

An Efficient Algorithm for Topic Ranking and Modeling Topic Evolution

Kumar Shubhankar, Aditya Pratap Singh, Vikram Pudi

Center for Data Engineering, International Institute of Information Technology,
Hyderabad, India

{shubankar, aditya_pratap}@students.iiit.ac.in, vikram@iiit.ac.in

Abstract. In this paper we introduce a novel and efficient approach to detect and rank topics in a large corpus of research papers. With rapidly growing size of academic literature, the problem of topic detection and topic ranking has become a challenging task. We present a unique approach that uses *closed frequent keyword-set* to form topics. We devise a modified *time independent PageRank* algorithm that assigns an authoritative score to each topic by considering the sub-graph in which the topic appears, producing a ranked list of topics. The use of citation network and the introduction of time invariance in the topic ranking algorithm reveal very interesting results. Our approach also provides a clustering technique for the research papers using topics as similarity measure. We extend our algorithms to study various aspects of topic evolution which gives interesting insight into trends in research areas over time. Our algorithms also detect hot topics and landmark topics over the years. We test our algorithms on the *DBLP* dataset and show that our algorithms are fast, effective and scalable.

Keywords: Closed Frequent Keyword-set, Topic Ranking, Citation Network, Authoritative Score, Evolution

1 Introduction

The ever growing size of academic literature and fast changing fields of research pose a challenging task for a researcher to identify *significant* topics of research over the timeline. Topic discovery has recently attracted considerable research interest [13], [14], [15]. In this paper, we propose a novel and efficient method to detect and rank research topics. Based on the intuition that a document is well summarized by its *title* and the title gives a good high-level description of its content, we use the keywords present in the title of a paper to detect the topics. We form *closed frequent keyword-sets* as topics from the phrases present in the titles of papers on a user-defined *minimum support*.

We propose a time independent, modified iterative *PageRank* [3] algorithm to assign an authoritative score to the papers. For a topic T , we consider all the research papers containing that topic and the citation edges of these papers. We then assign an authoritative score to each topic using the scores of the papers containing that topic. Our topic ranking algorithm is able to rank the topics based on their *significance* in research community rather than *popularity* of the topics, which only considers frequency of topics. All the papers sharing a topic form a natural cluster. It is to be noted that a paper could belong to a number of clusters forming *hierarchical, overlapping* clusters.

Considering the topics on year-wise granularity, we modeled the evolution of topics on timeline. We apply the evolution of the topics for First Topic Detection, finding Landmark Topics and Fading Topics. Our algorithms have many applications like topic recommendation systems for authors, trend analysis, topic search systems etc. We tested our algorithms on the *DBLP* dataset. Our experiments produced a ranked set of topics that on examination by field experts and based on our study match the prominent topics in the dataset over the timeline.

Related Work

Topic extraction from documents has been studied by many researchers. Most work on topic modeling is statistics-based like the work by Christain Wartena *et al.* [6], which use most frequent nouns, verbs and proper names as keywords. Our work is based on dissociation of phrases into frequent *keyword-sets*, which as discussed in section 2 is very fast and scalable. Topic summarization and analysis on academic documents has been studied by Xueyu Geng *et al.* [10]. They have used LDA model to extract topics which needs a pre-specified number of latent topics and manual topic labeling. In our study, no *prior* knowledge of topics is required. The work in [11] uses the correlation between the distribution of terms representing a topic and the links in the citation graph among the documents containing these terms. We have used frequent *keyword-sets* to form the topics and utilized the citation links to detect important topics among the topics derived.

Clustering documents based on frequent item-sets [1] has been studied in the algorithms FTC and HFTC [7] and the *Apriori*-based algorithm [8]. Both of these works consider the documents as bags of words and then find frequent item-sets. Thus, the semantic information present in the document is lost. We extract phrases from the titles of the research papers and derive its substrings as *keyword-sets*, maintaining the underlying semantics. We have used closed [2] frequent *keyword-set* rather than maximal frequent *keyword-set* as used by L. Zhuang *et al.* [9] in their work on document clustering. We cannot use maximal frequent *keyword-sets* as topics because then most of the information is lost as it considers only the longest possible *keyword-set*.

