ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 6, No. 3, 2011, *pp*.227-233

Dynamic Multi-Document Summarization Research based on Matrix Subspace Analysis Model

Mei-Ling LIU [1+2], De-Quan ZHENG [2], Tie-Jun ZHAO [2], Hong-e REN [1], Yang YU [1]

[1] Department of Computer Science and application, Northeast Forestry University, Harbin, China 150001
[2]Machine Intelligence & Translation Laboratory, Harbin Institute of Technology, Harbin, China 150001

(*Received June 15, 2011, accepted June 20, 2011*)

Abstract. In this paper, we described dynamic evolution of network information, as well as identify and analysis the document collection on the same topic in different stages. Dynamic summarization considers the different documents' temporal relationship in multi-document and analyzes the relationship between emerged information and emerging information. In order to construct a dynamic evolution of content differences, a dynamic multi-document summarization model was presented, called the Matrix Subspace Analysis Method model. On this basis, proposed some efficient dynamic sentence weighting methods, and experiments on the test data of Update Summarization in TAC2008, we showed effectiveness results.

Keywords: Multi-Document Summarization, Otherness analysis, Matrix model, D-TFIDF-T, Dynamic evolvement

1. Introduction

Traditional multi-document summarization technology [1] is a type of static summarization. It generates an abstract for a closed set of static documents without considering the external contact. In the Web2.0 era, the network information that has arisen through BBS、Blog、twitter、online reviews, and new media (such as network topics, hot events, collection expressed as a series of correlation articles) is dynamic. They appear, develop and have their demise as time passes. Topics have different emphases at different time section, but there remains relationship between the subject content.

The biggest difference between dynamic summarization and static summarization is that dynamic summarization needs to consider the different documents temporal relationship in multi-document and analyze the relationship between emerged information and emerging information, then makes a model on the dynamic evolution of the content.

This paper studied the dynamic summarization model based on the dynamic evolution of the environment, given a method for dynamic multi-document summarization model. This model is called the Matrix Subspace Analysis Method (MASM).

2. Related work

2.1. Related research

The basis of dynamic multi-document summarization is the temporal classifying of dynamic content. In the News Information Detection called NID [4], TDT [5] (Topic Detection and Tracking) and other fields, relevant research has been paid more attention. Time information acts as a very important role in Natural Language Processing (NLP) [6], and is the bases of lots of natural language processing tasks, for example, Multi-Document Summarization systems also need to order related information chronologically. The importance of time information makes the research of Temporal Expression Recognition and Normalization (TERN) attract wide attention. Related international evaluation is ACE [7] in the TERN evaluation and so on.

The using of time information vary the research of TDT in various forms, for example, Johan Makkonen added time information to the vector space model of report, and tried to transform the relative time into absolute time [8]. Ziyan Jia etc proposed similarity calculation based on time information and so on. Mani etc

[1] Corresponding author. *E-mail address*: mlliu.sandy08@gmail.com.

analyze the content of news events by using time-domain analysis [9].

Compared to traditional static multi-document summarization, dynamic multi-document summarization is confronted with two problems. One is how to select content and the other is how to control language quality. The difference is that dynamic multi-document summarization process relevant set of dynamic documents. The documents are highly dynamic and evolutional, which means that how to determine the importance, redundancy and coverage of abstract content base on the background of the new timing, and maintain the language quality of abstract will become the core of the problem.

2.2. Main evaluation methods

Currently, the evaluation system for temporal multi-summarization follows the evaluation system for traditional static multi-document summarization entirely, including automatic evaluation ROUGE, BE method and artificial evaluation method PYRAMID [2]. Evaluation of abstracts mainly focuses on to how to select content of abstracts and language quality. Automatic evaluation systems mainly evaluate content of the abstracts, and manual evaluation systems evaluate the choice of content for the abstracts, language quality and overall (considering the topic-oriented coverage and fluency). For the construction of the standard abstract, there are 8 official NIST evaluators writing abstracts for every topic, the topic of each time slice corresponds to four artificial abstracts. Thus, the quality of artificial abstracts performance as the upper limit, and the quality for abstracts of reference system (generally constituted by the first sentence in document) act as the lower limit of system performance. Abstracts content unit selection and comparison is two key issues.

