of citations, but also the PageRank (*quality*) of all citing papers. The rank of a node is determined by the following recursive formula, where $S(j)$ is the quantity of outgoing links from a node number j. are just sequence numbers and D is the set of nodes such that there is a path in the graph from them to node i.

$$P_i = \sum_{i \neq j}^{j \in D} \frac{P_j}{S(j)} \tag{4.1}$$

The formula can be seen in matrix form and the computation can be rewritten as an eigenvector problem:

$$\vec{r} = A\vec{r} \tag{4.2}$$

where A is the transition matrix, or stochastic Markov matrix.

This consideration exposes several potential problems in rank computation as discussed in [5, 54]. One of them is the presence of the nodes which have just incoming links but have no outgoing ones, called dangling nodes. In this case, equation 4.2 may have no unique solution, or it may have no solution at all (it will lead to zero-columns occurrence in the transition matrix). Such problem may be resolved with the introduction of a damp-factor d. The dump (or decay) factor is a positive double number

$0 < d < 1$:

$$P_i = (1 - d) \sum_{i \neq j}^{j \in D} \frac{P_j}{S(j)} + \frac{d}{N} \tag{4.3}$$

The damp factor was proposed by the PageRank inventors, Page and Brin. In their publication [9], Page and Brin give a very simple intuitive justification for the PageRank algorithm: they introduce the notion of "random surfer". Since in the specific case of web pages graph, the equivalent stochastic Markov matrix can be described as browsing through the links, we may imagine a "surfer" who makes random paths through the links. When the surfer has a choice of where to go, it chooses randomly the next page to visit among the possible linked pages The damp factor models the fact that surfers at some point get bored of following links and randomly starts another surf session. The damp factor therefore also eliminates the probability of surfers ending up in dangling nodes. Being in dangling node random surfer "teleports" itself to a random node and starts a new session of links surfing. The damp factor helps to achieve two goals at once:

- any single paper in a graph makes it's own influence on ranks of all papers
- ability to find the solution of linear system of equations in case of dangling nodes presence.

4.2.3 PageRank computation, simple cases

The most intriguing question about PR is how to compute it for the whole web? Whole internet contains terabytes of information, and being represented as a graph it exceeds modern computers memory. It is a creative engineering task to design fast access datastructures and algorithm to compute PR. But since Google works it is obvious that such problem is solved. Let us briefly outline major methods for PR computation

- The simplest one is the cyclic PR computation for all nodes in the graph one by one, using recursive formula 4.1 until convergence [15] (the simple implementation is available through [89] for free). This method takes unit vector (or vector with just one not zero component equal to constant) as initial rank approximation.

- PR authors, Brin and Page proposed polynomial convergence method [9], similar to Jacobi methods (will be further examined in details in this section).

- Method 2 above was improved by Kumvar *et al.*, 2003 [42] using quadratic extrapolations and was applied to 83 millions of web pages.

- In 2004 Langille [55] invented produced a specialized iterative aggregation algorithm for updating any Markov chain with any type of update, link or state. They reached speed up 5-10 times comparing with the previous results.

So the problem of computation has been found and it is easily scalable to the WWW dimension. Since we have much smaller dataset than [42] had, we can apply the simplest methods to get more or less good result.

Let us consider one of the simplest methods of computation of PageRank, or "power convergence method" in more details. In particular we may set initial approximation as a vector with all components equal to 1. Than to compute PageRank we should follow the recursive formula:

$$I_{k+1} = A \cdot I_k \tag{4.4}$$

, where A is a Markov matrix and I is an eigenvector. Lawrence and Page [9] reported that $k = 50 \div 100$ iterations are enough to converge to eigenvector with appropriate accuracy and it took up to 6 days for all available WEB pages graph those time (more than 10 years ago). Formula

4.4 is applicable to millions of nodes in a graph since it's simplicity and matrix sparsity. Reader may find concrete samples in [10] or in the web [2].

4.2.4 Paper Rank: PageRank Metric in the Digital Library Context

PageRank has been very successful in ranking web pages, essentially considering the reputation of the web page referring to a given page, and the outgoing link density (pages P linked by pages L where L has few outgoing links are considered more important than pages P cited by pages L where L has many outgoing links). Paper Rank (PR) applies page rank to papers by considering papers as web pages and citations as links, and hence trying to consider not only citations when ranking papers, but also taking into account the rank of the citing paper and the density of outgoing citations from the citing paper. From a computation perspective, PR is different from Page Rank in that loops are very rare, almost inexistent. Situations with loop where a paper A cites a paper B and B cites A are possible when authors exchange their working versions and cite papers not yet published but accepted for publication. In our dataset, we have removed these few loops (around 200 loops in our set). That is why the Markov matrix may be brought to diagonal form like it is explained in the work of Lee Giles [87], so the system will have solution even without damp factor. We decided to omit the damp factor and take the 4.1 for PR computation. Furthermore, considering that a citation graph has $N >> 1$ nodes (papers), each paper may potentially have from 1 to $N-1$ inbound links and the same quantity of outgoing ones. However, in practice citation graphs are extremely sparse, (articles in our ACM dataset normally have from 5 to 20 references[1]) and this impact the speed of the computation of

[1]for computer science domain in average

PR. We implemented 3 different methods of PR computation:

1. the one like it is reported in [87] (the most primitive),
2. second one is taken from the work [15], where we use formula 4.1 sequently for all nodes until convergence,
3. and the last one was the method proposed by Page and Brin [9], where sparse matrix is multiplied with vector of PR values until convergence. We used the damp factor $\alpha = 0.15$ like it was proposed by PR inventors.

Of course, results for PR with and without damp factor are different, but they correlate very well for most cited papers with $CC > 10$. So the plots below are made for the “PaperRank”, where graph has no loops and PR is computed without damping factor. We want to emphasize once again that the power of PageRank or PaperRank is in the fact that *every* paper may change the value of all nodes in a graph. PageRank is a unique and stable solution for a graph (www pages set, or scientific papers set) as *whole*. In other words every paper counts. It is very hard to judge which index is “better”, but world-acknowledged company Google [18] may be treated as a live proof that the PageRank is a “good” measure.

4.2.5 PR-Hirsch

One of the most widely used indexes related to author is the H-index proposed by Jorge Hirsch in 2004 [36] and presented earlier. The H-index tries to value consistency in reputation: it is not important to have many papers, or many citations, but many papers with many citations. We propose to apply a similar concept to measure authors based on PR. However, we cannot just say that PRH is the maximum number q such that an author has q papers with rank q or greater. This is because while for H-index it may be reasonable to compare number of papers with number of citations the papers have, for PRH this may not make sense as PR is for ranking, not to assign a meaningful absolute number to a paper. The fact that a paper has a CC of 45 is telling us something we can easily understand (and correspondingly we can understand the H-index), while the fact that a paper has a PR of 6.34 or 0.55 has little "physical meaning". In order to define a PR-based Hirsch index, we therefore rescale PR so that it gets to a value that can be meaningfully compared with the number of papers. Let's consider in some detail our set: we have a graph with N nodes (vertices) and n citations (edges). Each i-th node has PR equal to P_i, that expresses the probability for a random surfer to visit a node, as in the Page Rank algorithm. So let's assume that we run exactly n surfers (equal to quantity of citations), and calculate the most probable quantity of surfers who visited node i. If the probability to visit the node i for one surfer is p_i, expectation value Q_i for n surfers to visit the node i will be $p_i \cdot n$, which is most probable quantity of surfers, who visited node i. We multiply probabilities since all surfers are independent. To be precise we should first normalize PR for each node according to full probability condition: $\sum_i p_i = 1$. If the total sum of all PRs equals to M, the expected value for n surfers is as follows:

$$Q_i = P_i \frac{n}{M} \tag{4.5}$$

Where P_i is a Paper Rank of the paper i, n/M is the constant ≈ 5.9169 for our citation graph. So in other words we rescale PR to make it comparable with the quantity of citations. Indeed, Q_i is the most probable quantity of surfers who visited a specific paper i, whereas to compute Hirsch index we use quantity of citations for the paper i. It is interesting to compare the ranges of Q and citation count (see 4.1). Following the definition of H-index and the previous discussion, we define PR-Hirsch as the maximum integer number h such that an author has at least h papers with Q value (i.e. re-scaled PR following equation 4.5) equal or greater than h.

Average Q	Maximum Q	Average CC	Maximum CC
3.57	1326.77	3.57	1736

Table 4.1: Comparison of citation count and random surfers count mathematical expectation values for all papers in graph.

4.3 Exploring the Differences in Paper Matrics

This section explores the extent of the differences between paper metrics PR and CC when ranking papers, and their causes. As part of the analysis we introduce concepts and indexes that go beyond the PR vs CC analysis, and that are generally applicable to understanding the effects and implications of using a certain index rather than another for assessing papers' value.

4.3.1 Plotting the difference between paper metrics

The obvious approach to exploring the effect of using PR vs CC in evaluating papers would consist in plotting these values for the different papers. Then, the density of points that have a high CC and low PR (or vice versa) would provide an indication of how often these measures can give different quality indication for a paper. This leads however to charts difficult to read in many ways: first, points overlap (many papers have the same CC, or the same PR, or both). Second, it is hard to get a qualitative indication of what is "high" and "low" CC or PR. Hence, we took the approach of dividing the CC and PR axis in bands. Banding is also non-trivial. Ideally we would have split the axes into 10 (or 100) bands, e.g., putting in the first band the top 10% (top 1%) of the papers based on the metric, to give qualitative indications so that the presence of many papers in the corners of the chart would denote a high divergence. However the overlap problem would remain, and it would distort the charts in a significant way since the measures are discrete. For example the number of papers with 0 citations is well above 10%. If we neglect this issue and still divide in bands of equal size (number of papers), papers with the same measure would end up in different bands. This gives a very strong biasing in the chart (examples are provided in the companion page). Finally, the approach we took (Figure 4.2) is to divide the X-axis in bands where each band corresponds to a different citation count measure. With this separation we built 290 different bands, since there are 290 different values for CC (even if there are papers with much higher CC, there are only 290 different CC values in the set). For the Y-axis we leverage mirrored banding, i.e., the Y-axis is divided into as many bands as the X-axis, also in growing values of PR. Each Y band contains the same number of papers as X (in other words, the vertical rectangle corresponding to band i in the X axis contains the same number of papers q_i as the horizontal rectangle corresponding to band i of the Y-axis). We call a point in this chart as a square, and each square can contain zero, one, or many papers. The reasoning behind the use of mirrored banding is that this chart emphasizes divergence as distance from the diagonal (at an extreme, plotting a metric against itself with mirrored banding would only put papers in the diagonal). Since the overlap in PR values is minimal (there are thousands of different values of PR and very few papers with the same PR values, most of which having very low CC and very low PR, and hence uninteresting), it does not affect in any qualitatively meaningful way the banding of the Y-axis.

