International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 4
April 2017

Summarization of Changes in Dynamic Text Collections using Latent Dirichlet Allocation Model

Mr. Nilesh Nanaware
UG Student
Department of Computer Engineering
Pillai HOC College of Engineering and Technology, Rasayani,Raigad,India

Ms. Damini Daki
UG Student
Department of Computer Engineering
Pillai HOC College of Engineering and Technology, Rasayani,Raigad, India

Ms. Saikamal Dangeti
UG Student
Department of Computer Engineering
Pillai HOC College of Engineering and Technology, Rasayani,Raigad, India

Ms. Sucheta Nikam
Professor
Department of Computer Engineering
Pillai HOC College of Engineering and Technology, Rasayani,Raigad, India

ABSTRACT-

In this digital world, it is very difficult to retrieve information from where the documents are updated or modified usually. Let us take example of World Wide Web where the information changes both usually and importantly over time. Previous projects of abstraction of web documents simply reflects the latest version of each document discarding the dynamic nature of the web. This paper proposes a project with new challenge of the automatic abstraction of changes in dynamic text collections. Along with standard text summarization, this retrieval techniques displays a summary to the user by capturing the major points expressed in the most recent version of an entire document in a compressed form. A system based on Latent Dirichlet Allocation model (LDA) which is used to find the invisible topic formation of changes. The purpose of using the LDA model is to recognize different topics where the changes are made.

Keywords: Changes, summarization, Latent Dirichlet Allocation.

I. INTRODUCTION

Although the Internet is a major force in generating unstructured text (e.g., blogs, wikis, e-mails, chat-rooms, online surveys, etc.), businesses are also generating unstructured text data by streamlining their traditional document management practices to online operations. For many businesses, the mission seems to be to eradicate paper by streamlining their record management practices to maintaining electronic records of each transaction. Thus, unstructured text is constantly being produced, stored, and accessed through various means and it is generally being placed at our disposal through the click of a computer mouse.

Processing these documents by humans is a daunting task not only because of their large quantities but also because the length of many of these documents can be extremely long. Processing of these documents usually requires our cognitive abilities to perform content analysis, classification and deduce topics. [3]

However, as humans, we cannot possibly perform these tasks on the ever increasing volume of unstructured text that is being generated. To address this problem, a number of algorithms for text data processing are being developed. In this project, we tried to implement the **LDA (Latent Dirichlet Allocation)** which represents documents as **mixtures of topics** that spit out .

As our mutual knowledge continues to be digitized and stored—in the form of news, blogs, web pages, scientific articles, books, images, sound, video, and social networks—it becomes more hard to find and discover what we are looking for. We need new computing tools to help order, search and understand these huge amounts of information.

Right now, we work with online details using two main device search and links. We type keywords into a search engine and find a set of documents related to them.[1] We look at the documents in that set, possibly should be navigated to other linked documents. This is a high powered way of interacting with our online archive, but something is missing.

Visualize searching and exploring documents based on the topic that run through them. We might “zoom in” and “zoom out” to search specific or wide topics; we might look at how those topic modified through time or

how they are connected to each other. Rather than 1 finding documents through keyword search alone, we might first find the topic that we are interested in, and then examine the documents related to that topic.

II. LITERATURE SURVEY

With the help of LDA modelling technique we are able to make the summery of text. It helps to generate the sentences out of the whole document which will lead to easy reading & in advance it reduces the time required to read the whole document. It provides the helpful information to get known the content of the entire document. LDA reads the whole document & generates the result according to the user need which will result in increased speed.

We also provided ease access to load the document whom abstract is needed, In this the user can get the abstract of several language like france, english etc. The user can also put the data randomly.

III. EXISTING SYSTEM

The topic models developed in recent years have been classified into three categories:

Set theoretical methods (these methods represent the documents as set of words and use set theoretic operations to do further processing).

Set theory is a branch of mathematical logic that studies sets, which informally are collections of objects. Although any type of object can be gathered into a set, set theory is functional most often to objects that are related to arithmatic. The language of set theory can be used in the definitions of all mathematical stuff.

