International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 3
ISSN: 2088-5334

Using Ontology-Based Approach to Improved Information Retrieval Semantically for Historical Domain

Fatihah Ramli[a], Shahrul Azman Mohd Noah[b,1], Tri Basuki Kurniawan[b,2]

[a] *Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 93400 Kota Samarahan, Sarawak, Malaysia*
E-mail: rfatihah@unimas.my

[b] *Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia.*
E-mail: [1]shahrul@ukm.edu.my; [2]tribasukikurniawan@yahoo.com

***Abstract*— Searching and retrieving documents from large historical archives prove to be challenging for the information retrieval (IR) field as historians typically employ their knowledge, experience, and intuition. There are several works done on the application of IR in historical documents. As such, the conventional IR model is mostly used a simple Bag-of-Word (BOW) approach and usually unable to support precise document retrieval for the domain of history. We proposed an ontology-based approach to semantically index and ranked rich historical documents. The historical documents relating to the Vietnam War were chosen for this study. Several existing ontologies have been reviewed to identify the most suitable concepts and properties which contain rich information pertaining to relevant entities such as an event, time, and people. The domain ontology was developed by utilizing the existing Simple News and Press (SNaP) ontology and extended with concepts related to the Vietnam War. The ontology was then semantically mapped with concepts found in a collection of 133 documents relating to the Vietnam war. In this paper, we also proposed a simple ontology-based weighting mechanism derived from the classic *tf-idf* scoring scheme. Finally, 20 SPARQL queries are implemented to do the evaluation. The evaluation shows that the proposed ontological-based approach achieved better results as compared to the base-line BM-25 probabilistic retrieval model in terms of precision and recall metrics. The use of the ontology-based approach in document retrieval can compete with the keyword-based approach.**

***Keywords*— ontology; information retrieval; semantic search; historical documents; bag-of-word.**

I. INTRODUCTION

Information Retrieval (IR) research in various fields gives many new ideas for researchers to improve existing approaches in all areas. However, recently the field that receives special attention is history. Historians still expect a better approach for more accurate access to historical documents [1]. 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]. Hence, IR for historical documents is an essential issue to be studied.

Historical documents can be defined as those that keep information related to a time instant at which the documents were published at the same time that is still useful in the future [3]. Searching and retrieving documents from large historical archives 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 [4]. Hence, Elena et al. [1]suggest that historians need historical source repositories and building tools that will enable them to access comprehensive information rapidly. 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 are not sufficient as the documents contain rich semantic information relating to relevant entities such as an event, time, and people.

Therefore in this paper, we proposed an ontology-based approach to index and ranked [5], [6] semantically rich historical documents. The ontology developed centralized on the event-related elements, which are essential to the historical domain. An ontology-based approach to document retrieval is not new, as demonstrated. 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.

There are several works done on the application of IR in historical documents. Some applications of IR to historical documents mostly concern with spelling issues whereby users expect those modern keywords able to match with elements of words/spelling available in historical documents[5]. This is because there are too many spelling variants located in the large document of historical texts [3]. Full-text indexing of such documents is not sufficient 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, still not fully representing the fundamental characteristic of historical documents. The historical document can be defined as those that keep information related with time instant at which the documents were published while is still useful in the future [4]. 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] stated that historians need historical source repositories and building tools that will enable historians to access the comprehensive information rapidly. The 20th and early 21st centuries have transformed the way people obtained information.

Hence, users expected a wealth of historical information could be shared and reused through digital libraries that can provide the best-matched document for any search request in answering competency questions as well as providing support to a selected scenario [4], [7]. In order to fulfil the user request, Mirzaee, Iverson [4] and Corda [7] suggested the semantics of a historical document, which attempts to allow a richer representation of its embedded knowledge that should be captured rather than capable with standard text manipulation tools. Demner [8] also focus on the question and answer the problem and implemented a syntactic-semantic method for extracting the question frames from the free text topics. The used of semantics could be more useful if it is simplified through defining the time-based relations. Furthermore, the work by Schockaert, Cock [9] suggested that the documents should be sorted according to temporal aspects in the context to improve the IR systems. On the other hand, Alonso, Gertz [10] denotes that recognizing and the used of temporal information for IR applications was an important feature that can improve the functionality of search applications. Campos [11] also support that temporal can enhance the effectiveness of the IR method by exploiting temporal information in documents and queries.

