Study on Dynamic Multi-document Summarization System Framework Method

Meiling Liu[1,4], Wei Wei[2], Bing Xu[3], Huiqiang Wang[4]

[1] Department of Computer Science and application, Northeast Forestry University, China
[2] School of Computer Science and Engineering, Xi'an University of Technology, China
[3] Machine Intelligence & Translation Laboratory College of Computer Science and Technology, China
[4] College of Computer Science and Technology, Harbin Engineering University, China
lmling2008@163.com, weiwei@xaut.edu.cn, xb@mtlab.hit.edu.cn, wanghuiqiang@hrbeu.edu.cn

Abstract

This paper introduced a new Internet-based dynamic multi-document summarization system framework based on natural language processing and applied for the data management of wireless sensor networks that was capable of providing computational linguistics-related technical support. We mainly studied the related model of text mining for perceptual data. Wireless sensor networks had a large number of streaming data, with real-time characteristics. After basic node operation, it generated big data can be used for data mining. In order to apply the data mining technology to information processing of wireless sensor networks, this paper tried to find a similar such as DUC2008 abstract test samples, trained model and algorithm. The system integrated subsystems with different emphases to improve system performance, combined three innovative methods. Given that to date little research on dynamic multi-document summarization has been reported, this study had great significance. The results obtained by the new framework were compared with those from the TAC2008 evaluation, demonstrated that the new system's performance matched that of the best existing systems.

Keywords: Dynamic, Evaluation, Model, Summarization

1 Introduction

Wireless sensor network is a new research area, and any applications can not be implemented without processing and analyzing sensor data [1]. Wireless sensor network is a new information acquisition and processing technology, which has been widely used in real life [2]. At present, the wireless sensor network, as a new technology to obtain and process information, is being widely studied [3-4]. In order to apply the data mining technology to information processing of wireless sensor networks [5], this paper try to find a similar such as DUC2008 abstract test samples, trained model and algorithm.

People are now seeking to access the many and varied types and formats of information that are now available on the Internet. The growing awareness of the problem of dynamic summarization led the National Institute of Standards and Technology (NIST) [6] in the U.S. to initiate a series of Document Understanding Conferences (DUC) [7], a precursor of the ongoing Text Analysis Conferences (TAC). The work reported in this paper is the result of evaluations carried out for TAC2008 and TAC2009 in the summarization track.

It is realized based on the dynamic generated concept; the major contributions of the new application proposed here are as follows:

(1) The proposed summarization system utilizes dynamic evolution modeling, using similarity and centroid integer selection and sentence sorting weighted by the dynamic manifold method.

(2) Dynamic summarization system framework by modifying and adapting the multi-document summarization steps apply in existing systems and incorporating new modules capable of dealing with the dynamic nature of Internet documents.

(3) Three dynamic multi-document summarization models are combined to become complementary, thus preserving the dynamic evolution of the abstracts with high novelty and historical evolution and improving the overall performance of the dynamic abstract.

2 Related Work

The Document Understanding Conference (DUC), which is organized and funded by the U.S. National Institute of Standards and Technology (NIST), is the main driving force in the field of multi-document summarization research. Researchers from different countries, research institutions, and universities have presented their latest results and the new systems they have developed at the annual DUC meetings ever since 2001 [8-9]. In the early years, DUC promoted

*Corresponding Author: Wei Wei; E-mail: weiwei@xaut.edu.cn
DOI: 10.3966/160792642018091905006

technology development in the field, but as the Internet information age gathered pace DUC proposed a new evaluation task in order to adapt to users' rapidly changing needs in this area. The target of creating a dynamic multi-document summarization system is formally announced in 2007 [10]. The proposed evaluation task launch multi-document summarization as a new research area and it quickly became the new hot spot in the field. As technology research has advanced, the scale of DUC evaluation expanded to cope with it and in 2008 it is combined with the Text REtrieval Conference (TREC) and renamed to become the Text Analysis Conference (TAC) [11].

