I.J. Information Technology and Computer Science, **2013, 11, 1-12**
Published Online October 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2013.11.01

Utilizing Conceptual Indexing to Enhance the Effectiveness of Vector Space Model

Aya M. Al-Zoghby
Computer Science Department, Faculty of Computers & Information, Mansoura University, Egypt
E-mail: elzoghby.aya@gmail.com

Ahmed Sharaf Eldin Ahmed
Information Systems Department, Faculty of Computers & Information, Helwan University, Egypt
E-mail: profase2000@yahoo.com

Taher T. Hamza
Computer Science Department, Faculty of Computers & Information Science, Mansoura University, Egypt
E-mail: taher_hamza@yahoo.com

Abstract— One of the main purposes of the semantic Web is to improve the retrieval performance of search systems. Unlike keyword based search systems, the semantic search systems aim to discover pages related to the query's concepts rather than merely collecting all pages instantiating its keywords. To that end, the concepts must be defined to be used as a semantic index instead of the traditional lexical one. In fact, The Arabic language is still far from being semantically searchable. Therefore, this paper proposed a model that exploits the Universal Word Net ontology for producing an Arabic Concepts-Space to be used as the index of Semantic Vector Space Model. The Vector Space Model is one of the most common information retrieval models due to its capability of expressing the documents' structure. However, like all keyword-based search systems, its sensitivity to the query's keywords reduces its retrieval effectiveness. The proposed model allows the VSM to represent Arabic documents by their topic, and thus classify them semantically. This, consequently, enhances the retrieval effectiveness of the search system.

Index Terms— Semantic Web; Semantic Concepts; UWN; Vector Space Model; Arabic Language

I. Introduction

Search engines are the most indispensable tools used in navigating the information published on the Web. However, having the Web as the biggest global unstructured database complicates its machine understandability and processability. This, in turn, makes the ability to accurately acquire the desired information extremely difficult. Moreover, the query words may sometimes be ambiguous since different people may use different terminologies for the same concept, Synonymous, while, on the other hand, they may use the same words for different concepts, Polysemous [1]-[2]. Thus, most of the search engines face the problem of capturing the true purport of the user's query. Therefore, the main task of the search engines is to accurately interpret the users' needs, handle the relevant knowledge from different information sources, and deliver the authentic and relevant results to each user individually [3]-[4]-[5].

In terms of recall/precision, the traditional search engines can be described as: with high-recall, they have low precision. This is mainly caused due to the sensitivity of their results to the keywords, and the misinterpretation of the synonymous and Polysemous [6]. Therefore, even if all relevant pages are retrieved, some irrelevant documents are also retrieved, which is affecting the precision. On the other hand, if some of the relevant pages are missed, this leads to low recall. Therefore, the alternative is the Semantic Search Engines (SSEs) that use the Ontological concepts for indexing rather than the standard lexicons used in the traditional search engines. Thus, the semantic search engines aim to get pages referring to specific concepts, rather than collecting all pages that just mentioned the query's keywords; which may already be ambiguous [7]-[8]. This way, the problem of expressing the same semantic concept using different terminologies can be resolved since these terminologies will be recognizable via the Ontological representation of that concept. Moreover, the semantic search engines can utilize the Generalization / Specialization properties of Ontologies hierarchy. Therefore, if it fails to find any relevant documents, it may indicate more general answers, and if too many answers are retrieved, the search engine may suggest some specializations [7]-[8]. This way, the returned results will be more relevant, and those missed will be retrieved, which means higher recall with more precision [9].

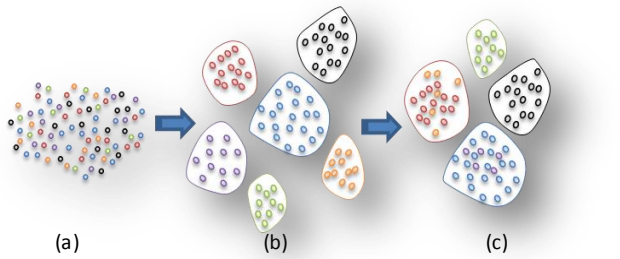
Unfortunately, however, the Arabic language is still not fully supported through SSEs [10]. Although Swoogle [1], Hakia [2], SenseBot [3], and DeepDyve [4] are among the top SSEs, they have low to no support of Arabic. As Arabic Web sites are increasing constantly, search systems that handle the semantics of Arabic Language come to be essential.

The Vector Space Model (VSM) is one of the most common information retrieval models for text documents searching due to its ability to represent documents into a computer interpretable form. Thus, VSM has a high degree of success and many researchers have focused on improving its traditional version [11]. In terms of VSM, the weight of term *t* in document *d* refers to its capability of distinguishing that document. The most frequent term in *d*, and least frequent in the entire document-space is the one most capable of distinguishing d. Sometimes, however, this may not be the case since the actual distinguishing concept c is defined via a set of terms scattered throughout the document in low frequency each.

This paper proposed a model for constructing an Arabic concept-space to be used as a VSM index. That enables the VSM to represent Arabic web documents by semantic vectors, in which the highest weights are assigned to the most representative concept. That permits a semantic classification of the Arabic web documents, and thus the semantic search abilities reflected in its precision and recall 0values can be obtained. The construction of the concept-space is based on the semantic relationships presented at the Universal WordNet (UWN) [5]. UWN is an automatically constructed multilingual lexical knowledge base based on WordNet. For over 1,500,000 words in over 200 languages, UWN provides a corresponding list of meanings and shows how such meanings are semantically related [12]. The evaluation of VSCAS system's retrieval effectiveness, a VSM search system, presented at [13], using the proposed concept-space index shows a noticeable enhancement against its performance using the traditional term-space.

The rest of this paper is organized as follows: Section II briefly describes how the use of concepts improves the performance of the VSM. Section III states the features of the proposed model in terms of the semantic expansion and then the concept-space generation. The next section represents the architecture of the model and its implementation details. The experimental results and discussion are detailed at Section V. Finally, the paper is concluded at Section VI.

