Clustering News Articles for Topic Detection

VAIDEHI PATEL[1], ARPITA PATEL[2]
[1,2] *Department of computer engineering, LDRP-ITR,Gujarat*

Abstract -- In this paper we presented an approach for detecting topics from news articles. Topic detection used in text mining process. Text mining is a field that extract previously unknown and useful information from unstructured textual data. The main purpose of topic detection and tracking is to identify and follow events presented in multiple news sources. Topic detection and tracking would be very helpful to have a system able to map out the data automatically finding story boundaries, determining what stories go with one another, and discovering when something new has happened. We would try to find the first story of new events, identifying all subsequent stories on a certain topic defined by a small number of sample stories, and detect the occurrence of new events. We are going to use agglomerative clustering based on average linkage for detecting the topics.

Indexed Terms — Topic detection, Text Mining, agglomerative clustering.

I. INTRODUCTION

News is an important source of information. Most newspapers and news agencies provide news on their web pages. News portals work as a news aggregator and gather, merge, and organize news articles obtained from various sources. Topic tracking and detection (TDT) is a relatively new challenge for information retrieval technology that can be used in the text mining process. It focuses on extraction of significant topics and events from news articles.

Topic detection task is considered as the problem of finding the most prominent topics in a collection of news articles. The task of topic tracking is to monitor a stream of news stories and find out what discuss the same topic described by a few positive samples. Topic detection and tracking is used to alert companies anytime a competitor is in the news, in medical industry and in the field of education to be sure they have latest references.[17] Our approach combines a variety of learning techniques. Topic detection is unsupervised task and topic tracking is supervised task. We are going to use agglomerative clustering to create topic clusters. To identify the major news, we identify the clusters of similar news items.

The corpus consists of news items gathered from a large number of internet news sites world-wide like Times of India and CNN, and of various subscription news wires. Thus the texts are from the different sources which often discuss the same events. Newspapers often publish the news they receive from press agencies with no or few correction. The corpus of news articles thus contains not only summaries of the same events written by different journalists, but also many duplicates and near duplicates of the same original text which need to be eliminated from the collection.

The motivation for research in TDT is to provide a core technology for a system that would monitor news and alert an analyst to new and interesting events occurring in the world. Analysts are keen to track and in particular to know the latest news about a story from a huge volume of information that arrives daily. It is important to provide a means for people such as journalists to understand and interpret what is happening in the news. News articles include more than one subject, but many NLP and IR techniques implicitly assume that articles have just one topic. Even in the presence of a single topic within an article, the article may address multiple subtopics and various aspects of the primary topic. Dividing articles into topically coherent units, discovering their topic could be quite valuable in many applications where people need efficient access to large quantities of information. For example, systems could alert users to new events and to new information about old events. [20]

The objective of research is to break the text down into individual news stories, to monitor the stories for events that have not been seen before, and to gather the stories into groups that each stories discuss a single news topic. The scope of the research is text in news articles obtained from the various news paper websites

TDT is used in text mining process. Text mining is a new field of computer science which fosters strong connections with natural language processing, data mining. Text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. The field of text mining has received a lot of attention due to the always increasing need for managing the information that resides in the vast amount of available documents. [6] The goal is to discover unknown information, something that no one yet knows. Text mining utilizes techniques from the field of data mining, combines methodologies from various other areas such as information extraction, information retrieval, computational linguistics, categorization, clustering, summarization, topic tracking and concept linkage.

The remainder of the paper is organized as follows. In Section 2 we have described general challenges in the domain of email classification. Section 3 presents related work. Section 4 concludes the paper.

II. RELATED WORK

Allan et al. (1998) identify new events and then track the topic like in an information filtering task by querying new documents against the profile of the newly detected topic. Topics are represented as a vector of stemmed words and their TF.IDF values, only considering nouns, verbs, adjectives and numbers. In their experiments, using between 10 and 20 features produced optimal results. Friburger & Maurel (2002) showed that the identification and usage of proper names, and especially of geographical references, significantly improves document similarity calculation and clustering. Hyland et al. (1999) clustered news and detected topics exploiting the unique combinations of various named entities to link related documents. However, according to Friburger & Maurel (2002), the usage of named entities alone is not sufficient. Topic detection is addressed in, where Sekiguchi et al. present a method which uses blogger's interests in order to extract topic words from weblogs. In this approach the authors assume that topic words are words commonly used by bloggers who share the same interests, and they use these topic words to compute similar interests between each two bloggers by using the cosine similarity measure. A topic score is assigned to each word. The processing time is also a problem in this approach, as they have pointed out, and the optimization for some of their calculations is needed.

