Sudan University of Science and Technology

Faculty of Graduate Studies

College of Computer Science and Information Technology

Title:

Designing Broken Plurals Processing Method for Enhancing the Performance of Arabic Information Retrieval Systems

تصميم طريقة لمعالجة جموع التكسير لتحسين أداء نظم إسترجاع المعلومات العربية

A thesis submitted in partial fulfillment of the requirement of M.Sc Degree of Computer science (Information Retrieval)

Submitted By: MohamedAlmoayed TajAlsir MohamedSaeed Mahmoud

Supervised by: Dr. Albaraa Abuobieda Mohamed Ali

November, 2015

Verse

وَيَسْـَٔلُونَكَ عَنِ ٱلرُّوحِ ۖ قُلِ ٱلرُّوحُ مِنْ أَمْرِ رَبِّى وَمَآ أُوتِيتُم مِّنَ ٱلْعِلْمِ إِلَّا قَلِيلًا ﴿٨٥﴾

سورة الإسراء الآية 85

Dedication

To My:

Parents

…

Sisters and Brothers

…

Teachers

…

Friends

ACKNOWLEDGEMENT

I would like to seize this opportunity to express of my deepest gratitude to all who have helped me directly and indirectly towards the successful completion of this research.

Foremost, I would like to express of my sincere gratitude to my supervisor **Dr. Albaraa Abuobieda** Assistant Professor in International University of Africa in Sudan, Dean of faculty of computer studies, he was agreed to supervise of this research, although he is busier. Also for his advice, constant, support and valuable suggestions throughout the course of this research.

Besides my supervisor, I would like to thank **Dr. Mohamed Mustafa Ali,** Assistant Professor, and **Dr. Ahmed Hamza,** for their helped us to understand Information Retrieval. Especially for me to choose this research topic. Also, I am grateful for entire teaching staff of SUST College of computer science and information technology for their help during my M.Sc.

Last but not least, I am thankful to my parent, sisters, brothers and colleagues (especially **Ebtihal Mustafa Alameen**) for their support and encouragement to pursue my interests.

The thanks firstly and lastly for Allah.

MohamedAlmoayd TajAlsir

ABSTRACT

Information Retrieval is one area of computer science highly associated with the field of the Internet. It concerned with the operations for indexing, searching and retrieving information and documents which are required by a user query. Search engines and E-library systems are examples of Information Retrieval System (IRS). IRS faces a fundamental challenges in some languages especially Arabic language because it is considered as a morphological language. A plurals in the Arabic language is divided into two types Sound Plurals (SP) and Broken Plurals (BP). IRS can identify the Sound plurals simply because it keeps the structure of words in its singular and plural form. Whereas IRS fails to recognize the BP because the structure of word is changed when the singular's form of the word is derived from its plural form and vice-verse. In addition, this is reflected negatively when implementing indexing in Arabic IR. For instance, if a user typed a query contains plural form, system can retrieve all documents contain plurals form as the result, while system misses documents which contain singular form for the same word which should be retrieved. BP identification represent one of challenges faces Arabic IRS and causes loss of relevant documents; this is therefore lead to reduce Arabic IRS accuracy as a result. This study aims to explore how Arabic BP represent challenge faces Arabic IRS, and suggests a methodology based on the analysis of words to resolve BP identification problem and retrieval. The proposed method consists of three stages which are: Preprocess, BP identification, Query expansion. This study covers three patterns of syntax of Montaha Jemoa (SMJ) which are (Tfaaeel تفاعيل – Faaeel فعاعيل – Fyaeel فياعيل). Method Results were compared with (System baseline) before applying the proposed method and with (System baseline) after applying the proposed method. As a research findings, this study has successfully able to identify Broken Plural words and enhance retrieval and precision.

.

المستخلص

إسترجاع المعلومات إحدى المجالات المهمة في علوم الحاسوب والذي إرتبط حديثاً بمجال الانترنت ، والذي يهتم بعمليات الفهرسة و البحث و إسترجاع المعلومات و المستندات التي يطلبها المستخدم. من أمثلة نظم إسترجاع المعلومات محركات البحث و المكتبات الالكترونية. نظم إسترجاع المعلومات تواجه مجموعة من التحديات في بعض اللغات خصوصاً اللغة العربية لأنها تعتبر لغة نحوية. الجموع في اللغة العربية تنقسم إلى قسمين: الجموع السالمة و جموع التكسير. نظم إسترجاع المعلومات يمكنها التعرف على الجموع السالمه لانها تحافظ على بنية الكلمة في حالة المفرد و الجمع ، بينما تفشل في التعرف على جموع التكسير لان بنية الكلمة تتغير في حالة المفرد و الجمع و العكس. و هذا يعكس سلبية عند تطبيق الفهرسة في نظم إسترجاع المعلومات. إذا قام المستخدم بكتابة إستعلام يحتوي على صيغة الجمع فإن النظام سيقوم بإسترجاع كل المستندات التي تحتوي على صيغة الجمع كنتائج ، بينما يتم فقد المستندات التي تحتوي على صيغة المفرد لنفس الكلمة و التي ينبغي استرجاعها. التعرف على جموع التكسير ايضاً واحده من التحديات التي تواجه نظم إسترجاع المعلومات العربية و تتسبب في فقد بعض المستندات وبالتالي عدم دقة النتائج. هذه الدراسة تهدف لتوضيح كيف ان جموع التكسير تمثل تحدي يواجه نظم إسترجاع المعلومات العربية، و إقترحت طريقة للتعرف على جموع التكسير و تحسين الاسترجاع. الطريقة المقترحة تتكون من ثلاث مراحل وهي: المعالجة المسبقة ، التعرف على جموع التكسير ، تمديد الاستعلام. هذه الدراسة غطت ثلاثة أنماط من انماط صيغ منتهى الجموع وهي: (تفاعيل – فعاعيل – فياعيل). تمت المقارنه بين التنائج قبل و بعد إستخدام الطريقة المقترحة. بناءا على النتائج هذه الدراسة نجحت في التعرف على حموع التكسير و تحيسن الاسترجاع و حسنت في دقة النتائج.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

IR - Information Retrieval

BP - Broken Plurals

IRS - Information Retrieval System

IRM - Information Retrieval Models

BM - Boolean Model

RRM - Ranked Retrieval Model

VSP - Vector Space Model

JG - Jemoaa Gellah (جموع القله)

JK - Jemoaa Katharah (جموع الكثرة)

SMJ - Syntax Montaha Jemoaa (صيغ منتهى الجموع)

LIST OF APPINDEXS

Chapter 1
Introduction

1.1 Introduction

With the growth of modern technology and people's reliance on them in their daily life, information and internet-users are grow increasingly. Therefore, the search for specific information became very difficult. Web search is became a preferred source of information[1]. Simple user of the Internet needs to write a query freely without being bound by specific structure, those made many researchers turning to work in the field of Information Retrieval. "Information Retrieval" (IR) aims to find information (documents) from unstructured source nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" [1].

Although the Information Retrieval science has existed from several decades, but recently was associated with the field of Internet and its uses in the Information Retrieval System such as search engine for example "Google" to retrieve information and files that is satisfy user query. Figure 1-1 shows Google search engine.

