Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Content Anatomy Approach and Ranking to Temporal Topic

[1]D M Abhinay Kanth, [2] S M Farooq Mtech

Dept of CSE, JNTUA

India

Abstract-- In the internet searching for a specific topic it contains some type of events and activities .The topic contains more documents in the search engine about a specific type of topic, topic anatomy is define summarizes and associates the core parts of a topic temporally so that readers can understand the content easily. The existing topic anatomy model, called TSCAN, evolves the major themes of a topic from the eigenvectors of a temporal block association matrix. Then, the similarity of events of the themes and their summaries are extracted by examining the constitution of the eigenvectors. Finally, the events which are extracted are associated through their temporal closeness and context similarity to form an evolution graph of the topic. Experiments found on the official TDT4 corpus demonstrate that the generated temporal summaries present the storylines of topics in a comprehensible form. The proposed framework contains ranking for a specific topics and contents, by this process it give usefulness of topic anatomy.

Keywords: searching topic, text mining, language summarization, ranking, graph construction

I. Introduction

With the diverse and explosive growth of Web information, how to coordinate and employ the information effectively and efficiently has become more and more critical. This is peculiarly important for Web 2.0 associated applications since user-generated information is more freestyle and less structured, which increases the problematic in mining useful information from these data sources. In order to satisfy the information needs of Web users and improve the user experience in many Web applications, Recommends Systems, have been well studied in academia and widely deployed in industry. The phenomenal growth in the number of documents posted on the Internet provides an abundant source of information as an alternative to traditional media. While present technologies are efficient in searching for appropriate documents to satisfy keyword search requests, users still have problematic assimilating needed knowledge from the overwhelming number of documents. The situation is even more obscure if the desired knowledge is related to a temporal incident about which many independent authors have published documents based on various perspectives that, considered unitedly, detail the development of the incident. To promote research on detecting and tracking incidents from Internet documents, the Defense Advanced Research Projects Agency (DARPA) initiated the Topic Detection and Tracking (TDT) project [1]. The project defines a topic as "a seminal event or activity, along with all pointed related events and activities." Its goal is to detect topics automatically and track related documents from several document streams, such as online news streaming. The TDT project has generated a great deal of interest due to the importance and practical implications of the problem. For instance, the Google News service employs TDT techniques to organize documents related to news topics from online news websites [2], [3]. While an effective TDT system can detect topics and track all related documents [1], [4], [5], users cannot fully comprehend a topic unless they read many of the tagged documents. For popular topics, such as "Kobe versus LeBron in NBA MVP race" shown in Fig. 1, the number of covered documents is simply too large for users to apprehend. Hence, there is an urgent need for effective summarization methods to extract the core parts of observed topics, as well as graphic representation methods to depict the relationships between the core parts. Applied together, the two types of techniques, called topic anatomy, can summarize essential information about a topic in a structured manner. Topic anatomy is an emerging text mining research paradigm that involves three major tasks: theme generation, event segmentation and summarization, and evolution graph construction. Generally, the content of a topic is comprised of several coincidental themes, each representing an episode of the topic [6]. The theme generation process tries to identify the themes of a topic from the related documents. Over the lifetime of a topic, the focus of the topic's content may shift from one theme to another to reflect the topic's development [6]. We define an event as a disjoint subdivide of a theme. The event segmentation and summarization process extracts topic events and their summaries by analyzing the intension variation of themes over time. Events may be related semantically because they are temporally close or share similar contexts, e.g., they may relate to the same named entities. By joining the associations, the constructed evolution graph reveals the storylines of the topic.

