IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Personalized Visual News Extraction and Archival Framework

Shine. K. George [1*], Jagathy Raj V. P [2]

[1] Department of Computer Applications, Cochin University of Science and Technology, Kochi, India
[2] School of Management Studies, Cochin University of Science and Technology, Kochi, India

[*]*Corresponding Author:* shineucc@gmail.com, *Tel.: 9447189662*

Available online at: www.ijcseonline.org

Accepted: 15/Jun/2018, Published: 30/Jun/2018

***Abstract*—** Enormous news contents are getting generated today which keeps on growing to a great extent. The Archiving of news items has become a tedious and challenging task because of its rich quantity. It also creates problems for journalists to find proper content using current search tools. Personalized news extraction helps the journalist to find the right news content without browsing through irrelevant news items. Semantic Web techniques improve personalizing the news content for information extraction. The proposed Ontology-based news extraction framework provides a high degree of semantically similar news contents for a search query. That helps the journalist to develop a news story within a short span of time. The Proposed framework is evaluated using YouTube – 8M dataset and results are positive. The lack of local knowledge concepts incorporated with the underlying ontology used in this framework can be addressed in the future enhancement. Further researches are needed to include the priority listing of news contents extracted for the same search query on geographical and news value specific.

***Keywords*—** Information Extraction, Knowledge Management, Ontology, Personalization, Semantic web techniques

I. INTRODUCTION

Nowadays visual news media is becoming an inevitable means for a large group of people for daily news updations. Visual news story or report is an edited sequence of semantically related visual scenes and relevant bite based on a voice description (news script). The visual scenes and bite are ordered as per the news script. TV news viewers always look forward to the news story which is content wise rich. News stories are built from a combination of related contents. The highlight of a news story is that it is capable to deliver relevant and recent updates on a particular topic as well as the past history also. Most of the archiving system in a typical news channel library is manual. Some of the news channel use computers to store rushes, bites, previously aired news etc, of various events with a short textual description [1].

The news channels are operating in 24 X 7 modes and the volume of news created in the world in a day is huge. The cataloguing, describing and ordering news and storing in a library is a challenging task. Hence this domain shares the problems and characteristics of World Wide Web and the importance of semantic web technologies can provide a solution to the news archiving issues [2]

In this paper we propose an ontology based framework for news extraction. The Ontology can effectively analyze and maintain the semantic relations of the words in a text. This research uses Open Calais, which is known to be an ontology that is news specific which is developed by Thomson Reuters. To evaluate the proposed framework YouTube-8M dataset [8] is used. The organization of this paper is as follows. Section I contains the introduction of news extraction, Section II contain the related work of news extraction framework, Section III contain the proposed framework process, Section IV contain experimental results and section V includes conclusion and the future scope.

II. RELATED WORK

When a detailed news story on a topic is assigned to the journalist, his/her natural choice would be channel library. As we know the news now is insignificant in the next minute, the time allotted to create a news story is very short for a journalist. Rather than searching through irrelevant news items, the journalist would love to get the information he needed from the first query. He/she is expecting a more personalized information extraction system. But conventional news archiving system does not provide this experience. Considering the massive size of news that is stored on everyday basis and lack of time to create the latest news update on a particular topic, semantic-based news extraction capability is essential [3]

In order to make news extraction in the channel library become effective, it is necessary to make machine aware of the underlying semantics of the search queries given by the

journalist. Ontologies are the powerful tool for knowledge representation and ontology-driven adaptation can give a significant contribution in the area of personalization. It has great potential in the information extraction frameworks [4, 5, 6, 7] and it can be used in visual news archiving framework to understand the semantics of the news stored. The Ontology is a knowledge base that even follows the rules of inference [12]. An ontology-driven framework can be used to solve the pitfalls of keyword-based news extraction and to provide more personalization. The ontology-based approach can easily locate the data and can also preserve the semantic structure in the results [11].