TAC [3] is the most influential international evaluation meeting in multi-document summarization area, which evolves from the DUC and the TREC evaluation that are sponsored by National Institute of Standards and Technology. TAC evaluation is founded by Intelligence Advanced Research Projects Activity and is hosted by the Information Retrieval Group in NIST Information Technology Laboratory each year. It supervised by advisory committee members come from government, businesses and academia. The goal of update summarization evaluation is to evaluate English summarization, and the test corpus mostly comes from the AQUAINT-2 data set in the TREC QA evaluation.

3. Dynamic modelling method

3.1. The basic concept dynamic model

In order to find a model to measure dynamic evolution of content, specifically a model for the difference of content between current document set *Di* and historical document set *D1,...,Di-1(1<=i<=n).* …

The key question of dynamic multi-document summarization is how to denote the evolution content of dynamic information, specifically, it is to find a model for the difference between the current document *Di* set and historical document set *D1,...,Di-1(1<=i<=n)* in the timing document set. For convenience, first this paper given the following definitions:

Definition 1: Current Information was denoted the information of the current document set in the temporal document sequence. We denoted the current information with *Ic*.

Definition 2: Historical Information was denoted the information of the historical document set in the temporal document sequence. We denoted the historical information with *Ih*.

Definition 3: *f* is the mapping from the document space to the abstract space, so the abstract of every document set *Di* in temporal document sequence can be written as *f (Di).*Thus the abstract of historical document set can be expressed as *f (Ih),* and the abstract of current document set can be expressed as *f(Ic).*

According to the definitions above, the dynamic summarization summary can be transformed to find a model for the difference of evolution content that between historical information and current information. The paper analyzed the relationship of historical information and current information, and using document filtering method to characterize the evolution of the dynamic content.

New information can be obtained by the method that gets contents of overlapping historical information *Ih* is filtered from the current information *Ic*, It can be expressed as *Ic-Ih.* Then generate the dynamic abstract *f(Ic-Ih)* by using the static multi-document summarization method. This dynamic summarization model extracted dynamic information to generate abstract by the document filtering method. Considering that an abstract is the representation of document content, in order to save computational cost, this paper can take historical abstract *f (Ih)* replace historical document *Ih.*

3.2. Idea of matrix subspace analysis

In order to grasp the dynamic evolution trend of content and analyze the difference and similarity between historical information and current information, it can start from two aspects. The first is filtering similar content to describe the dynamic evolution, and the second is extracting different content to describe the dynamic evolution. Due to the method of filtering similar content has been studied, this paper would find a model for extracting difference of dynamic content to implement difference model of dynamic content.

In the field of signal processing, optimization for solving many problems can be reduced to extracting a desired signal, and choking back all other interference, clutter or noise. Sub-space projection is an important mathematical tool to solve this problem.

A document set can be viewed as a matrix space, so historical collection of documents is regard as historical information space *A* and current collection of documents is regard as current information space *B*. Space *A* can be broken down into two disjointed subspaces; they are orthogonal subspace *P* and orthogonal complement subspace P_T. The direct sum of the two subspaces is the whole space.

Historical information subspaces can be divided into two mutually orthogonal subspaces *C* and *D*, where *C* is the historical information spaces matrix of the orthogonal subspace *A*, and *D* is the orthogonal complement space of the historical information spaces matrix *A*. Calculation of current information matrix *B* in the *D* projection can be eliminated on the *C's* component with *B*, which is historical information, achieves the purpose of filtering history information.

Therefore, we applied the matrix approach to filter the information of document. The main information subspace contains the abstract of documents, and noise space contains the redundant information.

The main information subspace *C* and noise information subspace *D* is denoted as follows (1) (2):

$$C = A < A, A >^{-1} A^T \tag{1}$$

$$D = I - A < A, A >^{-1} A^T \tag{2}$$

Where: *A* denote the historical information space matrix.

By computing the projection P_TB of current information space *B* on the orthogonal complement subspace P_T of historical information space *A*, we can decrease the weight (the frequency of keywords in the sentence) of current sentences that are highly similar with historical sentences, and the weight of current sentences that are different with historical sentences is keep same. Compared with the value in the *B*, then extract a specified number of sentences which weight in *B* and in P_TB are similar form a new collection of the current document set.

