International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

# Temporal Topic Summarization and Content Anatomy Based on Ontology Method

**B. L. Prabhu[1], M. Parveentaj[2]**

[1]Research Scholar, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore- 05, India

[2]Assistant Professor, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore- 05, India

**Abstract:** ***An anatomy based summarization method called Topic Summarization and Content Anatomy (TSCAN) was proposed to summarize the content of a temporal topic. TSCAN models the documents as a symmetric block association matrix, in which each block is a portion of a document, and treats each eigenvector of the matrix as a theme embedded in the topic. A temporal similarity (TS) function is applied to generate the event dependencies and context similarity to form an evolution graph of the topic. A unique feature of TSCAN is the introduction of the event segmentation process to extract the semantic construct event before summarization. An ontology database is used 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. Specifically the Natural language processing is the process of a computer extracting meaningful information from natural language input and/or producing natural language output. After identifying the main topics and determining their relative significance, rank the paragraphs based on the relevance between main topics and each individual paragraph. Depending on the ranks, we choose desired proportion of Para-graphs as summary.***

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

## 1. Introduction

Searching the web is a vital role in human life in last two decades. The users either search for exact information or just surf topics which interest them. Naturally, users enter a query in natural language, or as a set of keywords, and a search engine answers with a set of documents which are relevant to the query. Then, the users need to go through the documents to find the information that interests them. Only some parts of the document contain query-relevant information in the retrieved document. An advantage to the user would be if the system selected the relevant passages, put them together, made it summarizing and fluent, and returned the resulting text. Furthermore, if the resulting summary is not relevant enough, the user can refine the query. Thus, as a result, summarization is used as a technique for improving querying.

## 2. Methodology

### 2.1. 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 [10] that involves three major tasks: Theme generation, Event segmentation and summarization, and Evolution graph construction.

#### 2.1.1. 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.

#### 2.1.2. 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.

#### 2.1.3. 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.

### 2.2 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.

### 2.3. 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 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. [1] 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* [2][4].

### 2.4. System Architecture

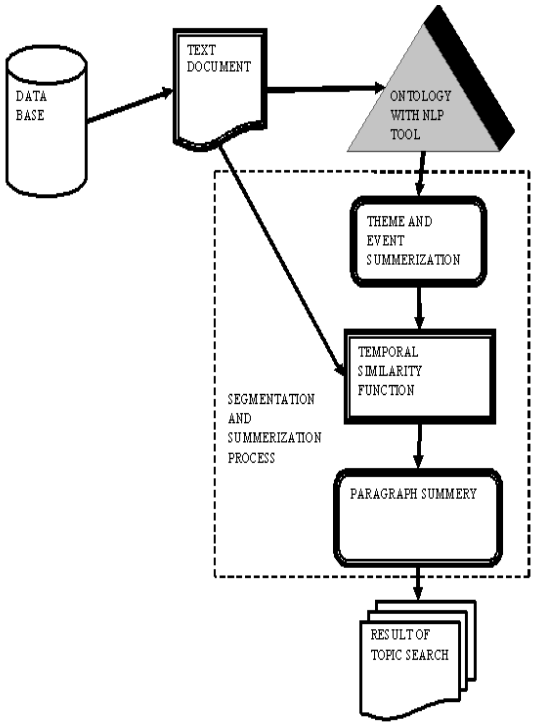



**Figure 1:** Segmentation and Summarization process

## 3. Ontology

Applications of ontology-related techniques have become increasingly popular in recent years. Nevertheless, there is no unique definition of ontology in literature yet. It uses 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 describes objects and the formalism of representation depends 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; use the ontology for identifying subtopics of articles, and encode each of these possible subtopics by a non-overlapping portion of the ontology.[3]

### 3.1. The Need for Ontology

It notices 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 [11].

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.

## 4. An Ontology based Anatomy Approach

In this work first, collect vocabularies and synonyms. Then, put those terms by the Data model of ontology. The first step of ontology based TSCAN approach is to determine the by comparing the terms of documents with terms in the ontology. If the term does not exist in the ontology, ignore it. Otherwise, record the number of times the word appears in the ontology [4]. The ontology decomposes the specific domain into several objects for describing them.

