Using Latent Dirichlet Allocation to Incorporate Domain Knowledge with Concept based Approach for Automatic Topic Detection

A.Mekala,
MCA, MSC, MPhil,
Research Scholar
Manonmaniam Sundaranar University,
Thruneveli.

C.Chandra Sekar, PhD.
Reader, Department of computer science
Periyar University, Salem

ABSTRACT

In the past couple of years multi-topic summarization is a research investigation that has expanded much attention. There has been a variety of effort on generating natural language summaries for variety of topics, but this is feasible only for a very small number of topics. In this research paper the method trying to provide automatic detection of topics to be summarized that is can determine how many topics should be chosen for automatic summarization. This is effectively done through the combined efficient framework of ontology based and non ontology based systems will be optimistic for the excellence of topic summary. To achieve this propose to apply latent Dirichlet allocation (LDA) model for capturing the semantic information on topic transcription. LDA is a generative model and defines a probabilistic method for generating a new document. The LDA model is utilized for estimating topic sharing in queries and word recorded topic documents, and the identical is performed at the topic level. Concept based topic matching between query words and topic documents are performed using ontology based and non ontology based matching algorithm. The results of topic level matching methods are evaluated between the automatic topic detection method and predefined topic detection mechanism along with the experimental results shown to be complementary.

Keywords: Latent Dirichlet allocation, Ontology based matching algorithm, Non ontology based matching algorithm, Topic summarization, and Automatic topic detection.

1. INTRODUCTION

Topic detection (TD) facilitates the automatic finding of new topics from any news corpus and the consequent task of input news group documents to discover topics. In general a new topic corresponds to a newsworthy incident for example the election in US 2008. Associations surrounded by the discovered topics can be unranked. Besides, as a topic is more specific than a news group such as sports or finance, the majority work on TD of course imagines a simple ungraded topical structure. Supplementary the TD process can be further partitioned into online (real-time) and offline (batch) modes than topic structural variations, which are as well known as new event detection [1] and retrospective event detection [9], subsequently. Online TD additionally inspects apiece incoming news document to assess whether it belongs to an existing topic or if a new topic should be created supported on it. Offline TD inspects the entire amount of news documents to concurrently unpick topics and their correlated news documents. Though, there are a number of understated TD features, which if not taken into due consideration, can unfavorably involve realistic clustering routines: 1) time plays a pivotal role, with every news document attitude a time stamp, 2) news topics are of course bursty, i.e., new topics are steadily generated while old topics die off, 3) news documents with semantically alike content but dissimilar time-frames most likely created from different topics. TD can thus be versioned as an extraordinary case of stream clustering [4], with the clustering collection at any time spot similar to a thought that floats with time [8], [5]. In spite of the reality that time cooperates a central position, the vast common of existing TD results [9], [2], [7], [10], [3], [6] do not clearly integrate time into their formulations; each news document is represented as a vector by means of time-doubter static weights, with just one trivial procedural alteration: news vectors are processed in time-stamp order, i.e., online TD, as opposed to batch TD, is used to switch the temporal factor.

Providentially, this simple procedural alteration over stationary manuscript symbol model appears to work moderately well in practice. A classic approach is using the term frequency–inverse document frequency (tf-idf) scheme (Salton and McGill, 1983). This frequency-based weighting approach is limited to the intra-document topic relations and provides little information on inter-document relations. The Latent Semantic Indexing (LSI) and probabilistic Latent Semantic Indexing (pLSI) have been proposed to address the limitation. But there also a limitation that we have to predefine the topic to be searched related to the input news group documents.

So in order to provide the automatic detection of topics to summarize use the LDA approach. One of the endeavors of LDA and comparable methods, including Probabilistic latent semantic analysis (PLSA) (Hofmann, 2001), is to construct low dimensionality demonstrations of texts in a “semantic space” such that most of their intrinsic algebraic features are protected. A diminution in dimensionality assists storage space as well as faster recovery. Modeling discrete data has many applications in classification, categorization, topic detection, data mining, information retrieval (IR), summarization and collaborative filtering (Buntine and Jakulin, 2004). The method has employed Latent Dirichlet allocation (LDA) (Blei et al., 2002) model as our main topic sculpting tool. The aim of this paper is to test LDA for creating the semantic consistency of a document supported on the premise that a real (coherent) document should discuss only a few number of topics, a property hardly granted for forged documents which are often made up of random grouping of words or topics. As a consequence, the coherence of a document may reproduce in the entropy of its subsequent topic distribution or in its confounded for the model. The entropy of the estimated topic allocation of a true document is expected to be lower than that of a bogus document. Furthermore, the matching is done based on the ontology concept and also based on the non-ontology concept with the combined framework of this research will produce the better information retrieval results.