Table 4.2 gives an indication of the actual citation and PR values for the different bands.

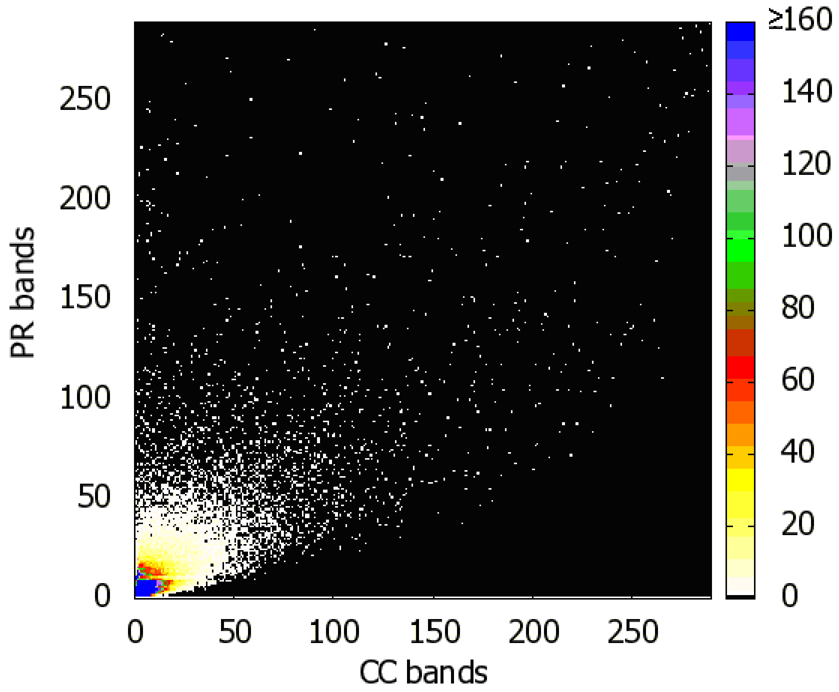

Figure 4.2: CC vs PR. X axis plots CC bands, Y axis plots PR mirror-banded by CC. The color corresponds to the number of papers within a band. (For actual values of PR and CC for each band see Table 4.2).

The chart in Figure 4.2 shows a very significant number of papers with a low CC but a very high PR. These are the white dots (a white color corresponds to one paper). Notice that while for some papers the divergence is extreme (top left) and immediately noticeable, there is a broad range of papers for which the difference is still very significant from a practical perspective. Indeed, the very dense area (bands 1-50) includes many excellent papers (CC numbers of around 40 are high, and even more considering that we only have citations from ACM papers). Even in that area, there are many papers for which the band numbers differ significantly if they are ranked by CC or PR.

To give a quantitative indication of the difference, Table 4.3 below shows how far apart are the papers from the diagonal. The farther away the

Number of band both for CC and PR	CC	PR
50	50	6.23
100	100	14.74
150	151	26.57
200	213	38.82
250	326	58.86
280	632	113.09
290	1736	224.12

Table 4.2: Mapping of band number to the actual value of CC or average actual value for PR.

papers, the more the impact of choosing an index over another for the evaluation of that paper.

Distance in bands from the diagonal	% of papers with this distance
0	36.83
1	24.30
2	13.02
3	5.76
4	5.43
5	2.50
6	1.70
7	1.34
8	1.86
9	1.57
10	0.79
≥11	4.89

Table 4.3: Deviation of papers around main diagonal.

The mean value for the distance from the main diagonal is 3.0 bands, while the standard deviation is 3.4. This deviation from the average is rather significant, i.e. in average the papers are dispersed through 3 bands around main diagonal. In the subsequent discussion, we will qualitatively refer to papers with high PR and high CC as popular gems, to paper with high PR and low CC as hidden gems, to papers with low PR and high CC as popular papers, and to papers with low CC and PR as dormant papers (which is an optimistic term, on the assumption that they are going to be noticed sometime in the future).

4.3.2 Divergence

The plots and table 4.3 above are an attempt to see the difference among metrics, but it is hard from them to understand what this practically means. We next try to quantitatively assess the difference in terms of concrete effects of using a metric over another for what metrics are effectively used, that is, ranking and selection. Assume we are searching the Web for papers on a certain topic or containing certain words in the title or text. We need a way to sort results, and typically people would look at the top result, or at the top 10 or 20 results, disregarding the rest. Hence, the key metric to understand divergence of the two indexes is how often, on average, the top t results would contain different papers, with significant values for $t = 1, 10, 20$. In the literature, the typical metric for measuring a difference between two rankings is the Kendall τ distance [43], measured as the number of steps needed to sort bi-ranked items so that any pair A and B in the two rankings will satisfy to the condition

$$sign(R_1(A) - R_1(B)) = sign(R_2(A) - R_2(B)) \tag{4.6}$$

where R_1 and R_2 are two different rankings. However, this measure does not give us an indication of the practical impact of using different rankings, both for searching papers and, as we will see later, for authors. What we really want to understand is to see the distance between two rankings based on the actual paper search patterns. Assume we are searching the

Web for papers on a certain topic or containing certain words in the title or text. We need a way to sort results, and typically people will look at the top result, or at the top 10 or 20 results, disregarding the rest. Hence, the key metric to understand divergence of the two indexes is how often, on average, the top t results would contain different papers, with significant values for $t = 1, 10, 20$. For example, the fact that the papers ranked 16 and 17 are swapped in two different rankings is considered by the Kendall distance, but is in fact irrelevant from our perspective. To capture this aspect, we propose a metric called divergence, which quantitatively measures the impact of using one scientometric index versus the other. Consider two metrics M1 and M2 and a set of elements (e.g., of papers) S. From this set S, we take a subset n of elements, randomly selected. For example, we take the papers related to a certain topic. These n papers are ranked, in two different rankings, according to two metrics M1 and M2, and we consider the top t elements. We call divergence of the two metrics, $Div_{M1,M2}(t, n, S)$, the average number of elements that differ between the two sets (or, t minus the number of elements that are equal). For example, if S is our set of ACM papers, and n are 1000 randomly selected papers (say, the papers related to a certain topic or satisfying certain search criteria), $Div_{CC,PR}(20, 1000, S)$ measures the average number of different papers that we would get in the typical 20-item long search results page. We measured the divergence experimentally for CC and PR, obtaining the results in the table 4.3.2 below. As a particular case, $Div_{M1,M2}(1, n, S)$ measures how often does the top paper differs with the two indexes.

The table 4.3.2 is quite indicative of the difference, and much more explicit than the plots or other evaluation measures described above. In particular, the table shows that more than almost 2/3 of the times, the top ranked paper differs with the two metrics. Furthermore, and perhaps even more significantly, for the traditional 20-element search result page, nearly

t	$Div_{PR,CC}(t, 1000, S)$, in %	$Div_{PR,CC}(t, 1000, S)$
1	62.40	0.62
10	49.94	4.99
20	46.42	9.28
40	43.29	17.31
60	42.51	25.5
80	41.75	33.39
100	40.52	40.52

Table 4.4: Experimentally measured divergence for the set of ACM papers.

half of the paper would be different based on the metric used. This means that the choice of metric is very significant for any practical purposes, and that a complete search approach should use both metrics (provided that they are both considered meaningful ways to measure a paper). In general we believe that divergence is a very effective way to assess the difference of indexes, besides the specifics of CC and PR. We will also see the same index on authors, and the impact that index selection can therefore have on people's careers. Details on the experiments for producing these results and the number of measures executed are reported in the companion web page.

4.3.3 Understanding the difference

We now try to understand why the two metrics differ. To this end, we separate the two factors that contribute to PR, see equation 4.1: the PR measure of the citing papers and the number of outgoing links of the citing papers (or numerator and denominator). To understand the impact of the weight, we consider for each paper P the weight of the papers citing it (we call this the potential weight, as it is the PR that the paper would have if all the citing papers P only cited P). We then plot (Figure 4.3) the average potential weight for the papers in a given square (intersection of a

CC and a PR band) in the banded chart. The estimation of the impact of outgoing links will be done in the following way.

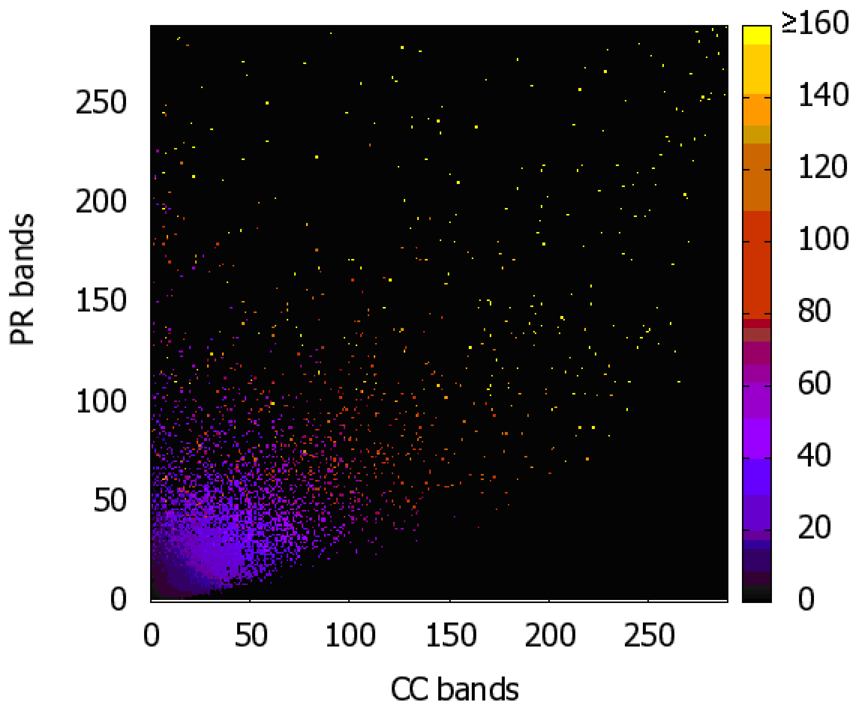

Figure 4.3: Average potential weight for all papers in a square The color in the Z-axis denotes the weight X axis plots CC bands, Y axis plots PR mirror-banded by CC.

What we want to see when examining the effect of outgoing links from citing paper, is the "weight dispersion", that is, how much weight of the incoming papers (i.e., how much potential weight) is dispersed through other papers as opposed to being transmitted to P. This is really the measure of the "damage" that outgoing links do to a Paper Rank. We compute the dispersed weight index for a paper P ($DW(P)$) as the sum of the PR of the citing papers $C(P)$ (that is, the potential weight of P) divided by the PR of P (the actual weight). Figure 4.4 plots the average dispersed weight for each square, as usual by CC and PR. The dark area in the bottom right corner is because there are no papers there.

