A LEXICON BASED APPROACH TOWARDS CONCEPT EXTRACTION

Anoud Shaikh
Mehran University of Engineering and Technology. Sindh (Pakistan)
E–mail: anoudmajid85@gmail.com

Naeem Ahmed Mahoto
Mehran University of Engineering and Technology. Sindh (Pakistan)
E–mail: naeem.mahoto@faculty.muet.edu.pk

Mukhtiar Ali Unar
Mehran University of Engineering and Technology. Sindh (Pakistan)
E–mail: mukhtiar.unar@faculty.muet.edu.pk

Recepción: 05/03/2019 **Aceptación:** 19/03/2019 **Publicación:** 17/05/2019

Citación sugerida:
Shaikh, A., Mahoto, N. A. y Unar, M. A. (2019). A Lexicon based Approach Towards Concept Extraction. *3C Tecnología. Glosas de innovación aplicadas a la pyme. Edición Especial, Mayo 2019*, pp. 50–67. doi: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.50–67

Suggested citation:
Shaikh, A., Mahoto, N. A. & Unar, M. A. (2019). A Lexicon based Approach Towards Concept Extraction. *3C Tecnología. Glosas de innovación aplicadas a la pyme. Special Issue, May 2019*, pp. 50–67. doi: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.50–67

ABSTRACT

The emergence of digital media has tremendously increased the amount of unstructured data. Recently 80% of data, generated over the web, is in an unstructured format. This immense amount of data is a great source for the knowledge discovery and thus, may be utilized for extracting purposeful information. This study adopted a lexicon–based approach for automatic concept extraction from online news stories and events. An application prototype has been developed to demonstrate the applicability and effectiveness of the adopted approach. The extracted knowledge about news stories, articles and blogs are essential in understanding in–depth information for news analysts. This knowledge plays a vital role in building societies since media is considered as an opinion maker for its audience.

KEYWORDS

Online news, Unstructured data, Concept extraction.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.50-67

1. INTRODUCTION

The digital age has provided an immense amount of data in terms of news articles, social media data, and web LaValle, Lesser, Shockley, Hopkins & Kruschwitz, 2014; Gharehchopogh & Khalifelu, 2011). Every day, a large amount of data is published on the news websites, micro–blogging websites and other information repositories (Lei, Rao, Li, Quan & Wenyin, 2014). The published news articles reveal the events happening around the world (Lei, *et al.*, 2014). The challenging issue, specifically, in the textual data format (i.e., news articles) is to extract purposeful information. Manually, it is a hard task to interpret a large collection of data (Lee, Park, Kim & No, 2013). Besides, the information hidden in unstructured data format inherently makes it difficult processing tasks, because it deals with natural language processing. Therefore, in the current era of information flow, media analysts and other researchers need an easily understandable and high–level summary of information. For instance, a media analyst may require searching news regarding a certain topic, events happening to a certain geo–location, and/or news events based on a timeline. These and other such queries are objectives, which requires an efficient method to answer such queries.

Text Analytics allows knowledge discovery and purposeful finding of information from such a massive amount of data for investigation. The extracted knowledge can be used for better decision–making strategies and effective resource management. Therefore, extracting purposeful knowledge from large data having natural language involvement is an open challenge, which acquires sophisticated methods and algorithms to deal with it. To this aim, this research study extracts concepts from a large number of news stories and articles. The concept extraction refers to a meaningful sequence of words that are used to represent objects, events, activities, entities (real or imaginary), topics or ideas, which are of interest to the users (Parameswaran, Garcia–Molina & Rajaraman, 2010; Szwed, 2015). The concept extraction technique is a very effective way of extracting all the possible useful and meaningful concepts from text documents. The extracted concepts, later, may be tagged as essential concepts and may be represented in an efficient

mechanism (Zhang, Mukherjee & Soetarman, 2013). The concepts, especially, present the understanding of the unstructured data format. The coverage and patterns of such concepts help in understanding in–depth about the news stories, news articles and inclination of the author's mindset. This knowledge about news stories, articles and blogs are essential for news analysts and plays a vital role in building societies, because media plays the role of opinion maker for the inhabitants of society.

