S.M. kadhem
Computer Sciences Department, University of Technology/ Baghdad/Iraq
suhad_malalla@yahoo.com

A.Q.Abd_Almeer
Computer Sciences Department, University of Technology/ Baghdad/Iraq
aseelqasim30@yahoo.com

Received on: 05/04/2016
Accepted on: 24/11/2016

Arabic Texts Classification Based on Keywords Extraction Technique

Abstract- *Keyword is useful for a various purposes including labeling, summarizing, indexing, categorization, searching, and clustering. In this paper we will extract keywords from the Arabic text in order to classify it. The proposed system classify any Arabic text through simple statistic and linguistic approaches by extracting the keywords of the text (with their frequency that appear in the text) depending on a Date Base of a particular field (in this work we choose computer science field). This Data Base is represented using one B^+ tree for keywords and the other DataBase for non-keywords. The proposed system was implemented using Visual Prolog 5.1, and after testing, it proved to be a valuable for Arabic text classification (From the viewpoint of accuracy and search time).*

Keywords *Keyword Extraction, Text Classification, Morphology, B^+ tree.*

How to cite this article: S.M. kadhem, A.Q.Abd_Almeer, “Arabic Texts Classification Based on Keywords Extraction Technique,” *Engineering and Technology Journal*, Vol. 35, Part B, No. 2, pp. 96-104, 2017.

1. Introduction

The Internet is rapid growth, so it has expands the number of documents available online. This has led to develop of automated text and document classification systems that are able to classify and organize documents automatically. Text classification (or it knows as categorization) is process of structuring a set of documents according to a group of structure that is known in advance [1] [2].

Applications of Text Classification are Categorize the newspaper articles and the newswires into topics, Organize Web pages into hierarchical categories, E-mail spam filtering, Sort journals and abstracts by subject categories (e.g., MEDLINE, etc.), International clinical codes are Assigning to the patient clinical records[3] .

Distance-based algorithms, Learning algorithms, and Bayesian classification methods are the previous work on Arabic text classification, which has used in developing automated text classification systems. When used N-grams for searching an Arabic text documents, they verification by using di-grams and tri-grams which is no stemming was performed. But they are concluded that the technique of using N-gram is not an efficient approach to corpus-based Arabic word conflation. Indexing Arabic documents used tri-grams without any prior stemming. The work of using N-grams with and without stemming for text searching, their results indicate that the use of tri-grams combined with stemming improved the performing of search retrieval, however, it was not statistically significant[4].

In our proposed system we will classify an Arabic text according to a simple statistical and keyword extraction by depending on a DataBase of a particular field represent by B+ tree in order to minimize search time.

2.Keyword extraction (KE)

The task of KE is to identify a small set of words, key phrases, keywords, or keysegments from the document. These small words can be described the document's meaning[5]. Since the keywords can be define as: it is a smallest unit that can be express the meaning of document. There are many applications of text mining which can take advantage of it, such as automatic: indexing, summarization, classification, clustering, filtering, topic detection and tracking, information visualization, etc. So we can considered the task of keywords extraction as the core technology of all automatic processing for documents[6].

I.Methods for extracting keyword

There are many different methods are using for extract the keyword:

1- Simple Statistics Approaches

Simple Statistics Approaches are simple. These methods do not need to training data. The statistics information of the words can be used to identify the keywords in the document. Cohen uses N-Gram statistical information to automatic index the document. N-Gram is language and domain-independent. There are many other statistics methods such as word frequency, TF*IDF, word co-occurrences, and PAT-tree, etc.[7]

2- Linguistics Approaches

https:// 10.30684/etj.2017.138549
2412-0758/University of Technology-Iraq, Baghdad, Iraq

Mainly in these methods are using the linguistics feature of documents, sentences and words. These approaches are include the syntactic analysis, lexical analysis, discourse analysis and so on.

