The Application of Weighted Co-occurring Keywords Time Gram in Academic Research Temporal Sequence Discovery

Shuqing Li
Department of Information Management and Information System, College of Information Engineering, Nanjing University of Finance & Economics, Nanjing 210046
leeshuqing@gmail.com

Ying Sun
Department of Library and Information Studies, Graduate School of Education, University at Buffalo, New York State University, Buffalo, U.S.A., 14260
sun3@buffalo.edu

ABSTRACT

The discovery and visualization of temporal sequence of personalized academic research can enhance the ability for discovering the latent trend of interests to information seekers. In this paper, we propose a definition of weighted co-occurring keywords time gram and use it as a basic unit to analyze the temporal information in existed keywords collection. We further propose a method to get the temporal sequence and temporal network based on these time grams. An application of the proposed method in discovering academic research temporal sequence is discussed, which includes techniques for acquiring extended keywords, assigning weight to each keyword and co-occurring weight to each keyword pair. A visualization tool is designed for browsing the temporal networks identified. Finally, we report an experiment in the area of library and information studies. The experiment results show the effectiveness of the proposed method in helping users analyzing and portraying the evolution pattern and developing trend of corresponding academic research.

Keywords

word co-occurrence, temporal sequence, visualization, academic research.

INTRODUCTION

It is very important for modern information recommending system to acquire more personalized and valuable information from users. Traditional methods often pay more attention to analyze the users' interests based on pure semantic information such as keyword or ontology. Along with the development of temporal analysis researches, more and more scholars begin to show more interests in mining the temporal information in all kinds of users' data and use them to enhance the ability for expressing the users' personalized models, so as to recommend more satisfactory results to users. In fact, temporal information is a very important criteria used by users to express their information needs or make relevance judgments. Traditionally, temporal information has been always treated as a document property. We propose a method to treat temporal information as a characteristic of the topics covered in documents. We can then portray a temporal network of topics for a subject area or a specific author using the temporal information and the co-occurrences of the keywords. We believe with a user-friendly visualization interface such a temporal network, it would help users examine and take full advantage of the information results.

We review related studies in the next section, followed by the introduction to our method. Then a pilot study on the field of library and information studies is reported to illustrate the proposed method. The paper concludes with some preliminary evaluation results.

RELATED WORKS

The research of how to obtain and analyze time information has primarily relied on statistical methods. However, these traditional statistical methods cannot provide the complex temporal information required in the area of information service Alkilany (2013) summarizes that temporal information involved in an information process can be classified into three types. First is the time information stored in records, the second one is the event time of records, and the last one is user-customerized time which often has different meanings in different applications. Many studies have been done on how to extract and use these types of temporal information. For example, Zhang and Li (2011) incorporate time information with semantic information in document classification. They include three types of time information in their analysis: exact time, vague time and decorative time. Xu and Li (2011) combine the machine-learning technology and rule-based methods to improve the effectiveness of identifying time expression.

Some studies have shown that combining semantic information with time information smoothly can provide extra perspective and usage of time information analysis (Radinsky & Svore, 2012; Leprovost & Abrouk, 2012). More and more scholars have begun to focus on this trend in the area of information retrieval and personalized service (Kim & Xing, 2013). It is believed that such technique can help reducing semantic ambiguity and exploring temporal trend of information resources such as academic researches and media news. For example, Hong and Zhang (2012) design a new method to identify the latent burst words in Internet news and academic documents based on the life cycle of keywords and their burst rules. Corresponding researches have formed some specific research topics such as trend analysis of disciplines, discovery of academic hotspot, etc. Ma and Lv (2012) use bibliometrics analysis with temporal analysis and draw conclusion that concepts in specific research area can expand to other areas with time passing by. Yin and Hirokawa (2013) design a cross tabulation search engine with the functions of conducting search, recognizing important authors and documents, and analyzing research trends.

Citation analysis has been used in the area of disciplines trend analysis and academic research trend discovery for a while since citation relationships can reflect temporal information. Farideh (2009) in his review of bibliometrics summarizes various methods and applications of citation analysis and co-citation analysis. Li and Ren (2010) use a citation timing visualization software to depict the citation chronological diagram in the research area of hybrid rice. They successfully identify more important literatures and the research trend in the area. We also have conducted some related research before such as the automatic recognition and visualization method of main-path in academic documents based on vibration algorithm and domain ontology, the research of automatic construction of domain ontology in library and information science based on weighted co-occurrence of citation keywords, and etc (Li, 2012; Li & Xu, 2012). Chen and Lv (2009) use knowledge mapping which combines the visualization with citation analysis to recognize important documents in mechanics research area. Li and Hou (2007) give a thorough review of citation timing visualization and co-citation analysis visualization.

However, citation is not a direct time indicator and carries limited time information. Using citation relations in temporal and trend analysis has its own problem and limitation. First, older documents tend to get more citing and newer ones get less. More importantly, citing relations only exist in the cited documents and citing documents.