However, the works mainly suggested the type of knowledge that should be extracted and modelled for describing historical documents. As such, the applications of such ontological knowledge to support semantic retrieval of historical documents are still open for further research.

II. Material and Method

The focus of this work is on historical documents. We chose to scope our work to ontology and documents relating to the Vietnam War.

A. The Domain Ontology

The development of our history ontology mainly focused on the aspects of events. This is due to the opinion of various researchers that the event is an essential element in history[12]. With this ontology, historical documents can be retrieved and analyzed based on events or other aspects related to the events.

In our work, the ontology development was executed semi-automatically and formalized by the domain experts and ontology developers. We reused the existing Simple News and Press Ontologies (SNaP) ontology and expanded it based on our vocabulary as shown in Figure 1. SNaP ontology consists of several ontologies that emerge in news content. Although it is meant for news document, it was found to be suitable in our case as it contains detailed representation about the event as well as documents (i.e. assets). The event ontology acquires from the public domain event Ontology. The object property of subEventOf is a rdfs:subPropertyOf event:sub_event with the addition of transitivity. Events are considered as composite entities in our domain (i.e. they are consisting of two or more interconnected entities with other entities, particularly people, organizations, locations and things both tangible and intangible).

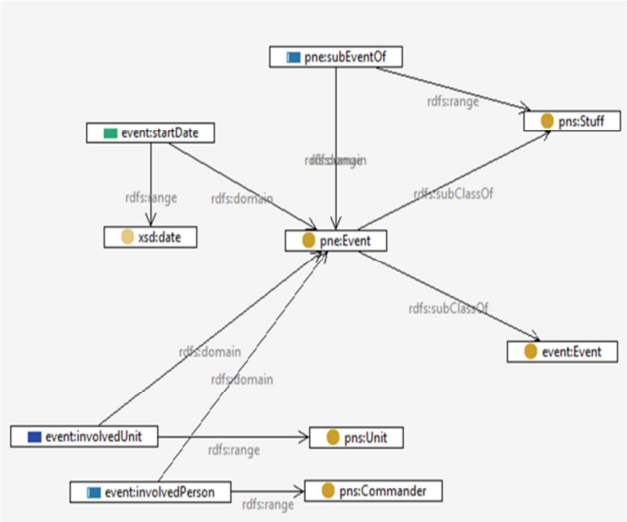

Fig. 1 Some of the concepts and properties on Historical Documents

Figure 1 above shows some of the classes and properties that were customized using Top Braid Composer. We have imported SNAP ontology into TopBraid Composer and started customizing it based on our vocabulary, i.e., historical domain. Among the basic classes that were matched to our domain were event, factor, person, spatial thing (location) and date. Then, we expanded the ontology by adding some classes like country and stuff. The country class was added to know the country involved in each war, whereas stuff class includes both tangible and intangible entities to assign people involved in a war with their country and organization. The details on ontologies review are elaborated in the next section.

B. Consider Reusing Existing Ontology

We considered reusing existing ontology developed by others for semantic annotation. Available resources had to be

checked whether they could improve and expand our particular domain and task. For our work, ontology reuse was beneficial as there was a time constraint in developing a new ontology from scratch especially in adapting and updating the necessary concept in a new ontology.

There are several existing ontologies were reviewed in historical domain. The first ontology is STOLE ontology. STOLE is a reference ontology which provides a vocabulary of terms and relations to explicitly model the domain specific. STOLE ontology used the history of Italian Public Administration as domain specific. The main aim of the STOLE Ontology is to have an explicit design model on historical concepts and seek views on particular areas.

STOLE aims to gather information about the most relevant journals on the history of public administration legislation in Italy that published between 1848 and 1946. The STOLE ontology's construction consists of three main phases: 1) Identification of key concepts, 2) Identification of the proper language and Tbox implementation, 3) Ontology population [13]. In the first phase, the key concepts involved in a specific domain must be defined by the domain expert. The domain experts provide manual semantic annotations that would be added to the ontology using JAVA program. Next, they classified all the data that are related to historical documents, and the results of all the concepts would be viewed in the form of a taxonomy that consists of three elements as shown in Table 1.