The evaluation of dynamic multi-document summarization systems is the major task assigned to TAC2009 and TAC2008. The dynamic nature of the content is important because timing lies at the heart of the dynamic abstracts that are searched for News Information Detection (NID), Topic Detection and Tracking (TDT) [12], and other such applications. Time information has thus been the focus of a great deal of attention and plays a very important role in Natural Language Processing (NLP) [13], making it the basis of many natural language processing systems. Many of these multi-document summarization [14] systems order the relevant information chronologically, but in their question answering systems the answer to "when" questions often lacks time information. Network information [15] has three main characteristics: large size, similar themes, and dynamic evolution. To account for these three characteristics, Ye [16] of Shenyang Aerospace University proposed an approach to multi-document summarization based on textual thematic analysis. Taking a different approach, Xu [17] of the Chinese Academy of Sciences has proposed a multi-document summarization system for web-oriented topics that models dynamic evolution in the documentation set, so the resulting abstract has a low redundancy when compared with older documents in the set.

3 Dynamic Summarization Systems

The dynamic multi-document summarization model proposed here thus included five sub-modules: document pretreatment, feature extraction, information filtering, sentence weighting, and sentence selection and ordering. The document pretreatment module applied the new system developed by our group and its task concluded sentence boundary detection, tokenization, part-of-speech tagging, morphological analysis and stop words filtering. The sentence selection was used to opt for the suitable sentence from the document set to construct an accurate and non-redundant abstract and the main function of ordering module was to approve the readability of the abstract. Both sub models applied a conventional sentence selection and ordering algorithm. The main innovation of this work was the creation of a new model that is capable of dealing with the evolution of information in multi-document sets. In this paper, the three modules related to the dynamic characteristics, namely those for feature abstraction, information filtering and sentence weighting were described and shown in Figure 1.

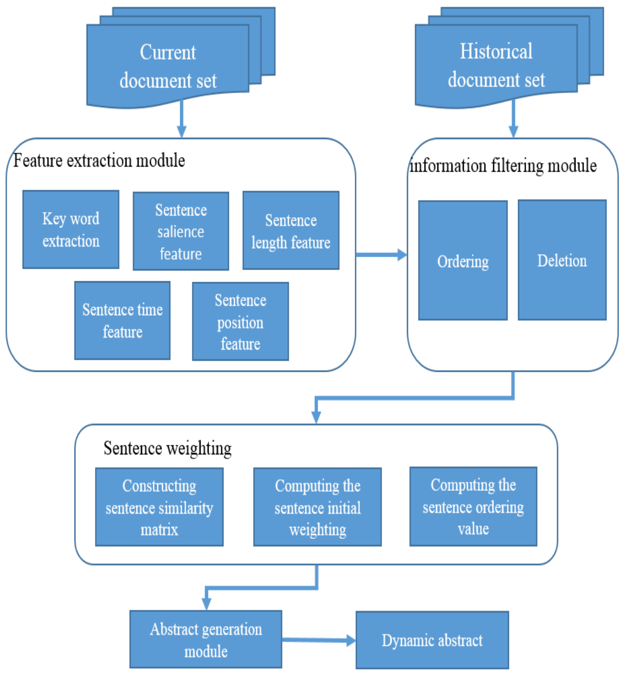

Figure 1. The frame diagram of the system proposed

3.1 Feature Extraction Module

In order to be an effective dynamic multi-document summarization system, the new system must not only take on all the characteristic of its predecessors but also incorporate dynamic characteristics. In addition to all the sub-modules included in traditional systems, such as sub-modules for extracting keywords, length features and position features, the new system must also incorporate three new sub-modules that extracted salient features, historical information features, and time features.

Extraction and weighting of keyword. Here, however, because the algorithm in a subsequent module needed all the words other than the stop words in a stop words list which contained all the common stop words, all the words were treated as keywords and the new algorithm proposed in this module computed the weighting of keywords using TF*IDF*ISF, where ISF was anti-sentence frequency.

