

Conceptual Search for Arabic Web Content

Aya M. Al-Zoghby¹ and Khaled Shaalan²

¹ Faculty of Computers and Information Systems, Mansoura University, Egypt
aya_el_zoghby@mans.edu.eg

² The British University in Dubai, UAE
khaled.shaalan@buid.ac.ae

Abstract. The main reason of adopting Semantic Web technology in information retrieval is to improve the retrieval performance. A semantic search-based system is characterized by locating web contents that are semantically related to the query's concepts rather than relying on the exact matching with keywords in queries. There is a growing interest in Arabic web content worldwide due to its importance for culture, political aspect, strategic location, and economics. Arabic is linguistically rich across all levels which makes the effective search of Arabic text a challenge. In the literature, researches that address searching the Arabic web content using semantic web technology are still insufficient compared to Arabic's actual importance as a language. In this research, we propose an Arabic semantic search approach that is applied on Arabic web content. This approach is based on the Vector Space Model (VSM), which has proved its success and many researches have been focused on improving its traditional version. Our approach uses the Universal WordNet to build a rich concept-space index instead of the traditional term-space index. This index is used for enabling a Semantic VSM capabilities. Moreover, we introduced a new incidence measurement to calculate the semantic significance degree of the concept in a document which fits with our model rather than the traditional term frequency. Furthermore, for the purpose of determining the semantic similarity of two vectors, we introduced a new formula for calculating the semantic weight of the concept. Because documents are indexed by their topics and classified semantically, we were able to search Arabic documents effectively. The experimental results in terms of Precision, Recall and F-measure have showed improvement in performance from 77%, 56%, and 63% to 71%, 96%, and 81%, respectively.

Keywords: Semantic Web (SW), Arabic Language, Arabic web content, Semantic Search, Vector Space Model (VSM), Universal Word Net (UWN), Wikipedia, Concept indexing.

1 Introduction

Search engines are still the most effective tools for finding information on the Web. Ambiguities in users' queries make searching the web content a challenge. Searching becomes more sophisticated when dealing with linguistically rich natural language

such as Arabic, which has a number of properties that makes it particularly difficult to handle by a computational system [1]. The use of terminological variations for the same concept, i.e. Synonymous terms, creates a many-to-one ambiguity. Whereas, the use of the same terminology for different concepts, i.e. Polysemous terms, creates a one-to-many ambiguity [2, 3]. It is desirable that search engines have three main features: accurately interpret the user's intension, handle the relevant knowledge from different information sources, and deliver the authentic and relevant results to each user individually [4, 5, 6]. Our goal is to address all these features in our proposed semantic-based search approach.

From the performance perspective, the traditional search engines are characterized by trading off a high-recall for low-precision. The reasons are that the results of these systems are sensitive to the input keywords, and the consequences of misinterpreting the synonymous and polysemous terminologies [7]. In other words, not only all relevant pages are retrieved, but also some irrelevant pages are retrieved as well, which impact the Precision. On the other hand, if some relevant pages are missing, this obviously leads to low Recall. One suggested solution is to use the Semantic Search Engines (SSEs) which rely on ontological concepts for indexing rather than lexical entries of standard lexicons that are commonly used by traditional search engines. Thus, SSEs aim to retrieve pages referring to specific concepts indicated by the query, rather than pages mentioning the input keywords, which will resolve the semantic ambiguity [8, 9]. The ontology should resolve the issue of semantic similarity of terminologies that comprise the keyword since they can be interpreted via the ontological representation of their related concept. Moreover, the SSEs can benefit from the inherent Generalization/Specialization properties of the ontological hierarchy. When a semantic search engine fails to find any relevant documents, it might suggest generic answers. On the other hand, if too many answers are retrieved, the search engine might suggest more specialized answers [8, 9]. As a comparison with traditional search engines, the returned results will be more relevant, and those missing documents will also be retrieved, which means higher accuracy and better robustness [10].

