ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 10, December 2012

A Survey on content Anatomy Approach to Temporal Topic Summarization

P.Sathyashree[1] **A.Mansoor ali[2]** **D.Mansoor hussian[3]**

[1] II year M.E SE, Department of Computer Science and Engineering (PG),
S.N.S College of Technology, Sathy main road,
Coimbatore 641035.

[2] II year M.E CSE, Department of Computer Science and Engineering (PG),
S.N.S College of Technology, Sathy main road,
Coimbatore 641035.

[3]Professor,. Department of Computer Science and Engineering,
Sri Krishna College of Engineering & Technology, covai pudur,
Coimbatore 641035.

Abstract

A content anatomy approach is an emerging research area in Content mining. In this approach, content anatomy approach defines a task called topic anatomy. Topic anatomy summarizes and associates the core parts of a topic temporally so that the readers understand the content easily. In This paper a survey on summarization techniques for content anatomy approach has been presented. Such as, forward method, backward method, SVD method, K-means method, Temporal summary (TS) method, frequent content word method (FCW), TSCAN.

***Index Terms*—text mining, content anatomy, pattern mining.**

INTRODUCTION

A text summarizer strives to produce a condensed representation of its input, intended for human consumption. It may condense individual documents or groups of documents. Text compression, a related area, also condenses documents, but summarization differs in that its output is intended to be human-readable. The output of text compression algorithms is certainly not human-readable, but neither is it actionable the only operation it supports is decompression, that is, automatic reconstruction of the original text. As a field, summarization differs from many other forms of text mining in that there are people, namely professional abstractors, who are skilled in the art of producing summaries and carry out the task as part of their professional life. Studies of these people and the way they work provide valuable insights for automatic summarization.

Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge. Just as data mining can be loosely described as looking for patterns in data, Text mining is about looking for patterns in text. However, the superficial similarity between the two conceals real differences. Data mining can be more fully characterized as the extraction of implicit, previously unknown, and potentially useful information from data [Witten and Frank, 2000]. The information is implicit in the input data: it is hidden, unknown, and could hardly be extracted without recourse to automatic techniques of data mining. With content mining, however, the information to be extracted is clearly and explicitly stated in the content. It's not hidden at all most authors go to great pains to make sure that they express themselves clearly and unambiguously and, from a human point of view, the only sense in which it is "previously unknown" is that human resource restrictions make it infeasible for people to read the content themselves. The problem, of course, is that the information is not couched in a manner that is amenable to automatic processing. Text mining strives to bring it out of the text in a form that is suitable for consumption by computers directly, with no need for a human intermediary.

Though there is a clear difference philosophically, from the computer's point of view, the problems are quite similar. Content is just as opaque as raw data when it comes to extracting information probably more so. Another requirement that is common to both data and text mining is that the information extracted should be

"potentially useful." In one sense, this means actionable capable of providing a basis for actions to be taken automatically. In the case of data mining, this notion can be expressed in a relatively domain-independent way: actionable patterns are ones that allow non-trivial predictions to be made on new data from the same source. Performance can be measured by counting successes and failures, statistical techniques can be applied to compare different data mining methods on the same problem, and so on. However, in many content mining situations it is far harder to characterize what "actionable" means in a way that is independent of the particular domain at hand. This makes it difficult to find fair and objective measures of success In many data mining applications, "potentially useful" is given a different interpretation: the key for success is that the information extracted must be comprehensible in that it helps to explain the data. This is necessary whenever the result is intended for human consumption rather than (or as well as) a basis for automatic action. This criterion is less applicable to content mining because, unlike data mining, the input itself is comprehensible. Content mining with comprehensible output is tantamount to summarizing salient features from a large body of content, which is a subfield in its own right content summarization.

The main problem in text mining is to find the accurate content in searching. In existing system, forward method, backward method, SVD method, K-means method, Temporal summary (TS) method, frequent content word method (FCW), TSCAN algorithm has been presented. These algorithms are used to find the close content for discovering text.

RELATED WORK

Earlier works on summarization methods has been expansively studied in text mining communities for many years. A Variety of efficient algorithm are used. Such as, forward method, backward method, SVD method, K-means method, Temporal summary (TS) method, frequent content word method (FCW), TSCAN has been proposed. The main problem in text mining is finding the closed pattern. These techniques are used for summarize the content.

Forward method

Forward method is a summarization method, for discovering content. In forward method, summarization is done by using initial block of content. In this method, it will consider only initial block of text. This is main drawback of this method.

Backward method

Backward method is a summarization method, for discovering content. In backward method, summarization is done by using end block of content. In this method, it will consider only end block of text. This is main drawback of this method.

