Vol. 1, No. 1 | January - June 2018

Estimating News Coverage Patterns using Latent Dirichlet Allocation (LDA)

Batool Zehra[1],Naeem Ahmed Mahoto[1],Vijdan Khalique[1]

Abstract:

The growing rate of unstructured textual data has made an open challenge for the knowledge discovery which aims extracting desired information from large collection of data. This study presents a system to derive news coverage patterns with the help of probabilistic model – Latent Dirichlet Allocation. Pattern is an arrangement of words within collected data that more likely appear together in certain context. The news coverage patterns have been computed as number function of news articles comprising of such patterns. A prototype, as a proof, has been developed to estimate the news coverage patterns for a newspaper – The Dawn. Analyzing the news coverage patterns from different aspects has been carried out using multidimensional data model. Further, the extracted news coverage patterns are illustrated by visual graphs to yield in-depth understanding of the topics which have been covered in the news. The results also assist in identification of schema related to newspaper and journalists' articles.

Keywords: *News Coverage Pattern; Probabilistic Model; data visualization; Multi-dimensional Data Model.*

1. Introduction

The rapid growth of the Web technologies has resulted in number of websites. According to statistics, there are 1800,047,111 active websites recorded in 2017 [1]. That means, tremendous volume of data is produced on these websites, which paves the way for information processing in order to extract knowledge and meaningful patterns. It can assist in decision-making regarding a number of scenarios and problems.

Out of total number of active websites, there are many e-news websites with a large share of visitors who regularly check news and read articles on these websites. For example, in US, around 70% of the population refer to the Internet for keeping up-to-date with news [2]. In addition, a large number of news articles are published on the daily basis covering various current topics. There exists a huge platform that can be explored with modern computing and data processing tools to find out interesting and useful knowledge.

In this research study, useful knowledge is extracted from the e-newspaper articles to obtain the news coverage patterns. News coverage pattern refers to finding issues or topics being discussed over a certain period of time. The aim is to find out the coverage given to a specific topic or issue in news. The coverage patterns would reveal that what topics or issues remained under discussed and trending subject in a newspaper. The news

[1] Mehran University of Engineering and Technology, Jamshoro, Pakistan
Corresponding Email: sayyid_zehra@yahoo.com

coverage patterns have been obtained by means of Latent Dirichlet Allocation (LDA), which is a probabilistic model for collection of discrete data [3]. LDA is a modeling technique, which automatically develops topics based on patterns of co-occurrence of words. LDA finds out a set of ideas or themes that well describe the entire corpus. The scope of this research is focused on the daily news and newspaper articles published on the website of a popular newspaper-The Dawn. During the course of this research, an application prototype as the proof of concept has been developed that performs the task of estimating the news coverage patterns. Text mining and analysis is performed to check coverage of certain news. Precisely, the goal of this research is to identify and highlight the topics or issues under discussion in numerous articles of the Dawn newspaper. Furthermore, the trending topics are statistically represented as knowledge to end-users. The information visualization has been considered an emerging field for the several application domains; for instance, structural information by injecting parameters of location has been represented in visual formats [7]. The paper is organized as follows: Related studies are discussed in section 2; the working principle of extracting news coverage patterns is reported in section 3; the outcomes of the study are described in the discussion section 4; finally, section 5 presents conclusions.

2. Literature

Maintaining the integrity of the specifications reference [4] described way to utilize LDA to journalism. The study [5] evaluated two novel approaches, one by using a video stream and second by using the closed caption stream. An LDA approach has been used to detect the text stream and the person shots. The research concluded that the individual system gave comparable results, however; a combination of the two systems provided a significant improvement as compared to the individual system.

Reference [6] treated groups of objects together as spatial visual words and investigated configurations of regions using LDA and invariant descriptors. In this research, computation of invariant spatial signatures for pairs of objects was based on a measure of their interaction inside the scene. A simple classification was used to define spatial visual words to extract new patterns of similar object configurations. The modeling of the scene into a finite mixture was in accordance with the spatial visual words by the use of latent dirichlet model. Statistical analysis was done to better understand the spatial distributions inside the discovered semantic classes. One case study for synthetic imagery and for real imagery was experimented using LDA and the results proved that this model has good performances with the small amount of training data. It has been concluded that scene level analysis can be done through LDA with minimal human interaction leaving behind the traditional approaches of pixel or region level analysis [6]. The reference [8] unearthed significant components, for instance, nouns and verbs given broadcast transcript. It further computed weights of components with the help of their frequency in the text.