The new study of set theory was initiated by Georg Cantor and Richard Dedekind in the 1870s.

Arithmetic models (These methods represent papers and queries by vectors/matrices/tuples).

Have you ever visited to the bank to deposit money into an interest bearing account, and the banker was able to tell you how much interest you will have after a certain number of years? If so, you may have imagined how they were able to compute this amount. One way of finding this is by using what's called an **arithmetic model**. An algebraic model takes a real-world situation described in words and describes that situation using algebra.

Probabilistic models (These models treat the process of topic modelling as a problem of probability. Theorems of probability are often used in these methods).

Probabilistic models have been the area of awareness for most of the examine in the last decade, as they are considered better than the other approaches of using sets and matrices for topic modelling. [2] The idea of the proposed topic model has also been taken from a probabilistic model named, Latent Dirichlet Allocation (LDA). LDA is the simplest and the most popular probabilistic model. As it comprises of all the basic steps involved in the extraction of the theme of a document, it serves as a basic reference framework for other probabilistic topic models.

IV. IMPLEMENTATION

The idea of the strategy being followed for topic modelling has been taken from LDA. If a collection of words of literature is available and these words are arranged into different lists in such a way that the words with similar meaning and falling into same category are placed in same list, then each of the word from the input text can be assigned a category to which they belong (per word topic assignment) leaving helping verbs etc. This process is called Generative Process[7], as it justifies how the input text would have been generated. After this process, the proportion of involvement of each topic in the input text is computed. This second process is termed as Statistical Inference Process. The collection of topics that are involved in the text (example 10% about education, 30% about health, and so on) can help to a great extent in finding out what theme the text contains. The idea of the proposed process is illustrated in Figure show below.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 4
April 2017

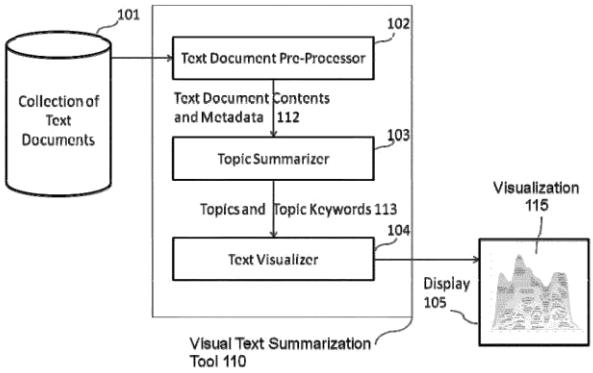
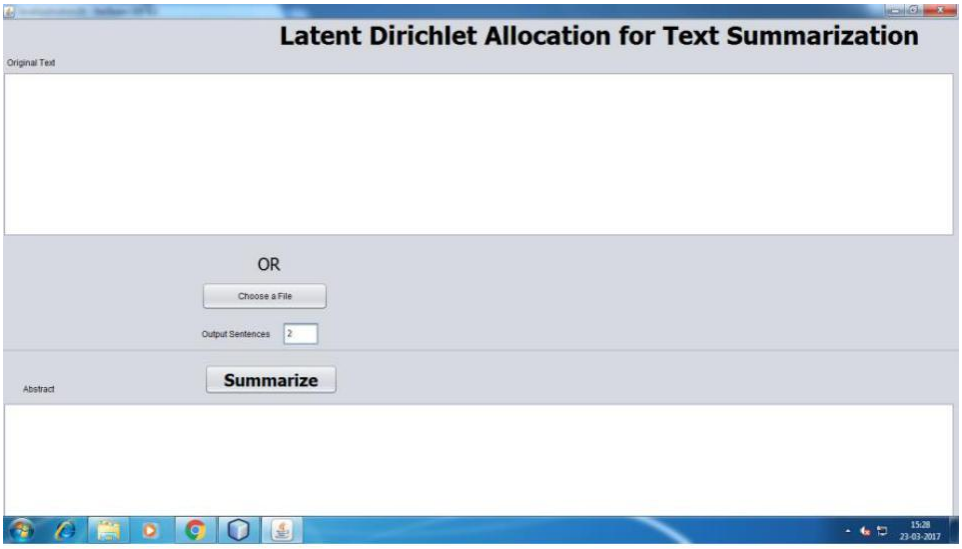
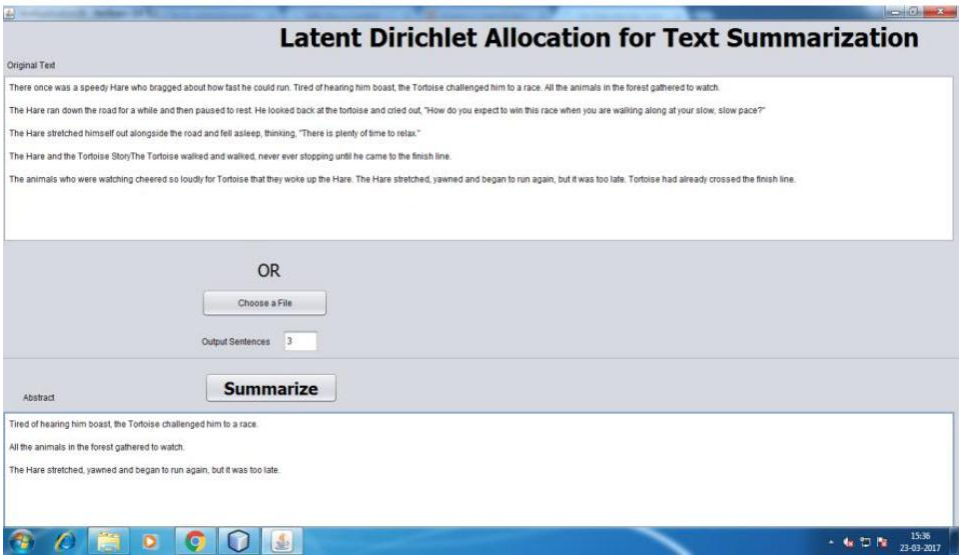


