EXTRACTION AND ANALYSIS OF INFORMATION IN NEWS DOMAIN USING SEMANTIC WEB

Smriti Arora and Niyati Baliyan
Information Technology Department
Indira Gandhi Delhi Technical University for Women
Delhi, India
smriti0696@gmail.com
niyatibaliyan@igdtuw.ac.in

Abstract— News-papers, blogs, and web-pages are a rich and diverse source of textual information. However, the information contained in these sources cannot be manually extracted, recorded, and indexed, mainly because it comes in a massive size. Moreover, the extraction of some information sometimes requires specific knowledge or technical background. This is the case in the news domain where we need to extract the relevant news from a lot of available information. In order to scale knowledge extraction to the large size of available textual information, and build extractors specific to a certain field various techniques are applied over the unstructured data so that it can be made available to the users. This could help the researchers and the news readers or users to find relevant information in less time and with great ease. This study aims to review all the approaches and techniques done so far, for the information retrieval, search ability and its analysis and it also proposed an idea for better searching that reduces the time complexity to extract the data and also reduces human intervention. This is a better idea to put forward which also helps in filtering of irrelevant data and thus integrates only the relevant data to create a better space for the news data.

Keywords: information retrieval, semantics, natural language processing, Ontology, Semantic Web**.**

1. INTRODUCTION

The present news search engines are created in such a way that they give the news to the users based on their rankings which is based on the relevance of the news. Mostly it is done by ranking those news to the best which are searched mostly from the repository. Hence, the retrieval system searches the repository for the document containing words matching with the keywords present in the search query.

We can improve this system by grouping the news from different medias all together so that when the user searches for news on same topic he does not need to surf it from various sources. The user can get all the information on the same platform. This approach can be done using the NLP which involves the following steps:

a. **Tokenization** - Segregation of the text into its individual constituent words.

b. **Stop words** - Throw away any words that occur too frequently as its frequency of occurrence will not be useful in helping detecting relevant texts (as an aside also consider throwing away words that occur very infrequently).

c. **Stemming** - combine variants of words into a single parent word that still conveys the same meaning.

d. **Vectorization** - Converting text into vector format. One of the simplest is the famous bag-of-words approach, where you create a matrix (for each document or text in the corpus). In the simplest form, this matrix stores word frequencies (word counts) and is oft referred to as vectorization of the raw text.

After, we get the classified data we can create RDF graphs over the data and create our own ontologies using that data using the semantic web technology. Thus, the output we get using the SPARQL queries will be the desired output.

The **Semantic Web** is an extension of the World Wide Web through standards by the World Wide Web Consortium (W3C). The standards allow simple data formats and allow exchange protocols on the Web, most commonly the Resource Description Framework (RDF). According to the W3C, The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. The Semantic Web is thus considered as an integrator

978-1-7281-1253-4/19/$31.00 © 2019 IEEE

for various contents, information applications and systems. In this paper you will know about the work done so far in second section of the paper, and in the third section you will get to know about the new proposed methodology for a better and improved system to gather information from all the possible sources and present it to the user on a single platform. The fourth section gives the data set used for this system. In the fifth section we will get to know how the limitations of the work done so far can be overcome using the proposed methodologyand it also gives the future scope of the work done. The sixth section gives the conclusion of the entire work done.

2. LITERATURE REVIEW

The papers published so far in recent years have been published using various methodologies and technologies. The literature review of those papers is given below:

The research objective is analyzing methods for connecting twitter posts with similar news for contextualization [1]. The research methodology used is performance evaluation with coverage and precision. The technology used are the URL-based strategies, Content based Strategies. The limitations of this work are discovering effect of user-modeling based profile construction on social web personalization. The scope of this work is for the registered users.

The research objective is news personalization based on semantic web [2]. The research methodology used are new framework developed, count of concepts. The technology used is rules based on patterns and actions. The limitations of this work is that the events and patterns may be increased in rule base. Also it has scope for the users keen to gather information.

The research objective Semantic Web Information Extraction (IE)[3].The research methodology used is combining IE based on the mature text engineering platform (GATE) with Semantic Web-compliant knowledge representation and management. The research technology used is KIM, GATE. The limitation is that this work does not provide enough evidence regarding the approach, technology, and resources being used. Also it has scope for the users keen to gather information.