1 swoogle.umbc.edu/
2 www.hakia.com
3 www.SenseBot.com
4 www.deepdyve.com
5 http://www.mpi-inf.mpg.de/yago-naga/uwn/downloads.html



(a) Spread Terms – Space
(b) Clustered Terms into Concepts-Space (Level 1)
(c) Shrieked Concepts-Space (Level 2)

Fig. 1: Term-space to Concept-space

II. Concepts and Semantic VSM

In Semantic Web, the terms are used to explain concepts and relationships, i.e., the concepts are identified by the meaning shared by the related terms [14]-[15]. When these terms are used separately to establish a term-space, as in the case of the traditional syntactic VSM, the definition of their corresponding concepts will be lost through the term-space, fig.1 (a), and thus the generated vectors will stray in the space between them, fig.2 (a) and (b). Moreover, the most frequent term in the document may not be the most expressive one. The documents normally have a central concept, indicated by a group of related-terms, which together can describe the document better. Therefore, the document's vector is accurately directed if its highest weight is assigned to its central concept. That can only be achieved if the concepts can be defined, and then used as a semantic index of the VSM. For that end, the term-space has to be compressed into its equivalent concept-space. That eliminates the dispersion and accumulates the weights of the separate terms to get effective weights for the corresponding descriptive concepts.

Table 1: Term-space vs. Concept-space VSM

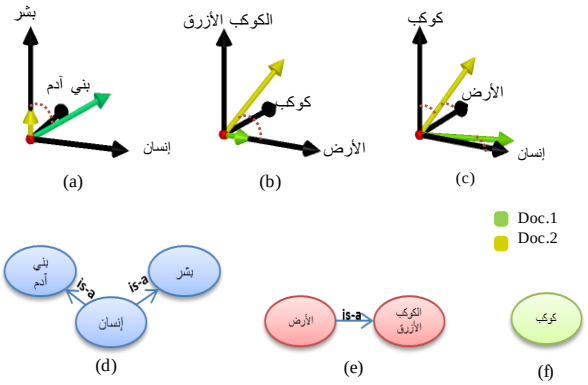
1. الإنسان ارقى الكائنات على الإطلاق. يتميز بنو ادم عن غيرهم بالعقل، ولذا فالبشر هم المكلفون بعمارة الأرض.
2. الأرض هي الكوكب الوحيد المأهول بالبشر. يشكل الماء 71% من الكوكب الأزرق.

	Term	Term frequency d1	Term frequency d2	Document frequency df	IDF = Log 4/df	Term weight d1	Term weight d2
Term-space VSM	الإنسان	1	0	1	0.6	0.6	0
	بني ادم	1	0	1	0.6	0.6	0
	البشر	1	1	2	0.3	0.3	0.3
	الأرض	1	1	2	0.3	0.3	0.3
	الكوكب الأزرق	0	1	1	0.6	0	0.6
	كوكب	0	1	1	0.6	0	0.6
Concept-space VSM	إنسان	3	1	2	0.3	0.9	0.3
	الأرض	1	2	2		0.3	0.6
	كوكب	0	1	1	0.6	0	0.6

For more clarification, follow the case presented at table 1. The first two documents, of a four-document

space, are presented at the top of the table. The terms presented at the table are instantiated only in these two documents. The first document is mainly related to the concept 'الإنسان' while the second is related to 'الأرض'. The results of a term-space-based VSM have the following defects:

- The highest weights of doc.1 are assigned to both: 'بني آدم' and 'الإنسان', which means that the vector's angle will mediate these two terms instead of being sharply aligned to the most relevant term, or hence concept, see Fig.2 (a). This is also the case of doc.2 with the two terms 'الكوكب الأزرق' and 'كوكب', Fig.2 (b).
- While the term 'بشر' is referring to the main topic of doc.1, but it takes the same value with the term 'الأرض'. That means it has the same power of representing the doc.1 as the word 'الأرض', which is a mistake. This is obvious in the vector of doc.1 at Fig.2 (a), which has an obtuse angle from the axis of 'بشر'. This case is also appearing in doc.2 for the terms 'الأرض' and 'كوكب', look at the vector of doc.2 at Fig.2 (b). That means that if the user uses the word 'بشر' to get results, so doc.1 will be ranked lower than that if he uses the word 'إنسان', which is directly affecting the *recall*. Moreover, the word 'الأرض' has the same weight in both documents, which means that if the user uses the word 'الأرض' in the query, both doc.1 and doc.2 will have the same rank while the doc.2 must achieve a higher rank. This case affects the *precision*.



(a) and (b) are two projections of 3D-Vector Space. That's for the simplification of representing 6-terms-space. (c) Represents the conversion of 6-terms-space into just 3-concepts-space. (d), (e), and (f) are Conceptual representations for the generated Concepts: 'إنسان', 'الأرض', and 'كوكب'

Fig. 2: Concepts and Vectors.

These problems are clearly avoided using the concept-space based VSM presented at the second part of the table. The six terms are grouped to produce just 3 concepts as presented at Fig.2 (d), (e), and (f). The representative concepts for each document receive the highest weights. Therefore, when the user asks for pages about 'البشر' or 'الإنسان' or 'بني آدم', then doc.1 will be returned as the most relevant. Also, doc.2 will be more associated to the concept 'أرض' than doc.1.

In terms of vectors, this is interpreted to a clear alignment of each document to its representative concept by means of both angle and length. Also, it is clear that the vector of doc.2 is mediated between the two concepts: 'الأرض' and 'كوكب' since both of them is representing it. By doing so, the concept becomes better able to represent the documents, which is the purpose of the research.

III. Features of the Proposed Model

In order to construct the concept-space dictionary, a traditional term-space is established first, and then converted into its corresponding concept-space. The essential tool used to extract the term-space from a given document-space is the RDI [6] Morphological Analyzer *(Swift®)*. *Swift* is mainly used to identify all morphological derivations occurrences of the currently processed term. These occurrences are stored at the database and then eliminated from the documents-space to not be encountered anymore, and thus, terms redundancy is avoided. Once the term-space is generated, terms will be expanded using the UWN into three expansion types; Synonyms, Generalization properties, and Specialization properties. The UWN had been navigated just for the first depth of each expansion type. Each expansion is returned along with its confidence, which is used to give the expansion an appropriate degree of relevance to the original term. The UWN had been invoked by 29 different Arabic dialects [7] in order to get the maximum possible expansions, since each dialect has its own set of expansions that may not be shared with other dialects. The redundant expansions are discarded with retaining those of the highest confidences.

3.1 Semantic Expansion

The first part of this section discusses how the synonyms affect the system while the second is dedicated for explanations about Generalization and Specialization properties.

3.1.1 Synonyms Usage

The proposed model exploited the synonyms in two main stages. The first stage is the semantic expansion of the term-space, as sampled at table 2. The system treats the synonyms as equivalents to the original term, and thus their morphological derivations have the same relevance degree to the term as those of the term itself,

6 http://www.rdi-eg.com/

7 The 29 dialects are those at: http://www.sil.org/iso639-3/codes.asp, which have the following codes:
aao, abh, abv, acm, acq, acw, acx, acy, adf, aeb, aec, afb, ajp, ajt, aju, apc, apd, ara, arb, arq, ars, ary, arz, auz, avl, ayh, ayl, ayn, and ayp.
Note that 'arb' is the code of the Standard Arabic.

see table 3. The synonyms are also used at the stage of concept-space generation, in which they are used as indicators to the terms that have a shared sense, and thus can be used to define a corresponding concept. This is declared in detail at the section of Concept-Space Generation.

