ISSN:0975-9646

Synchronizing Text Documents using Semantic Similarity for Topic Mining

Sakshi Jain
Computer Science of Dept.
Radha Raman Institute of Technology and Science, RGPV
Bhopal, India

Virendra Raghuwanshi
Computer Science of Dept.
Radha Raman Institute of Technology and Science, RGPV
Bhopal, India

Anurag Jain
(HOD)
Computer Science of Dept.
Radha Raman Institute of Technology and Science, RGPV
Bhopal, India

Abstract: Mining is a process of knowledge extraction with some meaningful information. Topic Mining also enables extraction of information from the set of text documents. Topic mining is a new way of categorizing text documents which are specially used in Big Organizations. Here in this paper a new and efficient technique is implemented for Topic mining which is based on the concept of synchronization between text documents using Semantic Similarity measures. The experiments are performed on two big text document sets of ICDE and SIGMOID on the basis of number of words extraction and computational time.

Index Terms: Topic Mining, Semantic Similarity, Synchronization, Topic Labelling, Jaccard Coefficient.

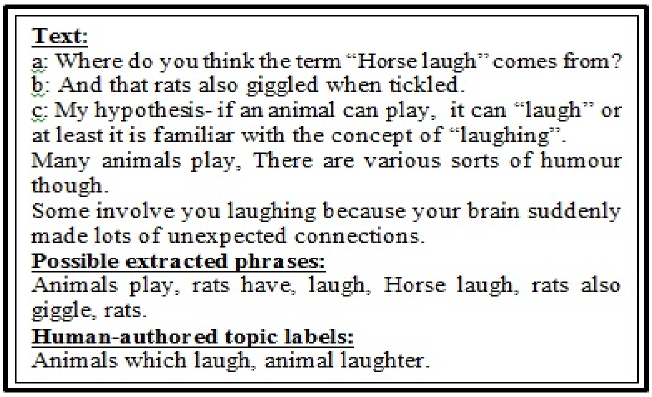
1. INTRODUCTION

There are a number of issues that Topic Mining (TM) researchers are currently addressing, including: accuracy, efficiency and effectiveness, privacy and security, and scalability. Accuracy is especially significant in the context of classification. Issues of efficiency and effectiveness pervade the discipline of TM. Issues of privacy and security centre around legal issues and the desire of many owners of data to maintain the copyright they hold on that data. The scalability issue is particularly significant as the amount of data currently available for TM is extensive and increasing rapidly year by year. One potential solution to the scalability issue is parallel or distributed TM, although this often entails a significant "communication" overhead.

Topic mining is the new born of data mining. In Topic Mining, patterns are extracted from natural language text rather than data bases. That means, automatically extracting information from a usually large amount of different unstructured textual resources. Topic mining is determined by the computer new, unfamiliar data by automatically extract data from various resources. Topic mining is unfamiliar from well-known within web search. The problem is brash onside all the material that currently is not applicable to your needs in order to find the applicable information. The goal of the text mining is to find the unknown data that are not available. Ranking can be used as a scoring function which scores the web pages or the document.

TOPIC LABELING

Given text segments about the same topic written indifferent ways (i.e., language variability), topic labelling deals with the problem of automatically generating semantically meaningful labels for those text segments. The potential of integrating topic labelling as a prerequisite for higher-level analysis has been reported in several areas, such as Synchronization [2] information extraction [3] and conversation visualization [4]. Moreover, the huge amount of textual data generated everyday specifically in conversations (e.g., emails and blogs) calls for automated methods to analyse and re-organize them into meaningful coherent clusters.



Text:
a: Where do you think the term "Horse laugh" comes from?
b: And that rats also giggled when tickled.
c: My hypothesis- if an animal can play, it can "laugh" or at least it is familiar with the concept of "laughing".
Many animals play. There are various sorts of humour though.
Some involve you laughing because your brain suddenly made lots of unexpected connections.
Possible extracted phrases:
Animals play, rats have, laugh, Horse laugh, rats also giggle, rats.
Human-authored topic labels:
Animals which laugh, animal laughter.