These two charts very clearly tell us that outgoing links are the dominant effect for the divergence between CC and PR. Papers having a high CC and low PR have a very high weight dispersion, while papers with high PR and low CC are very focused and able to capture nearly all potential weight. The potential weight chart (Figure 4.3) also tends to give higher numbers for higher PR papers but the distribution is much more uniform in the sense that there are papers in the diagonal or even below the diagonal and going from the top left to the bottom right the values do changes but not in a significant way (especially when compared to the weight dispersion chart). To see the difference concretely on a couple of example, we take a "hidden gem" and a "popular paper", see Figure 4.5.

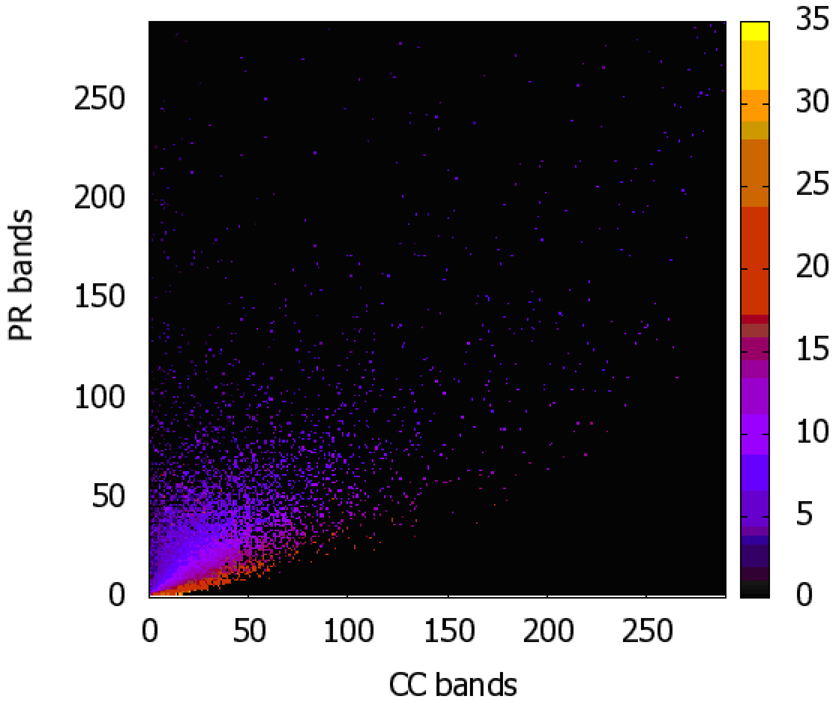

Figure 4.4: Average dispersed weight for all papers in a square The color in the Z-axis denotes the weight X axis plots CC bands, Y axis plots PR mirror-banded by CC.

The specific gem is the paper Computer system for inference execution and data retrieval, by R. E. Levien and M. E. Maron, 1967. This paper has 14 citations in our ACM-only dataset (Google Scholar shows 24 citations

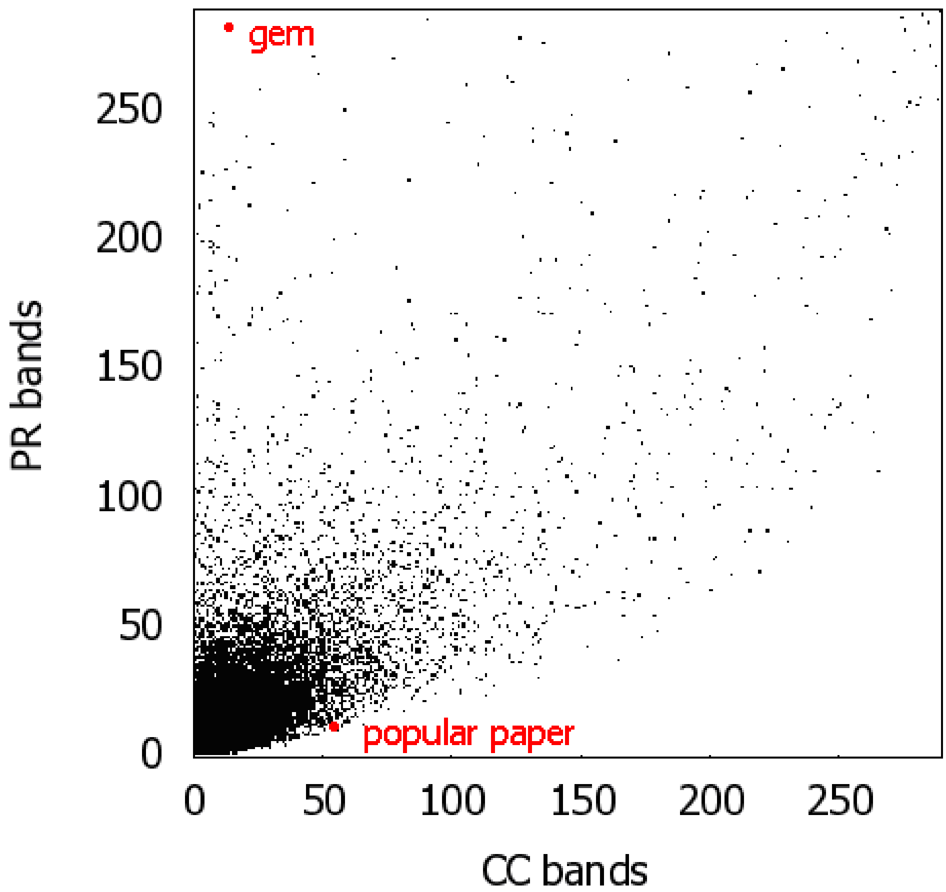

Figure 4.5: "Gem" and "popular paper" (or "stone") relative positions.

for the same paper). The PR of this "hidden gem" is 116.1, which is a very high result: only 9 papers have a greater rank. Let's go deep inside the graph to see how this could happen. Figure 4.6 shows all the incoming citations for this paper up to two levels in the citation graph. The paper in the center is our "gem", and this is because it is cited by an heavyweight paper that also has little dispersion: it cites only two papers. We observe that this also means that in some cases a pure PR may not be robust, meaning, the fact that our gem is cited by a heavyweight paper may be considered a matter of "luck" or a matter of great merit, as a highly respected "giant" is citing it. Again, discussing quality of indexes and which is "better" or "worse" is outside our analysis scope, as is the suggestion for the many variations of PR that could make it robust.

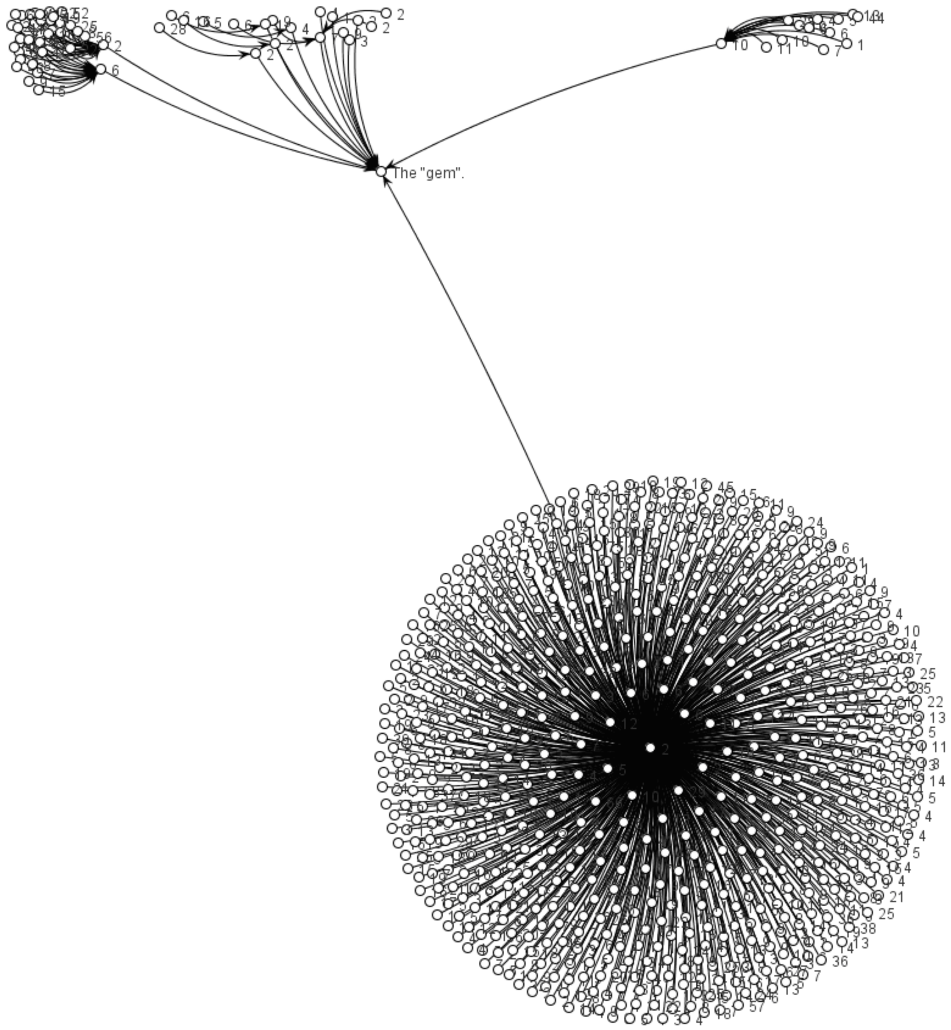

Figure 4.6: One of the “hidden gem” in the dataset, paper of E. Levien and M. E. Maron (in the center). Arrows refer to incoming citations. The digits near the papers refer to the quantity of outgoing links.