The study under consideration applies LDA in order to get news patterns for better understanding of trends of news topics/issues.

3. News Coverage Patterns Extraction Workflow

The workflow of identifying news coverage patterns is explained in this section that comprises of four steps. The steps are explained in the order of their execution as illustrated in Fig. 1. The workflow intends to find out the coverage of various issues/topics in multiple news articles and news. Following are the steps and their explanations.

3.1. Web Crawler

The primary source of data is online news websites. In order to fetch data from these websites, a crawler has been designed and developed. However, crawler targeted only one website – The Dawn (*www.dawn.com*). The crawler gathered daily news and news articles, which needed preprocessing before determining news coverage patterns. These collected news and news articles are stored in documents referred to as document database or corpus.

3.2. Data collection and preprocessing

Preprocessing stage involves the rectification of data such that the subsequent processes can be done. The refinement of data leads to better results, since unnecessary data gets removed and useful data elements are left behind. The data refinement procedure during preprocessing performs activities: *Tokenization, Stop word removal, Stemming and Vector Space Model.*

3.2.1. Tokenization

Tokenization is the process to partition the sentences contained in the textual data into its tokens (i.e., words). For example, consider a sentence 'This study aims at extracting news coverage patterns'; the tokenization results into tokens: {'This', 'study', 'aims', 'at', 'extracting', 'news', 'coverage', 'patterns'}.

3.2.2. Stop word removal

News and articles are read one-by-one by the system. The system removes stop words from the given document. Stop words are commonly used words in language such as *is, at, a, on, of etc*. Their presence can mislead the text search and text analysis. Therefore, stop words are removed from text during preprocessing stage.

3.2.3. Stemming

In the next stage of preprocessing, stemming takes place. Stemming is the procedure of determining the base word or root word of a given word. For example, *extract* is the root word of *extracting.* During stemming, each word is traced to its basic root word. This is an important process as it can help remove noise from data.

3.2.4. Vector Space Model

Vector space model represents text documents as vectors. Vector(s) corresponds to the dimension of the vector space. The text data of news and news articles has been represented in the form of bag-of-words.

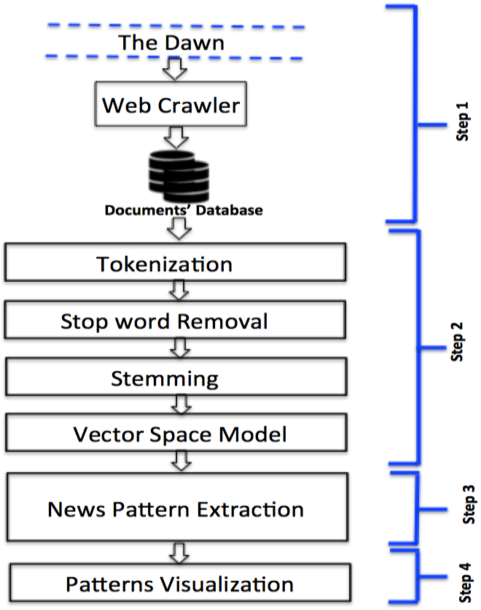

Fig. 1. Workflow of news pattern extraction and visualization.

3.3. News Pattern Extraction

News pattern extraction refers identifying patterns about coverage of certain topics in news and articles. LDA finds out what topics are discussed in a given article by observing processed news data (i.e., vector space model – bag-of-words) and produces a topic distribution. A prototype application developed in this study has implemented LDA for the news coverage pattern extraction.

Latent Dirichlet Allocation (LDA) – LDA algorithm reported in [3] presents the mathematical model. The principle concept of LDA describes that documents' database (or corpus) contains words referring to latent topics and thus relate to the overall theme of the documents.

Consider a set of ***M*** documents' database (***DD***) or corpus such that $DD = \{d_1, d_2, ..., d_m\}$ $\sum n = \{p_1, \dots, p_k\}$, where ***dᵢ*** represents set of documents in the corpus **DD** and each document ***d*** is a vector of ***N*** comprises of words ***wᵢ*** such that $d = \{w_1, w_2, ..., w_n\}$.