Table 2: Synonyms Sample

Term Id	Term	Synonyms
1	ضروري	أساسي, جوهري, هام, ملزم, حيوي, إجباري
2	وابل	مطرغزير, شؤبوب, طوفان
3	مطر	ماطر, ممطر
4	فيضان	طوفان, غمر, إغراق, زيادة تدفق
5	الأم	والدة, أم, الوالدة
6	تنافر	تنافر, تهكم, تعارض, سخرية
7	الملابس	ثياب, هدوم, كسوة
8	إجتهد	كافح, حرص, ناضل
9	الإنسان	آدمي, بشري, ابن آدم, ناس
10	الطاقة	كهرباء, طاقة نووية
11	عمل	شغل, مهنة, مشروع
12	مشروعات	مؤسسة, شركة, شراكة, عمل
13	شغل	يستعمل, يستخدم, مهنة, عمل
14	وظيفة	مهنة
15	جسد	جسد, جيفة, جثة, جثمان
16	هدية	هدية, التبرع, موهبة, قريحة
17	معدة	معدة, بطن, جوف

Table 3: Samples of Terms Synonymous Equivalents

Term	Synonyms	Equivalent Expansions
ملابس	ثياب, هدوم, كسوة	ملابس, لبس, ملبوس, ثوب, ثياب, أثواب, كسوة, يكسو, كساء, هدوم, ...
الإنسان	آدمي, بشري, ابن آدم, ناس	آدمي, آدميون, بني آدم, بنو آدم, ناس, بشر, بشري, ...
عمل	شغل, مهنة, مشروع	يشتغل, شغل, مشاغل, عامل, عمال, مهن, يمتهن, مشاريع, مشروع, ...

Table 4: UWN Generalization and Specialization Samples

Id	Term	Expansion Type	Expansions
1	الجو	Super-Classes	طقس
		Sub-Classes	ضباب
2	الإنسان	Sub-Classes	جيل, عالم, بشر
3	عمل	Sub-Classes	مهنة, مكان, وظيفة, بيت, مؤسسة, شركة تجارية, مشروع, شراكة, وكالة, تعاوني
		Super-Classes	فعل, مشروع, مأذون, مؤسسة, شراكة
4	هامبورغ	Instance-Of	عاصمة, ميناء
5	جزيرة	Has-Instance	قبرص, هايتي, برمودا, جامايكا, أوكينلوا, مالطا, جزر كناري, ...
6	مكان	Sub-Classes	منطقة, أدغال, الأرض, فضاء, مساحة
7	شراب	Sub-Classes	مشروبات غازية, لبن, حليب, شوكولاته, عصير, فنجان شاي, قهوة
		Super-Classes	محلول, أكل, مأكل

The synonyms vary in their degree of identification with the original term's meaning. Therefore, the confidence of each is used to precisely determine how their instances should increase the overall term's frequency tf_{ij}. For example, the synonym 'تعارض' of the term 'تنافر' is more relevant than the synonym 'تهكم'.

3.1.2 Generalization and Specialization Properties

Other types of semantic expansions are those for generalization properties (Super-Classes and Instance-of) and specialization properties (Sub-Classes and Has-Instances). These types of expansions are used as indicators of an indirect relevance to the term. I.e., if the

user searches for a term that is not directly found in the documents-space, the system can find generalized or specialized results using these kinds of expansion. These expanding types may also be used as Named Entities extractors. For example, the term 'هامبورغ', at table 4, is an instance of both 'عاصمة', and 'ميناء', which can be translated semantically into {'هامبورغ' is-a 'ميناء', 'هامبورغ' is-a 'عاصمة'}. Also, the word 'جزيرة' has the instances: 'قبرص', 'هايتي'…, which also can be translated into {'قبرص' is-a 'جزيرة', 'هايتي' is-a 'جزيرة' …}. However, the UWN is still not fully filled with Arabic Named Entities; that is why the Arabic Named Entity Extractor *ANEE* [16] is used instead.

3.2 Concept-Space Generation

The term-space is converted into a concept-space via two shrinking levels as described below.

3.2.1 Term-Space Direct Shrinking (Level 1)

This level of shrinking aims to integrate all terms sharing one or more synonyms into just one comprehensive concept. For more declarations, refer to items 2 and 4 at table 2, each of which has a different list of synonyms unless the shared one: 'طوفان'. That means that they can be merged to develop a concept gathering all synonyms of both terms as presented in table 5. The terms 11, 12, 13, and 14, at table 2, are another example of the direct shrinking process. The term 'عمل' has three synonyms as 'شغل', 'مهنة', and 'مشروع'. The synonym 'شغل' itself is another term in the term-space, so these terms along with their synonyms will be directly shrieked. The other synonym, 'مشروع', is also found in the term space as its derivation 'مشروعات'. Therefore, they will be shrieked also. The third synonym 'مهنه' is matched with a synonym of another term 'وظيفة'. Consequently, these four terms will be shrieked to establish the concept 'عمل' represented at table 5. This way, the terms can be merged to get a richer definition using the full set of their describing items as depicted in Fig.1 (b). That gives the concept a firm definition and thus a higher weight, which is exactly what is intended by this research. That high weight is the power that pulls the vectors of the germane topics to the right direction as presented in Fig.2 (c).

Table 5: Direct Shrinking Samples

New concept	Concept's Definition
وابل	غمر, إغراق, زيادة تدفق مطر غزير, شؤبوب, طوفان
عمل	شغل, مهنة, مشروع, مؤسسة, شركة, شراكة, عمل, يستعمل, يستخدم

3.2.2 Term-Space Indirect Shrinking (Level 2)

The subsequent shrinking level is applied on the case of the transitive expansions. For example, if term A is a synonym of term B, which in turn is a synonym of term C, then the terms A and C are belonging to the same concept. For example, the third term at table 6, 'لغة', is one of the synonyms of the fourth term 'لسان', so they will be directly shrieked. Likewise, the word 'كلام' is a synonym of the terms 'حديث', 'لغة', and 'بيان', so they are considered as direct synonyms to 'لغة', and indirect synonym to 'لسان', and so on. This process generates a list of distinct concepts-space in which no two concepts are overlapped as presented in Fig. 1 (c).

