IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
21-23 September 2009, Rende (Cosenza), Italy

Text Mining: Finding Hot Topics TF*PDF vs. LSI

Dr. J. Katyayani [1], A.V.Sriharsha [2] Dr. B. Sudhir [3]
[1] Sri Padmavathi Mahila Visva Vidyalayam, Tirupati, India
[2] Sree Vidyanikethan Engineering College, Tirupati, India
[3] Sri Venkateswara University, Tirupati, India

***Abstract* – With the vast amount of digital text materials available on the Net, it is almost impractical for people to absorb all related information in a timely manner. This problem has been overcome by erstwhile researchers and scientists of data mining. The efficiency in the methods and exploratory analysis has to be ascertained yet. Document wise term frequencies and inverted frequencies are available to calculate the statistical importance among the documents. Determining the time line importance of the documents plays very essential role than just finding the document's importance. LSI is a basic PCA approach, which is proposed with time-line approach and has been discussed comparatively in this paper.**

***Keywords* – Text mining, dimensionality reduction, latent-semantic indexing, IR.**

I. INTRODUCTION

Most previous studies of data mining have focused on structured data, such as relational, transactional and data warehouse data. The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail and the World Wide Web (which can also be viewed as a huge, interconnected, dynamic text database). The Web keeps growing and huge amount of new information are being posted on it continuously. Weekly, tens or hundreds of Megabytes of news stories can be added easily to the news archive of any newswire sources online. At the same time containing some influencing knowledge, this news archive may also be holding many uninteresting or trivial news. The influencing knowledge is desired but reading the news archive is rather a daunting task that will take us a lot of time and effort. And yet, this doesn't promise us that all the main topics will be discovered. So, it would be helpful if there is kind of system which would respond correctly to the generic queries such as "What's new?" or "What's important?" Unfortunately, traditional goal-driven retrieval system works well only for content-based queries.

It is very useful or efficient when the user knows precisely the goal or facts he/she is seeking. However, we are at a higher level of abstraction and creating the precise goals with zero knowledge of past week's news is rather unrealistic.

Hence, what would be desirable is an intelligent system that automatically summarizes us a weekly report of the main topics embedded in the archive of newswire sources on the Web.

Although timely access to information is becoming increasingly important in today's knowledge-based economy gaining such access is no longer a problem because of the widespread availability of broadband in both homes and businesses. Ironically, high speed connectivity and the explosion in the volume of digitized textual content available online has given rise to a new problem, namely information overload. Clearly, the capacity for humans to assimilate such vast amounts of information is limited. Topic detection has emerged as a promising research area that harnesses the power of modern computing to address this new problem. Topic Detection is a sub process of Topic Detection and Tracking (TDT) that attempts to identify "topics" by exploring and organizing the content of textual materials, thereby enabling us to aggregate disparate pieces of information into manageable clusters automatically. In the context of news, Topic Detection can be viewed as an event detection that groups stories into a corpus, wherein each group represents a single topic.

II. PREVIOUS WORKS

A. What is a Topic?

A Topic is defined as a seminal event or activity, along with all directly related events and activities [5]. A TDT Event is defined as something that happens at a specific time and place, along with all necessary preconditions and unavoidable consequences [6]. Such an event might be a car accident, a meeting, or a court hearing. A TDT activity is a connected series of events with a common focus or purpose that happens in specific places during a given time period [6].

B. What is a Hot Topic?

In [7], a "hot topic" is defined as a topic that appears frequently over a period of time. The "hotness" of a

topic depends on two factors: how often hot term appears in a document and the number of documents that contain those terms. Moreover, no topic can remain hot indefinitely; in other words, every topic goes through a life cycle of birth, growth, maturity and death. Hence, the "hotness" of each topic evolves over a given period of time. In the case of news, topics have different levels of popularity or "hotness" Some are so hot that every news channel broadcasts them and reports on them in great detail, whereas others that are not popular are only reported by a few channels.

Regardless of the peak level of "hotness", news topics eventually "cool off" and are replaced by other more up-to-date stories.