II. Related Work

2.1 Text Segmentation

The objective of text segmentation is to partition an input text into nonoverlapping segments such that each segment is a subject-coherent unit, and any two adjacent units represent different types of subjects [9]. Depending on the type of input text, segmentation can be defined as story boundary detection or document subtopic recognisation. The input for story boundary detection is usually a text stream, e.g., automatic speech identification of transcripts from online newswires, which do not contain discrete boundaries between documents. Generally, naive approaches, such as using cue phrases, can recognize the boundaries between documents efficiently [10]. For document subtopic recognize, the input is a single document, and the task involves recognizing paragraphs in the document that relate to a certain subtopic. Document subtopic recognisation enables many information systems to provide fine-grained services. For example, search the sub divided topics in a document are often similar; hence, salient cue phrases about subtopic boundaries are virtually nonexistent [10], [11]. However, one major problem with this approach is that the information in the blocks is usually insufficient to determine the blocks' interrelationships. Brants et al. [11] and Choi et al. [12] applied the concept of latent semantics to enrich the information in a block. Text Summarization is 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 carries many themes, generic summarization methods concentrate on extending the summary's diversity to provide wider coverage of the content [14]. In this study, we focus on extraction based generic text summarization, which composes summaries by taking out the informative sentences from the original documents. Topic Evolution Mining is introduced Kleinberg [26] developed a topic evolution mining technique that constructs a hierarchical tree from a series of topic documents. The technique uses a HMM-based, two-state transition diagram to model the status of topics and splits a topic into diverse themes, patterned as tree branches, if the topic contains splited information. Nallapati et al. [27] formalized the problem of topic evolution mining as a text clustering task in which the recognised clusters, i.e., the events of a topic, are connected chronologically to form an evolution graph of the topic. In addition to constructing a graph, Mei and Zhai [6] modeled the activeness trend of identified themes.

Feng and Allan [29] proposed an incident threading method that is similar to TSCAN system. The method first identifies incidents (i.e., events) from news documents, then, the semantic dependencies between the incidents are examined to produce an incident network. The authors also defines hand-crafted rules and an optimization procedure to assign types to network links. Experiments exhibit that link type assignment is a ambitious task, and better modeling of natural languages is required to improve the technique's accuracy. Swan and Allan [30] proposed a timeline system to display important topics in a document corpus graphically. The system uses a statistical feature selection method to identify terms that occur frequently in a specific time period. The recognized term-period groups, i.e., topics, are then coordinate sequentially to form the timeline of the corpus. This timeline system can be employed to a document corpus, so the results is similar to that of the TDT project. In contrast, we concentrate on a single topic and documents specifically related to that topic.

2.2 TSCAN System

It introduces Topic Model contains topic is a real world incident that comprises one or more themes, which are associated to a finer incident, a description, or a dialogue about a particular issue. During the lifetime of a topic, one theme may attract more attention than the others, and is thus described by more documents. We specify an event as a significant theme development that continues for a period of time. Naturally, all the events taken unitedly form the storyline(s) of the topic. Although the events of a theme are temporally disunite, they are considered semantically dependent in order to express the development of the theme. Moreover, events in varied themes may be associated because of their temporal proximity and context similarity. The TSCAN method identifies themes and events from the topic's documents, and connects associated events to create the topic's evolution graph. In addition, the recognized events are summarized to help readers better comprehend the storyline(s) of the topic. Fig.1 illustrates

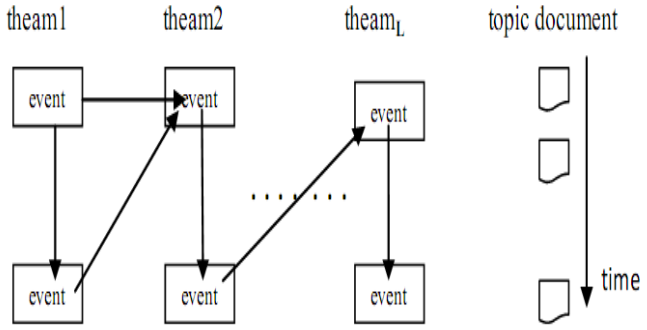

Fig.1. The relationships between themes, events, and event dependencies.

The relationships between the themes, events, and event dependencies of a topic in the proposed model. A topic is represented explicitly by a collection of chronologically arranged documents. In this study, we assume that the documents are published in the same order as the events of the topic reported by individual authors, and that there is no inconsistency between the contents of the documents. TSCAN disintegrates each document into a sequence of nonoverlapping blocks. A block can be several successive sentences, or one or more paragraphs. We define a block as w successive sentences. For a topic, let T = {t1, t2, . . . , tm} be a set of stemmed vocabulary without stop words [31]. The topic can then be described by an m×n term-block association matrix B in which the columns {b1, b2, . . . , bn} represent the blocks decomposed chronologically from the topic documents. In other words, for any two blocks, bi and bj, if i < j, then either the document containing bi was published before the document containing bj, or bi appears before bj in the same document.

