Journal of University of Babylon for Pure and Applied Sciences,Vol.(27), No.(4): 2019

Ontology-Based Model for Information Retrieval: an Application of Time Nouns in Nahj Al-Balagha

Ihab L. Hussein Alsammak[a] Humam M. Abdul Sahib[b]
Intedhar Shakir Nasir[c]

[a]*Ministry of Education, Directorate General of Education of Karbala*
[b]*Department of Electrical and Electronics EngineeringCollege of Engineering, University of Kerbala, Iraq*
[c]*Department of Family and Community Medicine,College of Medicine, University of Kerbala, Iraq*

[a]ehablaith@gmail.com
[b]humam.alkaabi@uokerbala.edu.iq
[c]intedhar.shakir@uokerbala.edu.iq

ARTICLE INFO
Submission date:30/5/2019
Acceptance date: 25/6/2019
Publication date: 18/9/2019

Keywords: Nahj Al-Balagha, information retrieval, time nouns, semantics, Ontology domain met.

Abstract

The internet plays a key role in life through the massive data that it provides. Currently, managing data and finding information on the internet is inaccurate because it depends on the form of the word rather than its meaning. Data representation and access are important factors when it comes to Information Retrieval (IR). In order to overcome the problem of document similarity, there are various similarity measurements in place that function according to weight, indexing and matching. Ontology is a data management infrastructure that gives precedence to the meaning of a word, the relationship between words and the domain of knowledge.
This paper presents a semantic system proposal based on a particular field of knowledge (time nouns) and relies on semantic input by indexing the search engine using a Vector Space Model (VSM). The aim of this work is to improve the retrieved semantic information by constructing a query based on the matching and similarity between the query words in the system. This paper builds upon previous work carried out in the same area [1]. The system was evaluated by using the similarity, average precision and recall of the experiments' results.

1.Introduction

The internet is the largest interconnected network of nodes that has a massive number of data warehouses, including academic and entertainment resources, to name but a few [2]. As a result of the decentralized links, there is a difficulty in making the machine have the ability to understand and process information without human interaction [3]. The structure of the traditional web depends on incoherent formats and most of the knowledge is unreliable because it is based on terms rather than concepts [4] . The retrieval of information is the cornerstone of the web. With billions of documents, it will become increasingly difficult to use keywords in searching because this type of search is based on the terms in documents rather than concepts. For example, the search results for "the best way in Spain" will show all documents containing one or more words in the sentence and other words will appear to be mismatched [5]. Restructuring links, accurate descriptions and the representation of knowledge according to domain and concepts will bring about the creation of conceptual knowledge and help demystify machines in the understanding of information [3]. Ontology can be defined as "the specification of conceptualizations used to help programs and humans to share knowledge"[6]. Ontological modeling is based on concepts and specific relationships between concepts, which can be enlisted in a hierarchical way. A semantic web uses ontological modeling to represent a domain and expand queries through understanding the natural language. Humans can easily understand semantic linking at the level of words (for example, the relation between monkeys and bananas). However, this understanding is difficult when it is related to sentences and becomes even more complex when these sentences depend on meaning and language philosophy. This task is impossible for traditional machines because it depends on the term rather than meanings. For example, the search results for the word "time"

in “Google” will be day, tomorrow, yesterday and morning, while there are also other types of time, such as biological time (e.g. childhood).There are currently four approaches that are used in a semantic web [7] :

1) Demystification by contextual analysis and definition. For example, the word "Galaxy" may refer to the system of stars or to a Samsung phone.
2) Facts and inferred additional facts from them. For example, each pen is located on the table and each table is located on the left of the room. The system will conclude that each pen is located on the left of the room.
3) Process content by defining objects and rules in the sentence, which is used in natural languages.
4) Domain representation and expand queries.

A Resource Description Framework (RDF) is a framework used to represent domain-objects and relationships between them [8]. Information Retrieval (IR) is the process of retrieving information that is relevant to the user query. In order to achieve this aim, there are three step: (1) indexing documents; (2) removing stop words; and (3) matching documents [9]. IR systems can be categorized into three types: (1) the Boolean model; (2) the probabilistic model; and (3) the Vector Space Model (VSM) [10]. The Term Frequency-Inverse Document Frequency (TF-IDF) determines the relationship between a word and documents [11].

