



## Building a Knowledge Base Shell Based on Exploring Text Semantic Relations from Arabic Text

**Amira M. Idrees<sup>1\*</sup>**      **Essam M. Shaaban<sup>2</sup>**

<sup>1</sup>*Information Systems Department, Faculty of Computers and Information, Fayoum University, Egypt*

<sup>2</sup>*Information Systems Department, Faculty of Computers and Information, Beni-Suef University, Egypt*

\* Corresponding author's Email: ami04@fayoum.edu.eg

---

**Abstract:** This research aims at focusing on building a knowledge base shell which task is to automatically build the knowledge base component from text sources. The proposed research aims at considering the Arabic text sources as one of the main challenges in due to the difficulty in processing Arabic text that suffers from the high inflection, shortcuts, and the varieties in letters' representation. The research restored the confidence in the term frequency method and the research adopts the semantic relations network representation. The research applied the experiment for building the weeds' identification facts' part in the knowledge-based system. The experiment applied the proposed framework on two sources, the precision, recall, and f-score have been measured which resulted in an average of 91.2%, 91.1%, and 91% respectively. The evaluation metrics provide a promising perspective in the proposed approach with more focus on the bottlenecks for enhancements.

**Keywords:** Natural language processing, Semantic relations, Knowledge based systems, Text mining, Information extraction.

---

### 1. Introduction

Knowledge bases are considered the most critical component in knowledge base systems [1, 2]. The reasoning of the knowledge base systems depends mainly on the knowledge accuracy and its representation [3, 4]. On the other hand, acquiring knowledge from text is currently one of the vital directions according to the richness of the text sources [5]. While it is crucial to deal with text in general, it is more crucial to consider Arabic text according to its characteristics [6, 7]. Extracting knowledge from text requires critically applying different techniques with considering all aspects of dealing with text from one perspective and the fact that it is an Arabic text from the other perspective [8]. Many researches have proved the effectiveness of semantic networks' representation [9]. Moreover, Ensuring knowledge base enrichment can be confirmed from different perspectives [10, 11], this research adopts the perspective of building a knowledge base shell with the aim of continuously

extracting recent knowledge from recent text sources. The scope of building a complete knowledge base shell is wide [12], therefore, this research focus on building the first part of the knowledge base which is the Facts, their attributes, their values. One of the significances of the proposed research is the minimum requirement of language resources as the research applies abstracted tasks including Part of Speech tagging (POS) tagging, lemmatization with respect to the text formatting. Moreover, although the indexing scheme approach was not previously adopted, the research highlights the significance role of enforcing the indexing attribute to the text segments. The remaining of this research presents the related work, the proposed framework architecture, the main stages, the experimental study and the results, and finally the conclusion in order.

### 2. Related work

Different research in many fields have considered text analysis, semantic networks [13],

semantic relations [14], as well as knowledge base systems [15-17] in different fields. In this section presents some of the research evolution considering these aspects in interrelated perspective.

In [18], a research is proposed for domain terms extraction from Arabic text. the research was based on TF-IDF evaluation method for ensuring the high accuracy with highlighting its advantage over the term frequency method which is confronted in the current research. Another research by [19] proposed an approach for entity relationship extraction from Arabic text with following the traditional approach of text manipulation. However, the research method highlighted the approach dependency on the domain context as the study depends on pre-defined relations with regular expressions for the specific domain. Moreover, the defined relations are determined by the researchers which raise the issue of neglecting other relations that could be more effective.

Another research in [20] proposed a model which used Wikipedia source for extracting ontologies, however, the proposed approach could be criticized by different perspectives. From one perspective, the method accuracy was 65% which is a low accuracy. From another perspective, the proposed model depended on the Wikipedia source which not only is a non-reliable source but also very general with no domain specification terms definition. This later criticism highlights the inadequacy of the proposed method for extracting domain knowledge bases as the domain terms usually have a its nature.

Another low evaluation research equal 75% in [21] which aimed at extracting ontologies based on translating sentences to first order logic relations. Another limitation of the study was the sentence by sentence translation, therefore, the proposed method is not suitable for a bulk of text represented in the textbooks.

Additionally, A research by [22] which aimed with extracting semantic relations from Arabic text similar to the present study. However, the research has the same issue as in [20] and relies on using Wikipedia as a source of entities' relations. The research also adopts an extensive processing method in building parse trees, weighted inverted index, as well as embedding DBpedia as the Arabic terms' source. This extensive processing and reliance on different sources provide more complication on the exploration process.

A more recent research by [23] for the same target, however, the studies had lower accuracy with f-score 70% and 79% respectively with following the safe road for similarity measurement by

applying TF-IDF method with no consideration of the other simpler methods.