III. METHODOLOGY

The main purpose of this framework is for archiving and extracting news in a news channel library. The proposed framework consists of two major two functions. First is how conceptual terms are generated from news and stored along with the news using ontology. The second important process is how the news has been retrieved on the semantically enriched query using ontology. Fig. 1 represents the proposed framework and its process.

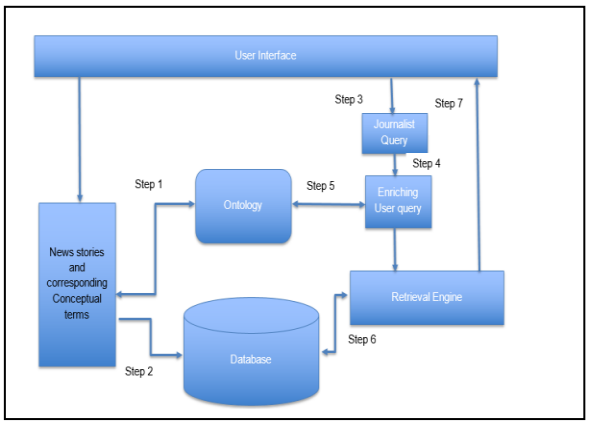

Figure 1. Ontology-Based News Extraction Framework

Framework activities start with storing news content along with corresponding conceptual terms derived from the ontology. The sequential steps involved in the news extraction process is described below

1) Journalists or news reporter enter the queries via User Interface (UI)
2) Tokenization of the search query is done. That is, splitting the words contained in the text
3) Eliminate the stop words like 'and', 'about', 'the 'etc. These words are not significant for searching process.
4) The Query is expanded by extracting equivalent terms from an ontology and run the query to the database
5) Concept terms in the query and corresponding keyword terms of each news content accumulated in the database is matched
6) News contents are displayed based on concept term matching with the search query.

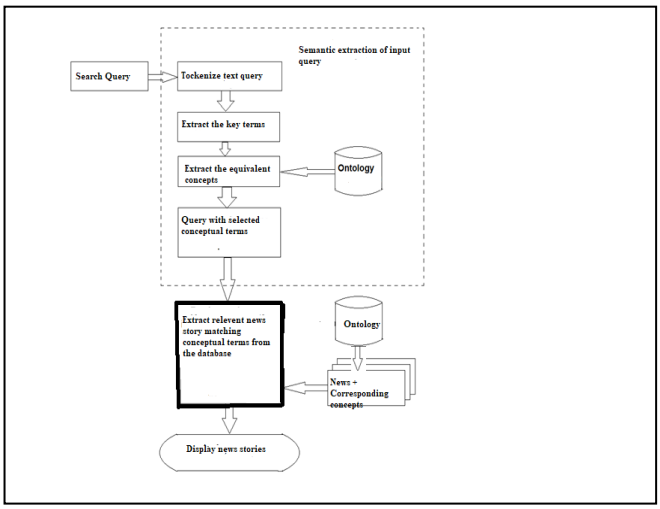

Figure 2. News Extraction process

The application to showcase the Proof of concept has been built in Drupal, an effective content management system is considering the availability of easily pluggable contributed modules. For the search efficiency apache solr, which is an efficient mechanism to index and retrieve contents. The ontology is used in this implementation is Open Calais and the dataset used is YouTube – 8M dataset

Basic Open Calais ontology is easily accessible and has implemented as a module in Drupal. It extracts the key concepts and returns tags considering Entities, Events, Topics, Relations, Social tags, etc. in the news content. One of the standardization frameworks particular to the news industry is formed by the well known International Press Telecommunications Council (IPTC), an international consortium of news agencies, the board of editors and newspapers suppliers and distributors. IPTC provides a formal subject categorization of news and events. Thomson Reuters launched this ontology in 2008 and it follows IPTC standards.

The YouTube-8M dataset consists of millions of YouTube video IDs. There are several categories of videos in this dataset. We considered categories like News broadcasting, CBS news, ABS-CBN news and current affairs, newscaster etc. From this category, we took videos only with English annotations. Sufficient text descriptions of news are required to generate concept terms using Open Calais ontology.