3- Machine Learning Approaches

The Extraction of Keyword is performing by using machine learning can be seen as supervised learning which can be learning by the examples. Machine learning approach uses the extracted keywords by training documents in order to learn a model, and applies this a model to extract keywords from new documents[6]. There is some tools keyword extraction, e.g. The Keyphrase Extraction Algorithm (KEA) is using the machine learning techniques and formula of naive Bayes for based extraction of keyphrases technical.

4- Other Approaches

Other keyword extraction approaches are mainly combine the methods above or use some heuristic knowledge in order for keyword extraction task for example the length, position, html tags around of the words, layout feature of the words, etc[7]. The proposed system we will combine between simple statistic and linguistic approaches for extract the keywords from an Arabic text.

3.Morphology

Morphology is concerned with the components that make up words. These include the rules governing the formation of words, such as the effect of prefixes and suffixes that modify the meaning of root words[8][9]. The responsible of morphology is to extracting the stem of the word by removing its suffix addition and/or its prefix addition. In Arabic language, there are three types of affixes: prefixes, suffixes, and infixes. See some examples of affix Arabic word in Table(1).

Table(1): Some examples of affix Arabic word

Kind of affix	Arabic word
Suffix	يكتب
Suffix and Prefix	يكتبون
Infix	كاتب
Suffix and Prefix	مكتبات

4. B^+ Tree[8]

It is the structure of the node, that is connected by the pointer which attached by a special node, this node is called the root, bounded by the leaves that has the unique path for each leaf, and all these paths have equal length. B^+ tree known as the index to the database, each the record in the database will store, its reference number and its key will be stored in the B^+ tree of this DataBase. B^+ tree is a balanced and arranged tree (see figure1), so it is fast when retrieving the data required.

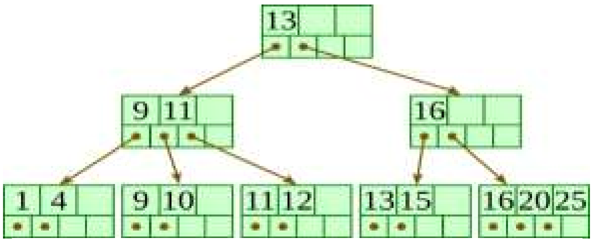


Figure (1): The example of B^+ tree

I.The proposed Text Classification

The processing of text classification involves two stages, first stage is the extraction of keywords and the second stage is actual classification of the text by using simple statistical approaches for these keywords as shown in figure(2).

We will be discussed in the following section the parts of the proposed system with more details.

1. The input to the proposed system is Arabic text, it will be consists of sentences (a sentence is a set of words, that separated by a stop mark such as "،", ".", "؟" or "!"), and sentences cutter will be responsible to producing these sentences.
2. Tokenization is a process of cutting the sentences, the tokenization part of the proposed system is using to convert a sentence to a list of tokens according to the spaces between Arabic words or any special characters.
3. Keyword extraction is a process to identify a set of words, keyphrases, keywords that describe the meaning of the input text.

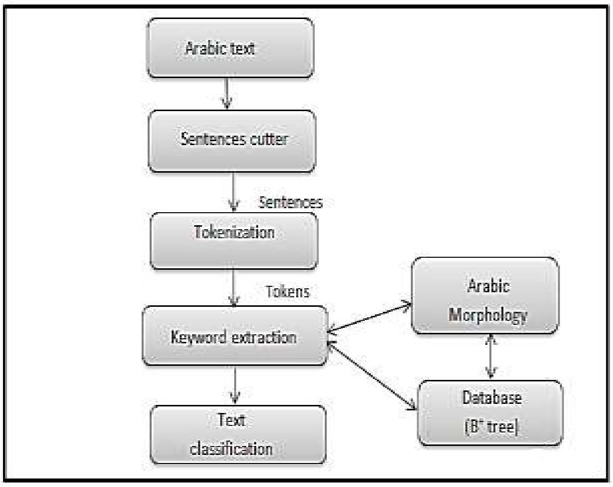


Figure (2): The block diagram of the proposed method

II.Arabic morphology

The morphology of Arabic language in this paper is dealing with the analyses an Arabic word and deal with the changes that occur by adding affixes to the Arabic words. The responsible of Arabic morphology is to extract the stem of Arabic word by removing its suffix and/or prefix.

