Scientometrics
DOI 10.1007/s11192-015-1571-0

A new method for automatically constructing domain-oriented term taxonomy based on weighted word co-occurrence analysis

Shuqing Li[1] · Ying Sun[2] · Dagobert Soergel[2]

Received: 10 October 2014

Abstract The automatically construction of term taxonomy can enhance our ability for expressing the science mapping. In this paper, we introduce the definition of weighted co-occurring word pair and corresponding improved method of word co-occurrence analysis. An application and evaluation of this proposed method in the library and information science is also discussed, which includes how to get the expanded effective keywords, how to calculate the weight of keywords and their relations, and how to abstract the hierarchical structures and other relations such as synonyms and etc. A visualization tool and a prototype search system are designed for browsing the term taxonomy identified. Finally, we report the experiment of evaluation and comparison. The experiment results prove that this proposed method in helping users doing semantic searches and expanding their searching results is effective and can meet the requirement of some specific domains.

Keywords Word co-occurrence analysis · Term taxonomy · Semantic search

Introduction

Along with the rapid increasing of scientific literatures, scientists have being challenged with the ability to find and digest exactly the right parts of the literature according to the specific requirements in some research domains. In many research areas of knowledge expression and data sharing, science mapping has been gaining more and more attentions from all kinds of researchers, especially when facing the big data in the Web pages and social network. Science mapping is put forward as an important method to handle these problems in the field of bibliometric (Van-Eck and Waltman 2010;

✉ Shuqing Li
leeshuqing@gmail.com

[1] College of Information Engineering, Nanjing, Jiangsu, China

[2] Department of Library and Information Studies, Graduate School of Education, University at Buffalo, The State University of New York, Buffalo 14260-1660, USA

Published online: 27 March 2015

Cobo et al. 2012). It is very promising to discover the hidden important knowledge and understand it comprehensively (López-Herrera et al. 2010). It can also express the meaning and semantic relations of terms more effectively (Zhang et al. 2011). Generally science mapping analysis can be divided into many different steps: data retrieval, preprocessing, network extraction, normalization, mapping, analysis and visualization (Börner et al. 2003). The method proposed here mainly focuses on three of these steps: preprocessing, network extraction and normalization. We also present a simple visual prototype for it.

The key in constructing science mapping is how to understand and express the meaning of terms and their semantic relations, which is called term taxonomy in this paper. Various methods have been explored by many researchers. However, there has been not a fully satisfied method by now (Nickerson et al. 2013). In fact, it is challenging to build term taxonomy with the full semantic information and well-organized structure in which the hierarchical semantic descriptions can be used for deducing wider or narrower meaning of the current terms, or to combine the temporal sequence analysis and visualization with which users can more easily discover the satisfied results. The main reason lies in the complexity of the natural languages, the broad scope and huge amount of scientific knowledge accumulated. The second challenge is how to construct the taxonomy automatically and efficiently. The most common method used currently is manual selection of terms with adding annotation as extra preprocesses. Some studies use human experts to assign the relations of different terms. However, it is always time consuming and more complex. Another challenge is the problem of time lag. We cannot keep updating taxonomy all the time while the new terms and their new relations are emerging continuously. As a feasible choice, the domain-oriented method has been widely proposed and well developed recently. In this way, the size of data can be in our control since it only aims at some special domains. These existing researches also give us a start point for our further research. We want to design an effective method for constructing the domain-oriented term taxonomy which we think will have the wider applicability than some artificial ones or provide a solid foundation for further necessary manual editing. The domain we choose is in library and information science.

The paper is organized as follows. We review the related studies in the next section, followed by the introduction to our method which include how to extract and weight different terms, how to weight term relations, how to identify the hierarchical relation and the highly related words. Then a pilot study on the field of library and information studies is reported to illustrate the proposed method. Finally the paper concludes with some preliminary evaluation results.

Related work

Machine learning and statistical methods are popular methods used to construct term taxonomy automatically or semi-automatically from the existing data sources. Based on the data sources, we can divide these means into two groups. The first group is to extract terms and their relations from some existing structural language corpuses such as dictionaries, knowledge bases (Buitelaar et al. 2009; Chen et al. 2009). This method usually has relatively high accuracy because dictionaries are well-organized and often with higher

quality. The computing complexity requirement is also not high. However, the biggest limitation of this method is the availability of data sources. It is too hard indeed to get high-quality dictionaries in some specific domains. The other limitation is the low extensibility of the constructed taxonomy. The second group is to extract the information from the raw texts directly. Comparing to the first group, most of existing researches in this group rely heavily on the natural language processing (NLP) technologies. The more popular methods include model-based methods, association-rules-based methods, concept-cluster-based methods, and so on. Without the limitation of the predefined term sets like dictionaries, a taxonomy built in this way has better extensibility. However, NLP methods usually involve the higher computing complexity, and often need the intervention and supervision of domain experts, even need experts to prepare some information corpus of term combination rules (Lim et al. 2009).

We propose here a method based on word co-occurrence analysis to automatically construct the taxonomy from the raw texts directly. We believe it is an efficient way keeping the balance between the effectiveness of construction and the extensibility of taxonomy. Of course, the idea of using word co-occurrence in taxonomy construction is not new. Soergel (1974a) describes a method on discovering concepts and their relations based on word co-occurrence analysis, which is a typical one of earlier researches. Fellbaum (1998) brings forward a valuable idea which can automatically extract the concepts and their relations based on the synonyms mode. More scholars begin to realize and reach a consensus that it is possible and feasible to construct term taxonomy with the word co-occurrence analysis (Hadzic and Chang 2005). For example, Morita et al. (2004) implement a DODDLE-OWL project with the specific lexicon based on word co-occurrence methods. Other similar works of constructing knowledge banks with word co-occurrence analysis can be also seen in the research area of knowledge management (Wang et al. 2006). Benz et al. (2010) describe an algorithm to obtain self-organized taxonomy (folksonomy) from the social tagging data using tag co-occurrence analysis.

Just as co-citation analysis and co-article analysis, word co-occurrence analysis is one of the co-occurrence analysis methods. Word co-occurrence analysis can quantitatively measure the relations of different words based on the co-occurrence rules of these words in the same document (Bordag 2008). Its methodological basis is the neighboring connecting rule, the principle of knowledge structure and mapping in psychology. The common process of word co-occurrence analysis often begins with a set of topic-related keywords extracted from the document-related fields such as title or abstract in the document collection. Then the associating degree of word pairs can be measured. Different effective methods can also be used to differentiate the relative term frequency and term distance (Geng and Geng 2006). These measuring algorithms often combine some statistics methods such as Dice Index, Cosine Index, Jaccard Index and etc (Egghe and Leydesdorff 2009). But the most important and fundamental method is still TF-IDF proposed by Salton (1983) and Jones (1972).