Table 6: Indirect Shrinking Samples

Term Id	Term	Synonyms
1	كلام	نقاش, مشاجرة, مناظرة
2	حديث	كلام, جدال
3	لغة	كلام
4	لسان	لسان, لغة, لهجة
5	بيان	كلام
6	سكن	منزل, بناء, عمارة
7	منزل	بيت, سكن
8	عائلة	أسرة, عائلة, بيت

IV. Architecture and Implementation Details

This section describes the architecture of the model used to extract an Arabic concept-space from Arabic Wikipedia. The architecture is depicted in Fig. 3.

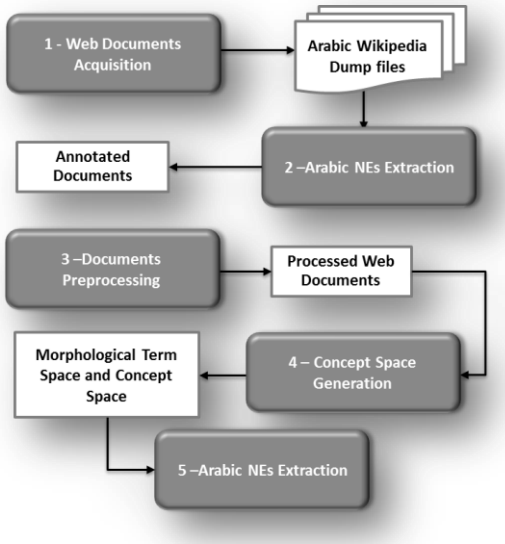


Fig. 3: Concept-space Generation Model

4.1 Web Documents Acquisition

This is the process of capturing the Web documents to be used as the source of knowledge. The acquired documents are extracted from a full Arabic Wikipedia dump [8]. The statistics of the acquired documents are described at table 7 below.

[8] The used Arabic Wikipedia dump is the version of 29-Aug-2012, and it is downloaded from: http://dumps.wikimedia.org/arwiki/

Table 7: Acquired Documents Statistics

	All Articles	Average per Article
Count of Articles	234,208	-
Count of Sentences	1952024	8
Count of Words	40111545	171
Count of letters	302817233	1293
Extracted Named Entities	2146884	9

4.2 Named Entities Extraction and Annotation

A named entity is a phrase that clearly identifies one item from a set of other items that have similar attributes. Examples of named entities are first and last names, geographic locations, ages, phone numbers, companies, organizations and addresses. ANEE [16] is capable to recognize Arabic Named Entities of types: Person, Location, Organization, Time, Number, Measurement, Percent, and File name. Nevertheless, just the first three types of annotations were used since the numbers cause some problems with the Arabic Morphological Analyzer *Swift*. The first main reason for extracting the named entities is to keep them out of any text preprocessing since some of them may include stop words or non-Arabic letters. The second reason is to send them out to the UWN to be expanded for their other aliases if they have any. Figure 4 presents a sample of the ANEE results. As presented in table 7, the overall number of extracted NEs is 2146884, with an average of 9 NEs per Wikipedia Article.

الاتحاد الدولى للاتصالات	Organization	فارسى	Location	جول	Person
مصر	Location	صافى	Person	الفيروسات	Location
القاهرة	Location	الموصل	Location	هيكل	Person
رام الله	Person	عصبية	Location	رنا	Person
فلسطين	Location	الأجرام السماوية	Location	ليفنى	Person
طرف	Person	أحمد محمد	Person	الصور	Location
أنطونيو	Person	صور	Location	جيمس	Person
الكونجرس	Organization	لأينشتاين	Person	القاعدة	Organization
رئيسة	Person	أينشتاين	Person	الكونى	Location
التلفونى	Location	اليابان	Location	كونى	Location
الكربونى	Location	الحياة	Organization	مسلم	Person
رقيقة	Person	مصدرا	Person	الميكرونى	Location
المملكة المتحدة	Location	ألبرت أينشتاين	Person	عالية	Person
أستراليا	Location	وتروس	Location	هانز	Person
وماليزيا	Location	أمريكا	Location	اندريه	Person
وجنوب أفريقيا	Location	هولندا	Location	ماري	Person
الهند	Location	المناطق	Location	مايكل	Person
هونج كونج	Location	فرنسا	Location	فيليب رايس	Person
ولندن	Location	بنى	Location	مصباح	Person
ونيويورك	Location	الحديثة	Location	نيكولا	Person
وطوكيو	Location	شمسى	Person	ألمانيا	Location
بالولايات المتحدة	Location	المبنى	Location	ماركونى	Location
ونيوزيلندا	Location	ثابت	Person	جون	Person
علم	Person	كواكب	Person	نيويورك	Location

Fig. 4: ANEE's Results Sample

4.3 Annotated Documents Preprocessing

This module is constructing the system's documents space *AWDS* as defined at Def.1.

Def.1*: *Arabic Wikipedia Document- Space (AWDS)* and Named Entities List (*NEs)

Given Arabic Wikipedia's dump files *Dump*, the Documents Space of the system ,*AWDS*, is produced by processing the *Dump* files as follows:

1. Converting *Dump* into .txt files using WP2TXT application,
2. Convert .txt files to *Annotated XML* files using *ANEE*,
3. Split XML files into set of separate articles

$A = \{a_1, a_2, \ldots, a_i, \ldots, a_n\}$; *where :*

n = *#articles in Dump, and* a_i = *article i.*

4. Clean *A* from any non-Arabic letters & Arabic stop-words, keeping the annotation tags.
5. Extract all NEs along with their categorization in NEs List.

That will generate both *AWDS* and *NEs*, which are defined as:

a. *AWDS* = {d_1, d_2, …, d_i, …, d_n},where d_i is an annotated preprocessed Arabic Wikipedia document.

b. *NEs* is the set of all Named Entities in AWDS, as

NEs = { Ne_1, Ne_2, …, Ne_i, …, Ne_n }, where:

- $Ne_i = \{ ne_{i1}, ne_{i2}, \ldots, ne_{ij}, \ldots, ne_{im} \}$
- m = # named entities in Article i
- $ne_{i,j} = ne(name, category)$
- $category = \begin{cases} P, & entity\ is\ a\ Person \\ L, & entity\ is\ a\ Location \\ O, & entity\ is\ a\ Organization \end{cases}$

4.4 Concept-Space Generation

This process is the core of the model, by which the concept-space index is produced. In order to be accomplished, the traditional Arabic Morphological Term-Space MTS is firstly constructed, as detailed at Def. 2 and Algorithm 1. Then the *MTS* is semantically expanded using the UWN as presented in Def.4. The terms are then grouped into concepts via two levels of shrinking as described in Def.5 and Algorithm 2. The main motive of using the UWN instead of the Arabic WordNet is its ability to return the expansions of a given word in many dialects. Moreover, its universality enables the use of translation tools in order to get the expansions of the words that have no expansions in Arabic. The last process is searching documents using the VSM based semantic Search System, VSCAS [13].