Affixes in this paper are deal with two type: suffix and prefix. In our work we will store the keyword so we will not deal with infix. Table(2) show some examples of prefixes and suffixes Arabic word(see algorithm1).

Table(2): some examples of analyzing an Arabic word.

Suffix	Prefix	Stem	Word
ين	ت	كتب	تكتبين
ات	-	مجال	مجالات
تم	-	ربط	ربطتم
-	ست	نقل	ستنقل
ات	ال	حاسب	الحاسبات
-	س	جسر	كجسر

Algorithm1 : ”Arabic Morphology”

```
Input :Arabic word(W), list of all expected prefix and suffix, list of all expected stem from (BT1).
Output :Stem of Arabic word(WS).
Process
Begin
Step1:Remove definite article(ال) from W if any.
Step2:Remove inseparable conjunction(و) from W if any.
Step3:While prefix list is not empty then do the following steps:
    3.1 if( length of W > length of prefix) then cut it from W, and check if it is found in BT1. If found then return WS and go to step6.
    3.2 if (length of W ≤ length of prefix) then discard this prefix and go to step3.
    3.3otherwise go to step4.
Step4:While suffix list is not empty then do the following steps:
    4.1 if( length of W > length of suffix) then cut it from W, and check it if it is found it in BT1. If found then return WS and go to step6.
    4.2 if (length of W ≤ length of suffix) then discard this prefix and go to step4.
    4.3 otherwise go to step5.
Step5:Remove prefix and suffix from W As following steps:
    Step5.1: while prefix list is not empty.
    Step5.2: while suffix list is not empty.
    Step5.3:if (length of prefix and suffix< length of word) then cut suffix and prefix from W and check it if found in BT1. If found it then return WS and go to step6.
    Step5.4: if (length of prefix and suffix ≥length of word) then discard this suffix and go to step5.2.
    Step5.5:otrerwise go to step6.
Step6:End.
```

5.Database (B^+ tree)

The keyword may be it is one word or may be it is a sequence of the words that stored in the form of tree (see figure(3)), but because it is difficult to represent all keyword or keyphrase in each node of the tree, it will be represented in numbers(see figure(4)) such as each number is represent specific keyword (see Table(3)).

We will store with each record of keyword its Synonyms, abbreviation if any, and all keywords belong to it but in another record have same path of that record.

For example “علوم الحاسوب” is have these keywords(“خزن”,” خزن البيانات”,”برمجة”,”البرنامج”,”حجم البرنامج”,...........), so it store in another record but have same path of “علوم الحاسوب”. And another example, the keywords are store with “معالجة اللغات الطبيعية” have these keywords (“اللغة العربية”, ”اللغة الانكليزية”, ...) , these are stored in another record that have same path of “معالجة اللغات الطبيعية”,and so on.

Each node in the tree will be represented through the path from the root(top) to keyword, this path is store as record with keyword. The Table (3) is represent some keyword and its records(paths).

When the word or sequence of words in the text have been found in the DataBase(BT1) then return its path. For example the keyword “الاقمار الصناعية” has found in (BT1) and its path is ([1,2,7,14,٥٤]).,this record is represent the path from root to this keyword, to explain each number is as following:

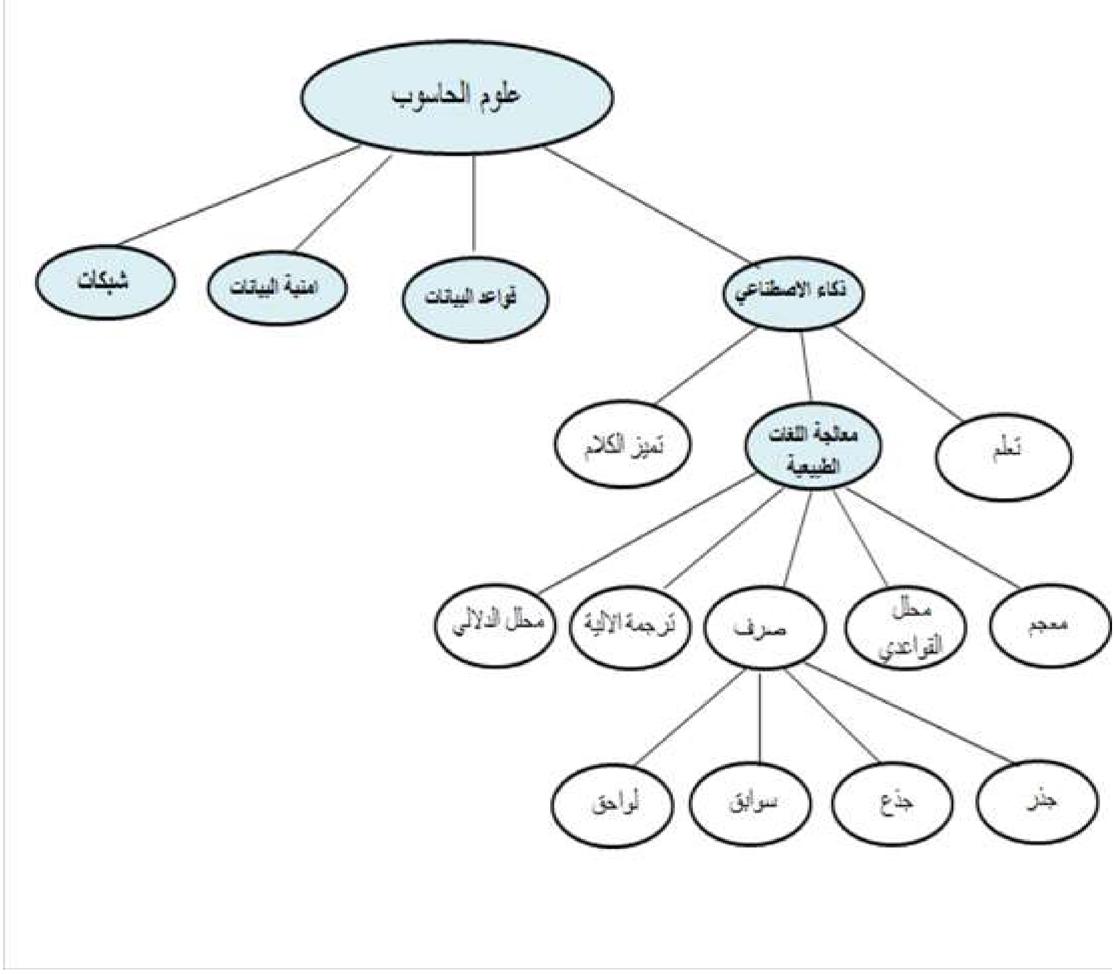
[1] is first node in first level“علوم الحاسوب” from the tree.

[1,2] is second node in second level “الشبكات ”.