The forward–backward algorithm computes a set of forward probabilities which provide, for all $k \in \{1, \ldots, t\}$, the probability of ending up in any particular state given the first k observations in the sequence, i.e. $P(X_k \mid o_{1:k})$. In the second pass, the algorithm computes a set of backward probabilities which provide the probability of observing the remaining observations given any starting point k, i.e. $P(o_{k+1:t} \mid X_k)$. These two sets of probability distributions can then be combined to obtain the distribution over states at any specific point in time given the entire observation sequence

$$P(X_k \mid o_{1:t}) = P(X_k \mid o_{1:k}, o_{k+1:t}) \propto P(X_k \mid o_{k+1:t}) P(X_k \mid o_{1:k})$$

The last step follows from an application of Bayes' rule and the conditional independence of $o_{k+1:t}$ and $o_{1:k}$given X_k.

As outlined above, the algorithm involves three steps:

1. computing forward probabilities
2. computing backward probabilities
3. computing smoothed values.

The forward and backward steps may also be called "forward message pass" and "backward message pass" - these terms are due to the *message-passing* used in general belief propagation approaches. At each single observation in the sequence, probabilities to be used for calculations at the next observation are computed. The smoothing step can be calculated simultaneously during the backward pass. This step allows the algorithm to take into account any past observations of output for computing more accurate results.

The forward–backward algorithm can be used to find the most likely state for any point in time. It cannot, however, be used to find the most likely sequence of states

SVD method

This method uses a particularly efficient algorithm for singular value decomposition that can handle even very large input matrices (of word counts and documents).

Assume matrix A represents an m x n word occurrence matrix where m is the number of input documents (files) and n the number of words selected for analysis. SVD computes the m x

r orthogonal matrix U, n x r orthogonal matrix V, and r x r matrix D, so that A = UDV', and so that r is the number of eigenvalues of A'A.

For most Text Mining problems, the SVD will be entirely appropriate to use. Without a data reduction technique, there will be more variables (terms) available than one can use in a data mining model. Some method must be applied to select an appropriate set from which a text mining solution can be built. Unlike term elimination, the SVD technique allows one to derive significantly fewer variables from the original variables. There are some drawbacks to using the SVD, however. Computationally, the SVD is fairly resource intensive and requires a large amount of RAM. The user must have access to these resources in order for the decomposition to be obtained **S**VD method is used to composes the summaries by extracting the blocks with the largest entry value in singular vectors. SVD method is using graph based summarization method. Note that the result derived by the SVD method is identical to that of the graph based summarization method,

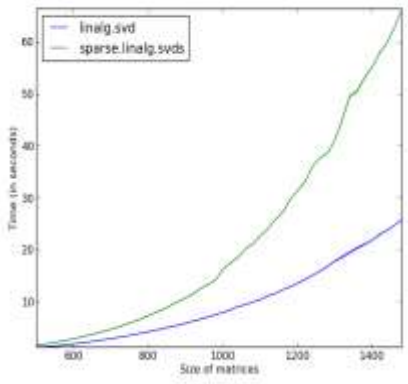

Figure1:SVD method algorithm

K-means method

The *k*-means algorithm is used for efficiency in clustering large data sets. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. The *k*-means algorithm (MacQueen, 1967; Anderberg, 1973), one of the mostly used clustering algorithms, is classified as a partition or non-hierarchical clustering method. it can be used to cluster texts. K-means algorithm is an algorithm to partition and classify the data based on attributes or features in to k-number of groups.The *k*-means algorithm has the following important properties:

1. It is efficient in processing large data sets.
2. It often terminates at a local optimum (MacQueen, 1967; Selim and Ismail,1984).
3. It works only on numeric values.

The K-means method which compiles summaries by selecting the most salient blocks of the resulting K clusters. This method's performance depends on the quality of the initial clusters. In this experiment, to ensure fair comparison of the K- means method, which provide the best result from 50 randomly selected initial clusters for evaluation.

Of the two key steps of the K-Means algorithm, the assignment step consists of assigning each data point to that cluster from whose center the data point is the closest. That is, during assignment, user compute the distance between the data point and each of the current cluster centers. This process assign the data sample on the basis of the minimum value of the computed distance. The second step consists of re-computing the cluster centers for the newly modified clusters.

Obviously, before the two-step approach can proceed, we need to initialize the both the cluster center values and the clusters that can then be iteratively modified by the two-step algorithmThe data points are then projected on to this direction and a histogram constructed from the projections. Centers of the smoothed histogram are used to seed the clustering operation. The other option, which is the older option, is to choose the cluster centers purely randomly. user get the first option if user set cluster_seeding to smart in the constructor, and user get the second option if user set it to random.