We now consider a paper in the bottom of the CC vs PR plot, a paper with high number of citations but relatively low PR. The corresponding citation graph is shown in Figure 4.7. This paper has 55 citations in our ACM dataset (158 citations in Google Scholar) and a relatively poor PR of 1.07. This result is not particularly bad, but it is much worse than other papers with similar number of citations. There are 17143 papers in

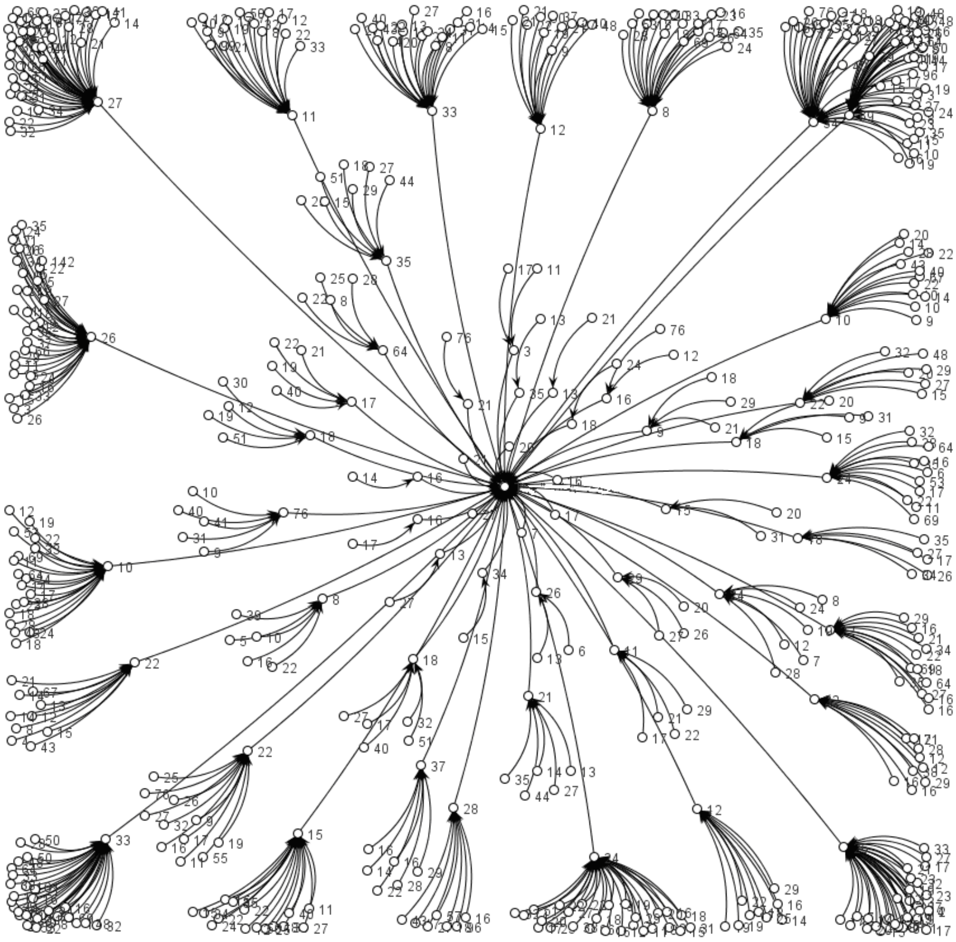

Figure 4.7: "Popular paper" (in the center): relatively highly cited but not very well-ranked.

the dataset that have grater Paper Rank and just 1394 papers with better citation count. Comparing with papers in the same CC and PR band, this paper has a weight dispersion factor that is over twice that of papers in the same CC band and three times the one of papers in the same PR band, which explain why the increased popularity with respect to papers in the same PR band did not correspond to a higher PR. As a final comment, we observe that very interestingly there are papers with very low CC and very high PR, but much less papers - almost none - with very high CC and very low PR. If we follow the dispersion plot this is natural, as it would assume that the dispersed weight should be unrealistically high (many papers with hundreds of citations) which does not happen in practice, while it is possible to have "heavyweight" papers with very few citations that make the presence of paper gems (papers in the top left part) possible. However, we believe that the absence of papers in the bottom right part and, more in general, the skew of the plot in Figure 4.2 towards the upper left is indicative of a "popularity bias". In the ideal case, an author A would read all work related to a certain paper P and then decide which papers to reference. In this case, citations are a very meaningful measure (especially if they are positive citations, as in the motto "standing on the shoulders of giants"). However this is impossible in practice, as nobody can read such a vast amount of papers. What happens instead is that author A can only select among the papers she "stumbles upon", either because they are cited by other papers or because they are returned first in search results (again often a result of high citation count) or because they are published in important venues. In any event, it is reasonable to assume that authors tend to stumble upon papers that are cited more often, and therefore these papers have a higher chance of being cited than the "hidden gems", even if maybe they do not necessarily have the same quality. We believe that it is for this reason that over time, once a paper increases with citation count, it necessarily increases with the weight, while gems may remain "hidden" over time. A detailed study of this aspect (and of the proper techniques for studying it) is part of our future work.

4.4 Focused PaperRank

The computational procedures to acquire PageRank or PaperRank have been described in Subsection 4.2.3, let us now try to look over this problem from another end. How PageRank or PaperRank may be changed and why. From the variety of possible PageRank modifications I would like to enumerate the following ones:

- PR Computation with or without dump factor (see formula 4.1, 4.3 above).

- Personalized Page Rank with some initial personalization vector is more common for web-search engines. Here all pages have their own personal weights before PR calculation.

- Focusing of PR, or redistribution of links to link probabilities in the stochastic Markov matrix. This means that core PR model of Random Surfer is no longer Random, it becomes focused. This model was successfully applied to the web pages ranking problem by Tony Abou-Assaleh *et al.* [1] and by Fuyong Yuan *et al.* [104] in 2007. Most recent application of Focused Random Surfer model is applied by Prof. Lee Giles for ranking items in Autonomous Digital Library CiteSeerX [88]. This is the closest research to the present one to the best of our knowledge.

- Double (or more) focusing of PR [22] takes into account more deep properties of citation graph entities during stochastic Markov matrix composition. For example, it may first focus on site name and then on site content.

4.4.1 Focused Surfer

The Random Surfer model is the basis of PageRank algorithm. PageRank of the certain node is proportional to the probability to reach this node by randomly riding the graph. At each step rider randomly chooses the link to follow. Focused Surfer decides which path is more preferable for him.

Formula 4.1 may be rewritten to better expresses this mathematically,

$$P_i = (1-d)\sum_{i\neq j}^{j\in D} P_j \cdot s(j|i) + \frac{d}{N} \tag{4.7}$$

where $s(j|i)$ is the probability to follow the reference i being at the place j. s is a function that may be arbitrary. We propose to use the simplest variant of it, which we show in formula

$$s(j|i) = \frac{C(i)}{\sum_{k\in D} C(k)} \tag{4.8}$$

where C(m) is paper m citations count, and D is the set of all references in paper C(j). This means that more cited nodes have advantage and they are more visible and attractive for further citation. More complex analogue of such focusing proposal I had been found after publishing of this contribution [50] in the paper of Prof. Lee Giles [88].

4.4.2 Evaluation, comparison with normal PR

Evaluation for the problem of Focused Paper Rank is performed for the dataset presented in the Section 4.2.1. We use the same mirrored-plotting methodology described in Subsection 4.3.1. The result is shown in Figure 4.8. In the top (sub-figure a)) we see the original Paper Rank, in the bottom b) the focused one.

Figure 4.8 b) (in the bottom) illustrates the Focused Surfer model and FPR algorithm instead of PR. Focused Surfer model gives better chances to more cited papers, at the same time stealing the part of the weight from their poorly cited neighbors. This idea leads us to the conclusion that in general total FPR rank remains the same as PR, it just gets re-distributed. This idea is supported by computation of average FPR and PR which are nearly the same: $< FPR >$=0.603 and $< PR >$=0.602. Now let us observe effects present in Figure 4.8 b). The points are located closer to the main diagonal (comparing with plot a)) and there is significantly less papers with big CC and small PR (reducing of the effect of outbound links). On the other hand we see that "gems"-effect is still noticeable. This means that FPR tends to be a "middle" between PR and CC.

4.4.3 Understanding the difference between PR and FPR

Focused Page Rank major strong points are:

1. It is the tradeoff between Page Rank and Citation Count. So it may serve as an agreement between the followers of pure citation count and Page Rank followers.

2. Proposed solution less suffers from the effect of outbound links.

3. It reflects one of the fundamental principles of Scientometrics, first time formulated by de Solla Price in 1976: *"Success seems to breed success. A paper which has been cited many times is more likely to be cited again than one which has been little cited. An author of many papers is more likely to publish again than one who has been less prolific. A journal which has been frequently consulted for some purpose is more likely to be turned to again than one of previously infrequent use"*.

4. It captures the power of Page Rank, where not only the quantity of citations, but also the quality of ones counts.

4.5 Exploring Author Metrics

4.5.1 Plotting the difference between author metrics

We now perform a similar analysis on authors rather than papers. For this, we initially consider PRH and Hirsch as main metrics, and then extend to other metrics. The plot to visualize the differences (Figure 4.9) is similar in spirit to the one for CC vs PR. The X-axis has Hirsch values, while the Y-axis has PRH values. A first observation is that applying “Hirsching” to CC and PR to get H-index and PRH smoothes the differences, so we do not have points that are closer to the top left and bottom right corners. This could only happen, for example, if one author had many papers that are hidden gems.

Since the authors with low Hirsch and PRH are dominant, a log scale was used plotting Figure 4.9. This increased similarity is also shown in Table 4.5.1, where many papers are on the diagonal (this is also due to the fact that we have a much smaller number of squares in this chart). The mean distance from the diagonal is 0.25 bands, while the standard deviation is 0.42 bands. Interestingly, as we will see, though at first look the differences seem less significant, the impact of using one rather than the other index is major.

Distance in bands from the main diagonal	Percent of authors with this distance
0	83.07%
1	12.23%
2	2.90%
3	0.99%
4	0.40%
5	0.19%
6	0.09%
7	0.05%
8	0.03%
9	0.02%
10	0.01%
$\geq$11	0.01%

Table 4.5: Deviation of authors around main diagonal.

4.5.2 Divergence

The same measure of divergence described for papers can be computed for authors (since divergence is a universal measure and may be applied to different ranking schemas). The only difference is that now the set S is a set of authors, and that the indexes are H-index and PRH instead of CC and PR. We also compute it for n=100, as the experiment we believe it is meaningful here is to consider replies to a typical job posting for academia or a research lab, generating, we assume, around 100 applications.

t	$Div_{PRH,H}(t)$ divergence for PR-Hirsch and Hirsch
1	59.3%
5	50.04%
10	46.13%
20	43.47%

Table 4.6: Divergence between PRH and H, $n = 100$.

Although nobody would only make a decision based on indexes, they are used more and more to filter applications and to make a decision in case of close calls or disagreements in the interview committees. The Table 4.5.2 tells us that almost two third of the times, the top candidate would differ. Furthermore, if we were to filter candidates (e.g., restrict to the top 20), nearly half of the candidates passing the cutoff would be different based on the index used. This fact emphasizes once again that index selection, even in the case of both indexes based on citations, is key to determining the result obtained, be them searching for papers or hiring/promotion of employees. Notice also that we have been only looking at differences in the elements in the result set. Even more are the cases where the ranking of elements differ, even when the t elements are the same. Another interesting aspect is that the divergence is so high even if the plot and Table 4.5.1 show values around the diagonal. This is because most of the authors have a very low H and PRH (these accounts for most of the reasons why authors are on average on the diagonal). However, and this can also be seen in the plot, when we go to higher value of H and PRH, numbers are lower and the distribution is more uniform, in the sense that there are authors also relatively far away from the diagonal (see the softer colors and the distributions also far from the diagonal towards the top-right quadrant of Figure 4.9). Incidentally, we believe that this confirms the quality of divergence as a metric in terms of concretely emphasizing the fact that the choice of index, even among citation-based ones, has a decisive effect on the result. We omit here the section on "understanding the difference" as here it is obvious and descends from the difference between CC and PR, described earlier and used as the basis for PRH and Hirsch respectively.