***Def. 2*: Morphological Term-Space (MTS)**

The *MTS* of the model is defined as the set of all distinct terms that belong to the *AWDS*.

MTS = { Mt_1, Mt_2, …, Mt_i, …, Mt_k }, where:

Mt_i = set of morphological derivations of item i in the *MTS* as:

$Mt_i = \{t_{i1}, t_{i2}, \ldots, t_{ij}, \ldots, t_{im}\}$;

k = # items in MTS;

m = # derivations of item i.

Algorithm1: MTS Generation

```
1.  Input:
2.    AWDS
3.  Output:
4.    MTS
5.  Functions and Variables:
6.    ne_j: Named Entity j in the set of all Named Entities NE defined in AWDS.
7.    Mt: list of the morphological derivations of the currently processing term.
8.    TSG: Term Space Generation function
9.    GetNextTerm: returns the next term in AWDS.
10.   Swift_Index: see Def.3.
11.   Swift_Search: see Def.3.
12. Steps:
13. Begin
14.   Swift_Index (AWDS);
15.   //MTS generation
16.   Foreach ne_j in NEs[9]
17.     If ne_j ∉ MTS
18.       x = ne_j;
19.       TSG(x); //GoTo 27
20.     EndIf
21.     Foreach d_i in AWDS
22.       x = GetNextTerm(d_i);
23.       TSG(x);
24.     EndFor
25.   EndFor
26.   Return MTS;
27.   Function: TSG(term)
28.     MTS.add(term);
29.     Foreach d_i in AWDS
30.       Pos = Swift_Search (term, d_i)[10];
31.       If Pos.count > 0
32.         Foreach p_k in Pos
33.           Mt.add(p_k.word);
34.           d_i.remove_word_at( p_k.index);
35.         EndFor
36.         MTS[j] = Mt;
37.       EndIf
38.     EndFor
39. End
```

***Def.3*: Swift Indexing and Searching**[11]

The indexing function of *Swift* is defined as:

$dx = Swift\_Index\ (scope)$.

Where, scope's indexes are saved at dx.

The searching function is defined as

$Pos = Swift\_Search(x,scope)$

The searching function finds morphological occurrences of *x* within the *scope*.

Pos: list of words/positions pairs that matched the term *x* in the *scope*.

***Def. 4*: Semantic Terms-Space (STS)**

Given *MTS*, a java application ,using UWN library [12], Princeton WordNet plugin [13], and UWN Core plugin, is built [14] in order to expand *MTS* to get the semantic

[9] The Multi-Words Named Entities are taken into consideration.
[10] Using RDI Swift Searcher.
[11] Using RDI Swift Indexer.
[12] Updated version that allows you to obtain statement weights and fixes a character encoding issue. It is updating according to Aya Al-Zoghby's report at (2012-11-23): http://www.mpi-inf.mpg.de/yago-naga/uwn/uwnapi.zip
[13] http://www.mpi-inf.mpg.de/yago-naga/uwn/wordnet.zip
[14] http://www.mpi-inf.mpg.de/yago-naga/uwn/uwn.zip

term-space *STS*. The expansion function is defined as follows:

$expand(t_i) = \{S_i, U_i, P_i, H_i, I_i\}$,

$St_i = expand(t_i) \cup M_i$

$STS = \{ St_1, St_2, \ldots, St_i, \ldots, St_k \}$, *where:*

k = # *items in MTS;*

$M_i = Mt_i$ *that is already defined at Def. 2,*

$S_i = \{s_1,\ldots,s_a\}$, *//synonyms*

$U_i = \{u_1,\ldots,u_b\}$, *//Sub-Classes*

$P_i = \{p_1,\ldots,p_c\}$, *//Super-Classes*

$H_i = \{h_1,\ldots,h_d\}$, *//Has-Instances*

$I_i = \{i_1,\ldots,i_e\}$, *//Instances-of*

Each expansion of s, u, p, h, & i, is represented as a pair of expansion-word and expansion-confidence on the form exp = (word, conf.)

Def. 5: concept-space (CS)

Given *STS*, all items that are representing the same concept are grouped in order to generate the concept-space which is defined as follows:

$CS = \{C_1, C_2, \ldots, C_q\}$, where:

q = # concepts extracted from c items of the *STS*;

C_i = the concept i that is defined by set of items of *STS* as follows:

$C_i = \{st_{i1}, st_{i2}, \ldots, t_{ij}, \ldots, st_{ic}\}$; where:

st_{ij} = the semantically expanded item j representing the concept i.

c = # items that defined the concept i.

Def. 6: Expansions-Merge

The concept-space CS is finally generated by the application of the function Expansions-Merge on the CS items to merge the expansions of all participating in the definition of the concept i in CS as follows:

$C_i = Expansions\text{-}Merge(\{st_{i1}, st_{i2}, \ldots, st_{ij}, \ldots, st_{ic}\})$

$= Expansions\text{-}Merge(\{\{M_{i1}, S_{i1}, U_{i1}, P_{i1}, H_{i1}, I_{i1}\}, \ldots, \{M_{ic}, S_{ic}, U_{ic}, P_{ic}, H_{ic}, I_{ic}\}\})$

$C_i = \{ M_i, S_i, U_i, P_i, H_i, I_i\}$

Algorithm2: Concept-Space Generation

```
1.  Inputs:
2.  STS.
3.  Outputs:
4.  CS.
5.  Functions and Variables:
6.  T: Set of all terms in STS
7.  S: Set of all synonyms of all terms in STS
8.  G: List of groups of terms that are directly clustered
9.  CG: List of groups of terms that are indirectly clustered
10. Expansions_Merge: see Def.6.
11. Steps:
12. Begin
13.   Swift_Index(S);
14.   //Shrinking Level1
15.   m=1; //groups counter
16.   Foreach term_j in STS
17.     x = [term_j, STS[j].Expansions.Synonyms];
18.     s = T.exclude(term_j) ∪
          S.exclude(STS[j].Expansions.Synonyms);
19.     relatedTermsIndexes = Search(x, s);
20.     If relatedTermsIndexes.coun >0
21.       Foreach rti in relatedTermsIndexes
22.         G[m].add(rti);
23.       EndFor
24.       m++;
25.     EndIf
26.   EndFor
27.   //Shrinking Level2
28.   Foreach g in G
29.     cg = g;
30.     Foreach g' in G ∩ g
31.       if g ∩ g' ≠ ∅
32.         cg = cg ∪ g';
33.       EndIf
34.     EndFor
35.   EndFor
36.   CG.add(cg);
37.   ForEach cg in CG
38.     c= Expansions_Merge(cg);
39.     CS.add(c);
40.   EndFor
41.   Return CS;
42. End
```

V. Results and Discussion

This section discusses the experimental results along with implementation limitations affecting them.