2 Topic Detection and Clustering

The method proposed by us is based on the formation of *keyword-sets* from titles of the research papers and finding closed frequent *keyword-sets* to form the *topics*.

Definition 1. Phrase: A phrase P is defined as a run of words between two stop-words.

Definition 2. Keyword-set: A keyword-set K is defined as an n -gram substring of a phrase, n being a positive integer.

Definition 3. Closed Frequent Keyword-set: A keyword-set K is said to be frequent if its count in the corpus is greater than or equal to a user-defined minimum support [12]. We define a closed frequent keyword-set as a frequent keyword-set none of whose supersets has the same cluster of research papers as it has.

2.1 Phrase Extraction and Keyword-set Formation

Given the title of a research paper R_i , we extract all its *phrases* P_{ij} , where P_{ij} represents its j^{th} phrase. Each research paper R_i is mapped to the corresponding phrases P_{ij} present in its title. We reverse map the problem domain, mapping each phrase P_i to the research papers R_{ij} it belongs to, in one scan of the dataset. In this domain, each phrase will be dissociated into *keyword-set* only once, giving the frequency of *keyword-set* in the second scan.

In our approach, we have considered only the substrings of the phrases as *keyword-sets* and hence the relative ordering of keywords is maintained, preserving the underlying semantics. Each *keyword-set* thus formed is a semantic unit that can function as a basic building block of knowledge discovery and hence is a potential *topic*. As an example of *keyword-set* extraction, consider the phrase *xml data base*, the potential frequent

keyword-sets are the set of all the ordered substrings giving the following *keyword-sets*: (*xml, data, base; xml data, data base; xml data base*). It is to be noted that finding all the substrings requires a simple implementation of queue in top-down fashion, taking $O(1)$ time at each level and $O(n)$ time overall. Deriving the substrings of a phrase rather than the power set of the keywords in the phrase which requires $O(n)$ time instead of $O(2^n)$.

2.2 Closed Frequent Keyword-sets as Topics

Frequent keyword-sets are formed on a user-defined *minimum support*. The supports of the *keyword-sets* are calculated during the generation of the *keyword-sets* from the *phrases* in the second scan. The length of the list of research papers corresponding to a phrase is its *support*. It is to be noted that in the first scan, we cannot eliminate the phrases whose support is less than the minimum support as two or more phrases can share the same *keyword-set* whose combined support might be greater than the minimum support. The elimination of non-frequent *keyword-sets* is done only after all the *keyword-sets*, along with their supports, have been generated in the second scan. The algorithm to increment the support and add research papers to a given *keyword-set* is shown below:

Procedure 1: Frequent Keyword-set Generation
Require: phraseKeys PK , minimum support min_sup

```

1: for each phrase in  $PK$ 
2:   keywordSetList $KSL$  = findAllSubstringOf( $P$ )
3: for each keywordSet  $K$  in  $KSL$ 
4:   keywordSetCount[ $K$ ] += 1;
5:   add paper  $R$  to keywordSetPaperList[ $K$ ]
6: for each keywordSet  $K$  in keywordSetCount
7:   if keywordSetCount[ $K$ ] <  $min\_sup$ 
8:     delete(keywordSetCount[ $K$ ])
9:     delete(keywordPaperList[ $K$ ])

```

In the procedure 1, all the frequent *keyword-sets* are derived along with their supports. From step 1 to step 5, all the *keyword-sets* of each phrase are extracted and their supports in *keywordSetCount* and the corresponding paper list in *keywordSetPaperList* are updated. From step 7 to step 9, we remove infrequent *keyword-sets*.

