Ontology based Framework for News Extraction in Visual Media

Shine K George
Department of Computer Applications
Union Christian College
Alwaye ,India
shineucc@gmail.com

Jagathy Raj V P
School of Management Studies
Cochin University of Science and Technology
Kalamassery, India
jagathy@cusat.ac.in

G Santhosh Kumar
Department of Computer Science
Cochin University of Science and Technology
Kalamassery, India
san@cusat.ac.in

Abstract—Anticipating the increase in video information in future, archiving of news is an important activity in the visual media industry. When the volume of archives increases, it will be difficult for journalists to find the appropriate content using current search tools. This paper provides the details of the study we conducted about the news extraction systems used in different news channels in Kerala. Semantic web technologies can be used effectively since news archiving share many of the characteristics and problems of WWW. Since visual news archives of different media resources follow different metadata standards, interoperability between the resources is also an issue. World Wide Web Consortium has proposed a draft for an ontology framework for media resource which addresses the intercompatiblity issues. In this paper, the w3c proposed framework and its drawbacks is also discussed.

I. INTRODUCTION

Information technologies used in the news industry has given a new dimension in journalistic activity. Information technology provides an ease to access, promptly and economical way to deliver information. The lifecycle of visual media news industry begins with journalists creating the news and ending with the broadcasting of the news. Journalists often search for information which is related to their latest news visual stories in the archival system available in the visual media. This search will be performed in tense situations, e.g., lack of time, lack of knowledge in relation to the archive system and standards used to store the information .The inability of the archival system creates a huge difficulty in the search given by journalists and makes the search more unsuccessful. The news archival process is normally done by the documentation department of the news media. The way of storing the information will be decided by the librarian in the documentation department by considering the possibilities of the software platform. The emerging Semantic Web technologies [1] provide a good approach to overcome the limitations of existing media extraction systems. The size and complexity of the stored news content, the time limitations for cataloguing, describing and ordering the incoming information, make news archives difficult to manage. In this sense, they share many of the characteristics and problems of the WWW.Therefore the solutions proposed in the Semantic Web vision are applicable here.

For example a journalist may wish to do a news story on the possibilities of monsoon tourism in Kerala. Suppose he/she gives queries like this 'return all the scenes in which rain is visualized as part of Monsoon tourism in Kerala'. Traditional search will not give an appropriate result for the above query and also provides irrelevant information.A semantic based extraction capability [2] is highly required in visual news media extraction system since the information stored from every day news is huge in volume and high domain knowledge is also required. It is generally agreed that Ontology-based information extraction has a lot of potential [3, 4, 5, 6].Ontology-based information extraction provides an automatic mechanism to generate semantic contents by converting the information into ontologies.

The rest of this paper is organized as follows. Section 2 presents the study we conducted about the media extraction systems of news channels in Kerala .Section 3 discusses interoperability issues among different media resources. Section 4 presents W3C approved draft of ontology for media resources and the proposed candidate recommendation of an API based on the proposed ontology. Section 5 presents the drawbacks of the proposed ontology. Section 6 summarizes the study we conducted and the future directions in this field.

II. EXISTING NEWS EXTRACTION SYSTEMS USED IN THE NEWS CHANNELS IN KERALA

The media extraction systems of popular news channels in Kerala were studied. Table 1 summarizes the News extraction systems used in news channels in Kerala.

978-1-4673-2149-5/12/$31.00 ©2012 IEEE

TABLE I. SUMMARIZES THE NEWS EXTRACTION SYSTEMS USED IN NEWS CHANNELS IN KERALA COVERED IN THIS STUDY

News Channel	Application Type and year	Main Categories	Sub Categories	Search Method	Advanced Filtering	Storage Device
Indiavision	Standalone (2006)	State National International	Nil	Keyword Based	No	Tape
Asianet	Standalone (1997)	State National International	Sports, events, Death etc... (around 30 sub categories)	Keyword Based	Yes	Tape
Manorama	Server Based (2006)	News (state) National Sports International	Archives Rushes Bite Event	Keyword Based	Yes	Server Computer
Jaihind	Standalone (2007)	State National International	Feed Sports Reuters	Keyword Based	No	Tape
People	Standalone (2004)	State National International	Nil	Keyword Based	No	Tape
Amrita	Standalone (2005)	State National International	Nil	Keyword Based	No	Tape

Among six news channels except Manorama everyone are using the same standalone application developed in Visual Basic which is keyword based. It was found that all the existing extraction systems are failed to take the advantage of the possibilities offered by the technologies .Aspects that can be improved include searching techniques, News Category list by considering the domain knowledge and lack of a commonly adopted standard representation for news archive etc.The drawbacks of the existing extraction systems of the news channels in Kerala can be overcome by developing an application which consists of a commonly adopted news media ontology.

III. INTEROPERABILITY ISSUES AMONG DIFFERENT MEDIA RESOURCES

Anticipating the increase in video information in future, it will become more difficult for journalists to find the appropriate content using current search tools. A wide range of media metadata formats are available for visual news author's to express their information. Several metadata solutions for media related content are also used in the news industry. The collections of visual archives in different news channels across the world are increasingly digitized. These archives are stored by often using domain specific or even proprietary metadata models. This creates difficulties in accessing these collections in a homogeneous or centralized way and linking them across collections. This is one of the key problems faced by recommendation service providers also. A common ontology spanning different metadata sets can allow recommendation systems to return a better relevant selection than when the metadata systems are unrelated. The interoperability between media resources are required to solve above mentioned issues.

