Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 7, July 2018, pp. 1570-1579
DOI: 10.23940/ijpe.18.07.p21.15701579

Automatic Generation of Comparative Summary for Scientific Literature

Yao Liu[a,*], Yuqing Yang[b], and Yi Huang[a]

[a]*Institute of Scientific and Technical Information of China, Beijing, 100038, China*
[b]*Peking University, Beijing, 100871, China*

Abstract

In this paper, we propose a comparative summary generation method and conduct key technologies research. We collect prior knowledge from the Internet via a light knowledge structure, extract core information from original literature, divide subtopics of two major topics with AGNES clustering to get the common and independent subtopics, and get comparative information with subtopics alignment and property alignment. We test the performance of each module to prove the validity of the proposed methods. Finally, we design and develop a comparative summary generation system, and the application in the nursing field shows that it can present users with useful information to facilitate the scientific research process.

Keywords: comparative summary; multi-document summary; knowledge mining

(Submitted on April 1, 2018; Revised on May 11, 2018; Accepted on June 25, 2018)

1. Introduction

A scientific database system can help researchers quickly obtain a list of relevant materials when searching for specific topics. With the rapid development of information technology in scientific literature, the growth of information has led to an increase in the comprehensiveness of potential information, and it has reduced the efficiency of information acquisition for research to a certain extent. Most of the existing services on automatic summary offered by a scientific database system are based on multi-document summary technology with a single-dimensional collection of documents, that is, the summary is generated based on the literature of the same field. Scientific literature, like news, is not isolated, but rather a huge flow of information networks. There are lots of connections among various fields in science, and researchers generally need to read a lot of literature to clarify the distinction and connections between different domains or topics. However, most of the existing scientific database systems do not carry out deep level mining for multi-topic literature collections to generate a comparative summary.

Summaries or abstracts can help us prejudge the information carried by a long or large number of documents. From the generating method, automatic summary generation can be divided into two types: extraction basis and abstraction based. The former abstract the key sentences directly in an article by sentence weight based on word frequency, clustering, graph, and language analysis. The sentences with higher weight and completeness are combined to generate a summary. The latter extracts the important text elements from a corpus, disorders the sentences, and reintegrates into the summary through sentence information extraction, structure compression, semantic reconstruction, information fusion and other technologies.

Zhai et al. [11] developed a "generative probabilistic mixture model" based on the study of comparative text mining (CTM), which performs interclass cluster and class clustering. The specific step is to give a set of m documents, assuming that there are k comparison points, each word a certain probability. k is a common topic, k * m are the characteristics of the mixed topic, and the document set is accumulated, and then the EM algorithm is used to find the probability distribution of the vocabulary of each common topic and characteristic subject on the document set. The similarities and differences of the topic are explored by the vocabulary in different clusters.

* Corresponding author.
E-mail address: liuy@istic.ac.cn

Campr and Jezek [1,2] used the Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) Model to dig out the independent topics of the two different collections and generate the summary. The concrete step is to model the two sets of documents, calculate the similarity of the subject by Jensen-Shannon divergence, and use the greedy algorithm to generate the summary sentence. But, the performance of the comparative summary generation by simply using LDA is not stable.

Chitra et al. [3] researched the user query comparison summary generation method, obtained comparative information from the URLs of search results, used DOM to divide pages into a concept block, combined it with query and feature keywords on the concept block, and then extracted important sentences to constitute a comparative summary.

There are also some other ways to generate a comparative summary. Shen and Li [9] proposed a new method of a multi-document summary based on sentence graph's minimum dominance map and applied it to comparative form generation. Wang et al. [10] proposed to use a sentence selection method based on a multivariate normal generation model that can represent independent information of different documents. The experimental results show that the method is superior to K-means clustering and other multi-topic non-redundant sentence selection methods.

Currently, studies in comparative summary mostly focus on mining relations between current literature and reference literature, but not the differences and relations between different topic collections. In addition, most of the studies do not take into account the semantic associations between topics in the comparative summary generation.