TABLE I
TAXONOMY OF STOLE ONTOLOGY

Elements	Examples
Data on the author of the article	Name, surname, biography
Data on the journal and the article	Article title, journal name, date and topic raised in the article. Persons, historical events, institutions
Data on the relevant facts and persons cited in the article.	

Table 2 shows the size of the STOLE ontology that was computed by PROTEGE. Finally, ontology populations are carried out to fill in missing entities in Abox with semantic annotations automatically. STOLE ontology is accessible to the public and can be considered as an expandable ontology.

TABLE II
TBOX STATISTICS ABOUT STOLE ONTOLOGY

Classes	14
Axioms	440
Object Properties	30
Data properties	29

Next ontology is event ontology. [14] stated that a semantic portal for cultural heritage required event ontology because of three reasons: 1) events need ontological identifiers (URIs) to build a metadata collection, 2) events are essential in creating a semantic relationship between cultural content and 3) Historical events are essential to shape the backbone of chronological history. [14] developed an event ontology using Finnish history as a domain specific. The historical event ontology was based on the timeline that was created by Agricola network and being utilized as part of the semantic portal "CultureSampo—Finnish Culture on the Semantic Web", a cross-domain follow-up system of Museum Finland. This portal is an application of semantic technologies toward the development of e-culture portals that providing multimedia access to distributed collections of cultural heritage objects[15]. The classifications of events were based on the temporal timeline and other dimensions such as event types, i.e. war, coronation or branch history, i.e. political history, history of science. They annotated manually 220 events between the years 1850–1920 utilizing the SAHA annotation tool combined with ONKI Ontology library servers for utilizing shared domain ontologies. As a result, history ontology defines URIs for events can be utilized for annotating other cultural objects and relating them to each other. However, the event ontology is not accessible to the public and cannot be considered as an expandable ontology.

Another ontology is FDR Historical Ontology. The primary goal of the FDR/Pearl Harbor project was developing applications that could help to improve searching and retrieving information from a set of documents. The documents was taken from the Franklin D. Roosevelt Presidential Library (FDRL). This project used a set of documents that referred to situations and events over the ten-year period, which was before the bombing of Pearl Harbor. The FDR/Pearl Harbor Project built the historical ontology based on the model presented using the entities and events in its document collection [16]. The FDR temporal ontology included only clearly defined endowment entities in the collection of documents, which comprised the general categories. The categories include geopolitical entities, geopolitical organizations, military organizations, military vehicles, geographical objects, geographical artifacts, documents, agreements, persons, and political organizations. Event and entity annotation of these documents used General Architecture for Text Engineering (GATE) to complete the manual semantic annotation. Next, automatic annotations are carried out using machine learning based on hand validated annotation. However, the FDR temporal ontology is not accessible to the public and cannot be considered as an expandable ontology.

Besides that, the Henry III project also has succeeded in producing a collaborative project between King's College London and the National Archives (UK). The primary aim of this project was to represent the complexity of historical documents known as the Fine Rolls [17]. The FRH3 ontology consists of several classes such as authority (Person, Place, and Subject) and Factoid (Role, Relationship and Role Relationship). The RDF/OWL had been chosen to do an authority list based on several reasons: 1) It is a W3C standard for the Semantic Web; 2) The number of existing tools is more significant for the RDF/OWL; 3) It can be expressed as XML, simplifying the process of data delivery and this makes it easy to index people, places and subjects using XSLT; 4) It can create the expression of relationship among the instances explained in the fine rolls source materials [17]. However, this ontology is not accessible to the public and cannot be considered as an expandable ontology.

In year 2005, [18] stated that Ontology-driven access to museum information can be represented as "core ontology"

that combines basic entities and relationship across the various metadata vocabularies. The core ontology is useful in helping to integrate information from multiple vocabularies and uniform processes across various sources of information. Core ontology is the basic core formal model for tools that integrate source data and perform a variety of functions [18]. There are several classes in this ontology such as E2 Temporal Entity, E52 Time-span, E3 Condition State, E4 Period and E5 Event. The ontology process was also helping in enriching knowledge[19]. Hence, higher levels of complexity are acceptable, and the design should be more motivated by logical correctness and completeness than human understanding. However, this core ontology is not accessible to the public and cannot be considered as an expandable ontology.

