INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN: 2278 – 0211 (Online)

An Ontology Based Anatomy Approach To Text Mining Summarization

Dr. Anadakumar. K
Assistant Professor (Selection Grade), Department Of Computer Application, Bannari Amman Institute Of Technology,India
Ms. Padmavathy. V
Research Scholar, Department Of Computer Science, Dr.S.N.S. Rajalakshmi College Of Arts And Science

Abstract:

Nowadays many events posted on the internet. It provides us abundant resources. The user may find some difficulties in extracting the most informative summary and the associative core parts of the topic which is defined temporary. Thus, as a result, summarization is used as a technique for improving querying. To ensure this technique an anatomy based summarization method called Topic Summarization and Content Anatomy (TSCAN) was proposed to summarize the content of a temporal topic in existing work. A temporal similarity (TS) function is applied to generate the event dependencies and context similarity to form an evolution graph of the topic. In this paper, we compare two methods for article summarization. The first method is mainly based on term-frequency, while the second method is based on ontology. We build an ontology database for analyzing the main topics of the article using NPL tool and protégé tool. Protégé can be customized to provide domain-friendly support for creating knowledge models and entering data.

Keywords: *Coherence, Text mining, Topic anatomy, TSCAN.*

1.Introduction

1.1.Topic Anatomy

A topic is a real world incident that consists of one or more themes, which are related to a finer incident, a description, or a dialogue of a certain issue. Topic anatomy is an emerging text mining research issue that involves three major tasks: Theme generation, Event segmentation and summarization, and Evolution graph construction.

1.1.1.Defining Themes

The content of a topic is comprised of several simultaneous themes, each representing an episode of the topic. The theme generation process tries to identify the themes of a topic from the related documents. A theme of a topic is derived from a collection of blocks.

1.1.2.Defining Events

An event is defined as a disjoint sub-episode of a theme. The event segmentation and summarization process extracts topic events and their summaries by analyzing the intension variation of themes over time.

1.1.3.Constructing Evolution Graph

Context similarities of all of the events and themes are calculated and an evolution graph is formed by associating all of the events and themes according to the temporal closeness of each of the events and themes of the document. From this we can analyze the performance, precision, recall rate etc.., by comparing the existing system and proposed system.

1.2.Text Segmentation

The objective of text segmentation is to partition an input text into non-overlapping segments such that each segment is a subject-coherent unit, and any two adjacent units represent different subjects. Depending on the type of input text, segmentation can be classified as story boundary detection or document subtopic identification. The input for story boundary detection is usually a text stream, e.g., automatic speech recognition transcripts from online newswires, which do not contain distinct boundaries between documents. Generally, naive approaches, such as using cue phrases, can identify the boundaries between documents efficiently. For document subtopic identification, the input is a single document, and the task involves identifying paragraphs in the document

that relate to a certain subtopic. Document subtopic identification enables many information systems to provide fine-grained services. Topic segmentation differs from document subtopic identification in a number of respects. First, the input for topic segmentation is a set of documents related to a topic, rather than a single document used in document subtopic identification. Second, the identified segments of a topic, i.e., the events of themes, have a temporal property rather than a textual paragraph or several contiguous paragraphs in a document. Finally, the segments of a document are disjoint textual units, but the events of a topic can overlap temporally.

1.3.Text Summarization

Generic text summarization automatically creates a condensed version of one or more documents that captures the gist of the documents. As a document's content may contain many themes, generic summarization methods concentrate on extending the summary's diversity to provide wider coverage of the content. In this study, we focus on extraction-based generic text summarization, which composes summaries by extracting informative sentences from the original documents. Their proposed method [8] allows the user to search for specific types of information (for example, opinion, fact or encyclopedic knowledge). Therefore, this proposed method produces summaries according to the type of information specified by the user as well as the topics of the documents. Text structure is also producing more balanced and coherent output summaries. The three main aspects of the problem in this dissertation are as follows: A. Extracting balanced contents of the source documents. B. Summarization to discriminate between types of information (fact, opinion, and knowledge) that the user's desire to know. C. Generating output summaries to improve the readability and reduce redundancy. We used text structure and document genre to extract the important sentences from the source documents in *A* and *B*, while we used text structure of output summaries to produce summaries in *C*.