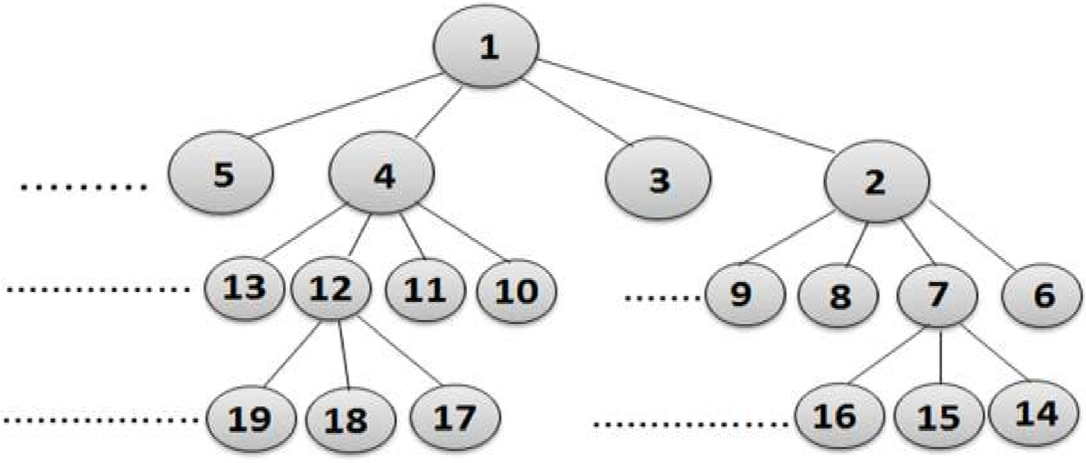
[1,2,7] is seventh node for in third level “ انواع الشبكات” .

[1,2,7,14] is 14th node for fourth level “الشبكة المحلية”

[1,2,7,14,54] is 54th node for fifth level “ اقمار الصناعية”.



Figure(3): An examples of keywords that stored using B^+ tree



Figure(4): keywords representation

Table(3): Some examples of node

keyword	Record No.	Record(Path)
علوم الحاسوب	1	[1]
شبكات الحاسوب	2	[1,2]
امنية البيانات	3	[1,3]
ذكاء الاصطناعي	4	[1,4]
المعالجة الصورية	5	[1,5]
قواعد البيانات	6	[1,6]
.		
.		
معالجة اللغات الطبيعية	60	[1,4,60]
التعلم	61	[1,4,61]
تميز الكلام	62	[1,4,62]
.		
.		
المعجم	69	[1,4,60,69]
الصرف	70	[1,4,60,70]
المحلل القواعدي	71	[1,4,60,71]
الترجمة الالية	72	[1,4,60,72]
المحلل الدلالي	73	[1,4,60,73]
.		
.		
.		
الجذر	80	[1,4,60,70,80]
الجذع	81	[1,4,60,70,81]
اللواحق	82	[1,4,60,70,82]
السوابق	83	[1,4,60,70,83]
.		
.		
قوانين	90	[1,4,60,71,90]
قواعد اللغة العربية	91	[1,4,60,71,91]
المحلل الدلالي	92	[1,4,60,71,92]
اعراب	93	[1,4,60,71,93]
القاعدة النحوية	94	[1,4,60,71,94]
.		
.		
.		
.		

6. The Propose Keyword Extraction

Each input word in the text will be as:

Nonkeywords then it will be discarded.

Keyword or keyphrase then it will be returned its record(path).

Word not found in keyword DataBaes(BT1) but it found in related record(path) to it, then it will be return the record(path) of keyword.

Otherwise discard it.

If the current word is found in nonkeyword database(DB2) then discard it. Otherwise if the word is not found then call the Arabic morphology (by removing prefix and suffix to return its stem(see algorithm1)), and try to identified if it nonkeyword.

Else if it is not found in(DB2) then translated to phrases to check if it will be keyword:

Take three words(we support that maximum length of key phrase is three that are store in B+ tree) and there is not have nonkeyword between them, then check current phrase is found in (BT1), if it found then it is keyword, return its record(path) and increase the counter of words(N).

Otherwise reduce number of words take two words and there is not have nonkeyword between them, then search again, if it found in (BT1) then it is a keyword, return its record(path) and increase the counter of words(N).

Else if it is not found take one word and try to find it in (BT1), if it is not found it then call the Arabic morphology and try to find it, if it is found then return its record(path) and increase the counter of words,otherwise discard this word but increase the counter of words(N).