4.5.3 Divergence between other indexes

The discussion above has focused on PRH vs H. We now extend the same analysis to other indexes. The table 4.5.3 below shows a comparison for PRH, H, G index, and the total citation count for an author (the sum of all citations for the paper by an author, denoted as TCC in the table).

t	PRH vs G	PRH vs TCC	H vs TCC	H vs G	G vs TCC
1	56.3	56.4	38.2	34.6	29.9
5	45.66	46.38	29.48	25.58	23.84
10	43.05	43.03	27.9	22.94	22.95
20	41.3	41.66	27.63	21.70	22.62

Table 4.7: Divergence for the different indexes in %, $n = 100$ (for simplicity the $Div()$ notation is omitted).

The first lesson we learn from the table is that no two indexes are strongly correlated. The higher correlation is between G and the total citation count, and we still get the top choice different in one out of four cases. The other interesting aspect is that PRH and H are the pair with the highest divergence, which makes them the two ideal indexes to be used (in case one decides to adopt only two indexes).

4.6 Discussion

We have explored the problem of ranking of scholarly papers in citation networks. We argue in favor of usage of invented modification of PageRank – PaperRank as the proper measure of an impact of scientific paper. The major argument here is that the PaperRank takes into account not just the quantity of citations but also the quality of ones. Another achievement of the Thesis is the adaptation of Focused Paper Rank for scientific citing. Of course, it cannot displace traditional measure: Citation Count, but can (and we believe should) be used in parallel.

The same thing is about PR-based Hirsch index, it is a try to grasp not just the broadness of an author in terms of quantity of citations per paper, but also an attempt to estimate the PR-based impact of a certain researcher.

The other interesting finding of the present Thesis is in understanding and visualizing the difference between various indexes. Here we mention the innovative methodology of plotting the difference and computation of divergence among different indexes.

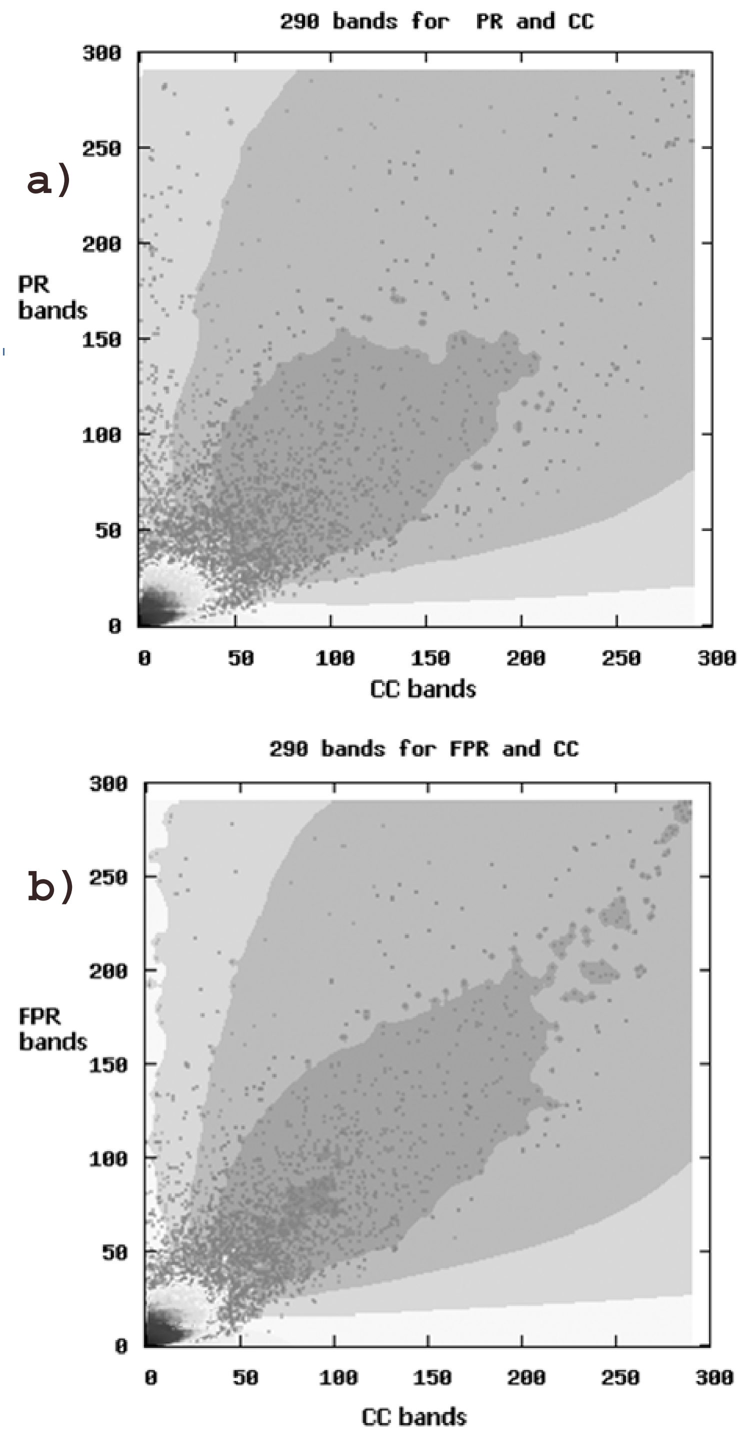

Figure 4.8: Diversity of PR a), FPR b) and Citation Count CC. White and black points in the bottom-left corner does not mean absence of papers. This is a gray-scale of colored map, where the major quantity of papers has small number of CC, and since lie exactly in the bottom-left corner and it is nearly the same for the both plots.

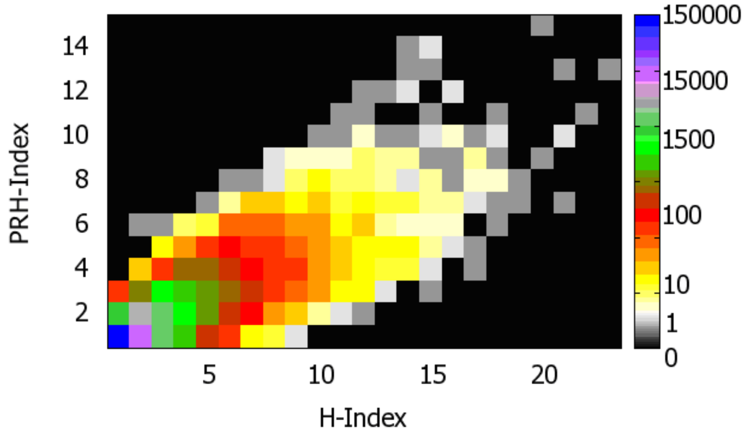

Figure 4.9: The gradient of Hirch and PRHirch in log scale. Author's density is plotted with colors: authors' number goes from 1 to 149170 of authors per square. PR-Hirch has been rounded.

Chapter 5

ResEval+: Embedding Improved Search, Navigation and Rank in ADL

This Chapter is dedicated to the description of a prototype application using both the keyphrases mining methods and ranking schemas described in the previous chapters.

5.1 PageRank in ResEval tool

Recently, we see a rapid evolution of web-based communities, blogs and social networks. Modern web is interactive and people connect to each other through social web application (also named "web 2.0" applications). These developments has made it possible to gather users' feedbacks and opinions very quickly. In this way web communities may pick-up interesting documents, artifacts, opinions, comments, short articles and news and share them with high speed around the world.

The idea of exploring these potential also in professional and - more specifically - in scientific domain is the the main research carried out in the European Union LiquidPub [11] project that started in 2008. Within the scope of this project, I have contributed in the design and implementation of a tool named ResEval for gathering information on scientific documents from Google Scholar [17] and computing a number of research impact metrics, like the Hirsch h index, g index and some other indexes.

A screen shot of an initial GUI for the web application of the tool is shown in Figure 5.1:

Figure 5.1: ResEval home.

More details may be found at URL `http://project.liquidpub.org/reseval/` together with information about the architecture, the overall service-oriented framework as well as comparison with other similar services like Google Scholar [17]. More specific functionalities are shown in a second screen-shot of the application in Figure 5.2.

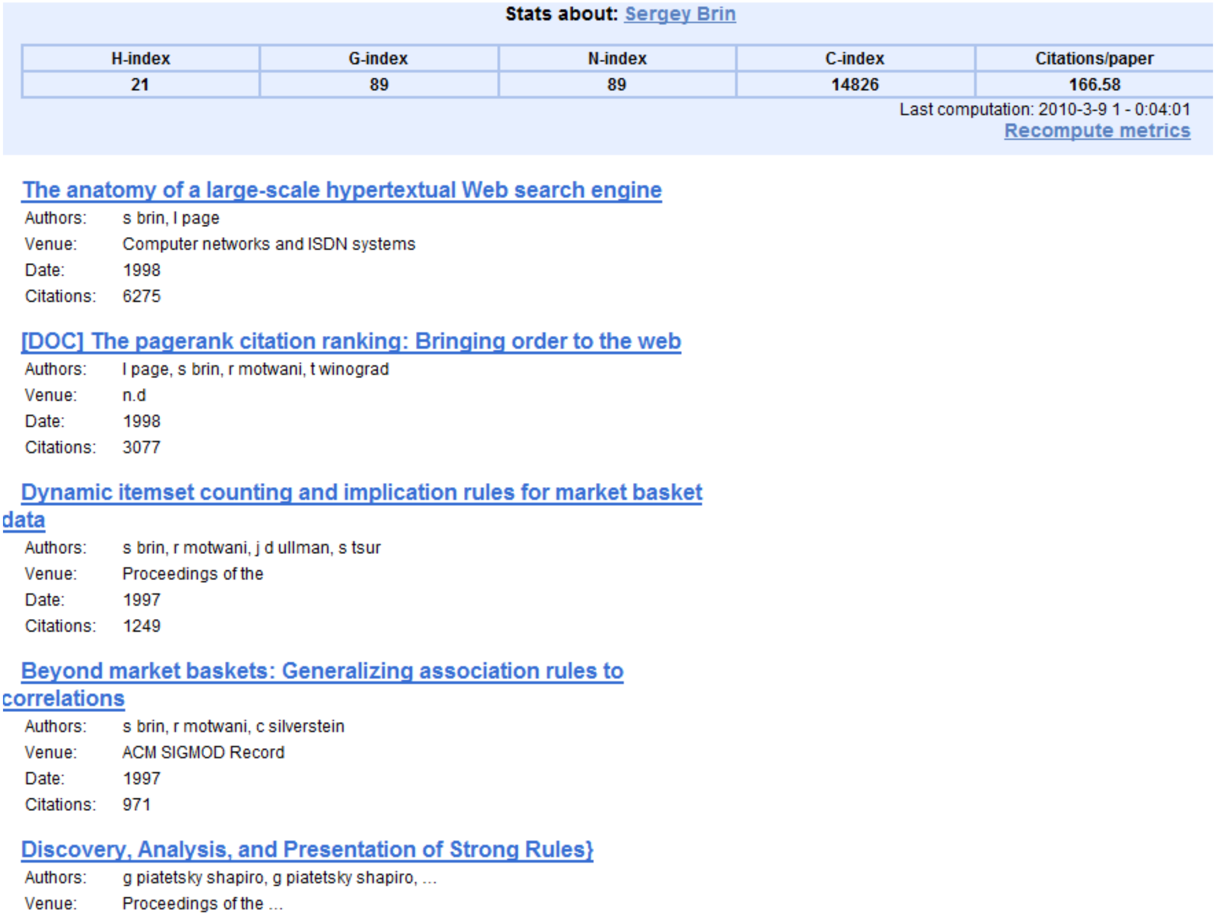

Figure 5.2: ResEval tool: with several indexes computed for an author.