The historical information feature of a sentence. The new system was thus able to estimate the evolution of the information held in the document set by extracting its historical information feature. The procedure for computing the value of the historical information feature was as follows: First, the system processed the historical document set and created an abstract by applying a traditional multi-document summarization approach. Second, the sentence set—

called the current sentence set—in the current document was examined. Taking the historical abstract as a reference, the system computed the document's information superposition by comparing each sentence in the current sentence set with its counterpart in the historical abstract.

The information salience feature of a sentence. The information salience feature of a sentence compared the information contained in the sentence to the integer information contained by all the data in the document set. Thus, the information salience feature could be used to measure the probability that the sentence was being examined should become a member of an abstract. The addition of a sentence to an abstract could improve the performance of the system markedly, so the information salience feature was an important part of the new system. According to the Vote Theory in mathematics, if something attracted the great majority of a vote, then it was more important than any of the other options. We could also apply this theory to the summarization process: if a sentence in a sentence set had the highest vote when compared with the other sentences in the set, it would be considered the most important sentence.

The time information feature of a sentence. Extracting the time information feature from a sentence was a difficult and time-consuming task. It constituted the main part of timing in multi-document summarization and was related to time expression recognition and normalization. State the number of documents in the document set and then applied the appropriate algorithm to compute the time information feature weighting of every document according to its Time Information Feature Ordering Value; the time information feature weighting of each document corresponds to the reciprocal of its Time Information Feature Ordering Value.

The length information feature of a sentence. The aim of any summarization system must also be to ensure the abstract contains as much important information as possible from the document set, enabling readers to gain useful information at a low time expense. This led us to conclude that as both very long and very short sentences decrease the abstract's information density, when the summarization model extracts a sentence to generate the abstract, a length range rule should be applied to select those sentences in which the ratio of information to length is highest.

The position information feature of a sentence. Research in linguistics had indicated that the majority of important sentences in a document are distributed at the beginning or end [18]; many sentences in the middle were complementary but contained much less information. Because the aim was to select sentences with high information when composing an abstract, the position information feature of a sentence was important for traditional and dynamic multi-document summarization systems.

3.2 Information Filtering Module

The object of the dynamic multi-document summarization system was to characterize the current document set, but it must take some information from the historical document set as information basis. In a general way the current document would contain some information that had been expressed by some corresponding document in historical document set. This information had been ignored, forgotten, or discarded by some readers, when they obtained the information from reading an historical document on the some subject. They needed not expend lots of time and energy to re-attain it.

3.3 Sentence Weighting Module

The algorithm applied to enhance the dynamic capabilities of the new model was the Dynamic Manifold Ranking model [19], which was a perfect sentence weighting algorithm. The ordering scores of the all nodes were determined by combining local information and context information in a manifold structure. In order to enable the manifold ordering algorithm to assign a dynamic score to every sentence, this paper proposed a new manifold ordering algorithm, the Dynamic Manifold Ordering Algorithm, which incorporated time and historical information features. The Dynamic Manifold Ordering Algorithm consisted of three main steps: the construction of a similarity matrix, and the computation of the initial sentence weighting and sentence ordering.

4 The New Framework

The emphasis of any new dynamic multi-document summarization system was on achieving a dynamic result in a way that was an improvement over traditional multi-document summarization systems. In this section the implementation algorithms would be described and the technique for resolving the problem of increasing dynamics would be discussed.