5.1 Experimental Results

5.1.1 Generated Concept-Space

From the *AWDS*, a term-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 102 derivations each. The term-space is then expanded via UWN with 191442 total expansions; 12% of them are in the form of phrases. The shrinking algorithm is then executed to generate a concept-space of 223502 concepts from the term-space, excluding NEs set, by compression ratio of 62 %. The ampler concept is defined by 66 merged terms while the tighter one is just of two. Some of the terms are never merged; most of them are NEs, strange, or misspelled words. The first shrinking level generated 299203 groups of terms from the 360486 terms. The second level then condensed them into just 223502 groups, each of which defines a distinct concept.

5.1.2 VSM Search System Results

The demonstrated model is evaluated by measuring its impact on the effectiveness of the VSM search

system, VSCAS [13], in terms of precision, recall, and F-Measure. To do so, the VSCAS is executed three times with the same query; 'ما هي مصادر الطاقة ؟', but different indices; *MTS*, *STS*, and *CS*. The query is preprocessed and then expanded to be:

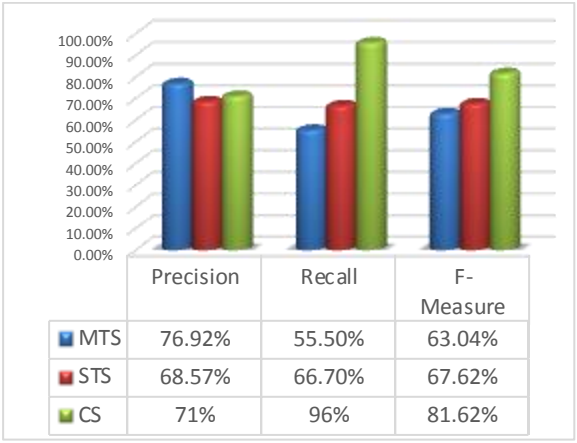
{(طاقة, كهرباء, طاقة نووية), مصادر,}

It is then conceptualized into:

{(طاقة, كهرباء, طاقة نووية, بترول, نفط, حرارة, خلايا شمسية), مصادر,}.

For calculating the system's recall, a sample of 200 documents is randomly selected, reviewed for relevancy, and then used as the documents-space of the three experiments. The experimental results showed the following:

- The highest precision's percentage was that of the experiment that used the MTS as its index. This is because, in this case, the unrelated, yet retrieved, pages are just those that have the words 'مصادر', and 'طاقة' in an irrelevant context, such as 'مصادرة الحقوق', 'مصادر المعلومات', 'المصادر المفتوحة', 'مصادر التشريع', 'حصن ', 'مدينة طاقة', and 'طاق بستان'. They are just 12 pages. However, despite its high precision, this experiment's recall is the lowest, since it ignored all those semantically relevant pages, which are 32.
- The system's precision decreased in the second experiment; when it used the *STS* indexing. The main reason of that is the phrases' handling shortage, detailed at B-6. Owing to that shortage, the expanding phrase 'طاقة نووية' is treated as two words, and hence all documents related to the word 'نووي' solely are also retrieved. That yields the retrieval of documents about 'أحماض نووية', 'برنامج نووي', 'الإمام النووي' ... etc. Therefore, the overall irrelevant retrieved documents are increased by other 10 to be 22 documents in total. Nevertheless, the recall is increased since the missed relevant documents are decreased to be just 24 instead of 32, as 8 of them are semantically retrieved due to the expansions 'كهرباء', and 'طاقة نووية'.
- As for the third experiment, which used the conceptual index *CS*, the precision increased again, but not to the extent of the first case, since the phrase handling problem still affecting it. This time, the phrase 'خلايا شمسية' caused the retrieval of additional 6 irrelevant documents related to the term 'خلايا' as ' خلايا جذعية', 'خلايا نباتية'... etc. However, the recall is increased to the extent that makes the F-Measure of this experiment the highest one. That is due to the semantic retrieval of most of the relevant documents. It just ignored 3 documents concerning 'إشعاع', and 'حركة المياه', as these terms are not present in the concept's definition itself.



	Precision	Recall	F-Measure
MTS	76.92%	55.50%	63.04%
STS	68.57%	66.70%	67.62%
CS	71%	96%	81.62%

Fig. 5: The results of VSM Search System using three dictionaries: MTS: Morphological term Space, STS: Semantic term Space, and CS: Concept Space

5.2 Implementation Limitations

Using the proposed model has enhanced the performance of the VSM search system; however, there are many difficulties and limitations that affect the overall effectiveness. These limitations are discussed below.

5.2.1 Arabic Language Processing Difficulties

The *Polysemy* problem is one of the main difficulties affecting the operation of concepts extraction. The Polysemy problem means that the same keyword may have different meanings depending on its context or its vowelization as the word 'طـ عـ م'. The non-vowelized form of the word 'طعم' is ambiguous between two words; 'طَعْم', and 'طُعْم'. Moreover, the vowelized word 'طُعْم' itself is contextually ambiguous. It may refer to 'طُعْم' for ' decoy', or 'طُعْم' for ' vaccine'.

At the semantic expansion stage, the ambiguous term may be expanded with a wrong sense, and; therefore, it will be categorized as a wrong concept. This, consequently, will affect the shrinking process since the word will be grouped with others belonging to that wrong concept. For example, while the word 'طعم', at table 8, contextually means 'طُعْم' for 'decoy', it is recognized by UWN as 'طُعْم' for 'vaccine' and thus expanded to 'لقاح'. *Swift* then aggravated the problem as it matched the morphological derivation 'طعميه' of the word 'طعم', means 'taste', to the word 'طعم' already expanded to 'decoy'. That generated a meaningless concept defined as {فلافل, طعميه, لقاح, طعم}.

Table 8: Polysemy and Ambiguity Problems

Id	Term	Synonyms
1	فلافل	طعميه
2	طعم	لقاح
3	كندي	-
4	سيد	أستاذ, سيدي, حضرتك, أفندي

The Prefixes and Suffixes uncertainty yield other matching errors as presented in the second group of the table. The terms 'سيد' and 'كندي' are grouped because the words 'كَنَدِي' and 'أَفَنْدِي' are considered as derivations of the stem 'نَدَى', with prefixes 'كَـ' and 'أَفَـ', which is wrong.