2.3 Theme Generation

A matrix A =B^TB, called a block association matrix, is an n ×n symmetric matrix in which the (i,j)-entry (denoted as ai;j) 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 interlock 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
presented as a vector v of dimension n, where each entry denotes the degree of correlation of a block to the theme.

III. PROPOSED SYSTEM

3.1 Graph construction

The exponential explosion of different contents produced on the Web, Recommendation techniques have become more and more crucial. Innumerable various types of recommendations are made on the Web each day, as well as movies, music, images, books recommendations, query suggestions, tags recommendations, etc. In this no issue what types of data sources are used for the recommendations, basically these data sources can be designed in the form of different types of graphs [6]. In graph building consider an undirected graph G={V,E},where V is the vertex set, and V={V_1,V_2,……,V_n} , E is the set of all edges. In this node contain query and edge contain user resource location (URL). The value on the edges is specify how many times a query is clicked on a URL. This module is responsible to take click-through data of American online (AOL) search engine [7] or Flickr data and extract bipartite graph from data which is undirected in nature. The bipartite graph is converted into another form of bipartite graph where each undirected edge becomes two directed edges. That web graphs are normally very huge, and so that generally algorithm will be performed on a sub graph extracted from the original graph. Hence, it is necessary to evaluate the size of the sub graph which affects the recommendation accuracy. The performance changes with different sub graph sizes. It is observed that when the size of the graph is very small, like 500, the performance of our algorithm is not accurate since this sub graph must ignore some very relevant nodes. However, when the size of sub graph is increasing, the performance also increases. It is also noticed that the performance on sub graph with size of 5,000 is very close to the performance with size of 1, 00,000. This indicates that the nodes that are far away from the query node are normally not relevant with the query node. The sub graph is designed by using depth-first search in the original graph. The search stops when the number of nodes is larger than a predefined number.

3.2 Data Corpus

In [37], two case studies using the official TDT (topic detection and tracking) topics are provided to demonstrate that the evolution graphs constructed by TSCAN can draw out the themes, events, and event dependencies of the examined topics successfully. In this research, we valuate our anatomy-based summarization technique by comparing the derived summaries with those of several text summarization methods. We utilize the official TDT4 topics for the executed evaluations. The Linguistic Data Consortium has compiled a series of TDT corpora for the annual TDT contends. The TDT4 corpus comprises 28,390 English news documents from eight well-known news agencies for the period 1 October 2000 to 31 January 2001. Among them, 70 news events with 1,926 associated documents were labeled by NIST annotators for various TDT evaluation tasks. The annotators also compiled factual descriptions of the topics, which are regarded as human-compiled reference summaries for summarization evaluations. Although Document Understanding Conferences1 (DUC) also use TDT topics for summarization contends, the average size of the topics is only 10 documents, which is too small for the purpose of topic anatomy. We therefore select 26 TDT4 topics, each containing more than 20 documents, for evaluation.

TABLE1: Statistics of Evaluated Topics

Number of topic	26
Number of news topics	1,211
Average no of topics documents per topic	46.6
Number of sentence	32,739
Average number of sentence per topic	1,259.5

Table 1 details the evaluated topics in our data corpus. In the preprocessing phase, each topic document is partitioned into blocks of sentences by using a simple Perl script2 supplied by DUC. The system parameters H and w mapped, respectively, the length of the sliding window used to aggregate the energy of a block in the R-S endpoint detection algorithm and the number of sentences in a block. To assess the influence of H and w on the summarization performance, they are set at {5, 7, 9} and {1, 3, 5}, respectively. The parameter L is critical to the quality of detected themes. The function U(L), defined in (12), is used to measure the underestimation of VL when approximating the interrelations of a topic's blocks U(L) is the average of the squared differences between A and VLDLV T L . A low U(L) value indicates that the selected L themes represent the interblock associations sufficiently well. From (6), it is clear that the larger the number of themes selected, the lower will be the value of U(L). However, a large L may be a drawback because the constructed evolution graph may have too many themes to be comprehensible. For summarization comparison, the evaluations are performed with L =1 to 10 in order to illustrate the influence of themes on the summarization performance. 26 evaluated topics for L =1 to 10. It is noteworthy that, contrary to expectations, blocks with little content information (i.e., w = 1) produce a low undervalue. This is because the block association matrix constructed with smallsize blocks is very sparse. Thus, the slight difference between A and VLDLV T L reduces the value of U(L) substantially.