The last ontology is SNaP ontology. SNaP ontology is a news ontology that consists of multiple ontologies, which describe assets (text, images, video) and the events as well as entities (people, places, organizations, abstract concepts, etc.). There are two categories of entities in SNaP ontology: simple entities i.e. stuff and complex entities i.e. event. The term stuff can be represented as abstract and intangible concepts as well as tangible things. The total numbers of concepts that are involved in event and stuff ontologies are 22 concepts. While it is intended for news documents, it is found to be appropriate in our case as it contains the detailed representation of events, people, organizations, locations, tangible and intangible things as well as documents [20]. SNaP ontology is accessible to the public and can be considered as an expandable ontology.

In conclusion, based on the above studies we have identified several important features for selecting an appropriate ontology to be expanded. The most important feature is availability whereby existing ontology must be accessible for reuse and subsequently developed based on domain specific. For instance, only STOLE and SNaP ontologies are available to the public. Besides, we also need to know the size and content of an ontology to facilitate the development of ontology. For example, SNAP and STOLE ontologies have the most number of concepts compared to other ontologies. With this, both ontologies have the potential to be reused for this study. Therefore, our scope is focused on the aspect of events; the SNaP ontology is chosen as existing ontologies to be used because it has many entities connected to events.

C. The Semantic Retrieval Framework

The overall framework is shown in Figure 2 which describes the whole retrieval process. As shown in the framework, the prototype use a formal SPARQL query as the input. The query is based on the knowledge base where the output consists of a list of semantic entity (instance) that meets the requirements of the query. The prototype then retrieves documents based on the matched entities.

The semantic retrieval framework has a knowledge base that associate to the historical domain (the document base) by using SNaP ontology that describes concepts emerging in the document text. The connection between the concepts in the knowledge base and the documents are connected precisely and stored in the form of annotations. These annotations are used to create an initial representation for retrieval and ranking processes. Figure 3 illustrates the annotation mechanism which starts with the system takes as input a set of documents from Wikipedia to do annotation and indexing. Then they will be a new annotation output and stored in the knowledge base. The implementation of document annotation process includes the steps as follow:

- Load the external resources (the historical documents) of basic terms which is extracting the text of the selected entities. The basic terms have extracted from Wikipedia on Battles and operations of the Vietnam War. Table 3 shows the list of basic terms.
- The linguistic analysis is utilized to clean by filtering basic terms and to identify all the suitable terms that can be used as concepts, instances and properties[21].
- Then the filtered basic terms acquiring the annotation of semantic entities.
- The annotations are weighted in accordance with the semantic entity frequencies in the historical documents.
- The annotations are included to the relational database for producing indexing list. Figure 4 shows the semantic indexing list.