The research objective A framework is proposed so as to minimize the amount of redundant intermediate results in SPARQL query processing [4]. The research methodology used is RDF Triple Filtering (R3F) technique that utilize the graph-structural information of RDF data. A path-based index called the RDF Path index (RP-index) is used to efficiently provide filter data for the triple filtering. The research technology used is SPARQL. The limitations of the work is extension of R3F to create the graph features of RDF data and to explore the application of R3F in parallel and distributed environments, such as Map Reduce. Also it has scope for the registered users.

The research objective developing an automatic metadata creation system using the information extraction technology for the Semantic Web [5]. The research methodology used is the components of an NLP software Architecture, GATE, as the processing engine and support all required language resources for the engine. The technology used are RDF, NLP. The limitations of the work done is wider coverage of extraction. Also it has scope for the users keen to gather information.

The research objective analyzing methods for linking “Twitter posts with related news articles in order to contextualize Twitter activities” [6].The research methodology used is data corpus crawled from Twitter and BBC, CNN and New York Times, performance evaluation with coverage and precision. The research technology used is URL-based strategies, Content based Strategies. The limitation of the work is the deepening the searching of how the profiles created by this type of user modeling techniques affect personalization on the Social Web. Also it has scope for the registered users.

The research objective is deploying Semantic Web technologies to the Spanish news agency EFE [7].The research methodology used are Semantic Web standards so as to create ontologies for the news industry. The technology used are HDDB, NITF. The limitations are scalability and smooth deployment. Also it has scope for the users keen to gather information.

The research objective is finding structured information from unstructured or semi-structured text [8].The research methodology is search based on properties, keywords, Ontology search. The research technology are OWL/RDF, SWRL .The limitations of the work are improving the efficiency of IE system to improve the precision. Also it has scope for the users keen to gather information.

The research objective is data model for capturing event types [9].The research methodology used is

978-1-7281-1253-4/19/$31.00 © 2019 IEEE

incorporation of CEVO ontology for constructing the background data model and capturing fine grained event types. The research technology used are an ontology for Linking Open Descriptions of Events (LODE) and Comprehensive Event Ontology (CEVO).The limitation is that there is a lack of a holistic view on event extraction from free text and subsequently developing a knowledge graph from it. Also it has scope for the users keen to gather information.

The research objective extracting important information from various sources in contrast to a list of topic 0073 [10] .The research methodology used is NLP, clustering and the research technology used is RDF triplets. The limitations are expand the technique to multimedia data, design more explained demonstration and compare the outputs with all other relatable extraction systems. Also it has scope for the registered users.

The research objective Information Extraction for Semantic Web [11]. The research methodology used is web content mining to create the ontology. The technology used are OWL, RDF. The limitations are wider coverage of extraction Also it has scope for the registered users

The research objective is finding structured information from unstructured or semi-structured text [12]. The research methodology used are search based on properties, keywords, Ontology search. The technology used is OWL/RDF SWRL. The future work that can be done in this work is improving the efficiency of IE system to improve the precision. The scope of this work is for the registered users.

The research objective is e-news system based on NLP.[13] The research methodology used are bag of words, extraction of information,domain clustering. The technology used is the NLP technique. The limitations of this work is the accuracy that may be improved be using suitable number of classes during classification. Also it has scope for the users keen to gather information.

The research objective is use of DBpedia by BBC and Linked Data so as to build connections [14].The research methodology used is integrating the data and connecting files across BBC domains. The technology used is Semantic web Technology. The future work of this study is the improvement of the system by taking in large data sets. It has scope for the BBC and their users.

The research objective an approach for knowledge discovery [15]. The research methodology used is a model is used for saving the data from web in an ordered and well-formed manner in RDF format. The technology used are RDF, SPARQL. The limitations of the work done so far is the use of machine learning algorithms for categorizing the files based on web and their data. Also it has scope for the users keen to gather information.

3. PROPOSED METHODOLOGY

This paper proposes a methodology in order to provide the relevant news to the user on a particular topic he wants. For this in the first step after the user enters his query, the news from different sources is given as aninput in the system as an unstructured data. After the input is given the next step is tokenization or token formation as per the Fig 1 to extract strings from the input.

In the next step we need to remove the stop words so as to reduce the size of the file we are working in and also to remove the irrelevant words. The next stage is to form the clusters of similar words or the related words in reference to their semantics, also known as Stemming. Vectorization is the next step in which the clusters formed in the previous step is used to count the frequency of the occurrence of a particular word in the news data which is also called as bag-of-words approach. After vectorization the data is classified using any classification algorithm and then the RDF graphs are drawn for the data over which ontologies can be applied to get the desired result.

