A Flexible System for a Comprehensive Analysis of Bibliographical Data

Sahar Vahdati, Andreas Behrend, Gereon Schüller and Rainer Manthey
Institute of Computer Science III, University of Bonn, Römerstr. 164, D-53117 Bonn, Germany

Keywords: Bibliographic Database, Digital Library, Citation Analysis.

Abstract: Scientific literature has become easily accessible by now but a comprehensive analysis of the contents and interrelationships between research papers is often missing. Therefore, a time consuming bibliographical analysis is usually performed by scientists before they can really start their research. This manual process includes the identification of the most important research trends, major papers, auspicious approaches, established conference series as well as the search for most active groups for a specific research topic. In addition, scientists have to collect related academic literature for avoiding reinvention of already published results. Although a large number of literature management systems have been developed in order to support researchers in these tasks, the offered analysis of bibliographical data is still quite limited. In this paper, we identify some of the missing analysis features and show how they could be implemented using data about author affiliations, reference relations and additional metadata, automatically generated from a set of research articles. The resulting prototypical implementation indicates the way towards the design of a general and extendible bibliographic analysis system.

1 INTRODUCTION

Bibliographic analysis is a very important aspect in scientific work but usually a time-consuming process. There are a lot of scientific literature management systems available by now supporting researchers in this process to a certain extent (e.g, ACM Digital Library, DBLP, CiteSeerX, SpringerLink, Google Scholar or Microsoft Academic Search, but a comprehensive analysis of bibliographical data is still missing. Most of the systems focus on a perfect keyword search or the computation of certain impact factors which do not help researchers in getting a good overview about a certain research field (Bakkalbasi et al., 2006). In contrast, researchers are often interested in getting a quick overview about a scientific topic rather than a list of papers ordered according to their citation ranks.

In particular researchers would like to find out: "What are the most active conferences, groups, institutes in the research area of my interest?" "Who was the most influential author and who is it now?" "What papers are the most important ones for a given research topic?" "Are there different schools or research directions and which direction has turned out to be a dead end, retrospectively?" "Which research topic is currently en vogue establishing a kind of trend?" "What conferences should I choose if I want to submit a paper about a certain topic?" . The basic technique needed to get such information is the exploration of related work which starts with navigating the citations provided in a set of research articles. To this end, bibliographical databases, digital libraries and search engines could be used which offer a citation analysis of publications (Lister and Box, 2008). Such systems offer different search methods but usually leave researchers alone with long lists of papers possibly matching the provided search criteria. For example, Google Scholar is good in finding papers which are relevant to a given list of keywords. In addition, the number of citations by other papers are displayed for each document which leads to another list of papers.

Despite of these fine-tuned results, the questions stated above cannot be answered that way and the user needs to perform various similar searches before he gets an overview about the research field he is interested in. Microsoft offers another search engine which additionally provides a graphical visualization for exploring the citation dependencies stepwise in one direction. The system, however, cannot provide a general overview over certain research areas and allows for investigating the interrelations of one selected paper, only. In this paper, we show how missing analysis features could be implemented using data about author affiliations, reference relations and additional metadata automatically generated from re-

Vahdati S., Behrend A., Schüller G. and Manthey R..
A Flexible System for a Comprehensive Analysis of Bibliographical Data.
DOI: 10.5220/0004799201430151
In *Proceedings of the 10th International Conference on Web Information Systems and Technologies* (WEBIST-2014), pages 143-151
ISBN: 978-989-758-023-9

search articles. All paper related data is stored in a relational database and SQL queries are employed allowing for a flexible and comprehensive article analysis. In contrast to other literature services, our approach supports

- the analysis of the complete citation graph allowing for detecting research schools, dead-end research directions or methodological differences
- the distinction between different types of citations which allows for a better understanding of relationships among papers
- the automatic detection of key papers as well as key conferences respectively journals for a given research school
- the analysis of author-related information (e.g. affiliations).

At the end of this paper, we provide an evaluation of our prototypical implementation using a sample data set which contains documents from the research field deductive databases over the last 37 years. We believe that our system indicates the way towards the design of a general bibliographic analysis tool.