TABLE III
BASIC TERMS

Item	Basic Term
1	event
2	location
3	person
4	date
5	cause
6	unit
7	belligerent

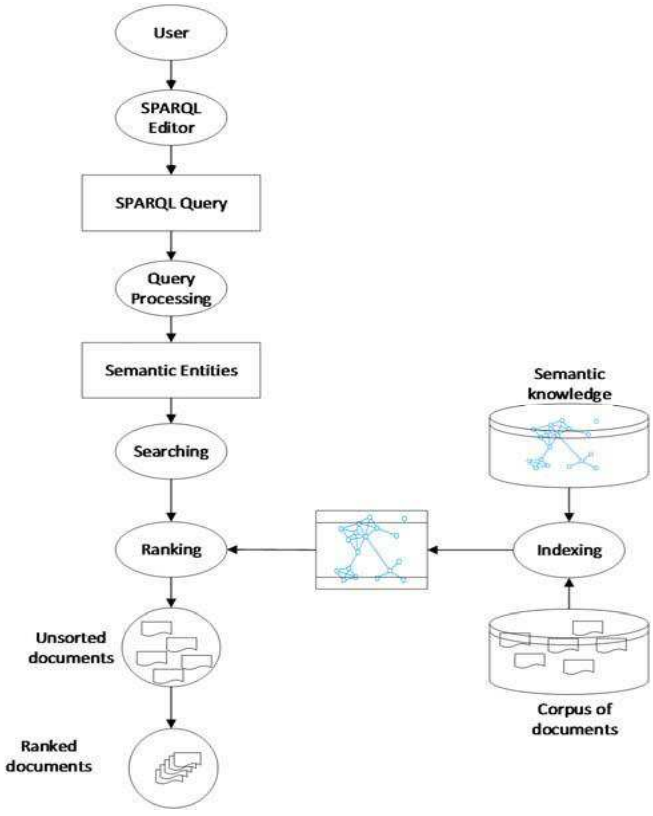

Fig. 2 Semantic Retrieval Framework

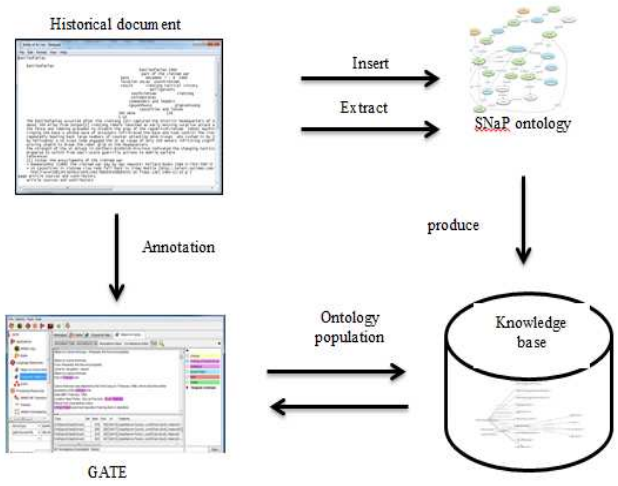

Fig 3 Document Annotation

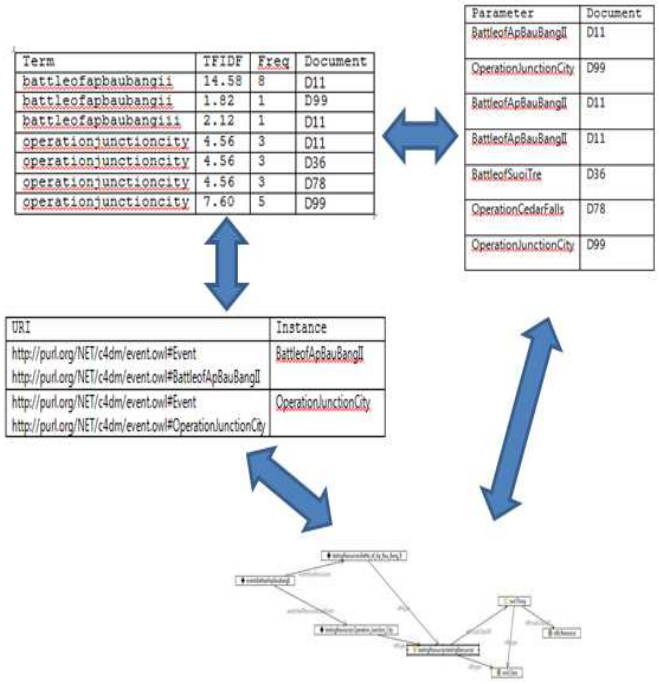

Fig 4 The implementation of semantic indexing list

The weighting used an accommodation of the classic information retrieval vector space model. This model shows keywords emerging in the document are allocated weight indicating the importance of the keywords for describing the content of document. Similarly, for this study, annotations are allocated weights that indicate the importance of instances regarding the documents. The *tf-idf* algorithm is used to calculate the weighting, which is based on the rate of instances occurrence in each document. In detail, the weight of term *dx* of an instance *x* for a document *d* is calculated as in[22], [23]:

$$d_X = \frac{freq_{x,d}}{\max_y freq_{y,d}} \cdot \log \frac{|D|}{n_x} \quad (1)$$

where $freq_{x,d}$ is the number of keywords attached to x that appeared in d, $max_y freq_{y,d}$ is the rate of occurrence the most repeated instance in d, n_x is the number of documents annotated with x and D is the set of all documents in the search area.

The query implementation produces a set of tuples that meet the SPARQL query. Then, the semantic entities extracted from the tuple and go into the semantic index to gather all the documents in the repository that have been annotated by the semantic entity. Once the documents lists are completed, the search engines calculate the semantic similarity value between the query and each document utilizing the classic vector space IR model. Finally, we sort and rank the documents in descending order according to the similarity values.

In this study, SPARQL queries are implemented directly through SPARQL query panel from JAVA application using the Jena library special method. The specific method of this library is named as onto.SPARQL library. The Jena libraries method can handle queries that have objects and relationships and reasoning. The similarity measure used to ranked documents was based on the conventional cosine similarity measure.

