Detection of Bursty and Emerging Trends towards Identification of Researchers at the Early Stage of Trends

Sheron L. Decker, Boanerges Aleman-Meza, Delroy Cameron & I. Budak Arpinar

Computer Science Department, University of Georgia
Athens, GA 30602-7404, USA
{decker, boanerg, cameron, budak}@cs.uga.edu

Abstract. Detection of trends is important in a variety of areas. Scientific research is no exception. While several methods have been proposed for trend detection, we argue that there is value on using semantics-based techniques. In particular, we demonstrate the value of using a taxonomy of topics together with data extraction to create a dataset relating publications to topics in the taxonomy. Compared to other approaches, our method does not have to process the content of the publications. Instead, it uses metadata elements such as keywords and abstracts. Using such dataset, we show that a semantics-based approach can detect "bursty" and "emerging" research topic trends. Additionally, our method identifies researchers involved at the early stage of trends. We use known lists of recognized and prolific authors to validate that many of the researchers identified at the early stage of trends have indeed been recognized for their contributions on important research trends.

Keywords: Trend Detection, Emerging Trends, Bursty Trends, Taxonomy of Computer Science Topics, DBLP, Data Extraction

1 Introduction

One way to keep up with the landscape of research in a field of study is to stay informed with the trends that are occurring in the area. Having knowledge of past, current and emerging trends is quite valuable. For example, a scientist might want to do research in an area that has not been touched on heavily or even in the sense of a business-person trying to evaluate the risks of investing in a new business. Trend detection has already been applied with the use of text documents, blogs and emails [8, 10, 11, 15]. The detection of trends could be very important for funding agencies such as NSF in order to determine or justify whether projects in new areas of research are to be funded. Automated approaches have been built for identifying funding agencies in acknowledgments section of papers [4]. From the stand point of identifying past trends, one could determine if there is any correlation to the amount of funding provided to a previous area of interest with respect to its success or impact. Identifying participants at the emerging stage of a trend is of importance because it will determine who were the influential people that aided or started the popularity of a given trend. For example, the ACM program recognizes and honors individuals for their achievements in the computer science and information technology fields. The identification

of researchers that are identified as "trend setters" could help in determining the individuals to consider for such awards. The goal of our work is to develop an approach that will detect two types of trends. The first are "bursty" trends, which have the characteristic of having one or more intense periods of activity. Second are "emerging" trends, which are characterized by having an increasing activity over a period of time but not necessarily with a "bursty" behavior.

Gruhl et al. studied detection of "bursts" in the blogsphere for cases where the total number of blog entries on a particular topic exceeded a formulated threshold [8]. They also examined whether these topics could be "de-spiked" to identify an underlying, probably unknown reason for the burst. We use these same ideas but with different approaches. First, we focus on identifying research topics using different data, namely metadata of publications such as keywords and abstracts. Second, we demonstrate that trends in research can also have "bursts," which we identify based on the total number of publications written on the topic using time intervals of days, months or years. Other work for trend detection in publication data has only managed to use years as the unit of time. Thirdly, we implement "de-spiking" on a research topic to identify other topics that might be the cause of the bursty behavior. This allows analysis to determine if other topics had any impact on the burst (if indeed there was a burst in the total number of publications on the topic). For example, how much of a contribution has the topic "PageRank" had towards the trend in the topic "Ranking"? We also focus on determining who were the authors that published on the research areas at the early stage of the trend. This approach builds upon the work of [8] where they adopted a simple set of predicates on topics that would allow them to associate particular blog posts appearing at different parts of a topic life cycle.

In other work [8, 11, 15], evaluation of bursts was accomplished using the construction of time graphs, whereas [10] took the approach of using a weighted automaton model. In the context of blogs, they have specific timestamps associated with them to identify when they were created in order to create time sequential graphs. Similarly, emails can be tracked based on arrival structure. In our work, we evaluate bursts in a research area using the time graph approach. In order to construct a graph for a research area we first have to create a dataset that relates papers to one or more research topics. Because DBLP (www.informatik.uni-trier.de/~ley/db/) is one of the largest websites that lists computer science bibliography, we decided to use it as data extraction source. We demonstrate how this type of dataset can be created with focused crawling and off-the-self techniques for term extraction (e.g., Yahoo! Term Extraction (developer.yahoo.com)). Extracting data relating topics to publications from DBLP is extremely challenging because DBLP does not contain information relating publications to research areas or topics. We developed methods that create such paper to topic relationships. A publication can then be explicitly related to one or more topics. One of our goals is in demonstrating how this is possible without having to process the content of documents, which in this case, are publications that exist in a variety of document formats (e.g., PDF, PostScript, and HTML). Similar datasets can be created for other research fields such as chemistry or biology. For the purposes of this paper, our approach is tested using a dataset that is focused on research areas of DB, Information Retrieval, Web and Semantic Web, AI and Data Mining. This dataset consists of 78K publications and 40K relationships connecting publications to topics in a taxonomy of Computer Science research areas.