The Arabic Text Classification

After we have extract the keywords from the input Arabic text and count the total number of words in the text, then classify input text. The proposed system of classification is by divide each record that have same path on total number of words(see Algorithm(2)).

Algorithm2 : ”The Proposed Arabic Text Classification”

Input : Arabic text T, BT1, DB2 for nonkeyword.
Output : Text Classification.
Process
Begin

```
Step1: Set KL as empty and set no. of word N as empty.
Sep2: Cut a sentences S from T.
Step3: If S is empty then go to step 7.
Step4: Convert S to list of tokens L.
Step5: While L <> [ ] do the following steps:
  Step 5.1: get the current word W from L
  Step 5.2: if W is found in nonkeyword (DB2) then discard it.
  Else if not found then Call Arabic morphology (Algorithm1) to get its stem SW, if SW is found in (DB2) then remove W from L.
  Else go to step5.
Step6: While L <> [ ] do the following steps:
  Step 6.1: if no. of words in L $\geq 3$ and there is no nonkeyword between them then do the following steps:
    Step 6.1.1: Get three sequence words from L
    Step 6.1.2: Concatenate these words to make candidate keyphrases(KP)
    Step 6.1.3: If KP found in (BT1) then increase N by one, get its corresponding record(path), put it in KL and remove these sequences words from L.
  Step 6.2: if no. of words in L $\geq 2$ and there is no nonkeyword between them then do the following steps:
    Step 6.2.1: Get two sequence words from L
    Step 6.2.2: Concatenate these words to make candidate keyphrases(KP)
    Step 6.2.3: If KP found in BT1 then increase N by one, get its corresponding record(path), put it in KL and remove these sequences words from L.
  Step 6.3: If no. of words in L $\geq 1$ then do the following steps:
    Step 6.3.1: Get current words W from L.
    Step 6.3.2: If W found in BT1 then increase N by one, get the corresponding record, put it in KL and remove this word from L.
    Step 6.3.3: if not found then call Arabic morphology (algorithm1) to get its stem WS. If it is found in BT1 then increase N by one, return its record(path), put it in KL, and remove it from L.
    Step 6.3.4: otherwise increase N by one and discard it.
Step7: If KL is empty then go to step 11.
Step8: Count the records that have same path.
Step9: Compute percent of each same path.
Step10: Print all the percent values.
Step11: End.
```

Example

Let the Arabic input text be:

" عادة تمثل قواعد اللغة الطبيعية على شكل قوانين داخل برنامج معالجة اللغات الطبيعية ،ولكن مع اللغة مثل اللغة العربية التي تمتاز بمرونتها النحوية، فان هذا التمثيل سينتج عنة عدد كبير من القوانين مما يودي الى زيادة وقت البحث وكذلك في حجم البرنامج. لذلك تم اقتراح طريقة وهي تمثيل قواعد اللغة الطبيعية كحدود منطقية في قاعدة المعرفة. ولقد تم استخدام الهيكل الشجري لخزن تلك الحدود لانها ملائم لخزن البيانات كبيرة الحجم."

Process

Step 1: set KL=[],N=0.

Step 2:cut sentences S from T.

“ عادة تمثل قواعد اللغة الطبيعية على شكل قوانين داخل البرنامج معالجة اللغات الطبيعية”

Step 4: Then convert the sentence to list of tokens:
(”عادة”,”تمثل”,”قواعد
”,”قوانين”,”داخل”,”البرنامج”,”معالجة”,”اللغات”,”الطبيعية”)
”اللغة”,”الطبيعية”,”على”,”شكل,

Step 5:Remove stop words from list of tokens:
”,”شكل”,”قوانين”,”البرنامج”,”معالجة”,”اللغات”,”الطبيعية”)
(”تمثل”,”قواعد”,”اللغة”,”الطبيعية

Step 6: Extract keywords, return their record(path) and count number of words or phrase:

Concatenate three words from list to make phrase:”تمثل قواعد اللغة”, then search in BT1 to check if it found or not. This phrase is not found then reduce number of words, take two words “ تمثل القواعد”, and search again. But this phrase is not found, take one word “تمثل”, and search. This word is found in the BT1, then return its record, store it in KL and remove it from the list of token.