III. Result and Discussion

A. Evaluation

We have compared the proposed approach with BM25 IR model using a corpus of 133 documents from Wikipedia and a total of 20 queries. We initially proposed the queries and the documents related to them (ground truth) were judged with the help from historians. BM25 IR model is considered as state-of-art in the IR community, and it has been widely used by IR researchers to improve search engine relevance[24, 25]. The documents relate to the event of Battles and operations of the Vietnam War. The precision and recall values for all queries are shown in Table 5 and Table 6. The queries are listed in Table 4. For the ontology-based approach, the queries were translated into the corresponding SPARQL query. For example, the query "What is the sub-event of the Battle of Ap Bau Bang II?" was translated to:

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

Table 5 and Table 6 show the results of the evaluation using the above queries. Based on Table 5 and Table 6, the MAP results are generated its clearly demonstrates the precision of all queries are increased for semantic retrieval approach. The MAP result of precision for semantic retrieval is 0.942 compared to the conventional keyword-based approach is 0.685. Hence, this result shows the uses of semantic retrieval approach to retrieve relevant historical documents are more effective and accurate. Besides, it also shows that the semantic retrieval approach can provide better search capabilities and help to improve the conventional keyword-based approach. This factor is discussed in this

observation result, which shows different levels of performance for both approaches in different situations.

TABLE IV
SAMPLE QUERIES

Query	Query
1	Find sub-event, start date and end date for Battle of Ap Bau Bang II
2	Find related event and person involved in Battle of Hamburger Hill.
3	Find related event and location for Operation Apache Snow.
4	Find sub-event and belligerent involved in Battle of Saigon 1968.
5	Find related event and unit involved in Bombing of Tan Son Nhut Air Base.
6	Find person involved in Battle of Ap Bau Bang II and its location.
7	Find location and cause for Battle of Hamburger Hill.
8	Find commander involved in Battle of An Lao and which country they represented.
9	Find unit involved in Operation Hong Kil Dong and which group they represented.
10	Find start date and location for Operation Frequent Wind.
11	Find cause and end date for Operation Crimp.
12	Find belligerent and unit involved during Battle of Suoi Chau Pha.
13	Find start date, end date and person involved in Operation Dewey Canyon.
14	Find related event and cause for Operation Babylift.
15	Find related event and sub-event for Bombing of Tan Son Nhut Air Base.
16	Find sub-event and person involved in Operation Coronado XI.
17	Find unit involved in Operation Union II and which country they represented.
18	Find location and start date for Battle of An Loc.
19	Find end date and related event for Battle of Hoa Da Song Mao.
20	Find person and unit involved in Operation Attleboro.

For example, in the first situation for Q5, Q14 and Q15, in one event A some other events such as B and C are related to event A. In this situation, the semantic retrieval approach gives better results than keyword-based search because knowledge base contains many instances of such Event Bombing of Tan Son Nhut Air Base, Operation Babylift and half of them match the query. Keyword-based search only identifies a document as relevant when the document contains words such as sub-event, start date, end date and the person involved in an event. While semantic search gets sub-event and date about the events once the sub-event name and date are specified in a document.

Next, the second situation involves Q3 which is "Find related event and location for Operation Apache Snow". In this query, the ontology knowledge base has only a few instances of Operation Apache Snow, so that not all the documents relevant to the query are annotated. This situation causes the precision to decrease to a lower value when the recall increases. Although the number of semantic search precision is low, it still has a good average precision of the document position that has been annotated with instances in the knowledge base. The results also showed that access to this query in the ontology approach could provide access to documents more accurate than the conventional approach based on keywords.