2. Idea and Framework

There are two characteristics of scientific literature: first, science development is a process of continuous accumulation, and some areas of the topic have been extensively studied, which can assist in mining the existing literature; second, scientific research is also a process of continuous development with lots of emerging topics. Only considering the topic with existing knowledge is not enough to fully meet the scientific literature mining needs. Domain concepts have semantic associations, and existing open encyclopedic knowledge or ontology structures can be regarded as a prior summary, which can quickly acquire the relevant information among them. Based on the existing research results, such as integrated crawler [5], automatic attribute extraction [8], automatic ontology construction technology [7], semantic annotation [6] and other technologies, according to the characteristics of the science literature combined with prior knowledge, we propose a method of mining the emerging topics and knowledge to generate a comparative summary. The overall idea and framework are shown in Figure 1.

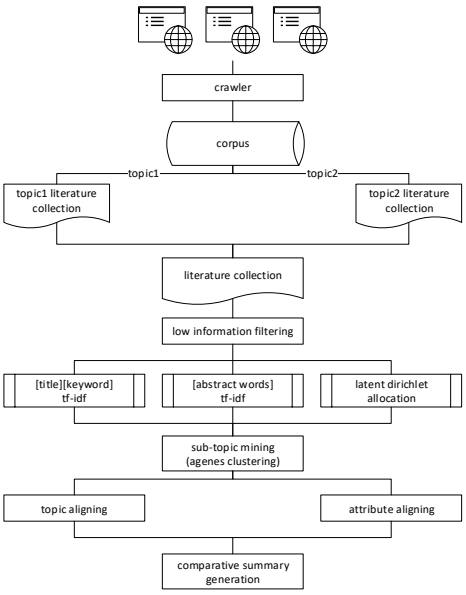

Figure 1. The overall idea and framework

3. Automatic Generation of Comparative Summary for Scientific Literature

3.1. Domain Prior Knowledge Acquisition

There are two parts of the corpus to be acquired. The first part is the transcendental light knowledge base of the domain topic, which mines subtopic and related knowledge words of a seed word and the domain related knowledge thesaurus involved in the literature. The second part is the scientific literature corpus, from which a comparative summary is generated.

Unlike the traditional domain knowledge base or domain ontology construction methods that require lots of human intervention, we propose a novel approach to automatically construct a lightweight knowledge base. By using an existing subject list and network encyclopedia list to obtain seed words, along with automatic screening and processing of these resources, common attributes of a topic and topic attributes related words can be extracted, and a prior semantic structure is formed. The semantic structure is used as a reference for information mining. At the same time, the related resources of a topic from the Internet are being processed, and the attributes of the topic are quickly formed based on the existing knowledge by attributes filtering, merging, and concepts screening. The resources acquired include structured domain subject headings, book content, semi-structured encyclopedic knowledge, unstructured industry websites, journal article, etc. A light knowledge structure is shown in Figure 2.

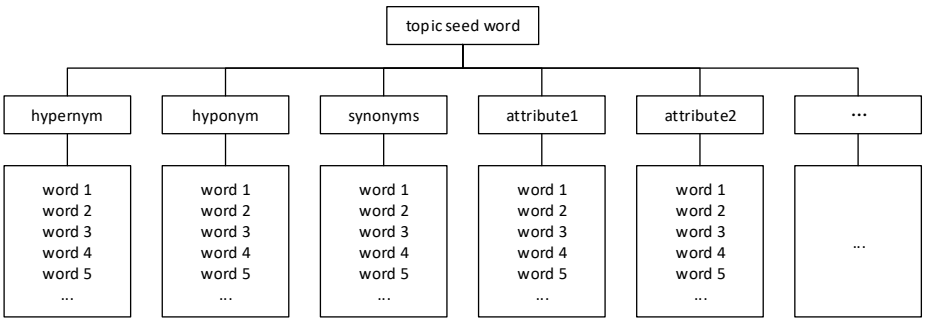

Figure 2. Light knowledge structure of a topic

3.2. Summary Core Information Acquisition

The original abstract of the scientific literature is long and unsuitable for direct information comparison, and it is necessary to extract sentences that are the most representative of the subject matter of the literature to ensure the readability and comprehensiveness of the information recommended. According to the characteristics of the abstract of scientific literature, we adopt a two-step compression method based on rules and clauses.

Rule-Based. Abstracts in scientific literature can be divided into two categories: one is structured, where the information can be divided into [object] [method] [results] [conclusion/discussion] four domains; the other is of the summary type.