How to take advantage of temporal information extracted from academic documents collection to serve academic research is an area getting more attentions from researchers. Ways of combining methods borrowed from other research areas such as bibliometrics indicators, machine-learning, and network analysis of literature co-citation clustering are tested (Wang & Wang, 2012; Yin, 2008). These methods can also be applied to personalized information retrieval and Web conception recognition (Al & Mills, 2012; Borth & Ulges, 2012). Recently, some scholars begin to use probabilistic graphical models to describe the evolution of scientific documents. Effective trends are obtained by extracting the topics in documents along a time series (Wang & Xu, 2009). Zhang and Liang (2010) try to identify the research hotspots and developing trends of a specific discipline by analyzing the full text of academic documents including time information, and clustering these documents based on topic clustering. Li and Liang (2011) analyze the general evolution trends of Chinese information system theory researches between 2000 and 2009 using topic classification.

Studies have shown that keywords co-occurrence method can enhance semantic analysis (Wang & Song, 2007). For example, Xia and Cheng (2012) design a multi-dimensional visualized clustering analysis method based on keywords co-occurrence in e-government platforms. We also have done some related researches based on keywords co-occurrence (Li, 2012; Li & Lv, 2012; Li & Bai, 2011). Built upon our researches on keyword co-occurrence, we propose this new method of integrating temporal information with term co-occurrence to discover and visualize the personalized academic temporal path.

TIME GRAM, TEMPORAL SEQUENCE, AND TEMPORAL NETWORK

Weighted Co-occurring Keywords Time Gram

The current methods for constructing semantic unit only pay attention to the semantic relationship of different keywords. Here we propose to assign a time element to each semantic unit in order to enhance its expressing effectiveness. We define the new expression of semantic time unit as in Formula 1:

$$\text{SemanticTimeGram}=<\text{SemanticUnit, Time}> \quad 1)$$

The semantic units can be any level from simplest keywords to more complex ontology items. In this paper, we use co-occurring keywords' pair as semantic unit and call it co-occurring keywords time gram, as shown in Formula 2:

$$\text{CooccurringKeywordsTimeGram}= <\text{Keyword}_1, \text{Keyword}_2, \text{Time}> \quad 2)$$

We can also weight each co-occurring keywords time gram. The weighted co-occurring keywords time gram is constructed as Formula 3:

$$\text{WeightedCooccurringKeywordsTimeGram}=<\text{Keyword}_1, \text{Keyword}_2, \text{Weight}_{\text{Keyword1, Keyword2}}, \text{Time}> \quad 3)$$

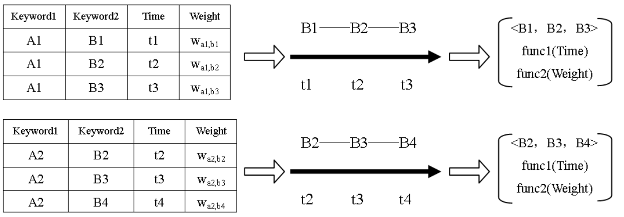


Figure 1: An example illustrating how keyword temporal paths are generated. (t1, t2, t3, t4 denote sequential time one by one)

Keyword Temporal Path

By lining the weighted co-occurring keywords time grams along a temporal sequence, we can obtain a semantic temporal path which will provide a good way to identify the evolution trend of corresponding semantic information. The basic expression of weighted co-occurring keyword temporal path is shown as Formula 4:

$$\text{CooccurringKeywordsTemporalPath}=<\text{WeightedCooccurringKeywordsTimeGram}_1, \text{WeightedCooccurringKeywordsTimeGram}_2, \ldots, \text{WeightedCooccurringKeywordsTimeGram}_n> \quad 4)$$

The grams in Formula 4 are ordered according their time.

To identify effective paths from the weighted co-occurring keyword temporal path, we start with each specific keyword, collect all of its co-occurring keywords, construct the weighted co-occurring keywords time grams, and arrange all these keywords to construct a complete keyword temporal path.

Figure 1 illustrates how keyword temporal path is obtained. In the example, two different keyword temporal paths are identified. They are < B1, B2, B3 > and < B2, B3, B4 >. Each path has its own weight value and time value which is calculated by the weight values and time values of the weighted co-occurring keywords time grams on this path. The weight value denotes the importance of path and the time value can be used to combine with other paths to formulate longer path. The two functions used to calculate these two values, func1 and func2, can be selected based on the application this model is used.

As for the size of keyword pairs, our previous study shows that three-keyword co-occurrence pair can provide better results for finding tightly related keywords (Li, 2013). However, to generate a meaningful three-keyword temporal semantic network requires large amount of data that it is rarely available, so it is more feasible to use two-keyword pairs in most of application areas.

To get longer keyword temporal path, we can further combine different keyword temporal paths together. In this study, we adopt a two-keyword coupling method. We start with identifying all paths with up to three keywords. Then any two different paths are examined. If the beginning two keywords in one path are the same as the last two keywords in the other path and the time value of the first path is later than that of the second path, we combine the two paths to one. For example, given the two paths in Figure 1, < B1, B2, B3 > and < B2, B3, B4 >, if func1(t1, t2, t3) is less then func1(t2, t3, t4), we can combine them to get a longer path <B1, B2, B3, B4>. The weight and time value of this path will be recalculated.