An application prototype has been developed in this study to demonstrate the automated concept extraction that works based on lexicon approach. On the contrary, the machine–learning approach (i.e., supervised learning) inherently possesses challenges due to unstructured data format. Whilst, the lexicon–based approach has produced comparatively better results. The developed prototype presents the applicability and effectiveness of the considered approach.

This paper is structured as follows: section 2 reports existing scientific literature about concept extraction, section 3 describes the architecture of the developed application. Results and discussion are reported in section 4, and finally, section 5 concludes.

2. RELATED WORK

The concept extraction has been remained focus of in the recent existing literature (Sˇili´c, *et al.*, 2012; Parameswaran, *et al.*, 2010; Villalon, *et al.*, 2009, Weichselbraun, *et al.*, 2013; Termehchy, *et al.*, 2014; Brin, 1998, Mahmood, *et al.*, 2018). Specially, concept extraction in the context of online news has become ta opic of interest. For instance, social emotions have been detected using a lexicon–based approach from news articles in (Lei, *et al.*, 2014). CatViz Temporally – Sliced Correspondence Analysis Visualization performs exploratory text analysis on large collection of textual data. The basis of CatViz is Correspondence Analysis (CA) and allows va isual analysis of different aspects of text data (Sˇili´c, *et al.*, 2012).

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.50-67

Extraction of concepts from query log data repository has been carried out in Parameswaran, Garcia–Molina and Rajaraman (2010), where sub–concepts and super concepts are pruned. The core concepts are taken into consideration, which is oriented on frequency, better meaning and idea. Similarly, automatic concept extraction from essays written by students in order to draw concept maps is reported in Villalon and Calvo (2009) for the concept map mining purpose. The limitations faced in machine–learning approach during training model have been addressed in Weichselbraun, Gindl and Scharl (2013). Two potentially efficient algorithms have been proposed Termehchy, Vakilian, Chodpathumwan and Winslett (2014), namely: 1) Approximate Popularity Maximization (APM) and Annotation–benefit Maximization (AAM). The patterns hidden in web documents have been explored in Brin (1998), where patterns are analyzed for concept determination.

The Dawn (newspaper) and The New York Times (newspaper) have been focused on Mahmood, Kausar and Khan (2018) for the purpose of textual analysis. `This study also focused on online news stories and events published at The Dawn (newspaper) as the data source in order to automatically extract concepts using lexicon–based approach. The dictionaries used for the understanding of concepts and meanings of the terms and/or concepts are WordNet and DBPedia.

3. LEXICON BASED CONCEPT EXTRACTION APPROACH

An application prototype has been developed for online news data in order to extract key concepts. The prototype is developed using C# (c sharp) programming language.

The application architecture of the developed prototype comprised of three layers: Layer 1: Data Source, Layer 2: Middleware and Layer 3: News Mining as shown in Figure 1. The purpose of each layer is reported in the subsequent sections.