2 ARCHITECTURE

The architecture of our system consists of three main components: Base Data Collection, Bibliographical Database System and User Interface. The first level is responsible for collecting, editing, storing and indexing research papers. To this end, paper-related information as well as a digital copy of each paper is stored in the repository after passing a digitization process. Data acquisition is one of the crucial steps as it determines the quality of the analysis results later. We have employed metadata extraction tools to our sample document set in order to obtain detailed information about title, authors, affiliation, publisher, references and covered research topics. Additional metadata could be extracted from other external systems such as digital libraries (DL), bibliographical catalogues or literature databases. In fact, we plan to develop our system towards a metasearch engine which integrates the analysis features of other DL. In order to enhance the paper classification process, we have annotated papers by information about the paper contents such as keywords as well as data model or programming language used in the presentation. Most of this information has been automatically extracted but manually checked in order to have a very clean data set. This additional data allowed for further refining the search for relevant articles with respect to a given

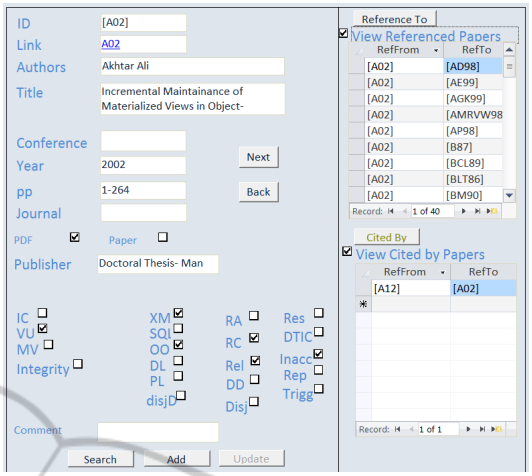

Figure 1: The interface of our prototype for advanced content search.

topic. Finally, the user interface provided is able to visualize the results using different types of graphs such as intensity maps, line or network graphs which transfer the analysis to a human readable format.

3 SAMPLE DATA COLLECTION

We have used a specific sample set of papers for showing the expressiveness of our bibliographic analysis tool. To this end, we used a collection of papers which are concerned with three classical deductive database topics, namely materialized view maintenance, integrity checking and view updating.

3.1 Conceptual Design

Although the analysis of paper-related information such as citations and topics covered were the main focus of our work, we additionally wanted to support the detection of the most influential conferences and journals for particular research fields. Therefore, we needed to carefully distinguish between the various forms of scientific publications such as proceedings, books or journals. Furthermore, we had to extract some additional metadata from the papers such as citations, keywords or author affiliations which are usually not provided by similar systems for scientific literature management. Metadata extraction can be automatically done using extraction tools such as TeamBeam. In our sample data set we could finally use more than 30 attributes for characterizing a paper including authorship, affiliations, citations, keywords and potentially covered research topics. Aside from the main paper table, further tables are employed

for storing the metadata determined such as reference lists or the information about authors (in order to handle multiple authorship). Based on that data set, we could provide a much deeper paper analysis using SQL queries than offered in similar bibliographic systems available today. Our data collections contains 1103 papers published from 1975 till 2012. The reference list basically forms the edge relation of the underlying citation graph and additionally contains a Boolean attribute indicating the "importance" of the citation relationship.

In fact, citations may play a different role in a paper. Some of them are used for indicating that the authors know other research work (or traditions) which is related to their own paper. Other citations are used to refer to scientific material (methods, approaches) which is really necessary in order to understand the presented approach of the paper under investigation. We consider the latter as much more important as they indicate research directions and even real scientific cooperations. During the process of evaluating the reference type we used the value "more important" as default which is automatically replaced if the citation appears only once and the paper keywords are considerably different. More details about the automatic detection of reference types is given in Section 4.2. Another important category of citations are self-references which are consequently omitted within the entire analysis process.

Many of the popular online services suffer from unreliable data. In order to avoid similar problems and to provide more reliable analysis results, we have applied a data cleaning and correction step (a step towards semi-automated curation). To this end, the stored information about publishers, page numbers and authorship has been verified using different online sources. Another problem we had to solve before analyzing the data was the generation of unique paper IDs. The primary ID of each paper is automatically generated from the authors' names and the publication year. In case that this method would lead to duplicates, the naming method is stepwise refined following the recommendations from (Han and et al., 2004) until all conflicts have been resolved. Although this represents no general solution, we could already obtain a clean sample set this way.