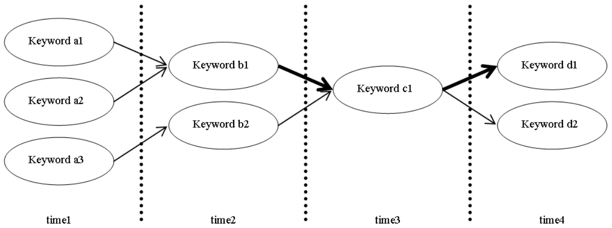


Figure 2: Keyword temporal network composed of different keyword timing paths

Keyword Temporal Network

The keyword temporal network can be constructed by combining many keyword temporal paths together. Assume there are three keyword temporal paths <a1, b1, c1, d1>, <a2, b1, c1, d2> and <a3, b2, c1, d1>. The nodes beginning with a, b, c and d belong to time1, time2, time3 and time4 respectively, and the weights of all the paths are identical, we can construct the keyword temporal network as shown in Figure 2.

The weight of a line linking two keywords denotes the weight of the relationship of the two keywords. In this figure, links <b1, c1> and <c1, d1> have higher weight than others. We use this method to depict more effective paths such as path <b1, c1, d1>.

A keyword temporal network generated using the proposed method has following important characteristics:

1. The links are ordered, and they show the evolution of corresponding keywords. Some keywords may appear more than once in the network, but in different time periods, which may denote the reappearance of academic research hotspots or the emergence of new complex research hotspots.
2. It has the flexibility of using different weighting methods on the nodes and links.

RECOGNITION OF ACADEMIC RESEARCH TEMPORAL PATH

One direct application of the proposed method is to identify academic research temporal path.

In order to get the academic research temporal path, we have to solve following problems:

1. Where and how to select keywords

2. How to measure the importance of different keywords on a temporal path
3. How to measure the relationship of different keywords on a temporal path
4. How to measure the effectiveness of final keyword temporal path

These problems and resolutions are tightly related. We first design a measure algorithm for keyword weight and get the measure algorithm for the weight of co-occurring keywords based on keywords' weights. The effectiveness of final keyword temporal path is also decided by the weight of co-occurring keywords.

Keywords Selection

The keyword list submitted by authors is a good source that reflects the subject content of the article. However, in each paper, the number of keywords is usually limited to 3 or 4. To build a temporal semantic network, we need more subject keywords from each document. We designed a method based on the assumption that a scholar's research works are generally focused on a certain area and related. So for each article, we mark not only the position and frequency information of the terms in the keyword list of the article, but also terms in keyword lists of the same author(s)' s other publications. We limit our analysis in title, abstract, highlight and keywords list fields. Highlight is a special segment which marks the most important ideas of each article in Elsevier database. The exacting process is described in the following pseudo code:

```
Input: The title, abstract, highlight and keywords list of each paper
Output: The extended keyword list and their term frequencies
// Iterate each scholar in collection
for each authori in authorList{
  // Get all the paper of current scholar
  Collection documents=getDocumentsByAuthor(authori);
  // Get all the keywords of the whole collection
  Collection keywords=getKeywordsByAuthor(authori);
  // Iterate each keyword of the whole collection
  for each keywordk in keywords{
    // Get the keyword and its frequency if exists in title
    if(existed(keywordk , getTitle(docj)))
      insertRecordInTitle(docj ,keywordk ,getFrequency(keywordk , getTitle(docj)));
    // Get the keyword and its frequency if exists in abstract
    if(existed(keywordk , getAbstract(docj)))
      insertRecordInAbstract(docj ,keywordk ,getFrequency(keywordk , getAbstract(docj)));
    // Get the keyword and its frequency if exists in highlight
    if(existed(keywordk , getHighlight(docj)))
      insertRecordInHighlight(docj,keywordk,getFrequency(keywordk,getHighlight(docj)));
  }
 }
}
```

Keyword Weighting

We use a weighting method similar to standard TFIDF method. Two components are calculated in each keyword's weight. The first part reflects the importance of a keyword itself or the resolving power of a keyword in the collection, the higher the keyword document frequency, the less its power in expressing the characteristics of document. So we assign this weight of each keyword as Formula 5 shown:

$$\text{weight1}_{\text{keywordi}}=\log(N/DF_{\text{keywordi}}) \quad 5)$$

N is the total number of documents in the collection. DF is the document frequency of keyword$_i$. We use logarithmic function to decrease the excessive impacts of high value.