The basic assumption of the word co-occurrence analysis is that two or more words relate with each other in semantic meaning if they co-occur in the same part of the same document, or vice versa (Peat and Willett 1991). But how do we measure the association degree between two co-occurring words? In fact, the high association between two words has many different meanings. For example, it can reflex the definitional relationship, or contextual contiguity, or even two terms from a multi-word term.

We should add some constraints to enhance the expressing ability of this method. For example, we should use the full texts instead of sentences in order to get the synonymy interpretation. But we should use the sentences as computation base if we want to make sense the interpretation as a multi-word (Soergel 1974b). As for algorithms, there are so many means we can use. In the earlier researches, Doyle (1962) compares some correlation coefficients and thinks it will be more meaningful if considering which applications use it. He also points out the importance of applying words co-occurrence analysis in constructing the association map of terms which can improve information retrieval system. It also can be displayed in such a way that users may select all appropriate terms rapidly and accurately (Gillum 1964). Callon et al. (1991) propose Equivalence Index which can measure the strength of association between different terms. The recognition of main topic can be achieved by term clustering based on the weighting analysis of links between terms.

Some other various direct similarity measures are being used in the literature. Van Eck and Waltman (2009) extensively analyze a number of well-known direct similarity measures and argue that the most appropriate measure for normalizing co-occurrence frequencies is association strength. Since many methods have been put forward recently, we can adopt some basic ideas from these researches for designing our new algorithm based on words co-occurrence analysis which will be explained in this paper.

All the co-occurring words then can be connected into a network based on the co-occurrence relation (Sheng and Li 2009). Yu and Zhou (2010) define a network of word co-occurrence which vertexes denote word entities and edges denote the co-occurrence relation of corresponding words. Obviously, this definition and description has wider commonality and this conclusion can be applied into other analyses and researches such as clustering analysis and automatic question answering system (Zhong et al. 2009; Xu et al. 2012; Lu et al. 2012). In order to enhance effectiveness of words co-occurrence analysis, many scholars begin to exploit it in more research areas. For example, some scholars explore the regularity of adjacent co-occurrence between semantic relations in the objective knowledge system but ignore the effect of part of speech to adjacent co-occurrence intensity (Qiu et al. 2012). Zhang et al. (2011) expand word co-occurrence analysis with authors' affiliation co-occurrence analysis. As for the design of weight assignment for co-occurring word, traditional methods do not distinguish the difference of co-occurrence of same words in the different documents and often only use term frequency as basic measure. Currently researchers begin to pay more attention to the relative word co-occurrence, that is to say, the conditional probability of word B in a document which has word A is not as same as conditional probability of word A in a document which has word B. This asymmetric weight assignment can help discovering more information in some specific applications. The popular methods of asymmetric weight assignment include bilateral similarity measure (BSM), multilateral similarity measure and Pearson's correlation coefficient (Choi et al. 2010). We choose BSM since it is mainly used for analyzing relations of two words. This measure includes Salton index, Jaccard coefficient and etc. (Cobo et al. 2012). Directional Affinity (DAff) method is also an effective method in which the DAff between terms A and B may be defined as the conditional probability of observing B, given that A was observed in a co-occurrence context. DAff may be the number of co-occurrence contexts that include both terms A and B, over the number of co-occurrence contexts that include term A (Labrou et al. 2012). In this paper, we use a revised DAff which emphasizes the bilateral relations of different words so that it is more suitable to meet the requirement of asymmetric weight assignment.

The construction of term taxonomy

We propose a two-step method to construct term taxonomy. These two steps are the extract of concepts and the extract of their relations.

Extracting and weighting concepts

We believe the set of keyword in academic publications is a good data source from which the effective concepts can be extracted. Just as what we have discussed above, since using only term frequency as the measure of word importance has its limitation, we combine term frequency with some other measures of word importance to enhance the effectiveness of expressing meaning. This step is very essential because accurate measuring the importance of words in the documents can also contribute to accurately recognize the relations of different words.

In order to use keywords as the source for concepts and further to discover the word relations, we have to expand the number of keywords in each document firstly. In reality, the number of keywords is usually limited to 3 or 4 in each academic article, which is not so many enough to analyze word co-occurrence. However, some words in the title or abstract of a document, even though not listed as the keywords of that document, may be used as the expanded keywords in this document. Just like Martínez et al. (2014) and Murgado-Armenteros et al. (2015), the method to add new keywords from the documents is something necessary which is usually carried out in the preprocessing step. We can expand the keyword list of each document by searching every keyword from the whole keywords collection in each document and add the word to the keywords list of the document if it also appears in the title or abstract. An occurrence of a keyword is weighted differently based on its position.

We design a weighting method based on the idea of TFI/DF. Two components are included in each keyword's weight. The first one reflects the importance of a keyword itself, the higher the keyword document frequency, the less its power in expressing the characteristics of a document. So we assign the first part of each keyword's weight as Formula 1 shown:

$$\text{Weight1OfWord}_{\text{keyword}i} = \log(N/\text{DF}_{\text{keyword}i}) \tag{1}$$

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 impact of high value.

The second part reflects the importance of a keyword in a document. Indeed, setting a different weight to the terms in the different parts of the document is so common that Google also use this similar approach in its PageRank algorithm. We assign the different weight coefficients to the different fields in which the keyword appears. Then, we calculate the second weight of keyword_i in doc_j as Formula 2 shown:

$$\begin{aligned}\text{Weight2OfWord}_{\text{keyword}i,\text{doc}j} = {} & \text{TFInAbstract}_{\text{keyword}i}, \text{doc}_j \times \text{coff}_{\text{abstract}} \\ & + \text{TFInKeywordslist}_{\text{keyword}i}, \text{doc}_j \\ & \times \text{coff}_{\text{keywordlist}} + \text{TFInTitle}_{\text{keyword}i},\ \text{doc}_j \times \text{coff}_{\text{title}}\end{aligned} \tag{2}$$

The TFInAbstract, TFInKeywordslist and TFInTitle mean the keyword frequency in the abstract, keyword list and title field respectively. In this study, we assign the field weight coefficients as follows: the weight coefficient of abstract field is 1, keyword list field is 2, and title field is 4. This assignment is based on the ad-hoc observation. The number of TFInAbstract is more likely larger compared with other fields so that the corresponding weight coefficient is set low. TFInKeywordslist only has two values which are 0 and 1. When a document has this keyword in its original keyword list, TFInKeywordslist is 1, otherwise is 0 and the corresponding keyword is an expanded keyword. The coefficient of TFInTitle is set highest given the short length of the field.

