# Semantic Search for Arabic

**Aya M. Al-Zoghby[1], and Khaled Shaalan[2]**

[1] Faculty of Computers and Information Systems, Mansoura University, Egypt

aya_el_zoghby@mans.edua.eg

[2] The British University in Dubai, UAE

khaled.shaalan@buid.ac.ae

**Abstract**

There is a growing interest in Arabic web content worldwide due to its importance for culture, religion, and economics. In the literature, researches that address searching 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). It uses the Universal WordNet ontology to build a rich concept-space index instead of the traditional term-space index. The proposed index is used for enhancing the capability of the semantic-based VSM. Moreover, the approach introduces a new incidence measurement to calculate the semantic significance degree of the document's concepts which is more suitable than the traditional term frequency measure. Furthermore, a novel method for calculating the semantic weight of the concept is introduced in order to determine the semantic similarity of two vectors. As a proof of concept, a system is applied on a full dump of the Arabic Wikipedia. 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.

## Introduction

The success of semantic web technology with Latin languages can also be used to bridge the gap in other underdeveloped languages such as Arabic (Al-Zoghby, Ahmed, & Hamza, Arabic Semantic Web Applications: A Survey, 2013; Shaalan, June 2014). However, there are linguistic challenges 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 (Hahash 2010; Al-Khalifa and Al-Wabil 2007; Elkateb et al. 2006).

This research proposes an enhanced semantic VSM-based search approach for Arabic Information Retrieval application and the like. 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. The construction of the concept-space is derived from semantic relationships obtained from the Universal WordNet (UWN).

As a proof of concept, a system is applied on a full dump of the Arabic Wikipedia. . 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.

The next sections will present the main aspects of the proposed approach, the architecture of the proposed system, experimental results, and the concluding remarks.

## The Main Aspects of the Proposed Approach

### Conceptualization and Concept Space Indexing

In semantic web notations, terms are used to explain concepts and relationships among them, i.e. the concept is identified by the meaning indicated by its related terms (Unni and Baskaran 2011; Renteria-Agualimpia et al. 2010). In the traditional syntactic VSM, terms are used separately in order to form a *term-space*, which does not consider the definition of their corresponding concepts. As a result, the generated vectors will stray in the space between the set of terms. More important, the measure of the most frequent term in the document might not indicate the most expressive one. Well-structured documents usually have one or more central concept(s) describing the document, indicated by a group(s) of related-terms, see Table 1. Therefore, the document's vector is accurately directed if its highest weight(s) is assigned to its central concept(s).

*Table 1: Concept Representation by Different Terms*

| Concept | | | Group of related Terms | | | | | |
|---|---|---|---|---|---|---|---|---|
| كتابة | / ktabh/ | Writing | كتابة | /ktabh/ | Writing | تسجيل | /tsjyl/ | Registration |
| | | | تدوين | /tdwyn/ | Notation | يكتب | /yktb/ | Writes |
| | | | تحرير | /thryr/ | Editing | يحرر | /yhrr/ | Edits |
| | | | توثيق | /twthyq/ | Documentation | … | … | … |

Hence, we decided to replace the traditional term-space by a *concept-space* that is semantically more expressive. This has the effect of overcoming the dispersion of terms and accumulating the weights of the individual terms to get effective weights for the corresponding descriptive concepts. The details of this process are fully described Al-Zoghby et al. (2013).

## A New Incidence Measurement: Semantic Significance Degree (SSD)

The weight of term *t* in document *d* refers to the term's capability of distinguishing the document. Traditionally, it is defined in terms of its frequency $tf_{d,}$. However, when the semantic-based conceptual search approach is adopted, this equation will no longer be accurate for the following three reasons. First, the term needs to be matched against all its inflected forms along with their word senses that occur in *d* rather than the input or primitive (normalized) form. Therefore, not all matches would have the same degree. Second, the semantic expansion of terms needs to be taken into consideration. The semantic expansions of terms can be classified into five different types: *Synonyms, SubClasses,* HasInstances *(Specialization expansions)*, *SuperClasses,* and *InstancesOf (Generalization expansions)***.** It is obvious that the expansions of *Synonyms* are semantically closer to the original term than either its generalized or specialized expansions. Moreover, the expansions of *SubClasses* and *SuperClasses* are worthier than those of *HasInstances* or *InstancesOf*, since the former represent classes that encompass a set of related concepts while the latter types are instances referring to specific related terms. Third, possible expansions of each type have different confidences because not all expansions are relevant to the original term at the same degree or weight.