The success and advances of Semantic Web technology with Latin languages can also be investigated in order to bridge the gap in other underdeveloped languages, such as Arabic. Statistics from WWW indicated that there are an increasing number of Arabic textual contents available on electronic media, such as Web pages, blogs, emails, and text messages, which make the task of searching Arabic text relevant. However, there are linguistic issues and characteristics facing the development of Semantic Web systems in order to be able to effectively search the Arabic web content. This is due to the richness of the Arabic morphology and the sophistication of its syntax. Moreover, the highly ambiguous nature of the language makes keyword-based approaches of the traditional search engines inappropriate [11, 12]. For example, the optional vowelization in modern written Arabic text gives different meaning to the same lexical form. Another example is the polysemous or multi-meaning of words which arise from terminologies that share the same orthography but differ in meaning [13, 14]. Nevertheless, there are specific issues that are related to the handling of Arabic text in computational systems. One of them is the differences in Arabic script encoding of web content [15]. Another issue is the availability of Arabic resources,

such as corpora and gazetteers [1]. Existing Arabic resources are available but at significant expense. In many cases these resources are limited or not suitable for the desired task. So, researchers often develop their own resources, which require significant efforts in data collection, human annotation, and verification.

This research proposes enhancements to the semantic VSM-based search approach for Arabic Information Retrieval application, and the like. VSM is a common information retrieval model for textual documents that has demonstrated its capability to represent documents into a computer interpretable form. Many researches have been successful in improving its traditional version [16]. In our proposed approach, we build a concept-space from which we construct a VSM index. This model has enabled us to represent documents by semantic vectors, in which the highest weights are assigned to the most representative concept. This representation permits a semantic classification within and across documents, and thus the semantic search abilities reflected in its Precision and Recall values can be obtained. The construction of the concept-space is derived from semantic relationships presented at the Universal WordNet (UWN). UWN is an automatically constructed cross-lingual lexical knowledge base from the multilingual WordNet. For more than 1,500,000 words in more than 200 languages, UWN provides a corresponding list of meanings and shows how they are semantically related [17]. The main reasons for choosing the UWN are: a) it is widely a standard, b) it is very powerful in supporting semantic analysis, and c) it has the ability to provide the meaning of missing Arabic terms from their corresponding translations from other languages. Moreover, the UWN would facilitate dealing with Arabic dialects, which is an active research topic. The proposed approach is used to develop a system that is applied on a full dump of the Arabic Wikipedia. The evaluation of the system's retrieval effectiveness using the constructed concept-space index resulted in noticeable improvements in the performance in terms of Precision and Recall as compared to the traditional syntactic term-space baseline. The experimental results showed an overall enhancement of the F-Measure score from 63% to 81% due to employing the semantic conceptual indexing.

The rest of this paper is organized as follows: Section 2 presents the architecture of the model and the implementation aspects. The experimental results are discussed in details at Section 3. Finally, the paper is concluded at Section 4.

2 System Architecture

This section describes the architecture of the proposed conceptual VSM-based search system and its components. The overall architecture is depicted in Fig. 1.

The '*Document Acquisition*' module acquires documents from the web which we use as the knowledge source.

The '*NE Extraction*' module extracts Arabic Named Entities from the acquired documents. The extracted "Named Entity" (NE) covers not only proper names but also temporal and numerical expressions, such as monetary amounts and other types of units [1]. ANEE [18], a popular rule-based named entity recognition tool that is integrated with the GATE development environment, is used for extracting the NEs. ANEE is capable of recognizing Arabic NEs of types: Person, Location, Organization,