Table 1: Topic labelling example

TOPIC LABELING FRAMEWORK

Each topic cluster contains the sentences that can semantically represent a topic. The task of clustering the sentences into a set of coherent topic clusters is called topic segmentation [5], which is out of the scope of this paper. Our goal is to generate an understandable label (i.e., a sequence of words) that could capture the semantic of the topic, and distinguish a topic from other topics by given a set of topic clusters. Among possible choices of word sequences as topic labels, in order to balance the granularity, we set as valid topic labels.

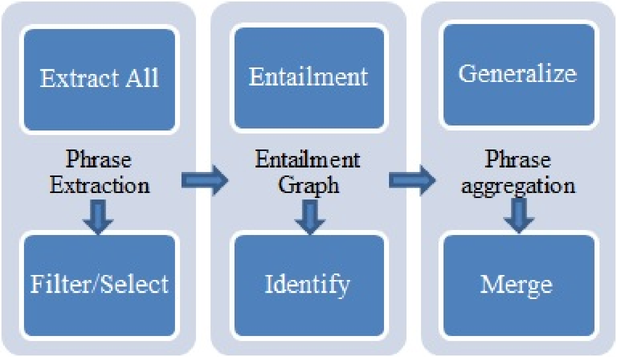


Figure 1: Topic labelling framework.

As shown in Figure 1, our framework consists of three main components that we describe in more details in the following sections.

Phrase extraction

We tokenize and pre-process each cluster in the collection of topic clusters with lemmas, stems, part-of-speech tags, sense tags and chunks. We also extract n-grams up to length 5 which do not start or end with a stop word. In this phase, we do not include any frequency count feature in our candidate extraction pipeline. Once we have built the candidates pool, the next step is to identify a subset containing the most significant of those candidates. Since most top systems in key phrase extraction use supervised approaches, we follow the same method [6].

AUTOMATIC TEXT SYNCHRONIZATION

Text Synchronization is the process of condensing text to its most essential points. When Synchronization process is carried out by machines, it is termed as Automatic Text Synchronization. Although the definition of Synchronization is obvious, it needs to be emphasized that Synchronization is a hard problem.

A Synchronization system has to interpret the source content, where content is a depiction of both information and emotion, and identify most relevant information. Classification of information as relevant or irrelevant is subjective to human nature. Synchronization is a challenging task for its inherent cognitive process, as an ideal Synchronization system has to mimic a human mind in the process of abstracting.

Synchronization is also interesting for its practical and real life applications. Researchers [7] have listed down Synchronization as a three phase process, consisting of following:

- Topic Identification: An initial exploration of text to identify its genre and topic.
- Relevance Assessment: Important and relevant topics are aggregated together, and presented in a new form. This formulation represents the actual concepts which may not be explicitly present in the text.
- Summary Generation: New formulation after relevance assessment is transformed into a coherent human readable format. It often involves cutting and pasting certain portions of text to form a summary.

2. LITERATURE SURVEY

Xiang Wang et. al's proposed a new and efficient technique of topic mining by providing asynchronizing between text documents [1]. Here in this paper a novel algorithm is implemented which is based on the concept of generative model. It consists of two steps in which the first step is the extraction of common topics from multiple text documents sequence based on the time stamp adjusted. The technique is implemented on six datasets and two news articles.

In this paper, [8] author has illustrate their work on a method capable of understanding these categories of characteristics to improve item for consumption of certain databases. This method gain knowledge of these characteristics by an appropriating text gain knowledge of this method to the given item for consumption explanations found on dealer web sites.

In this paper, [9] here the author has envision of the problem of decide on association rules as a classification job. A structure of a dual probabilistic classifier is established that bring into plays ontologies with the intention of approximation whether and to which quantity an imperative articulates a simple taxonomic association. In this paper here firstly in attendance the taken as a whole formal structure that predicts out our come within reach for deciding on association rules. Here they current examples that formulate use of straightforward ontologies which are built upon the hypernym-hyponym connection. Second, a probabilistic structure is initiated and is acquire to accomplish computations of the degree of appropriateness between each rule of a rule set to begin with extracted and a model of the domain chosen.

