International Journal of Applied and
Natural Sciences (IJANS)
ISSN (P): 2319-4014; ISSN (E): 2319-4022
Vol. 6, Issue 5, Aug – Sep 2017; 51-58
© IASET

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations
IASET

SEMANTIC RETRIEVAL OF HISTORICAL DOCUMENTS BASED ON IR APPROACH

POONAM YADAV

D. A. V College of Engineering & Technology, Kanina, Haryana, India

ABSTRACT

This paper proposes a new approach, namely Ontology-based approach for developing and designing a modern representation IR method over the existing keyword-based scheme. Moreover, these kinds of representation normally progresses the precision and the recall of manuscript retrieval as well. Experimentation is carried out by comparing both the proposed and conventional methods and the results demonstrate the proficiency of the proposed methodology.

KEYWORDS: Information Retrieval; Ontology; Semantic; Manuscript Retrieval, Keyword Based Approach

I INTRODUCTION

Information retrieval seems to be more familiar almost in all fields. This also grants more innovative ideas for the researches, especially for the enhancement of the conventional schemes in all areas. In the recent years, history is the area which is highly gets concentrated. However, the historians yet need an effective approach for exact accessing of historical manuscripts. From the recent study made by Australian National Library has found that, the visitors for searching the historical manuscripts are increasing gradually [2, 3]. Thus, the IR for historical manuscripts is a sensational concern that has to be studied.

Historical manuscripts are the documents, which are information based with times instant, that are more useful even in future [4]. Moreover, searching as well as retrieving the manuscript from the huge historical archive is a tedious task for IR field, since the historians purely employs their experience, knowledge, and the intuition for deciding which data is required for study and attempts the location which insist the information [8]. Thus, in [1], the authors have suggested that it is necessary for a building tools and historical source repositories, which enable the accessing of comprehensive data in a hasty way. The existing IR methods are commonly based on easy Bag-of Word (BOW) method in such a way that, terms-order are neglected and it asserts various texts which has varied semantic meaning into a sole custom. Thus, searching and ranking of historical manuscripts using BOW method is not an effective way, since the document comprises high semantic data with respect to significant objects like time, people and event.

Hence, in this paper, we introduced an ontology-based scheme for indexing and ranking [5] semantically abundant historical manuscripts. Even many ontology-based methods are demonstrated in [16], [17] and [18], still it is necessary for an innovative approach for historical manuscripts and the research on developing such methods will continue even in future. Apart from the proposed method, we also introduced a simple methodology named Ontology-based weighting method, which is derived from a scheme named classic tf-idf scoring method. The proposed method is evaluated over the BM-25 probabilistic model, which involves 133 manuscripts.

The rest of the paper is organised as follows: Section II gives the detailed description about the related works of IR system in historical manuscripts. Section III explains the overall process of the proposed methodology to the semantic historical manuscript retrieval. Further, section IV discusses the evaluation results of the proposed approach and Section V concludes the paper.

II. LITERATURE REVIEW

Many applications of IR to historical manuscripts are highly concentrated on the spelling issues and thereby, the users always expect that the new or modern keyword must have to match with the objects of spelling that are available in the manuscripts i.e. Historical manuscripts [3, 6, 7]. This is because of many spelling variants in huge documents of the historical texts [3]. However, full text indexing for these kinds of manuscripts are not enough, since the user uses the modern words in the queries, which are unable to be matched with the index. These mentioned issues are solved by introducing special matching ways and lexica for the historical language.

Even though the keyword matching approach is non-trivial, they are not completely expressing the major characteristics of the historical manuscripts. As explained before, the historical manuscripts are the documents which have data or information related with time instant in which, the manuscripts are published and meanwhile, they must be utilized in future too [4]. As per Elena, Katifori [1], the historians pay their experience, knowledge as well as intuition for deciding which data is needed to find and study and also attempts source location which contains the data. From the results of Elena, Katifori [1], it is specified that there is a need of additional building tools for the fast accessing of the comprehensive information. Further, there is a great change in information accessing of people from the 20th century.