Concatenate next three words after first word “ قواعد اللغة الطبيعية”, this phrase found in BT1 then return its record, store in KL and remove it from list of token.

Some word is take more than one record (it has more than one path) for example take the word “قوانين”. This word can be find in keyword of “ الذكاء الاصطناعي” and it find in the leave of “الصرف”.

So it has two record:

First record is [1,4].

Second record is [1,4,60,71,90].

The result of keyword extraction from text is: The first sentence:

Table(4): Keyword extraction from first sentences

Keyword	Record No.	Record
تمثل	4	[1,4]
قواعد اللغة الطبيعية	91	[1,4,60,71,91]
قوانين	4	[1,4]
	90	[1,4,60,71,90]
برنامج	1	[1]
	4	[1,4]
معالجة اللغات الطبيعية	60	[1,4,60]

For second sentences

Table(5): Keyword extraction from second sentences

Keyword	Record No.	Record
لغة	60	[1,4,60]
لغة العربية	60	[1,4,60]
نحوية	95	[1,4,60,71,94]

For third sentences

Table(6): Keyword extraction from third sentences

Keyword	Record No.	Path
تمثيل	4	[1,4]
قوانين	4	[1,4]
	90	[1,4,60,71,90]
وقت البحث	4	[1,4]
حجم برنامج	1	[1]
	4	[1,4]

For fourth sentences

Table(7): Keyword extraction from fourth sentences

Keyword	Record No.	Path
طريقة	4	[1,4]
تمثيل	4	[1,4]
قواعد اللغة الطبيعية	91	[1,4,60,71,91]
حدود منطقية	4	[1,4]
قاعدة المعرفة	4	[1,4]

For last sentences

Table(8): Keyword extraction from fifth sentences

Keyword	Record No.	Path
هيكل الشجري	4	[1,4]
خزن	1	[1]
	4	[1,4]
حدود	4	[1,4]
خزن بيانات	1	[1]
	4	[١,4]

Step8: Count and compute summation of keyword that have same path of keyword

Table(9): Count and summation of records

keyword	Record	Frequency
علوم الحاسوب	[1]	21
ذكاء الاصطناعي	[1,4]	15
معالجة اللغات الطبيعية	[1,4,60]	3
قواعد اللغة العربية	[1,4,60,71,91]	2
نحوية	[1,4,60,71,94]	1