The aforementioned reasons discourage the use of term frequency in semantic search. Hence, we introduce a new *incidence measurement* of the term *t*, called the *Semantic Significance Degree (SSD)*, and consequently, the concept *c*. The *SSD* is computed in terms of a new factor, called *association factor,* defined for each expansion of the term. It is a multiplication of two parameters: the *distance* of the expansion's type, and the *confidence* of the expansion itself. We defined the distance of the expansion's type as a multiplying coefficient, which is heuristically determined by the semantic closeness of the expansion. The set of *Synonyms* has the same degree, or an exact matching to the original term. Therefore, we decided to set its distance value to *1*. On the other hand, the *Sub/Super classes* are more related than *Instances* since they indicate a generic perspective of the meaning. Therefore, we decide to set the value of the distance of *Sub/Super* classes to *0.75* and the value of the distance of *Has-Instances/Instances* of classes to *0.5*. The value of the confidences is directly obtained from the UWN.

Formally, let $AF_i$ denotes the *Association Factor* of the expansion *i*.

$$AF_i = conf_i * dist_i \quad (1)$$

The *Term's Semantic Significance Degree (TSSD)* of the term *t* at document *d* is defined as:

$$TSSD_{d,t} = \sum_{i=1}^{n} tf_i * AF_i \quad (2)$$

Where *n* is the number of all expansions of the term *t* from all types, $tf_i$ is the frequency of expansion *i* instances occurred in document *d*, $conf_i$ is the confidence of expansion *i* in *UWN*, and $dist_i$ is the distance of expansion *i*.

Let Sem-$IDF_t$ *denotes the* semantic inverse document frequency of the term *t*. It is defined as follows:

$$\text{Sem-IDF}_t = \log \frac{|D|}{|\{e' \in d' \mid e' \in te \,\&\, d' \in D\}|} \quad (3)$$

Where, *te* is a set of all semantic expansions of the term *t*, *e* is certain expansion in the document *d*, and *D* is the entire documents-space. Thus, the semantic weight of term *t* in document *d* is:

$$\text{Sem-w}_{d,t} = TSSD_{d,t} * \text{Sem-IDF}_t \quad (4)$$

In terms of concepts, let $st_j$ denotes the set of the semantic expansions of the term *j* and $C_i$ denotes the concept *i* that is defined by a set of *m* semantically expanded terms. The expansion union of $C_i$ is defined by merging the sets of all semantic expansions of all definitive terms as follows:

$$C_i = \textit{Expansions-Merge}\,(\{st_1, st_2, \ldots, st_j, \ldots, st_m\}) \quad (5)$$

The Concept's Semantic Significance Degree (CSSD) is defined as follows:

$$CSSD_{d,c_i} = \sum_{j=1}^{m} TSSD_{d,st_j} \quad (6)$$

Finally, let Sem-$IDF_c$ denotes the Semantic Inverse Document Frequency of the concept c is defined as:

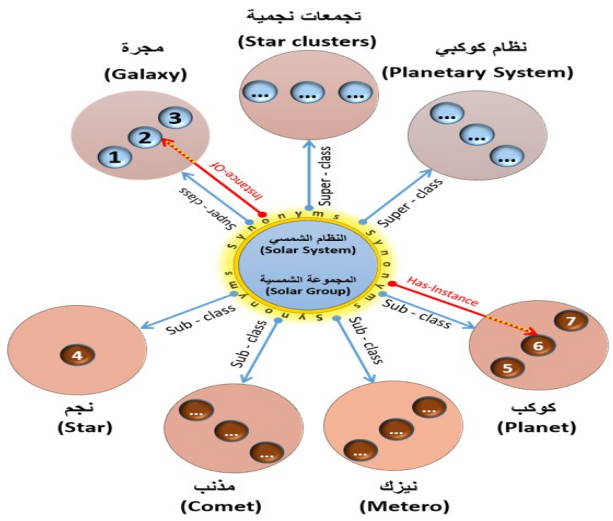
$$\text{Sem-IDF}_c = \log \frac{|D|}{|\{e' \in d' \mid e' \in ce \,\&\, d' \in D\}|} \quad (7)$$

Where, *ce* is the set of all semantic expansions of the concept *c*. The *Semantic Weight* of each concept in the new Concept Space *Sem-*$w_{d,c}$ is defined as:

$$\text{Sem-w}_{d,c} = CSSD_{\mathbf{d,c_i}} * \text{Sem-IDF}_c \quad (8)$$

### An Illustrative Example

Fig. 1 shows an example that clarifies the relationships surrounding a concept that is represented by the two synonyms: *النظام الشمسي/alnzam alshmsy/ (Solar System)*, and المجموعة الشمسية */almjmw'eh alshmsyh /(Solar Group)*. The *المجموعة الشمسية /almjmw'eh alshmsyh/ (Solar Group)* is closer to its synonym *النظام الشمسي /alnzam alshmsy/ (Solar System)* than its super-class *مجرة /mjrh/(Galaxy)* and its sub-class *كوكب /kwkb/(Planet)*. However, these sub/super classes are closer to the original term *الشمسية/alshmsyh / (Solar)* than its instances such as *الشمس /alshms/ (Sun)* and *الأرض/alard/(Earth)*.




| Node # | Node name | | |
|---|---|---|---|
| 1 | المجرة الإهليلجية | /almjrh alehlyljyh/ | (Elliptical galaxy) |
| 2 | مجرة المرأة المسلسلة | /mjrh almrah almslslh / | (Andromeda) |
| 3 | مجرة درب التبانه | /mjrh drb altbanh / | (Milky Way galaxy) |
| 4 | الشمس | /alshms/ | (Sun) |
| 5 | عطارد | /'etard/ | (Mercury) |
| 6 | الارض | /alard/ | (Earth) |
| 7 | زحل | /zhl/ | (Saturn) |

*Figure 1: An illustrative example of the different kinds of matching between expansions. The numbers in the figure refers to the Node # in the associated table.*

Assume that we have 20 documents in the space; five documents have occurrences of the مجموعة شمسية */mjmw'eh shmsyh / (Solar Group)* relationships. In this case the *Document Frequency (df)* is 5 and the *Inverse Document Frequency (IDF)* is Log(20/5) = 0.6. Table 2 presents three methods to calculate the weight of *Solar Group* term. Method 1 is the traditional way to compute the weights. Method 2 computes the weights using the *UWN Confidence (conf.)*. Method 3 computes the weights using the proposed *Association Factor (AF)*. The weight 9.435 is more accurate than 12.6 and 15, since it reflects the real significance of each semantic expansion, and thus the overall weight of the concept's representation in the document.

## Semantic Similarity

As far as semantic of concepts are considered, we need to accurately match each expanded concept in the input query with each concept in the space. However, not all matches between all instances of each semantic expansion type have the same matching consistency or degree. For example, one of the synonyms of the first query word, say *qw1s1,* may match one of the *super-classes* of a concept in the space $c_i$. This query word might also match one of the *synonyms* of another concept $c_j$. In this example, the second match is stronger than the first. Therefore, for efficiency reasons, each case has to be handled separately; otherwise, we might get unexpected results. As there is a possibility that weak matches might take weights similar to stronger matches, we derived formulas (9) and (10) for constructing entries of the *Matching-Sensitive* vector of the input query.