Thus, the users expect a wealth of historical data that facilitates in sharing and reusing via the digital libraries, which grant the matched manuscripts for all search request and must provide the support for the same scenario [8] [9]. To fulfill these users request, Mirzaee, Iverson [8] and corda [9] have suggested the historical manuscripts semantics that attempts for allowing higher representation of its embedded facts, which must be captured than the text manipulation tools. The semantic use could be more efficient, if it simplifies via time-based relations. Further, Schockaert, Cock [10] have suggested that the manuscripts must be sorted in accordance with the temporal characteristic for the improvement of the IR system. Alonso, Gertz [11] have denoted, the recognizing along with the use of temporal data for the applications of IR was considered as the vital feature, which could improve the functionalities of the search applications.

The above-mentioned works are highly concentrated on the type of knowledge which must be extracted and modeled for telling the historical manuscripts. However, the ontological knowledge applications for supporting the semantic retrieval of the historical manuscripts are yet in the open for future research.

III PROPOSED SCHEME TO SEMANTIC HISTORICAL MANUSCRIPT RETRIEVAL

This work focuses on historical manuscripts. The scope of our work related to the historical event, especially Vietnam War.

The Domain Ontology

The advance of this historical ontology, especially concentrated on characterizes of events. This is because, as pre researchers, events play a vital role in the history. With this ontology, one can retrieve and analyze the historical manuscripts which are on the basis events or other related stuff to the particular event.

In this work, we have reused the conventional Simple News and Press Ontologies (SNaP) ontology and expanded as per our vocabulary which is illustrated in Figure 1:. Moreover, SNaP ontology consist of various ontologies that explains assets (images, text and video) and entities as well as events (place, people, concepts, organisation etc.), which appear in the contents of the news. Even though it is for news manuscripts; it was identified to be more perfect for our case since it has complete representation about the assets like events and manuscripts. Further, the event ontology is completely inherited from the public domain event ontology. The property named object property of sub Event of is rdfs: sub Property of event: subevent in addition with transitivity. In our domain, events are referred as com-pound entities. In the sense, they are high entities made via the relations with related entities such as organisation, people, things that are tangible as well as intangible and locations). Figure 1: illustrates the entire classes which were customized with the use of Top Braid Composer. Additionally, we have also imported SNAP ontology into Top Braid Composer and customized as per our vocabulary, in the sense, historical domain. Among the classes, the classes which were matched with our domain were factor, event, location, person as well as time and date. After that, we have expanded the ontology by including certain classes like object or stuff and country. The class 'country' was included to identify the country that involved in each and every war and the class 'stuff' comprises entities, which was both tangible and intangible for assigning people who involved in the war for their organization and country.

Classes	
Resource	
	Owl:all things
	Countries: country (345)
	Event:event (245)
	Event: factor
	Event: related
	Geo: spatial thing

Figure 1: Classes for Historical Domain

The Framework of Semantic Retrieval

The complete frame work is illustrated in Figure 2: which gives the detailed description of the entire process. as illustrated in framework, the prototype proceeds with SPARQL query. The query is knowledge based and the output comprises the list of semantic instances (entity) which meets the query's requirements. On the basis of matched instance, the prototype retrieves the manuscripts.

This particular framework has knowledge base in association with manuscript basis (information source) that uses one or more domain ontologies which describes the concepts that appears in a manuscript text. Moreover, the instances and concepts that presents in the knowledge base are explicitly linked to the manuscript and saved in annotation form. Those saved annotations are used for the creation of the preliminary representation for both the process (retrieval and ranking).

Figure 3: shows the mechanism of annotation. In order to perform annotation and indexing, a set of manuscripts are taken as input from the Wikipedia. Later, the output will be a new-fangled annotation and saved in knowledge base. The manuscript annotation process comprises the following steps:

- Initially, load the manuscript based data of basic terms that extracts the textual representation of the chosen entity. Moreover, the basic terms, which were loaded, have been extracted from the Wikipedia on battles and operations of the Vietnam War. The list of basic terms is tabulated in Table 1.

- For filtering the basic terms and for the identification of those terms which could operate as properties, instances and concepts, the linguistic analysis is used.
- The terms that are filtered, obtains the subset of semantic entities for annotation purpose.
- The weight of annotations is given as per the frequencies of the semantic entity within the distinct manuscripts and entire group.
- In order to produce the indexing list, the annotations are included to relational database.