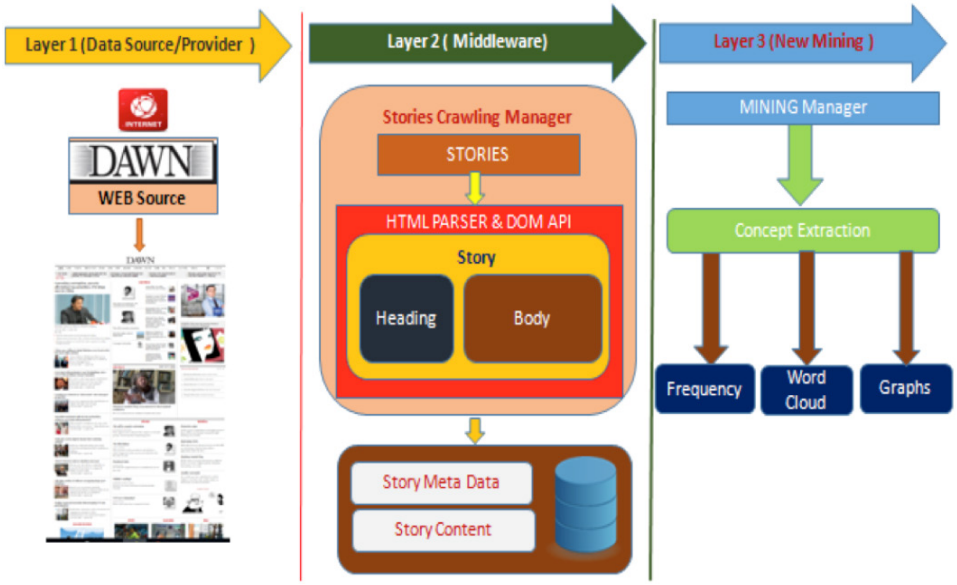

Figure 1. Prototype Application Architecture.

3.1. LAYER 1: DATA SOURCE/PROVIDER

Data source/Provider layer crawls online news events and stories published at ***The Dawn***[1] newspaper official website. The application, however, allows providing URL (Uniform Resource Locator) of a certain news website. This study has focused on the news stories and articles of The Dawn newspaper. This layer traverses the given URL to crawl its news events and stories available at its several webpages. The crawler uses the existing APIs (Application Programming Interfaces) for the traversing and retrieval of data from the source website (***The Dawn*** in this case).

3.2. LAYER 2: MIDDLEWARE

Middleware layer takes the news stories and articles and parses the given obtained news stories. In particular, HTML (Hypertext Markup Language) parser and DOM (Document Object Module) API has been used for the processing of news stories. The parsed and processed data is stored in the relational database.

1 The Dawn (www.dawn.com)

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.50-67

A relational database is the collection of data into table formats, which are logically related to each other. The news stories and articles comprised of several tags as represented in Figure 2.

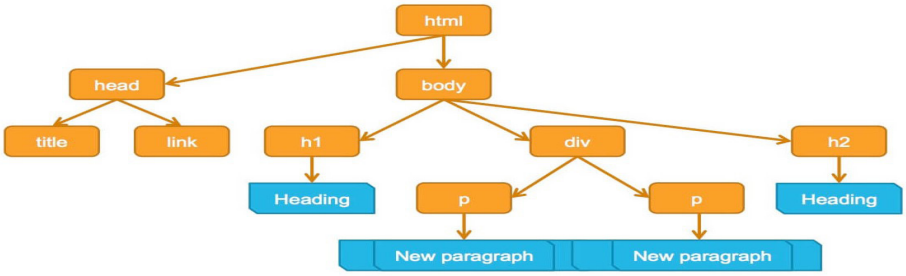

Figure 2. HTML webpage tags in a tree structure.

Middleware layer takes the news stories and articles and parses the given obtained news stories. In particular, HTML (Hypertext Markup Language) parser and DOM (Document Object Module) API has been used for the processing of news stories. The parsed and processed data is stored in the relational database. A relational database is the collection of data into table formats, which are logically related to each other. The news stories and articles comprised of several tags as represented in Figure 2.

HTML Parser is, basically, a library and it is used in parsing if text files formatted in HTML. Likewise, DOM API, written in JavaScript, is an object representation of webpage. The news stores and articles are provided as an input to the third layer of the developed prototype application.

3.3. LAYER 3: NEWS MINING

This layer is the key layer that actually automatically extracts concepts present in the collected news stories and articles. In particular, this layer comprised of Mining Manager, which performs necessary text preprocessing steps to transform the collected and stored news stories and articles into a suitable format for further processing.

Mining Manager: it performs tokenization, stemming, stopword removal operations before actual processing of automatic concept extraction.

Tokenization: this operation breaks given textual data into its tokens (i.e., terms or words). For instance, consider a sentence *'This study aims at automatic concept extraction using lexicon–based approach.'* The tokenization produces the following outcomes: ***'This', 'study', 'aims', 'at', 'automatic', 'concept', 'extraction', 'using', 'lexicon', 'based', 'approach', '.'.***

Stemming: stemming refers to an operation in which words (i.e., tokens) obtained from the previous step (i.e., tokenization) are acquired into their roots or base words. For instance, ***'Multiplying'*** becomes ***'Multipli'***, ***'Engineering'*** becomes ***'Engine'*** and many more. This step helps into a reduction of redundant terms used in textual data.