Let $n$ denotes the number of expansions that are matched between the query and the concepts in the space, and $m_x$ denotes the frequency of each match. The Semantic Significance Degree of the concept $c$ in query $Q$ *is:*

$$QSSD_{Q,c} = \sum_{x=1}^{n} m_x * cAF_x * qwAF_x \quad (9)$$

Where, $cAF_x$ is the *Association Factor* of the concept $c$ of the match #$x$, and $qwAF_x$ is the *Association Factor* of the query word of the match #$x$. The semantic weight of the concept $c$ in query $Q$ is:

$$\text{Sem-w}_{Q,c} = QSSD_{Q,c} * \text{Sem-IDF}_c \quad (10)$$

Thus, the semantic similarity between query $Q$ and document $d_i$ is calculated in terms of the semantic weights of the concept $c_j$ in both the query $Q$ and the document $d_i$ by the equation:

*Table 1: The weight of the semantic expansion of the term مجموعة شمسية /mjmw'eh shmsyh / (Solar Group)*[1]

| Semantic Relationship | Semantic Expansion | | | Confidence conf. | Distance dist. | Association Factor AF | Incidence Measurement | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Method 1 | Method 2 | Method 3 |
| | | | | | | | *tf* | *tf * conf.* | $TSSD_{d,t}$ |
| **Synonyms** | مجموعة شمسية | /mjmw'eh shmsyh/ | (Solar Group) | 0.9 | 1 | 0.9 | 5 | 4.5 | 4.5 |
| | نظام شمسي | /nzam shmsy / | (Solar System) | 0.7 | | 0.7 | 3 | 2.1 | 2.1 |
| **Super-Classes** | مجرة | /mjrh/ | (Galaxy) | 0.9 | 0.75 | 0.675 | 2 | 1.8 | 1.35 |
| | نظام كوكبي | /nzam kwkby/ | (Planetary System) | 0.7 | | 0.525 | 1 | 0.7 | 0.525 |
| | تجمعات نجمية | /tjm'eat njmyh/ | (Star Clusters) | 0.5 | | 0.35 | 1 | 0.5 | 0.35 |
| **Sub-Classes** | كوكب | /kwkb/ | (Planet) | 0.9 | 0.75 | 0.675 | 3 | 2.7 | 2.025 |
| | نيزك | /nyzk/ | (Meteor) | 0.8 | | 0.6 | 1 | 0.8 | 0.6 |
| | مذنب | /mdnb/ | (Comet) | 0.7 | | 0.525 | 0 | 0 | 0 |
| | نجم | /njm/ | (Star) | 0.9 | | 0.675 | 1 | 0.9 | .0.675 |
| **Has-Instance** | الشمس | /alshms/ | (Sun) | 0.9 | 0.5 | 0.45 | 2 | 1.8 | 0.9 |
| | الأرض | /alard/ | (Earth) | 0.9 | | 0.45 | 4 | 3.4 | 1.8 |
| **Instances – of** | مجرة درب التبانة | /mjrh drb altbanh/ | (Milky Way galaxy) | 0.9 | 0.5 | 0.45 | 2 | 1.8 | 0.9 |
| $CSSD_{d,c}$ | | | | | | | **25** | **21** | **15.725** |
| **Sem-$W_{d,c}$** | | | | | | | **0.6*25=15** | **0.6*21=12.6** | **0.6*15.725 = 9.435** |

[1] Note that the experiments are executed using Arabic text without vocalization symbols. Moreover, the UWN is designed to retrieve results in Arabic script.