Another recent research, a research by [24], aimed at exploring the news' credibility from Facebook, the research had to support the input text with the emotion for more accurate results according to the difficulty in detecting the required keywords. Additionally, TF-IDF method was again the key aspect in [25] which focused on applying approach for determining the opinion of the participants which was also introduced in [26, 27]. This approach proved its effectiveness, however, as mentioned earlier, one of the aims in the current study is to highlight the effectiveness of the term frequency that many researches have hindered including the research in [25]. Another earlier research in [28, 29] focused on the same target of detecting opinions' summary, however, the research contribution was for detecting the effective keywords for the summary.

To summarize, although the previous studies have proposed many methods for manipulating text including term frequency, TF-IDF, and Bayesian statistics. However, the complexity of these methods is increasing respectively based on the required extensive calculations for terms extraction. By previously discussing the researches' evolutions, it is highlighted that most of the studies in the field adopts the safe road by applying TF-IDF as the confirmed effective method for similarity measurement. However, the current study aimed at highlighting that adequate methods relies mainly on the surrounding environment and efficient manipulation. The current study restores the confidence in the term frequency approach and highlights its success for accurate term extraction with the least process complexity. Introducing term frequency as one of the effective methods when adapted with effectively surrounding framework and restoring the confidence in simpler term frequency methods is one of the main contributions in this research.

Moreover, as previously discussed, supporting resources play a vital role in many of the previous research, however, the current research adopts eliminating the required resources with depending on the proposed intelligent extracting approach. Finally, the proposed framework provides a generic approach for the exploration goal which can be applied in different fields.

### 3. Problem definition and research scope

Knowledge is the main key aspect of knowledge base systems. While extracting knowledge from

experts is a continuous exhaustive process, the success of this process relies on different factors including the accuracy in the extracted knowledge, the willingness of the experts' cooperation, and the required time. All these factors suffer from bottlenecks which raised the essential need for embedding intelligent techniques in the extraction process.

As the published textbooks are considered one of the most reliable sources, therefore, the success in extracting knowledge from textbooks is one of the vital solutions which has a positive impact on many of these factors. Extracting knowledge from textbooks reduces the extraction time, ensures the knowledge accuracy, hinder the reliance on human level and avoids the knowledge conflicts.

The need for intelligent approaches for knowledge extraction is exponentially growing especially with the knowledge sources' availability which increase the effectiveness for the targeted approach. Therefore, this research provides an intelligent framework for extracting knowledge from textbook sources for directly building the knowledge base. The proposed framework has succeeded in balancing between the required target and the minimal required tools and sources for extraction as well as adapting less processing complexity. The proposed framework adopted an intelligent method for successfully extracting the entities and exploring the semantic relations between entities. Automated entities and relations' extraction are one of the vital steps in automated knowledge base extraction which eliminates the main bottlenecks in knowledge base systems development field.

#### 4. The proposed framework architecture

The proposed framework includes a set of determined phases; illustrated in Fig. 1; which represents the knowledge base shell procedure for the knowledge base construction. The proposed framework considers building the weeds' identification knowledge base which is based on extracting knowledge from formal textbooks which has the traditional structure and contents as will be discussed in the following subsections. The following is the proposed framework significance.

- The proposed framework proposes applying Information extraction approaches with the support of mining and similarity techniques for Arabic text which highlights the main bottlenecks in manipulating the Arabic text.

- As indexing schemes have already proved to provide minimum process cost, therefore, the proposed framework adopts the indexing approach as one of the main pillars in building the required relations.
- Although many researches have highlighted the weakness of considering the term frequency as an approach for extracting keywords, however, the proposed framework is reconsidering this argument and adopts applying the term occurrence approach with relating the term frequency in documents with the term frequency in the corpus with no consideration to the term weight.

Finally, one of the main research significances is adopting the generic approach for the identification task as the proposed framework does not adopt specific domain terms. The proposed approach depends on successfully extracting the key terms from the textbook source and then exploring these terms' relations.

#### 5. Description of the knowledge base shell components

As Fig. 1 illustrates, the proposed framework includes four main phases, each phase describes the steps for building one of the required semantic relationships.

##### 5.1 Identify weeds' families' names component

The weeds' family names are extracted from the table of content (الفهرس). The component applies text analysis steps as follows:

- Identify the “table of content” scope in the text.
- Eliminate non-letters can be identified as follows:  
 $\forall x, x \in t, x \notin \{[\text{ي}-\text{أ}]^*, ([\text{A-Z}]^* [\text{a-z}]^*)^*\}, x \notin \{\text{\s}, \text{\n}\}$   
Where x is the token and t is the document
- Indexing approach is adopted for each token to preserve its place and order among other tokens.
- Remove prefix Identifiers in the list (a, an, the, ‘ال’)
- Remove stop words in the list (family, عائلة، عائلة)

##### 5.2 Identify weeds' types' names component

In this phase, the research adopts the hypothesis that each weed type has only one weeds' family parent. The following are the applied steps in this phase.