The main contribution work is as follows:

1. Prepare the input dataset and get the user query from the user
2. The concept matching is done with the use of ontology based scheme and also non ontology based scheme to provide semantic news group information's.
3. Detecting topics from a text corpus (including metadata) can be decayed into a series of tasks, including removal of words and/or entities, detection or reproductioning of topics by using algorithms, clustering or grouping features of topics, and evaluation of the results.
4. A supplementary current development is the Latent Dirichlet Allocation (LDA) algorithm which is a dimensionality decline technique at the same time as providing "proper underlying generative probabilistic semantics that make intelligence for the type of data that it models".
5. In this research the term "topic" refers to a semantic rank that consists of terms distribution some general subject relationships or similar meanings. The features of a topic are words, keywords or phrases that semantically fit in to the topic.
6. After this the classification is performed with the NLP tool sing the SVM classifier and the result will be summarized.

The remainder of this as follows. In section 2 the ontology based anatomy approach and the non-ontology based system is described. The LDA method is explained in section 3 and the method used for its training and practice of computing topic allocation for invisible text document. The experimental setup, database and results are discussed in Section 4, and the conclusions of this study are drawn in Section 5.

2. SUMMARIZATION SYSTEM

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. It mainly focused on extraction-based generic text summarization, which composes summaries by extracting informative topics from the original documents. In this section creating the framework which is based on the ontology based and non-ontology based schemes to summarize the topics. This will be discussed brefiely as follows.

2.1 Ontology Based Topic Summarization

In this module, first collect vocabularies and synonyms with the use of NLP tool. Next, place those words by the Data model of ontology. The primary step of our method is to determine the main subtopics of the article of interest. This is attained by comparing the words of articles with terms in the ontology. If the word does not exist in the ontology, we ignore it. Otherwise, we record the number of times the word appears in the ontology encode 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 our 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. Choose to ignore the subtopic if its count is less than 10. In addition, we choose only top three or required subtopics. After obtaining the subtopics, our system will use them for selecting paragraphs as the summary. Mainly, summarization system will give every paragraph significance attain and grade them by the scores. Higher scores involve that the paragraphs are extra probable to be selected into the digest. In the end, take out a desired portion of subsections as the summary. Using semantic information encoded in the ontology, our system determines which topics are useful for extracting paragraphs. Designing and constructing the ontology are the first two steps for building the summarization system.

Our system will use them for selecting paragraphs as the summary after attaining the subtopics. Rank the paragraphs maintained on their "closeness" to the top qualities subtopics. The collection route as follows.

1. Work out consequence among paragraphs and the picked subtopics. Evaluate the words of all selected subtopic through words in each paragraph, and connect with each paragraph the calculations of common words that appear in the paragraph and the selected subtopics using NLP tool. Let assume there are *n* selected subtopics, there will also be *n* scores associated with each paragraph, and these *n* scores stand for the relevance of the article with each selected subtopic.

2. Compute the score for each paragraph. The score of each paragraph is the sum of its weighted importance with subtopics with the semantic similarity measures. The weights are decided dynamically based on counts that utilized to selected main subtopics. The weight of each topic has a perceptive clarification. That is primary topic is further representative than other topics, so the weight should be superior to others.

3. Rank paragraphs, and select a needed quantity of the paragraphs as the topic summary.

Let denote the P_j is the score of the j^{th} paragraph. And the O_{ji} is the score of the i^{th} topic of P_j, w_i is the weight of the i^{th}. In summary, use the following formula for scoring paragraphs:

$$P_j = w_1 O_{j1} + w_2 O_{j2} \dots \dots + w_n O_{jn}$$

2.2 Non-ontology-based Summarization

The non-ontology based method is based on the following procedure to extract the feature from each of the documents which is described as follows:

- *Term-frequency computation:* Count up the term frequency of each word, and the select the majority frequent N words for scoring. After that, every paragraph is scored based on the appearance of these N words. The N value is chosen by user which is called threshold.