(1) Structured abstract compression rules. [object] contains the study object of the article and the [conclusion] generally includes the description and evaluation information of the study [method]. [purpose] and [conclusion] sentences can be directly extracted as summary candidates.

(2) Unstructured abstract compression rules. For the unstructured abstract, the abstract is divided into a set of sentences by commas and periods, and the sentence containing the important information can be synthesized as a sentence candidate by extracting the words that can express the research [object], [conclusion], [method] and so on. We use a word list shown in Table 1 to extract core information from the unstructured abstract:

Table 1. Characteristic Words for Core Information Extraction

Position	Words
Beginning of a sentence	conclude, report, review, in order to, explore, therefore, results show that, focus,
Middle of a sentence	adopt, conduct, be a kind of,
End of a sentence	be of great significance/importance

Clause-Redundancy-Based. Those core information still show a certain degree of redundancy. For example, [object] and [conclusion] may have some redundant information. For the first type, with [object] and [conclusion] as the sentence node, after segmented and stop words are filtered, we get word set A and word set B. Let $C = |A \cap B|$, $R1 = C / A$ and $R2 = C / B$. When its value is greater than the threshold 0.8, compare R1 and R2, and remove the smaller one. For the second category, with a comma as a sentence node, each group is processed the same way.

500 abstracts were selected as the experimental input data, and the validity of the methods was evaluated by readability, information integrity, and information compression ratio. Readability is given manually with [1,2,3]. Information integrity is composed of the coverage of the words in the final compression sentence in the manually generated sentences, and information compression ratio is the ratio of the number of words in the compressed sentence to the number of the original. The result is shown in Table 2.

Table 2. Summary Core Information Extraction Assessment

compression method	readability	information integrity	information compression ratio
rules based	2.7	0.932	0.642
clause redundancy based	2.4	0.891	0.547

It can be seen that after clause redundancy based compression, the information integrity and information rate didn't drop too much, and the method is capable of obtaining the core information of a summary.

3.3. Subtopic Mining Based on Joint Semantic Vector

Comparative summaries should be able to present core literature under a common subtopic of a double topic as well as the core literature of different sub-topics under a common attribute. To achieve this, we divide the document collection of a double topic into sub-topics, identify attributes, and extract core literature.

Bow Topic Associated Feature Word Selection. Based on semantic similarity, text can be expressed as a bag-of-words (BOW) model. We use the traditional methods (including the low information part of speech (POS) filtering, domain low information word filtering, single word filtering, synonym filtering, etc.) to carry out low-information word filtering to generate a text vector. The low information word filtering process is shown in Figure 3.

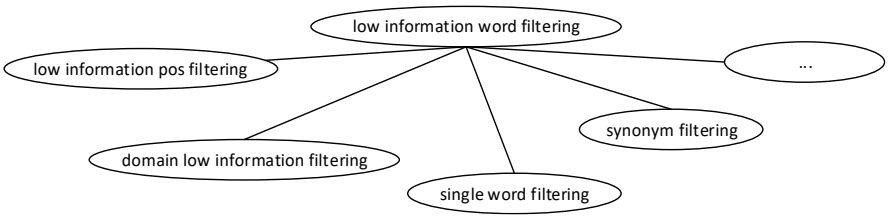

Figure 3. Low Information Word Filtering

Joint Semantic Vector Distance Representation. The basis of literature clustering is the similarity modeling between the literature, and the basis of similarity modeling is the vector representation of the lexical features. The similarity of the literature under the double topic has four characteristics:

(1) Participate in similarity modeling objects from two different topic areas. Double topics can be seen as two major topics with their own keywords.

(2) The similarity modeling objects come from titles, abstracts, and keywords of the scientific literature, and their contribution to the subject of the literature may be different. The vectors of the different positions of the scientific literature are respectively constructed and have a better effect than with the abstract alone.

(3) Extracting the abstract core information sentence should consider the contribution of the word to the topic.

(4) In addition to the polysemy phenomenon excavation, it is difficult to find these hidden common topics because of the small number of words involved in modeling.

Based on the above considerations, we choose to model the similarity of the literature as shown in Figure 4.