The weighing process is done on the basis of an adaptation of classic information retrieval vector space model. Here, the assigned weights are the keywords that appear in the document.

The weights reflect the importance of the keyword, which describes the document content. In the same way, annotations reflect the significance of the instances in correspondence with the manuscripts. Furthermore, tf-idf algorithm is adopted for the automatic computation of weights. This is on the basis of frequency of occurrence of instance in every manuscript.

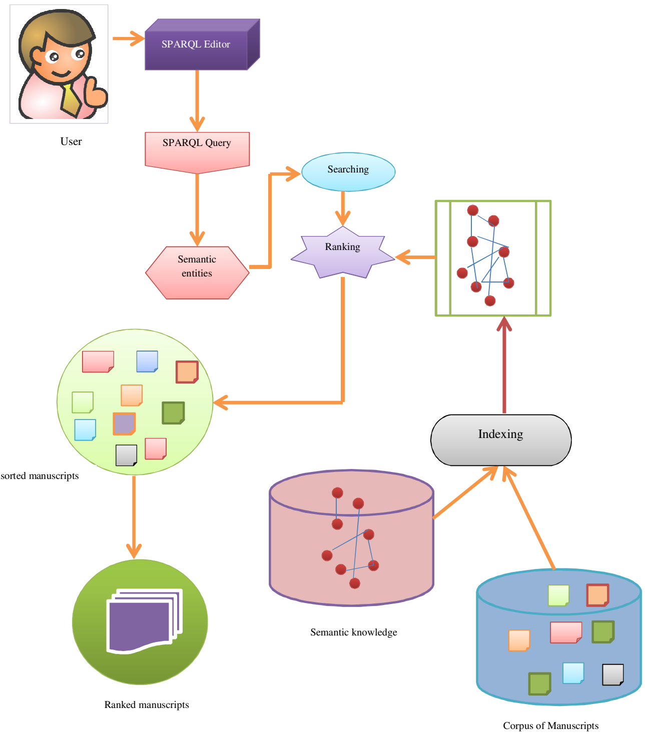

Figure 2: Framework of Semantic Retrieval

The weight of $d'_{x'}$ of an instance x' for the manuscript d' is formulated as given in eq(1).

$$d'_{x'} = \frac{fre_{x',d'}}{mx_{y'} fre_{y',d'}} \cdot \log \frac{|D'|}{n'_{x'}} \qquad (1)$$

where $fre_{x',d'}$ denotes the number of occurrence in d' of keyword that attached to x', $mx_{y'} fre_{y',d'}$ refers to the frequency of the repeated instance in d', $n'_{x'}$ represents the number of manuscripts that annotated with x' and D' denotes the set of the entire manuscripts in the search space.