1.4.System Architecture

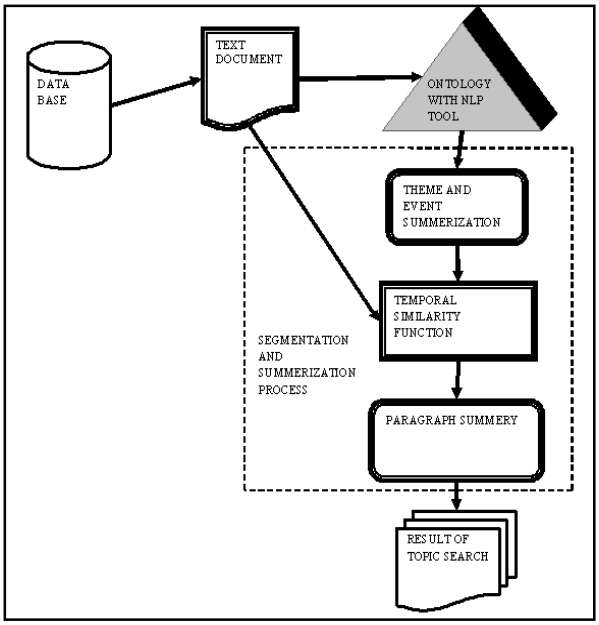

Figure 1: Segmentation and Summarization process

2.Ontology

Applications of ontology-related techniques have become increasingly popular in recent years. Nevertheless, there is no unique definition of ontology in literature yet. We use Gruber's definition of ontology: "ontology is an explicit specification of some topics. It is a formal and declarative representation, which includes the vocabulary (or names) for referring to the terms in a specific subject area and the logical statements that describe what terms are, how they are related to each other." Essentially, the ontology decomposes the world into several objects for describing them. The determination of the way we describe objects and the formalism of representation depend on individual applications. In this paper, the ontology is designed for analyzing and gathering the semantic information of a class of article. Assuming every article contains several subtopics; we use the ontology for identifying subtopics of articles, and encode each of these possible subtopics by a non-overlapping portion of the ontology.

2.1.The Need For Ontology

We notice that all the above mentioned work assumes that all information provided by different sources to be integrated is covered by a domain model. However, information is not necessarily presented in the same way. Due to this fact, information exchange is not

an easy task if different actors (producers or consumers of information) have not agreed on the semantic of data. It is necessary then to define an "alphabet" to ensure a good interpretation and understanding of exchanged data. The role of the ontology is to provide a common model that ensures the minimal requirements for this purpose. In fact, such a model allows one to construct a common view of different sources. The elements in the model are described in a way independent from the particularity of the data source. One has to note that the more an application domain is restricted, the more it is possible to elaborate a precise description of the domain with the help of an ontology, and the more the processing may be refined. This is achieved mainly with the help of a domain's meta-data.

Ontology is an explicit specification of some topic. Ontology is a way to discompose a world into objects, and a way to describe these objects. This is a partial description of the world, depending on the objectives of the designer and the requirements of the application or system. For each domain, there may be a number of ontologies. The use of ontology differs from an application to another, so are its design and its formalism of representation.

2.2.Construct Ontology

In this module, first we will collect vocabularies and synonyms. Next, we put those words by the Data model of ontology. The first step of our method is to determine the main subtopics of the article of interest. This is achieved by comparing the words of articles with terms in the ontology. If the word does not exist in the ontology, we ignore it. Otherwise, we record the number of times the word appears in the ontology we encode the ontology with a tree structure, and each node includes the concepts represented by the node's children. When the count of any node increases, the counts associated with their ancestors will also in-crease.