Stopword Removal: this operation prunes unnecessary words present in the text. These unnecessary words usually refer to auxiliary verbs and grammatically articles. For example, ***'the' 'is' 'am' 'are' 'was' 'and' 'a' 'an'*** and many more.

The processed tokens are further used as an input for automatic concept extraction. The application uses popular bag–of–words (BOW) as vector space representation model for the processed tokens. The words that are left after stopword removal operation are the bag–of–words, each word has its frequency in certain news story or article. The BOW is supplied to Concept Extraction module for determining concepts.

Concept Extraction: The concept extraction module determines the meaning and being concept state of terms, which have been processed at Mining Manager. The BOW is further supplied to concept extraction module as shown in Figure 3 that is connected with dictionaries: WordNet, DBpedia and Linked data to determine the meaning and concept for a given word of BOW.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.50-67

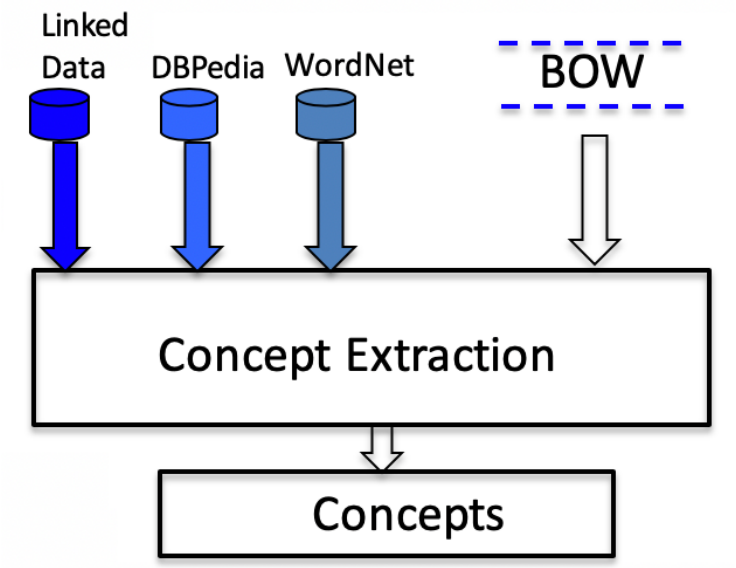

Figure 3. Lexicon based Concept Extraction Approach.

Each word in BOW undergoes for the concept extraction process. The outcomes of the concepts are later used for visualization. In particular, the frequency of the concepts is measured for a given article or news story. The word cloud is displayed for the concepts and graphs represent the trends of the concepts available in the news.

4. RESULTS AND DISCUSSION

This section discusses the outcomes of the developed application prototype. Figure 4 represents the crawled data. A certain URL of the newspaper website is provided to crawl its data and store into a database. The collected story is displayed at the user interface of the application as in Figure 4.

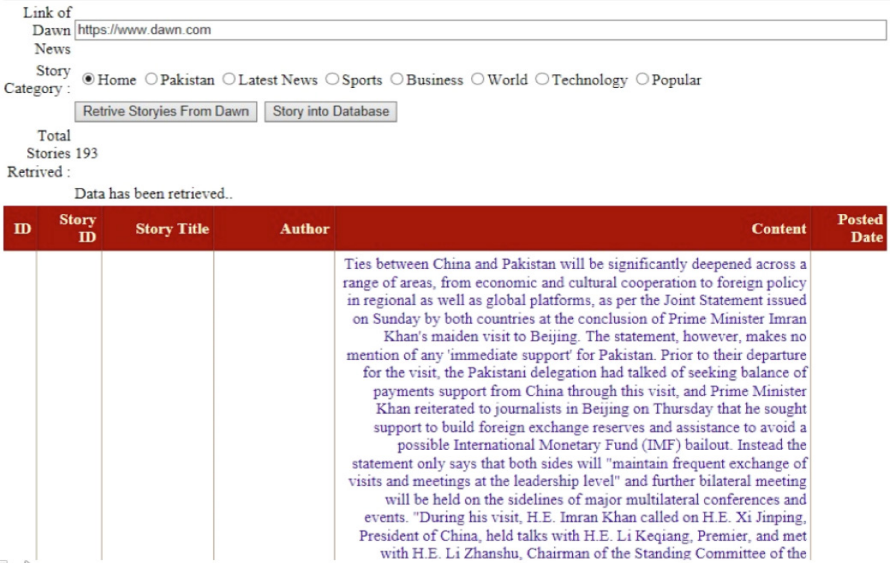

Figure 4. Developed Application Prototype.