- Identify Family name (recognized as header 1)
- Identify weed type (recognized as header 2)

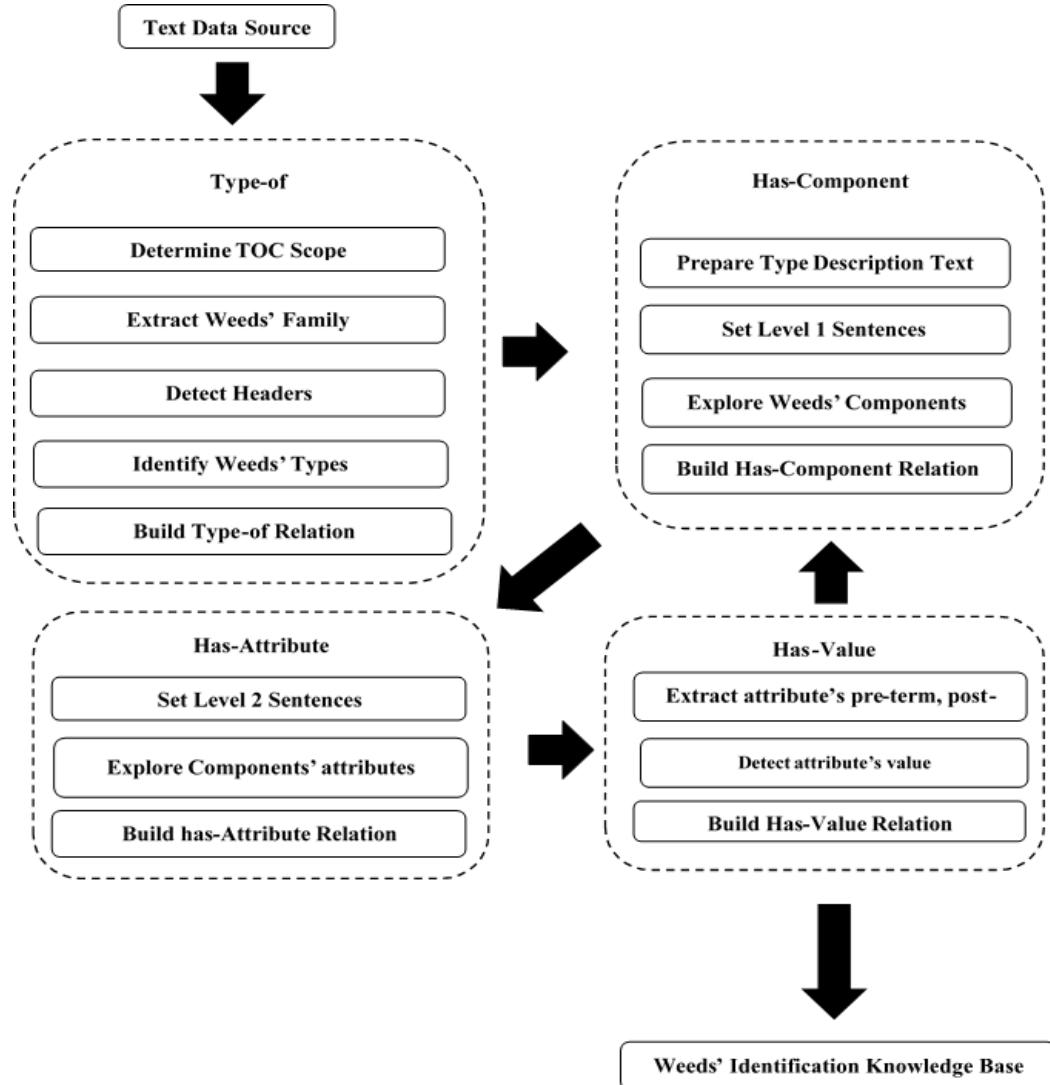


Figure. 1 The proposed framework for knowledge base shell components

- Remove redundant headers to ensure removing the redundant extracted entities.

### 5.3 Extract weeds' types' description component

In this phase, the weeds' type Arabic description is extracted from the text body, tokens' POS tagging is applied

- Apply token's POS tagging which is one of the main steps to focus on the nouns. The research argues that the components and attributes are of noun types.
- Remove stop words for preventing misleading nouns
- Unify the character presentation is applied. For example, the letter ‘ج’ has different presentation such as ‘جججج’. All of them should be unified into one appearance which is set to be ‘ج’ in this

research. The same situation is for other letters including ‘ب، الباء، الهاء، الياء’.