Fig 1. Implementation of text

V. RESULT





VI. CONCLUSION

The summarization of changes focuses on the generation of abridged and non-redundant accounts of document modifications in dynamic text collections. This research introduces a new framework for summarizing changes from a set of revisions made to a Wikipedia article during a given time period.

We implemented Latent Dirichlet Allocation (LDA) for topic modelling. We performed topic analysis on tweets and web page collections. Using our topic model, we extracted meaningful topics from all tweet and web page collections. Our evaluation results confirm the good quality of our topic model. We also devised an automated approach to label the topics that we obtained from LDA. In this project report, we have given complete information on background, implementation and evaluation details of our approach. We have developed an extensive developer's manual to facilitate someone to extend our work in future.

VII. REFERENCES

[1] Steyvers M, Smyth P, Rosen M, Griffiths T. Probabilistic author-topic models for information discovery. Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004. p. 306–15.
[2] Barbieri N, Manco G, Ritacco E, Carnuccio M, Bevacqua A. Probabilistic topic models for sequence data. Mach Learn. 2013; 93(1):5–29.
[3] Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Modeming general and specific aspects of documents with a probabilistic topic model. In *NIPS* (pp. 241–248).
[4] Ciglan, M., & Nørvåg, K. (2010). Wikipop: Personalized event detection system based on Wikipedia page view statistics.
In *Proceedings of the CIKM'10*.
[5] W. Buntine. Variational extensions to EM and multinomial PCA. In European Conference on Machine Learning, 2002.
[6] W. Buntine and A. Jakulin. Discrete component analysis. In Subspace, Latent Structure and Feature Selection. Springer, 2006.
[7] J. Chang and D. Blei. Hierarchical relational models for document networks. Annals of Applied Statistics, 4(1), 2010.
[8] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391{407, 1990.
[9] Wikipedia LDA: http://en.wikipedia.org/wiki/ Latent_Dirichlet_allocation.