The final weight of one keyword in a document is shown as Formula 3:

$$\text{weightOfWord}_{\text{keyword}i,\text{doc}j} = \text{Norm}(\text{weight1OfWord}_{\text{keyword}i} \times \text{weight2OfWord}_{\text{keyword}i,\text{doc}j}) \tag{3}$$

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

Weighting word relations

Word relations include hierarchical relations and other relations. More complicated relations exist between three or more words, and can have a network-like structure.

As for hierarchical relations, we need to identify the hypernyms and hyponyms of each keyword. The basic idea in our proposed method is also based on word co-occurrence analysis. In general, we can conclude that the more two word co-occur together, the closer their relation is. However, only considering the frequency of word co-occurrence is not good enough to draw conclusions of word relations. An occurrence of two words co-occurring in the title field can indicate the closer relationship than an occurrence of two words co- occurring far away from each other in a long document. We should assign a measurable weight to each word pair according to this difference.

The final weight of word relation is calculated as Formula 4 shown:

$$\text{weightOfRelation}_{\text{keyword}i,\text{keyword}j} = \frac{\sum_k \text{weightOfWord}_{\text{keyword}i,\text{doc}k} \times \text{weightOfWord}_{\text{keyword}j,\text{doc}k}}{\sum_k \text{weightOfWord}_{\text{keyword}i,\text{doc}k}} \tag{4}$$

Formula 4 is quite similar to Association Strength, but in this case it uses the sum of weights instead of term frequency. The numerator denotes the sum of all the weights' products of each word in word pairs and the denominator denotes the sum of words' weight. This formula is based on standard DAff method which is a traditional method for calculating the weight of co-occurring keywords. But standard DAff ignores the importance of keywords themselves and only pays attention to the co-occurrence of words at the document level. For example, keyword A and B co-occur in one document, and both of them appear only once in this document, keyword C and D also co-occur in only one document, but each of them appear 10 times in the document, we cannot distinguish these two keyword pairs using DAff because their document frequencies all equal to 1. But using this new weighting strategy, we can get more accurate relation of keywords since we have combined the weight of word importance into the calculation. We call this word pair as

weighted co-occurring word pair. We can get three types of special word pairs after this calculation. The first type includes those composed of the same word. They are no sense at all and should be deleted from the collection of word pairs. The second type is identical word pair. We aggregate their weights and use the summed weights to denote their final effectiveness. The third type is reversal word pair. That is to say, there will be a word pair of B and A if there is a word pair of A and B. Since the relation weight in Formula 4 is not symmetrical, we can keep all the pairs now and will decide which type should be kept finally according to the requirement of application.

Identification of hierarchical relation

Term taxonomy may include many different structures. Hierarchical structure is the most popular and important one, which has many applications such as semantic analysis and spreading activation in the personalization algorithm. Our method of identifying hierarchical relation includes two steps. In the first step, we construct the hierarchical structure by keeping only a half of all the word pairs in which the first word's frequency is greater than the second word's so that the word pairs left are ordered pairs. The basic idea lies in a common rule that the more general a word is, the more frequently it should appear in documents (Soergel 1974b). We then combine all the word pairs left to a hierarchical structure according to whether the second word in one word pair is identical to the first word in another word pair. In the second step, we try to eliminate as much unnecessary repetition as possible. Since some different hyponyms might have one same hypernym, we only keep its at most top n hypernyms with the highest relation weight for each word. Our experiment sets n as 5. Of course, this parameter n can be adjusted according to the requirement of applications.

Identification of highly related words

Highly related words are those words with the strong non-hierarchical relation such as synonyms, antonyms, abbreviations and etc. Adding these highly related relations to the hierarchical term taxonomy can produce a network-like structure. This network structure allows the ability to retrieve the related concepts not only through the up-and-down hierarchical structure, and but also through the relation links transversely. It will give us more traversal abilities in the term taxonomy.

There have been many algorithms proposed and tested for acquiring the highly related words and most of them are based on the form analysis and pattern similarity of words. These methods looks like good ways, however, they actually have many problems. For example, some synonymous words may be different greatly in their forms such as 'DBMS' and 'database', and some non-synonymous words may have similar forms such as 'diary' and 'dairy'. Here the method we proposed is based on the assumption that two words should be highly related if their co-occurrence rules with other words are identical or similar. We believe that the sequence of weighted co-occurring word pairs of two highly related words with other words should have the similar characteristics when ordered by the relation weight.

The algorithm is shown in the pseudo-code below:

```
BEGIN:
Input: word wordX and wordY
Output: the sequence similarity of weighted co-occurring word pairs of wordX and wordY

// get two sequences of weighted co-occurring word pairs of wordX and wordY
collectionX←getWeightedWordCooccurrenceSequence(wordX)
collectionY←getWeightedWordCooccurrenceSequence(wordY)

// order the co-occurring words on descending related weight in each sequence
orderByWeightDesc(collectionX)
orderByWeightDesc(collectionY)

// summarize the similarity of two sequences
// read each word in one sequence
For each wordI in collectionX {
    // get all words before current word in one sequence
    wordsBeforeInX←getWordsBefore(wordI, collectionX)

    // get all words before current word in another sequence for comparison
    wordsBeforeInY←getWordsBefore(wordI, collectionY)

    // read each word before current word in one sequence
    For each wordJ in wordsBeforeInX{
        // increase similarity if current word also exists in another sequence
        If existed(wordJ, wordsBeforeInY) Then
            similarity←similarity+1
    }
}

// normalize similarity value within 0 and 1
normalize(similarity)
END
```

Experiments

We have collected 28,848 academic publications in 19 journals in the area of library and information science from Elsevier and JASIST. These documents include articles, reviews, and letters published in these journals. The time span is about 60 years from 1950 to 2013. Each document has three parts: title, abstract and keyword list. The detailed information is summarized in Table 1.

We invite 20 users to evaluate the results of experiment. All of them major in library and information science with the rich comprehension of information retrieval.