As the amount day by day of electronic information amplifies, there is producing concentration in expanding tools to help people improved discover, filter, and control these supply. Text categorization [10] is the task of natural language texts to one or more predefined grouping based on their substance which is a significant factor in much information society and managing tasks. Machine learning techniques, including Support Vector Machines (SVMs), have incredible probable for facilitating inhabitants to efficiently organize the electronic supplies. Text mining frequently engages the mining of keywords concerning some measure of consequence. Weblog data is textual substance with an understandable and important as per temporal aspect. Text categorization [11] is the task of involuntarily arranging a set of files into grouping from a predefined set. So the existing task has a number of applications, together with programmed manifestation of systematic articles according to predefined sets of scientific terms, filing exclusive rights into patent directories, selective distribution of information to information consumers, programmed population of hierarchical catalogues of Web sources, person responsibility attribution, investigation coding, spam pass through a filtering process, recognition of document type, and even programmed essay grading.

Identifying relative rulings is also constructive in put into practice because direct evaluations are perhaps one of the most compelling methods of text appraisal, which may even be more important than opinions on each individual object. The relative judgment identification [12] problem first classify the relative sentences into unusual types, and then nearby a work of fiction put together pattern discovery and supervised learning come within reach of to recognizing relative sentences from text documents. Here this paper experimental outcomes using three types of documents, news articles, customer evaluations of products, and Internet environment situation, show a exactness of 79% and recall of 81%. Assessment is one of the largest part compelling methods for text assessment. Extracting relative sentences from text is constructive for many appliances.

In this existing system, [13] an efficient outline of discovering method established which first determines discovered specificity patterns and then estimates the expression influence according to the circulation of expressions in the discovered patterns to a certain extent than the distribution in manuscripts for solving the misconception of the existing difficulty.

Li and Huang [9] present a study that aimed at providing a more comprehensive picture of interactions in computer-supported collaborative learning. The authors propose a model of multidimensional analysis to investigate interactions based on techniques applied in content analysis, text mining, and social networking. Content analysis is used to investigate how students interact, thus discovering possible process patters (the discourse intention) in the conversation.

3. PROPOSED METHODOLOGY

1. Take an input text document whose topic mining is to be done.
2. Remove stop words and un-known words from the text document which is to be matched with the dictionary.
3. Group the remaining text document into a number of tokens.
4. Perform Semantic Similarity between words using:

Data content similarity (SimC)

It is the Cosine similarity between the term frequency vectors of
d1 and d2:

$$SimC(d1,d2) - \frac{V_{d1} * V_{d2}}{\left\|V_{d1}\right\| * \left\|V_{d2}\right\|}$$

where ,Vd is the frequency vector of the terms inside data unit d, ||Vd|| is the length of Vd, and the numerator is the inner product of two vectors.

Number of Common Neighbors

It is defined as the total number of nodes that are connected directly in relationship with node x and y for unweighted network,

$$CN(x,y) = \varphi(x) \cap \varphi(y)$$

Where, $\varphi(x)$ is the set of neighbors of nodes x.
$\varphi(y)$ is the set of neighbors of node y.

To calculate link prediction between nodes for unweighted network common neighbors can be calculated as,

$$CN(x,y) = \sum_{z \in \varphi(x) \cap \varphi(y)} w(x,z) + w(y,z)$$

Jaccard Coefficient

It is defined as the highest proportion of common neighbors to the total number of neighbors in the network. The jaccard coefficient can also defined for weighted as well for unweighted network. For unweighted network,

$$JC(x,y) = \frac{\varphi(x) \cap \varphi(y)}{\varphi(x) \cup \varphi(y)}$$

For weighted network,

$$JC(x,y) = \sum_{z \in \varphi(x) \cap \varphi(y)} \frac{w(x,z) + w(y,z)}{\sum_{a \in \varphi(x)} w(a,x) + \sum_{b \in \varphi(y)} w(b,y)}$$

5. Predict the most valuable words from the text documents having most similarity between words.
6. Applying parsing on the words to make a topic.

4. RESULT ANALYSIS

The table shown below is the analysis and comparison of number of words extracted from the **ICDE** dataset based on existing and proposed work. The table shows the performance of the proposed methodology.

ICDE Dataset	Base Work	Proposed Work
Doc1	1973	1957
Doc2	1701	1685
Doc3	1447	1431
Doc4	1992	1976
Doc5	1970	1954
Doc6	1436	1420
Doc7	1436	1420

Table 2. Analysis of words extracted on ICDE Dataset

The table shown below is the analysis and comparison of number of words extracted from the **SIGMOID** dataset based on existing and proposed work. The table shows the performance of the proposed methodology.