The contributions of our work are two-fold. First, we describe a methodology for building a dataset that contains relationships from publications to topics in a taxonomy of topics. The benefits of this type of dataset is that the papers to topics relationships connect topics in the taxonomy to publications in an existing ontology of publications that was created using DBLP data. Second, we demonstrate a semantics-based approach for determining "bursty" and "emerging" research topic trends together with the capability of identifying researchers at the emerging stages of research areas.

2 Method for Building a Publications-to-Topics Dataset

2.1 Using DBLP Bibliography Data

In the work of Tho et al., the majority of their dataset of scientific publications was retrieved from websites of academic institutions [15]. We argue that better results are possible when using larger datasets. DBLP is an excellent site that lists bibliography data of more than 885,000 computer science publications. Hence, it is a good dataset choice to demonstrate our approach. For the purposes of this paper, we used a subset of DBLP data that includes a variety of publications in research areas including Databases, Web, Semantic Web, Data Mining, AI, and Information Retrieval. However, the method of building a publications-to-topics dataset is not tied to these areas. A similar subset of DBLP data was used for finding connected researchers [5]. In a similar way as in such work, we list the conferences, workshops, and journals of the papers composing the subset we used, see Table 1. In fact, the list in Table 1 is a superset of that listed in [5]. The subset used in our approach was taken from DBLP data as of May 1st, 2007.

Table 1. Publication venues of the papers included in the dataset used (subset of DBLP)

Conferences, Symposiums, Workshops (113)
AAAI, ADB, ADBIS, ADBT, ADC, ARTDB, BERKELEY, BNCOD, CDB, CEAS, CIDR, CIKM, CISM, CISMOD, COMAD, COODBSE, COOPIS, DAISD, DAGSTUHL, DANTE, DASFAA, DAWAK, DBPL, DBSEC, DDB, DEDUCTIVE, DEXA, DEXAW, DIWEB, DMDW, DMKD, DNIS, DOLAP, DOOD, DPDS, DS, DIS, ECAI, ECWEB, EDBT, EDS, EFDBS, EKAW, ER, ERCIMDL, ESWS, EWDW, FODO, FOIKS, FQAS, FUTURE, GIS, HPTS, IADT, ICDE, ICDM, ICDT, ICOD, ICWS, IDA, IDEAL, IDEAS, IDS, IDW, IFIP, IGIS, IJCAI, IWDM, INCDM, IWMMDBMS, JCDKB, KCAP, KDD, KR, KRDB, LID, MDA, MFDBS, MLDM, MSS, NLDB, OODBS, OOIS, PAKDD, PDP, PKDD, PODS, PPSWR, RIDE, RULES, RTDB, SBBD, SDB, SDB, SDM, SEMWEB, SIGMOD, SSD, SSDBM, TDB, TSDM, UIDIS, VDB, VLDB, W3C, WEBDB, WEBI, WEBNET, WIDM, WISE, WWW, XP, XSYM
Journals (28)
AI, AIM DATAMINE, DB, DEBU, DKE, DPD, EXPERT, IJCIS, INTERNET, IPM, IPL, ISCI, IS, JDM, JIIS, JODS, KAIS, SIGKDD, SIGMOD, TEC, TKDE, TODS, TOIS, VLDB, WS, WWW, WWJ

We considered various RDF-encoded datasets of DBLP data, namely, the D2RQ-generated RDF data from DBLP [3], Andreas Harth's DBLP dataset in RDF (sw.deri.org/~aharth/2004/07/dblp/), and our own SwetoDblp ontology

(lsdis.cs.uga.edu/projects/semdis/swetodblp/). We selected SwetoDblp because it allows the possibility to exploit the benefits of representing and aggregating data in RDF. For example, SwetoDblp includes affiliation data based on heuristics using the homepage information of the authors. Hence, the individuals participating in trends could be listed together with their affiliation. The other side of the coin is that it is possible to determine all the trends in which a given university or organization is associated (through the people affiliated with it). Other research efforts not necessarily related to trend detection, have highlighted the value of using semantics for describing data of publications [1, 7].

Table 2 lists the number of instances in the major classes. Compared to SwetoDblp, this subset is around one tenth in terms of number of entities. SwetoDblp is a dataset of 845 MB file size; the subset we used is 95 MB file size. Throughout this paper, we will refer to SwetoDblp whenever a particular aspect of such ontology is highlighted, otherwise we will simply refer to DBLP.