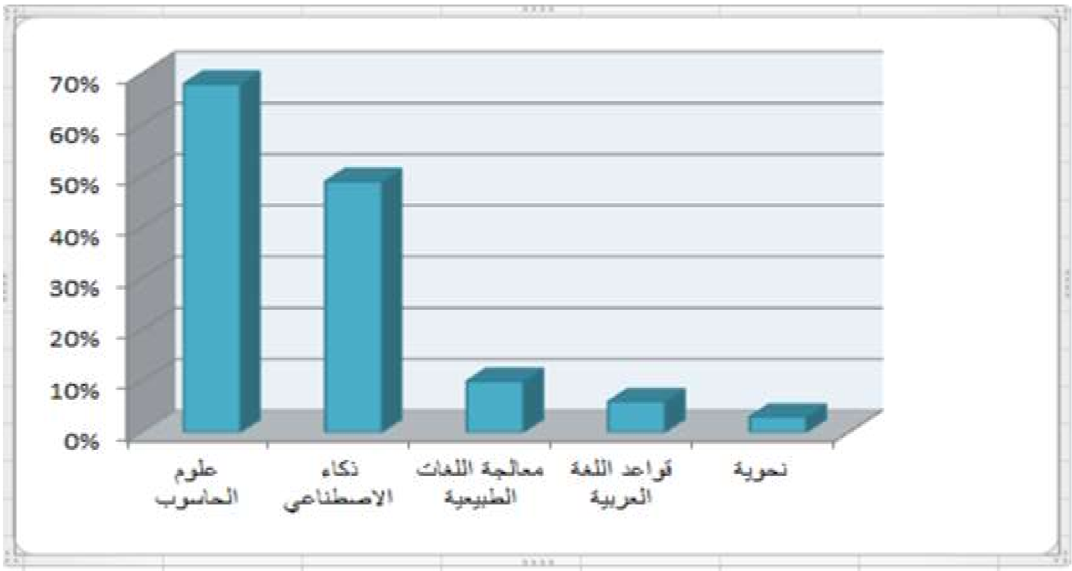


Figure.(5): The percent value of keywords input Arabic text

Divide each record on total number of word in text:

- “علوم الحاسوب” ⟶
- “ ذكاء الاصطناعي” ⟶
- “معالجة اللغات الطبيعية ” ⟶ 3/31=0.1
- “قواعد اللغة العربية ” ⟶ 2/31=0.06
- “نحوية “ ⟶

Step9:Compute the percent value (see figure(5)):

21/31=0.68
15/31=0.49

- “علوم الحاسوب” ⟶
- "ذكاء الاصطناعي” ⟶
- “ معالجة اللغات الطبيعية ” ⟶
- “قواعد اللغة العربية ” ⟶ 1/31=0.03
- “نحوية “ ⟶

Step10: Then print Arabic text classification is:

- 50% for "ذكاء الاصطناعي" 0.68 *100%= 68%
- 3٠% for "معالجة اللغات الطبيعية" 0.49 *100% =49%
- 10% for "قواعد اللغة الطبيعية" 0.1 * 100% = 10%
- 3% for "نحوية" 0.06 * 100%= 6%

0.03 * 100%= 3%

7. Conclusion

The following points in this paper can be concluded:

- The one of major application is Text classification. The proposed system is classified text based on extracted keywords by using B^+ tree.
- B^+ tree is a useful tool for Arabic text classification, and it will be minimize a search time.
- In this paper combine between two approaches the linguistic and simple statistic, these approaches a high accuracy will be provided for text classification by keyword extraction.
- Using the database base on the stem of Arabic the words will be provided an efficient memory usage of the memory.

References

[1] Matsuo, Y. M. Ishizuka, "keyword extraction from a single document using word co-occurrence statistical information", international journal on artificial intelligence tools, vol. 13, 2004.

[2] M. H. Dunham, "Data Mining: Introductory and Advanced Topics", Prentice Hall, 2003.

[3] Miss. Vidya Alone , Mrs .R.B.Talmale," Message Filtering Techniques for On-Line Social Networks: A Survey", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, March 2014.

[4] Laila Khreisat," Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study", DMIN, 2006 78-82:Jun 26; 2006.

[5] Vishal Gupta and Gurpreet Singh Lehal," Automatic Keywords Extraction for Punjabi Language", IJCSI International Journal of Computer Science Issues, Vol. 8, No 3, September 2011.

[6] Chengzhi ZHANG, Huilin WANG , Yao LIU , Dan WU, Yi LIAO , Bo WANG," Automatic Keyword Extraction from Documents Using Conditional Random Fields", Journal of Computational Information Systems, Vol. 4, 2008.

[7] Jasmeen kaur, Vishal Gupta, "Effective Approaches For Extraction Of Keywords",IJCSI International Journal of Computer Science Issues, Vol. 7, November 2010.

[8] Abdul Monem S. Rahma, Suhad M. Kadhem, Alaa Kadhim Farhan," Finding the Relevance Degree between an English Text and its Title", Eng. & Tech. Journal, Vol.30, No.9, 2012.

[9] Dr. Suhad M. Kadhem, Abdulraheem A. Abdulraheem," Using a Parser for Steganography Purpose", Eng. &Tech.Journal, Vol.33, No.4,2015.