Table 1 Collected journals and their articles' amount

Journal	Amount of articles
Journal of the American Society for Information Science and Technology	6804
Information Processing and Management	3327
Telecommunications Policy	2616
The Journal of Academic Librarianship	2512
Information and Management	2134
Government Information Quarterly	2077
International Journal of Information Management	1864
Government Publications Review	1382
Library and Information Science Research	1073
International Library Review	1053
Journal of Government Information	1050
Journal of the American Medical Informatics Association	733
The International Information and Library Review	659
Information Storage and Retrieval	454
Government Publications Review (1973)	442
Social Science Information Studies	319
Government Publications Review. Part A	240
Government Publications Review. Part B	70
Journal of King Saud University	39

Table 2 Keywords of the top 10 highest weights

Keywords	DF	Weight1OfWord
Information retrieval	196	4.9904
Internet	139	5.3327
Telecommunications	136	5.3566
e Government	124	5.4467
Information technology	108	5.5872
Information systems	105	5.6131
Knowledge management	92	5.7462
Decision support systems	84	5.8377
Regulation	78	5.9108
Evaluation	64	6.1092

Extracted concepts

We get 12,815 original keywords provided by authors in all the documents of this collection. Although this number is huge, the total number of keywords with document frequency greater than 1 is only 2782, 21.7 percent of the total. The total number of co-occurring keywords is 24,111 before expansion operation and 347,984 after expansion operation. The keywords with most document frequency are often related with 'Information' since the information is the main topic of this research area. Table 2 lists the top 10 keywords with the highest weight.

We also calculate the weight2OfWord of all keywords. For example, we choose one article "Image Retrieval from Scientific Publications: Text and Image Content Processing to Separate Multi-panel Figures" published in Journal of the American Society for Information Science and Technology, 2013. There are three original keywords in this paper, which are 'Automatic Indexing', 'Information Retrieval', and 'Image Retrieval'. After expansion operation, we get more related keywords and their weights shown in Table 3.

We distinguish the keywords from their substring and reduce the unnecessary duplicate counts. For example, we still assign 1 as the term frequency of 'Image' since the first 'Image' is just the substring of another keyword 'Image Retrieval', although 'Image' appears two times in the title. We can see the ability of expressing the content of a document has been enhanced through this expansion operation and these more words provide a good foundation for us to further explore the rules of words co-occurrence. We also calculate the relation weight of each word pair. Table 4 shows the top 10 words related to 'Information Retrieval' with the highest weights.

We use Prefuse toolkit to design a visual interface to demonstrate the relation of all words. Users can adjust the slider in the right to display wanted results. The Nodes slider can limit the number of nodes and Relation slider can limit the number of edges between nodes. When we set the Node slider to its max value and adjust the Relation slider to a suitable level, we can see the core of this data set which composes of many important keywords and their relations. It is apparent that the core has two clusters. One cluster is mainly about 'Telecommunications', 'Competition' and 'Regulation'. Another cluster is

Table 3 Expanded keywords with the descended order of weights

Keywords	Title	Abstract	Keywords list	TF	Weight2OfWord
Image	1	2	0	6	0.2143
Publications	1	2	0	6	0.2143
Image retrieval	1	2	1	8	0.2069
Content	1	2	0	6	0.1807
Images	0	2	0	2	0.0714
Indexing	0	2	0	2	0.0547
Automatic indexing	0	0	1	2	0.0536
Semantic	0	1	0	1	0.0357
Task	0	1	0	1	0.0357
Method	0	1	0	1	0.0357
Output	0	1	0	1	0.0357
Pattern	0	1	0	1	0.0357
Automatic segmentation	0	1	0	1	0.0357
Captions	0	1	0	1	0.0357
Information retrieval	0	0	1	2	0.0347
Retrieval	0	1	0	1	0.0319
Systems	0	1	0	1	0.0301
Research	0	1	0	1	0.0301
Recall	0	1	0	1	0.0289
Precision	0	1	0	1	0.0285
Development	0	1	0	1	0.0271

Table 4 The top 10 words related to 'Information Retrieval'

Co-occurring keyword	Relation weight
Information	3.762E−2
Retrieval	3.304E−2
Search	1.692E−2
Query	1.471E−2
Model	1.334E−2
User	1.263E−2
Use	1.250E−2
Research	1.204E−2
Method	1.169E−2
Systems	1.145E−2

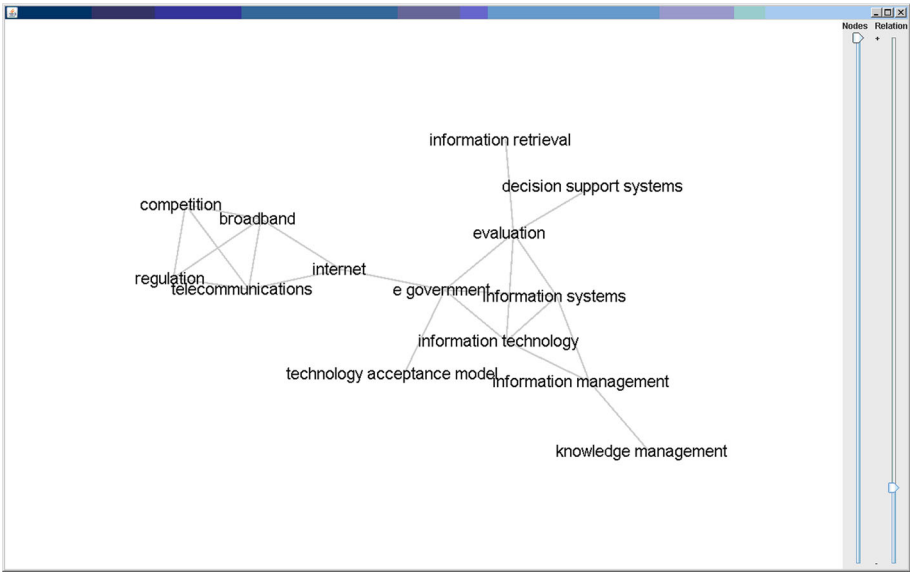

Fig. 1 The core of keywords collection

mainly about 'Information System' and 'Information Technology'. Two important keywords, 'Internet' and 'E-government' connect these two parts together (shown in Fig. 1).

Term taxonomy and its visualization

After removing the unnecessary repetition of keywords in the term taxonomy, the final total number of words in the term taxonomy is 12,815 and the final total number of hierarchical relation is 12,814. The node with the highest frequency is 'Information Retrieval' which has 5 hyponyms such as 'Evaluation', 'Database', 'Cross Language Information Retrieval', 'Relevance Feedback', and 'Interactive Information Retrieval'. Some more detailed information of other keywords in the term taxonomy can be shown in Figs. 2 and 3.