The second part reflects the importance of keyword in a document which is calculated by term frequency and the weight coefficient of the location the term appears. We assign different weight coefficients to different fields a keyword may appear. So we calculate the weight of keyword$_i$ in doc$_j$ as Formula 6:

$$\begin{aligned}\text{weight2}_{\text{keywordi,docj}}=&\text{TFInAbstract}_{\text{keywordi, docj}}\times\text{coff}_{\text{abstract}}+\\ &\text{TFInKeywordslist}_{\text{keywordi, docj}}\times\text{coff}_{\text{keywordlist}}+\\ &\text{TFInHighlight}_{\text{keywordi, docj}}\times\text{coff}_{\text{highlight}}+\\ &\text{TFInTitle}_{\text{keywordi, docj}}\times\text{coff}_{\text{title}} \quad 6)\end{aligned}$$

The TFInAbstract, TFInKeywordslist, TFInHighlight and TFInTitle mean the keyword frequency in abstract, keywords list, highlight and title field respectively. In this study, we assign field weight coefficients as follows: the weight coefficient of abstract field is 1, keyword list field is 2, highlight field is 3, and title field is 4. This assignment is based on ad-hoc observation. The number of TFInAbstract may be large compared with other fields, so the corresponding weight coefficient should be set lower. TFInKeywordslist only has two values which are 0 and 1. When a document has this keyword in its original keywords list, TFInKeywordslist is 1, otherwise is 0 and corresponding keyword is an extended keyword. The coefficients of TFInHighlight and TFInTitle are set higher given the short length of the fields.

The final weight of one keyword in a document is shown as Formula 7:

$$\text{weight}_{\text{keywordi, docj}}=\text{Norm}(\text{weight1}_{\text{keywordi}}\times\text{weight2}_{\text{keywordi, docj}}) \quad 7)$$

Norm is normalization function which divides each value with the maximum value and sets each value within 0 and 1.

Co-occurring Keyword Pair Weighting

The weight of co-occurring keyword reflects the effectiveness of their co-occurrence. Only paths connecting keywords with certain higher weight will be used in making final evolution discovery. Directional Affinity (DAff) is a traditional method for computing the weight of co-occurring keywords. But it ignores the importance of keywords themselves and only pays attention to the co-occurrence of keywords. For example, keyword A and B co-occur in one document, and both of them only appear once in the document, keyword C and D also co-occur in only one document, but each of them appears 10 times in the document, we cannot distinguish these two keyword pairs using DAff because their document frequencies all equal to 1. Based on DAff, we propose here a new method in which document frequency is replaced with keyword weight. The formula is shown in Formula 8:

$$relation_{keywordi,keywordj} = \frac{\sum_{dock}(weight_{keywordi,dock} \times weight_{keywordj,dock})}{\sum_{dock}(weight_{keywordi,dock})} \quad 8)$$

The value of Formula 8 is asymmetric for any two co-occurring keywords. That is to say, we have to calculate both the co-occurring weight of (A, B) and (B, A). Our experiment results show that this method is effective in differentiating keyword pair with similar semantic meanings, which is critical for academic research temporal path discovery.

The Discovery of Academic Research Temporal Path

In this part, we describe how to generate keyword temporal path given an academic publication collection. The detailed step includes four steps:

1) We first collect all weighted co-occurring keywords time grams. Then we use the co-occurring weight of keywords (formula 8) as the weight of each temporal gram and assign the publication year of the document to which keywords belong as time of each temporal path. Each temporal gram is shown as Formula 9:

$$<keyword1, keyword2, relation_{keyword1, keyword2,} year> \quad 9)$$

2) Each keyword in the collection of time grams will be used as the first keyword in keywords' pair. Based on these keywords, we can get all of keywords timing paths composed of the second keywords. Since we have to combine different paths to get longer timing paths, we start with getting the timing path having at most three keywords. The weight of each keyword timing path is the weights' sum of corresponding weighted co-occurring keywords time grams, and time value is average weight of these grams. Each keywords timing path is shown as Formula 10:

$$< keyword_1, keyword_2keyword_3, \sum_{i=1}^{3} relation_{keywordx,keywordi}, avg(year) > \quad 10)$$

The $keyword_x$ is the co-occurring keyword of both $keyword_1$, $keyword_2$ and $keyword_3$.

3) The identical keyword temporal paths will be merged. The final weight is the sum of the weights of paths merged and the time value is the average time value.

4) We generate the final whole keyword temporal network by combining different paths according to whether the front part of one path is identical to the end part of another path. The weight of each final path is the average weight of corresponding keyword temporal paths.

Document Number	Keywords	Year
2173	NULL	2000
2456	Web searching/Session duration/Query language/Search engine evaluation/	2005
2496	Automated assistance/Intelligent information retrieval systems/Explanation systems/Contextual help/Adaptive interfaces/Implicit feedback/	2005
2560	Web search engines/Web searching/Transaction log analysis/	2006
2561	NULL	2006
2624	Web searching/Web search engines/Web search engine evaluation/Ecommerce searching/Paid searching/Sponsored results/Organic results/Non-sponsored links/	2006
3995	NULL	2006
2645	Web search engine/Overlap/Google/Yahoo/MSN Search/Ask Jeeves/Dogpile/Infospace Inc/	2006
2841	Collaborative information behavior model/Collaborative information behavior/Healthcare teams/Healthcare information behavior/	2008
2917	User intent/Web queries/Web searching/Search engines/	2008
2984	ARIMA/Box-Jenkins model/Search engine/Time series analysis/Transactional log/	2009
3028	Web searching/Information searching/To Anderson and Krathwohl's taxonomy/Bloom's taxonomy/	2009
3141	Real time search/Real time content/Collecta/Twitter/Economic value of search/Search topics/	2011
3284	Sponsored search/Keyword advertising/Pay-per-click/PPC/Online advertising/Search engine marketing/Gender targeting/Demographic profiling/	2013

Table 1: The documents of one scholar in experiment

The scope of the collection can be flexible. It can be the document collection of a specific author. In that case, the temporal path generated will be a personalized academic research temporal path of this author. If the collection is an information user's search result set, the method can serve the user's information need by generated personalized temporal path of the documents the user selects.