- Lemmatization is applied [30].
- For all Alphabetic terms, determine the term occurrence in each document and the term occurrence across the whole corpus according to the following:
  - *Term Occurrence in the Document (TOD)*: the number of times that the term appeared in the defined document. For example, if the term  $t$  appeared in the same document 4 times, then its *TOD* ( $t$ ) equals 4.
  - *Term Occurrence in the Corpus (TOC)*: the number of documents in which the defined term appeared without considering the number of occurrences in each document. For example, considering the corpus included 50 documents, if the term  $t$  appeared in 40 documents, then *TOC* ( $t$ ) equal to 40/50 which equal to 80%.

## 5.4 Explore weeds' components and attributes

This phase apply splitting method for the required target as follows: *Full stop symbol/new line-based split* is applied as *level 1 splitting* forming complete separated sentences. The research argues that each one of these sentences describes a component. Then the *Comma, Semicolon symbols, and the letter ‘و’- based split* is applied as *level 2 splitting*. The research argues that these sub-sentences represent one of the components' attributes.

For each level 1 sentence, the term which is considered as the sentence primary key and identified as the weed component has the least TOD compared with the terms that belong to the same sentence, TOC is above 95%, and Max Level 1 Term Placement is 2. Moreover, for each level 2 sentence, the term which is considered the sentence foreign keys and is a weed attribute has the least TOD compared with the terms that belong to the same sentence, TOC is above 95%, max Level 2 Term Placement is 2, and is not identified as weed component (Primary Key)

## 5.5 Explore semantic relations' network

In this phase, the semantic relations are explored which build the facts part.

### 5.5.1. Build Has-component relation

The context of level 1 and level 2 sentences explicitly detects the has-a relation between the component and the attributes respectively. The detected terms which are considered as primary keys are tagged as a weed component with eliminating the redundant terms. The 'has- component' relation is represented as:  $\forall x, x \in PK(\text{weed}) \rightarrow \text{has-component}(x)$

### 5.5.2. Build Has-attribute relation

From the previous stage, the context of level 1 and level 2 sentences explicitly detects the has-a relation between the component and the attributes respectively. Following steps 3.4.3, the has-attribute relation members are the sentences' foreign keys with eliminating the redundant terms.

The 'has-attribute' relation is represented as follows:

$\forall x, \exists y, x \in PK(\text{weed}), y \in FK(x, y) \rightarrow \text{has-attribute}(x, y)$

The set of all attributes is described as follows:

Attributes (weed) = {  $<c_s, att_s(c_s)> | d, k \in N$  },  $att_s(c_s) = \{att_d, \dots, att_k\}$  Where  $c_s \in PK(\text{weed})$ ,  $att_d, \dots, att_k \in \text{has-attribute}(c_s, y)$

Consequently, the set of all attributes with eliminating redundancy is as follows:

$\text{ALLattributes} = \bigcup_{h=1}^s att_s(C_s) | att_f(C_f) \neq att_g(C_g), f, g \in \{1, \dots, s\}$

### 5.5.3. Build Has-values relation

In each level 2 sentence, the terms that are not identified as attributes (foreign keys), are highlighted to be examined for the values set membership. For each attribute, a double-side relationship is determined, each attribute has one of the two double side relationships, either with its prefix or with the postfix, (Ex: اللون أحضر OR أحضر اللون). The decision of the attribute's value follows the following rule:

$\forall x, \text{has-seq}(x) = \{d_g | (\text{PreTerm}(x, d_g) \parallel \text{PostTerm}(x, d_g)), d_g \in D_g, g \in N\}$  where  $D_g$  is the document under investigation.

$\forall v, \text{Calculate TOC}(v) \text{ where } v \in \text{has-seq}(x), v \in D_g$

$\forall x, x \in \text{ALLattributes}, \forall v, v \in \text{has-seq}(x), v \in D_g, \text{Max}(\text{TOC}(v)) \rightarrow \text{has-values}(x, v)$

To avoid missing values, a final verification step is recommended. This step targets the level 2 sentences which do not include an explicit attribute such as 'الأوراق عريضة ناعمة'. Investigating each term in the sentence is performed and an attempt to locate all the relationships is considered between the term and the explored attributes' set members in the documents' set based on the similarity with the previously explored values. Finally, the attributes' redundant values are eliminated, and the final set of members are the values' set for each attribute. Formally describing, a term  $d_g$  is highlighted to be a value for the attribute  $x$  in the document  $D$  if one of the following is satisfied:

$\forall y \in \text{has-component}(y), (\text{PreTerm}(y, d_g) \parallel \text{PostTerm}(y, d_g)), \text{has-values}(x, d_g)$

$\forall y \in \text{has-values}(x, y), \text{PreTerm}(y, d_g), \text{has-values}(x, d_g)$

## 6. Experimental case study

For confirming the proposed framework applicability, two experiments have been applied. The case study focused on utilizing the agriculture textbooks sources. The main data sources were two textbooks that focused on the weeds' families' identification. Fig. 2 illustrates a part of the textbook which identify one of the weeds' types, while Fig. 3 illustrates a part of the second textbook source

considering the same scope. The aim of applying two case studies is to confirm the applicability of the proposed framework.