TABLE V
EVALUATION RESULT FOR THE KEYWORD-BASED APPROACH (BM25 MODEL)

Query	Retrieved document (*n*)	Rel. ∩ Ret	Precision at *n* retrieved document	Recall at *n* retrieved document	Average Precision
Q1	119	3	0.025	1.000	0.691
Q2	106	26	0.245	0.929	0.536
Q3	128	23	0.180	0.920	0.398
Q4	117	97	0.829	0.898	**0.899**
Q5	112	5	0.045	0.833	0.611
Q6	117	7	0.060	0.700	0.433
Q7	117	90	0.769	0.891	0.785
Q8	115	105	0.913	0.868	0.935
Q9	126	54	0.429	0.931	0.534
Q10	127	43	0.339	0.915	0.445
Q11	127	1	0.008	1.000	1.000
Q12	105	96	0.914	0.793	**0.963**
Q13	127	1	0.008	1.000	1.000
Q14	111	3	0.027	0.750	0.448
Q15	105	5	0.048	0.833	0.587
Q16	115	5	0.043	1.000	**0.925**
Q17	124	116	0.935	0.936	0.975
Q18	116	76	0.655	0.916	**0.759**
Q19	119	24	0.202	0.960	0.358
Q20	115	20	0.174	0.870	0.421
Mean Average Precision (MAP)					0.685

TABLE VI
EVALUATION RESULT FOR ONTOLOGY-BASED IR APPROACH (ONTOLOGICAL MODEL)

Query	Retrieved document (*n*)	Rel. ∩ Ret	Precision at *n* retrieved document	Recall at *n* retrieved document	Average Precision
Q1	4	3	0.750	1.000	**1.000**
Q2	27	27	1.000	0.964	**1.000**
Q3	63	23	0.365	0.920	**0.731**
Q4	125	108	0.864	1.000	0.862
Q5	6	6	1.000	1.000	**1.000**
Q6	8	8	1.000	0.800	**1.000**
Q7	106	97	0.915	0.960	**0.956**
Q8	112	109	0.973	0.901	**0.981**
Q9	3	3	1.000	0.052	**1.000**
Q10	49	45	0.918	0.957	**0.954**
Q11	1	1	1.000	1.000	**1.000**
Q12	124	115	0.927	0.950	0.961
Q13	2	1	0.500	1.000	**1.000**
Q14	4	4	1.000	1.000	**1.000**
Q15	7	6	0.857	1.000	**0.976**
Q16	6	5	0.833	1.000	0.877
Q17	122	122	1.000	0.984	**1.000**
Q18	101	71	0.703	0.855	0.737
Q19	24	23	0.958	0.920	**0.916**
Q20	16	13	0.813	0.565	**0.896**
Mean Average Precision (MAP)					**0.942**

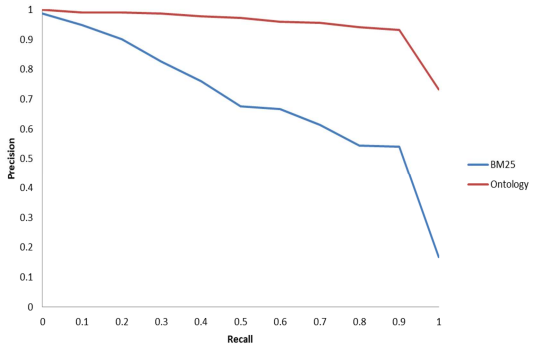

Fig. 5 Average precision at standard recall level for the ontological model and the conventional BM25 model

B. Discussion

In conclusion, the precision value of 20 queries using an ontology-based approach is much higher than the keyword-based approach. This is because the ontology-based approach is based on the concepts and relationships extracted from text-based semantics. Besides, the use of ontology as a database domain supports semantic search because semantic indexing methods allow data to be developed and linked more widely and in detail between each concept with the other concepts more practically. The effectiveness of the matching concept in every query depending on the annotation of documents in the ontology. In conclusion, the use of the ontology-based approach to information access and retrieval of documents is effectively able to compete with the keyword-based approach. Figure 5 clearly shows the better performance of the proposed ontology-based approach for historical documents. It provides an overall performance comparison between both approaches.

IV. CONCLUSIONS

We have discussed an ontology-based approach to support in designing and developing new representation IR-system in historical domain. Several experiments have been implemented on ontology-based approach and keyword-based approach to verify the retrieval of documents by using twenty queries. The evaluation results show that the purposed ontologies improve the precision and recall of the retrieval of the documents. As a conclusion of this work, we would like to focus on the semantic retrieval approach can contribute better search ability, thus achieving an advancement on keyword-based retrieval using the introduction and exploitation of ontologies. Future research works include further experiments by considering many documents, and it is also interesting to have a generic ontology and document processing which can be used for various other event-related documents.