4.1 Implementing the Feature Extraction Module

The keyword extraction algorithm. This paper proposed a new keyword extraction algorithm, the TF*IDF*ISF Based Keyword Extraction algorithm. The new algorithm differed from the standard TF*IDF [20] based keyword extraction algorithm in that it utilized additional sentence information. While IDF was used to measure the importance of the specified word in document level, ISF was applied to evaluate the meaning of the word for summarization in sentence level. ISF could remove many unimportant words that often occurred in sentences but had less contributions for the content of the document [21]. The new ISF term was the inverse sentence frequency, which could be computed as follows: First calculated the total number of sentences that contained the word and then computed its reciprocal, which was the inverse sentence frequency of word. We could compute the TF*IDF*ISF value for each word, namely the word weighting, using the following formula:

$$Wgt(w)=TF(w)*IDF(w)*ISF(w) \quad \textbf{(1)}$$

Once the word weighting of all the words has been computed, we could measure the importance score of every word based on their word weighting and thus determined the keyword set for the document.

The algorithm for the sentence historical feature. This algorithm could be expressed as follows: First, computed the similarity value of the sentence by comparing it with every sentence in the historical sentence set. This was done by applying an appropriate sentence similarity value algorithm to form a similarity value vector, and then computing the sum of all the values from the similarity value vector. The resulting algorithm could be expressed as:

$$NWgt(s)=\sum_{i=1}^{m}\left(\frac{\sum_{j=1}^{n} Wgt(w_j)}{length(s_i)}\right)/\,length(s)*count \quad \textbf{(2)}$$

where *NWgt(s)* denoted the historical information feature value of a sentence *s,* m was the number of sentences in the historical abstract, *n* was the count of the same words between the sentence *s* and sentence *si*, *Wgt(wj)* denoted the word weighting of word *wi* that appeared in both sentence s and sentence *si*, *s* and *si* were sentences in the current sentence set and historical sentence set, respectively, length*(si)* and length*(s)* were the numbers of words in the current and historical sentence sets, respectively, and count was the total number of sentences in the historical abstract.

The value of the sentence salience feature. The salience feature value of a sentence in the current sentence set was determined by computing the similarity value between it and all the other sentences in the set to form a similarity value vector. We could then compute the sum of all the values from the similarity value vector, which was the sentence salience feature value, by applying the following expression:

$$SWgt(s)=\sum_{i=1}^{m}\left(\frac{\sum_{j=1}^{n} Wgt(w_j)}{length(s_i)}\right)/\,length(s)*count \quad \textbf{(3)}$$

where *SWgt(s))* denoted the historical information feature value of a sentence s, m was the number of sentences in the historical abstract, and *n* was the count of the same words between sentence s and sentence s_i.

The value of the sentence time feature. The value of a sentence was computed by extracting the publication time ordering value of the document to which it belongs. This approach not only improved the system dynamics, but also did not increase the burden on the system, making it a very effective computing algorithm. The detailed algorithm was expressed as follows:

$$TWgt(s)=1/n \quad \textbf{(4)}$$

Where *n* was the ordering value of the document to which the sentence s belongs in the current document set.

The value of the sentence length feature. As too short sentences usually contained too less information and too long sentences need to occupy too much space of the abstract, the sentence with appropriate length should be given higher weight. The value of sentence could be written as follows:

$$LWgt(s)=\begin{cases}1/\left(Length(s)-0.5*MaxLength\right) & Length(s)>0.5*MaxLength\\ 2 & Length(s)=0.5*MaxLength\\ 1/\left(0.5*MaxLength-Length(s)\right) & Length(s)<0.5*MaxLength\end{cases} \quad \textbf{(5)}$$

where *Length(s)* was the length of sentence *s*, *s* denoted a sentence in the current sentence set, *LWgt(s)* denoted the length weighting of sentence *s* by the system, and *MaxLength* was the longest length allowed for a sentence and the value of *MaxLength* was set according to the longest sentence in the target document set. In order to avoid the denominator of the equation equaling to 0, the value of the *MaxLength* should be set to an odd number. This equation was an empirical formula, and more information could be accessed by referring to.

The value of the sentence position feature. The detailed implementation algorithm for this feature could be written as follows:

$$PWgt(s)=\begin{cases}1/j & j<n\\ 1 & j=1\end{cases} \quad \textbf{(6)}$$

where *j* denoted the sentence's position number in the document, *n* denoted total number of the sentences in the document and s was the sentence in the document.