- *Topic length computation:* Given a threshold for all paragraphs, pay no attention to the paragraphs that do not contain enough number of words.
- *Bonus words computation:* If one paragraph encloses bonus words, then the probability of the paragraph will be selected into the summary is higher as well. Use some amount of news articles as the training corpus to select the bonus words.
- *Proper nouns computation:* The significance of a paragraph is transmitted to the number of amount of proper nouns. For counting proper nouns, basically count the number of words with leading upper-case letters in each paragraph.

In this let assume the S_j is the score of the j^{th} paragraph. And the f_{ji} is the score of the i^{th} feature of the j^{th} paragraph, w_i is the weight of the i^{th} feature. L is 1 if the paragraph has sufficient number of words, otherwise 0. In summary, use the following formula for scoring paragraphs: After getting values of these features, score each paragraph with the following formula.

$$S_j = L(w_1 f_{j1} + w_2 f_{j2\dots} + w_n f_{jn}$$

3. LATENT DIRICHLET ALLOCATION (LDA) TO DETERMINE TOPICS

In this work, noisy topics are taken into relation as theme of attention in topic detection. This constructs the problem still additional complicated given that surrounded by a topic, a topic may not be always declared explicitly. For this motive, keyword mining and dictionary lookup approaches to conclude themes are unsuited in our framework. Further, the input documents postings can deal with very different topics which make creation of an appropriate word list difficult. For these reasons, LDA is selected as approach to theme detection since it ensembles best for the difficulty and data are dealt with. The LDA algorithm is on document level planning at estimating the concreteness of topic models. Especially, the confusion of a detained-out test set is calculated to evaluate the topic models. Results of LDA quality for topic detection in documents in general, and in topics of noisy are still unavailable.

3.1 Identification of Topics

In this paper, are considering the theme of a topic as its 'topic' which can be described by a set of words that do not have to be explicitly mentioned in a topic. Determine these topic describing terms (referred to as 'topic terms') using LDA. The original implementation of LDA is extended by a topic detection algorithm to apply it to topics. The entire process contains five steps explained in more detail in the following paragraphs.

Document preprocessing is the first step which is to avoid the stemming and stop word removal to provide the better result. Second topic and word normalization process in which, only nouns and proper nouns are stemmed and kept for further processing. For detecting word classes and to perform stemming, the Stanford NLP Tool is used. Limit the words to be judged by LDA to nouns to decrease computing time and to confine topic terms to those that are content-behavior.

In a third step the Topic Detection, topics beside with their probabilities are recognized for each topic using the LDA-algorithm and supported on the vector representation of topics. Each topic is believed to consist of a topic combination and each word's conception is attributable to one of the topic's themes. Therefore, a topic is explained by a set of words derived from the documents where to each word a probability is allocated that designates the relevance of this word for the topic. In this way, all topics are described by the same words, but with varying probability values for each word. The output of LDA is finally (1) the probability of each word for a topic and (2) the probability of each topic for a topic. Through previous experiments learned that it is necessary to pre-filter topics regarding their topical focus. In order to exclude topics without topical focus, our LDA modification considers only topics with at least four words (excluding stop words). A term is in turn considered relevant for clustering when it happens in at least 15 subjects.

To run LDA, the number of clusters to be formed requires to be fixed. Since it is still unidentified what the best number of clusters to be selected is, do a widespread evaluation to recognize associations between the number of topic clusters and the data set size. LDA also needs fitting two other parameters: The α parameter decides how dominant a topic is departure to be in a document. The hyper parameter β can be interpreted as the previous observation count on the number of times words are modeled from a topic previous to any word from the corpus is monitored. Steyvers and others have established α and β to work well with many different text collections. Choose these values in our experiments.