ACKNOWLEDGMENT

We are grateful to anonymous reviewers for their comments. We also would like to thank UNIMAS for appreciating this work.

REFERENCES

[1] T. Elena, A. Katifori, C. Vassilakis, G. Lepouras, and C. Halatsis, "Historical research in archives: user methodology and supporting tools," International Journal on Digital Libraries, vol. 11, no. 1, pp. 25–36, 2010.

[2] A. Gotscharek, A. Neumann, U. Reffle, C. Ringlstetter, and K. U. Schulz, "Enabling information retrieval on historical document collections," *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data - AND 09*, 2009.

[3] M. J. A. Cabo and R. B. Llavori, "A retrieval language for historical documents," *Lecture Notes in Computer Science Database and Expert Systems Applications*, pp. 216–225, 1998.

[4] 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.

[5] W. Frakes, *Introduction to information storage and retrieval systems.* Space, 1992. **14**: p. 10.

[6] S. Shekarpour, F. Alshargi, K. Thirunaravan, V. L. Shalin, and A. Sheth, "CEVO: comprehensive event ontology enhancing cognitive annotation on relations," in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 2019, pp. 385–391.

[7] I. Corda, " Ontology-based representation and reasoning about the history of science," M. Eng. thesis, The University of Leeds, 2007

[8] D. Demner-Fushman, S. Abhyankar, A. Jimeno-Yepes. A Knowledge-Based Approach to Medical Records Retrieval. in *TREC*. 2011.

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

[10] O. Alonso, M. Gertz, and R. Baeza-Yates, On the value of temporal information in information retrieval. *SIGIR Forum*, 2007. **41**(2): p. 35-41.

[11] R. Campos, G. Dias, A. M. Jorge, A. Jatowt. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 2015. **47**(2): p. 15.

[12] H. P. Blossfeld, G. Rohwer, and T. Schneider, Event history analysis with Stata, 2019: Routledge.

[13] G. Adomi, M. Maratea, L. Pandolfo, L. Pulina. An ontology for historical research documents. in *International Conference on Web Reasoning and Rule Systems*. 2015. Springer.

[14] E. Hyvönen., O. Alm, and H. Kuittinen. Using an ontology of historical events in semantic portals for cultural heritage. in *Proceedings of the Cultural Heritage on the Semantic Web Workshop at the 6th International Semantic Web Conference (ISWC 2007)*. 2007.

[15] D. Calvanese, A. Mosca, J. Remesal, M. Rezk, and G. Rull, "A 'historical case' of Ontology-Based Data Access," in *2015 Digital Heritage*, 2015, pp. 291–298.

[16] N. Ide and D. Woolner, "Historical Ontologies," in *Words and intelligence II*, K. Ahmad, C. Brewster, and M. Stevenson, Eds. Dordrecht: Springer Netherlands, 2007, pp. 137–152.

[17] J. M. Vieira and A. Ciula. *Implementing an RDF/OWL Ontology on Henry the III Fine Rolls.* in *OWLED*. 2007. Citeseer.

[18] O. Signore. *Ontology driven access to Museum Information*. in *Annual Conference of CIDOC Documentation and Users CIDOC*. 2005.

[19] C. d'Amato, S. Staab, A. G. B. Tettamanzi, T. D. Minh, and F. Gandon, "Ontology enrichment by discovering multi-relational association rules from ontological knowledge bases," in *Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16*, New York, New York, USA, 2016, pp. 333–338.

[20] F. Ramli and S. A. Mohd Noah, "Building an event ontology for historical domain to support semantic document retrieval," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, p. 1154, Dec. 2016.

[21] Gopnik, M., *Linguistic structures in scientific texts*. Vol. 129. 2018: Walter de Gruyter GmbH & Co KG.

[22] J. Pérez-Iglesias, J. R. Perez-Aguera, V. Fresno, Integrating the probabilistic models BM25/BM25F into Lucene. arXiv preprint arXiv:0911.5046, 2009.

[23] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016.

[24] G. L. Zúñiga, "Ontology: its transformation from philosophy to information systems," *Proceedings of the international conference on Formal Ontology in Information Systems - FOIS 01*, 2001.

[25] F. Jian, J. X. Huang, J. Zhao, T. He, and P. Hu, "A Simple Enhancement for Ad-hoc Information Retrieval via Topic Modelling," *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR 16*, 2016.