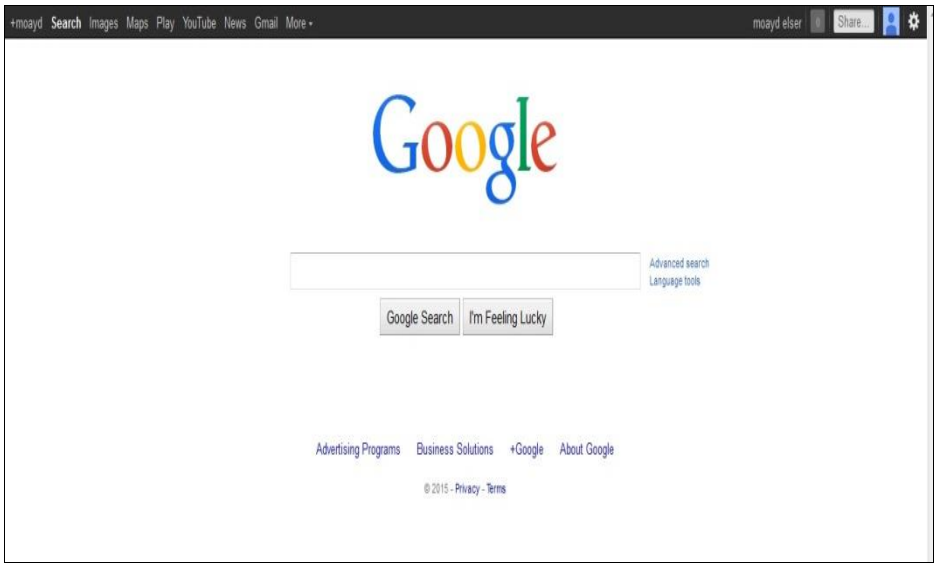

Figure 1-1: Google search engine

Figure 1.1 showed Google search engine which represent an example of Information Retrieval System (IRS). Google is probably the most powerful search engine on the market, or more precisely the most used one[2].

Information retrieval (IR) has become one of the most important areas there are a huge number of scientific studies and research have worked in this area, and the research efforts continuing to improve this domain. Information retrieval using was increased in the recent times because IR facilitate "semi- structured" search.

1.2 Problem Background

Web search engines provide quite good results for Latin characters-based languages. However, they still show many weaknesses when searching in other languages such as Arabic, and the Arabic internet users often prefer searching in these languages rather than in Arabic [2].Searching in Arabic language meets fundamental problems related to certain reasons one of them Arabic is morphological language. “Irregular (so-called broken) plural identification in modern standard Arabic is a problematic issue for Information Retrieval System (IRS) and represent challenge for language engineering applications”[3].

Although Broken plurals constitute 41% of texts in large Arabic corpora Boudelaa and Gaskell [4]. There are difficulties to identify Broken Plurals (BP) because there are no specific rule governed, that made it difficult to detect and retrieve properly [3]. There are a few studies proposed solutions for this problem. This research proposes solution for this problem to improve documents retrieval. Example (1) below illustrates this problem. Suppose that, there are two indexed text documents, first document D1 and the second document D2, and any document content sentence as:

D1: " امراض القلوب و كيفية علاجها".

D2: "أثبتت الدراسات بأن الرياضة تقلل من خطر الإصابه بأمراض القلب".

If a user type a query Q: “كيف نجعل القلب سليم ”, the system will retrieve the second document D2 only, because it contains the word (القلب,”heart”), but the system cannot retrieve the first document D1 although it contains the word (القلوب,”hearts”) which represents a plural of word (القلب), where it must be retrieve as a result too. Broken plurals have set of patterns and it is difficult to identify them. The words which come on these patterns may represent plural and it may be singular. Table 1.1 show some examples of words which match one of these patterns and represent Broken Plurals, and other words with same pattern but do not represent BP.

Table 1-1 : Broken Plurals patterns examples

No.	Patterns	Broken Plural	Not Broken Plural
1	افعال	اجيال Generations	احتار Puzzled
2	افعله	اعمده Columns	امسيه Evening
3	افعل	انهر Rivers	اسرع faster
4	افعلاء	اغنياء The rich people	اختباء Hide
5	فعول	قلوب Hearts	سلوك Behavior

1.3 Problem Statement

Simply, the problem statement of this research is to identify Arabic Broken Plurals and their singular derivation forms in order to enhance Arabic Information Retrieval Systems.

1.4 Research Objectives

The main objectives of this research are:

- To understand how Arabic *Broken Plural* (BP) represents challenge for IRS.
- To study some of Broken Plurals patterns for words identification.
- To retrieve back the singular form of BP words.
- To propose a method for increasing Arabic documents retrieval precision.

1.5 Research Hypothesis

Information Retrieval System fail to retrieve all documents especially if a query contain Arabic BP words. The proposed solution need to solve the BP identification and enhance a precision for document retrieval.

1.6 Research Significant

Contribute improving information retrieval for Arabic language, and the important goal of researches in this trend are to enhance retrieval for Arabic language by obtain a good result mainly for a query which contains Broken Plurals words.

1.7 Thesis Structures

Chapter one presents Introduction which highlight a brief history of Information Retrieval and research problem. Chapter two presents Background and Literature review. Chapter three presents the methodology used in this research, Chapter four introduces Result and discussion, Chapter five views Conclusion and Recommendations.

Chapter 2
Literature Review

2.1 Introduction

This Chapter previews basic concepts considered the keys of this research and related works in the same field. This chapter organized into six sections. Section 2.2 explain the principles of Information Retrieval IR (definition, basic concepts, Models, Application, IR for Arabic language), while section 2.3 explain the Arabic Language and Plurals. Section 2.4 explain the Data set corpus which this research depends on. Section 2.5 explain the Baseline Method 2.6 presents an Evaluation Measures that used to evaluate the results. Section 2.7 preview Related Works based on Broken Plural identification.

2.2 Information Retrieval Background

Information Retrieval (IR) is a branch of computer science concerned with indexing and searching operations for documents and files and web sites to facilitate and improve the search process. Information Retrieval has become one of the most important areas where most of the studies and researches have worked in this area and research efforts are continuing to improve this domain. The use of Information Retrieval are increased in recent times more than Databases because IR facilitate "semi structured" search. Although the Information Retrieval science has existed before many decades, but recently was associated with the field of Internet, and most important use in search engines to retrieve information and files that satisfy user query.

2.2.1 Information Retrieval Definition

Information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [1].

From the definition the collection usually stored in one or more computers, because the collection may be so large and also need hardware and software with

high properties. This collection may store in large server with high properties. For example search engines work in the World Wide Web (www). There are a set of search engines like Google, Yahoo, Orange, AltaVista,...etc., but the Google search engine is popular search engine in the world and most powerful and most people prefer to use it rather than other search engine because it provides better services and good results with high precision and speed and also have large collection. Google search engine has good hardware and software, Google data centers (in 2007) mainly contains commodity machines, Data centers are distributed all over the world, one million servers, three million processors/cores[5].

2.2.2 Information Retrieval Basic concepts

As mentioned previously Information Retrieval consists of two important processes. The first process is Indexing, and the second process includes the Searching. Figure 2.1 illustrates the typical Information Retrieval task which IRS follows[6].