Table 2. Instances in main classes in the subset we used compared to DBLP (as of May 2007)

Main Classes	**Subset**	**DBLP**
Proceeding (of conferences, etc)	857	8,665
Articles in proceedings	51,202	532,758
Articles in journals	25,973	328,792
Authors	67,366	539,301

The four most important classes that were used in our subset are proceedings within conferences, authors, articles in proceedings, and articles in journals. Each publication that is part of a proceeding is related to authors and a proceeding with an "inproceedings" and "author" relationship. Determining the proceedings with which a paper is located in is very important in our work for plotting data at different time intervals. Each publication contains the year the paper was published. Although this suffices for creating a time graph that represents the years of research papers, it is not enough information to plot papers by day and/or month. In order to overcome this dilemma, we extracted exact dates from each proceedings title. Our approach is an improvement over other approaches that are limited in plotting data only based on years [5, 15]. For example, the proceeding title "Graphics and Robotics, Dagstuhl Castle, Germany, April 19 - 22, 1993" has an exact date at which such meeting took place. We used methods that extracted these dates from each proceeding in the dataset (if indeed there was a date in the title). With this information we can explicitly relate many publications to an exact date, namely, the date at which the paper was presented in its corresponding conference, workshop, or symposium.

Out of the all the proceedings in our dataset, we were able to extract dates for 94% of the proceedings. For the papers that were not able to get associated to an exact date, a check was done to see if the year of the paper matched the year of the last-modified-date (metadata value in DBLP) for that paper and if so then such exact date was used.

2.2 Taxonomy of Topics

Building a taxonomy of research topics is a significant endeavor. In order to give structure to the research topics a taxonomy was created that contains Computer Science topics. The taxonomy of topics has very good coverage for the areas of Databases, Web, Semantic Web, AI, Information Retrieval and Data Mining. Other topics in computer science are also included but at lesser depth (e.g., Computer Architecture). We adapted our taxonomy of topics to that of CoMMA ontology, which has over 420 concept "arranged in a taxonomy with a maximal depth of 12 levels hops, more than 50 relationships and more than 630 terms to label these primitives" [6]. We also verified and adjusted the organization of the topics of the taxonomy based on the AKT ontology [13]. We created our taxonomy taking into account lessons learned from an earlier effort on creating a small ontology of topics in Semantic-Web (lsdis.cs.uga.edu/projects/semdis/iswcdemo2006/). Our taxonomy is comprised of 344 research topics from research areas and over 200 synonyms thereof. The taxonomy is available online (http://cs.uga.edu/~cameron/swtopics/taxonomy).

2.3 Paper to Topics Relationship

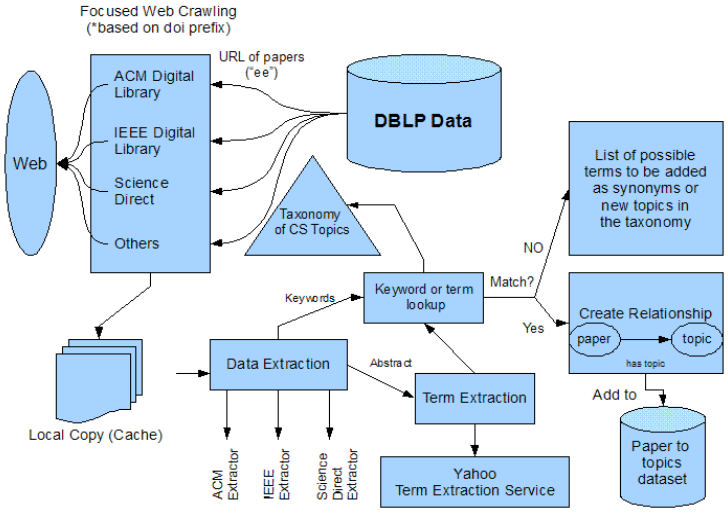

Figure 1. Overview of Creation of Papers-to-Topics Relationships

The information in DBLP is not sufficient to determine research topics of publications. For this reason, we developed methods to create paper-to-topic relationships. Creating these relationships was not a straightforward process (refer to Figure 1). The key aspect of our method is how we use the electronic edition "ee" URL literal value of individual papers (metadata value in DBLP) to retrieve additional information of publications. Based on such URL, we performed focused crawling for URLs having doi.acm.org, doi.ieeecomputersociety.org, or dx.doi.org/10.1016 prefixes.

Some publishers' sites, such as Springer, were difficult to extract documents for. If we could have extracted data from such large publisher site, then it would very likely improve the quality of our results. Crawled pages were stored in a local cache, from which data extraction methods obtained keywords and abstracts (when available).

The value of exploiting structure in different types of web content requires nontrivial data capture tasks [14]. In our case, the data scrapping we performed has the benefit that once data is extracted about a publication, such data is not expected to change. For example, the listed keywords and the abstract of a journal article will always be the same. In order to identify the key terms in abstracts, we used the Yahoo! Term Extraction API to determine, based on the input text, what are the most significant words or phrases. Each extracted term, keyword, and phrase of a paper was looked up in the taxonomy of topics to find matches with the name of research topics (or synonyms). If there was a match, a relationship from the paper to the research topic in the taxonomy was established. Otherwise, the terms were kept for possible consideration in improving the taxonomy (e.g., synonyms).