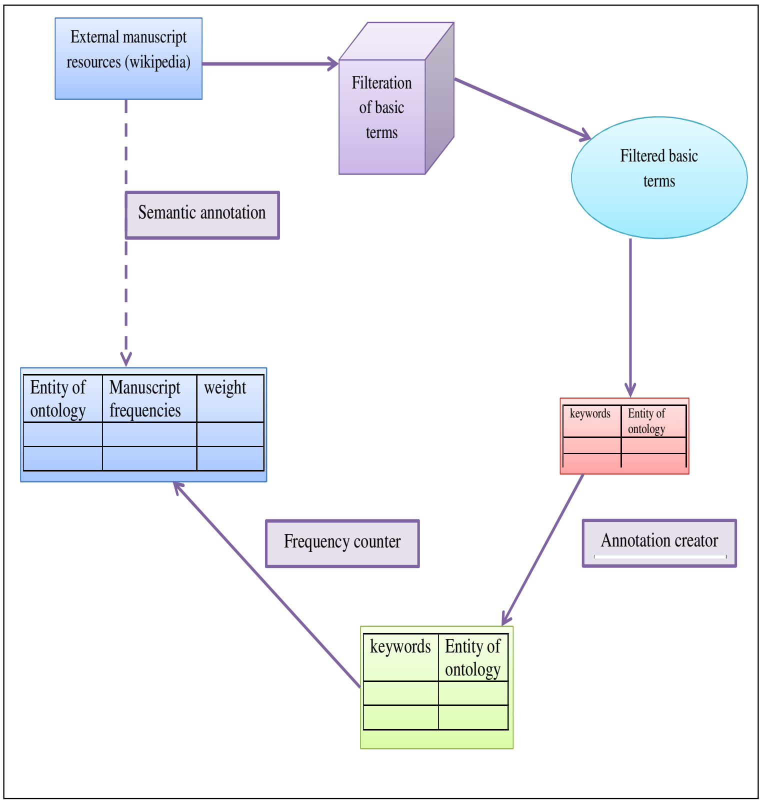

Figure 3: Manuscript Annotation

The execution of query generates a set of tuples which meets the SPARQL query. From the tuples, the semantic entities are mined and accessed the semantic index, for the collection of all the manuscripts that were annotated by the semantic entity. After the completion of document list, the semantic similarity values among the query and each document are calculated by the search engines with the use of classic vector space IR model. At last, the manuscripts are sorted and ranked in descending order, in accordance to the similarity values.

IV RESULT AND DISCUSSIONS

The proposed methodology is compared to the BM25 IR model with the aid of 133 manuscripts and five queries. The manuscripts are collected from Wikipedia. Basically, in the IR community, BM25 IR is the state of art and it has been used for the improvement of search engine [12]. The manuscripts are relevant to the events of battle as well as operations of Vietnam War. Moreover, this is the preliminary evaluation to extend with huge queries and manuscripts. The selected queries are listed in Table 2. In this ontology-based approach, the given queries will be translated to its corresponding SPARQL query. For instance, if the query is “mention the sub-event of the battle of Ap Bau Bang II?” would be translated to:

```
SELECT *
WHERE {
?event:BattleofApBauBangII pne:subEventOf?stuff.
                }
```

Table 3 and 4 shows the outcome of the above-mentioned queries. From the aforementioned tables, while comparing the existing keyword-based model, the semantic retrieval of the proposed model shows its excellence with MAP=0.9187, whereas the conventional method gives MAP=0.6269. Moreover, the results have also shown that the proposed ontology-based scheme retrieves fewer amounts of manuscripts, but the advantage is that, the retrieved manuscripts are relevant documents.

For instance, for query q1, three out of four retrieved manuscripts are relevant manuscripts and 27 out of 27 for q2 respectively. This result clearly shows the accuracy of the proposed ontology based model. Further, Figure 4: grants the overall performance of both the proposed as well as the conventional approaches. From the Figure, it is clear that the proposed model outperforms in retrieving historical manuscripts.

Table 1: Historical Domain's Basic Terms

Items	Basic Elements or Terms
1	Event
2	Sub event
3	Related event
4	Location
5	Person
6	Date
7	Time
8	Cause
9	Unit
10	belligerent

Table 2: Historical Domain's Basic Terms

Query#	Query
q1	Identify the related event as well the units which involved in the bombing of Tan Son Nhut Air Base
q2	Identify the sub event, initial date and final date for battle of Ap Bau Bang II
q3	Identify the related event and person who involved in battle of Hamburger Hill
q4	Identify the sub event and belligrant which involved in battle of Saigon 1968
q5	Identify the related event along with the location for operation apache snow.

Table 3: Evaluation Outcome of the Existing Keyword Based Scheme

Query	BM25 Approach				
	Retre.	Rela ∩ Retre.	Accuracy	Avg. Accuracy	Rec
q1	112	5	0.0446	0.6113	0.8333
q2	119	3	0.0252	0.6905	1.0000
q3	106	26	0.2453	0.5358	0.9286
q4	117	97	0.8290	0.8987	0.8981
q5	128	23	0.1797	0.3983	0.9200
		MAP	0.6269		

Table 4: Evaluation Outcome of the Proposed Ontology Based Scheme

Query	BM25 Approach				
	Retre.	Rela ∩ Retre.	Accuracy	Avg. Accuracy	Rec
q1	6	6	1.0000	1.0000	1.000
q2	4	3	0.750	1.0000	1.000
q3	27	27	1.000	1.0000	0.964
q4	125	108	0.864	0.8622	1.000
q5	63	23	0.365	0.7312	0.920
		MAP	0.918		