C. Text Indexing Techniques:

There are several popular text retrieval indexing techniques, including inverted indices and signature files. An inverted index is an index structure that maintains two hash-indexed or B+ tree indexed tables: document table and term table, where document table consists of a set of document records, each containing two fields:

- *doc id* and posting list, where posting list is a list of terms (or pointers to terms) that occur in the document, sorted according to some relevance measure.
- *term table* consists of a set of term records, each containing two fields: term id and
- *posting list*, where posting list specifies a list of document identifiers in which the term appears.

A *signature file* is a file that stores a *signature* record for each document in the database. Each signature has a fixed size of *b* bits representing terms. A simple encoding scheme goes as follows. Each bit of a document signature is initialized to 0. A bit is set to 1 if the term it represents appears in the document.

D. Dimensionality Reduction for Text:

With the similarity metrics specified in the literatures, we can construct similarity based indices on text documents. Text-based queries can then be represented as vectors, which can be used to search for their nearest neighbors in a document collection. However, for any nontrivial document database, the number of terms *T* and the number of documents *D* are usually quite large. Such high dimensionality leads to the problem of inefficient computation, since the resulting frequency table will have size $T \times D$. Furthermore, the high dimensionality also leads to very sparse vectors and increases the difficulty in detecting and exploiting the relationships among terms (e.g., synonymy).

To overcome these problems, dimensionality reduction techniques such as *latent semantic indexing*, *probabilistic latent semantic analysis*, and *locality preserving indexing* can be used.

E. Latent Semantic Indexing

Latent semantic indexing (LSI) is one of the most popular algorithms for document dimensionality reduction. It is fundamentally based on SVD (singular value decomposition). Suppose the rank of the term-document X is *r*, then LSI decomposes X using SVD as follows:

$$X = U \sum V^T$$

III. PROPOSED WORK

Information technology has been extremely geared to produce capacious information driven by computerized documents, the problem of how to reduce the tedious burden of reading them arises in the crest of information processing. Automatic text summarization is one solution to the problem, providing users with a condensed version of an original text. This is quite practicable with the techniques of finding inferences and importance of the text data from the text databases.

The work [1] has been discretely sequentiated with various procedures of preprocessing, statistical analysis of life time of a term, variance, hotness of word, identifying the hot sentences, sentence modeling, vectorization and agglomerative clustering. This is can be denoted as a conventional method for text miner to process text and identify the hot topics, whereas using LSI can simplify the job to an extent.

A. Latent Semantic Indexing vs. Conventional Method:

As described in the previous works, the most basic result of the initial indexing of words found in the input documents is a frequency table with simple counts, i.e., the number of times that different words occur in each input document. TDT is a mechanism to find the topics and understand the information provided by huge text databases. As systems are scored by comparing the system result to a manually composed *ground truth*. The systems must provide the satisfactory results of analysis on the text data, reading or comprehending the text. The categorical division of text with respect to the topic of the text is very essential to elucidate. Clustering is the technique to generate categorical collection of topics or concepts from the huge text databases. The cost of a (cluster) structure defines the 'distance'; a better structure has a lower cost. The ground truth is composed by annotators of the Linguistic Data Consortium and consists of manually labelled clusters containing news stories discussing a particular topic. A topic is defined as an event or activity, along with all directly related events and activities as referred in the previous section. The topics are selected from timeline collection of documents from the corpus.

Sentence analysis and extraction requires similarity of sentences. Sentence similarity can be calculated from gross weight between terms appearing in a sentence and other vectors. When we estimate similarity of sentences, we have to consider three problems, *how to estimate similarities of terms*, *how to identify the meaning of terms* and *how to calculate sentence similarity from them*. These problems cannot be addressed using TF*PDF and Hierarchical clustering techniques proposed in [1].

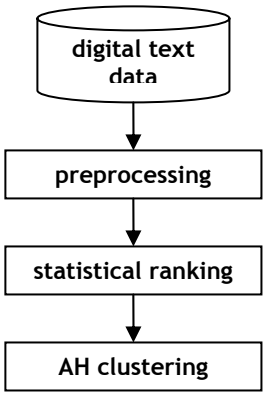

Fig. 1. Experimental procedure described by [1].