In the case of keywords of a paper, the process was similar but without need of term extraction. Two more methods were used. The first consisted of using the names of sessions in conferences as keywords for papers in such sessions. The second is a heuristics that assigns topics to all papers in a conference series but this is only applicable for very specialized conferences. At the end, 40,718 total relationships from paper to topics were determined. Table 3 lists a summary of how many such relationships were extracted from each site and by using keywords alone.

Table 3. Total number of paper to topic relationships created from extraction

Data source and/or data extraction method	**Relationships (paper to topic)**	**Papers with relationships to topics in taxonomy**
ACM (Keywords)	2,795	1,859
Science Direct (Keywords)	780	631
IEEE (Keywords)	617	454
ACM (Abstract/Terms Extraction)	5,641	3,574
Science Direct (Abstract/Terms)	2,330	1,688
IEEE (Abstract/Terms)	2,850	1,786
Crawling (session-names)	476	473
Conference Topics (heuristics)	25,229	23,083

3 Detection of Trends using Bibliography and Topics Data

Our method is able to detect two types of trends: bursty trends and emerging trends. In addition, it is possible to identify researchers at the emerging stages of a research topic. Figure 2 provides an overview of our approach. Several steps are taken in order to (1) retrieve all the information pertaining to a research area; (2) determining if a research topic is a bursty and/or emerging trend.

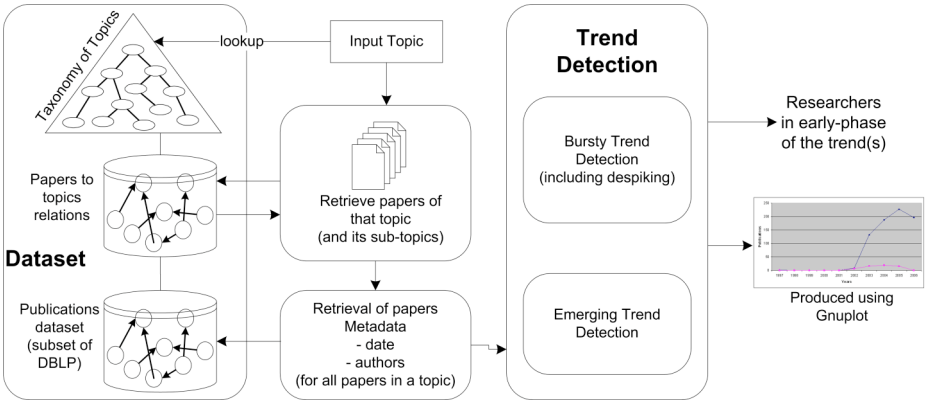

Figure 2. Overview of bursty and emerging trend detection and researcher identification

The information gathered on a research topic through our approach is very critical in our trend detection process. The first step is determining whether a given research topic is in the taxonomy. Second, the publications that are associated with the topic (and its sub-topics) are retrieved from our DBLP publications subset. The publications could not have been identified without the paper-to-topics dataset. Then, metadata such as authors and dates are used in detecting whether a research topic is a trend. The benefit of the taxonomy is that all subtopics of a topic are considered.

3.1 Detection of Bursty Trends

We adapted the work done by Gruhl et al. [8] for detection of bursty trends. They devised formulas for four predicates were devised in order to classify individuals to a region within a time graph of blog posts where they posted the most. We use a similar formula of one of the predicates for determining if a research topic is a bursty trend. Figure 3 is graph of an actual topic that illustrates how bursty trends are detected.

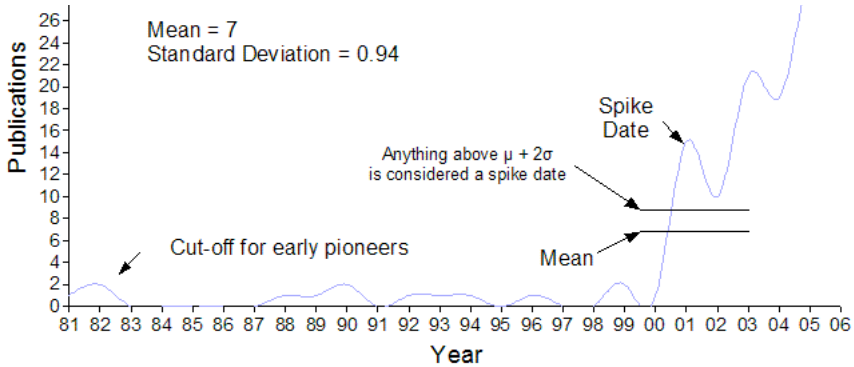

Figure 3. Bursty Trend Detection Overview

If the total number of publications for a particular research topic is greater than a threshold value ($\mu + 2\sigma$) for any day, month or year (depending on the time unit beging used), the research topic is considered to be a bursty trend. Figures 4 and 5 show examples of bursty trends that were detected by our approach. These are "Data Model" and "Semantics." Interestingly, both have had increased popularity in the last few years and both have also appeared in the literature over the last 30 years.