Document Number	Keywords and Their Term Frequencies	Document Number	Keywords and Their Term Frequencies
2173	excite(1:0:0:0), queries(3:0:0:1)	3995	transaction log analysis(4:0:0:0), web search engine(1:0:0:0), web search engines(1:0:0:0), web searching(1:0:0:0)
2456	query language(0:0:1:0), search engine evaluation(0:0:1:0), session duration(0:0:1:0), web searching(0:0:1:0), information system(1:0:0:0), queries(3:0:0:0), search topics(1:0:0:0), web search engine(1:0:0:0), web search engines(1:0:0:0)	2841	collaborative information behavior(0:0:1:0), collaborative information behavior model(0:0:1:0), healthcare information behavior(0:0:1:0), healthcare teams(0:0:1:0), healthcare(0:0:0:1)
2496	adaptive interfaces(0:0:1:0), automated assistance(0:0:1:0), contextual help(0:0:1:0), explanation systems(0:0:1:0), implicit feedback(0:0:1:0), intelligent information retrieval systems(0:0:1:0)	2917	search engines(0:0:1:0), user intent(0:0:1:0), web queries(0:0:1:0), web searching(0:0:1:0), classification(5:0:0:0), queries(5:0:0:1), search engine(1:0:0:0), web search engine(1:0:0:0)
2560	transaction log analysis(0:0:1:0), web search engines(0:0:1:0), web searching(0:0:1:0), queries(1:0:0:0), search engine(1:0:0:1), search engines(8:0:0:0), web search engine(10:0:0:0), world wide web(0:0:0:1)	2984	arima(0:0:1:0), box jenkins model(0:0:1:0), search engine(0:0:1:0), time series analysis(0:0:1:0), transactional log(0:0:1:0), queries(3:0:0:0), web search engine(1:0:0:1)
2561	interactive information retrieval(1:0:0:0), multitasking(5:0:0:1), queries(2:0:0:0), retrieval(1:0:0:0), web search(7:0:0:1), web search engine(3:0:0:0), web searching(1:0:0:0)	3028	bloom s taxonomy(0:0:1:0), information searching(0:0:1:0), to anderson and krathwohl s taxonomy(0:0:1:0), web searching(0:0:1:0)
2624	ecommerce searching(0:0:1:0), non sponsored links(0:0:1:0), organic results(0:0:1:0), paid searching(0:0:1:0), sponsored results(0:0:1:0), web search engine evaluation(0:0:1:0), web search engines(0:0:1:0), web searching(0:0:1:0), search engine(4:0:0:0), search engines(3:0:0:1), web search engine(0:0:0:1)	3141	collecta(0:0:1:0), economic value of search(0:0:1:0), real time content(0:0:1:0), real time search(0:0:1:0), search topics(0:0:1:0), twitter(0:0:1:0), google(2:0:0:0), keyword advertising(1:0:0:0), online advertising(1:0:0:0), search engine(4:0:0:0), search engines(1:0:0:0), web searching(1:0:0:0)
2645	ask jeeves(0:0:1:0), dogpile(0:0:1:0), google(0:0:1:0), infospace inc(0:0:1:0), msn search(0:0:1:0), overlap(0:0:1:0), web search engine(0:0:1:0), yahoo(0:0:1:0), queries(8:0:0:0), retrieval(1:0:0:0), search engine(4:0:0:0), search engines(2:0:0:0), web search(12:0:0:1), web search engines(12:0:0:1)	3284	demographic profiling(0:0:1:0), gender targeting(0:0:1:0), keyword advertising(0:0:1:0), online advertising(0:0:1:0), pay per click(0:0:1:0), ppc(0:0:1:0),search engine marketing(0:0:1:0), sponsored search(0:0:1:0)

Table 2: The extended keywords and their term frequencies (The four numbers after each word are count in abstract, count in highlight, count in keywords list, and count in title respectively)

EXPERIMENTS AND EVALUATION

We collected 22044 academic publications in 18 journals in the area of library and information science from Elsevier. The time span is 50 years from 1963 to 2013. Each document has four main parts which include title, abstract, highlight and keywords list.

Experiments on Personalized Academic Research Temporal Path of Authors

We use all the documents of one author in our data set to get the author's personalized academic research temporal path. Table 1 shows an example of an author's works:

The final extended keywords are listed in Table 2.