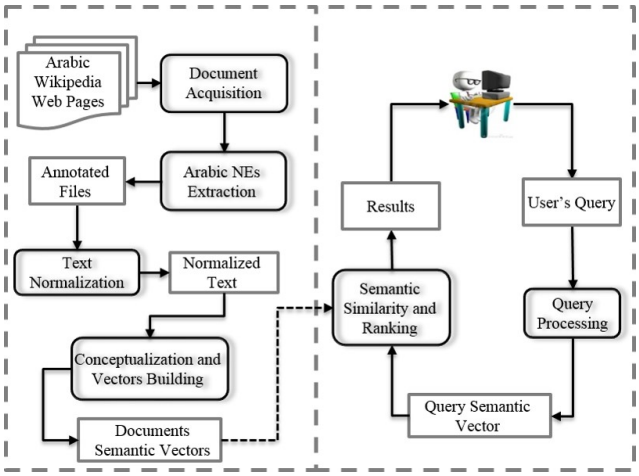



*Figure 2: System Architecture*

$$Sem\_Sim(d_i, Q) = \frac{\sum_{j=1}^{con} Sem - w_{Q,c_j} * Sem - w_{d_i,c_j}}{\sqrt{\sum_{j=1}^{con} Sem - w_{Q,c_j}^2} * \sqrt{\sum_{j=1}^{con} Sem - w_{d_i,c_j}^2}} \quad (11)$$

Where *con* is the count of concepts in the space.

## System Architecture

The overall architecture of the proposed system is depicted in Fig. 2. The *Document Acquisition* module acquires documents from the web which we use as a knowledge source. The *Arabic NEs Extraction* module extracts Arabic Named Entities (NEs) from the acquired documents using ANEE tool (Oudah and Shaalan 2012). Afterwards, these NEs are expanded, via the UWN, with their aliases[2]. The *Text Normalization* module splits the full dump of Wikipedia documents into separate articles. Then, the text is normalized by filtering out non-Arabic letters[3]. Finally, stop-words are eliminated. The *Conceptualization and Vectors Building* module is the core of the proposed system. The vectors are generated using three different indexing methods with different incidence measurements, see Table 3. The three resultant indices are gradually developed: *Morphological Term-Space (MTS), Semantic Term-Space (STS), and Concept-Space (CS),* which we use in evaluating the performance of the proposed approach. An entry of the *MTS* index considers all inflected forms of the term. Whereas, the entry of the *STS* index is generated from the semantic expansion of an *MTS* entry using *UWN*. An entry of the *CS* index consists of a main term that represents the concept and expansions of all encapsulated terms along with their morphological and semantic expansions. Table 4 gives examples that progressively show how an entry is generated from MTS to STS and from STS to CS, respectively. The *Query Preprocessing* module normalizes the query and semantically expands it. Finally, the *Semantic Similarity and Ranking* module uses equation 11 to calculate the similarity between user's query and documents-space. Then, the ranking is applied according to similarity.

## An Example of Utilization and Testing

### The Conceptualization Process

The *shrinking algorithms* (Al-Zoghby et al., 2013), are applied on the term-space to generate a concept-space, which results in a concept space of 223502 concepts. As a result, the number of the concepts-space entries shrank to 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 terms cannot be merged; most of them are NEs, out of vocabulary words, and misspelled words. 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 column of Table 3. It is noticeable that the weights of the semantic indexing (STS), are higher than those of the morphological indexing (MTS). Likewise, the weights of the conceptual indexing

*Table 2: The Indexing and Incidence Measurement methods, and the overall results of the Conceptualization Process.*

| Experiment | Indexing Method | Incidence Measurement | | #Entries | Type of Expanding | Average of Expansions / Entry | | | *Freq.* Ave. | *df* Ave. | *W* Ave. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Measure | Description | | | # | Classification | | | | |
| **V1** | MTS | Term's Morphological Frequency (TMF) | It is the count of the morphological inflections occurrences of the term. | 360486 | Morphologically | 4 | Morphological Inflections | | 0.5 | 2.1 | 1.1 |
| **V2** | STS | 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. | | Semantically | 11 | Inflections | 4 | 0.7 | 2.6 | 1.6 |
| | | | | | | | Synonyms | 2 | | | |
| | | | | | | | Sub-Classes | 1 | | | |
| | | | | | | | Super-Classes | 2 | | | |
| | | | | | | | Has-Instances | 1 | | | |
| | | | | | | | Instance-Of | 1 | | | |
| **V3** | CS | Concept's Semantic Significance Degree (CSSD) | It is the equivalent of the TSSD, but in terms of concepts. | 223502 | Conceptually | 33 | Concept of 3 STS entries | | 0.7 | 3.8 | 2.3 |

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

[3] 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.