4.2 Implementing the Information Filtering Module

The processing algorithm could expressed as follows: First, sort all the sentences in current sentence set in descending order according to the value of their sentence historical information feature, where the sentences ranked near the top of the order contain the most historical information. Second, set a threshold according to the information contained in the document. After it had been processed by the information filtering module, the initial sentence set became a dynamic sentence set that could then be processed by later modules. Setting the value of the sentence count filter was an important factor here, and may require some experimentation to determine an appropriate level. Here, the experimental value selected for this system is

50.

4.3 Implementing the Sentence Weighting

The implementation of the sentence weighting consisted of the following three steps:

Construction of the sentence similarity matrix. The construction of a sentence similarity matrix lied at the heart of the new Dynamic Manifold Ordering Algorithm proposed here. As the experimental tested reported in Section 5 confirm, the new sentence similarity computation algorithm developed for this study was an effective way of dealing with the key problem addressed by this research. The detailed algorithm was expressed as follows:

$$Sim(s_i, s_j) = \frac{\sum_{k=1}^{count} Wgt(w_k)}{length(s_i) + length(s_j)} \tag{7}$$

where *Sim(si,sj)* denoted the similarity between sentence *si* and sentence *sj*, and count was the number of same words between the two sentences. The similarity of any two sentences could be compared and computed by applying Eq. 7 and the n*m similarity matrix *S* constructed by integrating all the similarities between every possible pair of sentences, where *n* denoted the number of sentences in current documents set and m denoted the number of sentences in historical documents set.

Computing the sentence initial weighting. Computed the sentence initial weighting was another important step in the Dynamic Manifold Ordering Algorithm. An initial weighting for each sentence was used to embody its initial importance and this was calculated as follows:

$$FWgt(s) = \alpha * \sum_{i=1}^{n} Wgt(w_i) + \beta * LWgt(s) \tag{8}$$

where *n* was the length of sentence s, *LWgt(s)* was the length feature value of sentence *s*, and α and β were parameters whose values must be determined experimentally.

Computing the sentence ordering value. Computing the sentence ordering value was the third key step in the algorithm, as this controls several of the characteristics of the dynamic system. Here, the Dynamic Manifold Ordering Algorithm was applied to assign an ordering value to every sentence in the current sentence set. The proposed Dynamic Manifold Ordering Algorithm was a modified version of the traditional Manifold Ordering Algorithm and was expressed as follows:

$$f(t+1) = \alpha * Sim * f(t) + (1-\beta) * y \tag{9}$$

where f(t+1) was a vector known as the ordering value vector, where the element was the ordering value of corresponding sentence; *f(t)* was also a temporary vector as described above and had an initial value equal to the initial weighting of all sentences; Sim was a similarity matrix; and α and $1-\beta$ denoted the comparative contribution value of the current ordering value vector and were based on the previous ordering value and the ordering value of nearby sentences.

In order to enable the traditional Manifold Ordering Algorithm to take into account the dynamic nature of the data, time information and historical information features could be added to the formula, yielding the Dynamic Manifold Ordering Algorithm. This can be written as follows:

$$\begin{aligned} f(t+1) = {} & \alpha * TWgt(s) + \beta * PWgt(s) + \\ & \eta * SWgt(s) - \mu * NWgt(s) + \gamma * S * f(t) \end{aligned} \tag{10}$$

where *TWgt(s)* denoted the time information feature value of sentence *s*, *PWgt(s)* denoted the position information feature value of sentence *s*, *SWgt(s)* denoted the salience information feature value, *NWgt(s)* denoted the historical information feature value, *f(t)* and *f(t+1)* were as shown in Eq. (9), α, β, γ, η and μ were parameters that must be determined experimentally, and *S* was the similarity matrix. As this was an iterative algorithm, the number of iterations must also be set. The resulting vector consists of elements whose ordering values corresponded to the sentence weightings of the corresponding sentences.