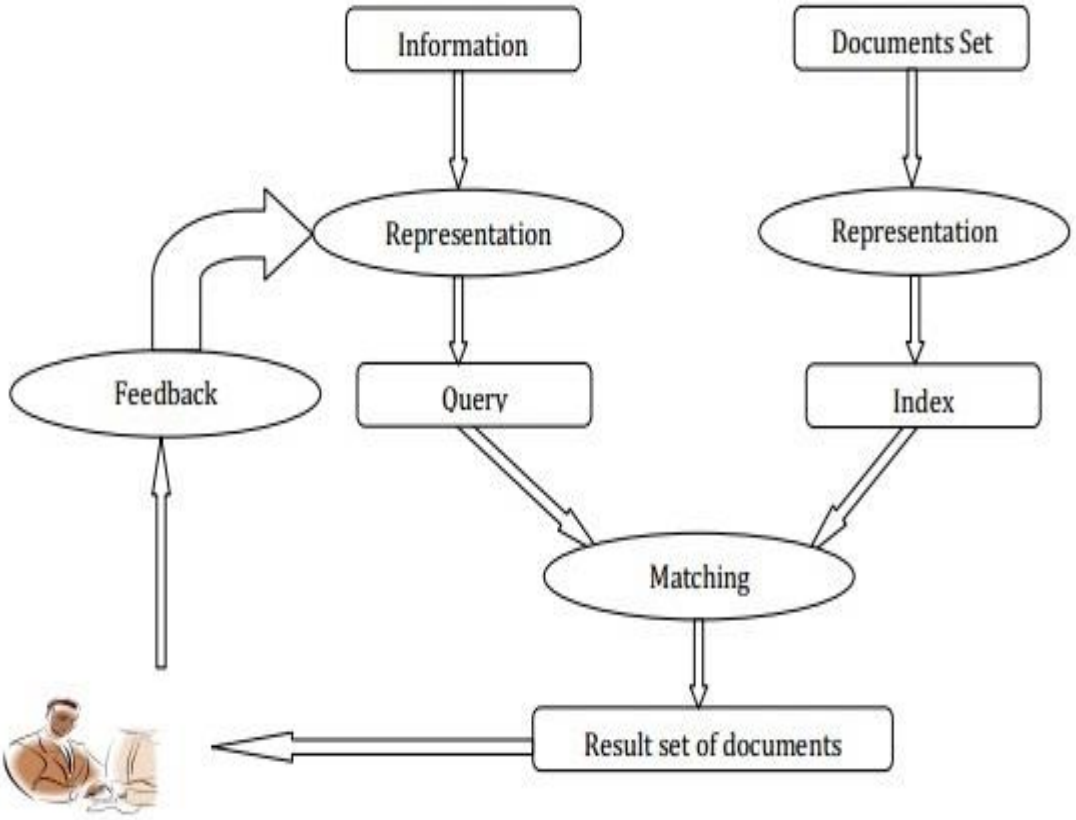

Figure 2-1: A typical information retrieval task

2.2.2.1 Indexing

The purpose of indexing is to optimize speed and performance to find relevant documents for a query statement. Without an index, the search engine must scan all documents in the corpus until finding the target document or file, which may require considerable time and computing power. For example *Inverted Index* one of methods which use for indexing purpose, in *Inverted Index* an index always maps back from terms to the parts of a document where they occur. Nevertheless, inverted index, or sometimes inverted file, has become the standard term in information retrieval [1].

2.2.2.2 Searching

Means how to search about information or documents using IRS and obtain good result from a query.

2.2.2.3 Pre-Processing

Pre-processing stage is important for Indexing and a Searching processes because it makes compatibility among the query statement and indexed data. The importance of this stage it allows the user to write query without restricting and make compatibility with existing index. Pre-processing contains several operations; in the next section will give briefly about the most important processes in this stage in the following paragraphs.

Tokenization: is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called token.

Lower case: usually is used for English language and related languages (use the same alphabet), it means changing all capital letters to small letters.

Stop words: very common words which would appear to be of little value helping select documents matching a user need. This process removes some word for indexing purpose. Stop words are different from one language to another, for example preposition in English language like these words (a, of, if,….etc.) denoted as "stop word" and it will be removed, and also (من ، عن ، على ، او ، ان) denoted as "stop word" for Arabic language. Appendix A stated an Arabic *Stopwords*.

Stemming: The users writes the query in different format, Stemmer can improve the search effectiveness so an Information Retrieval System (IRS) can match user's

queries with relevant indexed documents. Now an IRS should be able to change all these forms that have the same meaning to the standard form and thus grouping all these different formats in to standard format and this should be done on both sides, on user's queries and on index terms. There are many stemmer standards for each language, for example for English language use Portal Stemmer, and for Arabic use Light stemmer or Khoja stemmer (Root extraction)[7].

2.2.3 Information Retrieval Models (IRM)

The Information Retrieval Models define the internal representation of documents and queries as well as the score function. There are two types of IRM first one is called Boolean Model (BM) and the second one is Ranked Retrieval Model(RRM), BM as will show depends on matching between Query and indexed document, if there are matching then it will take it as result, the result list is a set of documents without order [8]. Ranked Retrieval Model lookup for how to calculate similarity between documents and query and returns ranking more relevant between them (Query and documents) by calculating scoring. Many IRM have been proposed and used in Information Retrieval System. In the following sections will describe the most commonly used ones [8].

- Boolean Model.
- Ranked Retrieval Model.

  Vector space model.

  Probabilistic model.

Most models above are built upon the notion of term. In IR term refers to a basic unit used in the representation, it can be a word (e.g., "automatic"), a word stem (e.g., "automat"), or a phrase (e.g., "computer system") depending on the indexing process used[8].

In the next sections will give a brief background about these models (Boolean Model, Ranked Retrieval Model).

2.2.3.1 Boolean Models

The Boolean model is the simplest model which was used in information retrieval system. This model work by comparing the term in the query and the terms in the documents if true (matching) then retrieve that document. In BM documents are represented by a conjunction of terms, such as D=t1^t2^t3 which means that the terms t1, t2, and t3 are present in the document D. Equivalently, this document can also be represented by a set of terms: D= {t1 , t2, t3}. Terms are not in the Boolean expression are assumed to be absent. BM can retrieve documents only if the terms in Query match terms in the documents. Then the problems of this model are feast or famine, and ranking, that means if terms in the query is found in all documents in the collection then this model can retrieve all documents collection without ranking what is the important document of them, and if the terms in a Query does not match terms in the collection it will return nothing. The main problem of this model it does not rank the result [8].