3.2 Content Description

In the following two sections we show how complex bibliographical questions can already be answered using the automatically collected paper data from above. For example, in Section 4 we show how the citation graph could be used to identify different research schools for a particular research topic. This kind of paper analysis, however, could be even improved if more information about the paper's contents would be present. We used additional classification data for each paper in order to provide better query results to the user. To this end, we employed the classification data from the digital library of ACM.

In order to provide even more classification details we determined characteristic keywords within the paper related to the employed data models, programming languages, and algorithms. This classification process was done using a pre-defined set of keywords (e.g. Relational Algebra (RA), Update Propagation (UP), Fixpoint (FP), SLDNF, Deductive Rules (DL), Transformation-based (TB),...) and text-mining tools provided by the underlying database system. The resulting data provide a more precise content description such that researchers can perform more advanced paper searches while reducing the number of false positives in the list of returned answers.

For example, in our user interface you can search for papers concerning integrity checking where the authors use the relational data model as well as the relational algebra in their presentation (see Figure 1). This way, the article "[Bro00]" with the title 'A general treatment of dynamic integrity constraints' and the article "[Dec02]" with the title 'Translating Advanced Integrity Checking Technology to SQL' could be easily identified as very related papers despite of the very different values in the ACM classification system (G.2.0 and H.2.1 vs. H.2.7 and H.2.3).

4 CITATION ANALYSIS

Despite of the fact that many scientific literature management tools provide a citation number for each paper, a comprehensive citation analysis is usually missing. Having the entire citation graph at hand, various interesting conclusions about the importance of paper ideas, authors, research schools and the success of research directions can be drawn.

4.1 Building a Citation Graph

Reference Depths: Citations form a relation between papers leading to a directed dependency graph. Papers may be connected via several relationships (paths) with different distance values. We call a relationship between papers "non-direct references" if a paper does not refer to a paper in its own reference list but indirectly via one or several referenced papers (i.e., the distance is larger than 1).

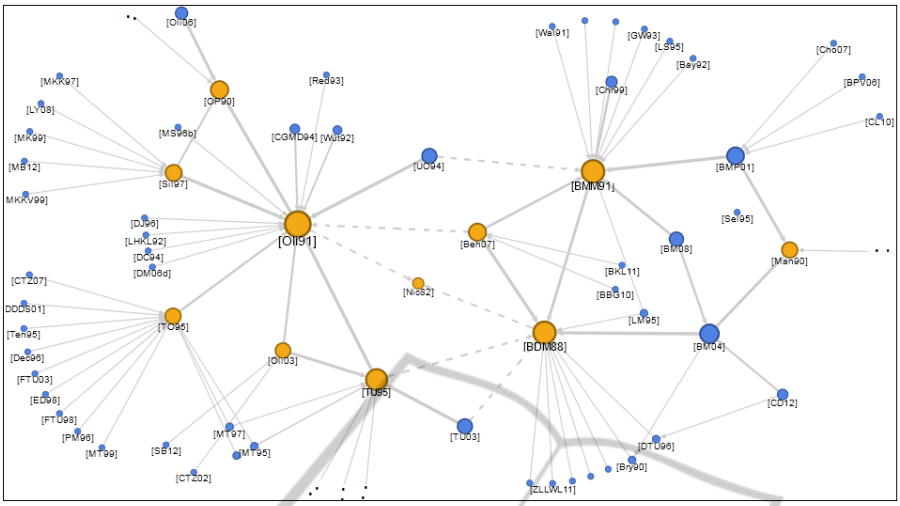

Figure 2: Citation relationships in form of a network graph view (Google Fusion Tables).

Self-References: The self-references may be justified if an author refines his previous work. However, self-references are often considered to be not that valuable. We developed a filter query for avoiding the analysis of such references and stored the remaining references using the materialized view "NewReferences".

Ancestors: All direct references which have been extracted from the documents have been added to the References table with distance number 1. These references can be used to determine indirect relations between papers that can be modelled as ancestors and descendants with different distances. For the subsequent analysis queries, it would be desirable to have the ancestors and descendants relation in materialized form. The paper relationship graph, however, is highly connected and a paper may be related to another paper by various paths of different lengths. Therefore, we created a new table named "NewAncestors" which contained all paths with the smallest distance. The corresponding paths have been determined using a fixpoint iteration starting from all edges in table "NewReferences" with distance 1.