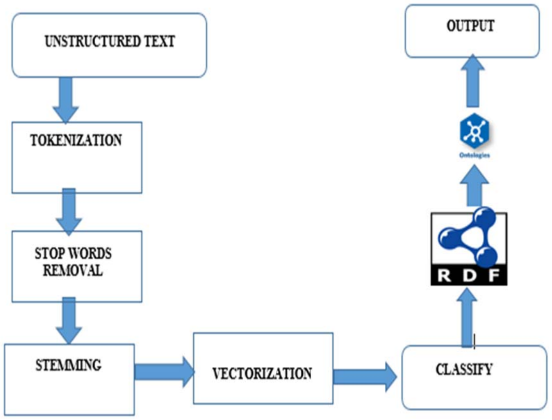

Fig 1: Proposed methodology

This methodology also includes the semantic analysis of the information we extract so that only positive and true news will be delivered to the user, and the fake news is discarded and not delivered to the user. This methodology is a very great way to present the data on a single platform and also only genuine information will be given. This method is different from the conventional methods because we are using the semantic web techniques. The results of this methodology will be better than that of the conventional methods as the semantic web techniques are better than those which were used earlier , as the ontologies leverage the technique used.

4. DATA SET

This dataset is a compilation of **2.7 million** news headlines published by Times of India from **2001 to 2017**, 17 years.

A majority of the data is focusing on Indian local news including national, city level and entertainment.

This News Dataset is a persistent historical archive of noteable events in the Indian subcontinent from start-2001 to end-2017, recorded in ReaL time by the journalists of India.

Content:

1. publish_date: Date of the article being published online in yyyyMMdd format
2. headline_text: Text of the Headline in English, very rare non-ascii characters

Start Date: 2001-01-01 End Date: 2017-12-31

Times Group as a news agency, reaches out a very wide audience across Asia and drawfs every other agency in the quantity of English Articles published per day. Due to the heavy daily volume (avg. 650 articles) over multiple years, this data offers a deep insight into Indian society, its priorities, events, issues and talking points and how they have unfolded over time.

It is possible to chop this dataset into a smaller piece for a more focused analysis, based on one or more facets.

- Time Range: Records during 2014 election, 2006 Mumbai Bombings
- One or more Categories: like Mumbai, Movie Releases, ICC updates, Magazine, Middle East
- One or more Keywords: like crime or ecology related words; names of political parties, celebrities, corporations

Table 1: Dataset Preview

publish_date (yyyy-mm-dd)	headline_text
20010101	win over cena satisfying but defeating undertaker bigger roman reigns
20010102	Raju Chacha
20010102	Status quo will not be disturbed at Ayodhya; says Vajpayee
20010102	Fissures in Hurriyat over Pak visit
20010102	America's unwanted heading for India?
20010102	For bigwigs; it is destination Goa
20010102	Extra buses to clear tourist traffic
20010102	Dilute the power of transfers; says Riberio
20010102	Focus shifts to teaching of Hindi
20010102	IT will become compulsory in schools
20010102	Parivar dismisses PM's warning
20010102	India; Pak exchange lists of N-plants

5. LIMITATIONS AND FUTURE WORK

In the work done so far some of the limitations that were noticed were:

- Improvement in efficiency of IE (information extraction), this limitation can be overcome in the methodology proposed as the system is using a better way to extract information from the unstructured data.
- Deepening the investigations of the information for better and efficient results,this limitation can be overcome in the methodology proposed as it deepens the extraction using the tokenization process.
- Improvement in scalability and smooth deployment,this limitation can be overcome in the methodology proposed as the system is a collection of distinct features all together in a single platform.

978-1-7281-1253-4/19/$31.00 © 2019 IEEE

- The coverage of extraction to be increased for efficient results, this can be done using this system as this system focuses on the extraction of words and then reducing the words of less importance.
- Semantically if the content is enriched it increases the ability of the audience to discover, navigate and share the content and the information, using this methodology we can also check the data using semantic analysis of the content.
- Easily correlate topics across documents, pages, blog posts, etc. to offer the most relevant content to users, this could be achieved by the extraction of data only if it is semantically approved.