3.3. Matrix Subspace Analysis Method

The essential idea is: according to the various types of training samples generated each corresponding to the subspace by the original model feature space, the basic vector obtained from subspace respectively described various types of mode distribution structural information, so each subspace and each categories correspondence. In this paper, we propose a model of dynamic summarization, which can be established by subspace using algebraic and statistical iterative learning approach.

Subspace method is essentially classification disposal; similarly, document filtering theory of dynamic summarization analyzes structure for the original data and extracts the most important features of each type to achieve feature extraction, data compression, high-dimensional space linear map to low dimensional space. Subspace method increases discrimination representation for every types, thereby the difference in the dynamic document content is identified more efficiently. Subspace method commonly uses inner product operation, so it greatly reduces calculation. Consequently, assume using matrix subspace method to improve document content filtering quality and dynamic.

Realization algorithm for the MSAM is shown as below algorithm 1.

Algorithm 1: Multi-document Summarization Based on MSAM algorithm.

1. First, generated historical abstract to extract the specified number of keywords setting *A* from it, and extracted the specified number of keywords setting *B* from the current document sets. Then the *A* merged the *B* to set *C*, then extracted a specified number of keywords from the theme *C* to Set *D* form this Keywords.

2. The keywords from sets *D* become matrix column, and the sentences collection of historical abstract became matrix row to assemble matrix *X*. The keywords from sets *D* became matrix column, the sentence

collection of each document in the current document sets became matrix row, and formed the current information matrix *Yn* As the number of documentation.

3. Subspace decomposition on the historical information matrix *X*, using the formula $P=X<X, X>^{-1}X^T$ and the formula $P_{\perp}=I-X<X, X>^{-1}X$ calculate the orthogonal space and the orthogonal complement space for the matrix *X*.

4. Fractionation calculated orthogonal projection from *N*-matrix of the current document to the history information they were the matrix *X*, that $Zn=P_{\perp}Xn$.

5. Then evaluated the similarity between the row of *Zn* and Corresponding the row of *Xn*, when the similarity less than a set value, deleted the matrix row corresponding sentences that in this matrix corresponding to the document.

6. Generated dynamic abstract to sets of documents processed using automatic multi-document summarization methods.

4. Sentence weighted method

This paper presented three new sentence weighing methods, which one was sentence weight method based on computation of sentence similarity and the second was dynamic TF-IDF-TIME (called D-TFIDF-T) sentence weighted method, and third was phrase information granularity representation. Moreover this experiment on the MSAM model validated algorithmic feasibility and effectiveness.

4.1. Sentence weight method based on sentence similarity computation

Similarity is a very complex concept and is widely discussed in the semantics, philosophy, and information theory communities. In different specific applications, the meaning of similarity is different. For example, in example-based machine translation, similarity in the main is used to measure the level of text words that can be replaced; in information retrieval, similarity is more a reflection of the degree of compliance in the sense of the text and the user's query; in question answering, similarity reflects the degree of matching between questions and answers; in multi-document summarization system, similarity can reflect the fitting degree of information on local topics.

Sentence weight method was based on a sentence similarity computation; the formula was as (3) shown:

$$Weight(S_i)=\sum_{j=1}^{count} Sim_{ij} \tag{3}$$

Where, Weight *(Si) (0<i<count)* represented weight of sentence *Si*; count represented the total number of sentences in document set.

The Sim_{ij} represented the similarity between *Si* and *Sj*. the formula was as (4) shown:

$$Sim_{ij}=\frac{Sim\_Length(S_i,S_j)}{Length(S_j)} \tag{4}$$

4.2. Sentence weighted method based on D-TFIDF-T

TF-IDF [10] (Term Frequency-Inverse Document Frequency) concept is considered the most important invention in information retrieval. TF-IDF is a common and effective word-weighted algorithm. When using the TF-IDF term carry out weight calculation, TF (Term Frequency) as word frequency, it is used to calculate the word ability to describe the document content. IDF (Inverse Document Frequency) as the anti-document frequency, is used to calculate the word ability of distinguish document. This paper researches dynamic summarization, so insert TIME parameter called *TimeWgt* into the scoring formula.