The respective query is iteratively applied as long as there are new insertions. With our test data, the iteration loop ended after 8 iterations, which means that the maximal distance between two related papers was 8. The total number of direct relations between two papers in our set is 4552, while there are more than 35000 indirect relations. These relations between papers form a graph in which papers are the nodes and the distance is the weight of each edge between two related papers. A small fraction of our reference relation can be seen in Figure 2. The yellow nodes indicate that this paper is referenced by another one. The more papers refer to a yellow node, the bigger (and more important) the node becomes. If a paper is not referenced at all, the node gets the color blue.

4.2 Analyzing Citations

The stored direct references together with the computed indirect ones allow for standard computations such as "Who is the most referenced author for a particular topic?" or "What paper about integrity checking has the highest number of references?". Figure 3 shows the list of papers with the highest number of citations, including the authors, conferences, and publication years. Here we have already used the total number of references including the computed indirect ones. Our data could also be used to find the most active author who is still publishing in the respective research fields. For example, the most active researcher over the last two years 2011 and 2012 in the field of integrity checking was Hendrik Decker according to the data from our collection.

The citation graph, however, even allows for a deeper paper analysis. Suppose you want to know whether there are different schools (or approaches) for checking integrity violations efficiently. In Figure 2, clearly two subgraphs can be identified which basically represent two research schools for this particular research area. As mentioned above, we have used two types of references in order to consider the different importance for the respective paper. Solid lines are used to indicate important citations (strong relationships) which are necessary to understand the presented approach while dotted lines (weak relationships) are employed for citations of minor importance. The paper "[TU95]", therefore, belongs to the subgraph centered by "[Oli91]" and not to the other one with the center "[BDM88]" because of the dot-

RefTo	NumRef	Authors	Conference	Journal	Year
[Nic82]	129	Nicolas		AI	1982
[BLT86]	116	Blakeley, Larson	SIGMOD		1986
[GMS93]	103	Gupta, Mumick,	SIGMOD		1993
[BS81a]	97	Bancilhon, Spyr		TODS	1981
[ZGHW95]	88	Zhuge, Garcia, H	SIGMOD		1995
[GM95]	87	Gupta, Mumick		DE	1995
[DB82a]	81	Dayal, Bernstein		TODS	1982
[GL95]	78	Griffin, Libkin	SIGMOD		1995

Figure 3: Papers sorted according to the number of references.

ted connection to the latter. For determining the two connected components of the depicted citation graph, current database systems already provide methods (e.g., in Oracle 11gR1 as part of the *Oracle Spatial and Graph* facility) which have to be applied to strong relationships, only.

In order to distinguish weak and strong relationships, the references have been weighted using a bonus system. To this end, we employed the following formula $bonus(c,k,o) := 10*c+20*k+20*o$, where c is the total number of common coauthors, k is the number of common keywords and o is the number of reference occurrences within the referencing paper. The first subterm is based on the assumption that common coauthors indicate related paper contents. The factor 10, however, leads to a relatively small impact on the total bonus number because coauthorship is only a very rough indicator. The main criterion is the number of common keywords used in the two papers as it provides the best indicator for the paper topic. We employed again a list of pre-defined keywords like UP, FP, SLDNF, RA, Deductive Rules and Transformation-based (cf. Section 3.2) each of them contributing to the total bonus by the weighting factor 20.

This rough estimation worked well for our test scenario but could be further improved by more sophisticated IR approaches (like the tf-idf method (Wu and et al., 2008)) for determining the importance of a keyword within a document. The third subterm evaluates the number of occurrences of a certain reference. If authors are referenced several times, we may assume that the relationship to this previous work is high. We simply added all bonus points and indicated a reference relation as weak if the resulting weight is below 300. In addition, weights above 700 are indicated by even thicker solid lines in order to show very close relationships between two papers. On that basis, different research schools could be easily identified as shown in Figure 2. The used interval [300,700] worked almost perfectly for our sample set but an extended evaluation is still to be done. This way, not only different research schools could be identified but also the importance of certain approaches may be estimated. For example, the subgraph on the left-hand side contains considerably more nodes than the one on the right-hand side, making it potentially more influential (active) than the other one. Furthermore, we could use our filter to omit any self-references which would allow to estimate the number of people in the respective research group working on that particular subject. Despite of these already promising results, the approach can be certainly further refined by applying more sophisticated methods for estimating paper relationship such as the vector space model from IR.

5 RESULTS

In the following, we provide a brief overview of our analysis results. While the following presentation remains rather sketchy, it already shows the general possibilities of a comprehensive paper analysis.