*Table 3: An example that progressively shows how an STS entry is generated from MTS and an CS entry is generated from STS*

| MTS Entry | STS Entry | CS Entry |
|---|---|---|
| ربح /rbh/ (Profit) | ربح /rbh/ (Profit) | ربح /rbh/ (Profit) |
| مربح/mrbh/(Profitable) | مربح /mrbh/ (Profitable) | مربح /mrbh/ (Profitable) |
| رابح/rabh/ (Winner) | رابح /rabh/ (Winner) | رابح /rabh/ (Winner) |
| | فوز /fwz/ (Winning) | فوز /fwz/ (Winning) |
| | فائز /fa'ez/ (Winner) | فائز /fa'ez/ (Winner) |
| | نصر /nsr/ (Triumph) | نصر /nsr/ (Triumph) |
| | ينتصر /yntsr/ (Conquers) | ينتصر /yntsr/ (Conquers) |
| | فائدة /fa'edh/ (Interest) | فائدة /fa'edh/ (Interest) |
| | استفادته /astfadth/(His Benefit) | استفادته /astfadth/(His Benefit) |
| | يستفيد /ystfyd/ (Benefits) | يستفيد /ystfyd/ (Benefits) |
| | كسب /ksb/ (Gain) | كسب /ksb/ (Gain) |
| | ينكسب /ytksb/ (Earn money) | ينكسب /ytksb/ (Earn money) |
| | | يغلب /yghlb/ (Overcomes) |
| | | يتغلب /ytghlb/ (Beat) |
| | | ينجح /ynjh/ (To succeed) |
| | | نجاح /njah/ (Success) |
| | | فلاح /flah/ (prosperity) |
| | | فالح /falh/ (Prosperous) |
| | | ثروة /thrwh/ (Riches) |
| | | ثراء /thra'/ (Wealthier) |
| | | أثرياء /athrya'/ (Wealthy) |
| | | رفاهية /rfahyh/ (Well-being) |
| | | ترف /trf/ (Luxury) |
| | | مترف /yghlb/ (Luxuriant) |

(CS) are higher than the weights of the semantic indexing (STS). The important observation is that the results have demonstrated the efficiency of the conceptual indexing in distinguishing documents by means of corresponding weights.

## Retrieval Capability

For the sake of evaluation, a gold standard test set of the first 200 documents from the *AWDS* is extracted, and classified according to their 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. The semantic expansion and the conceptualization of the query words are presented in Table 5. Three experiments were conducted that uses MTS, STS, and CS indexing methods. The performance of the system is shown in Table 6. From this table, we observe the following. The first experiment uses the MTS index. The reason of its high Precision is that the set of unrelated, yet retrieved, documents in this case is the smallest. 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, see for example Table 7. 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.

In the second experiment, using STS indexing lead to degrading the Precision. This is mainly due to issues in recognizing *multiword expressions, or phrases,* which is caused by some limitations imposed by the RDI[7]-Swift indexer. 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 /nwwy/*