Arabic is one of the six languages of the United Nations. It is the native language of Arab countries and the religious language of Muslims, as well as an official language in some other countries such as Eritrea. Arabic is characterized by a lot of vocabulary and each word has its own meaning. For example, the word “الازلي” (eternal) is different to “ الابدي” (eternal) which is sometimes known from the context of speech. Nahj al-Balagha is a collection of speeches and letters by Imam Ali ibn abi Talib that contains judgment, advice, sermons and knowledge. Henry Corbin said that “Nahj al-Balagha, after the Quran and Prophet Muhammad’s tradition, has prime importance”.

The main contribution of this work is to build an ontology based on the time nouns domain and to retrieve the relevant information. This work is based on previous work carried out by H. M. A. Sahib [1].

The first section in this work introduces the semantic web and information retrieval. The second section provides an overview of some related work. The third section introduces the proposed system and finally, the fourth section evaluates the proposed system.

2. Related Work

With the growth of the web, data content has grown significantly and one of the challenges that face machines is the retrieval of related data that can be integrated accurately. Semantic retrieval is based on concepts rather than textual keywords. According to Ishkewy, Harb and Farahat [6], lexical ontology is the first step for automatic text analysis. Zoghby, Ahmed and Hamza[12] presented a survey about Arabic semantics in different ontological domains and showed that there are deficiencies in Arabic semantics. They highlighted the fact that the tools used do not provide support for Arabic language. Saad et al. [13] proposed an approach to generate an ontology from unstructured documents by using a group of techniques such as natural language processing, text mining and information extraction. The case study used the Quran to extract an Islamic concept. Sudeepthi, Anuradha and Babu [7], on the other hand, carried out a survey on several features of some of the best search engines such as Hakia, Kngine, Kosmix, Powerset, DuckDuckGo and Sensebot. The results of the survey found that search engines have the power to search and retrieve data. Sudeepthi et al. [14] compared search engines based on keywords and semantics, comparing Google as a keyword-based search engine and DuckDuckGo as a semantic-based search engine. This comparison included seven main points, which entailed the mechanism of their work and the presentation of the results. This comparison concluded that semantic-based search engines have higher benefits than traditional search engines. Various approaches have been reviewed in the construction of Islamic etiology. A set of results has concluded, for example, that most of the Islamic field's ontology is manually constructed, and the process of solving the concept extraction problem is semi-automatic and based on limited linguistic lexicons. Furthermore, it takes a considerable amount of time and does not cover all linguistic concepts [15]. Ontology in the Qur'anic field is considered one of the most important applications of Arabic ontology because it is considered as the main source of Arabic language. Ullah Khan et al. [16] proposed a method based on living creatures and birds in the Qur'an. They used a set of tools such as the Protégé ontology editor and SPARQL Query. The integration and drawing of the methodology was discussed and it was recommended that a Quranic WordNet should be built. The aim of Azman Ta’a et al.’s research [17] was to propose an approach to represent the Al-Quran knowledge based on the themes in Syammil Al-Quran Miracle. The result of the work was that a thematic approach is easier to understand than a systematic approach. This work was validated by domain experts. Tashtoush et al. [18] have focused on social relations in the Quran by using OWL and RDF. The results were extracted by SPARQ and DL queries as retrieval tools. The methodology consisted of two phases: the first entailed main subject and document gathering, and the second involved the development, design and implementation of the ontology. This work supported Arabic, English and ArabEzi and presented two layers: verse text and hidden meaning from the Tafseer of the Verse. Research conducted by Al-Yahya et al. [18] presented a model of Arabic lexicon depending on "Time " in the Quran. The lexicon was based on an ontology. There were eighteen classes in this paper: seven for general use and others for time nouns in the Quran. The time domain consisted of ten sub-classes depending on the componential analysis such as temperature, period, range, embodiment and distribution, to name but a few. The results indicate that the model can be used for building an Arabic Semantic Web.