5.2.2 Morphological Analysis Limitations

The *Swift* module sometimes results in some indexing, and hence searching, errors. That causes the cancelation of several morphological occurrences of the terms, which yields redundant term-space.

However, the designed system solves this problem by grouping the redundant terms based on their shared synonyms. This case is sampled at table 9.

Table 9: Term-space Redundancy and synonyms Sharing Solution

Redundant term-space	Synonyms	Merging Redundant terms
ضرب	ضرب	ضرب
الضربة	ضرب	
نضرب	ضرب	
بضربات	ضرب	
بحرص	اجتهد كافح حرص ناضل	بحرص
حرصت	اجتهد كافح حرص ناضل	
وتحرص	اجتهد كافح حرص ناضل	

5.2.3 UWN Limitations

UWN has some limitations that affect the system's results. The firs is related with the confidence value, which may give inaccurate or wrong impression about the relevance between a term and its expansion. It may give a low value for a highly confident expansion and the vice versa. Moreover, while the valid confidence value ranges between 0.0 and 1.0, it sometimes exceeds that limit. That is due to the combination of information from two sources, according to Gerard de Melo [12], which together give more than 1.0.

Another issue is that not all UWN expansions are accurate. They sometimes are incorrectly categorized, while other times they are overlapped in different categories. Moreover, they may not be an expansion of the term at all. For example, at table 10:

1. The words 'جمارك' and 'استخدام' are not expansions to the term 'تقاليد'.
2. Even the word 'استخدام' is not a true expansion; it takes the highest confidence, which, to make things worse, exceeds 1.0.
3. The word 'أكثر' is duplicated as a synonym for the term 'زيادة'. That comes from the use of different dialects; however, the redundant expansions must have the same confidence.

TABLE 10. UWN LIMITATIONS

term	Expansions	Confidence
تقاليد	جمارك	0.78932524
	استخدام	1.0747153
زيادة	أكثر	0. 5178884
	أكثر	0.50702417

5.2.4 NEs Extraction Limitations

The performance of ANEE as a whole is acceptable; however, it has some deficiencies. The first is the extraction of non-NE as sampled in table 11-A. Also, the NEs may sometimes be misclassified, see table 11-B. Moreover, it suffers from the contextual ambiguity, as presented in table 11-C. The last problem is the overlapping of categorization such as:

ويستعمل أرباب الصناعة حوالي 10 لترات من الماء لتكرير لتر واحد من النفط .
تسْحَب المصانع في
الولايات<Person><Location><Organization>
المتحدة حوالي</Organization></Location></Person> 600 مليار لتر
من الماء يوميًا من الآبار والأنهار والبحيرات

TABLE 11. ANEE LIMITATIONS

	Named Entity	ANEE Category	The Actual Category
A- Miss-Extraction	العربي الوحيد	Person	
	معمر هي	Person	
	تفصيلي زراعة علم	Person	
	رنين أو	Person	
	نصيب الأسد	Person	
	حركة الاطراف ويستعمل عادة الجهة اليمنى	Organization	
	الحكومة الإيرانية تسعى	Organization	
	مجموعة صور توضيحية	Organization	
	للكيلوغرام الواحد	Location	
	بالكيروسين	Location	
	للأورام السرطانية	Location	
	بحوالي	Location	
B- Miss-Classification	رام الله	Person	Location
	مجلس الامن	Person	Organization
	يهوذا	Location	Person
	قنا	Organization	Location
C- Contextual Ambiguity	الخليل	Location	lover
	القاعدة	Organization	rule, base
	الفجر	Organization	daybreak
	إعمار	Organization	reconstruction
	رقيقة	Person	thin, fragile
	صافي	Person	net, pure
	كواكب	Person	planets
	هيكل	Person	structure
	صور	Location	photos

The NE 'الولايات المتحدة' is categorized as Person, Location, and Organization. In this case, the outermost annotation is taken as the categorization result, which may sometimes be incorrect.

5.2.5 Wikipedia Misspellings

Since Wikipedia is originally based on the community efforts, it sometimes contains misspelling errors, which are unrecognizable by neither UWN nor *Swift*, and thus cannot be conceptualized, even if they are crucial terms. These misspellings may be caused by the adhesion of two or more words as 'وأكسجين', 'والأعاصير', 'ومناشير', 'الخامسةوذلك', 'ميكانيكاالكم', or by the word's spelling itself as 'جولوجيا', 'بمغنطيس', 'كهرمائية'.

5.2.6 Implementation Deficiencies

The main defect of the system implementation is the capability of indexing and analyzing phrases. This shortage is the main reason for *CS* and *STS* precisions' dropping. The set of all expansions of such term are collected in one sentence as a set of words separated by spaces to be indexed by *Swift*. This index is then used for searching for their morphological occurrences within the documents-space. Actually, this approach caused a problem in dealing with the phrasal expansions, since the phrase will not be identifiable from the rest of other words. So, for example, when the word 'طاقة' was expanded to the phrase 'طاقة نووية', it was handled as two separate words. Consequently, the irrelevant documents about 'نووي' were retrieved, and; therefore, the precision is affected. In order to solve that problem, each expansion must be indexed as a standalone object, and thus the phrases will be discriminated, and will be searched for as a whole. However, the main hindrance of doing so is the excessive memory consumption caused by this indexing approach, which in turn hangs up the overall search process.

VI. Conclusions and Future Work

Arabic Language is the mother tongue for 23 countries and more than 350 million persons. Nevertheless, it is still far from being semantically searchable. This paper proposes a model for producing an Arabic Concepts-Space to be used as the index of Semantic 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, we proposed a model for extracting a concept-space dictionary, using the semantic ontology UWN, to be used as a semantic index instead of the traditionally used term-space. The proposed model enables a conceptual representation of Arabic documents space, which in turn permits the semantic classification of them and thus obtaining the semantic search benefits. The system's experimental results showed an enhancement of the F-Measure value to be 81.62% using the semantic conceptual indexing instead of 63.04% using the standard syntactic one. Still, the model's implementation suffers some limitations. Consequently, the above results will certainly be improved if those limitations are overcome. And so, we are working on solving the ambiguity problem by discriminating the meaning contextually. Moreover, we are working on refining the processing of the phrasal expansions. That will improve the results noticeably since 12% of the semantic expansions are in the form of phrases.