IV. WORLD WIDE WEB CONSORTIUM APPROVED DRAFT OF ONTOLOGY AND API FOR MEDIA RESOURCE

World Wide Web Consortium (w3c) [7] has approved a draft for an ontology for media resource. The ontology which is proposed by W3c addresses the intercompatiblity issues between different media standards by providing a common set of properties to define the basic metadata needed for media resources and the semantic links between their values in different existing vocabularies. It will help to manage different types of current video metadata formats by providing full or partial translation and mapping from properties in respective formats to a common set of properties in the proposed ontology. W3c is also proposing an API [7] for media resource that provides uniform access to all elements defined by the ontology.ie.The ontology will define mappings from properties in respective formats to a common set of properties. Then the API will define methods to access heterogeneous metadata, using such mappings.Fig.1 represents w3c approved ontology framework for media resources.

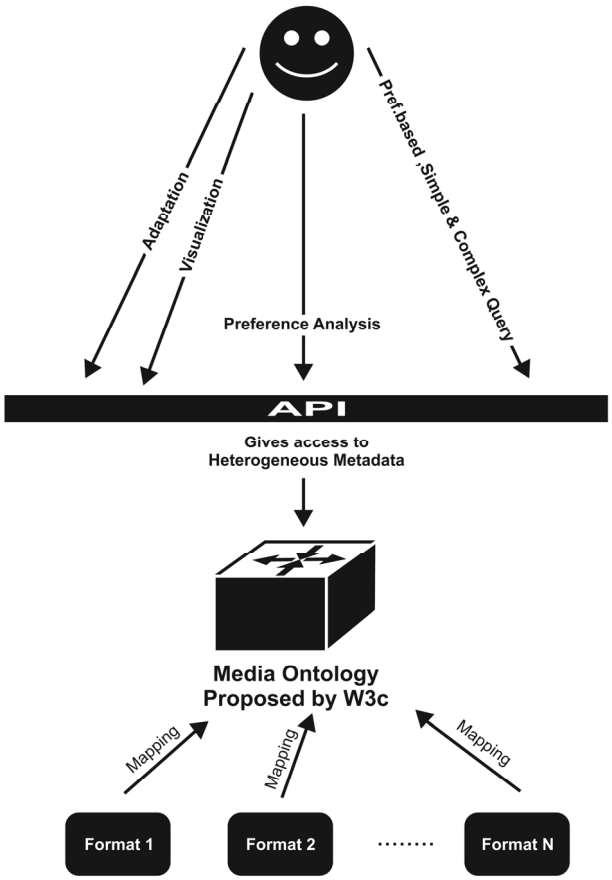

Figure 1. w3c approved ontology framework for media resource

An important aspect of the above mentioned framework is that everything visualized above the API is left to user defined applications. For example languages for simple or complex queries, analysis of user preferences and other methods for accessing metadata etc.The ontology and the API provide a basic interoperability for applications.

V. DRAWBACKS OF W3C PROPOSED ONTOLOGY FRAMEWORK FOR MEDIA ACCESS

Access to user-defined metadata to media resources is not supported by the W3c proposed ontology frame work. User defined metadata means metadata that is not defined in a standardized format and is being created by the user. This is highly relevant when we consider the geographical scope of different visual media news. For example if we want to get visuals for a news on Monsoon tourism in Kerala ,the metadata information regarding monsoon in Kerala need to be more detailed in the proposed ontology. Since this information is regional wise specific, it can't be standardized.

If the proposed ontology has support for user defined metadata, the above situation can be managed by incorporating the details of monsoon in Kerala as user defined. The mappings defined in the proposed ontology to preserve the semantics of a metadata item across different metadata formats cannot be easily achieved. Mapping between the Ontology's property and the elements from two different formats that have such a difference will not allow a semantic-preserving mapping. The property dc: creator from the Dubling Core [8] and the property exif: Artist defined in the EXIF [9] is both mapped to the property: creator, in the proposed media Ontology framework. So there is every possibility of having a certain loss in semantics. Mechanisms for correction for this loss are not incorporated in the proposed frame work.

VI. CONCLUSION

In this paper we have reviewed news extraction systems of different news channels in Kerala. It is noted that archiving systems covered in this study do not follow a common standard for archiving news items. The introduction of semantic based technologies can improve the visual archiving system and its exploitation. The lack of a commonly adopted standard for archiving visual news items is a problem in the usage of different media resources. In this article we have discussed the World Wide Web consortium approved ontology based framework for media resources and its drawbacks. Further studies can take place in this area to incorporate the user defned metadata information without losing the semantics.

REFERENCES

[1] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American, 2001, p. 29-37

[2] 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. (1 June 2010), pp. 306-323

[3] F. Wu and D. S. Weld, "Autonomously semantifying wikipedia". In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, (ACM, New York, 2007), pp. 41-50

[4] J. Kietz, A. Maedche, and R. Volz, "A method for semi-automatic ontology acquisition from a corporate intranet", In: Proceedings of the EKAW'00 Workshop on Ontologies and Text, (Springer, Berlin, 2000)

[5] P. Cimiano, S. Handschuh, and S. Staab, "Towards the self-annotating web", In: Proceedings of the 13th International Conference on World Wide Web, (ACM, New York, 2004).

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

[7] W3C Consortium Reference
http://www.w3.org/standards/semanticweb/

[8] DCMI Metadata Terms. January 2008.
http://dublincore.org/documents/2008/01/14/dcmi-terms/

[9] EXIF version 2.3.. Standard 26, April 2010.
http://www.cipa.jp/english/hyoujunka/kikaku/pdf/DC-008-010_E.pdf