2.2.3.2 Vector Space Model

The Vector Space Model (VSM) Salton and McGill, 1983; Salton et al., 1975 uses a vector to represent a document or a query [8]. The VSM is formed by all terms the system recognizes in the documents. In the document vector and the query vector, each element (di or qi, $1 \leq i \leq n$) represents the weight of the corresponding term in the document or the query[8]. For instance, see the vectors below:

$$\begin{array}{lccccc} \text{vector space:} & \langle t_1 & t_2 & t_3 & \ldots & t_n \rangle \\ \text{document:} & \langle d_1 & d_2 & d_3 & \ldots & d_n \rangle \\ \text{query:} & \langle q_1 & q_2 & q_3 & \ldots & q_n \rangle \end{array}$$

The weights of di or qi could be binary, i.e., 1 representing the presence, or 0 representing the absence of term in the document or the query. However, the most commonly used method is the tf *idf weighting schema [8]. tf means term frequency within the document (or the query), and idf means inverse document frequency, which is usually calculated as follows:

$$idf(ti) = log \frac{N}{n(ti)} \qquad (2.1)$$

Where t_i is a term in the vocabulary, n is the number of documents in the whole document collection, and n (***ti***) is the number of documents containing ***ti***(also called document frequency). The general idea behind ***tf * idf*** weighting is that, the more a term appears in a document (or a query) is important (the ***tf*** factor); the less term is common among all the documents in the collection, the more it is specific, thus important (the ***idf*** factor).

Given vector representations, the score of relevance is estimated by a similarity between the vectors. In the first similarity calculated by Euclidean distance but this method may give large distance between q and d although they are more similar, when the length of vector is more differ [5]. As the example in Figure 2.2 bellow shows that distance between q and d_2 is larger than q,d_1 and q,d_3 although q and d_2 is more similar than other.

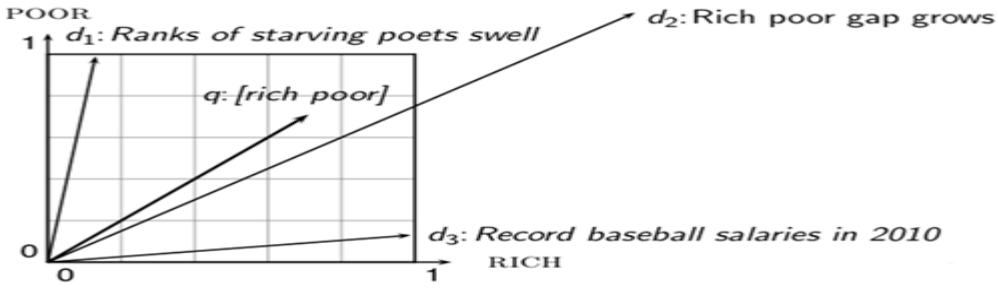

Figure 2-2: VSM using Euclidean distance and bad result

To calculate similarity there are other equation used it is called **cosine similarity**, which use angle instead of distance it defined as follows:

$$sim(\bar{D},\bar{Q}) = \frac{\bar{D} \bullet \bar{Q}}{|\bar{D}| \times |\bar{Q}|} \qquad \textbf{(2.2)}$$

Where $|\bar{D}|$ it is the length of the vector. Figure 2.3 illustrates The Vector Space Model (VSM) by using cosine similarity. Cosine similarity is a monotonically decreasing function of the angle for the interval (0°, 180°).

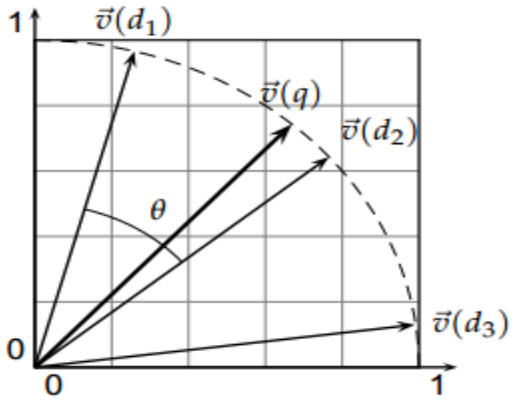

Figure 2-3: Vector Space Model (VSM) using cosine similarity

2.2.3.3 Probabilistic Model

Probabilistic methods are one of the oldest formal models in IR. Already in the 1970s they were held out as an opportunity to place IR on a firmer theoretical footing, and with the resurgence of probabilistic methods in computational linguistics in the 1990s, that hope has returned, and probabilistic methods are again one of the currently hottest topics in IR. Traditionally, probabilistic IR has had neat ideas but the methods have never won on performance. Getting reasonable approximations of the needed probabilities for a probabilistic IR model is possible, but it requires some major assumptions [1].

Documents in a collection could be ranked according to their probabilities of relevance or their similarity degrees with regards to queries. The key issue here is how to compute a probability of each term in a query and how to assign the final probability that a document is relevant to the query. Broadly speaking, a probabilistic retrieval model employs the absence or the presence of a term in a document to predict a weight for that term. This weight corresponds to the estimated probability of relevance of that term and the combination of all the query terms weights is there by used to determine whether the document is relevant or not.

2.2.4 Information Retrieval Application

Information Retrieval System has large number of application, most searching in Internet use unstructured query, in next section below remind the most used application.

2.2.4.1 Search engine

Search engine indexing collects, parses, and stores data to facilitate speed and accuracy of information retrieval. Index design incorporates interdisciplinary concepts from linguistics, cognitive psychology, mathematics, informatics, physics, and computer science. Search engines designed to find web pages on the internet which need to make indexing for web pages.

2.2.4.2 Digital Library

A special library (also referred to as digital library or digital repository) is focused collection of digital objects that can include text data.

2.2.4.3 Media Search

It include image retrieval, music retrieval, News retrieval, Speech retrieval, video retrieval.

2.2.4.4 Desktop Search

Which use for searching about files and document that it saved in computer, it provide search with the name of file or search with contain file.

2.3 Arabic language and plurals

Arabic language one of the most widespread and commonly use, which used by millions of users in the Internet daily basis. To spread Arabic culture must be exploit the Internet and web site and facilitate the search processes for Arabic web sites. Arabic language faces many challenges especially with Information Retrieval System applications such as search engine [2].

2.3.1 Arabic Language and Information Retrieval System Challenges

Most Arabic internet users master a second language for example English because the information on the Web is widely available in this language, the Arabic internet users often prefer searching by these languages rather than searching by Arabic language.

Web search engines provide quite good results for Latin characters-based languages however; they still show many weaknesses when searching in other

languages such as Arabic. Arabic information retrieval still faces many difficulties due to the Arabic linguistic features[2].

Information Retrieval System for Arabic is a popular area of research because the nature of Arabic as an inflectional, tri-consonantal roots-based and morphological language. Arabic has much richer morphology than English that make it difficult to use with Information Retrieval System, also Arabic language presents significant challenges to many natural language processing applications.

Arabic Broken Plurals identification represent one of these challenges which faces IRS. These challenges will explore in detail in this chapter.

2.3.2 The Arabic number system

Arabic has two genders: feminine and masculine, and three numbers: singular, dual, and plural, and three grammatical cases: nominative, genitive, and accusative. A noun has the nominative case when it is a subject; accusative when it is the object of a verb, and genitive when it is the object of a preposition[9]. Figure 2-4 shows the Arabic number system hierarchy[10]. The concept of Plural in Arabic is differ from English, in English a plural noun can refer to two or more of something, but in Arabic a plural noun refers to three or more of something and dual refer for two something.