The generated files are transformed to HTML format for easier parsing. This section aims to illustrate the weeds' family identification knowledge base.

### 6.1 Extract weeds' families

The first stage identifies the “table of content” scope in the text. According to HTML format, the table of content is identified with the tag following the regex = <a href="#\_Toc[0-9]+>. TOC is then processed as follows:

- separate each line. The line starts by <p dir="RTL"> tag, end by </p>
- eliminate the lines which do not include the word “عائلة”，“العائلة”，“عائلاً”，“العائلاً”. Then remove these words.
- Each line represents one family with the Arabic and English names

### 6.2 Extract weeds' types

The main families' names are set from the table of content of the textbook. As a result of this step, a total of 32 documents are generated which represent 32 weeds' families. The same approach of transforming text into HTML is followed, the generated files are transformed to HTML format for easier parsing. Unifying characters is applied, then similarity between the families' names and the document content is applied to identify the weed types of the weeds families.

It should be highlighted that each weed type should have at least one of the names' attributes, English name or Arabic name. Weed Type Index is determined to be a concatenation of the Family Index, the English letter that has the order of the family index, and an auto number. For example, 3C1 indicates the family Index = 3, the third letter in English = C and an auto number = 1. would be the responsibility of the field expert. An example is the type ‘زربيح’ follows the ‘chenopodium album, L’ family, has an English name ‘white goosefoot’ and an index ‘1D4’. A sample of a document is in Fig. 2.

### 6.3 Explore the facts' knowledge Base

In this stage, after indexing each term, parsing text is applied with tagging each term with its POS. Then eliminating stop words is applied by removing all tokens except nouns, adjectives, and adverbs.

The sentences are then separated following the level 1 and level 2 separation as discussed earlier, and Lemmatization is applied. Calculating the token level 1 and 2 sentences' placement is performed with the TOC and TOD occurrences, which is the final step to explore the required components and types. Tables 1 and 2 present the results of applying the discussed steps with highlighting the explored components and types for one of the documents. The steps are repetitive in all documents, the TOC is calculated for all components.

For more clarification, as shown in Table 2, with considering that the whole corpus includes 135 documents, each document represents one of the weeds' types. It is found that the term “نظام” appeared in 40 documents of them, therefore TOC percentage equals  $40/135 = 30\%$  which below the

## Cyperaceae

*Cyperus alopecuroides* Rottb.  
Cyperaceae

**الوصف:**

من النباتات المعاصرة ذات ساق قلماً لونها أحمر لامع، خالية من الزغب، الأوراق شريطية ذات حافة كاملة لونها أحمر، الساقات عريضة يبلغ عرضها من 2 – 3 مم تقريباً قاربة الشكل المقنطر ، الميالس عادة 2، التمرة محدبة القمة .

**البذور:**

مستحلبة الشكل لها ثلاثة زوايا ومحاطة بشبكة من الخطوط الرمادية يبلغ حجمها  $0.5 \times 1$  مم بنية اللون .

**Description:-**

Perennials plant, stem erect, slightly green, glabrous, leaves linear margin entire, green, broader spikelets, 2 – 3 mm, hardly keeled glumes with incurved margins, stamens and stigmas normally 2 – nut plano – convex .

**Seedling:**

**Seeds:**  $0.5 \times 1$  mm size, oblong, brown color.

Figure. 2 A sample of a weed's Type Document (First Experiment)

***Cyperus rotundus***  
تحت هذا الحسينية المصنف رقم واحد في قائمة أخطر حشائش العالم (شكل 1). وينتمي هذا النوع إلى العائلة السعفية Cyperaceae. وهو نبات معمر تميزه أوراقه الحضراء الداكنة وساقه متلائمة المقاطع. والنبات عادة قصيرة لا يتجاوز ربع المتر، إلا أنه قد يصل إلى المتر في التربة الرطبة. كما أن له نظاماً جذررياً رزرياً درنياً كثيفاً تحت سطح التربة، وله زهر حمراء بنياء أو ينقسمية اللون فليلاً ما تنتج بذوراً لاذعة، وينحصر غالباً إنتاج تلك البذور تحت ظروف خاصة، فرغ ندرة ظهور تلك البذور في الولايات المتحدة مثلاً، فإنها تظهر عادة في مناطق زراعة القطن في منطقة الجزيرة بالسودان خلال التسهر الثلاثة المطيرية التي تمتد من شهر يونيو حتى شهر سبتمبر. وقد تظهر ساقان طويلة مزهرة للحسينية في شهر أغسطس في الحقول المنخفضة المعرضة للرياح. وتنتشر البذور الناتجة بواسطة الرياح لتغزو حواف القرارات المالية والحقول الأخرى، إلا أن كرة إناث تلك البذور تحتاج إلى وجودها على عمق لا يتجاوز بضعية سنتيمترات من سطح التربة. ورغم إمكانية إنتاج البذور، فإنه لا يهد هاماً، حيث يندر بصفة عامة إنتاج البذور القابلة للإناث، وهذه بذورها، نادرًا ما يربو متوسط نسبة إناثها عن بضعة أجزاء من المائة.