5.1 Publishers and Venues

The documents in our collection had been published in 314 different conferences and journals. There are 199 conferences related to the three research areas and five top active conferences could be determined according to the number of references to papers published there. The conferences in which most of the papers were published over the years are VLDB, SIGMOD, ICDE, PODS, and DEXA. In particular, around 1/5 of all papers from our collection appeared in these conferences.

In Figure 4 the line graph of the number of papers published in the top 5 conferences is shown. During the mid 90s there was a peak in the interest for the respective three topics. Then, between 2003 and 2007 there was a renewed interest in these topics with publications at ICDE, VLDB and DEXA.

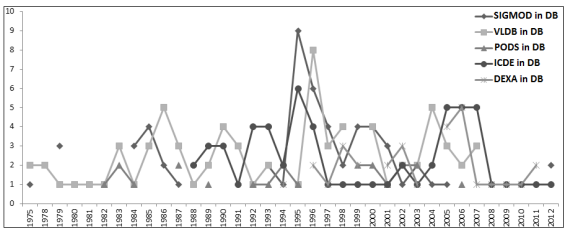

Figure 4: Number of papers from our collection published in the top 5 conferences.

Afterwards, the number of published papers at these conference decreased significantly, showing that the research topics under investigation have lost their dynamics. This analysis is equivalent to the observations made by (Mayol and Teniente, 1999). During the last two years 2011 and 2012, the confer-

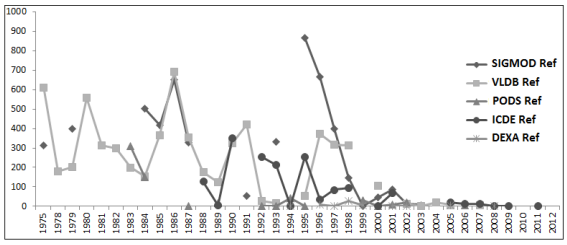

Figure 5: No. of papers referencing conferences per year.

ence series KES (Conference on Knowledge-Based and Intelligent Information & Engineering Systems) became more important for publishing results from our three sample topics. So, if a researcher is interested in the topics integrity checking, materialized view maintenance or view updating, he or she ought to consult older papers from the identified top 5 conferences in order to get an overview about the proposed research approaches. For publishing new results, the conferences KES, DEXA and ICDE appear to be more appropriate than VLDB or PODS.

5.2 Active Conferences

Let us now consider the effectiveness role of conferences for the topics under investigation. To this end, we considered the top 10 conferences and determined how often papers in these conferences have been referenced by papers from our collection. Two queries are employed over the table Papers and NewAncestors which comprise direct as well as indirect references while avoiding the counting of self-references. The first query returns the number of references from different papers to papers published by these conferences and journals. The second one illustrates the number of papers published by a certain conference with respect to the year of publication. The number of references to different conferences is counted in the first query which performs an inner join of NewAncestors on the table Paper. Afterwards the result is aggregated in order to sum up the number of references for a particular conference with respect to each year. The result was that the most active conferences are also the most referenced ones, namely VLDB, SIGMOD, ICDE, PODS, and DEXA.

In Figure 5 the number of references to the 5 most active conferences per year is shown. The oldest and still active conference in our interest research area is VLDB. For example, papers from the VLDB 1990 were cited by 100 different papers from our collection. The conferences which have been mostly referenced by papers from our collection are SIGMOD and VLDB indicating that most of the underlying methodological framework was presented at these conferences. Another interesting aspect is the aging of citations. More recent conferences receive very few citations indicating that new publications on our preselected topics do not use results published in these conferences anymore but rather refer to classical standard papers from the past. Note that there are other ways of analyzing co-authorship which have been proposed, e.g., (Nascimento et al., 2003; Smeaton and et al., 2002), which could be integrated in our system, too.

5.3 Active Countries

Generally, a paper arises from a work or project which has been done in a research institute or university. The origin of a paper is usually indicated in its metadata. The information about the organizations in which the papers have been produced allows for determining the amount of research activity of that organization or country in a certain way. Having such information stored in the database enables us to identify the most active countries for a particular research area.

Another interesting result is the identification of topic movements over countries. By publishing the original papers in a conference or journal, other researchers become aware about that new topic and start working on that. Consequently, a topic may become a trend in a country's research for a particular time interval which is reflected by the intensity of the respective publication numbers.