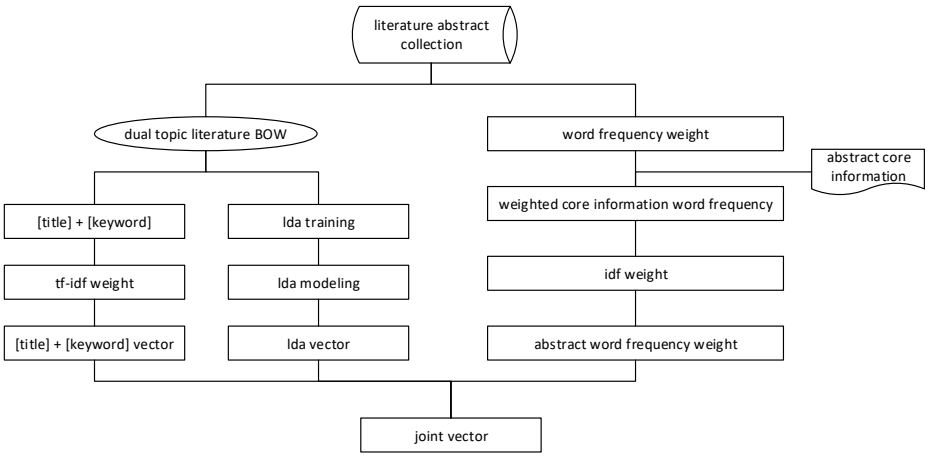

Figure 4. Joint Vector Construction Method

According to the characteristics of the double topic literature, the formula of the basic word weight is modified, and the word weight is weighted by the normalized word frequency in the core information. We combine the inverse document frequency and the inverse document into a single topic frequency for joint calculation. Finally, the formula for calculating the weight of word i in document d is shown in Equation (1).

$$\text{Weight}_{i,d} = \frac{n_{i,d}}{\sum_k n_{k,d}} * \left(1 + \frac{n_{i,c}}{\sum_m n_{i,c}}\right) * \log\left(\frac{|D_{TOTAL}|}{\{j: t_i \in d_j\}} + \frac{|D_{a|b}|}{\{m: t_i \in d_m\}}\right) \quad (1)$$

In the formula, $n_{i,d}$ is the number of occurrences of words I in document d. $\sum_k n_{k,d}$ is the total number of words in document d, $n_{i,c}$ is the number of words i in the core part c of document d. $\sum_m n_{i,c}$ is the length of the core part c,$|D_{TOTAL}|$is the total number of documents,$\{j: t_i \in d_j\}$ is the number of documents with word i under the double topic, $|D_{a|b}|$is the number of documents under the topic, and $\{m: t_i \in d_m\}$ is the number of documents with word i under the topic of current document.

For the existence of a large number of topics in the literature similar to the semantic similar terms, we choose the title words, keywords, and abstract words composed of BOW as the LDA model input.

By combining the word frequency vector and LDA as the composition of the literature similarity calculation, the word frequency vector measures the importance of the vector based on a co-occurrence degree but ignores the subject relation between the words. The LDA addresses the subject relation. The cosine distance is used as the basic method of vector distance calculation. According to the composition of the joint vector, the similarity calculation is carried out for the three vectors respectively, and the normalized linear weighting coefficient is used for the result. The formula for calculating cosine similarity is shown in Equation (2), and the formula for calculating vector distance is shown in Equation (3).

$$\cos(D_A, D_B) = \frac{\sum_{i=1}^{n}(A_i * B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2 * \sum_{i=1}^{n}(B_i)^2}} \quad (2)$$

$$\text{Similarity(A, B)} = \alpha * \cos(TK_A, TK_B) + \beta * \cos(S_A, S_B) + (1 - \alpha - \beta) * \cos(T_A, T_b) \quad (3)$$

In the formula, TK_A denotes the [title + key] vector of the literature. S_A denotes the abstract vector of the literature. T_A denotes the subject vector of the literature. α and β are the linear parameters adjustment weight distribution, and the linear parameters are determined by experiment.

Hierarchical Clustering based on Joint Vector. The number of documents in each cluster is not fixed, and it is difficult to delimit the number of topics in advance. Moreover, in the topic mining of double topic documents, the importance of clusters does not only depend on size. If a small cluster contains documents from a double topic, then the cluster is still important. Therefore, the improved AGNES algorithm is used to cluster the documents. The process shown in Figure 5.