The determination of the way describes objects and the formalism of representation depends on individual applications. In this paper, the ontology is designed for analyzing and gathering the semantic information of a class of article. Assuming every document contains several subtopics; use the ontology for identifying subtopics of document, and encode each of these possible subtopics by a non-overlapping portion of the ontology. After selecting the blocks using ontology construct symmetric block association matrix and finding the Eigen value from that matrix. Many different contents and structures exist in constructed

ontologies, including those that exist in the same domain. If extant domain ontologies can be used, time and money can be saved. However, domain knowledge changes fast. In addition, the extant domain ontologies may require updates to solve domain problems. The reuse of extant ontologies is an important topic for their application. Thus, the integration of extant domain ontologies is of considerable importance. Through the method, two extant ontologies can be converted into a fuzzy ontology. The new fuzzy ontology is more flexible than a general ontology. The experimental results indicate that our method can merge domain ontologies effectively. A Temporal Similarity (TS) function is applied to generate the event dependencies and context similarity to form an evolution graph of the topic.

### 4.1 Construct ontology

In this module, first will collected vocabularies and synonyms. Next, 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, ignore it [12]. Otherwise, record the number of times the word appears in the ontology encodes 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, 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 system will select multiple subtopics. There are limited topics an article can contain, and a reasonable summary probably should include fewer. Therefore, only choose a limited number of subtopics and ignore others [6]. It chooses to ignore the subtopic if its count is less than 10. In addition, choose only top three or required subtopics. After obtaining the subtopics, our system will use them for selecting paragraphs as the summary.

### 4.2 Topic Detection and Tracking (TDT)

Topic detection and Tracking is a multi-site research project, now in its third phase, to develop core technologies for a new understanding system. Specifically, TDT systems discover the topical structure in unsegmented streams of topics as it appears across multiple topics and in different themes and events. It is used to develop automatic techniques for finding topically related material in streams of data [5]. This could be valuable in a wide variety of applications where efficient and timely information access is important. It would be very helpful if computers were able to map out topics automatically finding story boundaries, determining what stories go with one another, and discovering when something new (unforeseen) has happened.

In this project use an ontology based approach to improve the TSCAN approach. The idea of using ontology is to analyze documents and obtain semantic information before constructing symmetric block association matrix. It constructs an ontology database for analyzing the main topics of the document [25]. After recognizing the main topics and determining their relative significance, rank the paragraphs based on the relevance between main topics and each individual paragraph. Depending on the ranks, decide preferred proportion of paragraphs as summary.

## 5. NPL Tool Overview

Natural language processing (**NLP**) is a field of computer science, artificial intelligence (also called machine learning), and linguistics concerned with the interactions between computers and human (natural) languages. In theory, natural language processing is a very attractive method of human–computer interaction. Natural language understanding is sometimes referred to as an AI-complete problem because it seems to require extensive knowledge about the outside world and the ability to manipulate it [6].

Whether NLP is distinct from, or identical to, the field of computational linguistics is a matter of perspective. The Association for Computational Linguistics defines the latter as focusing on the theoretical aspects of NLP. On the other hand, the open-access journal "Computational Linguistics", styles itself as "the longest running publication devoted exclusively to the design and analysis of natural language processing systems".

Modern NLP algorithms are grounded in machine learning, especially statistical machine learning. Research into modern statistical NLP algorithms requires an understanding of a number of disparate fields, including linguistics, computer science, and statistics. For a discussion of the types of algorithms currently used in NLP, see the article on pattern recognition.

## 6. 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 [7]. A topic is represented explicitly by a collection of chronologically ordered documents. In this study, 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. It defines 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.

## 7. 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 [8]. Given the constitution of a vector computes the theme's association to the topic's content. The objective function of theme generation process determines entry values so that the acquired theme is closely associated with the topic.

## 8. 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.

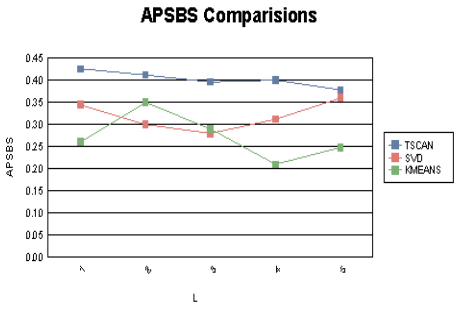
Here, S2 represents the word "one" and comprises a sequence of points with large positive and negative amplitudes. Therefore, adopt Rabiner and Sambur's R-S endpoint detection algorithm for event segmentation. To segment events, the R-S algorithm examines the amplitude variation of an eigenvector to 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 [10]. To calculate the energy, adopt the square sum scheme, which has proved effective in detecting endpoints in noisy speech environments.

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

## 10. Experiment Results

Summary-to-document content similarity (SDCS) is defined as the average cosine similarity between an evaluated summary and the topic documents. Both components are represented by TF-IDF term vectors. A high similarity score implies that the summary is representative of the topic and can effectively replace the original topic documents for various information retrieval tasks. Parameter w controls the granularity of topic blocks. In the preprocessing phase of the experiments, observed that the sentence segmentation program supplied by DUC sometimes segments sentences incorrectly when dealing with noun abbreviations followed by a period. The superior SDCS scores achieved by our method demonstrate the advantage of using event segmentation for temporal topic summarization.