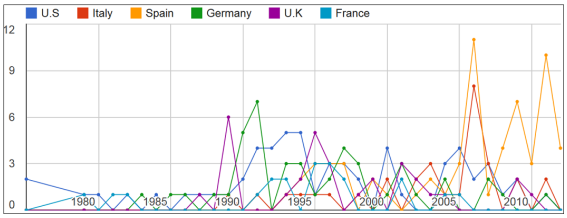

Figure 6: Topic movement for integrity checking over time and country

Figure 6 illustrates the top 6 active countries which have published the highest number of papers over years. In 1975, the first paper on integrity checking appeared from the USA. In parallel, researchers in France were active in this research field and published papers in the same year. Afterwards, other countries like Germany and Spain became interested in this topic and are even still working on them. Another interesting point is that Spain represents the most active country nowadays for the field of Integrity Checking, although it was the last country in the list starting to publish paper in this research area. It is also visible that this topic was an active research field during the 90s but it regained interest recently.

5.4 Citation Genealogy

The impact of scientific publications (Aksnes, 2005) is often estimated by the number of direct citations they receive. On the other hand, indirect references also indicate the relationships of two papers. Suppose a user wants to know whether a paper indirectly refers to a very important work in a certain research field. A possible way to answer this question could be to determine all indirectly referenced papers and to select the one with the highest impact.

In fact, public bibliographic services such as the academic research system by Microsoft allow for exploring indirect citation references by hand. However, this may soon turn into a very tedious task and a query about the most influential paper which is indirectly cited can hardly be answered that way. Indirect citation references may also show the footprints of a publication topic through the evolution of its respective research area. The same technique can be used for analyzing the cited_by relationship. Suppose you want to know what is the most important paper (directly or indirectly) citing the publication under investigation. Again, a query like this could be easily implemented using the ancestor relationship among paper citations

```
SELECT NA.RefFrom, NA.RefTo, max(NA.#DC)
FROM   NewAncestors AS NA
GROUP BY NA.RefTo;
```

where the attribute #DC refers to the number of direct citations. This type of query is currently not supported by any of the public bibliographic services available so far. In fact, most of the systems just focus on direct relationships neglecting the influence of transitive connections. The exploration of indirect references could be used to refine the impact values of a paper but it is not sufficient just to consider the total number of direct and indirect ones. The reason is that once a paper is cited by a key paper with many direct references, this paper inherits all this references which may lead to unjustified high impact scores.

In Fig. 7 the ranking of papers on the topic 'materialized view maintenance' is depicted according to their total number of direct as well as indirect references. We have highlighted the papers we consider to be key research papers in the respective research area. Obviously, both measurements provide valuable information about possible key papers. The number of indirect references in particular indicates how active a research direction has been followed and whether the ideas of a highly cited paper have really spread. The computation of a general impact factor, however, is a difficult task and various other paper related values should be additionally considered. At least, in our

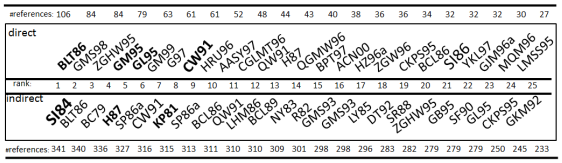

Figure 7: Citation Genealogy for Selected Papers.

particular case Data mining techniques could be used in order to discover a formula for computing impact factors based on direct and indirect references (Chen, 2006). The resulting formula should also take an aging factor into account. For example, the length of the time interval in which direct references occurred may indicate the relevance of a paper and its addressed research topic.

The ancestor relationship could also be used to form a kind of scientific genealogy for any tracked publication. In this way one can see all the work that directly or indirectly influenced a given paper. In Figure 8 we present a sample case in which the ancestor relationship for four different papers is depicted. To this end, we have chosen four papers with a very high number of (in)direct citations (cited_by as well as cited_to references). In the sample study, we have chosen four papers with more than 100 references each, as this border allowed to determine almost every key research paper in our sample collection. It is interesting to notice that the ancestors and children for the rightmost key paper (root) showed almost no interconnections with ancestors and children from another root paper.

