Ontology-Based Information Retrieval for Historical Documents

Fatihah Ramli[1], Shahrul Azman Noah[2], Tri Basuki Kurniawan[3]
[1]Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
[2, 3]Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia 43600 UKM Bangi, Selangor, Malaysia.
[1] rfatihah@unimas.my, [2] shahrul@ukm.edu.my, [3]tribasukikurniawan@yahoo.com

***Abstract*— This article presents an ontology-based approach to designing and developing new representation IR system instead of conventional keyword- based approach. Such representation improves the precision and recall of document retrieval. Experiments carried out on the ontology-based approach and keyword-based approach demonstrates the effectiveness of the proposed approach.**

Keywords-ontology; information retrieval; semantic retrieval;

I. INTRODUCTION

Information Retrieval (IR) research in various fields gives many new ideas for researchers to improve existing approaches in all fields. However, recently the field that gets special attention is history. As discussed by [1], the historians still expect a better approach for more accurate access to historical documents. For example, a recent study of the Australian National Library found that the numbers of visitors increased radically when they provided historical documents as searchable full text index [2, 3]. Hence, the IR for historical documents is an essential issue to be studied.

Historical document can be defined as those that keep information related with time instant at which the documents were published at the same time that are still useful in the future [4]. Searching and retrieving documents from large historical archive prove to be challenging for IR field as historians typically employ their knowledge, experience and intuition to decide which information they will need to find and study, and attempt to locate sources that contain the information [8]. Hence, Elena et al. [1] suggest that historians need historical source repositories and building tools that will enable them to access the comprehensive information in a rapid manner. Conventional IR approaches are mostly based on a simple Bag- of-Word (BOW) approach whereby terms-order are ignored and it conflates many texts that have very different semantic meanings into a single form. As a result, searching and ranking of historical documents based on the BOW approach is not suffice as the documents contain rich semantic information relating to important entities such as event, time, and people.

Therefore in this paper we proposed an ontology-based approach to index and ranked [5] semantically rich historical documents. The ontology developed centralised on the event related elements which are important to the historical domain. Ontology-based approach to document retrieval is not new as demonstrated in the work in [16- 18]. However, the applications of such an approach to historical documents are still scarce and are still open for further research and development. Apart from the ontology-based approach, we also proposed a simple ontology- based weighting mechanism mainly derived from the classic tf-idf scoring scheme. We evaluated our proposed approach against the BM-25 probabilistic model involving 133 documents.

The paper is organized as follows: an overview of the environment in which ontology has been used is presented. In section 2, describe in detail about related work of IR in historical document. Section 3, illustrates the overall process of semantic retrieval while section 4 discusses the results obtained from the evaluation of the approach. Finally, section 5 concludes.

II. RELATED WORK

Some applications of IR to historical documents mostly concern on spelling issues whereby users expect that modern keywords able to match with elements of words/spelling available in historical documents [3, 6, 7]. This is due to the fact that there are too much spelling variants located in large document of historical texts [3]. Full-text indexing of such documents is not suffice as modern words are used in users' queries unable to match with the index. Two popular approaches to solve the issues are by proposing special matching procedures and lexica for historical language.

Keywords matching procedures although are non- trivial, they still not fully representing the fundamental characteristic of historical documents. Historical document can be defined as those that keep information related with time instant at which the documents were published at the same time that are still useful in the future [4]. A response from Elena, Katifori [1], stated that historians employ their knowledge, experience and intuition to decide which information they will need to find and study and attempt to locate sources that contain the information. The result from Elena, Katifori [1] obviously stated that historians need historical source repositories and building tools that will enable historians to access the comprehensive information in a rapid manner. The 20th and

978-1-5090-2954-9/16/$31.00 ©2016 IEEE