Measuring semantic similarity is an important part in assessing the accuracy of information retrieval frameworks; one of the popular semantic similarity measures is latent semantic analysis (LSA) [9, 10] which is best suitable to identify the relationships between the document sets. LSA analyzes the relationship between text contents and its terms. It is a good technique to capture text semantics. Drupal has semantic similarity module which uses LSA algorithm. In this study, we use this module to identify and measure the semantic similarity which lies between the search query and the resultant extracted news stories.

IV. RESULTS AND DISCUSSION

We prepared a total of 9 queries as displayed in Table 1. For each query, we created a query code for representation purpose and key terms generated using open Calais ontology.

Table 1. Query Code Representation

Query Code	*Query*	*Ontology Terms*
Q1	Obama and Iran	Obama ,Iran
Q2	Obama about Israel nation	Obama,Israel,nation
Q3	Obama and Iraq	Obama,Iraq
Q4	Syria war and Obama	Syria,Obama,war
Q5	Bush and Obama	Bush,Obama
Q6	Obama and Cheney	Obama,Cheney
Q7	Romney and Obama	Romney,Obama
Q8	President election and Obama	President,Obama,election
Q9	Senator Barack Obama	Senator,Obama,Barack Obama

Table 2 presents the average semantic similarity of extracted news content in percentage of two different extraction techniques with the search query. We tried Normal Keyword Search extraction and Ontology-Based Search for each query code. Clearly, it shows that ontology based news extraction is giving a better performance.

Table 2. Average Semantic Similarity

Semantic Similarity (In Percentage)		
Query Code	*Ontology Based Search*	*Normal Keyword Search*
Q1	89.73	57.11
Q2	83.53	59.43
Q3	87.12	39.13
Q4	88.56	41.23
Q5	84.32	67.68
Q6	93.28	53.12
Q7	91.77	49.71
Q8	90.21	47.71
Q9	94.19	36.33

Figure 3 gives a bar chart representation of the evaluation result. It clearly shows the difference in the two approaches. It is clear that the proposed framework works out well and produces efficient results than traditional normal search. The Ontology based news extraction produces best results as shown below

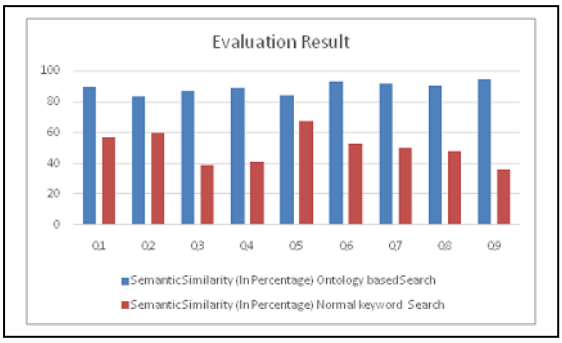

Figure 3. Evaluation Result

V. CONCLUSION AND FUTURE SCOPE

In this work, we presented a personalized framework for news extraction and archiving. Semantic Web techniques like ontologies are used to provide better personalization. The journalist can extract highly semantically related news contents such as news story, bite, news footages from a single query in the proposed system. So he/she can build a news story within a short time. From the evaluation result, it is clear that ontology-based proposed framework provides better extraction result than the traditional approach.

Further studies can take place in this area. Localization is a challenging issue in this domain. Nowadays local news is also getting importance in news broadcasting around the world. There will be the tremendous amount of local knowledge or local concept terms are referred to local news. The currently available news ontologies do not understand these local concepts and that may reduce the effectiveness of the proposed system.

Nowadays centralized news archiving systems are becoming popular since it reduces a lot of manpower and cost of infrastructure. For example, a single news desk may handle a plenty of news bulletins for geographically different locations and the priority of news is geographic location

specific. Proposed framework does not focus on the news priority of the extracted news content for a search query.