Images, tables and the grids in the data are not considered for generating set of information as of now in this paper. It primarily focuses on the textual content of the news, and that is why it could be extended in the future work. This paper also do not take into account the automatic finding of the data from sources available to provide as input. This idea behaves perfectly for users struggling to gather information from various sources with inclusion of their extra time and effort to easily get all relevant information related to the specified topic at one place in one blog.

6. CONCLUSION

This paper focuses on the need of gathering news from different result sets of the query and integrating the relevant result obtained at one single place which reduces the human efforts and also the time that is spent in finding the relevant news. This paper also proposes a model or the steps to achieve the desired target. The aim is to provide the user with relevant news he wants from all the sources at just one place. This can be done by initially forming the tokens from the data and then removing the stop words which make it easier to work with large amount of data and also helps to get rid of the undesired or irrelevant data. Later the words are grouped together so as to find the frequency of their occurrence and their relevance too. Further they are classified and drawn into RDF graphs. Finally the data can be retrieved using various queries over the ontologies created.

REFERENCES

[1]Sakaki, T., Okazaki, M., Matsuo, Y.: "Earthquake shakes Twitter users: real-time event detection by social sensors",2010.

[2] Anantharangachar. Raghu & Srinivasan. Ramani, (2012) "Semantic Web techniques for yellow page service providers", *International Journal of Web & Semantic Technology (IJWesT)* Vol.3, No.3.

[3] O. Vikas, A. K. Meshram, G. Meena, and A. Gupta, "Multiple document summarization using principal component analysis incorporating semantic vector space model," *Computational Linguistics and Chinese Language Processing*, vol. 13, no. 2, pp. 141–156, 2008

[4]Kim, K., Moon, B., & Kim, H.-J. (2013). "R3F: RDF triple filtering method for efficient SPARQL queries processing."*World Wide Web, Springer,* 2013.

[5] Jayatilaka A.D.S, Wimalarathne G.D.S.P, "Knowledge Extraction for Semantic Web Using Web Mining*". The International Conference on Advances in ICT for Emerging Regions –ICTer*2011

[6] "Implicit Entity Linking in Tweets, Heraklion, Crete, Greece," 06/2016 2016. Springer.

[7] Arndt.Richard, RaphaëlTroncy, Steffen Staab, Lynda Hardman, and MiroslavVacura. "COMM: Designing a Well-Founded Multimedia Ontology for the Web." *Springer Berlin Heidelberg, 2007.*

[8] Gerber.Daniel, Hellmann.Sebastian, Bühmann.Lorenz, TommasoSoru, Usbeck.Ricardo, and Axel CyrilleNgongaNgomo. "Real-time RDF extraction from unstructured data streams*." In The Semantic Web - ISWC 2013 - 12th International Semantic WebConference, Sydney, NSW, Australia, October 21-25, 2013,* Proceedings, Part I, pages 135–150, 2013.

[9]Cheng Li, Bendersky.Michael, Garg.Vijay, and Sujith Ravi. "Related event discovery*". In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017.*

[10]Sören Auer, Jens Lehmann, and Sebastian Hellmann. "LinkedGeoData - adding a spatial dimension to the web of data". *In Proc. of 8th International Semantic Web Conference (ISWC),* 2009.

[11] Boicea, A., Radulescu, F. and Agapin, L.I. (2012) "MongoDB vs Oracle-database

978-1-7281-1253-4/19/$31.00 © 2019 IEEE

Comparison", *in EIDWT*, pp.330–335.

[12] Chen, C-L., Tseng, F.S.C. and Liang, T. (2010) "Mining fuzzy frequent item sets for hierarchical document clustering, Information Processing & Management,"Vol. 46, No. 2, pp.193–211.

[13] Kamath, S.S. and Kanakaraj, M. (2015) 'Natural language processing-based e-news recommender system using information extraction and domain clustering', Int. J. Image Mining, Vol. 1, No. 1, pp.111–125.

[14] Dakka,W., Cucerzan, S.: Augmenting Wikipedia with Named Entity Tags. In: Proceedings of the 3rd International Joint Conference on Natural Language Processing (2008).

[15] Hassanzadeh, O., et al.: A Declarative Framework for Semantic Link Discovery over Relational Data. Poster at 18th World Wide Web Conference (2009).

978-1-7281-1253-4/19/$31.00 © 2019 IEEE