Figure. 3 A sample of a weed's type document (second experiment)

Table 1. A Sample of terms' placement status

Suspected Explored Type	TOD minimum rank	Level 2 placement	Level 1 placement	Term
Attribute	2	10	13	سطح
	5	12	15	ورق
	1	2	6	طرف
	1	3	7	مسنن
	1	1	1	نظام
Component	1	2	2	زهر
	1	5	5	دورة
	1	1	1	بادرة
	5	1	1	ورق
	1	1	1	بذر
	1	3	3	بذر

Table 2. A sample of explored facts

Suspected Explored Type	TOC Percentage	Term
Attribute	70/135 = 55%	طرف
Component	40/135 = 30 %	نظام
Component	130/135 = 97 %	زهر
Component	134/135 = 99 %	بادرة
Component	134/135 = 99 %	بذر

Table 3. Evaluation results

Experiment	Extracted Part	Recall (%)	Precision (%)	F-Score (%)	Average (%)
First Experiment	Families	99	100	99.5	99.5
	Types	97	97	97	97
	Components	97	92	94	94
	Attributes	85	80	82	82
	Values	79	82	81	81
	Average	91.4	90.2	91	91
Second Experiment	Types	96	100	98	98
	Components	93	90	91	91
	Attributes	90	93	91	91
	Values	83	85	84	84
	Average	91	92	91	91
	Average	91.2	91.1	91	91

threshold, therefore it is rejected as a component. On the other hand, the term “بذر” appeared in 134 documents of the corpus, therefore, TOC percentage is  $134/135 = 99.2\%$  which is above the threshold and accepted as a component with respect to the previously stated conditions. This sample represents the terms extracted from the previously discussed document.

Furthermore, considering the term “بذر”, Table 2 presents that the term has appeared in 134 documents of the corpus out of 135 documents with

a TOC percentage equal 99.2%. It has appeared once in each document (TOD). Moreover, the term has appeared in the first placement of the sentence in some of these documents. This description highlighted that the term “بذر” satisfied all the conditions to be tagged as a component. Following this approach, each term has been tagged to be either a component, an attribute, or a rejected term. These stages are repeated in all the weeds’ types documents which finally integrated to present all the components and attributes, with illustrating the attributes’ values.

Following the approach in [1], the post/pre terms which has the type nouns/adjective of the attributes are extracted and considered as the attributes’ values. For example, a part of a phrase “مستدقة الطرف مسننة” illustrate that the prefix and the postfix of the attribute “الطرف” are considered two of its possible values. This approach has been successfully presented in [1] with minimal scope for extracting the first concepts’ level. However, the authors of the current research believe that the approach is promising and can contribute in different perspectives.

Evaluating the explored Facts’ part in the Weeds’ identification knowledge base is performed through two experts in the field. The evaluation has been presented in terms of three metrics, they are precision, recall, and f-score. Table 3 presents the evaluation for the two experiments.

The extraction process included four stages which are extracting weeds’ families’ names, weeds’ types that belong to each weed family, the weeds’ plant components, the attributes of each component, and finally the values for each of these attributes. The evaluation metric calculation has been applied for each extracted part individually which is illustrated in Table 3 and the average score is then presented. The illustrated percentages highlight the success of the proposed approach with an average of 91% in the exploration process. It could be monitored that the main part that negatively affected the evaluation percentages is the attributes’ values. This negative impact could be explained as the followed approach for the values which was proposed in [1], targeted a different task which highlight the requirement of more investigation to adapt the proposed approach in [1] targeting to explore more accurate results.

## 7. Conclusion and future work

In this research, the authors aimed at focusing on building knowledge bases from Arabic text documents, this target is a step towards the success

for building a continuously updated knowledge bases which is considered the pillar of knowledge-based systems. The research presented a staged framework representing a knowledge base shell which could develop a weeds' family identification knowledge base based on extracting semantic relations from Arabic text. The proposed approach restored the confidence of the term frequency method of terms' evaluation which was strongly criticized in previous studies. The detailed stages' steps are presented which was applied on two of the Arabic textbooks in the agriculture field of weeds' subject.