The average term frequency in title is 1.0206, 1.8795 in abstract, and 1.4884 in highlight. There are a total of 12706

Keyword	Document Frequency	Weight1
information retrieval	178	4.8122
internet	138	5.0689
telecommunications	136	5.0876
e government	124	5.1761
information technology	108	5.3181
information systems	104	5.3519
knowledge management	87	5.5334
decision support systems	84	5.5683
regulation	78	5.6419
competition	64	5.84061

Table 3: The weight1 values of keywords in top 10 document frequency

Keyword	Weight2
web search	0.5739
multitasking	0.5294
queries	0.2581
classification	0.2295
web search engines	0.2236
web search engine	0.1916
search engines	0.1837
healthcare	0.1811
transaction log analysis	0.1773
transactional log	0.1765

Table 4: The weight2 values of keywords in top 10 average term frequency

keywords collected in the whole collection. Parts of keywords and their weight1 (Formula 5) values are shown in Table 3. For the selected scholar, we get 65 keywords in total. Top 10 are shown in Table 4.

As for the co-occurring weight of keywords, the total number of keywords' pair is 119038. In order to compare the co-occurring weight algorithm in this paper with traditional Directional Affinity (DAff) algorithm, we list a comparison result in Table 5. Given keyword Knowledge Management, the top 5 words selected by our weighting method and top 5 obtained by traditional DAff are listed.

824 co-occurring keywords are included in the final network. In order to reduce the complexity and time of computation, we only use the top 200 keyword pairs with highest weight and temporal paths with top 500 highest weights. Some temporal path results are shown in Table 6.

co-occurred weight		DDA weight	
keyword	weight	keyword	weight
km	0.0117	knowledge sharing	0.1034
knowledge sharing	0.0115	information management	0.0804
knowledge	0.0101	information systems	0.0804
information management	0.0080	knowledge	0.0690
knowledge transfer	0.0069	innovation	0.0575

Table 5: Keywords in top five highest co-occurred weights compared their traditional DDA weights about Knowledge Management

Timing Path	Weight
web search engines/queries/ search engine	56.0288
web search engines/queries/ web search engine	47.7074
web search/queries/ web search engine	45.6189
web search engines/queries/ search topics	35.1557
web search engines/queries/ keyword advertising	35.1557

Table 6: The timing path results with top 5 highest weights

The Visualization of Results

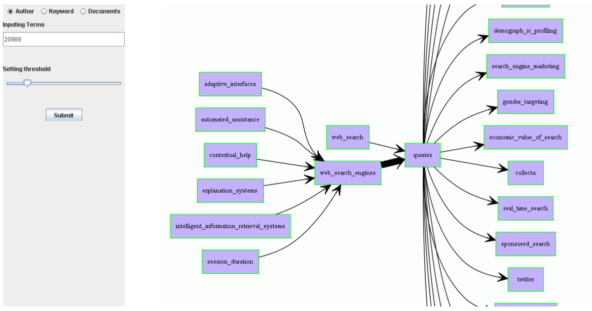
We have implemented a prototype system Personalized Academic Research Temporal Path Discovery for visualizing the temporal path result. This system uses GraphViz as drawing toolkit and provides a good presenting interface for users to interact with the results. It provides three different functions:

1) Query for a specific author. The data is mainly from all the documents of one author in this data set. To resolve author name ambiguity, we design an effective method which can recognize one author based on the similarity of names and addresses. We assign one unique ID to each different author. In this system, a user can choose the Author radio button and set the threshold to limit the number of nodes displayed. The result of experiment in 5.1 is shown as Figure 3.

In Figure 3, the weight of arrowed lines reflects the relationship between two corresponding keywords. We can see the research interests of this scholar shift from Web Search Engine to more specific research topics such as

Topic	Satisfaction	Number	Average Satisfaction	Topic	Satisfaction	Number	Average Satisfaction
information retrieval	3	6	2.125	digital library	3	4	2.286
	2	6			2	1	
	1	4			1	2	
citation analysis	3	5	2.2	others	3	4	1.857
	2	2			2	2	
	1	3			1	1	
information literacy	3	7	2.7	Total	3	26	2.32
	2	3			2	14	
	1	0			1	10	

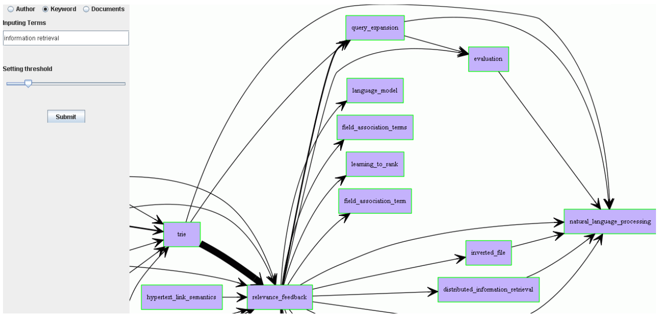
Table 7: The weight1 values of keywords in top 10 document frequency


● Author ○ Keyword ○ Documents

Inputing Terms: 20988

Setting threshold

Figure 3: Querying timing path by author


○ Author ● Keyword ○ Documents

Inputing Terms: information retrieval

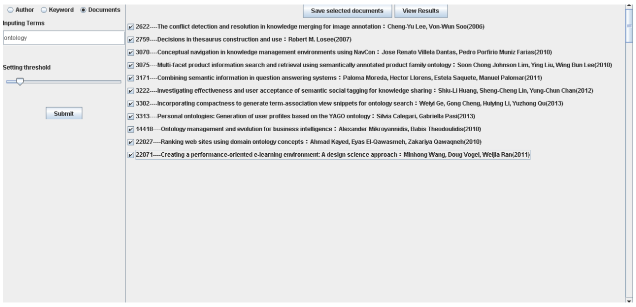
Setting threshold

Figure 4: Querying timing path by keyword

Query (in Web Searching area). And all the detailed research interests are shown in this picture.