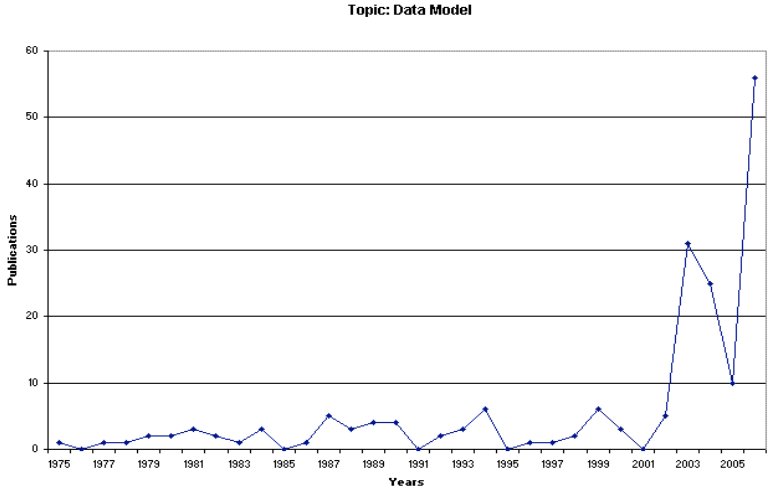

Figure 4. Example of Bursty Trend for Topic: Data Model

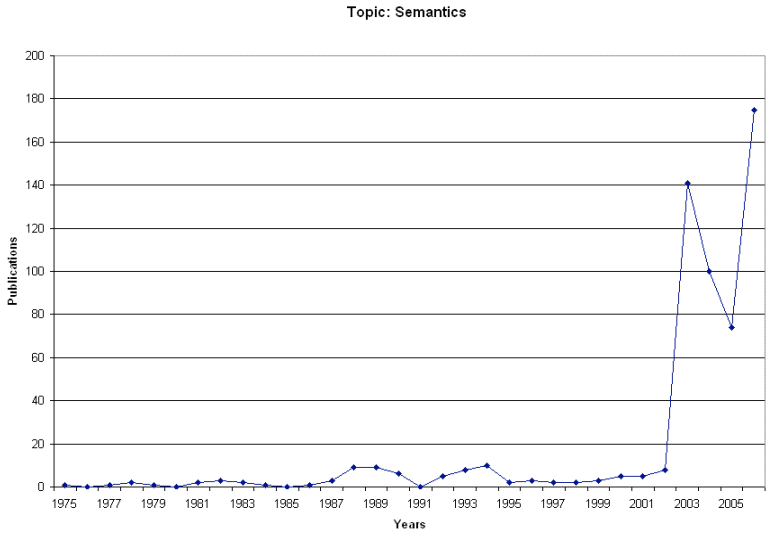

Figure 5. Example of Bursty Trend for Topic: Semantics

3.2 Detection of Emerging Trends

A method for detecting emerging trends used statistical information of documents published in research areas [15]. We implemented their algorithm for identification of emergent trends to apply it with our dataset. Their method determines whether there has been a significant increase in the total number of publications within recent years. Emerging trends do not necessarily exhibit a bursty behaviour. Figures 6 and 7 show examples of emerging trends that were detected with our approach. We purposely excluded the current year (2007) from our data for the reason of it not being a complete year as of yet.