Topic detection is an unsupervised and topic tracking is supervised. In our approach we are going to use hierarchical agglomerative clustering for topic detection based on average linkage using document similarity vector. For document similarity we have to use cosine similarity based on TF.IDF.

III. PROPOSED APPROACH

Goal: Regarding the current events, a system is required to detect topics within news articles. We would be choosing any one or two domain from politics, sports, science and discovery, education, entertainment etc. Focusing on chose domain goal is to implement a system that gives quite satisfying results about current events with all related stories using the optimal approach. We would be using clustering for topic detection. [16]

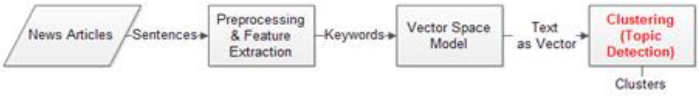

Figure 1 Flow of System

In our approach first step is preprocessing on collected text news. Following are the steps of preprocessing:

- First of all tokenization will be applied on texts of news articles. Here in tokenization sentences will be broken in to words.
  Example: Text Mining is used to extract knowledge from unstructured data. After applying Tokenization it will be like Text, Mining, is, used, to, etc.

- Then from the set of all words stop words will be removed. Here it will remove non informative words like the, more, And, when, etc.

- Then stemming will be applied on the words to get root word. Here it will remove suffix to generate word stem.
  Example: walking, walked, walks will become walk.

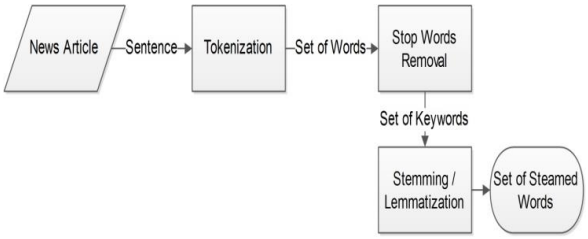

Figure 2 Preprocessing

Topic Detection:

The topic detection task evaluates technologies that detect novel, previously unknown, topics. As in the tracking task, topics are defined by associating together stories that discuss the topic. However, topic detection systems are not given a priori knowledge of the topic. Therefore, systems must embody an understanding of what constitutes a topic, and this understanding must be independent of topic specifics. The systems detect clusters of stories that discuss the same topic. The concept of clustering is easily applied to news stories, but the assessment of performance is difficult because stories frequently discuss multiple topics. This phenomenon not only means the topic clusters are dependent on previously processed stories, but also that decomposition of performance into casual subsets is misleading. [10]

Agglomerative clustering has been used successfully for topic detection. It is a sequence of partitions in which each partition is nested into the next partition in the sequence. It is defined by disjointed clustering, which individualize each of the N documents within a cluster. This process is repeated in which the number of clusters decreases as the sequence progresses until a single cluster containing all N documents. An important feature of creating topic clusters based on keywords is the presence of data overlap between clusters. If one story contains five different keywords describing its content, then the text for the story will appear in five different clusters. When using agglomerative clustering to create a topic tree, the effects of data overlap on the measure of cluster similarity need to be considered. We will be using agglomerative clustering with average distance measures. Advantage of average system measure is it can handle categorical as well as numerical data. Use of this measure overcomes the outlier sensitivity problem.

Algorithm for Agglomerative Hierarchical Clustering:

Algorithm Agglomerative(D)

Make each data point in the data set D a cluster,

Compute all pair-wise distances of X1, X2, …. Xn є D;

Repeat

Find two clusters that are nearest to each other;

Merge the two clusters from a new cluster c;

Compute the distance from c to all other clusters;

Until there is only one cluster left

Algorithm steps for Topic Detection

Algorithm for Agglomerative Hierarchical Clustering (average linkage base):

1. Initially put each keyword in its own cluster.
2. Among all current clusters, pick two clusters with the smallest distance.
   Here for distance we would be using average distance:

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d(a,b).$$

   Where A and B are two sets of observation and a and b are two points.

3. Find two clusters that are nearest to each other.
4. Replace these two clusters with a new cluster, formed by merging the two original ones.

5. Repeat the above two steps until there is only one remaining cluster in the pool.

IV. OUR APPROACH WITH EXAMPLE

Given one Paragraph like:
Your resume is nice. It includes all requirement that we need. Congratulations! You are selected for the interview process. You can come at our office any time after 11 a.m.