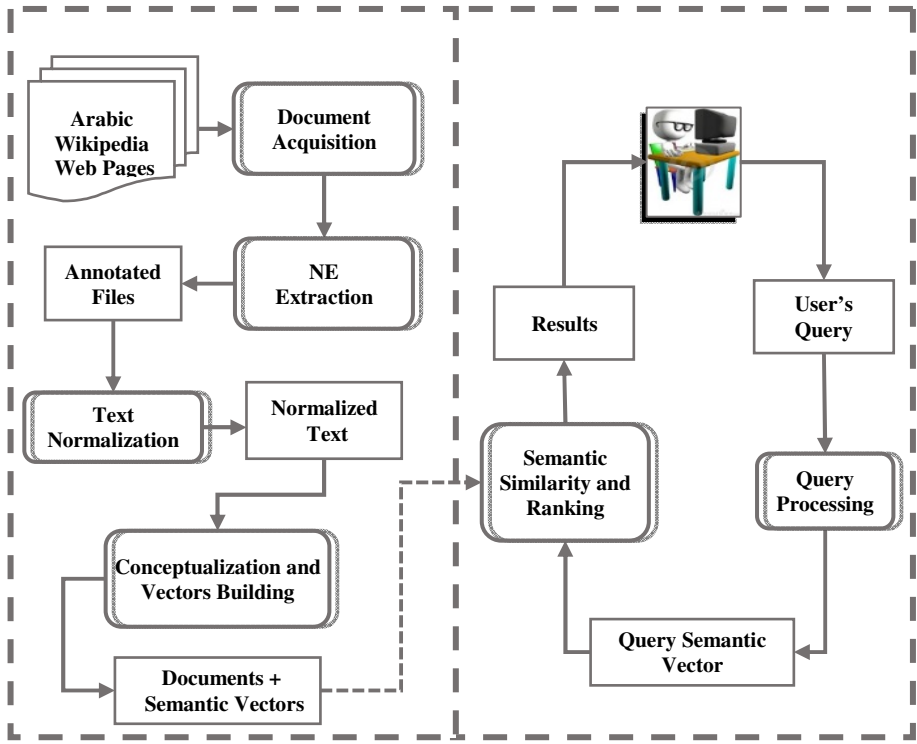


Fig. 1. The System Architecture

Measurement, Time, Number, Percent, and File name. A number of scholars have used the ANEE tool in their research studies on Arabic NER, including Maynard et al. (2002), Elsebai et al. (2009), Elsebai et al. (2011), and Abdallah et al. (2012) [1]. In our research, we also used ANEE tool to annotate the Arabic text with Person, Location, and Organization NEs. In the literature, these types are so called ENAMEX. Afterwards, these named entities are expanded, via the UWN, with their aliases¹.

The 'Text Normalization' module performs four preprocessing steps. First, it splits the full dump into separate Wikipedia articles; each one has a designated title. Then, the text is normalized by filtering out non-Arabic letters². Next, the RDI-Swift indexer³ is applied on the cleaned texts to perform indexing. This indexing would facilitate eliminating all inflected forms of the Arabic stop-word⁴ by searching for it within the text. Finally, stop-words are eliminated; excluding those that are part of NEs.

¹ For example: **Yasser Arafat** (Arabic: ياسر عرفات, *Yāsir `Arafāt*) and his aliases **Abu Ammar** (Arabic: أبو عمار, *'Abū `Ammār*).

² This is specific to the current version of the RDI-Swift indexer that is available to us which requires normalizing the text using the RDI morphological analyzer.

³ http://www.rdi-eg.com/technologies/arabic_nlp.htm;
<http://rdi-eg.com/Demo/SwiftSearchEngine/Default.aspx>;
<http://www.rdi-eg.com/Downloads/Swift2WhitePaper.doc>

⁴ For example: ...من, منه, منها, منهما, لمن...

The '*Conceptualization and Vectors Building*' module, which is the core of the proposed system, performs progressive development of the Arabic semantic vectors. The vectors are generated using three indexing methods that differ in their incidence measure, see Table 1. The three indices gradually developed: Morphological Term-Space (MTS), Semantic Term-Space (STS), and Concept-Space (CS) are used to evaluate the performance of the proposed model.

An entry of the MTS dictionary considers all inflected forms of the term. An entry of the STS dictionary is a semantic expansion of an MTS entry using UWN. In other words, each term in the space is represented by a set of related inflected forms of their semantic expansions. An entry of the CS dictionary consists of a main keyword that represents the concept and expansions of all encapsulated terms along with their morphological and semantic expansions. For the purpose of incidence measurement, we developed five different measurements that progress gradually from the *Term Frequency (tf)* to the *Concept Semantic Significance Degree (CSSD)*. These measurements show the improvement we made in system's performance.

Query processing is handled by the '*Query Preprocessing*' module. Firstly, the query is normalized by removing stop words and non-Arabic words or letters. The query is then semantically expanded using the UWN. Finally, the '*Semantic Similarity and Ranking*' module calculates the similarity between user's query and documents-space. Then, if it finds similar results, the ranking is applied according to similarity.

Table 1. Indexing and Incidence Measurement methods






Indexing Method	Incidence Measurement	
	Measure	Description
Morphological Term-Space (MTS)	Term's Morphological Frequency (TMF)	It is the count of the morphological inflections occurrences of the term.
Semantically Expanded Term-Space (STS)	Term's Semantic Frequency (TSF)	It is the frequency of the morphological derivations of the term itself and its semantic expansions as well.
	Term's Semantic Significance Degree (TSSD)	In this measurement, the variation of the association degree of each expansion of the term is taken into consideration.
Concept-Space (CS)	Concept's Semantic Frequency (CSF)	This type is equivalent to that of TSF but in terms of Concepts instead of terms, so it will be computed as the count of matching occurrences of all concepts' ingredients.
	Concept's Semantic Significance Degree (CSSD)	It is the equivalent of the TSSD, but in terms of concepts.