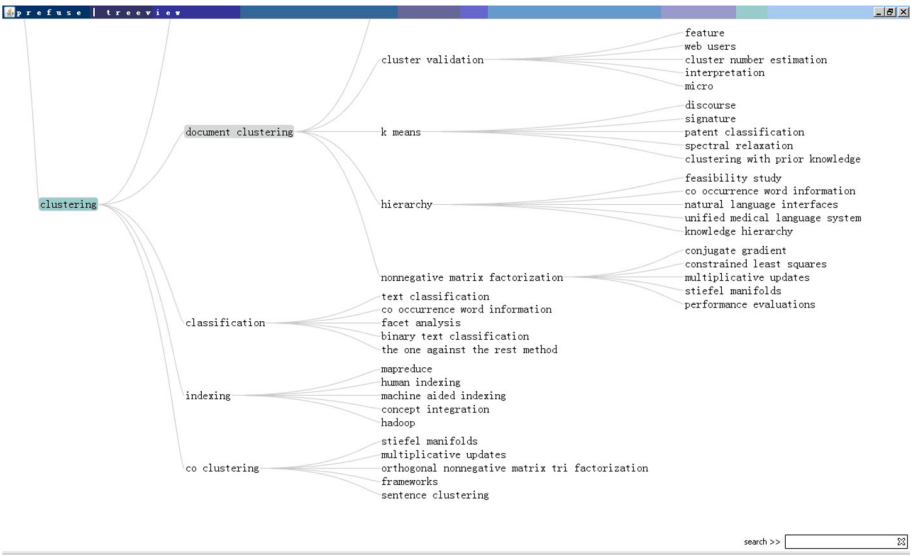

Fig. 2 The hyponyms of term 'Clustering'

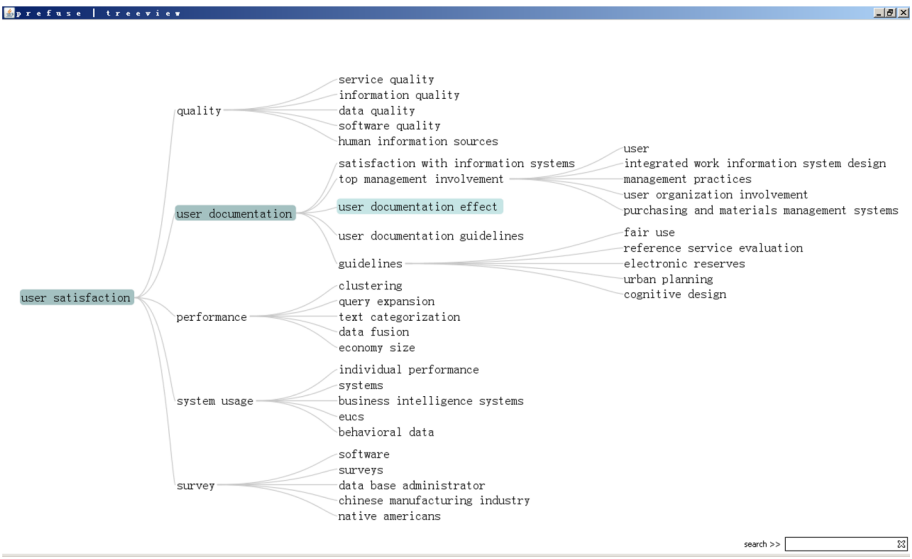

Fig. 3 The hyponyms of term 'User Satisfaction'

From these figures, we can see that the hierarchical relation includes many semantic meanings. Some are the part-whole relation such as 'Clustering' and 'Document Clustering', 'Co-clustering', etc. Some can reflect the relation of contextual contiguity such as 'Clustering' and 'Indexing', 'Classification' and 'Facet Analysis'. Some are class-inclusion or topic-inclusion relation such as 'Indexing' and 'Map-Reduce', 'Hadoop'. All of these

can be interpreted as the hierarchical association between words and they can help users to refine queries in the information retrieval systems. However, we also see something not very satisfactory in them. For example, the relation of 'Survey' and 'Software' is not typically hierarchical. The reason mainly lies in the size and content of our dataset. The number of some keywords is not so many enough to support getting a reasonable conclusion. It is obvious that the more keywords we have, the more reasonable its result becomes. For example, the frequencies of keyword 'Clustering' and 'Document Clustering' are 22 and 14 respectively, while the frequencies of keyword 'Survey' and 'Software' are 17 and 12 respectively.

Highly related words and their results

We use the method of extracting highly related words to get more word relations. Some relations with the highest weights are shown in Table 5.

In this experiment, our results show some interesting findings. The first one is that the relations with the highest weight often are symmetry. That is to say, if word A and word B have a strong relation in this highly related relation, word B is often related with word A with same weight. All the records in Table 5 have these characteristics. The second one is that these relations mainly include synonyms, antonyms and acronyms. All of these findings reflect the effectiveness of this algorithm.

Table 5 The top 20 highest weights of keyword relations

Keyword1	Keyword2	Weight
Intangible assets	Intellectual capital	0.8426
Cim	Computer integrated manufacturing	0.7685
Decentralization	Centralization	0.7500
Precision	Recall	0.7222
Strategic information systems	information systems planning	0.7130
Bpr	Business process reengineering	0.6389
First responder	Public safety	0.5926
Rfid	Radio frequency identification	0.5741
Information systems planning	Strategic information	0.5648
Edemocracy	Eparticipation	0.5093
Accessibility	Disability	0.4537
Executive information systems	Executive support systems	0.4444
Public safety	Public private partnership	0.4259
It infrastructure	Integration	0.4167
Perceived ease of use	Perceived usefulness	0.3889
Bpr	Reengineering	0.3796
Graphs	Web	0.3704
Crm	Customer relationship management	0.3333
Graphs	Web pages	0.3333
Recommender systems	Collaborative filtering	0.3241

User evaluation of experiments

It is not a trivial task to judge the similarity between a constructed concept hierarchy and reference hierarchies, especially regarding the absence of well-established and universally accepted evaluation measures. Two crucial questions need to be solved: first, which suitable reference taxonomy to choose, and second which measure to use to compute the similarity between the constructed and the reference taxonomy.