3.The Proposed System Architecture

This system aims to propose a methodology to extract time noun concepts from the Nahj Al-Balagha domain ontology and retrieve the data with more accuracy according to the domain rather than terms. This paper is based on previous work carried out by H. M. A. Sahib [1] with expert help in the Nahj Al-Balagha involved throughout all steps. The core function of this work is to retrieve ontological information based on the time noun concepts because time is the dimension which is used for measuring events, thinking, feeling and actions [19]. For example, expectation is faith in what happens in the future. This system consists of three main stages as shown in Figure (1):

3.1 First stage: This stage focused on extracting the time nouns from Nahj Al-Balagha, indexing by VSM, obtaining approval by an expert and building the ontology. It included four sub-phases:

3.1.1 Word tokenization: Text is a set of characters. In order to process it, it must be broken down into meaningful sentences or tokens. NLTK Word Tokenize was used in this step.

3.1.2 Stop words removal: To extract words, stop words should be removed. The top ten number of stop words in Nahj Al-Balagha included " من " (from) " الى" (to)" كل" (all)" على" (over) " اذ" (if) and " في " (in)

3.1.3 Word weights: This is a discriminating measure which is used to specify a word's weight to categorize words and enhance retrieval performance. In this case, TF-IDF will be used.

3.1.4 Domain extraction: This phase was supervised by an expert to extract the knowledge of time nouns from Nahj Al-Balagha because some words have the same syntax but different meanings. For example, the word " اليوم"may refer to "day" or "doomsday". There are 37-time nouns in Nahj Al-Balagha, as shown in Table (1). They consist of three main classes:

- Physical time: Physical time is the traditional time that is used in daily life, such as years, months, weeks, days and hours.
- Psychological time: Psychological time is the time associated with activity, thinking, ambition and human actions, such as hope and action.
- Biological time: Biological time is the time associated with its effect on the body and it represents the stages of human growth, such as childhood, youth and adulthood.

3.1.5 Ontology Verification: In this stage, experts were consulted to review, verify and approve the vocabulary in the domain. This was necessary because sometimes there is a word that depends on the context of the sentence. For example, the word "الفناء" (cessation of being) "sometimes refers to psychological time (death) and sometimes to the attribute of "doomsday".

3.1.6 Indexing: To present the results according to how well documented and query matched they were, this arrangement was good for similarity computing. VSM will be used.

Table 1: Time Nouns in Nahj Al-Balagha

Biological Time	Psychological Time	Physical Time
العمر، نطفة ، الاعمار ، الشباب ، ظلمات الارحام ، علقة ، محاقا ، جنينا ، راضعا ، وليد ، يافعا ، السن	الموت، الامل ، السبقة، المدة ، الغاية ،الأوان، الفناء ، الابد ، المنايا ، اليوم ،الدهر ، سرمدا ، الاجل	اليوم، الصباح ، الغد ، النهار ، الليل ، الوقت ، امس ، الشهر ، اللحظة ، المساء ، الظلام ، الساعة