There are two general types of queries supported in ResEval: a) query for author's impact, b) query for single paper's impact. A query may have the following options: certain publication years, broad branch of science like: "Physics" or "Computer Science" and the year of publication. Figure 5.2 illustrates the tool search result by query for Google co-founder, Sergey Brin [1].

In Figure 5.2 we see five indexes, namely H-index (see [36]), G-index, total quantity of papers, total quantity of citations and average quantity of

[1] http://portal.acm.org/author_page.cfm?id=81100070777&coll=GUIDE&dl=GUIDE&trk=0&CFID=79158426&CFTOKEN=71473616

citations by paper It is not possible to compute RageRank or PaperRank since we have not access to the whole citation graph present in Google Scholar. In case we would have access to it, it would be feasible to sort search result not just by citation count (as it is implemented at present) but also by PaperRank. The more queries users do, the more information we collect in our database, so we hope soon we will be able to reconstruct a significant piece of Google graph which will give us a chance to compute the PageRank.

This is - very briefly - the current usage of various indexes for assessing scientific progress of authors and individual papers in ResEval.

5.2 ResEval+ tool

We now introduce a prototypal tool for evaluating the new metrics proposed in the present Thesis. For constructing this tool we have extended the ResEval (see the enhanced tool at `http://demo.liquidpub.org:8081/ResevalGUI`) tool described above with the novel citation network-based ranks presented in the previous Chapter. In Figure 5.4 we show the same query for the Sergey Brin's scientific indexes, but on a different source, namely the dataset used in Chapter 4 and consisting of 266K of scientific papers present in the ACM library from 2003 to 2005. Obviously this is not the complete citation graph for Sergey Brin's publication but just a subset of the whole graph present in the ACM data set at our disposal. We think that the availability of different ranking metrics may support better search and navigation of content, improving in particular the diversity of the retrieved information. Since - as we have seen in the analysis presented in Chapter 4 - different metrics captures different dimensions of impact/quality of the infromation, the availability of diverse metrics support the navigation of a more varied content. PageRank-based indexes in combination with Citation Count or separately may be adopted for seeking for best papers in a domain and in some case will identify different contributions that would not have been found using only one metric. Some with the search of domain experts using both traditional Hirsch or G-indexes or novel PR-H index.

Moreover, on the same data set, we also exploited the methods and approach for semi-supervised key phrases extraction - presented in Chapter 4 - to be used for the tagged search. Tagging is a technique evolving from community-based web, and in brief it means assigning a certain word (of several words) to a content, article, news, blog post or other. It may be done by the author of a content or by an accidental reader. In other words, tagging is a kind of user-created "classification" of text information. It is widely spread in various web communities-oriented sites like youtube [2], twitter [3], linked-in [4], livejournal [5] and facebook [6]. Each tag is a short representation of a topic user writes about. Site visitors may see tags made by other people, make their own tags, select information by a certain tag. Same thing may be implemented in scientific Autonomous Digital Libraries. Intuitively we see that a keyphrase or a keyword carries the same sense as a tag. Thus both types of keyphrases, assigned by user or extracted automatically may improve the usability of large amounts of information. Being self-descriptive, they make easier not the search only, but also may be useful in academia domain, particularly for:

- State-of-the-art domain detection.
- Seeking for experts/best papers in a small sub-domain.
- Define classifications and categories more precisely.

[2]http://www.youtube.com/
[3]http://twitter.com/
[4]http://www.linkedin.com
[5]http://livejournal.com
[6]http://www.facebook.com/

In the real use-case of digital libary, e.g. ACM data set, some documents have already human assigned key phrases. However, a large number have not. As an example, let's take the contribution with ACM ID equal to 637415 and located by URL `http://portal.acm.org/citation.cfm?id=637415`. The details about this paper are:

- Title: *A conceptual architecture for semantic web enabled web services.*
- Abstract: *Semantic Web Enabled Web Services (SWWS) will transform the web from a static collection of information into a distributed device of computation on the basis of Semantic technology making content within the World Wide Web machine-processable and machine-interpretable. Semantic Web Enabled Web Services will allow the automatic discovery, selection and execution of inter-organization business logic making areas like dynamic supply chain composition a reality. In this paper we introduce the vision of Semantic Web Enabled Web Services, describe requirements for building semantics-driven web services and sketch a first draft of conceptual architecture for implementing semantic web enabled web services.*

The best of our machine learning approach (Random Forest with linguistic feature set) is capable to extract (once trained) the following keyphrases: "web enabled", "semantic web", "ontology". All extracted keyphrases seem to be reasonable and related to the topic of the specific contribution and can be used to described - as tags - the main concepts of the article. We also think that extracted key phrases can be used as a kind of bootstrapping tagging for new contribution that have not yet been tagged by the community.

In our prototype for a ResEval+ tool, we have explored the usage of a combination of both approaches: proper ranks with tagged content. This paradigm is implemented in the mockup of ADL search tool located by

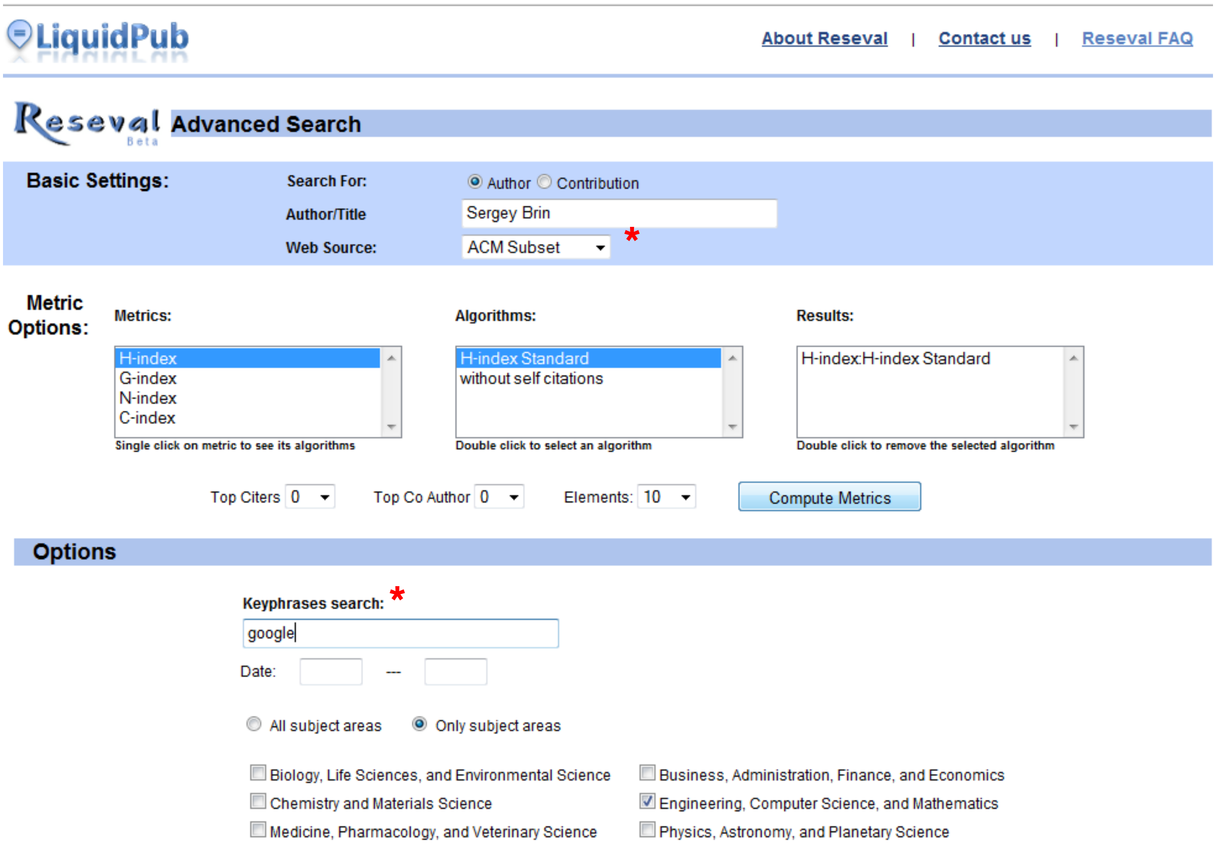

Figure 5.3: Enhanced ResEval tool: “advanced search” screen. The new features are marked with red asterisk.

URL `http://demo.liquidpub.org:8081/ResevalGUI`. To try out the tool one should click “advanced options” link and follow to the screen like in Figure 5.3, Here a user chooses all option available for usual Google Scholar-based search plus keyphrases. We emphasize that “Web Source” combobox must be set to a different data source, namely “ACM subset”. The results of such search are shown in Figure 5.4, where all new features are marked in red.

Let us briefly enumerate the new features, they are:

- Full authors names and surnames as compared with the short name obtained from Google Scholar. These full name can be better used for disambiguating authors.