3 Experimental Analysis

3.1 The Conceptualization Process

From the AWDS⁵, a terms-space of 391686 terms is extracted, 31200 of which are NEs. Each extracted term is enclosed with its set of derivations that occurred in the document space, with an average of 16 root-based derivations each. The term-space is then expanded via UWN with 793069 in total; redundant expansions forms about 12% and takes the form of phrases. The shrinking algorithms are applied on the term-space, excluding the NEs set, to generate a concepts-space, which results in of 223502. As a result, the count of the concepts-space entries is shrunk to be just 62 % of that of the morphological-space. The ampler concept is defined by 66 merged terms while the narrower one is just of two. Some of the terms are never merged; most of them are NEs⁶, out of vocabulary, or misspelled words. This, first level of, shrinking generated 299203 groups of terms from the 360486 terms. The second shrinking level then condensed them into just 223502 groups, each of which defines a distinct concept. This leads to the increment of the representation power of each item in the space, since the average of items weights is increased as shown by the last two columns in Table 2. Note that the weight average of V3 and V5 are lower than those of V2 and V4, respectively. However, it is noticeable that the weights of the semantic indexing (STS), are higher than those of the morphological indexing (MTS). Moreover, the weights of the conceptual indexing (CS) are higher than the weights of the

Table 2. The overall results of the Conceptualization Process

Exp.	Indexing	#Entries	Expanding Type	Average of Expansions / Entry		Freq. Avg	df Avg	Weight Average		
				#	Classification					
V1	MTS	360486	Morphologically	4	Morphological Inflections	0.514	2.02 ≈ 2	1.02		
V2	STS		Semantically	11	Inflections	4	0.704	2.6 ≈ 3	2.11	
					Synonyms	2				
					Sub-Classes	1				
					Super-Classes	2				
V3				Has-Instances	1			1.65		
				Instance-Of	1					
V4	CS	223502	Conceptually	33	Concept of 3 STS entries, each with its semantic expansion	0.7076	3.8 ≈ 4	2.83		
V5								2.32		

 Morphological Based Weights Average
 Semantic Based Weights Average
 Conceptual Based Weights Average

⁵ AWDS stands for Arabic Wikipedia Documents Space.

⁶ Note that we excluded the already extracted NEs, however, other NEs are still impeded in the space and couldn't be recognized by the ANEE.

semantic indexing (STS). The important observation is that these results demonstrate the efficiency of the conceptual indexing in distinguishing documents by means of corresponding weights. The conceptualization process is fully described at [19].

3.1.1 Retrieval Capability.

The F-measure is very popular for calculating the performance of the retrieval system based on the harmonic mean of Precision and Recall scores, which we used to evaluate the retrieval accuracy of the proposed system. It is defined as follows:

$$F_Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (1)$$

Where, *Precision* is the fraction of retrieved instances that are relevant, and *Recall* is the fraction of relevant instances that are retrieved.

In order to conduct evaluation experiments, we have to determine beforehand the correct answer of the test query that will be applied on a dataset from AWDS. As an example, consider the query: مصادر الطاقة /*msadr altaqh/ (Energy Resources)*. Then, we need to determine the correct relevant documents, i.e. the actual Recall.

For the sake of evaluation, a gold standard test set of the first 200 documents from the AWDS is extracted, and classified according to relevancy to the test query: مصادر الطاقة /*msadr altaqh/ (Energy Resources)*. 77 documents, out of the 200 documents, are syntactically/semantically/conceptually relevant, whereas the rest are irrelevant. These documents are then used as the document-space of the search process. The query is preprocessed and semantically expanded using the UWN.