How to specify K is one of the most vexing issues in any approach to clustering. In some case, we can set K on the basis of prior knowledge. But, more often than not, no such prior knowledge is available. When the programmer does not explicitly specify a value for K, the approach taken in the current implementation is to try all possible values between 2 and some largest possible value that makes statistical sense. We then choose that value for K which yields the best value for the QoC (Quality of Clustering) metric. It is generally believed that the largest value for K should not exceed sqrt(N/2) where N is the number of data point to be clustered.

How to set the QoC metric is obviously a critical issue unto itself. In the current implementation, the value of QoC is a ratio of the average radius of the clusters and the average distance between the cluster centers. But note that this is a good criterion only when the data exhibits the same variance in all directions. When the data variance is different directions, but still remains the same for all clusters, a more appropriate QoC can be formulated using other distance metrics such as the Mahalanobis distance.

Every iterative algorithm requires a stopping criterion. The criterion implemented here is that we stop iterations when there is no re-assignment of the data points during the assignment step.

Ordinarily, the output produced by a K-Means clusterer will correspond to a local minimum for the QoC values, as opposed to a global minimum. The current implementation protects against that when the clusterer constructor is called with the random option for cluster_seeding, but only in a very small way, by trying different randomly selected initial cluster centers and then selecting the one that gives the best overall QoC value.

Temporal summary (TS) method

Temporal summary method is one of the summarization methods for content discovery. The temporal summary (TS) method take on the useful2 and novel1 techniques proposed by the authors to compute the in formativeness score of a topic block. we do not take on 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.

Frequent content word method (FCW)

Frequent content method is used to construct the summaries by using selecting the block with frequent terms This method's performance is comparable to that of state-of-the-art summarization methods. In addition, we adopt Nenkova et al.'s context adjustment technique to increase the summary diversity.

TSCAN

TSCAN method stands for Topic Summarization and Content Anatomy (TSCAN), which organizes and summarizes the content of a temporal topic by using set of documents. 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. The eigenvectors are then examined to extract events and their summaries from each theme.

The eigenvector are used for calculate the probability for extracting the content. Then, temporal similarity (TS) function is applied to generate the event dependencies, which are then used to construct the evolution graph of the topic. The results of experiments on the official TDT4 corpus demonstrate that our anatomy-based summaries are highly representative. Moreover, they are more consistent with human composed summaries than those derived by other text summarization methods. hat involves three major tasks: theme generation, event segmentation and summarization, and evolution graph construction.

TSCAN method are used to help the internet users graph grasp the gist of a topic covered by a large number of topic documents, text summarization methods have been proposed to highlight the core information in the documents. Most summarization methods try to increase the diversity of summaries to cover all the important information in the original documents.

CONCLUSION

Text summarization methods are used to summarize the text document for extracting the content. Text summarization methods have been proposed to highlight the core information in the documents. Most summarization methods are tried to increase the diversity of summaries to cover all the important information in the original documents. The main issue in Text summarization are to find the related content. Most summarization methods are proposed to temporal properties of the topic, which will have the storylines. In this paper, Forward method, Backward method, SVD method, K-means method, Temporal summary (TS) method, frequent content word method (FCW), TSCAN has been discussed. Forward method is only considering the initial block of the text. Backward method is only considering the end block of the text. SVD method The K-means method, SVD method, TS method, and frequent content word method focus on increasing summary coverage by using clusters, singular vectors, block novelty, and context adjustment, correspondingly. TSCAN method is used to find important events in themes to achieve both summary diversity and narrative tracing properties. TSCAN algorithm is used to increase the performance better than other summarization methods. Ontology are used in the study of concern about what kinds of things exist what entities are in universe. TSCAN algorithm was developed by using ontology, which will improve the better performance in text summarization.

REFERENCES

[1]. J. Allan, R. Gupta, and V. Khandelwal, "Temporal Summaries of News Topic," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 10-18, 2001.

[2] Chien Chin Chen and Meng Chang Chen, "TSCAN: A Content Anatomy Approach to Temporal Topic Summarization" IEEE transactions on knowledge and data engineering, VOL. 24, NO. 1, January 2012.

[3] T. Nomoto and Y. Matsumoto, "A New Approach to Unsupervised Text Summarization," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 26-34, 2001.

[4] Russ Albright, Ph.D, "Taming Text with the SVD", SAS Institute Inc., Cary, NC, January, 2004
[5] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic Text Structuring and Summarization," Advances in Automatic Text Summarization, The MIT Press, 1999.
[6] Somya Srivastava R. Uday Kiran P. Krishna Reddy, "Discovering Diverse-Frequent Patterns in Transactional Databases", 2011
[8]. Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, no. 1, August 2009.
[9] H. Zha, "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering," Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 113-120, 2002.
[10] Zhexue huang, "Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values" Data Mining and Knowledge Discovery 2, 283–304 (1998)°c 1998 Kluwer Academic Publishers. Manufactured in The Netherlands.
.