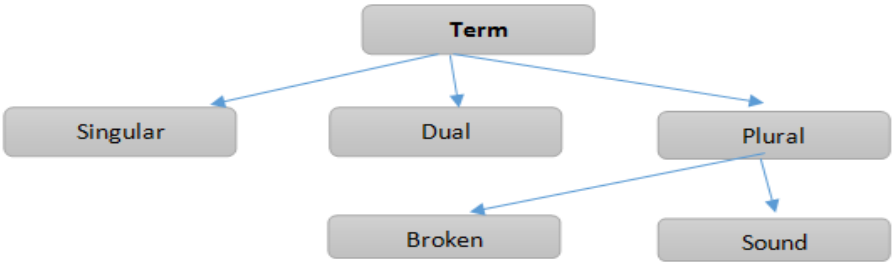

Figure 2-4: Arabic Number System

2.3.3 Arabic Plurals

Plurals in Arabic language divided into two types *Sound Plurals* and *Broken Plural*, in the following paragraph will illustrate briefly the *sound plural*. As stated before, study focus more on Broken Plurals[3]. Figure 2.5 illustrate exactly the area of BP which conceder by this study.

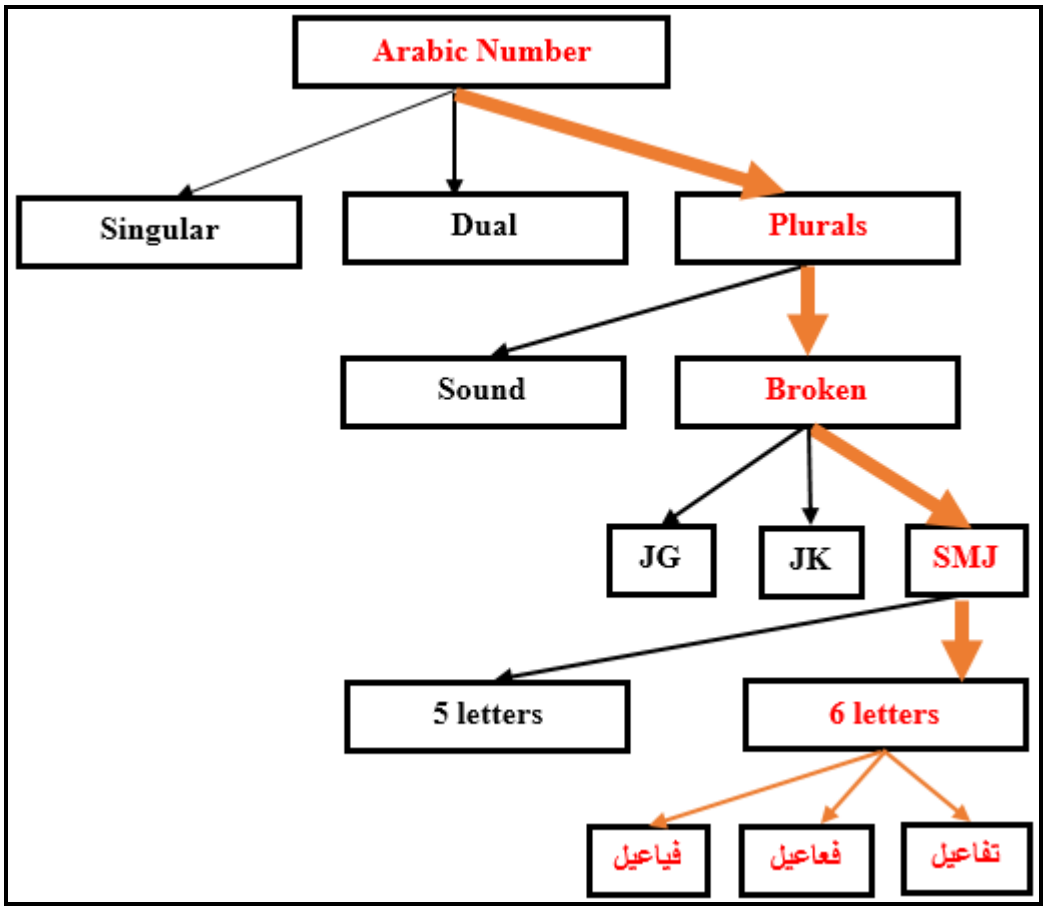

Figure 2-5: The area of BP conceder by this study

2.3.3.1 The sound plurals

The sound plurals is a type of plurals in the Arabic language and so-called correct plurals (الجموع السالمه), so named because it keeps the proportion of the singular structure without changing, where this type divided into two parts which are Masculine plural(جمع المذكر السالم) and Feminine plurals(جمع المؤنث السالم).[4]

The Masculine Plurals: is Sound plural which called for more than two, remained unscathed changed single, called this plurals of a male or a recipe for same masculine by adding suffix (ين , ون) to singular nouns, for example the word (باحث,"researcher") is a singular it can be pluralized by adding suffix (ون) to form the masculine plural (باحثون, "researchers") another example the word (جالس, "set") pluralized to (جالسون), in the nominative case the plural suffix by add (ين) to form masculine plural (باحثين, "researchers") also (عامل,"worker") pluralized to (عاملين) in the genitive and accusative cases[11].

The Feminine Plurals: is Sound plural which called for more than two ,by adding suffix (ات) to singular, for example the word (باحث, "researcher") is a singular it can be pluralized by adding the suffix (ات) to form the feminine plural (باحثات , "researchers") also the word (طالب,"student") pluralized to (طالبات,"students").

Based on above, The Sound plurals keeps the structure of the singular, and get the plural by adding certain suffix, this feature has made getting to plurals easy process, as well as access to a single too. This feature made facilitate the search and retrieve process in IRS with high accuracy by using Arabic Stemmer in pre-processing stage. For example if there are three documents D1 and D2 and D3, any document contain some of words as:

D1:"ما روي عن المسلمين في رمضان".

D2: "حصن المسلم".

D3:" يعتبر البروفسير عزالدين محمد عثمان اشهر الباحثين في مجال علوم الحاسوب في السودان".

assuming these documents indexed using Arabic analyzer which use light stemming, and then the user type the following query Q1:" حديث اخرجه البخاري و مسلم", IR system will retrieve D1and D2, although the word ("مسلم",Muslim) in a Query and (المسلمين,"Muslims") in D1 are not identical letters, also the system will retrieve D2 although (المسلم) which find in a Query and (مسلم) in D2 are not identical letters, but the system recognizes that the word (المسلمين) is a plural of words (مسلم) because there are fixed rules governing the Sound Plural and the light stemming known how to deal with this type of plural by removing prefix and suffix from word, in our example the prefix is (ال) and suffix is (ين) after stemming the word result is (مسلم), without suffix or prefix.

If user types query Q2: "باحث سوداني", for this query the system can retrieve D3 although the word (الباحثين,"researchers") in document is not identical letters with word in query (باحث,"researcher").

2.3.3.1.1 Sound Plurals and Information Retrieval system

Information Retrieval System such as search engines can handle this type of plurals, that because Sound plurals has a particular rule governed, in addition the structure of the singular remain unchanged, all this feature made this type of plurals (Sound plural) easy to identify and facilitated the understanding and it is applicable

to Information Retrieval Systems, but another type of plurals which called Broken Plurals have not rule governed, and the structure of the singular will change, and the Information Retrieval Systems cannot handle this type of plural, this study shows how to recognize Broken plurals, and discusses in detail in the following paragraphs about this type of plurals[3].