2) Users can query the collection using specific keywords by choosing the Keywords radio button. The initial results include all the papers having this keyword. It is a flexible tool for users to explore academic research trends on a certain area.

Figure 4 shows the network generated for Information Retrieval. We can see that the main researches focus of Information Retrieval begin with the retrieval methods for improving performance, and most of them shift to Relevance Feedback. The latest focus of this research area is about Natural Language Processing which is believed to be able to improve the effectiveness of modern information retrieval.


○ Author ○ Keyword ● Documents

Inputing Terms: ontology

Setting threshold

☑ 2622----The conflict detection and resolution in knowledge merging for image annotation：Cheng-Yu Lee, Von-Wun Soo(2006)

☑ 2759----Decisions in thesaurus construction and use：Robert M. Losee(2007)

☑ 3070----Conceptual navigation in knowledge management environments using NavCon：Jose Renato Villela Dantas, Pedro Porfirio Muniz Farias(2010)

☑ 3075----Multi-facet product information search and retrieval using semantically annotated product family ontology：Soon Chong Johnson Lim, Ying Liu, Wing Bun Lee(2010)

☑ 3171----Combining semantic information in question answering systems：Paloma Moreda, Hector Llorens, Estela Saquete, Manuel Palomar(2011)

☑ 3222----Investigating effectiveness and user acceptance of semantic social tagging for knowledge sharing：Shiu-Li Huang, Sheng-Cheng Lin, Yung-Chun Chan(2012)

☑ 3302----Incorporating compactness to generate term-association view snippets for ontology search：Weiyi Ge, Gong Cheng, Huiying Li, Yuzhong Qu(2013)

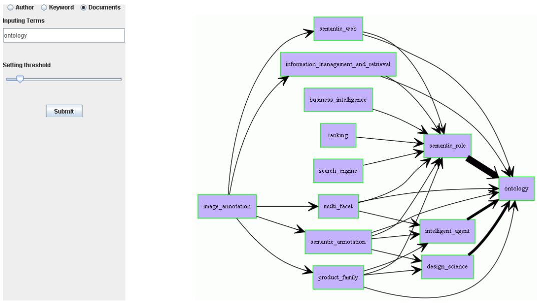
☑ 3313----Personal ontologies: Generation of user profiles based on the YAGO ontology：Silvia Calegari, Gabriella Pasi(2013)

☑ 14418----Ontology management and evolution for business intelligence：Alexander Mikroyannidis, Babis Theodoulidis(2010)

☑ 22027----Ranking web sites using domain ontology concepts：Ahmad Kayed, Eyas El-Qawasmeh, Zakariya Qawaqneh(2010)

☑ 22071----Creating a performance-oriented e-learning environment: A design science approach：Minhong Wang, Doug Vogel, Weijia Ran(2011)

Figure 5: The documents about Ontology fitting one user's information need


○ Author ○ Keyword ● Documents

Inputing Terms: ontology

Setting threshold

Figure 6: Querying timing path by documents

3) A user can also query for any group of documents which reflects the user's current interests. In other words, a user can formulate his/her own document collection and generate a temporal semantic network of the collection. For example, a user can save some documents about Ontology as shown in Figure 5:

Figure 6 shows the main trend of research topics about ontology discussed in these papers.

Preliminary User Evaluation and Future Work

We conduct a pilot evaluation to measure users' satisfaction with the system. 10 users participate the test. Each user is asked to use the system to do five queries in his/her interested research area. After reviewing the results, they are asked to rate their satisfaction using a 3 point Likert scale in which 3 is satisfied and 1 is not-satisfied. After reviewing the queries submitted by the users, we classify them into five groups based on their broad topic area. They are information retrieval, citation analysis, information literacy, digital library, and others. The results are summarized in Table 7.

It is interesting to notice that the average satisfactions in research areas of library science are often higher than the average satisfactions in other sciences such as information science. This may be resulted from the document types of this experimental data set. The types of journal we have chosen are mainly about library science so that the corresponding effective number of documents and keywords bring in this difference in evolution results. But we see that average satisfaction in total is still 2.32 which tell us these results can meet users' information need better.

For future work, we plan to expand our data collection in library and information studies and conduct a more formal user evaluation. We will also test the application of the system in other research areas.

ACKNOWLEDGMENTS

This work has been supported by Chinese National Natural Science Foundation 71103081 (2011), and Natural Science Foundation of Jiangsu Province 11KJB630001 (2011).