At first, we want to evaluate by comparing our constructed taxonomy and existing taxonomies about library and information science. However, it is so hard to get a satisfied and available one. Most of these taxonomies are just mentioned in the papers and cannot be easy to access publicly. We have to consider other available taxonomies. Although WordNet is a structured lexical database of the English language (Miller 1995), it is not specialized for library and information science. Meantime, it can only be searched by a single word and does not provide the search function of one term including many words which are very common in our database. We choose Wikipedia as the final reference taxonomy since it is the world's largest collaboratively built online encyclopedia. The related terms can be acquired from the content in Web articles and hyponyms can be acquired from its category hierarchy. We only evaluate hyponyms not hypernyms since it is impossible to get the full hypernyms of one term. In fact, better hypernyms can also be got from the taxonomy if the structure of its hyponyms is better. However, there are still some limitations of using Wikipedia because its included fields are also so wide and not very specialized for library and information science. Some specialized term cannot be found in the content of Web articles, especially in the category. For this reason, we adopt two complementary methods to conduct the full evaluation. Firstly we use Wikipedia as reference taxonomy to evaluate occurring terms. For those terms which cannot be found in Wikipedia, we then measure their semantic relevance with their corresponding concepts in Wikipedia based on the search results of Google and Google Scholar. The full evaluation of entire taxonomy can be achieved in this way.

Secondly, how to select evaluation measures and judge the similarity between them is also essential for this evaluation. More two taxonomies overlap, more similar they should be. Dellschaft and Staab (2006) propose two measures, namely taxonomic precision and taxonomic recall for this purpose. The basic idea is to find a concept present in both structures and then to extract its corresponding sub-concepts and super-concepts from both structures. If both excerpts are very similar, we can say that these structures themselves are judged to be similar. In our experiment, we only consider the precision not recall because we cannot know what the full related terms of one term should be, especially in Wikipedia.

For example, the calculation of precision of 'Information Retrieval' is shown in Table 6.

Each term in hyponyms and highly related terms will be checked whether it occurs in the corresponding part of Wikipedia. In this way, all the precisions can be calculated.

We collect the evaluation result of 134 terms in the taxonomy, which include 67 evaluations of hyponyms and 67 evaluations of highly related terms. The total precision is 0.3276. The precision of hyponyms is 0.2507 and the precision of highly related terms is 0.4045. The reason that the second value is higher than the first one is because the category in Wikipedia is not very complete so that the hyponyms of many terms cannot be found correctly.

For terms which hyponyms or highly related terms cannot be found in Wikipedia, we cannot measure the precision of these terms directly but we can measure the semantic

Table 6 The calculation of precision of 'Information Retrieval'

	Terms in taxonomy					Precision
Hyponyms	Evaluation	Relevance feedback	Information seeking	Query expansion	Clustering	0.8
Is occurring in the categories?	1	1	0	1	1	
Highly related terms	Hypermedia	Object oriented approach	Distributed information retrieval	Ontology	Ontologies	0.4
Is occurring in the content?	0	0	0	1	1	

Table 7 Average semantic relevance

	Hyponyms found	Hyponyms not found	Highly related terms found	Highly related terms not found	Total
From Google	0.2857	0.2170	0.4036	0.3653	0.3179
From Google scholar	0.6143	0.4871	0.6647	0.3593	0.5314
Total	0.4500	0.3521	0.5342	0.3623	0.4246

relevance between these terms with their corresponding concepts in the taxonomy. If they are similar semantically, it also can prove the effectiveness of entire taxonomy.

We design an indirectly measure of semantic relevance shown as Formula 5:

$$\text{semanticRelevance}_{\text{termi,termj}} = \frac{2 \times \text{amountOfSearchResult}(\text{'termi termj'})}{\text{amountOfSearchResult}(\text{'termi'}) + \text{amountOfSearchResult}(\text{'termj'})} \quad (5)$$

If two terms is more similar semantically, the value of semantic relevance will be higher. The same two terms can get 1 which is the highest value. We get 120 search results of 60 terms from Google and Google Scholar, which each returns 60 results. The detailed information of average semantic relevance is shown in Table 7.

From Table 7, we can see the average semantic relevance of terms found in Wikipedia is higher than terms not found which are less semantically relevant to their corresponding concepts. However, the total average semantic relevance of them is still satisfactory. And we can see another interesting result from Table 7 which Google Scholar gets higher value than Google since it is more specific to research domains and related to academic articles.

Finally, we design a prototype search system of academic articles. In this system, all search results are from Google Scholar. But users can expand their query terms more easily so that they can get more satisfied results quickly.

Figure 4 shows an example. A user submits a query about 'Search Engine' and gets the original result from Google Scholar. On the top of results, we add an expansion list of keywords. It includes hyponyms, hypernyms and highly related terms. User can expand current query by clicking desired terms with either OR operation or AND operation. The main purpose of this system is not to provide a full and sophisticated function but an evaluating environment. It remains all kinds of original expanded terms and displays them

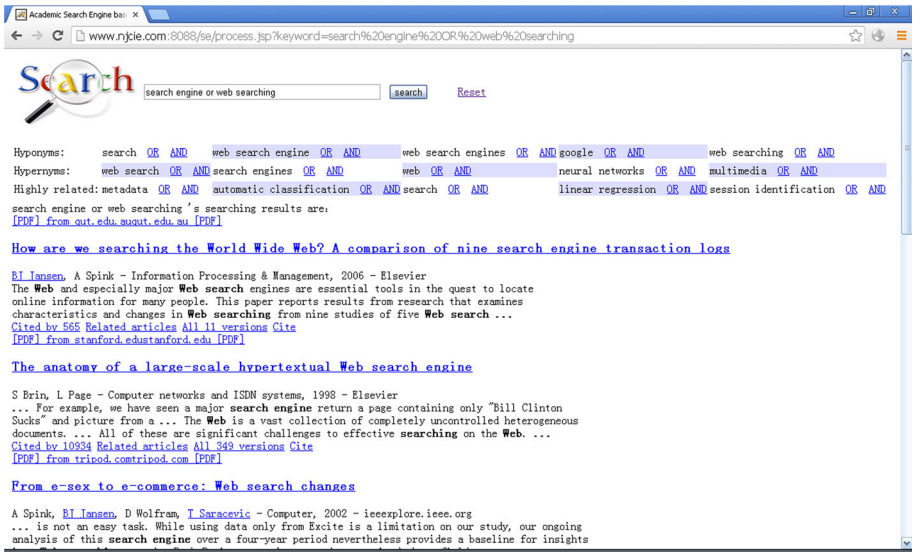

Fig. 4 The interface of searching 'Search Engine'