2.3.3.2 Arabic Broken Plural (BP)

The Broken Plurals (جموع التكسير) is another type of plurals in Arabic language, in this type of plurals there are no a specific rules governed, each word treated quite differently based on some patterns as will show in this section, where it is difficult to identify them. Broken plurals similar to irregular nouns in English (e.g.: tooth/teeth), but this type of plurals are very common in Arabic, it represents more than 40% of the plurals in Modern Standard Arabic, while the remaining percentage 60% is assigned to the other types of plurals, sound masculine and feminine plurals[3, 12].

Although BP does not have fixed rule depend on, but it comes in a set of patterns where these patterns were divided into three categories, namely: (جموع قله ، جموع كثره ،صيغ منتهى الجموع), each of this category has a set of patterns.

2.3.3.3 Jemoa Gela (جموع القلة)

Jemoa Gela JG one of Broken Plurals patterns, this category has several patterns which can be called for plurals from three to ten[13]. This category has four patterns as in table 2.1.

Table 2-1: Jemoa Gela Patterns(انماط جموع القلة)

No.	Pattern	Examples/plural	singular	Plurals
1	أَفْعِلَةٌ	أَجْهِزَةُ	جهاز	3<=10
2	أَفْعُلْ	أنْفُس	نفس	3<=10
3	أَفْعالٌ	أقمار	قمر	3<=10
4	فِعْلَةٌ	فِتْيَةٌ	فتى	3<=10

2.3.3.4 Jemoa katharh (جموع الكثرة)

This category has several patterns which can be called for plurals more than ten [13]. This category have several patterns, some of these patterns are shown in the Table 2.2 below.

Table 2-2: Jemoa katharh patterns(انماط جموع الكثرة)

No.	Pattern	Examples /Plural	singular	Plurals
1	فَعَلَه	جَهَلَه	جاهل	>10
2	فُعَلَه	رُمَاه	رامي	>10
3	فُعْل	حُمَر	حمار	>10
4	فُعُل	زُرُع	زراع	>10
5	فُعَل	صُغَر	صغرى	>10
6	افعلاء	اغنياء	احتواء	>10

2.3.3.5 Syntax of Montahaa Jemoa SMJ (صيغة منتهى الجموع)

SMJ is the Broken Plurals which the third letter is Alef (ا) after this letter three or two letter which middle letter is vowel. This category has several patterns which can be used for plurals shaped of more than or equal to three[14]. This category have several patterns. Some patterns have five characters and other have six characters shown in the Table 2.3.

This research concerns analyzing this category of patterns(صيغ منتهى الجموع) it investigates how to identify patterns and make analysis to extract rules help to distinguish is the word represent a plural for singular or not. Then get all singular forms of word if it represent plurals. Chapter 3 discusses the following patterns **tfaaeel(تفاعيل)**, **fyaeel(فياعيل), faaeel(فعاعيل).**

Table 2-3: Syentax of Montahaa Jemoa patterns (انماط صيغ منتهى الجموع)

No.	Pattern	Examples/plural	singular
1	فواعل	لوائح	لائحه
2	فعائل	رسائل	رساله
3	فعالل	دراهم	درهم
4	مفاعل	مسارح	مسرح
5	افاعل	اصابع	اصبع
6	مفاعيل	مناشير	منشار
7	فعاليل	عصافير	عصفور

2.3.3.5.1 Broken Plurals challenge for Information Retrieval Systems

The broken plurals identification represent problem especially for Information Retrieval Applications. It is difficult to deal with Arabic broken plurals and reduce them to their associated singulars, because that there is no a specific rule governed, each word treated quite differently where it is difficult to identify them, and no obvious rules exist, also there are no standard stemming algorithms that can deal with this type of plural [11].

There are some patterns are analyzed to extract some rules which are used to identify and determine if a word represents a plural or not, and another patterns that were difficult to be analyzed. To denote the importance of this study below are some examples derived to explain some of these challenges.

Challenge 1: Some patterns are difficult to be identified due to their plural structure have three letters. The root of words in Arabic language is (Fa-al,"فعل") word and that make identify is this word represent plural is difficult process for IRS. For instance the word (جلس) is singular, and the word (صُغر) represent plurals.

Lucene is one of IRS, it use Arabic Analyzer to preprocess the input words first. Then, it stripes and return the origin word without suffix, prefix or special characters (التشكيل) supplemented to the word.

Challenge 2: Some words that are match a plural form but they are not. For instance the suitable pattern for word (اختباء) is (افعلاء), since this word does not represent a plural. For more explanation in contact, the word (اغبياء) represent a plural [3].

Challenge 3: stemming process deleted some of the original characters for some words as a suffix, therefore the meaning of the word is missing.

This study is mainly focuses on challenge 2. Next section explore simply the problem statement this research which is based on challenge 2. When re viewing some State-of-art works, current Arabic_based IRS failed to recognize the BP. For instance, when write the term (تقارير), documents that include (تقرير) are not retrieved. So, the precision of these search engine are low. Therefore, this study is proposed to increase such precision. See the example below:

D1: "يعتبر الوليد بن طلال من أغنياء العالم وهو من المملكة العربية السعودية ".

D2: "أسامه داوٗود غني لانه يمتلك مجموعة من الشركات و يعتبر من أثرياء السودان".

assuming that was indexed these documents using Arabic Analyzer, if the user write query Q1: "اغنياء رجال الاعمال " and then press the search button, the system will retrieve the first document D1 only, because it contains the word (اغنياء,"Rich"), but the system does not retrieve the second document D2 although it contains the word (غني,"Rich") which represents a singular of the word (اغنياء) where it is must be retrieve this document as result, but do not been retrieved because failure to recognize that word (غني) represent the singular of word (اغنياء), which the word (اغنياء) is the broken plural.

2.4 Information Retrieval Data Set

The data set is very important component for information Retrieval systems, it contain large number of documents and files, and our analysis depend on word which extracted from this data set. This research depend on Watan-2004 corpus.

2.4.1 Watan-2004 corpus Description

Corpus of Arabic newspaper texts which was developed for topic identification. The corpus contains 20291 articles downloaded from the web. The punctuation was deleted from the corpus. The topics covered are culture, religion, economy, local news, international news and sports. The corpus can be downloaded for free from the

webpage, but it should be used only for research purposes[15]. Appendix B show an example of watan-2004 documents.

2.5 Baseline Method

In order to obtain the essential information needed for the corpus analysis, and also for experiments reported in this thesis, the Lucene IR system was used. Lucene is an experimental information retrieval system that has being extensively used in previous editions of the CLEF, NTCIR and TREC joint evaluation experiments [6]. ***Lucene*** used for Information Retrieval System purpose which work under java to provide IR processes such as searching and indexing.

2.6 Evaluation Measure

In pattern recognition and information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

To evaluate is the research achieve their objectives researchers need to evaluate the result. There are some measure used for evaluation. Precision and Recall and F-measure is most measure used for this purpose based on related studies. The performance of an IR system can be measured in different ways, depending on retrieval task and used relevance judgment. If the binary relevance judgments are employed for assessing Documents, then precision and recall measures can be used.

In the next section, will give a definitions for this measures.

Precision:

Define as the ratio of the number of retrieved relevant documents over the total number of documents retrieved[6].

Recall: The recall is defined as the fraction of relevant documents that are retrieved.