This paper presented D-TFIDF-T sentence weighted method as the following (5).

$$Score(senti)=\alpha * fWordWgt+\beta * fPosWgt+\gamma * TimeWgt \tag{5}$$

Where: *Score (senti)* was sentence *senti* score; and the *fWordWgt* was as the following (6):

$$fWordWgt=\sum_{k=0}^{count(senti)} TF(word_k)*IDF(word_k) \tag{6}$$

fWordWgt was sentence word weight of *senti*, and $TF(word_k)$ was the frequency of word $word_k$ in multi-

document, $IDF(word_k)$ was the anti-document frequency of word $word_k$, *count(senti)* denoted the count of word in *senti*, *f(w)* was function of frequency statistaics, *SF(w)* was the number of sentences containing word *w* in whole document set. These Parameters were shown as the following (7):

$$TF = f(w) \quad ISF = 1/SF(w) \quad Position\_Weight(senti) = 1/i \tag{7}$$

fPosWgt represented sentence *senti* position weight. *TimeWgt* represented time information value of sentence *senti*, where time represents sort value that the document published date containing sentence *senti* in the document collection, *count (D)* represented document number in the document collection. The formula was as (8) shown:

$$TimeWgt = time / count(D) \tag{8}$$

In the actual calculation process, if SentLength_Weight(senti)>x, SentLength_Weight(senti)=0; if SentLength_Weight(senti)<x, then deleted this sentence.

4.3. The information granularity representation based on the phrase

Information granularity is a concept that reflects the level of information detail level. In order to adapt the detail levels of different subsystems' information needs and set up a different granularity, it is can be describe classification for the knowledge is divided on domain. Information granularity refers to the relative size of an information unit, or roughness, for abstract content that information granularity can be a chapter, paragraph, sentence, events, phrases, keywords, sub-topics and so on.

This paper presented the information granularity representation based on the phrase formula as (9) shows.

$$Weight(senti) = \sum_{j=1}^{lenth(senti)} Phrase\_Weight(j) + SentLength\_Weight(senti) \tag{9}$$

Where: *Phrase_Weight(j)* was phrase weight of sentence, calculation method as (10) shown.

$$Phrase\_Weight(j) = \frac{FR(Phrase)}{MaxFR} \tag{10}$$

Where: *FR (Phrase)* was phrase Frequency, *MaxFR* was the Maximum phrase Frequency, and *SentLength_Weight(senti)* was sentence length weight.

In the actual calculation process, *if SentLength <x,* then deleted this sentence, don't participate in operations; if *SentLength >= x,* then *SentLength_Weight(senti)=0*. Here, *x* was the length threshold of dynamic sentence and can be adjusted according to actual needs.

5. Experimental results and analysis

5.1. experimental data and Evaluation

In the TAC 2008, Update Summarization task test corpus came from the 48 topics in AQUAINT-2; each topic contains two time slices, and was composed of 10 documents. Topic "D0801A" by the two time slices "D0801A-A" and "D0801A-B" components, the 10 within documents were represented by their ID, the topic itself is described by the <title> and <narrative>.

Evaluation criteria used by well-known summarization ROUGE tools; the two most important evaluation targets are the ROUGE-2 and ROUGE-SU4*. Tests were carried out on the Update Summarization test data in TAC2008. The dynamic summarization scoring (R-2) and (R-SU4*) compared with the actual system scores in TAC 2008 Update, and the results showed that the dynamic multi-document summarization method has a good performance.

5.2. Experimental results

Tests did two groups of experiments, experimental group 1 compared evaluation results on three different weighting methods of dynamic multi-document summarization basing on the MSAM model strategy; experimental group 2 compared performance with TAC2008 test system.

Experiment 1 was the ROUGE evaluation results on three different weighting methods of dynamic multi-document summarization basing on the MSAM model strategy. Table 1 showed, in three of the ROUGE scores comparison of weighted algorithms, the ROUGE_2 and ROUGE_SU4 score of the MSAM_2 higher than the score of basis system MSAM_1, indicating that D-TF-IDF-T performance was

better than when using the similarity cumulative weighting approach. The MSAM_3‘s two item scoring were more raise, indicating that the sentence weighted method has better performance based on the phrase-level. Therefore, the sentence weighting method based on the phrase-level was better than these two; it can be more effective in extracting sentences, consequently the abstract was of better quality.