External methods for finding the similarity between the sentences are used. The statistical ranking includes finding the variance, FO and VO and New Weights and subsequently finding the hot terms and hot topics. This is restricted to the modeling of sentences and terms consisting appropriate new weights. By using LSI more hidden topics are unveiled which are universal importance (weights) and more accurate clusters are formed. Using LSI the text can be clustered with all its rich properties even. Where in the earlier work they are considered as five vectors, limited to Name Entity Vector, Hot Term Vector, Direct Concept Vector and Kind of Vector and Part of Vector; which are limited to some directions of the subjects in the text. These vectors are again used in finding the similarity of the sentences. In the proposed method the Similarity metric [4] is applied class-wise: "*comparing names in one document with names in the other*". A semantic similarity is confided for string matching which results doing a decisive comparison.

B. Aging Theory

Aging theory [3] is a general temporal phenomena prevailing among the living objects of the real world. This specifies capturing variations in the distribution of key terms on a time line, which is a critical stage in the process of extracting the hot topics. Therefore, it is essential to track the topics to determine what stage of their lifecycle they are in. To model life span of a news event aging theory is suggested, that flows through birth, growth, decay and death. Frequency of terms and topics are critically tracked with their variations in their life.

C. Proposed Experimental Scheme

As LSI uses SVD, A common analytic tool for interpreting the "meaning" or "semantic space" described by the words that were extracted, and hence by the documents that were analyzed, is to create a mapping of the word and documents into a common space, computed from the word frequencies or transformed word frequencies (e.g., inverse document frequencies). In general, here is how it works:

Suppose a collection of end user reviews are indexed, every time in the review the terms are found incrementally as they get familiar (improve their frequency of occurrence). Such terms are identified as dimensions; the idea of LSI is to generate dimensions into which the words and documents can be mapped. As a result, it is possible to identify the underlying (latent) themes described or discussed in the input documents, and also identify the documents that mostly deal with economy, reliability, or both. Hence, mapping of the extracted words or terms and input documents into a common latent semantic space is carried on. The SVD is in order to extract a common space for the variables and cases (observations) and reduce the overall dimensionality of the input matrix to a lower-dimensional space, where each consecutive dimension represents the largest degree of variability between words and documents possible. Ideally, two or three salient dimensions are identified, which are accounting for most of the variability (differences) between the words and the documents and, hence, identify the latent semantic space that organizes the words and documents in the analysis. In some way, one such dimensions can be identified, the underlying meaning can be extracted of what is contained in the documents (described, discussed). Indexing Exhaustivity and Term Specificity [2] may implemented with parametric qualifications to asses the recall and precision of the method.

IV. CONCLUSIONS

In this paper a comparative study of the intensity of TF*IDF, TF*PDF and LSI have been done. These techniques reduce the dimensionality of the text databases semantically, parametrically. These are considered to be as fundamental text processing techniques; they play an operative role during the process of finding the hot topics in the text data. For this work, the analysis took place in two methods. First, we compared hot terms extracted by the TF*IDF weighting scheme with our proposed method. This experiment validates the effectiveness of our method over TF*PDF. Secondly, we compared hot terms extracted by applying LSI in reducing the database size and then using TF*PDF weighting scheme. The lifecycle model of the terms' quality and efficiency is also used to determine the hot topic in the sentence modeling. The LSI model is implemented using SVD and upon rigorous comparison of

documents, semantic similarities are obtained. Direct machine learning, fast machine learning methods may be used to identify the hot topics in the text databases which may be undertaken as future work.

REFERENCES

[1] Kuan-Yu Chen, Luesak Luesukprasert, and Seng-cho T. Chou, "Hot topic extraction based on timeline analysis and multidimensional sentence modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 8, August 2007.

[2] G. Salton and C.S. Yang, "On the specification of term values in automatic indexing," *J. Documentation*, pp. 351-372, 1973.

[3] C.C. Chen, Y.T. Chen, Y. Sun, and M.C. Chen, "Life cycle modeling of news events using aging theory," *Proc. 14th European Conf. Machine Learning (ECML '03)*, pp. 47-59, 2003.

[4] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Simple semantics in topic detection and tracking," *Information Retrieval*, vol. 7, no. 3–4, pp. 347-368, 2004.

[5] The 2004 Topic Detection and Tracking (TDT '04) Task Definition and Evaluation Plan, http://www.nist.gov/speech/tests/tdt/, 2004.

[6] TDT 2004: Annotation Manual Version 1.2, http://www.nist.gov/speech/tests/tdt/, Aug. 2004.