The proposed framework succeeded in parsing the text, identifying all the knowledge base facts' part components with the aim of extending the research in future to include the rules' building part. The results confirmed the successful approach in exploring the weeds' families, the weeds' types that belong to each family, their attributes and the values of these attributes. The results highlighted the promising direction in extending the framework for exploring the production rules. The evaluation metrics included the precision, recall, and f-score which has been measured for each part individually, then an average have been calculated as a general framework performance. The highest evaluation score was for the families' names which equal 99.5% and the types' names which has an average for the two experiments equal 94%, while the lowest evaluation score was for the attributes' values' part with 83%. Moreover, an average evaluation for each of the evaluation metric score to the whole exploration process along the two experiments was equal 91.2%, 91.1%, and 91% respectively which leads to an overall average score equal to 91%.

It could be confirmed that the proposed framework succeeded in effective exploration for the first three parts with highlighting the low percentage in the values part. This situation could be explained as the values could have different appearance, therefore, one of the future work directions is investigating different methods for extracting the attributes' values with all possible appearance targeting more accurate knowledge exploration.

Future work of the research could also follow other targets. According to the illustrated results, future investigation is required to enhance the evaluation metrics' results. Moreover, as knowledge base includes facts, relations, and rules, therefore, extending the proposed knowledge base shell to include the rules' part is one of the main future aims. The research can also be extended to prove its generality by applying more experiments in different

fields such as industry and veterinary and others which provide higher benefits for the applied fields to both inexperienced persons in the field and beneficiary. Finally, the proposed framework is aimed to be generic for types' identification, however, more knowledge base systems could get benefits of the proposed approach such as diagnosis, treatment and others. This situation highlighted a future research direction to include further enhancement in providing a generic knowledge base shell for different knowledge-based systems tasks.

## Acknowledgments

The authors would like to express their sincerest gratitude to Prof. Ayman E. Khedr who was leading all the research stages and provided insight and invaluable expertise that greatly assisted the research.

## References

- [1] M. Y. Dahab, A. M. Idrees, H. A. Hassan, and A. Rafea, "Pattern Based Concept Extraction for Arabic Documents", *International Journal of Intelligent Computing and Information Sciences*, Vol. 10, No. 2, pp. 1-14, 2010.
- [2] M. Sayed, R. Salem, and A. E. Khedr, "Accuracy evaluation of Arabic text classification", In: *Proc. of the 12th International Conference on Computer Engineering and Systems (ICCES)*, pp. 365-370, 2017.
- [3] A. E. Khedr, A. Khalil, and M. A. Osman, "Enhanced Liver Tumor Diagnosis Using Data Mining and Computed Tomography (CT)", *The International Conference on Computing Technology and Information Management (ICCTIM)*, pp. 254-261, 2014.
- [4] A. E. Khedr, S. Kholief, and F. Saad, "An Integrated Business Intelligence Framework for Healthcare Analytics", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 7, No. 5, pp. 263-270, 2017.
- [5] M. Sayed, R. Salem, and A. E. Khedr, "A Survey of Arabic Text Classification Approaches", *International Journal of Computer Applications in Technology*, Vol. 95, No. 3, pp. 236-251, 2019.
- [6] W. S. Ismail, T. I. Sultan, M. M. Nasr, and A. E. Khedr, "Semantic Conflicts Reconciliation as a Viable Solution for Semantic Heterogeneity Problems", *International Journal of Advanced Computer Science and*