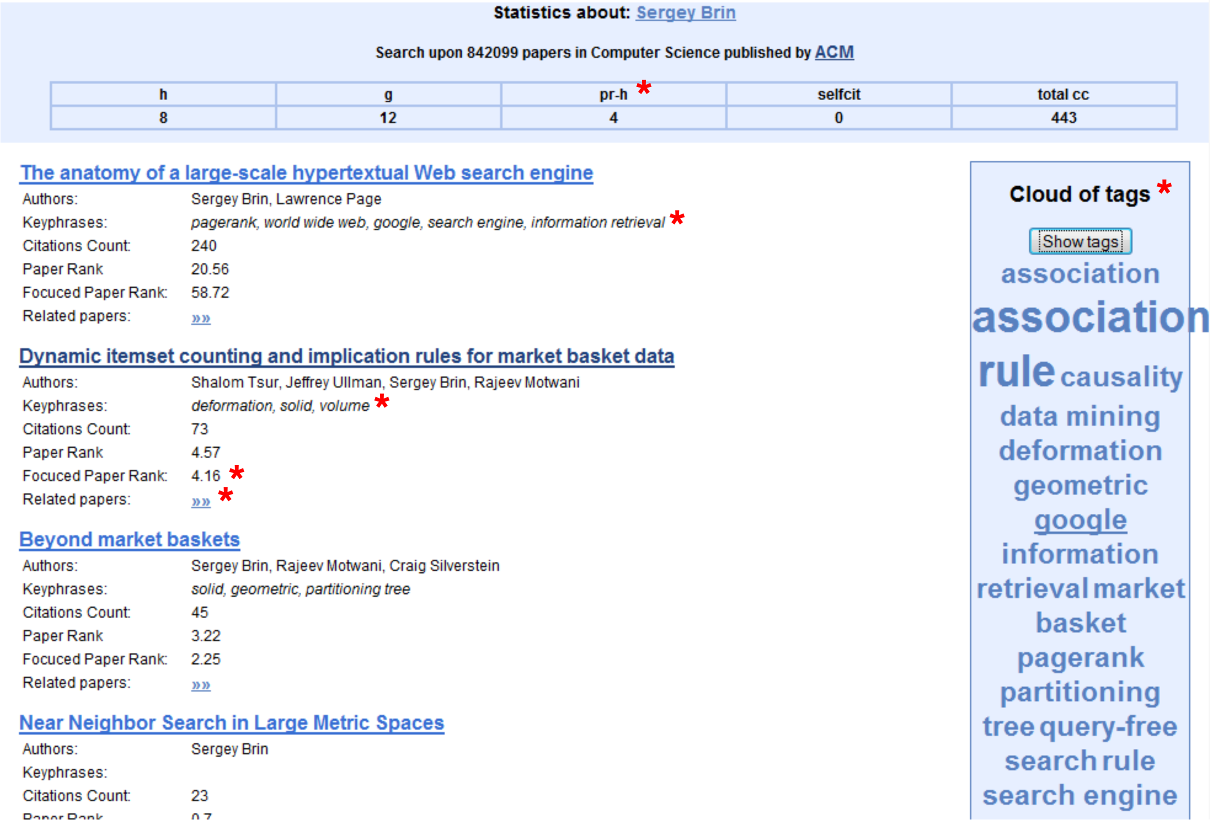

Figure 5.4: Enhanced ResEval tool: all new parts are marked with red asterisk.

- The titles of the contributions are “clickable” and lead to acm portal where all information may be verified.
- The “cloud of tags” box, where all important keyphrases which describe areas of expertise of an author are presented. The more frequent is the keyphrase, the bigger font we take to plot it.
- Co-occurrence of particular keyphrases may establish connections between documents and so we can find “related papers” (marked in red in the bottom), which may improve the navigation and search for bounds of a state-of-the-art.
- For author assessment novel Pr-Hirsch index is used, for an individual paper assessment we use novel citation-based indexes like PaperRank

and Focused PaperRank.

Such kind of tools may be exploited as an effective instrument for finding more interesting and relevant information about top people in domain, seeking for bounds of state-of-the-art, assessing candidates for promotion or candidates for hiring in academia. Evaluating scientific progress with the wide spectrum of ranks is more expressive, and keyphrases may effectuate navigation between domains, authors or related articles easier. The problem of search is getting more sharp since nowadays just one particular problem, for instance PageRank studied in Chapter 4, has thousands papers devoted to. It is getting simply impossible to read at least 30% of them. But with help of appropriate tools this problem could be alleviated, when most "noisy" (duplicated, incremental, erroneous) papers may be cut with help of proper ranking.

Chapter 6

Conclusion

6.1 Discussion of results

In the present Thesis we have addressed two open problems in modern autonomous digital libraries.

In the first part of our work, we focused on the ranking of a scholarly papers. We have explored and tried to understand and explain the differences among citation-based indexes. In particular, we have explored a variation of Page Rank algorithm specifically design for ranking papers - that we have named Paper Rank - and compared it to the standard citation count index. We have also analyzed related novel indexes for authors, in particular the Paper Rank Hirsch-index and compared it with the traditional H-index. We have explored in details the impact they can have in ranking and selecting both papers and authors. The following are the main findings of this line of work:

i) PR and CC are quite different metrics for ranking papers. A typical search would return half of the times different results.

ii) The main factor contributing to the difference is weight dispersion, that is, how much weight of incoming papers is dispersed through other papers as opposed to being transmitted to a particular paper.

iii) For authors, the difference between PRH and H is again very significant, and index selection is likely to have a strong impact on how people are ranked based on the different indexes. Two thirds of the times the top candidate is different, in an average application/selection process as estimated by the divergence.

iv) An analogous exploration of divergence between several citation-based indexes reveal that all of them are different in ranking papers, with g-index and total citation count being the most similar.

In addition to the findings, we believe that:

i) Divergence can be a very useful and generally applicable metric, not only for comparing citation-based indexes, but also for comparing any two ranking algorithms based on practical impact (results).

ii) There are a significant number of "hidden gems" while there are very few "popular papers" (non gem). The working hypothesis for this fact (to be verified) is that this is due to citation bias driven by a "popularity bias" embedded in the author's citation practices, i.e. authors tend to stumble upon papers that are cited more often, and therefore these papers have a higher chance of being cited.

The second problem that we have explored in the Thesis is the extraction of keyphrases from scientific papers. Here we created - and made publicly available - a high-quality data set that is - to the best of our knowledge - the largest public available for the specific domain of scientific documents (it includes *c.a.* 2000 full text documents).

The evaluation on the prepared dataset shows that:

i) The best results of NLP-based ML methods always outperform KEA in all quality parameters: Precision, Recall and overall F-Measure; in particular, it improves the average F-Measure from 22% (KEA) to 30% (Random Forrest) without the use of controlled vocabularies;

ii) A combination of KEA and RF may produce interesting result because both capture different keyphrases.

iii) Feature removal leads to stable decrease of F-Measure for all considered machine learning methods.

iv) Training set size increase improves F-Measure and reaches a “plateau” with training set size around 400 documents.

v) Random Forests is a good tradeoff between quality of keyphrases extraction and computational speed.

vi) NLP helps to avoid “strange” keyphrases candidates often extracted by a purely statistical systems; For instance KEA in some cases recognizes two keyphrases: “ad” and “hoc” instead of capturing “ad hoc” as the whole phrase.

vii) NLP is computationally expensive, but it provides more accurate keyphrases and proposed NLP heuristics cut search space by 50%.

We think that both problems: information extraction and ranking of more coarse-grained pieces of information (*c.a.* documents) are very related for good quality digital libraries construction. To support this opinion we have constructed two research tools: ResEval, the tool for observing major indexes for a particular author using web crawling; and the second one is the prototype of scientific ADL’s search system with both paradigms implemented: search by keyphrase and ranked search with PaperRank usage.

6.2 Future work

6.2.1 Data mining part

Apart from using syntactic knowledge to improve keyphrases extraction there is another powerful method of statistical texts analysis: each phrase should be taken into account in its own context. This means the piece of text around a phrase is very important for understanding how valuable is the particular phrase. This leads us to the area of *unsupervised* statistical learning. We do consider this method as a promising machine learning-based information mining for key phrases extraction.

6.2.2 Ranking

Ranking papers and researchers is a very delicate problem. We know for sure that different co-authors make different impact in the same paper; some people usually cite other works just because they have to provide a certain number of citations, so some works may be cited just by chance. Separating "noisy" papers, papers cited "by chance" from really breakthrough papers seems to be a very challenging and interesting research field. Having certain experience in the domain we do believe that it is statistically possible to draw interesting conclusions about paper impact just investigating graph structure. We think that the same importance papers have similar graph structure "around". So it is going to be similar to observing a "trace" of a paper in a net of citations. This is another dimension of the problem that we would like to explore.

Bibliography

[1] T. Abou-Assaleh, T. Das, G. Weizheng, M. Yingbo, P. O'Brien, and Z. Zhen. A link-based ranking scheme for focused search. In *In Proc. WWW2007*, pages 668–677, New York, USA, 2007. ACM Press.

[2] David Austin. Page rank simple samples, computaion, explanation. http://www.ams.org/featurecolumn/archive/pagerank.html.

[3] K. Bharat and G. A. Mihaila. When experts agree: Using non-affiliated experts to rank popular topics. In *Tenth International World Wide Web Conference*, 2001.

[4] Jinbo Bi and Vladimir Vapnik. Learning with rigorous support vector machines. In *COLT*, pages 243–257, 2003.

[5] M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Transactions on Internet Technology*, 5(1), 2005.

[6] Enrico Blanzieri and Farid Melgani. Nearest neighbor classification of remote sensing images with the maximal margin principle. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1804–1811, 2008.

[7] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[8] Leo Breiman. Random forests. *Machine Learning*, pages 5–32, 2001.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[10] Kurt Bryan and Tanya Leise. The 25,000,000,000 eigenvectors the linear algebra behind google. *SIAM Review*, 48(3):569–582, 2006.

[11] Fabio Casati, Fausto Giunchiglia, and Maurizio Marchese. Publish and perish: why the current publication and review model is killing research and wasting your money. *Ubiquity*, pages 1–1, January 2007.

[12] Fabio Casati, Maurizio Marchese, and Fausto Giunchiglia. Liquidpub project. http://www.liquidpub.org/.

[13] C.C. Chang and C.J. Lin. *LIBSVM: a library for support vector machines*, 2001.

[14] Chao Chen, Andy Liaw, and Leo Breiman. Using random forest to learn imbalanced data. Technical report, Department of Statistics, UC Berkeley, 2004.

[15] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with google. *Journal of Informetrics*, 2007.

[16] TREC conference. Trec collection. `http://trec.nist.gov/data.html`.

[17] Google Corporation. Google scholar autonomous digital library. http://scholar.google.com/.

[18] Google Corporation. Google search engine. http://www.google.com/.

[19] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[20] Isaac G. Councill, C. Lee Giles, Ernesto Di Iorio, Marco Gori, Marco Maggini, and Augusto Pucci. Towards next generation citeseer: A flexible architecture for digital library deployment. In *Research and Advanced Technology for Digital Libraries*, pages 111–122, 2006.

[21] D. deSolla Price. *Little Science - Big Science*. Columbia Univ. Press, New York, 1963.

[22] M. Diligenti, M. Gori, and M. Maggini. Web page scoring systems for horizontal and vertical search. In *World Wide Web Conference (WWW'2002)*, Honolulu, Hawaii, USA, May 2002.

[23] Elsevier. Elsevier. http://www.elsevier.com/.