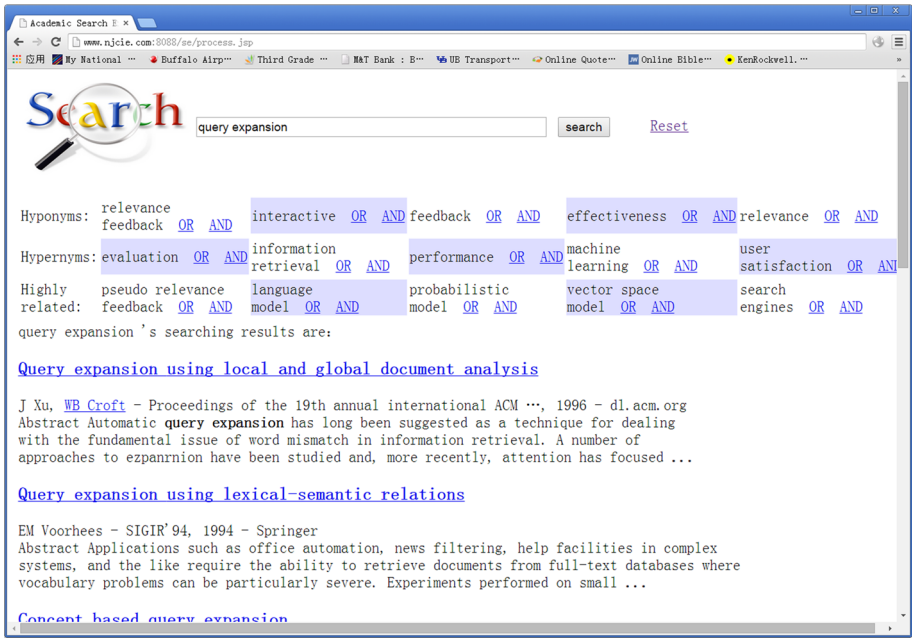

Fig. 5 The interface of searching 'Query Expansion'

in a direct way. And it also does not change the search results from Google Scholar. Users can understand how it helps improving searching with this term taxonomy.

Here is another example. Figure 5 shows what the interface looks like when submitting 'Query Expansion'.

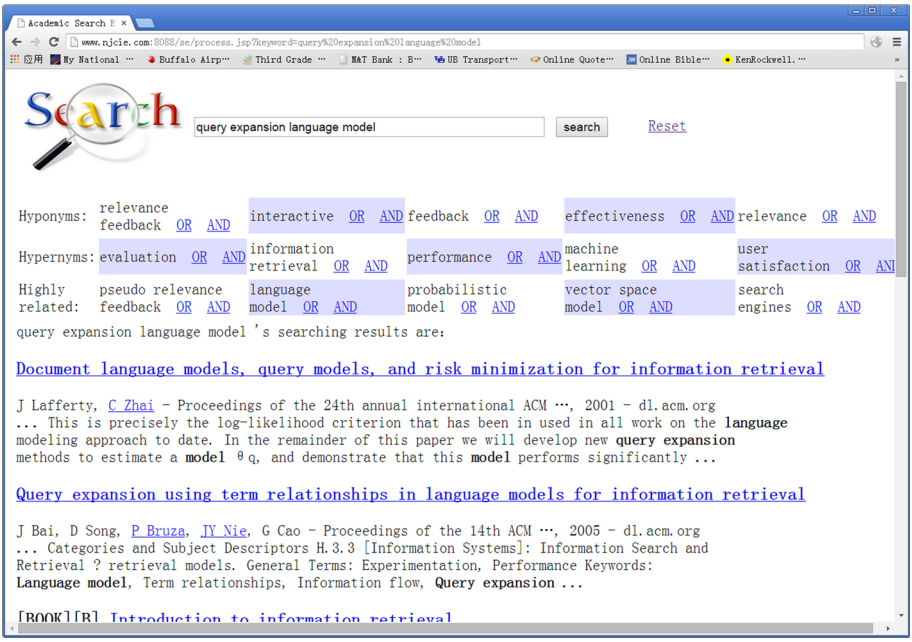

Fig. 6 The interface of query reformulating

The hyponyms include 'Relevance Feedback', 'Interactive', 'Feedback', 'Effectiveness' and 'Relevance'. The hypernyms include 'Evaluation', 'Information Retrieval', 'Performance', 'Machine Learning', 'User Satisfaction'. The highly related terms include 'Pseudo Relevance Feedback', 'Language Model', 'Probabilistic Model', 'Vector Space Model', 'Search Engines'. When a user want to know more about 'Language Model', he/she can click link 'AND' after 'Language Model'. Then the refined search results from Google Scholar can be got as Fig. 6.

We invite these 20 users to use the system and ask each user to search 5 terms to which they are familiar. They are also asked to mark their satisfaction for each search term. They can try many different expansions for one search term until they want to stop. We use a Likert scale of 5 in which 5 is satisfied and 1 is not satisfied. We group the queries into five groups based on their broad topic areas. They are 'Information Retrieval', 'Citation Analysis', 'Information Literacy', 'Digital Library', and 'Others'. The evaluation result of five groups is shown in Table 8.

We can conclude that the average satisfaction of our users in the area of library science is higher than in other areas such as information science. The reason lies in the journal types of our collection. The types of journals we have chosen are mainly about library science so that the corresponding amount of documents and keywords lead to this difference in the evaluation result. But we also see that average satisfaction in total is still 3.51 which tell us in general, our users are satisfied with the terms and their relations in the taxonomy constructed with our proposed method.

Table 8 The evaluation results

Topic	Satisfaction	Number	Average Satisfaction	Topic	Satisfaction	Number	Average Satisfaction
Information Retrieval	5	6	3.1304	Digital Library	5	4	3.1579
	4	3			4	4	
	3	4			3	5	
	2	8			2	3	
	1	2			1	3	
Citation Analysis	5	8	3.5652	Others	5	9	3.7600
	4	6			4	7	
	3	3			3	5	
	2	3			2	2	
	1	3			1	2	
Information Literacy	5	5	4.3000	Total	5	32	3.5100
	4	4			4	24	
	3	0			3	17	
	2	1			2	17	
	1	0			1	10	

Conclusions and outlook

We have presented an approach of automatically constructing term taxonomy and demonstrated that the relations of terms based on the weighted keyword co-occurrence analysis can be acquired effectively. In our future work, we plan to continue evaluating the stability and expand our data collection in other areas for conducting a wider user evaluation. As mentioned above, purely utilization of the semantic information of keywords can cause some difficulties when the data collection in some domains is not large enough to extract more effective relations of terms. Our future work will also involve designing some methods that will allow us to combine more useful information such as time sequences, and use some other approaches to enhance the ability of this algorithm.