This disconnection indicates two very different research fields (integrity checking for the first three papers and query evaluation for the one on the right) as the authors did not refer to the other branch. Thus, different research schools or research fields could be identified even though no keywords nor specific terms were known in advance. The resulting tool could be used to identify unknown research schools and to automatically determine a representative collection of keywords for them (just like reverse engineering). We have called the respective graph *Citation Genealogy*. It has been prepared using the TouchGraph[1] software. In principle, four different cases may occur depending on possible overlaps between the ancestors and/or the children subgraphs. In case that we have connections between children and ancestors of two different key papers (see the two key papers from left-to-right in Fig. 8) we may assume that we are dealing with just one research field from which more than one key paper originated. The case that we have no connections between ancestors nor children is a strong indication

[1] http://www.touchgraph.com/navigator

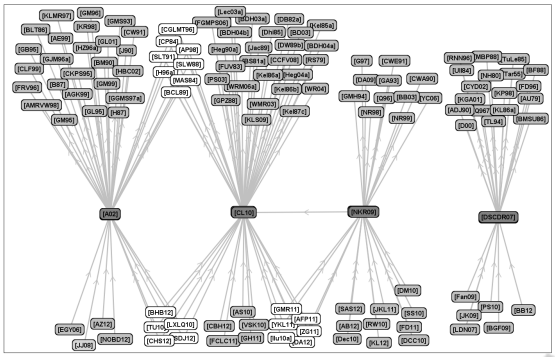

Figure 8: Citation Genealogy for Selected Papers.

for two different research schools. Another case is that we have some connections between the children and even the key papers but none between the ancestor papers (as it is the case for the two key papers in the middle of Fig. 8). This constellation may indicate that two different research directions have developed over time from one common research field. Note that this kind of analysis could even be refined by omitting weak reference types as proposed in Section 4. In this way, however, already identified keywords for different research topics are implicitly used again.

6 RELATED WORK

For decades, citation counts and impact factor scores ,e.g., the h-index or Eigenfactor have been the primary currency in the entire business of publication (Rahm, 2005). Of course, the corresponding metrics could be easily incorporated into our analysis queries. The problem of evaluating the value of paper references has been tackled in various publications ,e.g., (Lacasta and et al., 2013) and (Newman, 2001). In particular, these authors try to find criteria which help to discover very similar scientific work or even duplicated results. An approach for automatically extracting topic classifiers from a paper text has been proposed in (Klampfl, 2013) that could be used in our system for refining the distinction of weak and strong references. The problem of reference reconciliation and methodes for detecting similar references and improving keyword search has been proposed in (Dong and et al., 2005), (Tejada and et al., 2002) and (Harzing, 2013) which allow to incorporate similarity measures. In (Falagas and et al., 2008) and (Jacso, 2005) comparisons between four popular online bibliographical databases has been done using specific keyword searches and analyzing the utility of the retrieved information. The problem of search engines returning lists of results which needs additional manual mining is also addressed with an enriching approach in (Khazaei, 2012). All these approaches, however, solve particular search problems only. A system that provides a comprehensive and extensible citation analysis, however, is still missing. We decided against the usage of a graph database because of the extensive multi-user support we need in our system and the lack of support for large scale application in graph databases (Vicknair and et al., 2010).

7 EVALUATION

For evaluating our system two aspects have to considered: performance and accuracy of the results. With respect to performance, the only critical part is the determination of the transitive closure of the reference relationship. As the resulting graph is materialized, all queries using this information (see samples in Subsections 5.3 and 5.4) could be executed in less than a second using a standard PC. This is not surprising because of the relatively small amount of around 35.000 paths for the 1103 papers and applicable index structures which would scale well for even bigger collections such as DBLP. The determination and materialization of the transitive closure took less than 5 Minutes by exploring link connections of at most 8 steps. Another data set from DBLP has been tested which contained more than 2300 papers with only 6000 [2] indirect references and detected 8 again as the maximal distance of two related papers. Even if the maximal distance number were considerably higher in different research fields, the computation of the transitive closure remains to be strongly limited and we can assume that the total number of indirect relations may considerably stay below $|collection| \times |collection|$.