The performance of the system in terms of Precision, Recall, and F-Measure are shown in Table 3. As shown in the table, five experiments were conducted, i.e. V1 through V5 experiments, which uses MTS, STS, and CS indexing. Notice that both V2 and V3, have achieved the same values of the Precision, Recall, and F-Measure. Likewise, both V4 and V5 have achieved same scores. The reason is that these scores are not sensitive to the incidence measurement which differentiate between these experiments. However, the incidence measurement has an effective role in ranking results which we discuss in details in the subsequent section. From the results shown at Table 3, we observe the following:

First, the highest Precision is achieved by the V1 experiment, which uses the MTS index. They were just 12 irrelevant documents of a total of 123. It only contains the documents that have either the word مصادر /*msadr/ (Resources)* or the word طاقة /*taqh/ (Energy)* in an irrelevant context. However, despite its high precision, this experiment's *Recall* is the lowest, since some of the semantically and conceptually relevant documents are not considered; they are 32 out of the total 77 relevant documents.

Second, the *Precision* is degraded while using the STS indexing with experiments that use V2 and V3. This is mainly due to some issues in recognizing *multiword*

Table 3. The Retrieval Capabilities of the indecies: MTS, STS, and CS

Experiment	Indexing Method	Precision	Recall	F-Measure
V1	MTS	76.92%	55.50 %	63.04%
V2, V3	STS	68.57%	66.70%	67.62%
V4, V5	CS	71%	96%	81.62%

expressions, or phrases, by the RDI indexing system⁷. For example, the word طاقة/taqh/ (energy) is expanded to the expression طاقة نووية/taqh nwwyh/ (Nuclear energy), which is handled as two individual words. Consequently, documents related only to the word نووي/nwwyh/ (nuclear or atomic) should have been considered irrelevant. In this case, when V2 and V3 experiments are used, the count of the irrelevant documents becomes 22, which affects the Precision⁸ to be lower than that obtained by V1. The Recall, on the other hand, is increased since the missing relevant documents are decreased from 32 achieved by V1 experiment to 24 achieved by V2 and V3 experiments. This improvement is due to the semantic expansions that included the words such as كهرباء/khrba' (Electricity), and the phrases such as طاقة نووية/taqh nwwyh/ (Nuclear energy).

Third, as for the V4 and V5 experiments, where the conceptual index CS is exploited, the Precision increased again, but not to the extent of the V1 experiment, since the issue of recognizing multiword expressions or phrases is still affecting it. This time, the individual words of an expression such as خلايا شمسية/khlaya shmsyh/ (Solar Cells) caused the retrieval of 6 additional irrelevant documents related to the term خلايا/khlaya/ (Cells). The Recall, however, is increased to the extent that makes the F-Measure of these experiments the highest. This means that the conceptual retrieval caught most of the relevant documents. Only three documents are escaped. They concern the terms: إشعاع /esh'ea'e/ (Radiation), and حركة المياه/hrkh almyah/ (Water flow), since they are not covered in the concept's definition itself. The results showed an enhancement of the F-Measure value to be 81.62% using our proposed semantic conceptual indexing as compared to 63.04% achieved when using the standard indexing method.











3.2 Ranking Accuracy


In addition to measuring the proposed system's performance in terms of the retrieval capability, the ranking accuracy are also evaluated, see Table 4. The evaluation of





⁷ The set of all expansions of a term is collected as a set of individual words; separated by a space. The RDI indexing system is used to search for their inflected forms within the documents-space. This approach caused a problem in dealing with the multiword expansions, since the sequence of words forming a multiword expression is not recognized.

⁸ In order to resolve this problem, each expansion is indexed such that the multiword expression is discriminated. However, this solution is favorable if the indexing algorithm is not efficient in using the memory space.

Table 4. The Ranking accuracy of each Incidence Measurement

Exp.	Incidence Measurement	W Average	Ranking Distance Average ⁹
V1	TMF	1.028692 	4.625359 
V2	TSF	2.113503 	4.402382 
V3	TSSD	1.653511 	4.397337 
V4	CSF	2.830778 	4.165544 
V5	CSSD	2.322311 	4.154426 



-  Enhancement caused by using STS indexing
-  Extra Error caused by using TSF Incidence Measurement
-  Extra Error caused by using CSF Incidence Measurement
-  Enhancement caused by CS Indexing

document ranking is aimed at measuring the accuracy of assigning the congruent weight that exactly represent the association degree of each entry of the index with a document in the space. This should indicate how the *Incidence Measurement* factor is directly affecting the capability of the accurate ranking.