SIGMOID Dataset	Base Work	Proposed Work
Doc1	1993	1977
Doc2	1997	1981
Doc3	1990	1974
Doc4	1997	1981
Doc5	1999	1983
Doc6	1967	1951

Table 3. Analysis of words extracted on SIGMOID Dataset

The Figure shown below is the analysis and comparison of number of words extracted from the **SIGMOID** dataset based on existing and proposed work. The table figure the performance of the proposed methodology.

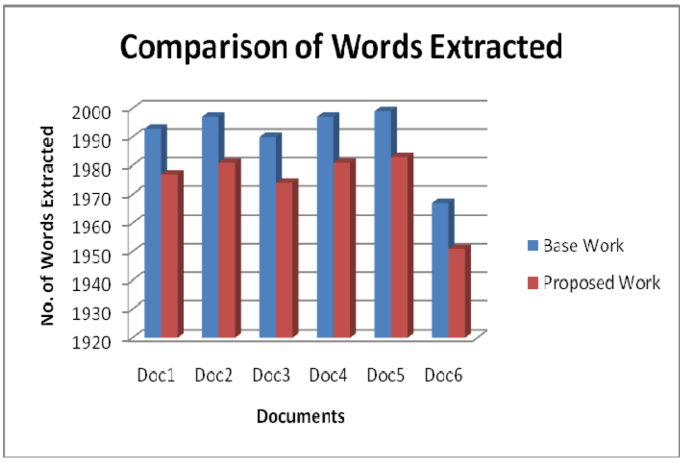


Figure 2. Comparison of words extracted on SIGMOID Dataset

The figure shown below is the analysis and comparison of number of words extracted from the **ICDE** dataset based on existing and proposed work. The figure shows the performance of the proposed methodology.

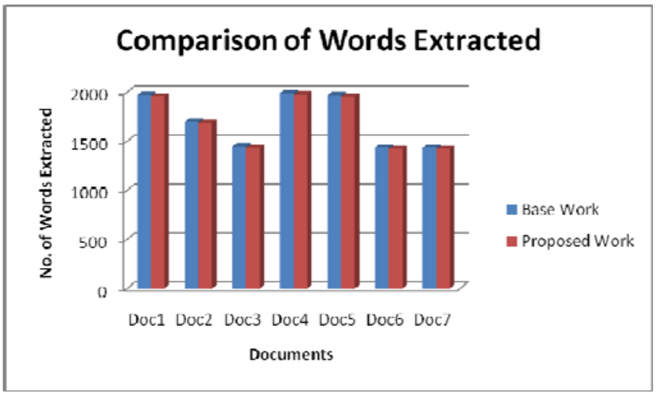


Figure 3. Comparison of words extracted on ICDE Dataset

The table shown below is the analysis and comparison of Computational time from the **ICDE** dataset based on existing and proposed work. The table shows the performance of the proposed methodology.

ICDE Dataset	Base Work	Proposed Work
Doc1	70000	62652
Doc2	30000	22928
Doc3	50000	42602
Doc4	120000	110904
Doc5	130000	104990
Doc6	50000	46354
Doc7	40000	37588

Table 4. Analysis of Computational time on ICDE Dataset

The table shown below is the analysis and comparison of Computational time from the **SIGMOID** dataset based on existing and proposed work. The table shows the performance of the proposed methodology.

SIGMOID Dataset	Base Work	Proposed Work
Doc1	80000	76744
Doc2	70000	67434
Doc3	60000	52868
Doc4	50000	47124
Doc5	90000	88654
Doc6	60000	55133

Table 5. Analysis of Computational time on SIGMOID Dataset

The figure shown below is the analysis and comparison of Computational time from the **ICDE** dataset based on existing and proposed work. The figure shows the performance of the proposed methodology.

