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.*