The F-score measure:

Is used to balance system performance on both “Precision” and “Recall” measures.

These measures will use these metrics to evaluate the results obtained after the implementation in the next chapter “Result and discussion”.

2.7 Related work

As mentioned in Section 1.1 IR has existed from several decades, recently was associated with the field of Internet and become one of important area of research. Information Retrieval System (IRS) provide quite good results for Latin characters-based languages. However, they still show many weaknesses when used to search for other languages such as Arabic. Searching in Arabic language meets fundamental problems related to certain reasons one of them Arabic is morphological base language. Arabic BP represent one of important challenges faces IRS application.

There are only few studies addressing the problem of broken plurals, they differ from each other[9]. Some of these studies work on deriving broken plurals from their singulars or roots, while others aimed at extracting singulars from plural forms [3, 16]. To the best of our knowledge, there are two studies proposed approaches to identify Arabic Broken plurals, and some studies used these approaches for other topics such as translation as shown in the following paragraph.

Goweder, et al. 2004 and Goweder, et al. 2005 proposed three approaches to identify Broken Plurals (BP). The first approach is the ***Simple Broken Plurals Matching Method***. The basic idea is to get a word, use light-stems to produce morphological information such as stemming prefix and suffix, then returns TRUE if the word match one BP patterns in the list or FALSE. This method identifies plurals by match the word with broken plurals patterns by checking some letters in the word with equivalent letters on same positions in pattern. Although it is simple method that made it easy to implement but the main problem with the simple BP matching approach is that the BP patterns are too general to achieve a good performance, which means there are several words have the same pattern but they do not represent BP. The results showed that the ***Simple Broken Plural Matching*** approach has low precision (13.73%) - on a test set of about 187,000 words.

To improve the performance of identification the same authors proposed ***Restricted Broken Plural matching Method***. In this approach, the development of

Simple Broken Plural method it increase the precision which obtain more specific BP patterns by restricting the original one. The main idea is to allow only a subset of the alphabet to be used in the meta characters f (ف), (ع), and (ل) positions of the patterns was the restricted matching method, in which the broken plural patterns are used to detect broken plurals according to sets of rules that govern their applicability. This method makes some tests to identify a word that represent plural or not. Those steps are summarized as the following. First check the word with BP patterns, if it matches one of these patterns, then checks the position of character based on rule that obtained from analyzed patterns. The results of this approach showed an increase in the precision reaching about 75%. The third approach for identifying broken plural was built on the top of the previous. This approach used a dictionary which lists broken plural stems. This dictionary was constructed automatically by extracting all instances of broken plural stems that match broken plural patterns. Next, sets of rules, as in the previous approach, were extracted. Results showed that a significant improvement in precision, reaching 92%, compared to other two approaches [3].

Abduelbaset,et al,2008 they used the Simple Broken Plurals Matching approach to identify Arabic Broken Plurals for translation purpose. The proposed method consists of four main phases, these are: The Simple Matching phase, and Arabic to English Translation phase, and English stemming phase, and finally phase is English to Arabic Translation phase. But this study focus only in this paper on what they used to identify an Arabic broken plurals. It will explore the first phase. In Simple Matching phase, they have used a simple matching algorithm as referred in previous section to recognize all words to see if it is possibly broken plural, as a first phase. In the simple BP matching approach, they use Arabic light stemmer to light-stem all the words by removing prefixes and/or suffixes attached to a word and ignores any infixes encountered. The resulting stem is compared against a set of 41 broken plural patterns found in traditional grammar books of Arabic. The stem matches a BP pattern if and only if they have the same number of letters and the same letters in the same positions, excluding the consonants f (ف), Q (ع), and l (ل) of the basic root f Q l (فعل , "to do") found in the pattern. If the stem matches one of the BP patterns in the list, then it is initially classified as Broken Plural[9].

Rammal Mahmoud, et al, 2009 they explained that a broken plurals are not handled by current Arabic stemmers. One technique to address this problem is to use N-grams character. Although broken plurals are not derived by attaching word affixes, many of the letters in broken plurals are the same as in the singular forms (though sometimes in a different order). If words are divided into character N-grams, some of the N-grams from the singular and plural forms will probably match. This technique can also handle words that have a stem but cannot be stemmed by a stemmer for various reasons. The retrieval scores show that stem-based N-grams are more effective than word-based N-grams for retrieval. The probable reason is that some of the word-based N-grams are prefixes or suffixes, which can cause false matches between documents and queries[17].

The use of character n-grams to detect the broken plural is a solution that is proposed by (Xu, et al., 2002). In this approach, the developers implemented ***n-grams*** created from stems as well as n-grams from words. Results concluded that stemming by the use of n-grams with the stemmed word is better than n-grams with the word-base. The reason behind that is some of the word-based n-grams are prefixes or suffixes. They claimed that there may not be a straight-forward algorithm to handle the broken plural in Arabic [18]. As stated in appendix C.

2.8 Summary

In this chapter researchers reviewed an Information Retrieval concepts, Models, Applications. Also gave a brief about Arabic language and challenges which faces Arabic language for Information Retrieval System. Additionally reviewed an Arabic number system which is explained the types of plurals in Arabic Sound plurals and Broken Plurals. Challenges which face Broken Plurals identification are also discussed. This chapter reviewed the selected Data set, the evaluation measures, and reviewed the related work to this research.

The next chapter explains the proposed methodology and its stages which are followed to solve the research problem.

Chapter 3
Research Methodology

3.1 Introduction

This chapter presents the methodology used in this research, and discusses the stages taken to carry out this research. It describes the design and implementation of chosen methods in achieving the goal and objective of this research. This research aims to identify Broken Plurals (BP) and analyzes BP patterns, which analyzed patterns based on words which extracted from watan-2004 corpus. One of the objectives of this study is to identify word which represent broken plural and then get the singular of that word to enhance retrieval.

Base line system is ***Lucene*** packet which work under java to implement Information Retrieval processes (IRP) such as indexing and preprocessing, and Searching. The result will calculated by compare the Base line system and proposed methodology.

There are six section in this chapter where section 3.1 is for the introduction. Section 3.2 present the stages that methodology followed. Section 3.3 explains pattern which analyzed. Finally, section 3.5 is about evaluation and reporting of the result.

3.2 Research Framework

Figure 3.1 shows the general framework of the proposed method. In next sections, and going to describe each component found in the framework.

3.2.1 Query Statement

Query Statement, usually user typing a query to get an information or documents he need. Already there are a set of documents were indexed as in Figure 3.1. The query passes through some stages which will show in detail in the following section.