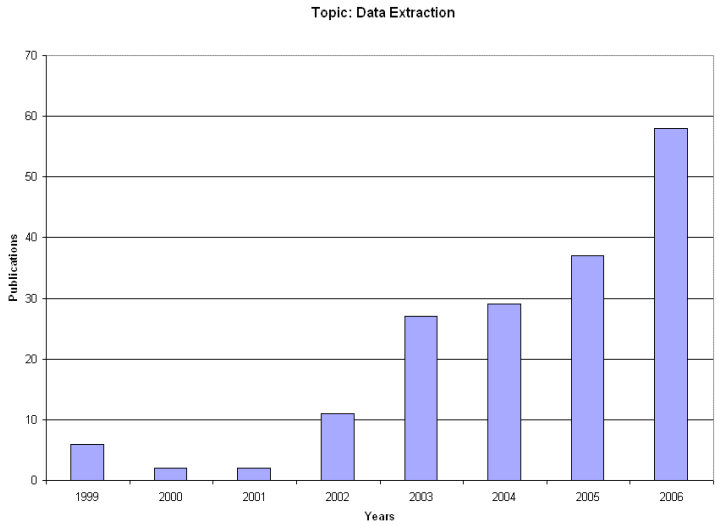

Figure 6. Example of Emerging Trend for Topic: Data Extraction

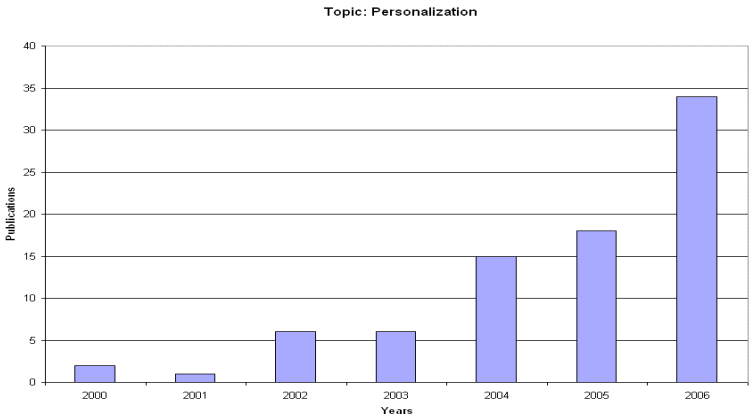

Figure 7. Example of Bursty Trend for Topic: Personalization

4 Evaluation and Results

4.1 Influential Researchers

After identification of trends, it is possible to determine the people involved at the early stage of the development of a trend. Evaluating whether we found influential researchers at the emerging stages of a research topic is challenging. We chose to compare the individuals appearing at the early stage of a trend with respect to four existing lists that contain highly recognized or prolific computer scientists. The lists are as follows. (1) ACM Fellows, (2) DBLP People that are in Wikipedia, (3) H-Index, and (4) Prolific Authors. (1) The ACM fellows list (fellows.acm.org/) includes members recognized in the Computer Science and Information Technology for their professional, technical and leadership contributions. (2) Researchers that have a Wikipedia page and also appear in DBLP can be arguably considered as persons that have had some sort of impact or recognition in Computer Science. We extracted URLs for such persons by processing a dataset of such persons made available by a recent effort on extracting semantics from Wiki content [2]. (3) The h-index is defined as a measure to characterize the scientific output of a researcher, where *h* is the number of papers with citation number higher or equal to *h* [9]. We used an existing list of computer science researchers with h-index of 40 or higher (www.cs.ucla.edu/~palsberg/h-number.html). (4) Prolific authors from DBLP data (www.informatik.uni-trier.de/~ley/db/indices/a-tree/prolific/index.html). For the ACM Fellows and those in the list of h-index, we looked them up manually to determine their URL in DBLP, which is the URI used for researchers in SwetoDblp.

We measured the overlap in these four lists and found somewhat little overlapping among the acknowledged people in these lists. Table 4 shows the results of the total number of recognized people who appeared in each list.

Table 4. Comparing overlap of lists of recognized/prolific researchers

	# Individuals in appearing in	**Percentage of total**
1 list	4470	91.97%
2 lists	322	6.63%
3 lists	64	1.32%
4 lists	4	0.08%

The individuals detected by our method appearing in the early stage of trends can then be compared to the lists before mentioned. However, before such comparison, a process was executed to exclude researchers that do not necessarily publish a lot based on using a measure of collaboration strength [12]. We found that a threshold of 1.0 was sufficient for excluding authors that, for example, have just one or two papers.

Table 5 shows a comparison of the overlap of the four lists plus the list of all researchers at the early stage of research trends identified by our method. This shows that our method detects many of the recognized/prolific authors. In fact, the relative percentages of both lists are very similar.

Table 5. Comparing our list with overlap of lists of recognized/prolific researchers

	# Individuals in appearing in	Percentage of total
1 list	4548	89.44%
2 lists	451	8.87%
3 lists	74	1.45%
4 lists	12	0.24%
5 lists	0	0.0%

Table 6 shows an example of researchers detected by our approach in the emerging stages of a research topic. These are cases where there is exact match of a recognition they have been given with respect to the topic where they were identified as possible "trend setters." The column *Contribution* in the table contains verbatim text from the corresponding list (either ACM Fellows site or a description from Wikipedia).

Table 6. Recognized researches from trend detection

Topic	Person	Appears in List	Contribution
Association Rules	Rakesh Agrawal	ACM Fellow H-Index Prolific Author (167)	"... contributions to data mining"
Database	E.F. Codd	ACM Fellow	"... contributions to the theory and practice of database management systems"
Information Extraction	Steve Lawrence	Prolific Author (58) Wikipedia Person	"Among the group ... responsible for the creation of the Search Engine/Digital Library CiteSeer"
Knowledge Discovery	Jiawei Han	ACM Fellow H-Index Prolific Author (274)	"For contributions in knowledge discovery and data mining"
Artificial Intelligence	Raymond Reiter	ACM Fellow Prolific Author (71)	"... contributions to artificial intelligence..."
Data Mining	Ming-Syan Chen	ACM Fellow Prolific Author (172)	"... contributions to query processing and data mining"
Information Extraction	C. Lee Giles	ACM Fellow Prolific Author (144)	"... contributions to information processing and web analysis"
Knowledge Acquistion	Rudi Studer	Prolific Author (130) Wikipedia Person	"Head of the knowledge management research group at the Institute AIFB"