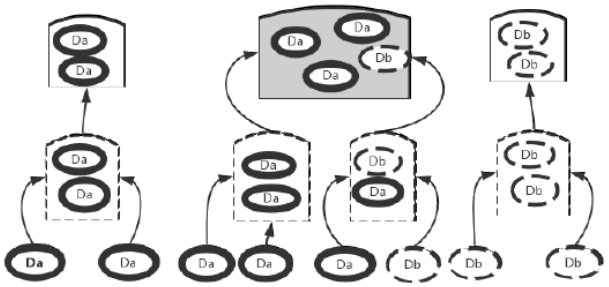

Figure 5. Document Clustering of Sub-Topics in Double topic

The average purity (Purity) and normalized mutual information (NMI) were chosen as the evaluation indexes. Purity between the calculations was designated to the correct cluster number of documents and also other documents in proportion to the number to determine the effect of clustering; the NMI value is used to measure the differences between the two cluster distributions to affect the balance of the one-way evaluation of the Purity index. The calculation of Purity and NMI are shown in Equation (4) and Equation (5).

$$\text{Purity} = \sum_j \frac{n_j}{n} argmax\ P(i,j) \quad (4)$$

$$\text{NMI} = \frac{MI(C,S)}{\sqrt{H(C)H(S)}} = \frac{\sum_{c,s} n_{c,s} \log \frac{n n_{C.S}}{n_C n_S}}{\sqrt{(\sum_c n_c \log \frac{n_c}{n})(\sum_s n_s \log \frac{n_s}{n})}} \quad (5)$$

1085 papers were chosen from the nursing field, including "gestational hypertension care" (246) "gestational diabetes care" (256) "elderly diabetes care" (259), "elderly care" (324). These 4 topics construct four topic pairs on [pregnancy induced hypertension in elderly hypertensive nursing care], [nursing care of senile diabetes gestational diabetes], [gestational hypertension], [gestational diabetes nursing care], and [nursing in elderly hypertensive elderly diabetes care].

They are divided into four topic pairs according to the demand for subtopic characteristics mentioned at the beginning of this section. In other words, we put a priority on hot subtopics and also on those that appear in double topics. So, we divide the documents into three collections, common subtopic collections and independent subtopic sets.

Through experimental analysis, weights in the formula of [title, keywords] and [abstract] are set to 3:7. The proportion of TFIDF weight vector is set to 4/5, and the a-value and p-value in Equation (3) are set to 0.24 and 0.56, respectively.

We made the comparison among the following methods: the method without low information (Method A), the word weight calculation method with low word information filtering and the use of only a traditional term frequency-inverse document frequency (Method B), the word weighting method based on comparative summary with low word information filtering (Method C), the method only using implicit topic vector clustering (Method D), and the combined joint vector method (Method E). The results are shown in Figure 6. It shows that the value of NMI is greatly improved after filtering low-information words. The improved calculation method of weight that is proposed is closer to the manual classification compared to other traditional methods. Finally, the joint vector method is slightly better than the method only using frequency vector.

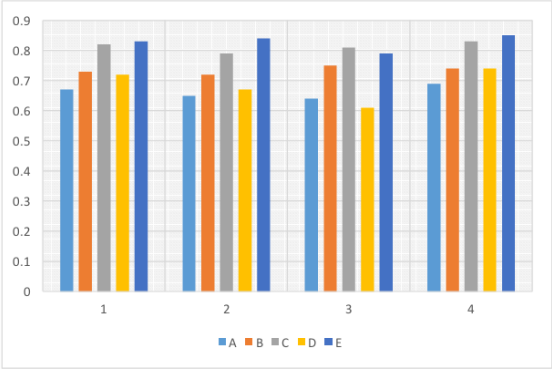

Figure 6. Result of Clustering Comparison

3.4. Mining of Subtopic Attributes

In order to get comparison information based on attribute alignment, the subtopic attributes should be identified first. In this paper, the mining of subtopic attributes is transformed into the process of subtopic core word classification: first, the representative words of the subtopic are identified; second, the subjective semantic knowledge is combined to classify the subtopic collections based on the classification of attributes.