Traditional association rule mining algorithms like *Apriori* that require one scan of the dataset to calculate the supports of the item-sets at each level take too much time and space. In our algorithm, we require only 2 scans of the dataset to calculate the supports of all the candidate *keyword-sets*. Since our algorithm runs in *linear* time compared to *exponential Apriori* like algorithms, our algorithms are fast and highly scalable. Also, in *Apriori* like algorithms which build higher length item-sets from smaller ones, the relative ordering between the item-sets is lost. In our method, relative ordering of keywords is maintained preserving the underlying semantic of the phrases.

At this point, we have the frequent *keyword-sets*. In our algorithm, we may derive non-closed frequent *keyword-sets* as well. Our topic should consist of the maximal number of common keywords present in all the papers in the cluster, so we remove the non-closed frequent *keyword-sets*. Thus, we have *closed frequent keyword-sets* as topics.

2.3 Clustering Research Papers based on Topics

Till now, we have *closed frequent keyword-sets* as topics which act as the similarity measure to cluster the research papers. These topic clusters are complete in the sense that we have the maximal length *keyword-set* shared by all the papers represented by that topic. In the mapping *keywordSetPaperList*, corresponding to each topic, we have a list of papers, forming several hierarchical, overlapping clusters. The cluster representing a broader topic is essentially a combination of several clusters representing its sub-topics. For example, *databas* is a broad topic and *imag databas*, *distribut databas*, etc. are its sub-topics. Each level of the hierarchy represents a different level of data description, facilitating the knowledge discovery at various levels of abstraction.

3 Ranking of Topics

Our next step is to order the topics. At this stage, we have a comprehensive list of topics from various fields of research and on varied levels of abstraction. For a researcher looking for new topics for research, it becomes a very cumbersome task to go through the entire list of topics and decide upon which topics are *important*.

To determine the *importance* of a topic, we introduce an approach which is based on the intuition that the topic's *importance* should be determined by not only its frequency in the corpus but also the quality of papers in which the topic lies and quality of citations those papers have. To this end, our approach assigns authoritative scores [4] to the topics producing a ranked list of topics. For each topic we have a cluster of papers in which the topic lies. To find out which papers are of good quality, we have developed a time independent, modified *PageRank* algorithm using the citation network of the papers.

Definition 4. Citation Graph: We define the citation graph $G = (V, E)$ comprising a set V of nodes, which each node N_i representing a research paper R_i and a set E of directed edges, with each edge E_{ij} directed from the citing node N_i to the cited node N_j .

Definition 5. Citation Sub-graph: For a topic T , its citation sub-graph $G_T = (V_T, E_T)$ comprises the set V_T of nodes, where the topic T lies in each node and the edges citing these nodes (G_T can be collection of many sub-graphs not necessarily a connected-graph).

Definition 6. Outlinks: From a given node N , link all the nodes N_i that the node N cites.

Definition 7. Inlinks: To a given node N , link all the nodes N_j that cite the node N .

The iterative formulae for calculating the *PageRank* score is:

$$PR(P) = (1-\theta) + \theta * \sum PR(P_i) / OC(P_i) . \quad (1)$$

Here $PR(P)$ is the *PageRank* score of the paper P . The *PageRank* algorithm is based on the fact that the quality of a node is equivalent to the summation of the qualities of the nodes that point to it. The *inlink* scores $PR(P_i)$ are divided by $OC(P_i)$ which is the number of *outlinks* of the *inlink* P_i . This takes care of the fact that if a paper cites more than one paper, it depicts that it has drawn inspiration from various sources and hence its effect on the score of the paper it cites should diminish by a factor equal to the number of paper it cites. The damping factor θ in the algorithm prevents the scores of research papers that do not have any *inlinks* from falling to zero. For the experiments we set the damping factor to 0.85 [3] which gave satisfactory results.