Wherever the values of the weights averages of V2 and V4 experiments are greater than those of V3 and V5 experiments respectively, the ranking results show that V3 experiments is better than V2 experiment, and V5 experiments is the best. The reason is the extra error ratios caused by using the *frequency Incidence Measurements*, *TSF* and *CSF*, instead of the *Semantic Significance Degree Incidence Measurements*, *TSSD* and *CSSD*. However, these extra ratios are truncated via exploiting the *TSSD* and *CSSD* at the V3 and V5 experiments. This is directly reflected by the ranking efficiency of these experiments. Still, the V4 experiment gives better ranking that the V3 experiments since it is based on the *conceptual indexing CS*, even it suffers from the extra error ratio caused by using the *CSF Incidence Measurements*.

The *Ranking Accuracy* is measured by calculating the *Distance Average* between the experiment's ordering and that of a human specialist. *Distance Average* is defined as following:

$$Distance-Average(V_i) =$$

$$\frac{\sum_{j=1}^n \left| \frac{1}{Standard_Rank_j} - \frac{1}{Experimental_Rank_j} \right|}{n} \tag{2}$$

⁹ The smallest the distance the closest to the correct ranking.

Where,

$n = \#$ retriever documents as a result to the user's query.

Standard_Rank_j = The standard rank of the document #j.

Experimental_Rank_j = The rank of document #j at V_i .

The closest ranking is obtained from the V5 experiment, which assures the capability of its features to rank the retrieved result more accurately.

4 Conclusion and Future Work

This study sheds the light on the inefficiency in the handling the Arabic language semantically, which in turn results in an unqualified Arabic Web content for being a suitable environment of the Semantic Web technology implementation. This inefficiency may be ascribed to the sophistication of the Arabic language that makes it complex to the extent that hinders its treatment electronically. Nevertheless, this should not stop more effective efforts for achieving the best possible solutions that enable the Arabic language and its users to take the advantages of the new electronic technologies generally, and the Semantic Web particularly. This might not be achieved unless the Arabic electronic infrastructure is adequately built. Briefly, we have to construct all of the necessary Arabic resources, such as the Ontologies, Gazetteers, WordNet, as well as the complementary utilities such as morphological analyzers and named entity extractors. Likewise, the support of the Arabic script in the Semantic Web tools such as OWL/RDF editors must be taken into consideration. If that infrastructure is achieved and is sufficiently strong, the development of wider applications will be better facilitated, and the obtained results will be more reliable and trustworthy.

In an attempt to take a step in that long pathway, we proposed an Arabic semantic search system that is based on the Vector Space Model. The Vector Space Model is one of the most common information retrieval models for textual documents due to its ability to represent documents into a computer interpretable form. However, as it is syntactically indexed, its sensitivity to keywords reduces its retrieval efficiency. In order to improve its effectiveness, the proposed system has extracted a concept-space dictionary, using the UWN ontology, in order to be used as a semantic index of the VSM search system instead of the traditionally used term-space. The proposed system enables a conceptual representation of the document space, which in turn permits the semantic classification of them and thus obtaining the semantic search benefits. Moreover, we introduced a new incidence measurement to calculate the semantic significance degree of the concept in a document instead of the traditional term frequency. Furthermore, we introduce a new formula for calculating the semantic weight of the concept to be used in determining the semantic similarity of two vectors. The system's experimental results showed an enhancement of the F-measure value to 81.62% using the semantic conceptual indexing instead of 63.04% using the standard syntactic one.

Still, the model's implementation suffers from some limitations. Consequently, the presented results will certainly be improved if those limitations are overcome. Therefore,

as a future work, we have to solve the ambiguity problem by discriminating the meaning contextually. Also, we may work on refining the processing of the multiword expression expansions. That will improve the results noticeably since 12%¹⁰ of the semantic expansions are in the form of multiword expressions. We also will try to expand the named entities in order to use them in the environment of the linked data. Moreover, the improvement of the Arabic knowledge representation in the UWN will help to overcome its limitations that directly affects the search results. Another open research area is to solve the problems of the Arabic language morphological analysis in order to prevent the consequent errors occurred in the indexing process, and hence, the construction of the search dictionary. We also may try to use Google Translation API with the UWN in order to find results for these terms that have results in languages other than Arabic.