Figure 5 represents the concepts extraction and the frequencies used for the BOW as an input for concept extraction module discussed in section 3.3. Since the BOW is large in number, the prototype allows increasing or decreasing the number terms in BOW based on their frequencies.

Figure 5. Prototype – Concepts and their Frequencies.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.50-67

The graphs and concepts in terms of word cloud are presented in the developed prototype for a better understanding of concepts present in the news stories and articles as reported in Figure 6.

Figure 6. Prototype – Concepts and their Word Cloud.

The outcomes of the approach help in understanding in–depth news stories and articles, which may be used as a baseline for the decision–making strategies. The news media has been used widely for opinion making purposes. Thus, the extracted concepts help in getting an insight into the news events, articles and mindset of the journalists.

5. RESEARCH CHALLENGES AND LIMITATIONS

To acquire data for this study, The Dawn newspaper has been targeted due to its popularity and neutrality. This could be considered as the limitation of the study since the emphasis of the study remained over concept extraction using lexicon approach. However, the developed approach may also be provided a dataset of any other newspaper.

The challenges that have been encountered during the course of the research study is PakistaniEnglish words. The injection of *Urdu* words in English has been referred to as Pakistani English. For instance, *chai–wala, ziaism, Sahab* and *Naya Pakistan* are some of the PakistaniEnglish vocabulary. The challenge is to determine the concepts from this derived vocabulary. PakistaniEnglish vocabulary has been not addressed in the study due to lack of its lexical chains and thorough grammatical aspects that help in understanding words.

6. CONCLUSION

This study reported a lexicon–based approach for concept extraction. In particular, a working prototype has been developed to demonstrate the applicability and effectiveness of the approach. The application automatically crawls news events and stories, which are stored in a relational database. The focus remained on The Dawn newspaper for the data source due to its neutrality and popularity in the region.

The collected news data has been gone through necessary text processing phases in order to transform it for further process. The concepts have been extracted with the help of WordNet, DBPedia and Linked Data. The extracted concepts, later, have been displayed with visualization techniques such as Word Cloud and charts. Generally, media has an influential role over the minds of its audience. Thus, the extracted concepts may help in understanding the core concepts available in the news events and stories that may lead to strategic decision–making. The outcomes of this study may assist and help media analysts to have an in–depth understanding of media personnel and general public opinion about news and facts on the ground.

The deviation of lexical chains in terms of PakistaniEnglish words will be considered as future work. In particular, developing PakistaniEnglish corpus to tackle the limitations of this study would be a focus in future work.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.50-67

ACKNOWLEDGEMENTS

This research has been performed under the Institute of ICT Mehran University of Engineering and Technology, Pakistan and funded by the ICT Endowment for Sustainable Development.

REFERENCES

Brin, S. (1998). Extracting patterns and relations from the world wide web. In *International Workshop on the World Wide Web and Databases*, 1998, pp. 172–183. Springer, Berlin, Heidelberg.

Gharehchopogh, F. S., & Khalifelu, Z. A. (2011). Analysis and evaluation of unstructured data: text mining versus natural language processing. *Application of Information and Communication Technologies (AICT)*, 5th International Conference on, 12–14 October, 1–4. doi: http://dx.doi.org/10.1109/ICAICT.2011.6111017

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2014). *Big data, analytics and the path from insights to value*. MIT Sloan Management Review, 21.

Lee, J. E., Park, H. S., Kim, K. J., & No, J. C. (2013). Learning to predict the need of summa– rization on news articles. *Procedia Computer Science 24(0)*, pp. 274 – 279. 17th Asia Pacific Symposium on Intelligent and Evolutionary Systems, (IES2013).