After marking the counts of the nodes in the ontology, we select second-level nodes that have higher counts as the main subtopics of the article. Generally speaking, one article is composed of several subtopics, so our system will select multiple subtopics. There are limited topics an article can contain, and a reasonable summary probably should include fewer. Therefore, we only choose a limited number of subtopics and ignore others. We choose to ignore the subtopic if its count is less than 10. In addition, we choose only top three or required subtopics. After obtaining the subtopics, our system will use them for selecting paragraphs as the summary.

3.Topic Model

A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. A topic is a real world incident that comprises one or more themes, which are related to a finer incident, a description, or a dialogue about a certain issue. During the lifespan of a topic, one theme may attract more attention than the others, and is thus reported by more documents. The proposed method identifies themes and events from the topic's documents, and connects associated events to form the topic's evolution graph. In addition, the identified events are summarized to help readers better comprehend the storyline(s) of the topic. A topic is represented explicitly by a collection of chronologically ordered documents. In this study, we assume that the documents are published in the same order as the events of the topic reported by independent authors, and that there is no inconsistency between the contents of the documents.

TSCAN decomposes each document into a sequence of non overlapping blocks. A block can be several consecutive sentences, or one or more paragraphs. We define a block as w consecutive sentences. For a topic, be a set of stemmed vocabulary without stop words. The topic can then be described by an m _ n term-block association matrix B in which the columns represent the blocks decomposed chronologically from the topic documents.

3.1.Theme Generation

A matrix, called a block association matrix, is symmetric matrix in which the entry is the inner product of columns i and j in matrix B. As a column of B is the term vector of a block, A represents the inter block association. Hence, entries with a large value imply a high correlation between the corresponding pair of blocks.

A theme of a topic is regarded as an aggregated semantic profile of a collection of blocks, and can be represented as a vector v of dimension n, where each entry denotes the degree of correlation of a block to the theme. Given the constitution of a vector computes the theme's association to the topic's content. The objective function of our theme generation process determines entry values so that the acquired theme is closely associated with the topic.

3.2.Event Segmentation And Summarization

The tasks of our event segmentation and speech endpoint detection are similar in that they both try to identify important segments of sequential data. In addition, it is the

amplitude of sequential data that determines the data's importance. For example, given the speech utterance, the speech endpoint detection task involves distinguishing the significant segment S2 from the insignificant silent segments mixed with background noise. Here, S2 represents the word "one" and comprises a sequence of points with large positive and negative amplitudes.

To segment events, the R-S algorithm examines the amplitude variation of an eigenvector to find the endpoints that partition the theme into a set of significant events. In the R-S algorithm, every block in an eigenvector has an energy value. To calculate the energy, we adopt the square sum scheme, which has proved effective in detecting endpoints in noisy speech environments.

3. 3.Evolution Graph Construction

Automatic induction of event dependencies is often difficult due to the lack of sufficient domain knowledge and effective knowledge induction mechanisms. However, as event dependencies usually involve similar contextual information, such as the same locations and person names, they can be identified through word usage analysis.

Our approach, which is based on this rationale, involves two procedures. First, we link events segmented from the same theme sequentially to reflect the theme's development. Then, we use a temporal similarity function to capture the dependencies of events in different themes. For two events, ei and ej, belonging to different themes, we calculate their temporal similarity between these two events and providing the graph description from the result.

4.1.Performance Evaluations

4.1.Summarization Evaluations

We compare the summarization performance of Summarization with Ontology method with the following six well-known summarization methods.

- The forward method, which generates summaries by extracting the initial blocks of a topic.
- The backward method, which extracts summaries from the end blocks of a topic. This is frequently used as the baseline method in DUC contests.

- The SVD method [14], which composes summaries by extracting the blocks with the largest entry value in singular vectors.
- The K-means method, which compiles summaries by selecting the most salient blocks of the resulting K clusters. Generally, this method's performance depends on the quality of the initial clusters.
- The temporal summary (TS) method, where we adopt the useful2 and novel1 techniques proposed by the authors to compute the in formativeness score of a topic block. We do not adopt the novel2 technique because the authors have shown that the performance difference between using novel1 and using novel2 is not significant. In addition, novel2 requires a training corpus to derive an appropriate number of clusters (i.e., parameter m), but the training corpus is not available.
- The frequent content word (FCW) method, which constructs summaries by selecting blocks with frequent terms. This method's performance is comparable to that of state-of-the-art summarization methods.