References

1. Khaled, S.: A Survey of Arabic Named Entity Recognition and Classification. *Computational Linguistics* 40(2), 469–510 (2014), doi:10.1162/COLIA00178.
2. Saleh, L.M.B., Al-Khalifa, H.S.: AraTation: An Arabic Semantic Annotation Tool. In: *The 11th International Conference on Information Integration and Web-based Applications & Services* (2009)
3. Tazit, N., Bouyakhf, E.H., Sabri, S., Yousfi, A., Bouzouba, K.: Semantic internet search engine with focus on Arabic language. In: *The 1st International Symposium on Computers and Arabic Language & Exhibition* © KACST & SCS (2007)
4. Cardoso, J.: *Semantic Web services: Theory, tools, and applications*. IGI Global (March 30, 2007) ISBN-13: 978-1599040455
5. Hepp, M., De Leenheer, P., de Moor, A.: *Ontology management: Semantic web, semantic web services, and business applications*. Springer (2008) ISBN: 978-0-387-69889-1
6. Kashyap, V., Bussler, C., Moran, M.: *The Semantic Web: Semantics for Data and Services on the Web (Data-Centric Systems and Applications)*. Springer (August 15, 2008) ISBN-13: 978-3540764519
7. Panigrahi, S., Biswas, S.: Next Generation Semantic Web and Its Application. *IJCSI International Journal of Computer Science Issues* 8(2) (March 2011)
8. Unni, M., Baskaran, K.: Overview of Approaches to Semantic Web Search. *International Journal of Computer Science and Communication* 2(2), 345–349 (2011)
9. Renteria-Agualimpia, W., López-Pellicer, F.J., Muro-Medrano, P.R., Noguera-Iso, J., Zarezaga-Soria, F.J.: Exploring the Advances in Semantic Search Engines. In: de Leon F. de Carvalho, A.P., Rodríguez-González, S., De Paz Santana, J.F., Rodríguez, J.M.C. (eds.) *Distrib. Computing & Artif. Intell., AISC*, vol. 79, pp. 613–620. Springer, Heidelberg (2010)
10. Kassim, J.M., Rahmany, M.: Introduction to Semantic Search Engine. In: *International Conference on Electrical Engineering and Informatics, ICEEI 2009* (2009)
11. Habash, N.Y.: Introduction to Arabic Natural Language Processing. *Association for Computational Linguistics* 30 (August 2010) ISBN 978-1-59829-795-9
12. Al-Zoghby, A.M., Eldin, A., Hamza, T.T.: Arabic Semantic Web Applications: A Survey. *Journal of Emerging Technologies in Web Intelligence*, 52–69 (2013)

¹⁰ This percentage is computed from the retrieved UWN expanding results.

13. Elkateb, S., Black, W., Vossen, P., Farwell, D., Pease, A., Fellbaum, C.: Arabic WordNet and the challenges of Arabic: The Challenge of Arabic for NLP/MT. In: International Conference at the British Computer Society (October 23, 2006)
14. Al-Khalifa, H.S., Al-Wabil, A.S.: The Arabic Language and the Semantic Web: Challenges and Opportunities. In: International Symposium on Computers and the Arabic Language, Riyadh, Saudi Arabia (November 2007)
15. Omar, D.: Arabic Ontology and Semantic Web. al-Mu'tamar al-duwali lil-lughah. al-lughat al-'Arabiyah bayn al-inqirad wa al-tatawwur, tahaddiyat wa tawqi'at, Jakarta, Indonesia July 22-25 (2010),
-الأنطولوجيا العربية و الويب الدلالي: المؤتمر الدولي للغة العربية (اللغة العربية بين الانقراض والتطور-
التحديات والتوقعات) ، جاكرتا، إندونيسيا: 22- 25 يوليو 2010.
16. Zhao, Y.-H., Shi, X.-F.: Shi: The Application of Vector Space Model in the Information Retrieval System. Software Engineering and Knowledge Engineering: Theory and Practice 162, 43–49 (2012)
17. de Melo, G., Weikum, G.: UWN: A Large Multilingual Lexical Knowledge Base. In: Annual Meeting of the Association of Computational Linguistics (2012)
18. Oudah, M.M., Shaalan, K.: A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012) (2012)
19. Al-Zoghby, A.M., Eldin Ahmed, A.S., Hamza, T.: Utilizing Conceptual Indexing to Enhance the Effectiveness of Vector Space Model. International Journal of Information Technology and Computer Science (IJITCS) 5(11) (2013) (October 2013) ISSN: 2074-9007