Time Invariant Factor: The basic *PageRank* algorithm does not take into consideration the time factor. It is observed that the newer papers do not get sufficient time to be cited

compared to the older papers and thus fall behind in the ranking even if they are *important*. To counter this, we introduce a time-dependent metric which reduces the bias against the older papers to make the ranking time-independent. This metric Average Year Citations Count, *AYCC* is a time dependent metric and directly reflects the varying distribution of citations over the years. We observe that this metric captures the time bias against the newer papers well and has high values for older papers and low values for newer ones. It is calculated as:

$$AYCC(Y) = \sum(P_I(P_Y))/N(P_Y). \quad (2)$$

$AYCC(Y)$ is the metric score for year Y . $P_I(P_Y)$ is the inlink count for papers published in year Y and $N(P_Y)$ is the total number of papers published in the year Y . Considering the year of publication of all the research papers, we pre-compute the total number of citations for each year and the number of research papers published in each year. Using them, the average number of citations per paper for each year is determined. Its inclusion normalizes the biased distribution of citations on the timeline. We use this metric in calculation of the modified *PageRank* score by the following formulae:

$$MPR(P) = (1-\theta) + \theta * (\sum MPR(P_i)/OC(P_i))/AYCC(Y_p). \quad (3)$$

The *modified PageRank* score of paper P , $MPR(P)$ incorporates the metric *AYCC* for its year of publication Y_p . Till now, we have *topic clusters* T each consisting of a number of research papers R_T dealing with research in the field represented by that topic. For ranking the topics, we use an authoritative score that takes into account the authoritative scores of the individual papers in the cluster. The formulae for topic score is as follows:

$$TS_T = (\sum MPR(P_T))/N_T. \quad (4)$$

The *topic score* TS_T of a topic T is the mean of the authoritative scores of all the research papers P_T present in the topic cluster, where N_T is the count of papers in the topic cluster T . Our algorithm is able to rank the topics based on their *significance*, considering the citation information as well as eliminating the time bias against the newer papers. Thus it is able to detect topics which may not be popular yet but may become popular afterwards. This information is not captured if we consider only the *frequency* of topics. For ex, if we analyze year-wise topics in the section 5.3 we find *mine associ rul* as top topic in 1993, which shows that even if it had just emerged and had low frequency, it still was a *significant* topic due to important citations it received over the time.

4 Evolution of Topics

Every topic has its time-span. Topics evolve over time. It is important for a researcher to know how the topics are evolving, which topics are on the surge, which are on the decline and so on. Also, since we have the cluster of research papers corresponding to each topic and we have the scores of these papers, the papers with high scores can be labeled as the *important* papers of the corresponding topic. Topic evolution has following two notions:

- **Topic Year-wise:** We assign authoritative scores of all the papers in each topic to their year of publication and then calculated the average score of a topic for each year. This gives a clear idea how a topic has evolved over the years.
- **Year-wise Topics:** We assign each topic's scores for all the years as calculated above and for each year, sort the scores, taking only the top few topics. The results give a clear picture of how the top ranked topics vary over the years.

5 Experiments and Results

5.1 Dataset Description

To show the results of our algorithms, we used the *DBLP XML Records* [5] dataset. The *DBLP* dataset contains information about 1,632,442 research papers from various fields published over the years. It is to be noted that the dataset contained papers with citation information till the year 2010 only. As part of data pre-processing, the keywords present in the titles of the research papers were stemmed using the *Porter's* Stemming algorithm.

5.2 Results of Topic Ranking

An objective and quantitative evaluation of the results obtained is difficult due to the lack of standard formal measures for topic detection tasks. But, the ranked list of topics produced by our experiments on examination by field experts and based on our observations match the prevailing topics in the dataset. We tested our algorithms on various values of minimum support. Upon implementing the topic detection algorithms with minimum support 100, we obtained 12,057 topics constituting 5,476 1-length topics, 5,766 2-length topics, 748 3-length topics, 62 4-length topics and 5 5-length topics.