REFERENCES

Al Bawab, Z., Mills, G. H., & Crespo, J. F. (2012). Finding trending local topics in search queries for personalization of a recommendation system. In P. of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 397-405). ACM.

Alkilany, A. (2013). An overview: temporal-side of sequential patterns discovery. International Journal Data Mining & Knowledge Management Process 3(1),1-12.

Borth, D., Ulges, A., & Breuel, T. M. (2012). Dynamic vocabularies for web-based concept detection by trend discovery. In P. of the 20th ACM international conference on Multimedia (pp. 977-980). ACM.

Chen, L., Lv, Z., & Liang, L. (2009). Visualizing scientific frontiers and emerging trends in the branches of mechanics. Journal of the China Society for Scientific and Technical Information 28(5), 736-744.

Hong, N., & Zhang, Z. (2012). A method for detection of latent burst words based on decision tree. Journal of the China Society for Scientific and Technical Information 31(3), 228-241.

Kim, G., & Xing, E. P. (2013). Time-sensitive web image ranking and retrieval via dynamic multi-task regression. In P. of the Sixth ACM International Conference on Web Search and Data Mining (pp. 163-172). ACM.

Leprovost, D., Abrouk, L., Cullot, N., & Gross-Amblard, D. (2012). Temporal semantic centrality for the analysis of communication networks. In P. of Web Engineering (pp. 177-184). Springer Berlin Heidelberg.

Li, D., Li, C., & Li, J. (2011). Current status and trends of information systems research in china:2000 ～ 2009. Journal of the China Society for Scientific and Technical Information 30(11), 1209-1218.

Li, S. (2012). Research on automatic construction of domain ontology in library and information science based on weighted co-occurrence of citation keywords. Journal of the China Society for Scientific and Technical Information 31(4), 371-380.

Li, S. (2013). Recognition of scholars' main research interests and implementation of personalized foreign documents recommendation service based on three-word co-occurrence analysis. Journal of the China Society for Scientific and Technical Information 32(6), 629-639.

Li, S., & Bai, Y. (2011). The analysis of research trend in library and information science based on hotspot recognition of timing keywords. New Technology Of Library And Information Service 27(5), 69-76.

Li, S., & Lv, X. (2012). The matching algorithm of heterogeneous user personalized profile based on centripetal spreading weighted xml model. New Technology of Library and Information Service 28(5), 32-40.

Li, S., Xu, X., Qian, G., & Han, W. (2012). A method for automatic recognition and visualization of main-paths in academic documents based on vibration algorithm and domain ontology. Journal of the China Society for Scientific and Technical Information 31(7), 676-685.

Li, Y., & Hou, H. (2007). Study on visualization of citation analysis. Journal of the China Society for Scientific and Technical Information 26(2), 301-308.

Li, Y., Ren, Y., He, L., & Du, H. (2010). Mining the regularity in disciplinary development through citati on timeline Web visualization. Journal of the China Society for Scientific and Technical Information 29(5), 880-888.

Ma, F., Lv, P., & Ye, M. (2012). Study on global science and social science entropy research trend. In P. of Advanced Computational Intelligence (ICACI), IEEE Fifth International Conference (pp. 238-242). IEEE.

Osareh, F. (1996). Bibliometrics, citation analysis and co-citation analysis: A review of literature I. Libri 46(3), 149-158.

Radinsky, K., Svore, K., Dumais, S., Teevan, J., Bocharov, A., & Horvitz, E. (2012). Modeling and predicting behavioral dynamics on the web. In P. of the 21st International Conference on World Wide Web (pp. 599-608). ACM.

Wang, J., Xu, C., & Geng, X. (2009). Study on research topic evolution based on probabilistic graphical models. Journal of the China Society for Scientific and Technical Information 28(3), 347-355.

Wang, X., Wang, Z., & Xu, S. (2012). Tracing scientist's research trends realtimely. ArXiv Preprint: 1208.1349.

Wang, Y., Song, S., Lu, N., & Zhu, Y. (2007). Application of co-occurrence analysis in text knowledge mining. Journal of Library Science in China 33(2), 59-64.

Xia, L., Cheng, X., & Gui, S. (2012). Clustering and multidimensional visualization of co-occurrence query keywords in e-government platform. Journal of the China Society for Scientific and Technical Information 31(4), 352-361.

Xu, X., & Li, B. (2011). Recognition of time expressions based on conditional random fields and rules. Journal of the China Society for Scientific and Technical Information 30(10), 1065-1071.

Yin, C., Hirokawa, S., Yau, J., Hashimoto, K., Tabata, Y., & Nakatoh, T. (2013). Research trends with cross tabulation search engine. International Journal of Distance Education Technologies 11(1), 31-44.

Yin, S. (2008). Analysis of the methods for detecting emerging trend. Information Science 26(4), 536-540.

Zhang, C., & Li, Y. (2010). Detecting hotspot and trend of disciplines using topic clustering. Journal of the China Society for Scientific and Technical Information 29(2), 342-349.

Zhang, X., & Li, B. (2011). Time information extraction of event detection and characterization. Journal of the China Society for Scientific and Technical Information 30(4), 395-401.