Acknowledgments This work has been supported by Social Science Foundation of Jiangsu Province 2014SJB144 (2014), and Chinese National Natural Science Foundation 71103081 (2011).

References

Benz, D., Hotho, A., & Stumme, G. (2010). Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In *Proceedings of the 2nd web science conference (WebSci10)*, Raleigh, NC, USA.

Bordag, S. (2008). A comparison of co-occurrence and similarity measures as simulations of context. In *Computational linguistics and intelligent text processing* (pp. 52–63). Berlin: Springer.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179–255.

Buitelaar, P., Cimiano, P., Haase, P., & Sintek, M. (2009). Towards linguistically grounded ontologies. In *The semantic web: research and applications* (pp. 111–125). Berlin: Springer.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics, 22*(1), 155–205.

Chen, H. P., He, L., Chen, B., & Gu, J. G. (2009). Design and implementation of ontology generator based on relational database. *Computer Engineering, 35*(5), 34–36.

Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics, 8*(1), 43–48.

Cobo, M. J., López-Herrera, A. G., Herrer-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology, 63*(8), 1609–1630.

Dellschaft, K., & Staab, S. (2006). On how to perform a gold standard based evaluation of ontology learning. *The Semantic Web-ISWC 2006* (pp. 228–241). Heidelberg: Springer.

Doyle, L. B. (1962). Indexing and abstracting by association. *American Documentation, 13*(4), 378–390.

Eck, N. J. V., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.

Egghe, L., & Leydesdorff, L. (2009). The relation between Pearson's correlation coefficient r and Salton's cosine measure. *Journal of the American Society for Information Science and Technology, 60*(5), 1027–1036.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Geng, Q., & Geng, C. (2006). Concept extraction in automatic ontology construction using words co-occurrence. *New Technology of Library and Information Service, 22*(2), 43–45.

Gillum, T. L. (1964). Compiling a Technical Thesaurus. *Journal of Chemical Documentation, 4*(1), 29–32.

Hadzic, M., & Chang, E. (2005). Ontology-based support for human disease study. In *proceedings of the 38th annual hawaii international conference on system sciences, HICSS'05, IEEE.*

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation, 28*(1), 11–21.

Jung, Y., Ryu, J., Kim, K. M., & Myaeng, S. H. (2010). Automatic construction of a large scale situation ontology by mining how to instructions from the web. *Web Semantics: Science, Services and Agents on the World Wide Web, 8*(2), 110–124.

Labrou, Y., Stergiou, S., Adler, B. T., Marvit, D. L., & Reinhardt, A. (2012). U.S. Patent no. 8,280,886. Washington, DC: U.S. Patent and Trademark Office.

Lim, E. H., Tam, H. W., Wong, S. W., Liu, J. N., & Lee, R. S. (2009). Collaborative content and user-based web ontology learning system. In *Fuzzy systems, 2009. FUZZ-IEEE 2009. IEEE international conference* on (pp. 1050–1055). IEEE.

López-Herrera, A. G., Cobo, M. J., Herrera-Viedma, E., & Herrera, F. (2010). A bibliometric study about the research based on hybrid-dating the fuzzy logic field and the other computational intelligent techniques: A visual approach. *International Journal of Hybrid Intelligent Systems, 7*(1), 17–32.

Lu, X. B., Meng, X., & Zhang, J. (2012). Visualization of hot topics in social tagging based on co-words analysis method. *Journal of the China Society for Scientific and Technical Information, 31*(2), 204–212.

Martínez, M. A., Cobo, M. J., Herrera, M., & Herrera-Viedma, E. (2014). Analyzing the scientific evolution of social work discipline using science mapping. research on social work practice, 1049731514522101.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Morita, T., Shigeta, Y., Sugiura, N., Fukuta, N., Izumi, N., & Yamaguchi, T. (2004). DODDLE-OWL: On-the-fly ontology construction with ontology quality management. In *Proceedings of the 3rd international semantic web conference (ISWC).*

Murgado-Armenteros, E. M., Gutiérrez-Salcedo, M., Torres-Ruiz, F. J., & Cobo, M. J. (2015). Analysing the conceptual evolution of qualitative marketing research through science mapping analysis. *Scientometrics*, *102*(1), 519–557.

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems, 22*(3), 336–359.

Peat, H. J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42*(5), 378–383.

Qiu, J. G., Zhang, B., Wang, H. L., & Zhang, K. (2012). Research on regularity in adjacent co-occurrence between semantic relationsin objective knowledge system. *Journal of the China Society for Scientific and Technical Information, 31*(2), 126–135.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sheng, L., & Li, C. (2009). English and Chinese languages as weighted complex networks. *Physica A: Statistical Mechanics and its Applications, 388*(12), 2561–2570.

Soergel, D. (1974a). Automatic and semi-automatic methods as an aid in construction of indexing languages and thesauri. *International Classification, 1*(1), 34–38.

Soergel, D. (1974b). *Indexing languages and thesauri: construction and maintenance*. Los Angeles, CA: Melville Publishing Company.

Van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology, 60*(8), 1635–1651.

Wang, Y. F., Song, S., & Miao, L. (2006). Application study of co-occurrence analysis in knowledge service. *New Technology of Library and Information Service, 22*(4), 29–34.

Xu, S., Qiao, X. D., Zhu, L. J., Zhang, Y. L., & Xue, C. X. (2012). A novel approach for co-occurrence clustering analysis: maximal frequent itemset mining. *Journal of the China Society for Scientific and Technical Information, 31*(2), 143–150.

Yu, C. M., & Zhou, D. (2010). The complexity analysis of the emotional word co-occurrence network. *Journal of the China Society for Scientific and Technical Information, 29*(5), 906–914.

Zhang, Y. F., & Cai, J. J. (2011). Research on the user interest ontology learning based on web mining technology. *Journal of the China Society for Scientific and Technical Information, 30*(4), 380–386.

Zhang, Z. L., Zhang, Z. Q., & Li, X. Y. (2011). Co-occurrence analysis between research institutes and keywords based on 2-mode network. *Journal of the China Society for Scientific and Technical Information, 30*(12), 1249–1260.

Zhong, M. J., Wan, C. X., & Liu, A. H. (2009). Question answering system based on frequently asked questions using co-occurrence word model. *Journal of the China Society for Scientific and Technical Information, 28*(2), 242–247.