3.3 Summary-to-Document Content Similarity

Summary-to-document content similarity is defined as the average cosine similarity between an evaluated summary and the topic documents. Both components are mapped by TF-IDF term vectors. A high similarity score involves that the summary is representative of the topic and can effectively replace the original topic documents for various information recollect tasks. Fig. 8 shows the micro average summary-to-document content similarity scores derived by the compared methods. As shown in the figure, our method outperforms the compared methods with small L values.

When w=1, the TS method yields superior SDCS scores. This is because the useful2 technique determines the in formativeness of a block by calculating the probability that the block is generated by the language model of the block's document. Therefore, the choosed summary blocks are highly similar to the documents they are drawn out from, which increases the SDCS scores. It is interesting to note that the SDCS score of TS decreases as w increases. This is because topic documents generally report the latest information about a topic, so the contents of the documents would be dissimilar. Even though the summary blocks of the TS method are highly similar to the documents they are drawn out from, they are somewhat dissimilar to the other topic documents. A large w increases the content of a block and this also magnifies the dissimilarity. The K-means method achieves a higher similarity score because its summary provides better coverage of the topic's contents. Our method simply selects the top L significant themes (L <<r) to correspond a topic, whereas the K-means method partitions all of a topic's content into K clusters and extracts the most salient block from each cluster to map the topic. As a result, summaries constructed by the K-means method provide better content coverage, and the similarity score increases as more clusters are used to partition the content. However, without an efficient mechanism, such as the structure of themes and events, to leverage and forms the summarized results, large K values indicate that the summaries are unstructured; therefore, they would be effortful for users to understand [37]. Compared to the SVD method, which is also a vector-based summarization method, the higher-ranking SDCS scores achieved by our method demonstrate the advantage of using event segmentation for temporal topic summarization. The SVD method does not reckon the temporal information (i.e., events) of topics and strives to increase the diversity of summaries by including a lot of side information extracted from minor singular vectors. As a result, the constructed summaries deviate from the core content of the topics, which impacts the SDCS performance. The summaries extracted by the frequent content word method are based on a multinomial model constructed from all the topic documents. In the preprocessing phase of the experiments, we noticed that the sentence segmentation program supplied by DUC sometimes segments sentences incorrectly when dealing with noun abbreviations followed by a period. As a result, the setting w ¼ 1 yields inferior SDCS performances (see Fig. 10) because each summary block contains a short segmented sentence that may not convey complete topic information. The virtually equal SDCS performance scores under w ¼ 3 and w ¼ 5 also show that the summary size or block size does not affect the coverage of the drawn out summaries provided that the extracted blocks contain enough topic information.