In last step of topic selection a, the probabilities decided by LDA is used to pick the topic and topic words for a topic. The probability per topic and topic computed by LDA indicates to what quantity the topic goes to the topic. If for a topic all themes have the identical probability, they are equally allocated with a probability of 1/k (with k = number of topics). In this case, the probability does not permit illustrating any conclusions on the most possible topic of a topic. For this reason, in our approach, topics with a probability larger than 1/k are believed as topics of a topic. Topics with a smaller probability are excluded given that their support for describing the content of a topic is too low. In this way, up to based on the threshold value topics are selected for each topic. One reason for choosing LDA is these probability values that allow us to filter out irrelevant topics and also topics without relevant focal point.

```
INPUT: b, a, CL,$\omega^{\delta}$, $S^t, t \in \{1, \dots N_{stream}\}$
For $t = 1$ to $N_{stream}$ do
If $t = 1$ then
$\beta_k^t = b, k \in \{1, \dots k\}$
Else
$\beta_k^t = B_k^{t-1}\omega^{\delta}, k \in \{1, \dots k\}$
End if
$\propto_d^t = a, d = 1, \dots \dots M^t$
Initialize $\emptyset^t$ and $\theta^t$ to zeros
Initialize the topic assignment, $z^t$ for all word tokens in $S^t$
$[\emptyset^t, \theta^t, z^t] = LDA(S^t, \alpha^t, \beta^t)$
$B_k^t = B_k^{t-1} \cup \emptyset_k^t, k \in \{1, \dots k\}$
```

```
If $t > 1$ then
$[topics(t), topicsA(t)] = detect(CL)$
End if
End for
```

4. RESULTS AND DISCUSSION

4.1 Precision vs. Number of datasets

This graph shows the precision rate of existing and proposed system based on two parameters of precision and the number of datasets. From the graph can see that, when the number of number of datasets is advanced the precision also developed in proposed system but when the number of number of datasets is improved the precision is reduced somewhat in existing system than the proposed system. From this graph can say that the precision of proposed system is increased which will be the best one. The values are given in Table 1:

Table 1: Precision vs. Number of datasets

SNO	Number of datasets	Proposed system	Existing system
1	10	0.3	0.21
2	20	0.52	0.45
3	30	0.63	0.56
4	40	0.69	0.6
5	50	0.75	0.64

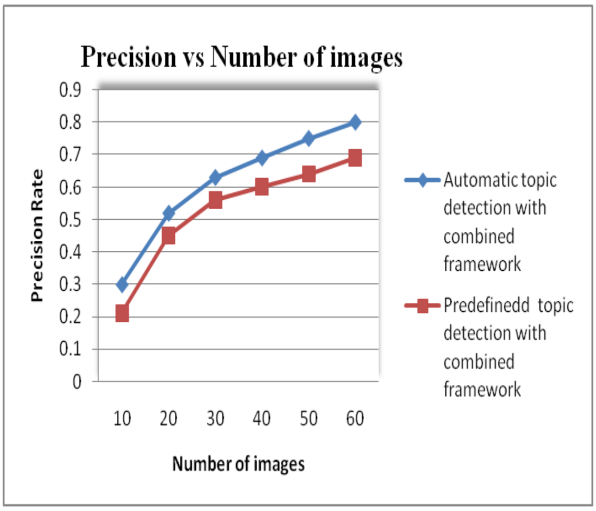

Fig 6: Precision vs. Number of datasets

In this graph have chosen two parameters called number of datasets and precision which is help to analyze the existing system and proposed systems. The precision parameter will be the Y axis and the number of datasets parameter will be the X axis. The blue line represents the existing system and the red line represents the proposed system. From this graph see the precision of the proposed system is higher than the existing system. Through this can conclude that the proposed system has the effective precision rate.

4.2 Recall vs. Number of datasets

This graph shows the recall rate of existing and proposed system based on two parameters of recall and number of datasets. From the graph can see that, when the number of number of datasets is improved the recall rate also improved in proposed system but when the number of number of datasets is improved the recall rate is reduced in existing system than the proposed system. From this graph can say that the recall rate of proposed system is increased which will be the best one. The values of this recall rate are given below:

Table 2: Recall vs. Number of datasets

SNO	Number of datasets	Proposed system	Existing system
1	10	0.35	0.23
2	20	0.43	0.32
3	30	0.5	0.42
4	40	0.59	0.49
5	50	0.72	0.6
6	60	0.87	0.69

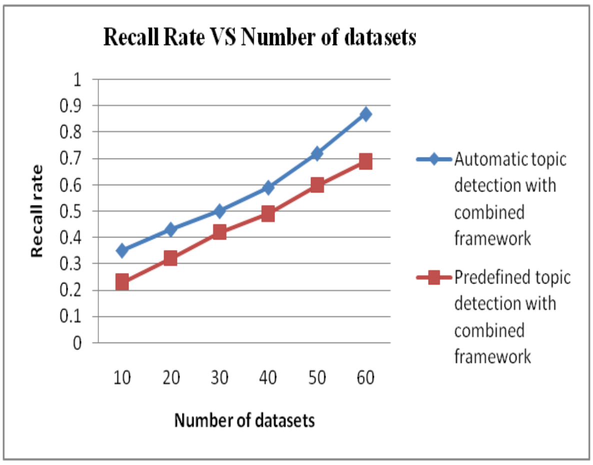

Fig 7: Recall vs. Number of datasets

In this graph have chosen two parameters called number of datasets and recall which is help to analyze the existing system and proposed systems on the basis of recall. In X axis the iteration parameter has been taken and in Y axis recall parameter has been taken. . From this graph see the recall rate of the proposed system is in peak than the existing system. Through this can conclude that the proposed system has the effective recall.