In the subtopic core word acquisition, most of the traditional extraction methods are based on rules or syntactic analysis because scientific literature expression is uniform. However, exhausting rules require a certain time cost. Therefore, we consider the acquisition of the subtopic core word as the process of extracting the core features of a document set and construct the MI (mutual information) between the two random variables (W, T) centered on the literature to describe the relationship between the word and type. These two random variables are dimensioned by the number of documents under a double topic, and if the term contains words, then W is 1, otherwise, it is 0. Similarly, if the document is included in the subject, the value of T is 1, otherwise, it is 0.

The formulas for the calculation of mutual information and the extended formula using the maximum likelihood estimation are shown in Equation (6) and Equation (7).

$$I(W;T) = \sum_{e_k \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(W = e_t, T = e_c) \log_2 \frac{P(W=e_t, T=e_c)}{P(W=e_t)P(T=e_c)} \tag{6}$$

$$I(W;T) = \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_1.N_{\cdot 1}} + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_1.N_{\cdot 0}} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_0.N_{\cdot 1}} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_0.N_{\cdot 0}} \tag{7}$$

In the formula, N_{10} represents the number of literature that contain W but that are not in subject C. N_1.represents the number of literature that contain the subject word W. N is the total number of literature. In this paper, we sort words according to mutual information from high to low and select the highest value of mutual information as the type of keywords.

The experiment uses a double topic corpus with sub-topics and compares with the machine classification results. It uses the precision rate P, and the recall rate R and F to determine the effectiveness of this method. The result is shown in Table 3.

Table 3. The Result of Subtopic Attributes Mining

	P	R	F
Subtopic attributes mining classification	0.888	0.636	0.741

3.5. Subtopic Sorting and Core Sentence Extraction

Subtopic Sorting with Prior Knowledge. Since the displayed number of text is limited, it is impossible to present all candidate information sentences. Therefore, the sub-topics need to be sorted according to its importance, and the importance of the sub-topics is presented in the following way.

(1) Priority of field. Give priority to the main topics of literature main field.

(2) Priority of topic prior knowledge. Give priority to the sub-topics that appear in the topic semantic knowledge structure.

On the basis of the first two priorities, the sub-topics are sorted by the subject popularity. The subject popularity is the literature coverage of the subject in the topic, the wider the coverage of the subject, the greater the information load in the topic and the more important it is. The calculation method of the importance of the final subtopic is shown in Equitation (8), where N_t represents the number of literature in the sub-topic, and $\sum_i^{|T|} N_i$ represents all the literature in the set of sub-topics. In particular, when the topic is sorted in the common subtopic set, the number of literature is N_{t_a} and N_{t_b}, respectively, considering the balance of the two-topic in the common sub-topic. The number of documents in Topic A and Topic B in the subtopic is N_{t_a} and N_{t_b}, ie, $N_t = \overline{N_t} = \mathrm{N}_{t_a} + \mathrm{N}_{t_b} + \left(\mathrm{N}_{t_a} - \frac{\mathrm{N}_{t_a}+\mathrm{N}_{t_b}}{2}\right) + \left(\mathrm{N}_{t_a} - \frac{\mathrm{N}_{t_a}+\mathrm{N}_{t_b}}{2}\right)$, and the denominator is the sum of N_t.

$$\mathrm{Importance}_t = \frac{N_t}{\sum_i^{|T|} N_i} \tag{8}$$

Ranking of Candidate Sentences Combined with Document Extension Information. The core documents are selected to represent a sub-topic. In this process, we consider the following factors to measure the importance of core information in topic sentences.

(1) The measure of the information coverage of the subtopic in the candidate sentence. Within a topic, the most representative sentences should be able to cover the information that is expressed in most sentences within the subject. If a sentence is similar to other sentences in a topic, then the sentence in the subtopic can be considered more important, which can be obtained from the average distance between the literature and the other documents in the sub-topic. The higher the average distance is, the stronger the representation of the document in the subtopic is.