In the results, we show only those topics for which the number of papers in their cluster is more than a threshold η . This threshold is used so that the clusters suffice a minimum number of papers for authoritative score calculation. It should be noted that the threshold η considers only those papers in a cluster that have at-least one citation. The following table shows the top ten topics, where $\eta = 10$.

Table 1. Top 10 Topics with their Respective Authoritative Scores and Cluster Supports

Topic	Score	Support
<i>congest avoid</i>	0.0791	112
<i>blind deconvolut</i>	0.0758	152
<i>learn tool</i>	0.0728	104
<i>sequenti process</i>	0.0719	112
<i>trecvid</i>	0.0716	111
<i>mine associ rule</i>	0.0710	197
<i>locat system</i>	0.0665	103
<i>hyperlink</i>	0.0662	200
<i>automat text</i>	0.0635	121
<i>large databas</i>	0.0623	346

We see that the quality of a topic is dependent on both the quality of individual papers as well as the number of papers in the cluster. A topic with few but good quality papers can have a high ranking. Also, the topics that appear at the bottom of the ranking are the ones which have not been/could not be researched much. Some of such bottom-ranked topics are *radio access*, *ant coloni optim algorithm*, *x rai imag*, *ipv6 network*, etc. These topics can be of special interest to the new researchers looking for new dimensions of research.

5.3 Evolution of Topics

Topic Year-wise: The evolution of a topic is informative in itself. We can infer the birth of the topic, its period of *significant* impact and its end. Here, we present two graphs for the evolution of some selected topics showing their *average* and *cumulative* topic scores.

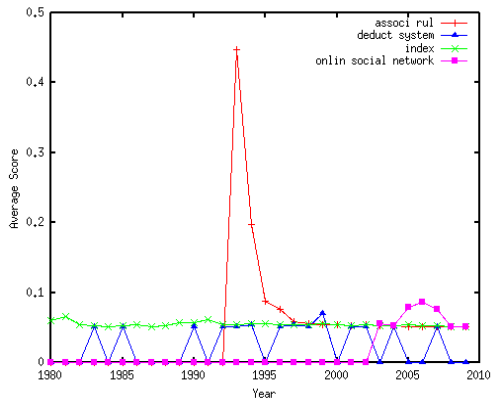


Fig. 1a. Graph showing average scores of the topics on year-wise granularity

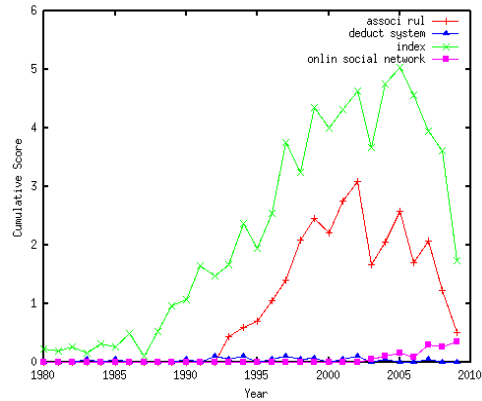


Fig. 1b. Graph showing cumulative scores of the topics on year-wise granularity

From the above two graphs, the following observations were obtained:

- If the average score of a topic is higher than that of another topic but its cumulative score is less, it means that the former topic has good quality papers in the cluster though few in number. For example, from 2005 to 2007, the topic *onlin social network* has better average score than the topics *index* and *associ rul* but its cumulative score is less than both of them. This is because *onlin social network* was a new-born but *significant* topic, while *index* and *associ rul* were already known fields and considerable work was being done on them. Thus, both average and cumulative scores need to be considered to get a clear picture.
- The significance of the topic *associ rul* between 1993 and 1996 is clear as shown in Fig 1a by its high average score which becomes similar to the score of *index* from 1997 onwards. From Fig 1b, it can be seen that the cumulative score of *index* was always higher than that of *associ rul*. Thus after 1997, the quality of *index* is similar to *associ rul*, but it was relatively more popular.