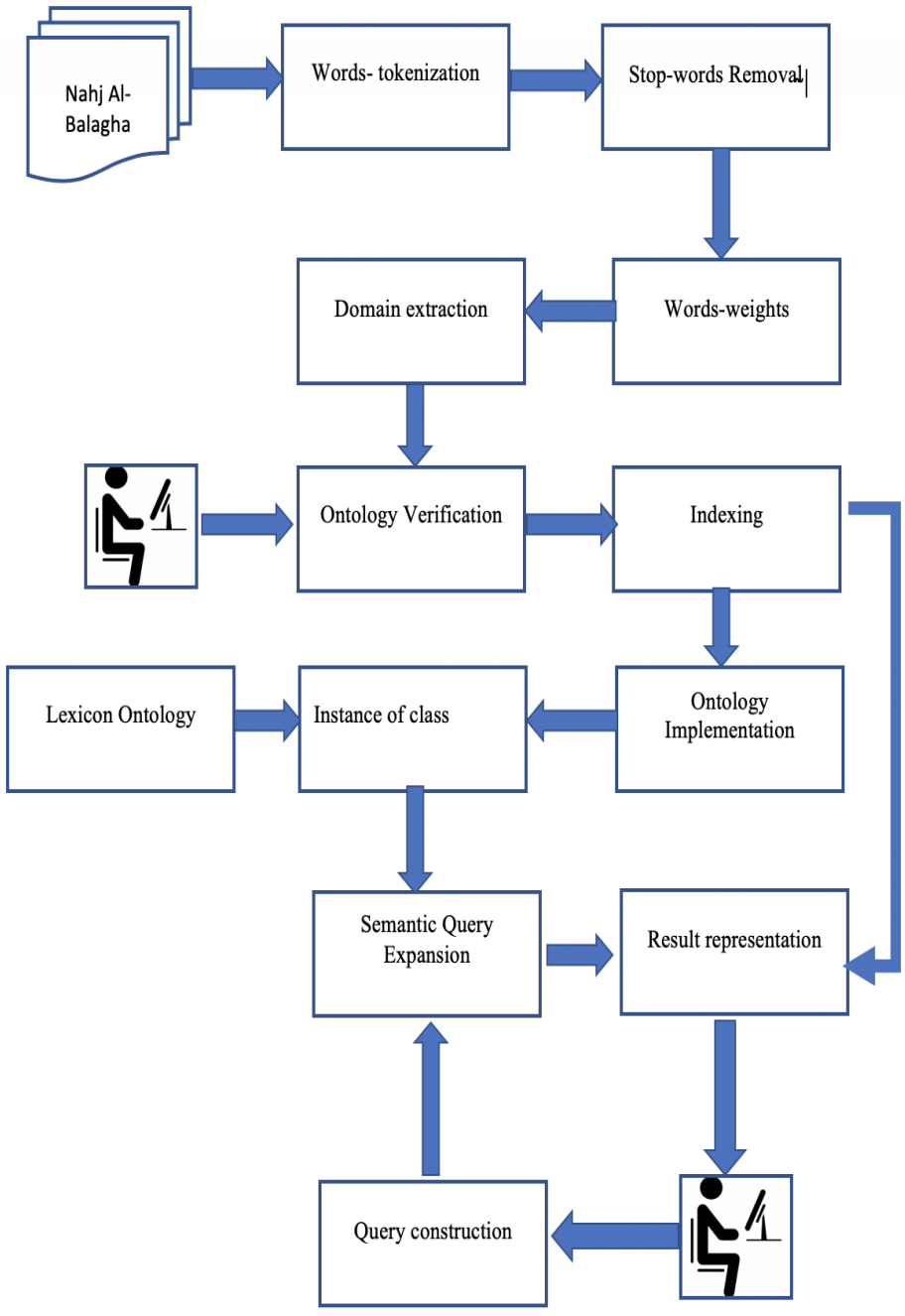

Figure 1: The Proposed System

3.1.7 Ontology Implementation: In this step, Web Ontology Language (OWL) was used to implement the time noun ontology. The protégé 4.3 has been used for this purpose, as shown in Figure (2), and carried out through a comprehensive examination. Figure (3) shows an example of the ontology represented in RDF.