Acknowledgment

Special thanks to the RDI's team; Prof. Dr. Mohsen Rashwan, Mrs.Asma's Rashwan, and Eng. Ala'a Badr for their assistance to use the Swift module under x64 bit framework. The authors would also like to convey thanks to Prof. Dr. Khaled Sha'lan and Dr. Mai Ouda for their effort to provide the Arabic Named Entities of a full Arabic Wikipedia dump. The first author would would especially like to thank her sister, Eng. Eman Al-Zoghby, for here support in reviewing this article.

References

[1] AraTation: An Arabic Semantic Annotation Tool. Layan M. Bin Saleh and Hend S. Al-Khalifa. s.l. : The 11th International Conference on Information Integration and Web-based Applications & Services (iiWAS2009, 2009.

[2] Semantic internet search engine with focus on Arabic language. Naima Tazit, El Houssine Bouyakhf, Souad Sabri, Abdellah Yousfi, Karim Bouzouba. s.l. : The 1st International Sysmposium on Computers and Arabic Language & Exhibition 2007 © KACST & SCS, iscal.org.sa, 2007.

[3] Jorge Cardoso. Semantic Web services: theory, tools, and applications. s.l. : IGI Global, Mar 30, 2007. ISBN-13: 978-1599040455.

[4] Martin Hepp, Pieter De Leenheer, and Aldo de Moor. Ontology management: semantic web, semantic web services, and business applications. New York ; [London] : Springer, 2008. ISBN: 978-0-387-698899-1.

[5] Vipul Kashyap, Christoph Bussler, and Matthew Moran. The Semantic Web: Semantics for Data and Services on the Web (Data-Centric Systems and Applications). s.l. : Springer , 15 Aug 2008. ISBN-13: 978-3540764519.

[6] Next Generation Semantic Web and Its Application. Soumyarashmi Panigrahi and Sitanath Biswas. s.l. : IJCSI International Journal of Computer Science Issues, , March 2011, Vols. 8, Issue 2,.

[7] OVERVIEW OF APPROACHES TO SEMANTIC WEB SEARCH. Meena Unni , K. Baskaran. s.l. :

International Journal of Computer Science and Communication, July-December 2011, Vols. 2, No. 2, pp. 345-349.

[8] Exploring the Advances in Semantic Search. Walter Renteria-Agualimpia, Francisco J. López-Pellicer,Pedro R. Muro-Medrano, Javier Nogueras-Iso, and F.Javier Zarazaga-Soria1. s.l. : International Symposium on Distributed Computing and Artificial Intelligence, 2010.

[9] Introduction to Semantic Search Engine. Junaidah Mohamed Kassim and Mahathir Rahmany. Selangor : International Conference on Electrical Engineering and Informatics ICEEI '09, 2009.

[10] Lilac Al-Safadi, Mai Al-Badrani, and Meshael Al-Junidey. s.l. : International Journal of Computer Applications, April 2011, Vol. 19 No. 4.

[11] The Application of Vector Space Model in the Information Retrieval System. Yao-hong Zhao, Xiao-feng Shi. s.l. : Software Engineering and Knowledge Engineering: Theory and Practice, 2012. Vol. Volume 162, pp. pp 43-49 .

[12] UWN: A Large Multilingual Lexical Knowledge Base . Gerard de Melo, Gerhard Weikum. s.l. : Annual Meeting of the Association of Computational Linguistics , 2012.

[13] VSCAS: Vector Space Model- Conceptual Arabic Semantic Search System. Aya M. Al-Zoghby, Ahmed Sharaf Eldin Ahmed, Taher T. Hamza. s.l. : [Submitted for publication].

[14] Beyond Concepts: Ontology as Reality Representation. Smith, Barry. 2004 : IOS Press, 73--84.

[15] Karin Breitman, Marco Antonio Casanova, and Walt Truszkowski. Semantic Web: Concepts, Technologies and Applications. s.l. : Springer London Ltd, 28 October 2010. ISBN 13: 9781849966214.

[16] A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach. Oudah, M. M. and Shaalan, K. . s.l. : Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)., 2012.

[17] Arabic model for semantic web 3.0. Omar Isbaitan, Huda Al-Wahidi. s.l. : International Conference on Intelligent Semantic Web-Services and Applications, 2011.

[18] Samhaa R. El-Beltagy. Technology : Semantic Search. s.l. : ARABIC LANGUAGE TECHNOLOGY CENTER (ALTEC) : The Pre-SWOT Analysis, Feb 2010.

[19] Ontology Based Annotation of Text Segments. Samhaa R. El-Beltagy, Maryam Hazman, and Ahmed Rafea. Seoul, Korea : SAC '07 Proceedings of the 2007 ACM symposium on Applied computing , March 11-15, 2007.

[20] Arabic Semantic Web Applications – A Survey. Aya M. Al-Zoghby, Ahmed Sharaf Eldin Ahmed, Taher T. Hamza. s.l. : Journal of Emerging Technologies in Web Intelligence, Feb 2013. Vols. Vol 5, No 1, pp. 52-69. doi:10.4304/jetwi.5.1.52-69.

Authors' Profiles

Aya M. Al-Zoghby: A PhD student at the Faculty of Computer and Information Sciences, Mansoura University. She works as a faculty staff since 2001. She holds a Master degree from the Menofeya University; and a Bachelor degree, with excellent grade and first honor, in Computer Science, Faculty of Computer and Information Sciences, Mansoura University.

Ahmed Sharaf Eldin Ahmed: A recognized Prof. in CS and IS in Egypt. He authored more than 150 papers in international, national journals and conferences. He is the founder and coordinator of the B. Sc. Software Engineering academic program at HU. He is also the founder and manager of the Student Assessment Centre at HU. He was also the founder and manager of the quality assurance centre at HU. He is also one of the Egyptian members of the "Bologna Promoters" formed and funded by EU/Tempus. Its aim was to promote the Bologna process among the partner countries (Egypt is one of them). He was the coordinator of two Tempus III projects (M024A04-2004 and 31053-2003). He has a broad experience in curriculum development according to the European standards.

Taher T. Hamza: Assis. Professor of Computer Sciences, Faculty of Computers and Information-Mansoura University, Vice Dean For Graduate Studies and Research.

How to cite this paper: Aya M. Al-Zoghby, Ahmed Sharaf Eldin Ahmed, Taher T. Hamza,"Utilizing Conceptual Indexing to Enhance the Effectiveness of Vector Space Model", International Journal of Information Technology and Computer Science(IJITCS), vol.5, no.11, pp.1-12, 2013. DOI: 10.5815/ijitcs.2013.11.01