- Applications*, Vol. 4, No. 4, pp. 93-102, 2013.
- [7] A. E.-F. Hegazy, A. E. Khedr, and Y. Al Gedawy, "An Adaptive Framework for Applying Cloud Computing In Virtual Learning Environment at Education aCase Study of“AASTMT”, in *International Conference on Communication, Management and Information Technology (ICCMIT 2015)*, pp. 450-458, 2015.
- [8] A. E. Khedr, A. M. Idrees, A. E.-F. Hegazy, and S. El-Shewy, "A proposed configurable approach for recommendation systems via data mining techniques", *Enterprise Information Systems*, pp. 196-217, 2017.
- [9] E. Afify, A. Sharaf Eldin, A. E. Khedr, and F. K. Alsheref, "User-Generated Content (UGC) Credibility on Social Media Using Sentiment Classification", *FCI-H Informatics Bulletin*, Vol. 1, No. 1, pp. 1-19, 2019.
- [10] Y. Helmy, A. E. Khedr, S. Kolief, and E. Haggag, "An Enhanced Business Intelligence Approach for Increasing Customer Satisfaction Using Mining Techniques", *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 17, No. 4, pp. 159-176, 2019.
- [11] N. Sultan, A. E. I. A. M. Khedr, and S. Kholeif, "Data Mining Approach for Detecting Key Performance Indicators", *Journal of Artificial Intelligence*, Vol. 10, No. 2, pp. 59-65, 2017.
- [12] E. Azhary, A. M. Idrees, and A. Rafea, "Diagnostic expert system using non-monotonic reasoning", *Expert Systems with Applications*, Vol. 23, No. 2, pp. 37-144, 2002.
- [13] A. E. Khedr and J. N. Kok, "Adopting Knowledge Discovery in Databases for Customer Relationship Management in Egyptian Public Banks", In: *Proc. of IFIP World Computer Congress, TC 12*, pp. 201-208, 2006.
- [14] M. M. Nazier, A. E. Khedr, and M. Haggag, "Business Intelligence and its role to enhance Corporate Performance Management", *International Journal of Management & Information Technology*, Vol. 3, No. 3, pp. 8-15, 2013.
- [15] A. E. Khedr and A. M. Idrees, "Adapting Load Balancing Techniques for Improving the Performance of e-Learning Educational Process", *Journal of Computers*, Vol. 12, No. 3, pp. 250-257, 2017.
- [16] A. Mostafa, A. E. Khedr, and A. Abdo, "Advising Approach to Enhance Students' Performance Level in Higher Education Environments", *Journal of Computer Science*, Vol. 13, No. 5, pp. 130-139, 2017.
- [17] M. A. Osman, A. Darwish, A. E. Khedr, A. Z. Ghalwash, and A. E. Hassanien, "Enhanced breast cancer diagnosis system using fuzzy clustering means approach in digital mammography", *Handbook of Research on Machine Learning Innovations and Trends*, IGI Global, pp. 925-941, 2017.
- [18] A. M. Al-Thubaity, M. Khan, S. Alotaibi, and B. Alonazi, "Automatic Arabic Term Extraction from Special Domain Corpora", In: *Proc. of International Conference IEEE on Asian Language Processing (IALP)*, pp. 1-5, 2014.
- [19] M. G. Al Zamil and Q. Al-Radaideh, "Automatic Extraction Of Ontological Relations From Arabic Text", *Journal of King Saud University-Computer and Information Sciences*, Vol. 26, No. 4, pp. 462-472, 2014.
- [20] N. I. Al-Rajebah and H. S. Al-Khalifa, "Extracting Ontologies from Arabic Wikipedia: a linguistic approach", *Arabian journal for Science and Engineering*, Vol. 39, No. 4, pp. 2749-2771, 2014.
- [21] R. Ribeiro de Azevedo, F. Freitas, R. G. Rocha, J. A. Alves de Menezes, C. M. de Oliveira Rodrigues, and G. D. F. Silva, "An Approach for Learning and Construction of Expressive Ontology from Text in Natural Language", In: *Proc. of IEEE/WIC/ACM International Joint Conferences on the Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pp. 149-156, 2014.
- [22] K. M. Nagi, "ArabRelat: Arabic Relation Extraction using Distant Supervision", In: *Proc. of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015)*, pp. 410-417, 2015.
- [23] A. Benabdallah, M. Alaeddine, M. Ei, and A. Abderrahim, "Extraction of Terms and Semantic Relationships from Arabic Texts for Automatic Construction of an Ontology", *International Journal of Speech Technology*, Vol. 20, No. 2, p. 289–96, 2017.
- [24] A. M. Idrees, F. K. Alsheref, and A. I. ElSeddawy, "A Proposed Model for Detecting Facebook News' Credibility", *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 10, No. 7, pp. 311-316, 2019.

- [25] A. M. Mohsen, A. M. Idrees, and H. A. Hassan, "Emotion Analysis for Opinion Mining From Text: A Comparative Study", *International Journal of e-Collaboration*, Vol. 15, No. 1, pp. 38-58, 2019.
- [26] A. M. Mohsen, H. A. Hassan, and A. M. Idrees, "A Proposed Approach for Emotion Lexicon Enrichement", *International Journal of Computer Electrical Automation Control and Information Engineering*, Vol. 10, No. 1, pp. 242-251, 2016.
- [27] A. M. Mohsen, H. A. Hassan, and A. M. Idrees, "Documents Emotions Classification Model Based on TF-IDF Weighting", *International Journal of Computer Electrical Automation Control and Information Engineering*, Vol. 10, No. 1, pp. 252-258, 2016.
- [28] M. Othman, H. Hassan, R. Moawad, and A. M. Idrees, "A Linguistic Approach for Opinionated Documents Summary", *Future Computing and Informatics Journal*, Vol. 3, No. 2, pp. 152-158, 2018.
- [29] M. Othman, H. Hassan, R. Moawad, and A. M. Idrees, "Using NLP Approach for Opinion Types Classifier", *Journal of Computers*, Vol. 11, No. 5, pp. 400-410, 2016.
- [30] M. Boudchiche and A. Mazroui, "A hybrid approach for Arabic lemmatization", *International Journal of Speech Technology*, Vol. 22, No. 3, pp. 563-573, 2019.