4.2 De-spiking

De-spiking is the notion of figuring out whether there was some other topic(s) that substantially contributed towards a burst, which are also call spikes. For example, the

topic like “Ranking” can be used to relate to several types on ranking. De-spiking removes highly published subtopics that were used in the statistical information of a primary research topic for the purposes of analyzing what the cause of bursts was in a topic. This is achieved with the same method used for detecting a bursty trend. For each subtopic of a research topic, if it is determined that there is a spike in the total number of publications for a given day, month or year (depending on the time unit), then that subtopic is removed from the primary research topic and then re-plotted. Figures 8 and 9 show topics that were detected as bursty trends and the results after the subtopics were de-spiked. It is interesting to see that PageRank is indeed a topic that substantially contributes to the topic Ranking. In the case of de-spiking the topic Service, the contribution of topic Web Services is even more noticeable.

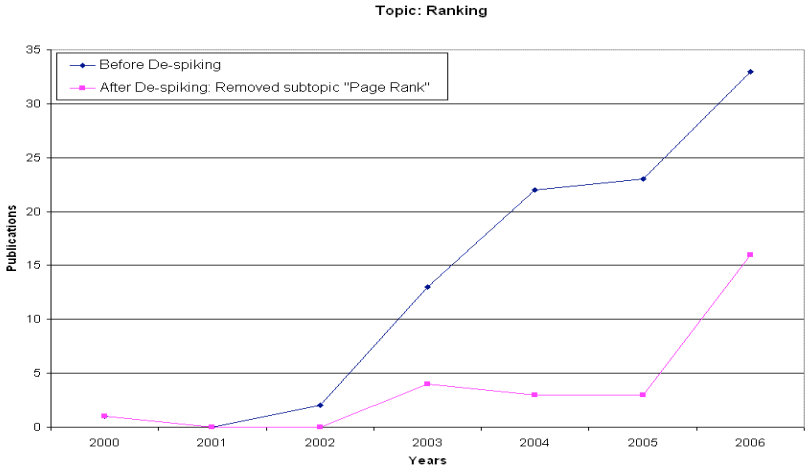

Figure 8. Example of De-spiking for Bursty Trend Topic: Ranking

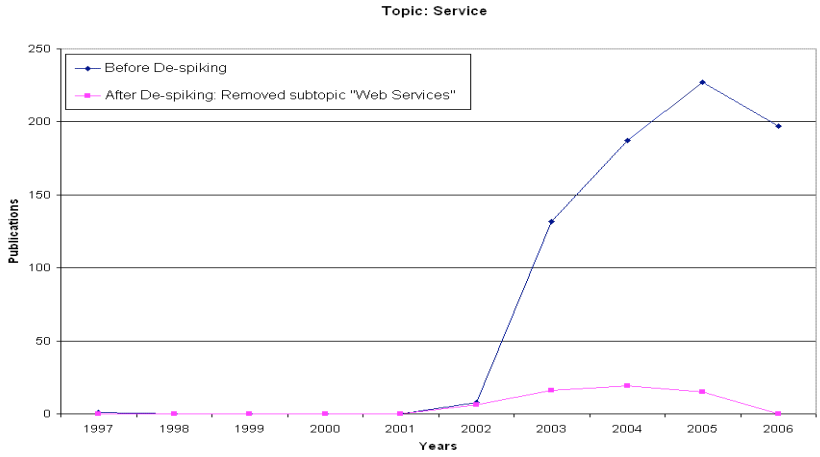

Figure 9. Example of De-spiking for Bursty Trend Topic: Service

5. Related Work

The identification of trends has been addressed with techniques such as mining and the use of social networks to name a few. For example, Tho et al. use a web mining approach for identifying research trends and emergent trends [15]. Their dataset was obtained using Indexing Agents that search for research web sites to download scientific publications. In contrast, our approach uses publications from the DBLP bibliography. Additionally, we use semantics in order to explicitly establish relationships between publications and topics.

Detection of bursts has also been studied in the work of [11]. Subsequent work, detects "bursts" in topic areas based on data extracted from blog feeds through the social network representation by the space of all weblogs [8]. Although our work is similar in the sense that bursts are detected on topics, we deviate from their work when it comes to using a different dataset. Our approach uses metadata of publications. Their approach relies on blog data, which has date/time information at more specific time units than research publications.

In the recent work of Zhou et al. trends of research topics can be found together with indication of how authors impact the topics [17]. However, their main concern is determining how topics are related and where and when these topics evolve. Specifically, they address the question of "Is a newly emergent topic truly new or rather a variation of an old topic"? Our work can complement their work by providing a collection of known "emergent trends" to evaluate.