*Table 4: Expansion of the query مصادر الطاقة /msadr altaqh/ (Energy Resources)*

| Expansion Type | User's Query Keywords | Semantic Expansion | | |
|---|---|---|---|---|
| Semantic Expansion | مصادر/msadr/ (Resources) | مصادر | /msadr/ | (Resources) |
| | الطاقة /altaqh/ (Energy) | طاقة | /taqh/ | (Energy) |
| | | وقود | /wqwd/ | (Fuel) |
| | | كهرباء | /khrba'/ | (Electricity) |
| | | طاقة نووية | /taqh nwwyh/ | (Nuclear Energy) |
| Conceptualization | Synonyms | تيار | /tyar/ | (Electricity) |
| | | طاقة كهربائية | /taqh khrba'eyh/ | (Electrical energy) |
| | | وقود | /wqwd/ | (fuel) |
| | | طاقة نووية | /taqh nwwyh/ | (Nuclear Energy) |
| | | شعاع الشمس | /sh'ea'e alshms/ | (Sunbeam) |
| | | وقود أحفوري | /wqwd ahfwry/ | (Fossil fuels) |
| | | خلايا شمسية | /khlaya shmsyh/ | (Solar Cells) |
| | | حرارة الشمس | /hrarh alshms/ | (Heat of the sun) |
| | | بترول | /btrwl/ | (Petroleum ) |
| | | نفط | /nft/ | (Petrol) |
| | | أشعة الكاثود | /ash'eh alkathwd/ | (Cathode ray) |
| | | الزيت الخام | /alzyt alkham/ | (Crude Oil) |
| | | نووي | /nwwy/ | (Nuclear) |
| | Sub-Classes | كهرباء | /khrba'/ | (Electricity) |
| | | ديزل | /dyzl/ | (Diesel) |
| | | بنزين | /bnzyn/ | (Gasoline) |
| | | سولار | /swlar/ | (Solar) |
| | | كيروسين | /kyrwsyn/ | (Kerosene) |
| | | فحم | /fhm/ | (Coal) |
| | | زيت البترول | /zyt albtrwl/ | (Petroleum oil) |
| | Super-Classes | زيت | /zyt/ | (Oil) |
| | | موجه كهرومغناطيسية | /mwjh khrwmghnat-ysyh/ | (Electromagnetic Wave ) |

*Table 6: The Retrieval Capabilities of the indices: MTS, STS, and CS*

| Experiment | Indexing | Precision | Recall | F-Measure |
|---|---|---|---|---|
| **V1** | MTS | **77%** | 56 % | 63.04% |
| **V2** | STS | 69% | 67% | 68% |
| **V3** | CS | 71% | **96%** | **81.62%** |

*Table 7: Irrelevant context to the words: مصادر /msadr/ (Resources) and طاقة /taqh/ (Energy)*

| طاقة /taqh/ (Energy) | مصادر /msadr/ (Resources) |
|---|---|
| مدينة طاقة /mdynh taqh/ (Energy City)[4] | مصادر المعلومات /msadr alm'elwmat/ (The information resources) |
| | مصادرة الحقوق /msadrh alhqwq/ (Confiscation of rights) |
| حصن طاقه /hsn taqh/ (Taqah Castle)[5] | المصادر المفتوحة /almsadr almftwhh/ (The open sources) |
| طاق بستان /taq bstan/ (Taq Bostan)[6]- | مصادر التشريع / msadr altshry'e/ (The sources of legislation) |

[4] Energy City Qatar (ECQ) is the first ever integrated energy hub in the GCC and MENA regions; http://www.energycity.com

[5] Taqah Castle (حصن طاقة) is the most popular castle to visit in the region of Dhofar in the South of Oman; http://www.omantripper.com/taqah-castle-dhofar/

[6] Taq wa San or Taq-e Bostan(Kurdish: تاق وه سان, "arch of Stone"),(Persian: طاق بستان, "arch of Bostan") is a site with a series of large rock relief from the era of Sassanid Empire of Persia, the Iranian dynasty which ruled western Asia from 226 to 650 AD; http://en.wikipedia.org/wiki/Taq_Bostan

[7] 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

*(nuclear or atomic)* should have been considered irrelevant, some examples are shown under Group A in Table 8. In this case, the number of the irrelevant documents of the second experiment becomes 22, which changes its *Precision* to be lower than that of the first experiment.