4.4 Abstract Generation Module

This paper proposed an improved Maximum Marginal Relevance (MMR) [22] based redundancy deletion algorithm that adequately considers the information relationship between two sentences and adds dynamic information into the system. This algorithm was constructed based on keyword weighting and was written as follows:

$$\begin{aligned} AZWgt(s) = {} & \alpha * (BZWgt(s) \\ & -\beta * \sum_{i=1}^{n} \frac{\sum_{j=1}^{simcount} Wgt(w_j)}{\sum_{k=1}^{count(s_i)} Wgt(w_k)} \end{aligned} \tag{11}$$

where *AZWgt(s)* was the ordering value of a sentence after it has been processed by the algorithm, *BZWgt(s)* was the ordering value of a sentence before it had been processed by the algorithm, *n* denoted the current number of sentences in the abstract, Simcount was the number of words that appear in both abstract sentence *si* and candidate abstract sentence *s*, and α and β were parameters whose values must be determined experimentally.

5 Experimental Results and Analysis

The corpus of documents used to test this system consisted of the standard set provided by TAC2008. This body of documents consists of 48 topics, each of which was composed of a document set containing 20 documents. Each set of 20 documents was divided into two sub-document sets according to their date of publication, with the first set of ten being used here as the historical information, referred to as the historical document set, and the remaining ten documents being the current document set used here as new information to compare with the information contained in the historical document set. Because the corpus of documents used in this study contains both historical and current information synchronously, it provided a useful source of dynamic data for our study of information evolution. In this paper, R-2 and R-SU4* are used as the evaluation metrics and there were calculated by applied the ROUGE tool [23], where R-2 was the co-occurrence rate of binary phrase between the abstract constructed by computational method and the standard abstract produced by human and R-SU4* was the co-occurrence rate of binary phrase with 4 space between the abstract constructed by computational method and the standard abstract.

This experimental test of the proposed algorithm consisted of six sub-experiments. In order, Experiments 1 through 5 were used to determine the optimum values of the five parameters α, β, γ, η and μ, respectively. Experiment 6 and experiment 7 were used to find the optimum value of n. The results are shown in Table 1.

Table 1. The parameter effect influence on the results

Exp	α	β	γ	η	μ	R-2	R-SU4*
1	0.2	0.2	0.2	0.2	0.2	0.069	0.104
2	0.1	0.1	0.2	0.1	0.3	0.079	0.098
3	0.3	0.2	0.3	0.1	0.1	0.100	0.125
4	0.3	0.1	0.3	0.1	0.2	0.069	0.104
5	0.2	0.1	0.2	0.2	0.2	0.101	0.137
6	0.1	0.1	0.2	0.2	0.2	0.060	0.098
7	0.1	0.2	0.1	0.3	0.1	0.100	0.126

The data shown in Table 1 reveal that this system exhibits the best performance when the five parameters are set to be 0.2, 0.1, 0.2, 0.3 and 0.2. Our analysis of the seven different test runs suggest that the value of all the parameters must be carefully balanced to achieve the best performance for the system as a whole. TAC2008 not only provides a suitable standard body of documents for testing new summarization systems, but also records the evaluation scores of the best of the candidate summarization systems proposed. We are thus able to use the evaluation data provided by TAC2008 to make further improvements. The performances of the three best summarization systems from TAC2008 are compared with that of our proposed new dynamic summarization system in Table 2.

Table 2. Comparison with TAC2008

SYSTEM	R-2	R-SU4*
The proposed system	0.101	0.137
Rank_1	0.101	0.137
Rank_2	0.097	0.134
Rank_3	0.092	0.132

As the data in Table 2 indicate, the performance of our proposed system exceed that of the second and third ranked systems from TAC2008, and equaled the score achieved by the highest ranked system. This suggested that the system presented in this paper is at the forefront of international research in this area.