REFERENCES

[1] Shine K George, V P Jagathy Raj, G Santhosh Kumar, "Ontology based framework for news extraction in visual media", In:Proceedings of IEEE international conference on Data Science & Engineering (ICDSE) ,pp. 220-222, 2012

[2] Berners-Lee, T. Hendler, J. Lassila, "The Semantic Web.Scientific American, pp. 29-37, 2001

[3] Daya.C.Wimalasuriya, Dejing Dou,"Ontology based information extraction: an introduction and a survey of current approaches", Journal of Information Science, Vol. 36, No. 3. pp.306-323,2010

[4] F. Wu and D. S. Weld, "Autonomously semantifying wikipedia". In:Proceedings of the Sixteenth ACM Conference on Information andKnowledge Management, pp. 41-50, 2007

[5] J. Kietz, A. Maedche, and R.Volz, "A method for semi-automaticontology acquisition from a corporate intranet", In: Proceedings ofthe EKAW'00 Workshop on Ontologies and Text, 2000

[6] P.Cimiano, S. Handschuh, and S. Staab, "Towards the self annotating web", In: Proceedings of the 13th InternationalConference on World Wide Web, 2004.

[7] D. Maynard, W. Peters, and Y. Li, "Metrics for evaluation ofontology-based information extraction",. In: Proceedings of the WWW 2006 Workshop on Evaluation of Ontologies for the Web,(ACM, New York, 2006)

[8] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici,B. Varadarajan, and S. Vijayanarasimhan. YouTube-8M: Alarge-scale video classification benchmark. arXiv preprint,arXiv:1609.08675, 2016

[9] Anna Rozeva1, Silvia Zerkova1 "Assessing Semantic Similarity of Texts – Methods andAlgorithms", Proceedings of the 43rd International Conference Applications of Mathematics in Engineering and EconomicsAIP Conf. Proc. 1910, 060012 ,2017

[10] Chelsea Boling, "Semantic Similarity of Documents Using Latent Semantic Analysis", Proceedings of the National Conference On Undergraduate Research (NCUR) 2014 University of Kentucky, Lexington, KY April 3-5, 2014

[11] Apurva Dube, Pradnya Gotmare, "Semantics Based Document Clustering", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.4, pp.25-30, August 2017

[12] M.Chahbar, A.Elhore, Y.Askane, "PERO2: Machine Teaching based on a Normalized Ontological Knowledge Base", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.5, pp.63-74, October 2017

Authors Profile

Shine K George is currently pursuing Ph.D in Computer applications, at Cochin University of Science and Technology, Kochi, Kerala, India. He is also working as Associate Professor in Department of Computational applications, Union Christian College, Aluva, Kerala, India. His main research work focuses on Ontology, Knowledge management, machine learning and deep learning. He has 14 years of teaching experience and 7 years of Research Experience.

Dr. Jagathy Raj V. P., currently working as Professor in Operations and Systems Management at School of Management Studies, Cochin University of Science and Technology, Kochi, is a Ph.D holder in Industrial Engineering and Management from Indian Institute of Technology (IIT), Kharagpur.Dr. Jagathy Raj did his B.Tech in Electrical Engineering from University of Kerala and M.Tech (Electronics with Communication as specialization) and MBA (Systems and Operations Management) from Cochin University of Science and Technology. He is equally well-versed in both Engineering and Management related areas have sufficient experience in both the fields. He has more than 28 years of teaching and 22 years of research experience in Engineering and Management related areas. He is also guiding research scholars both in Engineering and Management related fields. He has produced 10 Ph.Ds through research in the above areas. He has more than 185 research publications in National and International Journals and Conference Proceedings to his credit in Engineering and Management related areas with good citations. He has also presented number of research papers in National and International conferences. He has also delivered number of invited talks. He is acknowledged to be expert in computer simulation and modeling and has undertaken several consulting assignments for leading business houses. At present, he is serving as consultant to several Government and Non-Governmental organizations.