The *Recall,* on the other hand, is increased since the missing relevant documents are decreased from 32 that are achieved by the first experiment to 24 at the second experiment. This improvement is achieved 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 third experiment, where the conceptual index CS is adopted, the Precision is increased again, but not to the extent of that of the first 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)* cause the retrieval of 6 additional irrelevant documents related to the term خلايا/khlaya/ (Cells), some examples are also shown under Group B in Table 8. 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 81.62% using our proposed semantic conceptual indexing as compared to the baseline 63.04% that is achieved when using the standard indexing method.

*Table 8: Sample of concepts related to words but not to phrases*

| | Phrase | | | # Related Doc.s |
|---|---|---|---|---|
| A | أحماض نووية | /ahmad nwwyh/ | (Nucleic acid) | 4 |
| | برنامج نووي | /brnamj nwwy/ | (Nuclear program) | 4 |
| | الإمام النووي | /alemam alnwwy/ | (Imam Al-Nawawi)[9] | 2 |
| B | خلايا نباتيه | /khlaya nbatyh/ | (Vegetable cells) | 4 |
| | خلايا جذعية | /khlaya jd'eyh/ | (Stem Cells) | 2 |

## Conclusion and Future Work

The inefficiency in handling the Arabic Language semantically 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.

In an attempt to take a step in that long pathway, we proposed an Arabic semantic search approach that is based on VSM. VSM 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 is extracting a concept-space index, using the UWN ontology, to be used as a semantic index of 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 introduced a new formula for calculating the semantic weight of the concept in order to determine the semantic similarity of two vectors. The system's experimental results showed an enhancement of the F-measure value to be 81% using the semantic conceptual indexing instead of 63% using the standard syntactic one.

As a future work, we plan to solve the ambiguity problem by discriminating the meaning contextually. Also, we need to address refining the processing of the multiword expression expansions. That should improve the results noticeably since they form 12%[8] of the semantic expansions.

## References


Al-Khalifa, H. S.; and Al-Wabil, A. S. 2007. The Arabic Language and the Semantic Web: Challenges and Opportunities. The International Symposium on Computers and Arabic Language & Exhibition (1): 27-63. Riyadh, Saudi Arabia.

Al-Zoghby, A. M.; Ahmed, A. S.; and Hamza, T. H. 2013. Arabic Semantic Web Applications: A Survey. Journal of Emerging Technologies in Web Intelligence 5(1): 52-69.

Al-Zoghby, A. M.; Ahmed, A. S.; and Hamza, T. H. 2013. Utilizing Conceptual Indexing to Enhance the Effectiveness of Vector Space Model. International Journal of Information Technology and Computer Science (IJITCS) 5(11): 1-12.

Elkateb, S.; Black, W.; Vossen, P.; Farwel, D.; Pease, A.; and Fellbaum, C. 2006. Arabic WordNet and the Challenges of Arabic. In proceedings of Arabic NLP/MT Conference/ London,U.K,

Hahash, N. Y. 2010. Introduction to Arabic Natural Language Processing. Association for Computational Linguistics.

Oudah, M.; and Shaalan, K. 2012. A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach. Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), 2159-2176. Mumbai, india.

Renteria-Agualimpia, W.; López-Pellicer, F. J.; Muro-Medrano, P. R.; Nogueras-Iso, J.; and Zarazaga-Soria1, F. J. 2010. Exploring the Advances in Semantic Search Engines. International Symposium on Distributed Computing and Artificial Intelligence 2010 (DCAI'2010). Advances in Intelligent and Soft-Computing. Springer, vol. 79, 613-620.

Shaalan, K. 2014. A Survey of Arabic Named Entity Recognition and Classification. Computational Linguistics 40 (2): 469-510. USA: MIT Press.

Unni, M.; and Baskaran, K. 2011. Overview of Approaches to Semantic Web Search. International Journal of Computer Science and Communication (IJCSC) 2: 345 – 349.


[8] This percentage is calculated from the UWN expansion results.
[9] http://en.wikipedia.org/wiki/Al-Nawawi