Furthermore, the reference list of already processed papers hardly changes anymore such that the transitive closure of them really has to be computed only once. New paths within the transitive closure induced by new papers added to the collection can be efficiently determined using incremental maintenance methods as proposed, e.g., in (Behrend, 2011). In addition, the determination of the transitive closure could be the application of recursive views as supported by many database systems like PostgreSQL by now. In contrast to performance issues, the evaluation of the system's accuracy is much more problematic due to the specific data collection we currently work on. Having preselected papers from the research

[2]The low number is explained by the fact that this was a collection with randomly chosen papers. Thus, the number of related papers was quite small in contrast to our topic specific collection.

fields materialized view maintenance, integrity checking and view updating, we can hardly evaluate the accuracy of finding other research fields. For this we would need to evaluate a broader collection of papers. The detection of research schools, however, was highly accurate for our collection but an evaluation of our rules for distinguishing weak and strong reference types is still to be done. In particular, the trade-off between the detection of as many different schools as possible while avoiding the generation of false positives has still to be investigated. Nevertheless, the results we have achieved so far already indicate that the automatic detection of research schools and/or research fields is possible in a feasible way.

8 CONCLUSION

We have presented a system for bibliographical analysis which is desired by many researchers. Our analysis results indicate that such a system is able to answer the bibliographical questions which researchers might encounter while searching for specific papers in a particular area. Our tool can become a powerful and influential information system supporting researchers in other scientific communities to collect, manage, access the bibliographical and citation analysis of documents. The specificity of our system proposal is the consequent application of a database system with a full-fledged query language at hand. In this way, a flexible analysis tool could be programmed which can be easily extended by new paper related measurements. While the analysis part as well as data acquisition is constantly extended, we additionally plan to develop a web-based interface for supporting a flexible querying in the future.

REFERENCES

Aksnes, W. (2005). *Citation and Their Use as Indicators in Science Policy. Study of Validity and Applicability Issues with a Particular Focus on Highly Cited Papers*. PhD thesis, University of Twente.

Bakkalbasi, N., Bauer, K., Glover, J., and Wang, L. (2006). Three options for citation tracking: Google scholar, scopus and web of science. *BDL*, 3(1).

Behrend (2011). A uniform fixpoint approach to the implementation of inference methods for deductive databases. In *LNAI*, pages 1–16.

Chen, C. (2006). Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST*, 57(3):359–377.

Dong and et al. (2005). Reference reconciliation in complex information spaces. *SIGMOD Rec.*, pages 85–96.

Falagas and et al. (2008). Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *FASEB.*, 22(2):338–342.

Han, H. and et al. (2004). Two supervised learning approaches for name disambiguation in author citations. In *JCDL*, pages 296–305.

Harzing (2013). A preliminary test of google scholar as a source for citation data: a longitudinal study of nobel prize winners. *Scientometrics.*, 94(3):1057–1075.

Jacso (2005). As we may search-comparison of major features of the web of science, scopus, and google scholar citation-based and citation-enhanced databases. *CURRENT SCIENCE-BANGALORE.*, 89(9):1537–1547.

Khazaei, H. (2012). Metadata visualization of scholarly search results: supporting exploration and discovery. In *i-KNOW*, pages 1–8.

Klampfl, K. (2013). An unsupervised machine learning approach to body text and table of contents extraction from digital scientific articles. In *TPDL*, pages 144–155.

Lacasta and et al. (2013). Design and evaluation of a semantic enrichment process for bibliographic databases. *DKE*, 88(1):94–107.

Lister, R. and Box, I. (2008). A citation analysis of the sigcse 2007 proceedings. In *SIGCSE*, pages 476–480.

Mayol, E. and Teniente, E. (1999). A survey of current methods for integrity constraint maintenance and view updating. In *ER Workshops*, pages 62–73.

Nascimento, M. A., Sander, J., and Pound, J. (2003). Analysis of sigmod's co-authorship graph. *SIGMOD Rec.*, 32(3):8–10.

Newman (2001). The structure of scientific collaboration networks. In *PNAS*, pages 401–409.

Rahm, T. (2005). Citation analysis of database publications. *SIGMOD Rec.*, pages 48–53.

Smeaton, A. F. and et al. (2002). Analysis of papers from twenty-five years of sigir conferences: What have we been doing for the last quarter of a century. *SIGIR Forum*, 36:39–43.

Tejada and et al. (2002). Learning domain-independent string transformation weights for high accuracy object identification. *SIGMOD Rec.*, pages 350–359.

Vicknair and et al. (2010). A comparison of a graph database and a relational database: a data provenance perspective. *ACMSE.*, pages 1–6.

Wu and et al. (2008). Interpreting tf-idf term weights as making relevance decisions. 26(3):1–37.