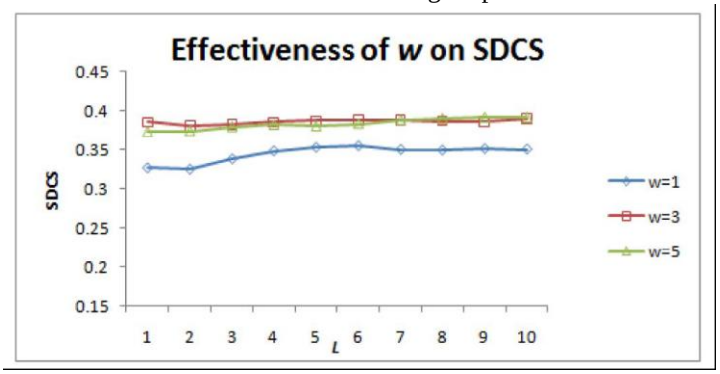

Fig. 2 The effect of parameter w on the SDCS metric's performance

3.3 RANKING

Table2: Result of searching query through ranking

Testing Queries	Suggestions				
	Top 1	Top 2	Top 3	Top 4	Top 5
michael jordan	nba	nike	jordan xi	air jordans	michael jordan bio
java	sun java	java download	java updates	virtual machine	sun microsystems
apple	itunes	ipod	quicktime	apple ipod	apple stores
fitness	exercise	fitness magazine	muscle and fitness	mens fitness	weight loss
solar system	planets	jupiter	saturn	neptune	pluto
sunglasses	chanel sunglasses	oakley	maui jim sunglasses	designer sunglasses	oakley sunglasses
flower delivery	flowers	florist	gift baskets	cheap flowers	proflowers
wedding	wedding channel	wedding dresses	the knot	wedding plans	wedding poems
astronomy	apod	star charts	planets	solar system	skyandtelescope
real estate	remax	realtor	homes for sale	coldwell banker	houses for sale

Through the existing process follows the ranking process for every searching content topic, documents and also update topics. By this process we can easily retrieved by the search engine and satisfy the user needs. In this process click through data process is important concept for searching. After retrieving results from existing process applied the ranking to result. In this process is done by click-through data process. This process is applied in any search engine lick Google, yahoo,…..In this process after TSCAN ,calculate the weigh(click-through data) of the results ,which node contains high weight ,that node display in top of the result and contains rank1 for that result

4. RESULT

By using ranking for the existing system we easily gather the searching results. In this process it require less time and also efficacy more. Here semantic results also involved in ranking process. In this process first data sets are constructed to graph, the graph is directed graph .Second after graph construction TSCAN is applied for every node and URL for results. Third after TSCAN process apply ranking for each and every result in second step. The results of ranking give good performance in searching process and text anatomy .The process provide hidden documents and contents in searching process.

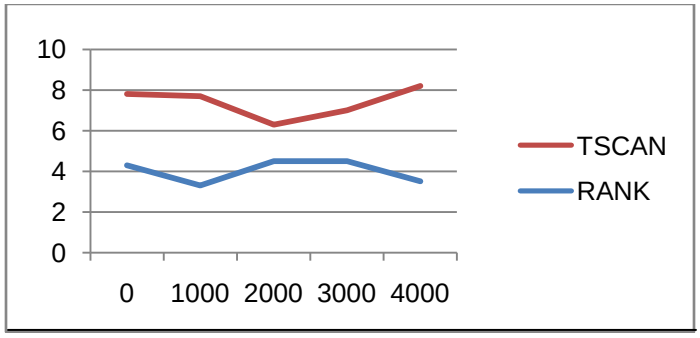

Figure3: Comparison Graph

IV. Conclusion

We discuss the various types of Experiments based on the official corpus demonstrate that the generated temporal summaries present the storylines of topics in a comprehensible form. The proposed framework contains ranking for a specific topics and contents, by this process it give usefulness of topic anatomy.

References

[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 Investiga- tion 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. C arbonell, "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.

[7]. L E. Spence, A.J. Insel, and S.H. Friedberg, Elementary Linear Algebra, a Matrix Approach. Prentice Hall, 2000.

[8] 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.

[9] 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, pp. 322-329, 2003.

[10] T. Brants, F. Chen, and I. Tsochantaridis, "Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis," Proc. 11th Int'l Conf. Information and Knowledge Management, pp. 211-218, 2002.

[11] F.Y.Y. Choi, P. Wiemer-Hastings, and J. Moore, "Latent Semantic Analysis for Text Segmentation," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 109-117, 2001.

[12]. E. Agichtein, E. Brill, and S. Dumais, "Improving Web SearchRanking by Incorporating User Behavior Information," SIGIR '07:Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 19-26, 2006.

[13] E. Auchard, "Flickr to Map the World's Latest Photo Hotspots," Proc. Reuters, 2007.

[14] R. TiberiBaeza-Yates and A. Tiberi, "Extracting Semantic Relations from Query Logs," KDD '07: Proc. 13th ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining, pp. 76-85, 2007.

[15] R.A. Baeza-Yates, C.A. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Current Trends in Database Technology (EDBT) Workshops, pp. 588- 596, 2004.

[16] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," KDD '00: Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 407-416,2000.

[17] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimension- ality Reduction and Data Representation," Neural Computation, vol. 15, no. 6, pp. 1373-1396, 2003.

[18] J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI), 1998.