6 Conclusions

This paper described a new Internet-based dynamic multi-document summarization system framework based on natural language processing for modeling the evolution of dynamic data using the sub-space matrix method, information filtering using the similarity and centroid integer selection method, and sentence weighting using dynamic manifold sorting method, leading finally to a new abstract generation system. By modifying and adapting the multi-document summarization steps applied in existing systems and incorporating new modules capable of dealing with the dynamic nature of Internet documents, we were able to improve the abstract system performance in several ways. Three dynamic multi-document summarization models were combined to become complementary, thus preserving the dynamic evolution of the abstracts with high novelty and historical evolution and improving the overall performance of the dynamic abstract. The test results for this system were very promising, equaling those of the best of the TAC2008 systems and confirming that the combination of these model algorithms exhibited good performance and stability. In addition to having a high-application value, the innovative nature of the models and algorithms will be promoted the further development of the dynamic multi-document summarization field.

Acknowledgments

Supported by the “National Natural Science Foundation of China”, No 61702091; Supported by the “Provincial Natural Science Foundation”, No. F2015037, Heilongjiang, P.R.China; Supported by the “Fundamental Research Funds for the Central Universities” No 2572018BH06; Northeast Forestry University, P.R.China; Supported by the “National Natural Science Foundation of China”, No 61402134, P.R.China.This work is also supported by the Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20136118120010)

References

[1] W. Wei, Q. Xu, L. Wang, X. H. Hei, P. Shen, W. Shi, L. Shan, GI/Geom/1 Queue based on Communication Model for Mesh Networks, *International Journal of Communication Systems*, Vol. 27, No. 11, pp. 3013-3029, November, 2014.

[2] W. Wei, X.-L. Yang, P.-Y. Shen, B. Zhou, Holes Detection in Anisotropic Sensornets: Topological Methods, *International Journal of Distributed Sensor Networks*, Vol. 8. No. 10, pp. 135054, October, 2012.

[3] W. Wei, X. Fan, H. Song, X. Fan, J. Yang, Imperfect Information Dynamic Stackelberg Game Based Resource Allocation Using Hidden Markov for Cloud Computing, *IEEE Transactions on Services Computing*, Vol. 11, No. 1, pp. 78-89, January-February, 2018.

[4] H. Song, M. Brandt-Pearce, A 2-D Discrete-time Model of Physical Impairments in Wavelength-division Multiplexing Systems, *Journal of Lightwave Technology*, Vol. 30, No. 5, pp. 713-726, March, 2012.

[5] C. M. Bowman, P. B. Danzig, U. Manber, M. F. Schwartz, Scalable Internet Resource Discovery: Research Problems and Approaches, *Communications of the ACM- Association for Computing Machinery-CACM*, Vol. 37, No. 8, pp. 98-107, August, 1994.

[6] I. Mani, *Automatic Summarization*, John Benjamins Publishing Company, 2001.

[7] A. Hickl, K. Roberts, F. Lacatusu, LCC's GISTexter at DUC 2007: Machine Reading for Update Summarization, *Document Understanding Conference*, Rochester, NY, 2007, pp. 333-342.

[8] M. L. Liu, D. Q. Zheng, T. J. Zhao, Y. Yu, Dynamic Multi-document Summarization Model, *Journal of Software*, Vol. 23, No. 2, pp. 289-298, February, 2012.

[9] M. L. Liu, H. E. Ren, Y. Yu, D. Q. Zheng, T. J. Zhao, Web-based Dynamic Multi-document Summarization System Framework, *Journal of Software*, Vol. 24, No. 5, pp. 1006-1021, May, 2013.

[10] M. L. Liu, T. J. Zhao, D. Q. Zheng, Y. Yu, Research on Dynamic Multi-document Summarization by Topic Detection and Tracking Technology, *Journal of Harbin Institute of Technology*, pp. 1767-1770, November, 2010.