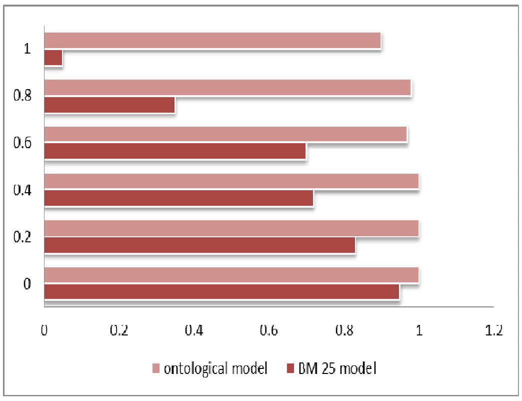

Figure 4: Manuscript Annotation

V.CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel approach for designing and developing a representation IR system, namely ontology based IR approach, exclusively for the retrieval of the historical manuscripts. The proposed model is compared to the existing keyword-based approach, with respect to the validation of retrieval of manuscripts using five respective queries.

Initial experimental outcomes have shown the extreme improvement in the accuracy, and the recall of the manuscripts retrieval. Moreover, we have proved that this proposed semantic retrieval scheme work could provide better capabilities in searching of manuscripts. Thus, the proposed scheme achieves the improvement over the existing keyword-based scheme in terms of introduction as well as the manipulation of the ontologies.

In future, we have planned to do further experiments with the consideration of the huge number of manuscripts, as well as, to improve the number of query coverage. It is even better to have the generic manuscript processing that could be used for numerous related event manuscripts.

REFERENCES

1. T. Elena, A. Katifori, C. Vassilakis, G. Lepouras, and C. Halatsis,"Historical research in archives: user methodology and supportingtools," International Journal on Digital Libraries, vol 11(1), 2010, p.25-36.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed.,vol. 2. Oxford: Clarendon, 1892, pp.68-73.
2. A. Gotscharek, A. Neumann, U. Reffle, C. Ringlstetter, and K.U.Schulz, "Enabling information retrieval on historical documentcollections: the role of matching procedures and special lexica,"Proc. The Third Workshop on Analytics for Noisy UnstructuredText Data2009, ACM: Barcelona, Spain, 2009, p. 69-76.

3. A. Gotscharek, U. Reffle, C. Ringlstetter, K.U. Schulz, and A.Neumann, "Towards information retrieval on historical document collections: the role of matching procedures and special lexica," International Journal on Document Analysis and Recognition (IJDAR), vol 14(2), 2011, p. 159-171.

4. M. Cabo, and R. Llavori, "A retrieval language for historicaldocuments, in Database and Expert Systems Applications,"G.Quirchmayr, E. Schweighofer, and T.M. Bench-Capon, Editors.Springer Berlin Heidelberg, 1998, p. 216-225.

5. Frakes, W., Introduction to information storage and retrievalsystems. Space, 1992. 14: p. 10.

6. M. Koolen, F. Adriaans, J. Kamps, and M. De Rijke, "A Cross-LanguageApproach to Historic Document Retrieval," in Advances in Information Retrieval, Springer Berlin Heidelberg, 2006, p. 407-419.

7. T. Pilz, W. Luther, N. Fuhr, and U. Ammon, "Rule-based Search in Text Databases with Nonstandard Orthography," Literary and Linguistic Computing, vol 21(2), 2006, p. 179-186.

8. V. Mirzaee,, L. Iverson, and B. Hamidzadeh, "Towards ontological modelling of historical documents," in The 16th International Conference on Software Engineering and Knowledge Engineering (SEKE), 2004.

9. I. Corda, "Ontology-based representation and reasoning about the history of science," The University of Leeds, 2007.

10. S. Schockaert, M. Cock, and E. Kerre, "Reasoning about fuzzy temporal information from the web: towards retrieval of historical events," Soft Computing, vol 14(8), 2010, p. 869-886.

11. O. Alonso, M. Gertz, and R. Baeza-Yates, "On the value of temporal information in information retrieval," SIGIR Forum, vol 41(2), 2007, p. 35-41.

12. J. Pérez-Iglesias, J. R. Pérez-Agüera, V. Fresno, and Y. Z. Feinstein, "Integrating the probabilistic models BM25/BM25F into Lucene," arXiv preprint arXiv:0911.5046, 2009.