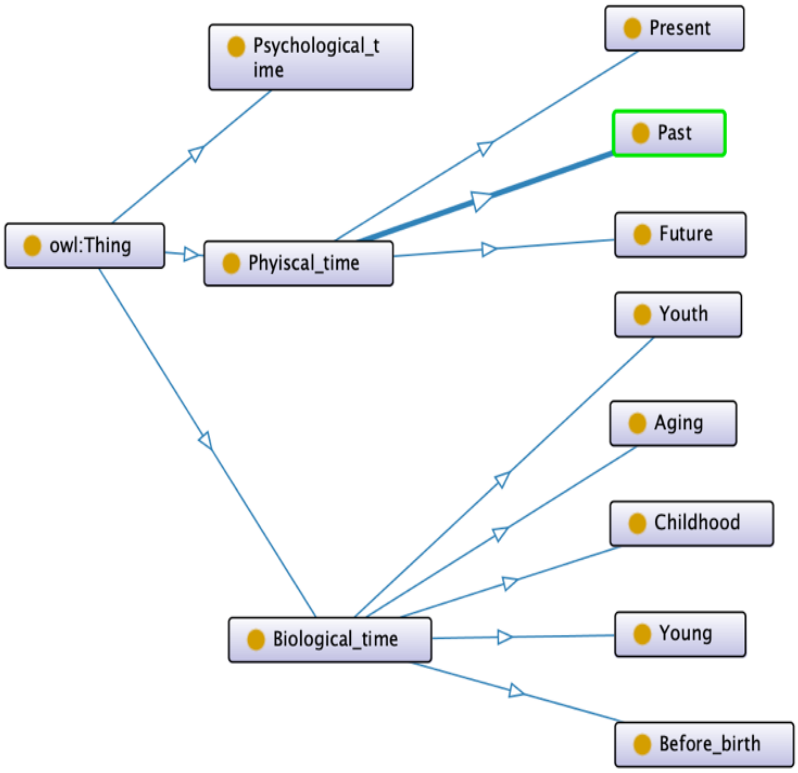

Figure 2: Time Nouns Ontological Classes

```
<owl:Timenoun rdfabout="&timont; الغد>
 <rdftype rdfresource="&imont;time_noun"/>
< imont:is_partOf rdfresource="& imont; Physical "/>
< imont:is _a rdfresource="& imont; Future "/>
</owl: Timenoun >
```

Figure 3: Sample of RDF

3.2 Second stage: This stage aims to link the OWL with the lexical ontology in [1] by using the OWL Linker plugin of Protégé.
3.3 Third stage: This stage consists of three steps:
3.3.1 Query construction: This step receives the query, word tokenization, stop removal and word weights.
3.3.2 Semantic query expansion: This step aims to expand the query based on the Nahj Al Balagha ontology. In order to retrieve the word query that was related to the concepts in the time nouns ontology, Owlready2 0.16 was used, which converted the RDF to a hash map.
3.3.3 Results representation: The core function of this step is based on cosine similarity to compute the matching between the query and the document vectors.

4. Experimental Results and Evaluation

The proposed system has been designed and implemented using Python3.6 and Spyder editor. Figure (4) shows the system user interface, which enables the user to input the query. The verification of semantic qualities was carried out by using the Protégé editor.

Query evaluation: This step evaluates the retrieved results and depends on the time nouns domain. A DL query was used for this purpose. The result page is shown in Figure (5).