[24] C. Ercan and I. Cicekli. Using lexical chains for keyword extraction. *Information Processing and Management*, 43:1705–1714, 2007.

[25] Christiane Fellbaum. *Wordnet: An Electronic Lexical Database*. Bradford Books, 1998.

[26] Association for Computing Machinery. Acm digital library. http://portal.acm.org/portal.cfm.

[27] E. Garfield. *Citation Indexing*. ISI Press, 1979.

[28] E. Garfield. The agony and the ecstasy - the history and meaning of the journal impact factor. In *International Congress on Peer Review And Biomedical Publication*, pages 605–612, Chicago, USA, 2005.

[29] C. Lee Giles. Citeseer: Past, present, and future. In *Atlantic Web Intelligence Conference (AWIC)*, 2004.

[30] Lee Giles. Citeseer autonomous digital library. http://citeseer.ist.psu.edu/.

[31] Lee Giles. Citeseerx autonomous digital library. http://citeseerx.ist.psu.edu/.

[32] Fausto Giunchiglia, Mikalai Yatskevich, Paolo Avesani, and Pavel Shvaiko. A large dataset for the evaluation of ontology matching. *Knowledge Eng. Review*, 24(2):137–157, 2009.

[33] Fausto Giunchiglia, Ilya Zaihrayeu, and Uladzimir Kharkevich. Formalizing the get-specific document classification algorithm. In *ECDL*, pages 26–37, 2007.

[34] Apache Group. Nutch crawler. http://lucene.apache.org/nutch/.

[35] Leximancer group. Leximancer. `http://leximancer.com/`.

[36] J[orge]. E. Hirsch. An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences*, pages 16569–16572, November 2005.

[37] H. Hui, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, , and E. A. Fox. Automatic document metadata extraction using support vector machines. In *in Proceedings of the 3rd ACM/IEEE-CS JCDL)*, pages 37–48, 2003.

[38] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing)*, volume 10, pages 216–223, 2003.

[39] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing)*, volume 10, pages 216–223, 2003.

[40] IEEE. Ieee digital library. http://www.ieee.org/.

[41] Alexander Ivanyukovich and Maurizio Marchese. Unsupervised free-text processing and structuring in digital archives. In *in Ist International Conference on Multidisciplinay Information Sciences and Technologies*, 2006.

[42] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. Extrapolation methods for accelerating pagerank computations. In *WWW*, Budapest, Hungary, 2003.

[43] M. Kendall and J. D. Gibbons. *Rank Correlation Methods.* Edward Arnold, London, 1990.

[44] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *In Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms*, pages 668–677, New York, USA, 1998. ACM Press.

[45] M. Krapivin, M. Marchese, A. Yadrantsau, and Yanchun Liang. Automated key-phrases extraction from scientific papers using domain and linguistic knowledge. In *Digital Information Management. ICDIM 2008*, pages 105–112, London, GB, Nov 2008.

[46] Mikalai Krapivin. Focused page rank application for scientific papers ranking in digital libraries. In *VLDL workshop, ECDL conference*, Aarhus, Denmark, 2008.

[47] Mikalai Krapivin, Aliaksandr Autayeu, and Maurizio Marchese. Large dataset for keyphrases extraction. Technical Report DISI-09-055, DISI, Trento, Italy, May 2008. http://eprints.biblio.unitn.it/archive/00001671/01/disi09055-krapivin-autayeu-marchese.pdf.

[48] Mikalai Krapivin, Aliaksandr Autayeu, Maurizio Marchese, Enrico Blanzieri, and Nicola Segata. Improving machine learning approaches for keyphrases extraction from scientific documents with natural language knowledge. Technical Report DISI-10-003, Trento, DISI, January 2010.

[49] Mikalai Krapivin, Aliaksandr Autayeu, Maurizio Marchese, Enrico Blanzieri, and Nicola Segata. Keyphrases extraction from scientific documents: Improving machine learning approaches with natural language processing. In *ICADL*, Gold Coast, Australia, 2010.

[50] Mikalai Krapivin and Maurizio Marchese. Focused page rank in scientific papers ranking. In *ICADL*, pages 144–153, 2008.

[51] Mikalai Krapivin, Maurizio Marchese, and Fabio Casati. Exploring and understanding citation-based scientific metrics. Technical Report DISI-08-022, DISI, Trento, Italy, May 2008.

[52] Mikalai Krapivin, Maurizio Marchese, and Fabio Casati. Exploring and understanding scientific metrics in citation networks. In *Complex (2)*, pages 1550–1563, 2009.

[53] Mikalai Krapivin, Maurizio Marchese, and Fabio Casati. Exploring and understanding citation-based scientific metrics. *Advances in Complex Systems (ACS) Journal*, 2010.

[54] A. N. Langville and C. D. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.

[55] Amy N. Langville and Carl D. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1:2004, 2004.

[56] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(5):67–71, 1999.

[57] Fei Liu, Feifan Liu, and Yang Liu. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 181–184, Dec. 2008.

[58] Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *The 2009 Annual Conference of the North American Chapter of the ACL*, pages 620–628, June 2009.

[59] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(5):22–31, 1968.

[60] Del Corso G. M., Gull A., and Romani F. The anatomy of a large-scale hypertextual web search engine. *Internet Mathematics*, 2(3):251–273, August 2005.

[61] A. McCallum. Rexa scholar autonomous digital library. http://rexa.info/.

[62] Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML*, pages 591–598, 2000.

[63] O. Medelyan and I.H. Witten. Thesaurus based automatic keyphrase indexing. In *JCDL*, 2006.

[64] H. Moed. *Citation Analysis in Research Evaluation.* Springer, 2005.

[65] Thomas Morton. *Using Semantic Relations to Improve Information Retrieval.* PhD thesis, University of Pennsylvania, 2005.

[66] Thuy Dung Nguyen and Min yen Kan. Keyphrase extraction in scientific publications. In *In Proc. of International Conference on Asian Digital Libraries ICADL T07*, pages 317–326. Springer, 2007.

[67] Joakim Nivre, Johan Hall, and Jens Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *In Proc. of LREC-2006*, pages 2216–2219, 2006.

[68] E. Osuna, R. Freund, and F. Girosi. Support vector machines: Training and applications. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR T97)*, page 130, Washington, DC, USA, 1997. IEEE Computer Society.

[69] F. Peng and A. McCallum. Accurate information extraction from research papers using conditional random fields. In *Proceedings of Human Language Technology Conference*, 2004.

[70] Fuchun Peng and Andrew McCallum. Information extraction from research papers using conditional random fields. *Inf. Process. Manage.*, 42(4):963–979, 2006.

[71] M. F. Porter. An algorithm for suffix stripping. In *Program*, pages 130–137, 1980.

[72] D. Radev, S. Blair-Goldensohn, and Z. Zhang. Experiments in single and multi-document summarization using mead. In *In Proceedings of The First Document Understanding Conference*, 2001.

[73] A. Ratnaparkhi. A maximum entropy part of speech tagger. In *Proceeding of the Conference on Empirical Methods in Natural Language Processing*. University of Pennsylvania, 1996.

[74] A. Ratnaparkhi. A simple introduction to maximum entropy models for natural language processing. Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania, aug 1997.

[75] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[76] Sam Scott and Stan Matwin. Feature engineering for text classification. In *ICML*, pages 379–388, 1999.

[77] R[oberto] Scrinzi. Sviluppo di un tool per l'esplorazione semi-automatica all'interno del servizio di google scholar. Master's thesis, University of Trento, Italy, 2008.

[78] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[79] Nicola Segata. FaLKM-lib v1.0: a Library for Fast Local Kernel Machines, 2009.

[80] Nicola Segata and Enrico Blanzieri. Fast Local Support Vector Machines for Large Datasets. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition: 6th International Conference (MLDM 09)*, number 5632 in Lecture Notes in Artificial Intelligence, pages 295–310, Leipzig, Germany, 2009. Springer.

[81] Nicola Segata and Enrico Blanzieri. Fast and Scalable Local Kernel Machines. Technical report, University of Trento, Trento, Italy, 2009. Submitted to a journal.

[82] David Sheskin. *Handbook of parametric and nonparametric statistical procedures.* CRC Press, Canada, 2004.

[83] Andrew E. Smith. Automatic extraction of semantic networks from text using leximancer. In *HLT-NAACL*, 2003.

[84] Springer. Cite you like. http://www.citeulike.org/.

[85] Springer. Springer digital library. http://www.springer.com.

[86] A. Stolke. Srilm an extensible language modeling toolkit. In *In Proceedings of ICSLP*, page 901904, 2002.

[87] Y. Sun and C. L. Giles. Popularity weighted ranking for academic digital libraries. In *29th European Conference on IR Research (ECIR'2007)*, pages 605–612, Rome, Italy, 2007.

[88] Yang Sun, Huajing Li, Isaac G. Councill, Jian Huang, Wang-Chien Lee, and C. Lee Giles. Personalized ranking for digital libraries based on log analysis. In *10th ACM International Workshop on Web Information and Data Management (WIDM2008)*, pages 133–140, New York, USA, 2008. ACM Press.

[89] JUNG team. Java universal network/graph framework. http://jung.sourceforge.net/.

[90] Mendeley Team. Mendeley. http://www.mendeley.com/.

[91] Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[92] Stephen Tomlinson. Compared the snowball stemmers with the hummingbird lexical stemming (lemmatization) system. In *CLEF*, 2003.

[93] P. Tourney. Coherent keyphrase extraction via web mining. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 434–439, 2003.

[94] Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 63–70, 2000.

[95] P. Turney. Learning to Extract Keyphrases from Text. Technical report, NRC/ERB-1057, feb 1999.

[96] A. F. J. Van Raan. Scientometrics: State-of-the-art. *SCIENTOMETRICS*, 38(1):205–218, 1997.

[97] Vladimir N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[98] X. Wan, J. Yang, and J. Xiao. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *In Proceedings of ACL*, page 552559, 2007.

[99] J. Wang and H. Peng. Keyphrases extraction from web document by the least squares support vector machine. In *IEE/WIC/ACM International Conference on Web Intelligence)*, 2005.

[100] Marcus Watson, Andrew Smith, and Scott Watter. Leximancer concept mapping of patient case studies. In *KES (3)*, pages 1232–1238, 2005.

[101] Christopher K. I. Williams and David Barber. Bayesian classification with gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1342–1351, 1998.

[102] I. H. Witten, G. W. Paynte, E. Frank, C. Gutwin, and C. G. Nevill-Manning. Kea: Practical automatic keyphrase extraction. In *Proceedings of DL '99*, pages 254–256, 1999. http://www.nzdl.org/Kea/.

[103] Ian .H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2005.

[104] Fuyong Y., Chunxia Y., and Jian L. Improvement of pagerank for focused crawler. In *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, pages 797–802, 2007.

[105] David Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *COLING*, pages 454–460, 1992.

[106] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, pages 189–196, 1995.