States are individual sentences from given paragraph after pre-processing step. For above example states are:

1. You resume nice.
2. Include all require we need
3. Congratulate you select for interview process
4. You can come our office any time am

Observations are position of each word in each sentence. For same example position (you) is: 1st in sentence 1, 2nd in sentence 3, 1st in sentence 1.

Initial probability (I.P.) of each word (uni-gram): occurrence of particular word in that paragraph (frequency of that word). For same example I.P. of you is 3

Transition probability, also called as bi-gram is frequency of words w1 and w2 occurring together. For above P (select | you) = 1.

Observation probability is count of number of times word W1 at same position say O1. For same example word "you" occur 2 time on 1st position. So P (1|you) = 2.

Inference: we choose four categories for our research, that's why we need to build 4 HMM models, one for each category. Each statement of paragraph associated with one label (category). Whichever label is more associated with statements in paragraph that label is given to paragraph. This process remains same during training and testing phase. For example, in one paragraph 2 statements is of "educational" label and one statement is of "entertainment" label then that paragraph labelled as "educational". So in general paragraph belongs to which label defined as:

Maximum (educational score, social score, entertainment score, personal score)

And if it following in none of the above category, then it following in miscellaneous category.

V. CONCLUSION & FUTURE WORK

In this paper, we have combined machine learning approaches. We have used Agglomerative hierarchical clustering using average distance measure for topic detection. We would be applying system for sports domain, In future we would check our approach on politics, entertainment, science and discovery, etc.. For future work we can add Topic Tracking module within our approach.

REFERENCES

[1] A. A. Kumar, "Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering Sri Sivani College of Engineering Sri Sivani College of Engineering," vol. 1, no. 5, pp. 1–6, 2012.

[2] A. Krause and C. Guestrin, "Data Association for Topic Intensity Tracking," 2006.

[3] A. Saha, "Learning Evolving and Emerging Topics in Social Media : A Dynamic NMF approach with Temporal Regularization Categories and Subject Descriptors."

[4] B. Acun, A. Ba, O. Ekin, M. İ. Saraç, and F. Can, "Topic Tracking Using Chronological Term Ranking," vol. 25, 2011.

[5] B. Pouliquen, R. Steinberger, C. Ignat, E. Käsper, and I. Temnikova, "Multilingual and cross-lingual news topic tracking," 1998.

[6] C. Elkan, "Text mining and topic models The multinomial distribution," 2013.

[7] C. Aksoy, F. Can, and S. Kocberber, "Novelty Detection for Topic Tracking," vol. 63, no. 4, pp. 777–795, 2012.

[8] C. S. Series, R. A-, and A. Xii, Semantic Classes in Topic Detection and Tracking Juha Makkonen. 2009.

[9] C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel, "Large , Multilingual , Broadcast News Corpora For Cooperative Research in Topic Detection And Tracking : The TDT-2 and TDT-3 Corpus Efforts," no. January 1998, 1999.

[10] D. Eichmann, M. Ruiz, P. Srinivasan, N. Street, C. Culy, and F. Menczer, "A Cluster-Based Approach to Tracking , Detection and Segmentation of Broadcast News."

[11] F. Perez-tellez, D. Pinto, J. Cardiff, and P. Rosso, "Clustering Weblogs on the Basis of a Topic Detection Method," pp. 1–10.

[12] F. Fukumoto and Y. Yamaji, "LNAI 3651 - Topic Tracking Based on Linguistic Features," pp. 10–21, 2005.

[13] I. De, "Experiments in First Story Detection," pp. 1–8, 2005.

[14] I. Z. B. Bigi, A. Brun, J.P. Haton, K. Smaili, "Dynamic Topic Identification: Towards Combination of Methods.," pp. 7–9.

[15] J. Allan, S. Harding, D. Fisher, A. Bolivar, S. Guzman-lara, and P. Amstutz, "Taking Topic Detection From Evaluation to Practice," pp. 1–10, 2004.

[16] J. M. Schultz and M. Liberman, "Topic Detection and Tracking using idf-Weighted Cosine Coefficient," pp. 2–5.

[17] K. Kaur, "International Journal of Advanced Research in A Survey of Topic Tracking Techniques," vol. 2, no. 5, pp. 384–393, 2012.

[18] K. Kaur and V. Gupta, "Racking for punjabi language," vol. 1, no. 3, pp. 37–49, 2011.

[19] K. Megerdoomian and A. Hadjarian, "Automatic Topic Detection in Persian Blogs," 2011.

[20] M. Mohd, "Design and Evaluation of an Interactive Topic Detection and Tracking Interface," 2010.