[11] J. Zhang, X. Cheng, H. Xu, Dynamic Summarization: Another Stride Towards Summarization, *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops*, Silicon Valley, CA, 2007, pp. 64-67.

[12] X. Tang, C. C. Yang, Following the Social Media: Aspect Evolution of Online Discussion, *4th International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction,* College Park, MD, 2011, pp. 292-300.

[13] D. R. Radev, H. Jing, M. Budzikowska, Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies, *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization- Volume 4*, Seattle, Washington, 2000, pp. 21-30.

[14] F. L. Wang, C. C. Yang, X. Shi, Multi-document Summarization for Terrorism Information Extraction, *International Conference on Intelligence and Security Informatics*, San Diego, CA, 2006, pp. 602-608.

[15] S. Ye, T. S. Chua, M. Y. Kan, L. Qiu, Document Concept Lattice For Text Understanding and Summarization, *Information Processing & Management*, Vol. 43, No. 6, pp. 1643-1662, November, 2007.

[16] N. Ye, J. Zhu, Y. Zheng, M. Y. Ma, H. Wang, B. Zhang, A Dynamic Programming Model for Text Segmentation Based on Min-Max Similarity, *Asia Information Retrieval Symposium*, Harbin, China, 2008, pp. 141-152.

[17] J. Zhang, X. Cheng, H. Xu, X. Wang, Y. Zeng, ICTCAS's ICTGrasper at TAC 2008: Summarizing Dynamic Information with Signature Terms Based Content Filtering, *Proceedings of the First Text Analysis Conference*, Gaithersburg, Maryland, 2008, pp. 1-9.

[18] G. Salton, A. Singhal, M. Mitra, C. Buckley, Automatic Text Structuring and Summarization, *Information Processing & Management*, Vol. 33, No. 2, pp. 193-207, March, 1997.

[19] J. Allan, Introduction to Topic Detection and Tracking, in: J. Allan (Ed.), *Topic Detection and Tracking*, Springer, Boston, MA, 2002, pp. 1-16.

[20] Y. Yang, J. O. Pedersen, A Comparative Study on Feature Selection in Text Categorization, *ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning*, Nashville, TN, 1997, pp. 412-420.

[21] X. Wan, J. Yang, J. Xiao, Manifold-Ranking Based Topic-Focused Multi-Document Summarization, *IJCAI'07 Proceedings of the 20th International Joint Conference on Artifical Intelligence*, Hyderabad, India, 2007, pp. 2903-2908.

[22] J. Goldstein, V. Mittal, J. Carbonell, M. Kantrowitz, Multi-document Summarization by Sentence Extraction, *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization- Volume 4*, Seattle, Washington, 2000, pp. 40-48.

[23] D.-J. Chiang, Real-Time Data Delivering Based on Prediction Scheme over Internet of Things, *Journal of Internet Technology*, Vol. 18, No. 2, pp. 395-405, March, 2017.

Biographies

Meiling Liu She graduated from Harbin Institute of Technology with doctor of engineering in the field of computer science. Then she research on Information Retrieval in Harbin Engineering University for postdoctoral. She was done the research in USA, during 2014-2015 as a visiting scholar in Virginia Tech.

Wei Wei is faculty of School of Computer Science and Engineering，Xi'an University of Technology, Xi'an 710048, China. He received his Ph.D. and M.S. degrees from Xian Jiaotong University in 2011 and 2005, respectively. His research interest is in the area of wireless networks, WSNs Application, Mobile Computing, Distributed Computing, etc.

Bing Xu, associate professor. She is a member of Chinese Information Processing Society (CIPS). Her research fields include: natural language processing (NLP), sentiment analysis.

Huiqiang Wang received his received M.E. and Ph.D. degrees from HEU in 1985 and 2005, respectively. From 2001 to 2002, he was at Queen's University, Ontario, Canada, as a senior visiting scholar. Now, he is engaged in teaching and researching as a professor and a doctoral advisor at HEU.