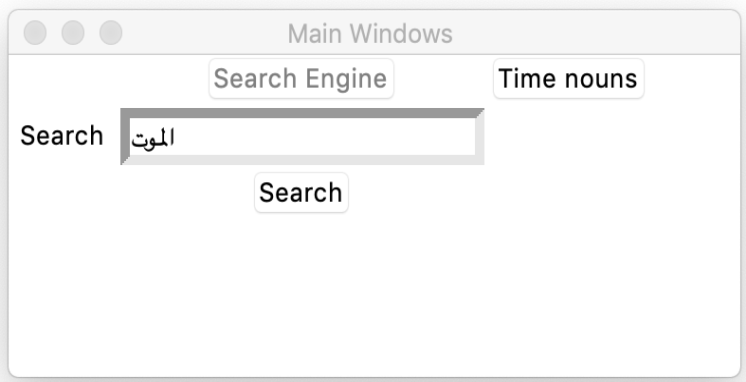

Figure 4: Main Page

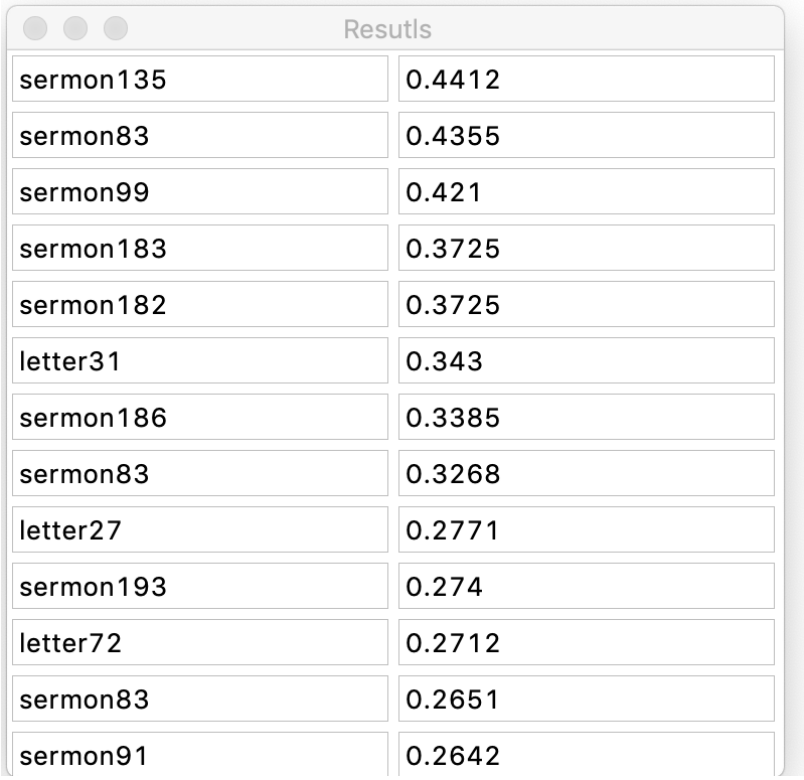

Figure 5: Result Page

The average recall and precision of ten executed queries in the proposed system was 95% and 89.7%, respectively, as shown in Table 2.

Table 2: The Average Recall and Precision

Object	Word search	Recall	Precision
الموت	الفناء	0.85	0.7
المستقبل	الغد	1	1
يوم القيامة	الساعة	0.94	0.8
مراحل عمر الانسان	الجنين	1	1
نهاية عمر الانسان	الغاية	1	0.9
الموت	الاجل	1	1
عمر الشباب	الشباب	1	1
الدجى	الظلام	0.8	0.75
أجزاء اليوم	النهار	1	1
يوم القيامة	اليوم	0.91	0.82

5. Conclusions and Recommendations for Future Work

This work presents an example of Arabic information retrieval based on domain ontology. This work is based on previous work by H. M. A. Sahib [1] and has used VSM to index the time nouns and implement them using OWL. The system consists of three stages: knowledge extraction and ontology implementation, link with the lexical ontology and finally, query and results representation. The results have shown that there is an improvement on the previous work in terms of the proportions of accuracy, retrieval and similarity. The conclusion of this work has indicated that the ontological domain is a good method of data representation because it enables relationships to be more clearly specified. Word weight and indexing can also be used to manage the relevant information finding. Future work will attempt to improve the ontology by linking it with other Arabic resources such as poems.

CONFLICT OF INTERESTS

There are no conflicts of interest.

References:

[1] H. M. A. Sahib, “Nahj Al-Balagha Semantic Search Engine (NSSE),” vol. 15, no. 3, pp. 40–49, 2017.
[2] E. Siow, “Analytics for the Internet of Things : A Survey,” vol. 1, no. 1, pp. 1–35, 2018.
[3] P. Ristoski, H. Paulheim, and H. Paulheim, “Semantic Web in Data Mining and Knowledge Discovery : A Comprehensive Survey,” 2016.
[4] J. Paralic and I. Kostial, “Ontology-based Information Retrieval,” 14th Int. Conf. Inf. Intell. Syst. (IIS 2003), pp. 23–28, 2003.
[5] M. Unni and K. Baskaran, “Overview of approaches to semantic web search,” Int. J. Comput. Sci. …, vol. 2, no. 2, pp. 345–349, 2011.
[6] H. Ishkewy, H. Harb, and H. Farahat, “Azhary: An Arabic Lexical Ontology,” Int. J. Web Semant. Technol., vol. 5, no. 4, pp. 71–82, 2014.
[7] G. Sudeepthi, G. Anuradha, and M. Babu, “A survey on semantic web search engine,” Int. J. Comput. Sci., vol. 9, no. 2, pp. 241–245, 2012.
[8] A. Shamsuzzaman Sadi et al., “Applying Ontological Modeling on Quranic "Nature" Domain,” 2016.
[9] A. Roshdi and A. Roohparvar, “Review : Information Retrieval Techniques and Applications,” vol. 3, no. 9, pp. 373–377, 2015.
[10] B. Abu-Salih, “Applying Vector Space Model (VSM) Techniques in Information Retrieval for Arabic Language,” 2018.
[11] J. Ramos, J. Eden, and R. Edu, “Using TF-IDF to Determine Word Relevance in Document Queries.”
[12] A. M. Al-Zoghby, A. S. E. Ahmed, and T. T. Hamza, “Arabic semantic web applications - A survey,” J. Emerg. Technol. Web Intell., vol. 5, no. 1, pp. 52–69, 2013.
[13] S. Saad, N. Salim, H. Zainal, and S. A. M. Noah, “A framework for Islamic knowledge via ontology representation,” Proc. - 2010 Int. Conf. Inf. Retr. Knowl. Manag. Explor. Invis. World, CAMP’10, no. April, pp. 310–314, 2010.
[14] A. Malve and P. P. M. Chawan, “A Comparative Study of Keyword and Semantic based Search Engine,” pp. 11156–11161, 2015.
[15] M. Aman, A. B. M. Said, S. J. A. Kadir, and B. Baharudin, “A review of studies on ontology development for Islamic knowledge domain,” J. Theor. Appl. Inf. Technol., vol. 95, no. 14, pp. 3303–3311, 2017.
[16] H. Ullah Khan, S. Muhammad Saqlain, M. Shoaib, and M. Sher, “Ontology Based Semantic Search in Holy Quran,” Int. J. Futur. Comput. Commun., vol. 2, no. 6, pp. 570–575, 2013.
[17] A. Azman Ta’a, Syuhada Zainal Abidin, Mohd Syazwan Abdullah and and M. A. Bashah B Mat Ali, “Al-Quran Themes Classification Using Ontology,” Icoci.Cms.Net.My, no. 074, pp. 383–389, 2013.
[18] M. Al-Yahya, H. Al-Khalifa, Maha Al-yahya, Hend Al-khalifa, and Nawal Al-Helwah, “an Ontological Model for Representing Semantic Lexicons : an Application on Time Nouns in the Holy Quran,” Arab. J. Sci. Eng., vol. 35, no. 2, pp. 21–35, 2010.
[19] D. Christopoulos, “A simple definition of Time,” no. June, 2014.

استرجاع المعلومات لنموذج قائم على الانطولوجيا: تطبيق لأسماء الزمن في نهج البلاغة

الخلاصة

نت دوراً اساسياً في الحياة من خلال كمية واهمية المعلومات التي يوفرها. حاليا، تعتبر إدارة البيانات وإيجاد المعلومات غير دقيقة وذلك لأنها تعتمد على شكل الكلمة وليس معناها. ان عملية تمثيل البيانات والوصول لها من اهم العوامل التي تساهم باسترجاع المعلومات والتغلب على مشكلة التشابه بين المستندات. توجد وسائل لقياس التشابه مختلفة تعمل وفقاً للوزن والفهرسة والمطابقة. الانطولوجيا هي البنية الأساسية لإدارة البيانات لأنها تستند الى معنى الكلمة والعلاقة بين الكلمات ومجال المعرفة.

يقدم هذا البحث اقتراحاً لنموذج نظام دلالي مبني على مجال معرفة محدد (في هذا البحث أسماء الزمن في نهج البلاغة) ويعتمد على المدخلات الدلالية عن طريقة فهرسة محرك البحث باستخدام Vector Space Model (VSM). الهدف من البحث هو تحسين المعلومات الدلالية المسترجعة عن طريق إنشاء استعلام يستند الى المطابقة والتشابه بين كلمات الاستعلام في النظام. هذا العمل مبني على عمل سابق [١].

تم تقييم النظام باستخدام معدل التشابه والدقة والاسترجاع لنتائج التجارب.

الكلمات الدالة: نهج البلاغة، استرجاع المعلومات، أسماء الزمن، النظام الدلالي، الانطولوجيا.