The creation of taxonomies using web documents has been addressed towards detecting emerging communities and their associated interests [16]. A difference of our work from such approach is that we do not focus on the building of a taxonomy as the goal. Instead, we aim to demonstrating the value of the paper-to-topics relationships that connect the topics of a taxonomy to research papers.

6. Conclusions and Future Work

In this paper we were able to detect bursty trends and emerging trends using a semantic approach. Both methods used for detection were effective, resulting in detecting 118 research topics as bursty trends and 75 topics as emerging trends from among the listed 344 research topics in our taxonomy. Based on these results, the Computer Science area is indeed evolving in 34% of the topic areas listed in our taxonomy of topics. We were also able to pinpoint several topics that contributed in the burst of specific research topics by means of de-spiking. Our method for identifying researchers in the early stages of a research area was very effective in finding many exact matches of researchers that had major contributions within the research area being identified. We also demonstrated the potential of a semantic approach using only metadata of publications (i.e., abstracts and keywords). It was possible to detect trends without using all content in a document.

For future work, our approach for trend detection could be extended to use the terms of a trend in determining emails that match are related to the terms and the email data could then be used for social network analysis (e.g., identification of com-

munities) possibly relating it back to authors of papers. In addition, terms of trends of interest could be used for mining or processing other datasets such as intranets, blogs, forums and email corpus. For example, Kleinberg [10] described a scenario of grouping emails by topic of identified trends. Moreover, names in emails could be matched against authors of papers that are related to at trend.

References

1. Al-Sudani S., Alhulou R., Napoli A., Nauer E.: OntoBib: An Ontology-Based System for the Management of a bibliography, *17th European Conference on Artificial Intelligence*, Riva del Garda, Italy (August 28 - September 3, 2006)
2. Auer, S. and Lehmann, J.: What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. *4th European Semantic Web Conference*. Innsbruck, Austria (2007)
3. Bizer, C.: D2R MRP—a database to RDF mapping language, *12th International World Wide Web Conference*, Budapest, Hungary (2003)
4. Councill, I. G., Giles, L., Han, H., Manavoglu, E.: Automatic Acknowledgement Indexing: Expanding the Semantics of Contribution in the CiteSeer Digital Library. *K-CAP* (2005)
5. Elmacioglu, E., Lee, D.: On Six Degrees of Separation in DBLP-DB and More. *SIGMOD Record,* 34(2):33-40 *(June 2005)*
6. Gandon, F. Engineering an ontology for a mulit-agent corporate memory system. *Proceedings of ISMICK'01*, 209-228, (2001)
7. Golbeck, J., Katz, Y., Krech, D., Mannes, A., Wang, T.D., Hendler, J.: PaperPuppy: Sniffing the Trail of Semantic Web Publications, *Semantic Web Challenge at ISWC-2006*, Athens, GA, USA (November 2006)
8. Gruhl, D., Guha, R., Liben-Nowell, D., Ding, L., Tomkins, A.: Information Diffusion Through Blogspace. *WWW-2004,* New York, New York (May 17-22, 2004)
9. Hirsch, J.E. (2005) Random samples: Data point – Impact factor. *Science* 309, 1181
10. Kleinberg, J.: Bursty and Hierarchical Structure in Streams. *ISIGKDD '02*, Edmonton, Alberta, Canada (2002)
11. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the Bursty Evolution of Blogspace. *WWW2003*, Budapest, Hungary, (May 20-24, 2003)
12. Newman, M.E.J., *Phys. Rev. e Stat. Phys. Plasmas Fluids Relate. Interdiscip. Top.* 64, 016132, 2001
13. Nigel Shadbolt, Nicholas Gibbins, Hugh Glaser, Stephen Harris, Monica M. C. Schraefel: CS AKTive Space, or How We Learned to Stop Worrying and Love the Semantic Web. *IEEE Intelligent Systems,* 19(3):41-47 (2004)
14. The Yahoo! Research Team, "Content, Metadata, and Behavioral Information: Directions for Yahoo! Research". *IEEE Data Engineering Bulletin,* 31(4):10-18, (2006)
15. Tho, Q. T., Hui, S. C., Fong, A.: Web Mining for Identifying Research Trends. *ICADL 2003,* Berlin Heidelberg (2003) 290-301
16. Velardi, P., Cucchiarelli, A., Michaël Petit, M.: A Taxonomy Learning Method and its Application to Characterize a Scientific Web Community, *IEEE Transactions on Knowledge and Data Engineering*, 19(2):180-191 (February 2007)
17. Zhou, D., Ji, X., Zha, H., Giles, C.L.: Topic Evolution and Social Interactions: How Authors Effect Research. *CIKM-2006*, Arlington, Virginia, USA, pp. 248-257 (2006)