LDA accomplishes steps as reported below for each ***w*** in corpus ***DD***.

i. Select N from Poisson Probability (ξ)

ii. Select q from Dir(a), where a shows per-document topic distribution

iii. For each of the ***N*** words ***wₙ***:

   a. Select a topic ***zₙ*** ~ Multinomial(θ)

   b. Select a word ***wₙ*** from $p(w_n | z_n, b)$, a multinomial probability conditioned on topic ***zₙ***; $b_{ij} = p(w^j = 1 | z^i = 1)$ probability of ***wₙ*** towards the topic ***zₙ***

Having parameters a, b, the joint distribution of a topic mixture q, a set of N topics **z**, and a set of N words **w** is given by:

$$p(q, z, w | a, b) = p(q | a) \prod_{n=1}^{N} p(z_n | q) p(w_n | z_n, b)$$

3.4. News Pattern Extraction

Visualization is the final output of this study. The extracted news patterns are presented statistically using graphs, line charts and bar charts. The statistical representation gives a comprehensive view of how certain topics are discussed in news and articles in a newspaper.

4. Discussion

LDA helps to find the main theme of an article and discover the coverage of specific news theme. The discovered theme is then chronologically ordered and presented in this section as bar graphs and line charts in order to show the trend of various patterns of news issues and topics.

Figure 2 and 3 show the trends of topics mentioned in newspaper. These topics are frequently discussed and have been in news for the given period of time. The frequencies are plotted on y-axis while x-axis possesses timings. Each topic has its own line with different colors to indicate the change in trends. For instance, the topic named *NAB* in Fig. 2 remained top in news during 15 days. Similarly, trends about *corruption* and *politicians* have been the most frequent topics during the given dates. It can also be observed that these topics have almost similar trend with slight variations. This trend leads to understanding the media coverage in newspapers. In other words, media targets news about *NAB*, *corruption* and *politicians*. The similar trends are depicted in Fig. 3.

In Fig. 4, the topics (i.e. army, military, Pakistan) discussed in news articles by different columnist of The Dawn newspaper are depicted. The topics being covered by these writers show a trend of certain issues. These trends represent that who among the considered writers focuses which of the issues and topics in their writings.

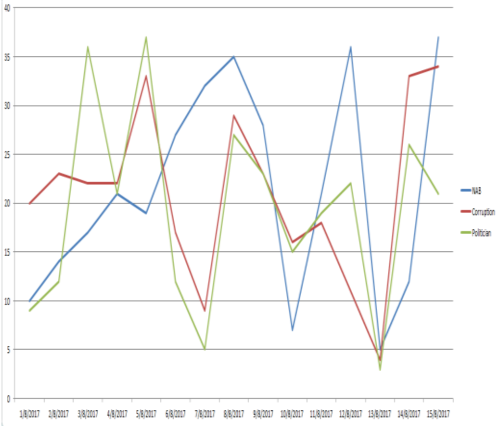

Fig. 2. News coverage trend of 3 topics - 15 days.

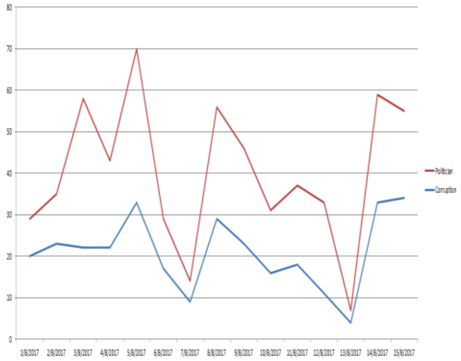

Fig. 3. Number of news regarding certain topics.

For example, referring to Fig. 4, Cyril Almeida mostly talked about Pakistan, military and army in his news articles. These news coverage patterns and topics covered by the writers in their writings clearly yield the direction and mindset of the media personnel, since media has been considered as opinion maker for the societies.

Fig. 5 provides information regarding available collected news and news articles in the database. The prototype application offers to search for the topics or terms in the specified range of time within news or in news articles of columnists.

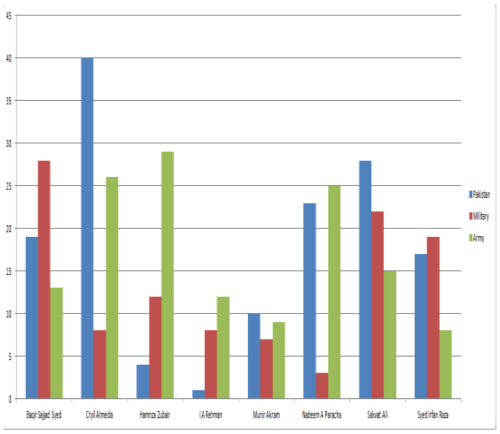

Fig. 4. Frequency of topics covered in articles written by news analysts.