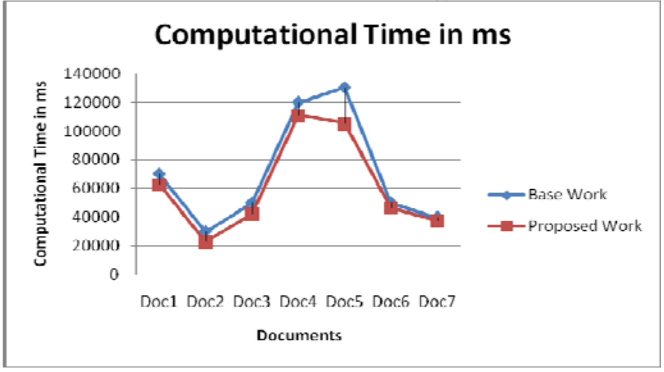


Figure 4. Comparison of Computational time on ICDE Dataset

The figure shown below is the analysis and comparison of Computational time from the **SIGMOID** dataset based on existing and proposed work. The figure shows the performance of the proposed methodology.

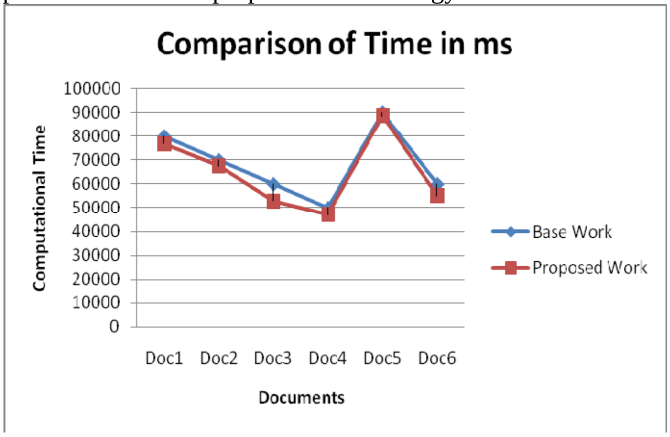


Figure 5. Comparison of Computational time on SIGMOID Dataset

5. CONCLUSION

The proposed technique implemented here for the topic mining using semantic similarity is efficient as compared to the existing technique implemented here for the topic mining using asynchronization. The experiments are performed on two text documents ICDE and SIGMOD. The results are performed and compared on the basis of computational time and number of words extracted. The proposed methodology is efficient as compared to the existing technique.

REFERENCES

[1] Xiang Wang, Xiaoming Jin, Meng-En Chen, Kai Zhang, and Dou Shen," Topic Mining over Asynchronous Text Sequences", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 1, January, 2012.

[2] Sanda Harabagiu and Finley Lacatusu,"Using topic themes for multi-document summarization", *ACM Trans. Inf. Syst.*, 28:13:1–13:47, July, 2010.

[3] James Allan,"Topic detection and tracking: event-based information organization", 2002.

[4] Shixia Liu, Michelle X. Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian, "Tiara: Interactive, topic-based visual text summarization and analysis", *ACM Trans. Intell. Syst. Technol.*, 3:25:1–25:28, February, 2012.

[5] Shafiq Joty, Gabriel Murray, and Raymond T. Ng,"Supervised topic segmentation of email conversations", *In ICWSM11*, 2011.

[6] Su Nam Kim, OlenaMedelyan,Min-Yen Kan, and Timothy Baldwin ,"task 5: Automatic keyphrase extraction from scientific articles", *In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, pages 21–26, Stroudsburg, PA, USA. Association for Computational Linguistics, 2010.

[7] K. S. Jones, "Automatic summarising: factors and direction", *In Advances in automatic text summarisation*, pages 1–12. MIT Press, 1998.

[8] Rayid Ghani and Andrew E. Fano," Using Text Mining to Infer Semantic Attributes for Retail Data Mining", 2003.

[9] Dietmar Janetzko ,Hac` ene Cherfi , Roman Kennke, Amedeo Napoli and Yannick Toussaint," Knowledge-based Selection of Association Rules for Text Mining" 2003.

[10] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *Proc. European Conf. Machine Learning (ICML '98)*, pp. 137-142, 1998.

[11] M.F. Caropreso, S. Matwin, and F. Sebastiani. Statistical Phrases in Automated Text Categorization, Technical Report *IEI-B4-07- 2000, Instituto di Elaborazione dell'Informazione*, 2000.

[12] N. Jindal and B. Liu. Identifying Comparative Sentences in Text Documents, *Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06)*, pp. 244-251, 2006.

[13] K. Mythili, K. Yasodha, "A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining" *International Journal of Science and Applied Information Technology, ISSN No. 2278-3083* Volume 1, No.3, July – August 2012.