(2) The selection of candidate sentences based on external parameters of documents. The traditional automatic summarization methods are based on scientific literature, which only refers to the semantic information of sentences. But, in this demand, the double topic information recommendation service needs to consider the importance of literature. In addition to the semantic information requirements that are associated with the query words, the system should also predict the quality of the recommended reading documents in order to get a better recommendation effect. In addition to the semantic features of texts, scientific documents also contain much non-semantic information, which includes citations, journal names, the time of publication, references and so on, and can serve as a key to the importance of a reference in the document summary.

a. Impact factor of source literature (I). The usual impact factor refers to the ratio of the number of documents cited in the past two years, which are published in the journal. The influence factor scores of nursing magazines were obtained from The Journal of Nursing.

b. Cited quantity (C). In addition to the overall influence factor of the journal, the higher the single article cited, the higher the contribution of the literature to this topic, which means the importance of the subject in the subject is higher.

c. Time published (T). The later the document is published, the more important information it contains for the present research.

Finally, the importance of the score sentence is shown in Equation (9).

$$\text{Importance}_{S_k} = \frac{\sum_i^m sim(S_k,S_i)}{|T|} * (1 + \alpha * N(I) + \beta * N(C) + \gamma * N(T)) \quad (9)$$

In the formula, $sim(S_k, S_i)$ is generated by calculating the distance between the core sentences in the summary. The normalized weighting factors of external information of documents are the adjustment coefficient, whose value is 1.

By obtaining the subtopic collection from each topic group and consulting relevant researchers, 15 common topics can be extracted. The selection of sentence is then compared with the results of the machine. In this paper, we use the extraction method to get the final abstract rather than modify abstract sentences. Thus, we adopt the Edmundson [4] evaluation method to explore the effect of extraction of the three-topic collection of core document sentences. There is no absolute standard for the importance of the paper. Therefore, this paper takes the union of the core collection of manual selection to match the results obtained by machine.

The method proposed is compared with the traditional SumBasic abstract extraction method. The traditional SumBasic kernel sentence is used to score the sentence by accumulating the word weight in the sub-topic, and no other semantic features are taken into account. The final comparison results are shown in Table 4.

Table 4. Comparison of Core Paper Extraction Methods

Candidate Sentence	Topic Combination 1	Topic Combination 2	Topic Combination 3	Topic Combination 4
SumBasic	0.725	0.745	0.757	0.765
SumBasic+External Patameter	0.735	0.765	0.775	0.790
avgSimi	0.745	0.754	0.738	0.782
avgSimi+External Parameter	0.755	0.778	0.806	0.812

The results show that this method is closer to manual literature selection, because SumBasic only considers word frequency information while our method can recognize the semantic similarity relation between sentences using the average similarity calculation method. Some low-quality literature can be filtered by external parameters.

4. System Development and Application

For the framework, we used JAVA as the programming language, and MongoDB and MyEclipse as the development environment to make the comparative summary generation system. We can enter two words, and then get common /independent subtopic distribution information and attribute alignment information, as well as comprehensive abstract comparison information and common/independent sub-subject core paper information. As shown in the figure, the common/independent subtopic information mainly includes three pieces of content, which are subtopic heat distribution map, subtopic list, and subtopic attribute map. We can click on different labels to obtain different forms of comparison information.