Lei, J., Rao, Y., Li, Q., Quan, X., & Wenyin, L. (2014). Towards building a social emotion detection system for online news. *Future Generation Computer Systems, 37,* pp. 438–448.

Mahmood, T., Kausar, G., & Khan, G. Z. (2018). A Critical discourse analysis of the editorials of "Dawn" and "The New York Times" in the aftermath of Army Public School attack. The "Us" versus "Them" ideology. *Journal of Research in Social Sciences (JRSS)*, *6(2)*, pp. 1–17.

P**arameswaran, A., Garcia–Molina, H., & Rajaraman, A.** (2010). Towards the web of concepts: Extracting concepts from large datasets. *Proceedings of the VLDB Endowment, 3*(1–2), pp. 566–577.

Ramirez, P. M., & Mattmann, C. A. (2004). ACE: improving search engines via Automatic Concept Extraction. In *Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference* on pp. 229–234. IEEE.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.50-67

Sˇili´c, A., Morin, A., Chauchat, J. H., Dalbelo Baˇsi´c, B. (2012). Visualization of temporal text collections based on correspondence analysis. *Expert Systems with Applications 39*(15), 12143–12157.

Szwed, P. (2015). Enhancing concept extraction from polish texts with rule management. In Beyond Databases, Architectures and Structures. *Advanced Technologies for Data Mining and Knowledge Discovery*, pp. 341–356. Springer, Cham.

Termehchy, A., Vakilian, A., Chodpathumwan, Y., & Winslett, M. (2014). Which concepts are worth extracting? *ACM international conference on Management of data SIGMOD, 2014*, pp. 779– 790.

Villalon, J., & Calvo, R. A. (2009). Concept extraction from student essays, towards concept map mining. *Ninth IEEE International Conference on Advanced Learning Technologies, ICALT 2009*, pp. 221–225.

Weichselbraun, A., Gindl, S., & Scharl, A. (2013). Extracting and grounding contextualized sentiment lexicons. *IEEE Intelligent Systems, (2),* pp. 39–46.

Zhang, Y., Mukherjee, R., & Soetarman, B. (2013). Concept extraction and e–commerce applications. *Electronic Commerce Research and Applications, 12*(4), pp. 289–296.

AUTHORS

Anoud Shaikh

Ms. Anoud Shaikh is lecturer in the Department of Software Engineering at MUET, Pakistan. She received her M.E degree from MUET Pakistan in 2011 and is presently pursuing her PhD working on Text Analytics. Her research interests include Software Engineering, Databases and Data Analytics.

Naeem Ahmed Mahoto

Dr. Naeem Ahmed Mahoto is an Associate Professor and Chairman of the Department of Software Engineering, MUET Pakistan. He received his Master degree in Computer Engineering from MUET, Pakistan and Ph.D in Information Engineering from Politecnico di Torino, Italy, in 2013. His research interests are focused in the field of data mining and bioinformatics. His research activities are also devoted to summarization of web documents, sentiment analysis, data visualization and data mining.

Mukhtiar Ali Unar

Prof. Dr. Mukhtiar Ali Unar is the Dean Faculty of Electrical, Electronics and Computer Systems Engineering and a meritorious Professor at the Department of Computer Systems Engineering, MUET, Pakistan. He did his B.E in Electronic Engineering from MUET in 1986, M.Sc in Electrical and Electronic Engineering in 1995 and Ph.D in Artificial Intelligence from University of Glasgow, UK in 1999. He also remained the pro vice chancellor of MUET, S.Z.A.Bhutto campus, Khairpur Mir's and Director Institute of Information & Communication Technologies MUET, Pakistan. He has 30 years of teaching, research & management/admin experience. He is the author of more than 60 journal/conference papers of national/international repute.

His research interests include Artifical Intelligence, Control System Design, Digital Signal Processing and Knowledge Discovery. Dr. Unar is a member of IEEE (USA), an affiliate of International Federation of Automatic Control, a member of Pakistan Institute of Engineers and a member of Pakistan Engineering Council.

Edición Especial Special Issue Mayo 2019
DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.50-67