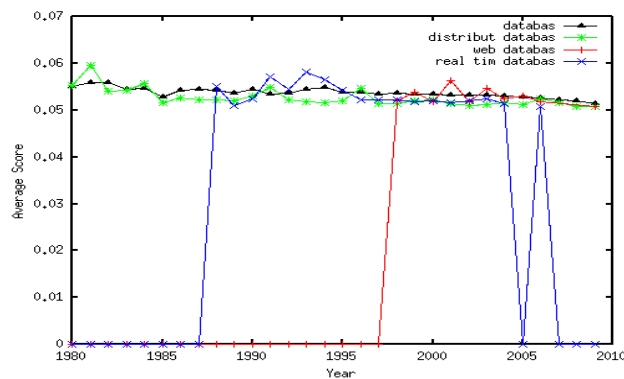


Fig 2. Graph showing the evolution of *databas* and some of its sub-topics

Another aspect of topic evolution could be studying the evolution of sub-topics of a topic. These sub-topics share a common *keyword-subset*. In the above graph in Fig 2, we show the evolution of the topic *databas* along with some of its sub-topics viz. *distribut databas*,

web databas and *real tim databas*. As a topic consists of various sub-topics, at all points, some sub-topics lie above the topic graph while others below it. The topic score gets contribution from all its sub-topics in addition to its own topic cluster. Sub-topics like *distrbut databas* span a major part of timeline while some of the topics give way to other topics or evolve into other topics as is the case with *web databas* and *real tim databas*.

Year-wise Top Topics: In this case, we compare all the topics for a given year. The following table shows the top three topics on year-wise granularity.

Table 2. Top Three Topics for each Year from 1993 to 2009

Year	Topic 1	Topic 2	Topic 3
1993	<i>mine associ rule</i>	<i>machin learn</i>	<i>larg databas</i>
1994	<i>associ rule mine</i>	<i>collabor filter</i>	<i>wordnet</i>
1995	<i>exchang blind deconvolut</i>	<i>data wareh environ</i>	<i>sequenti pattern</i>
1996	<i>data cluster</i>	<i>access control model</i>	<i>data hide</i>
1997	<i>adapt distribut</i>	<i>semi-structur data</i>	<i>collabor filter</i>
1998	<i>web search engine</i>	<i>wireless ad hoc network</i>	<i>anatomi</i>
1999	<i>learn tool</i>	<i>wireless sensor network</i>	<i>hyperlink</i>
2000	<i>instant messag</i>	<i>xml databas</i>	<i>evalu methodolog</i>
2001	<i>condit random field</i>	<i>dirichlet</i>	<i>peer to peer system</i>
2002	<i>stream system</i>	<i>cancer classif</i>	<i>k anonym</i>
2003	<i>transact memori</i>	<i>spatial correl</i>	<i>automat imag</i>
2004	<i>imag feature</i>	<i>network program</i>	<i>delaitoler network</i>
2005	<i>multi touch</i>	<i>object orient approach</i>	<i>onlin social network</i>
2006	<i>onlin social network</i>	<i>internet access</i>	<i>cyberspac</i>
2007	<i>evalu methodology</i>	<i>multimod interact</i>	<i>boltzmann machin</i>
2008	<i>distribut storage</i>	<i>trecvid</i>	<i>buffer manag</i>
2009	<i>fir filter</i>	<i>vision system</i>	<i>web portal</i>

Landmark Topics: We define *landmark topics* as those topics which gained extreme popularity within a short span of their emergence in the research domain. The year associated with a *landmark topic* is the year in which the topic *first* emerged. It is to be noted that these topics may not span sufficient number of papers but still our time-independent modified *PageRank* algorithm is able to derive these topics. The following table shows the *landmark topics* that have emerged in the last fourteen years:

Table 3. Landmark Topics that Emerged Between 1996 And 2009

Year	Landmark Topics		
1996	<i>java</i>	<i>data cube</i>	<i>visual cryptographi</i>
1997	<i>xml</i>	<i>firewall</i>	<i>robocup</i>
1998	<i>web search engin</i>	<i>mobil ad hoc network</i>	<i>cellular neural network</i>
1999	<i>xml base</i>	<i>sensor network</i>	<i>dynam web</i>
2000	<i>mine frequent pattern</i>	<i>e busi</i>	<i>open sourc softwar</i>
2001	<i>multipath rout</i>	<i>intuitionist fuzzi set</i>	<i>multi hop wireless network</i>
2002	<i>pagerank</i>	<i>agil method</i>	<i>mobil learn</i>
2003	<i>gpu</i>	<i>microarray gene express</i>	<i>spam filter</i>
2004	<i>blog</i>	<i>bpel</i>	<i>multiplay onlin</i>

2005	<i>bit torr</i>	<i>cross layer design</i>	<i>3d face recognit</i>
2006	<i>cell broadband engin</i>	<i>folksonomi</i>	<i>web 2 0</i>
2007	<i>social media</i>	<i>ieee 802 16e</i>	<i>time delai system</i>
2008	<i>cuda</i>	<i>cloud comput</i>	<i>svc</i>
2009	<i>microscopi imag</i>	<i>schrodinger equat</i>	<i>reson tunnel</i>

6 Conclusion and Future Works

In this paper, we proposed a method to derive topics, cluster papers into these topics and rank the topics using the authoritative scores of the constituent papers calculated by our time independent modified iterative *PageRank* algorithm. The topics were identified by forming *closed frequent keyword-sets* as proposed by our algorithms, which works better than traditional approaches like *Apriori*. We also studied the evolution of topics over time. We also analyzed the results of topic ranking and evolution of topics in detail.

As mentioned above, our algorithms have a variety of applications. In future, we would like to build topic recommendation systems. We would also like to examine statistical approaches for topic correlation and explore other domains like web site clustering, document clustering, etc. in which our algorithms can be applied.

References

1. Agarwal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. of the 20th VLDB Conference (1994)
2. Pasquier, N., Bastide, Y., Taoull, R., Lakhal, L.: Efficient Mining of Association Rules Using Closed Itemset Lattices.: Information Systems (1999)
3. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proc. of the 7th International Conference on World Wide Web (1998)
4. Klienber, J.: Authoritative sources in a hyperlinked environment. In: Proc. of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (1998)
5. The DBLP Computer Science Bibliography. <http://dblp.uni-trier.de/>
6. Wartena, C., Brussee, R.: Topic Detection by Clustering Keywords. In: Proc. of the 19th International Conference on Database and Expert Systems Applications (2008)
7. Beil, F., Ester, M., Xu, X.: Frequent Term-Based Text Clustering. In: Proc. of the 8th International Conference on Knowledge Discovery and Data Mining (2002)
8. Krishna, S.M., Bhavani, S.D.: An Efficient Approach for Text Clustering Based on Frequent Itemsets. European Journal of Scientific Research (2010)
9. Zhuang, L., Dai, H.: A Maximal Frequent Itemset Approach for Web Document Clustering. In: Proc. of the 4th International Conference on Computer and Information Technology (2004)
10. Geng, X., Wang, J.: Toward theme development analysis with topic clustering. In: Proc. of the 1st International Conference on Advanced Computer Theory and Engineering (2008)
11. Jo, Y., Lagoze, C., Giles, C.L.: Detecting Research Topics via the Correlation between the Graphs and Texts. In: Proc. of SIGKDD (2007)
12. Agarwal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proc. of the 1993 ACM SIGMOD Conference (1993)
13. Griffiths, T.I., Steyvers, M.: Finding Scientific Topics. In: Proc. of the National Academy of Sciences (2004)
14. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.I.: Probabilistic Author-topic Models for Information Discovery. In: Proc. of SIGKDD (2004)
15. Mei, Q., Zhai, C.: Discovery Evolutionary Theme Patterns from Text – An Exploration of Temporal Text Mining. In: Proc. of SIGKDD (2005)