Figure 7. Double Topic Comprehensive Abstract

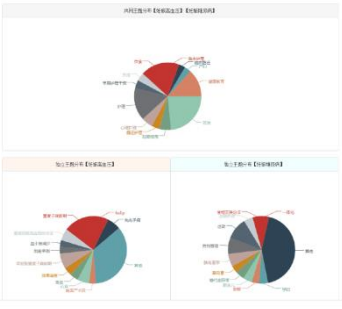

Figure 8. Common Subtopic and Independent Subtopic Distribution Pie Chart

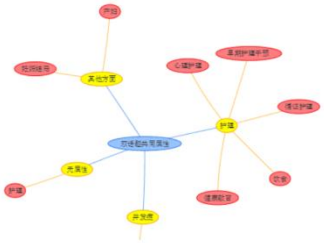

Figure 9. Common Attributes Map

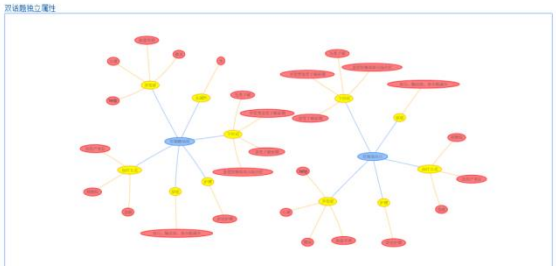

Figure 10. Comparison of Independent Attributes of Double topic

5. Conclusions

In this paper, we propose a method for comparative summarization generation of domain scientific topics, which gathers comparative information with subtopic alignment and property alignment based on web knowledge mining and multi-document summarization. The key steps and technologies are discussed. We collect prior knowledge from the Internet via a light knowledge structure, extract core information from original literature, divide subtopics of two major topics with AGNES clustering to obtain common and independent subtopics, and get comparative information with subtopics alignment and property alignment. The performance of each module is evaluated to prove the validity of the proposed methods. Finally, we design and develop a comparative summary generation system, and the application in the nursing field shows that it can present users with useful information to facilitate the scientific research process. However, this paper only explores the evolution of a priori knowledge and topic knowledge, and it does not excavate the knowledge at the attribute level. How to automatically obtain a more comprehensive and accurate knowledge semantic structure will be the focus of future research.

Acknowledgements

This work is partially supported by the National Key Project of Scientific and Technical Supporting Program (No. 2013BAH21B02). The authors thank the anonymous reviewers for their helpful suggestions.

References

1. M. Campr and K. Jezek, “Comparative Summarization via Latent Semantic Analysis,” In WSEAS International Conference. Proceedings. Recent Advances in Computer Engineering Series, no. 7. WSEAS, 2012.
2. M. Campr and K. Jezek, “Comparative Summarization via Latent Dirichlet Allocation,” in CEUR Workshop Proceedings, 2013, vol. 971, pp. 80–86.
3. P. Chitra, R. Baskaran, and K. Sarukesi. “Query Sensitive Comparative Summarization of Search Results Using Concept Based Segmentation,” arXiv preprint arXiv:1201.2304, 2012.
4. H. P. Edmundson, “New Methods in Automatic Extracting,” Journal of the ACM (JACM), vol. 16, no. 2, pp. 264–285, 1969.
5. X. Gong and Y. Liu, “Research on Construction of Integrated Semantic Crawler,” ICIC Express Letters, Part B: Applications, vol. 7, no. 7, pp. 1591–1598, 2016.
6. Y. Liu, H. Shi, D. Zheng, and Y. Huang, “Study on Semantic Annotation for Professional Literature,” ICIC Express Letters, Part B: Applications, vol. 5, no. 5, pp. 1383–1389, 2014.
7. Y. Liu and R. Wang, “Research on Semantic Metadata Online Auxiliary Construction Platform and Key Technologies,” ICIC Express Letters, Part B: Applications, vol. 4, no. 4, pp. 897–904, 2013.
8. Y. Liu, D. Zheng, and Z. Guo, “Research on Feature Acquisition and Key Expression Technology of Knowledge-Intensive Text,” ICIC Express Letters, Part B: Applications, vol. 5, no. 1, pp. 57–64, 2014.
9. C. Shen and T. Li, “Multi-Document Summarization via The Minimum Dominating Set,” In Proceedings of the 23rd International Conference on Computational Linguistics, vol. 2, pp. 984–992, 2010.
10. D. Wang, S. Zhu, T. Li, and Y. Gong, “Comparative Document Summarization via Discriminative Sentence Selection,” ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 3, pp. 1–18, 2012.
11. C. Zhai, A. Velivelli, and B. Yu, “A Cross-Collection Mixture Model for Comparative Text Mining,” in Proceedings of the 2004 ACM SIGKDD International Conference On Knowledge Discovery And Data Mining - KDD ’04, 2004, p. 743.

Yao Liu graduated from the School of Information Science and Technology, Peking University. He entered the Post-doctoral mobile station of the Institute of Computational Linguistics in Peking University, as an assistant researcher from 2005 to 2007. Now, he is a research fellow at the Institute of Scientific and Technical Information of China (ISTIC), and the Deputy Director of the Engineering Research Centre of ISTIC, Beijing, China. He is also a distinguished member of the China Computer Federation (CCF). His current research interests include natural language processing, knowledge engineering, and artificial intelligence.

Yuqing Yang graduated from the School of Software and Microelectronics, Peking University with her Master's degree. Her current research interests include natural language processing and machine learning.

Yi Huang graduated from the School of Software and Microelectronics, Peking University with his Master's degree. His current research interests include natural language processing and knowledge engineering.