For instance, Fig. 6 represents that term *PPP (Pakistan People's Party)* has been trending more in the news on 5th March 2017 as compared with rest of the days of March 2017 in the available database of the prototype application. The news coverage patterns and their trend helps in understanding the behavior and mindset of media personnel, who play an essential role in building opinions of the people. The patterns not only will serve government officials to have in-depth information about directions of media and its agenda. Likewise, the writings of certain columnists will help in understanding the targets and their priorities in building the nation.

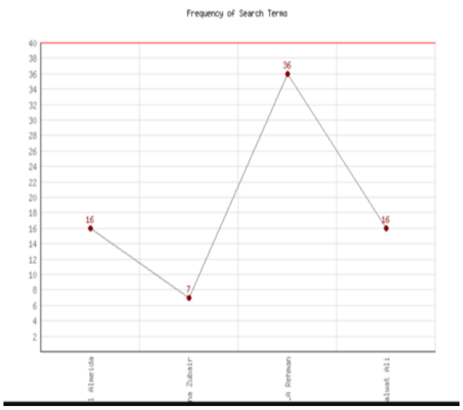

Fig. 5. The news and news articles in the corpus.

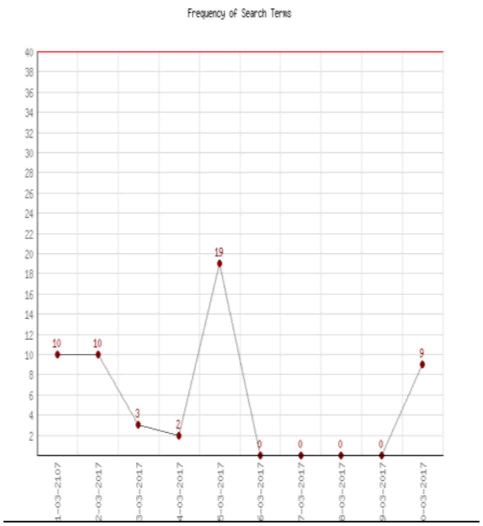

Fig. 6. Term PPP (Pakistan People Party) coverage trend.

5. Conclusion

This paper presents an approach to apply Probabilistic Topic Model (i.e., LDA) in finding significant patterns from newspapers. The data, crawled from The Dawn newspaper's official website, has been processed to uncover the news coverage in certain time limits. To validate the potential of understanding news coverage patterns, a prototype application has been built, which performed the necessary steps to reveal the patterns. These patterns have been presented with the help of visualization methods.

It has become a fact that media influences perception of general public and mold their sentiments and thinking about certain issues and events. This research would assist in identifying the narrative of media groups regarding certain issues and events. The point of view of article writers can be stipulated by applying the analytical approach presented in this study. Consequently, the inclination of a newspaper can be judged based on the coverage they are giving to various issues. The in-depth knowledge about media groups and their news coverage patterns may assist government authorities like PEMRA (Pakistan Electronic Media Regulatory Authority) to regulate the electronic print media.

As future works, we plan to compare the applied approach with the state-of-the-art methods and included other newspapers as well as research articles to determine the topical coverage of the scientific research articles.

REFERENCES

[1] D. S. Fowler, "Tek Eye How Many Websites Are There In The World?," Online: tekeye.uk/computing/how-many-websites-are-there (Accessed 30-1-2018).

[2] Fuller, Steve, "Topic: News Industry", Online: www.statista.com/topics/1640/news/ (Accessed 20-1-2018).

[3] D. M. Blei, Y. Ng. Andrew, and Michael I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, pp. 993-1022, Jan 2003.

[4] C. Jacobi, W. V. Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling", *Digital Journalism*, vol. 4, no. 1, pp. 89-106, 2016.

[5] H. Misra, F. Hopfgartner, A. Goual, P. Punitha, and J. M. Jose, "News Story Segementation Based on Semantic Coherence and Content Similiarity", *In MMM*, pp.347-357, January 2010.

[6] C. Vaduva, I. Gavat, and M. Datcu, "Latent Dirichlet Allocation for spatial analysis of satellite images," *IEEE Transactions on Geoscience and Remote sensing*, vol. 5, no. 15, pp. 2770-2786, 2013.

[7] S. Shah, V. Khalique, S. Saddar, and N. A. Mahoto, "A Framework for Visual Representation of Crime Information," *In Indian Journal of Science and Technology*, vol. 10, no. 40, pp. 1-8. ISSN: 0974-6846, 2017.

[8] M. J. Pickering, L. Wong, and S. R¨uger, "ANSES: Summarisation of news video," Image and Video Retrieval, LNCS 2728, pp.425-434, 2013.