4.3 F-measure vs. Number of datasets

This graph shows the Fmeasure rate of existing and proposed system based on two parameters of Fmeasure and number of datasets. From the graph can see that, when the number of number of datasets is improved the Fmeasure rate also improved in proposed system but when the number of number of datasets is improved the Fmeasure rate is reduced in existing system than the proposed system. From this graph can say that the Fmeasure rate of proposed system is increased which will be the best one. The values of this Fmeasure rate are given below:

Table 3: F-measure vs. Number of datasets

SNO	Number of datasets	Proposed system	Existing system
1	10	0.89	0.76
2	20	0.82	0.71
3	30	0.74	0.62
4	40	0.65	0.52
5	50	0.54	0.42
6	60	0.42	0.32

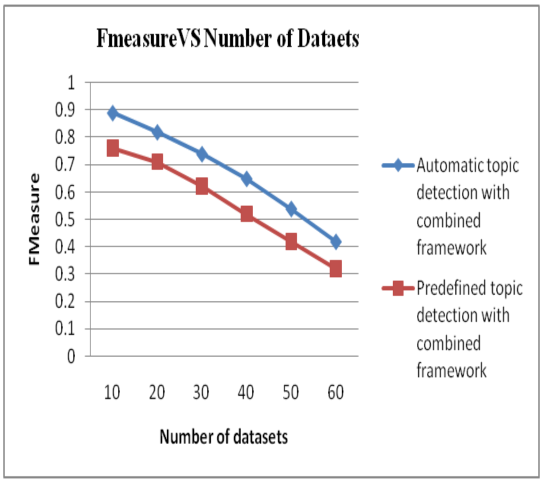

Fig 8: F-measure vs. Number of datasets

In this graph have chosen two parameters called number of datasets and recal which is help to analyze the existing system and proposed systems on the basis of Fmeasure. In X axis the iteration parameter has been taken and in Y axis recal parameter has been taken. . From this graph see the recal rate of the proposed system is in peak than the existing system. Through this can conclude that the proposed system has the effective recal.

5. CONCLUSION

This manuscript studies the automatic detection of topic evolutions for the input documents. Our experimental results show that incorporating the related topics with the topic model LDA improves the performance of topic detection on both manual and automatic transcripts over a baseline that uses slides alone. Incorporating topic summarization also makes the detection task more robust. This is done efficiently with the combined framework of ontology based and non ontology based topic summarization.

REFERENCES

[1] James Allan, Carbonell, George Doddington, Jonathan Yamron and Yiming Yang, “Topic Detection and Tracking Pilot Study: Final Report”, In Proceedings of the Broadcast News Understanding and Transcription Workshop, 1998.

[2] James Allan, Victor Lavrenko, Hubert Jin, First story detection in TDT is hard, In CIKM’00.

[3] T. Brants, F. Chen and A. Farahat, “A system for New Event Detection”, In SIGIR’03.

[4] Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani and Liadan O’Callaghan, “Clustering data streams: Theory and practice”, J. TKDE 2003.

[5] Qi He, Kuiyu Chang and Ee-Peng Lim, “A Model for Anticipatory Event Detection”, In ER’06.

[6] G. Kumaran and J. Allan, “Text classification and named entities for new event detection”, In SIGIR’04.

[7] Nicola Stokes and Joe Carthy, “Combining semantic and syntactic document classifiers to improve first story detection”, In SIGIR’01.

[8] Alexey Tsymbal, “The problem of concept drift: Definitions and related work”, Technical report, Department of Computer Science, Trinity College, 2004.

[9] Y. Yang, T. Pierce and J. Carbonell, “A Study of Retrospective and On-Line Event Detection”, In SIGIR’98.

[10] Y. Yang, J. Zhang, J. Carbonell and C. Jin, “Topic-conditioned Novelty Detection”, In SIGKDD’02.