3/5/2012

| Alnam | TSCA | SVD | KMEA |
|---|---|---|---|
| 1 | 0.43 | 0.35 | 0.26 |
| 2 | 0.41 | 0.30 | 0.35 |
| 3 | 0.40 | 0.28 | 0.29 |
| 4 | 0.40 | 0.31 | 0.21 |

**Figure 2:** APSBS Comparisons

Although our summaries are not as diverse as those of the K-means method, they are more coherent. A popular measurement frequently used to judge the content coherence of a set of documents is the average pair wise document similarity TSCAN achieves superior APSBS (Average Pair wise Summary Block Similarity) scores[9]. The reason is that our summaries focus on events in the first few significant themes; therefore, summary blocks have similar contexts. By contrast, the summaries compiled by other approaches try to cover diverse themes, so they are less coherent than our summaries. For all the summarization methods, APSBS decreases as the size of summaries increases. As a large summary covers many themes, its content is more diverse than that of a small summary. Hence, the average pair wise similarity will be low

## 11. 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. Even without the summary, just looking at the nodes and relationships in the graph gives us an idea about what the document is taking about. A summary however gives us the actual details of the topic search. This is the first system that uses ontological knowledge in this manner to obtain extractive summaries of topics. After identifying the main topics and determining their relative significance, we rank the paragraphs based on the relevance between main topics and each individual paragraph.

Depending on the ranks, we choose desired proportion of Para-graphs as summary. Experimental results indicate that both methods offer similar accuracy in their selections of the paragraphs. In the future, we will research other method to determine the relationships between concepts more accurately instead of the above simple method and improve the method of ontology construction in a large data set.

## References

[1] Wenjie Li, Mingli Wu and Qin Lu (The Hong Kong Polytechnic University) and Wei Xu and Chunfa Yuan (Tsinghua University). Extractive Summarization using Inter- and Intra- Event Relevance, 2006
[2] Ani Nenkova. Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference, 2005.
[3] Maciej Janik, Krys J. Kochut. Wikipedia in action: Ontological Knowledge in Text Categorization.
[4] Leonhard Hennig, Winfried Umbrath , Robert Wetzker. An Ontology-based Approach to Text Summarization, 2008
[5] Hele-Mai Haav. A semi-automatic method to ontology design by using FCA, Edited by Vaclav Snasel, Radim Belohlavek, In: Proc. of the CLA 2004 lnfl. Workshop on Concept Lattices and their Applications Ostrava, Czech Republic, Sept. 2004, 13-24.
[6] Lixin Han, Guihai Chen. A fuzzy clustering method of construction of ontology-based user profiles,Advances in Engineering Software, 2009, 535-540,
[7] Mingli Feng. Construction of User-Query Semantic Ontology (UQSO) for Personalized Topic Search Engine. Xihua University, The thesis of master degree, 2010.
[8] P. Cimiano, G. Stumme, A. Hotho, J. Tane, Conceptual knowledge processing with formal concept analysis and ontologies. In: Proc. of the Second Intl. Conf. on Formal Concept Analysis (ICFCA 04), Springer, 2004, 189-207
[9] E.A. Kendall, “Role models – patterns of agent system analysis and design,” In: British Telecom (BT) Technical Journal, Vol. 17, No. 4, Springer, 1999, pp. 46-56.
[10] J.O. Kephart, D.M. Chess, “The vision of autonomic computing,” In: IEEE Computer,Vol. 36, No. 1, IEEE, 2003, pp. 41-50.
[11] B.B. Kristensen, “Object-oriented modeling with roles,” In: Proceeding of the International Conference on Object-Oriented Information Systems (OOIS), Springer, 1995, pp. 57-71.
[12] H. Knublauch, M.A. Musen, A.L. Rector, “Editing description logic ontologies with the Protégé-OWL plugin,” In: Proceedings of the International Workshop on Description Logics (DL), http://CEUR-WS.org/Vol-104/, CEUR-WS.org, 2004, #8.

## Author Profile

**B. L. Prabhu** has completed MSc (CS) and pursuing M. Phil (CS) in Sri Jayendra Saraswathy Maha Vidyalaya CAS. His areas of Interest are Networking, Data Mining, and Software Testing.

**M. Parveentaj** M.C.A., ADCA, M.Phil, working as an Associate Professor in Department of Computer Science – Sri Jayendra Saraswathy Maha Vidyalaya CAS for past 8 years. Area of Interest: Networking, Data mining, Software Engineering. She has presented 2 papers in an International Conference, 1 paper in National Level.