Table 1 MSAM three algorithm performance comparison

System tab	model	Sentence weighting method	R-2	R-SU4*
MSAM_1	Dynamic Subspace	Sentence similarity cumulative	0.04222	0.09033
MSAM_2	Dynamic Subspace	D-TFIDF-T	0.05227	0.10420
MSAM_3	Dynamic Subspace	phrase-level weighting	0.09252	0.12154

Experiment 2 was the comparison that best performance of the MSAM model to the top one, the tenth, and the twentieth of actual system in TAC 2008 Update evaluation task. Table 3 showed that, where the R-SU4*score of sentence weighting method of phrase information granularity in the MSAM performance was very close to the first system, and better than the twentieth. In the MSAM model, R-2 scores of D-TFIDF-T and phrase-level weighting method slightly worse than the first system. Overall performance was in the top 3 of all evaluation systems and is in the forefront of 73 Evaluation systems. It showed that this method has good performance and potential.

Table 2 Performance contrast with TAC2008 system

SYSTEM	R-2	R-SU4*
MSAM_1	0.042	0.090
MSAM_2	0.052	0.104
MSAM_3	0.092	0.121
Rank 1	0.101	0.137
Rank 10	0.089	0.127
Rank 20	0.081	0.119

6. Conclusion

The MSAM started from the public, its emphases was information space constructed by document set. Historical document set and current document set are proposed as historical information space and current information space, respectively. This model applied matrix Subspace theory to identify overlapping space and otherness space. Then filtered the overlapping space and leave over the otherness space.

Dynamic model of multi-document summarization is a new subject, and is currently in its initial stages. This paper carefully studied the latest developments in the field of multi-document summarization at home and abroad, carries through different analysis to the evolution relationship of dynamic content, using content filtering method described evolution content, thereby proposing a dynamic summarization model. Sentence weighing has been improved, and the phrase-based information granularity had a better performance. Experiments on test data of Update Summarization in TAC2008 showed that the proposed model of dynamic multi-document summarization and generation method are valid. The next step will be continuing to research

different models and methods, and the static summarization algorithm performance should be improved in order to get higher scores in dynamic summarization.

7. Acknowledgements

Here, we thank the teachers and schoolmates who have given support and advice, especially, the guidance of the Harbin Institute of Technology Computer Science and Technology Associate Professor Zheng Dequan and Professor Zhao Tiejun.

SUPPORTED: by the National Natural Science Foundation of China under Grant No. 61073130; the National High-Tech Research and Development Plan of China under Grant Nos.863, 2006AA010108; the Fundamental Research Funds for the Central Universities No.DL10BB27

8. References

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

[2] Shu Zhang, Tiejun Zhao, Hao Yu, Hua Zhao. The Research on the Influence of the Types of Document Sets on Multi-document Summarization. *Journal of Computational Information Systems*. 2007, **3**(3): 1201~1206.

[3] H. Dang, K. Owczarzak. Overview of the Tac 2008 Update Summarization Task[C]//Text Analysis Conference. 2008.

[4] J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Lavrenko, R. Hoberman, and D. Caputo. *Topic-based novelty detection. Baltimore: Center for Language and Speech Processing*. Johns Hopkins University. Technical report: ws99, 1999.

[5] J. Allan, R. Papka, and V. Lavrenko. *On-line New Event Detection and Tracking.* In Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, Melbourne, Australia. 1998, pp. 37~45.

[6] I. Mani. *Recent Developments in Temporal Information Extraction (draft)[J]*.

[7] Nicolov N. and Mitkov, R. Proceedings of RANLP, 2004, 3.

[8] http://projects.ldc.upenn.edu/ace/intro.html

[9] J. Makkonen. Investigations on Event Evolution in TDT. *Proceedings of Student Workshop of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.* Edmonton, Canada, 2003: 43~48

[10] I. Mani and G. Wilson. Robust temporal processing of news. *In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Hong Kong. 2000, pp. 69~76.