4. 2.Scalability And Time Comparisons

We evaluated the execution time of the compared summarization methods on an AMD AthlonTM 64 Processor 3200++ PC with the Windows XP Service Pack 3 operating system and a 2 GB main memory. For each method, we recorded the time required to generate the summaries of the 26 evaluated topics under a specific parameter setting. We employ MATLAB to calculate a matrix's eigenvectors. The time complexity is O(n2I2), where I is the number of eigenvectors to be computed. the linear summarization methods run faster than the other methods. In terms of time complexity, the Eigen-based **method** run slower than the Ontology method.

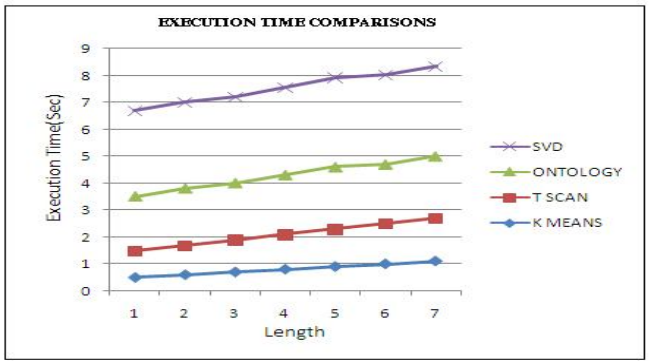

Figure 2: An Execution time comparison

5.Conclusion

The system we have built is a knowledge-based summarization system with the knowledge of topics coming from ontology. In this project, the ontology knowledge approach was presented, the approach based on feature appraisal and NLP application in summarization The knowledge is composed of not only in recognizing important topics in the document, but also in recognizing the relationships and the relationship types that exist between them. This extracted knowledge is represented in the form of evolution graph.

6.Reference

1. J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang,"Topic Detection and Tracking Pilot Study: Final Report," Proc. US Defense Advanced Research Projects Agency (DARPA) Broadcast News Transcription and Understanding Workshop, pp. 194-218, 1998.
2. V. Hatzivassiloglou, L. Gravano, and A. Maganti, "An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering," Proc. 23rd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 224-231, 2000.
3. C.D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge Univ. Press, 2008.
4. Y. Yang, T. Pierce, and J. Carbonell, "A Study on Retrospective and Online Event Detection," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 28-36, 1998.
5. C.C. Chen, M.C. Chen, and M.S. Chen, "An Adaptive Threshold Framework for Event Detection Using HMM-Based Life Profiles," ACM Trans. Information Systems, vol. 27, no. 2, pp. 1-35, 2009.
6. Q. Mei and C.X. Zhai, "Discovering Evolutionary Theme Patterns from Text—An Exploration of Temporal Text Mining," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining, pp. 198-207, 2005.182.
7. S. Strassel and M. Glenn, "Creating the Annotated TDT4 Y2003 Evaluation Corpus, "http://www.itl.nist.gov/iad/mig/tests/tdt/2003/papers/ldc.ppt, 2003.
8. L E. Spence, A.J. Insel, and S.H. Friedberg, Elementary Linear Algebra, a Matrix Approach. Prentice Hall, 2000.
9. M.A. Hearst and C. Plaunt, "Subtopic Structuring for Full-Length Document Access," Proc. 16th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 59-68, 1993.
10. X. Ji and H. Zha, "Domain-Independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, p*p. 322-329, 2003*
11. FISCUS, G. J. AND DODDINGTON, G. Topic detection and tracking evaluation overview. In Topic Detection and Tracking: Event-Based Information Organization. Kluwer Academic Press, 17–30, 2002.

12. C. Mario, D. C. Luigi, and S. Claudio. Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation. In IWMDM, pages 1–10, 2010.
13. I. S. Dhillon and D. S. Modha. Concept Decompositions for Large Sparse Text Data using Clustering. Machine Learning, 42(1):143–175, 2001