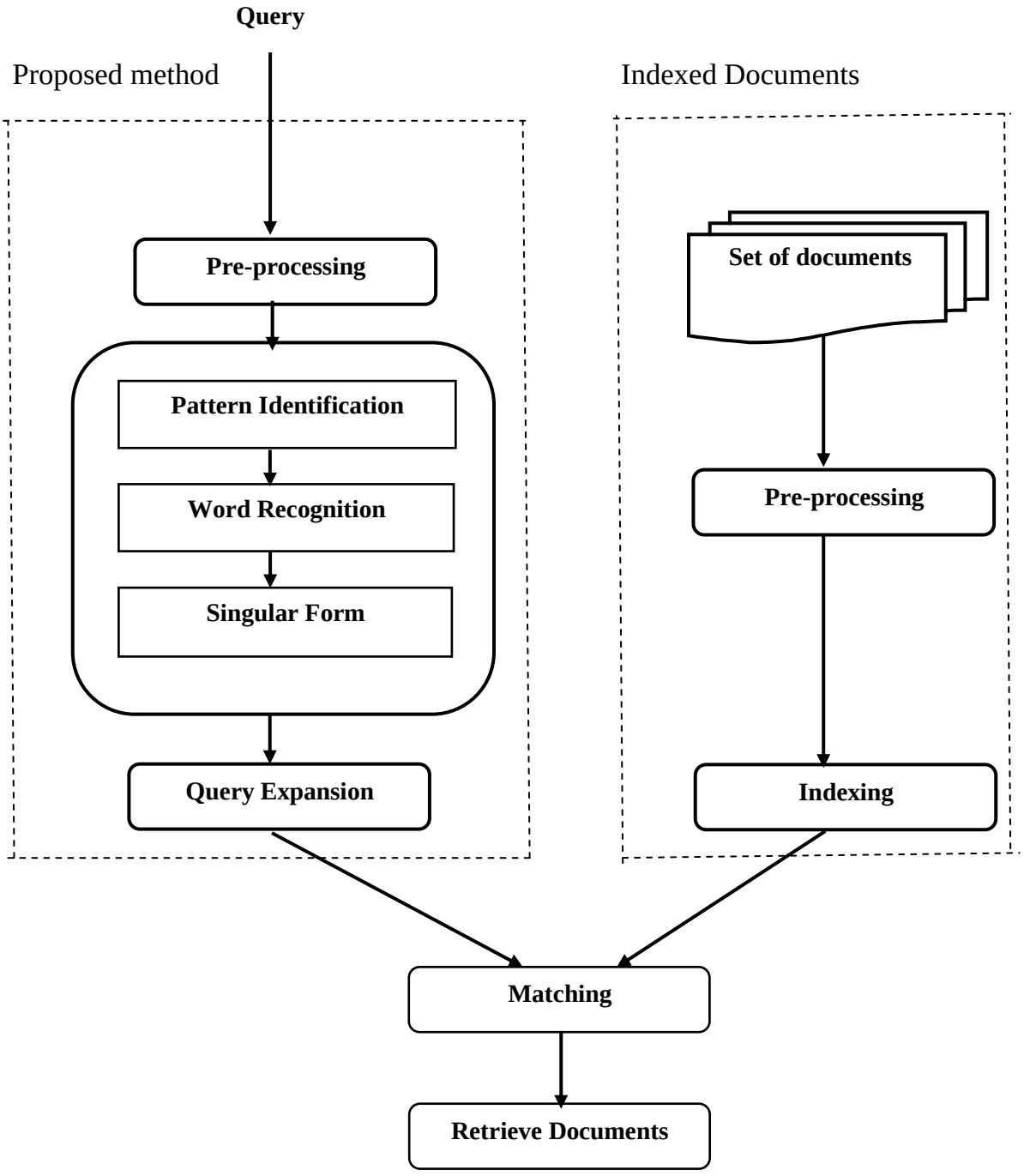

Figure 3-1 : Methodology Framework

3.2.2 Preprocessing

Include set of processes such as "***Tokenization***" which divides the query statement into set of word, and "***Stopword***" which concern to remove some words, Stopwords differ from one language to another. There are set of words should be

removed such as "من". For instance, if a query is "من هو عبدالله الطيب" the query would be edited to "عبدالله الطيب".

This process removes all stop word as stated in appendix A. Also capital letters should converted to small letters, this process called "***Lowercase***", for example if a query is "Sudan University" it will be changed to "sudan university". Other processing steps which is called "***Stemming***"used to remove prefix and suffix and return original form. This preprocessing steps used for both *User query* and *Indexing*.

3.2.3 Broken Plural Identification

To identify whether the word represent Broken Plural should follow three steps. **Step 1**: Pattern Identification, **Step 2**: word Recognition, **Step 3**: Singular form[3]. Figure 3.2 shows these steps (Black-Box Framework)

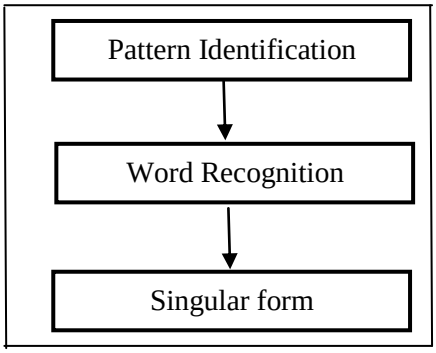

Figure 3-2: Black-Box Framework

Broken plural identification is difficult process because the word may come on one of BP patterns and it may represent a plural or a singular as referred early in Chapter 1 (see Table 1.1).

The difficulty in identifying the BP is depends on the phonetic patterns, which is difficult to be measured. The word structure was differ from the singular form and plural form by add or remove some characters. For example: the word (غني) which mean rich is singular form, the plural form of this word is (اغنياء) which mean (rich people), from above noted that, the singular form changed to plural form by adding three character to the singular.

This methodology proposed a solution to identify some of these patterns. As referred in previous chapter (see Table 2.3), a BP have a set of patterns, some of

these patterns have five characters and other patterns have six characters. This study focus on some patterns which have six character it called by Arabic (صيغ منتهى الجموع) and extract some rules for to identification. After identifying them, then should check is the word represent plural or not. If the word represent plural, then get the singular form and then make a query expansion to add a singular form to the query. For example if the user typed a query " تقارير الإدارة ” which mean rich people in the world the query must change to " تقارير تقرير الإدارة ". So, the query is expanded to three words rather than two. This research can enhance information retrieval when query include broken plural words.

3.2.3.1 Step one: Pattern Identification:

As referred in the previous chapter (see Section 2.3.3.2) the plurals in Arabic divided into two types Sound Plurals and Broken Plurals. Research focuses on BP patterns exactly focus on Syntax Montaha Jemoa (SMJ) صيغ منتهى الجموع with six letters. There are a set of patterns with six letters. To identify pattern of each word must do some check for letters of the word.

There are some letters distinct the word belong to SMJ (صيغ منتهى الجموع) patterns and some letters determined the specific pattern because there are some pattern which have six letters.

3.2.3.1.1 Identify the word on SMJ patterns

There are three following letters must checked to determine whether the word belong to SMJ with six letters or not. The first check is to specify the number of letters of the word, it must be consist of six characters. The second check done for third letter of the word, it must equivalent to (alef,"ا"). The third check the character in position number five and it must equivalent to (yaa,"ي"). If these three checked are true, then the word on the BP and it represent one of the SMJ with six letters. Table 3.1 illustrate SMJ pattern identification.

Table 3-1: SMJ pattern identification

6	5	4	3	2	1	Position الموقع
	ي		ا			Letter الحرف

Table 3.1 showed that, to identify is the word belong to SMJ patterns, then must do 3 checks (word contains 6 letters, the character in position 3 equivalent to (ا), the character in position 5 equivalent to (ي)).

From the above noted that, all six patterns must be achieved in these checks, and if one check is false that means the word does not belong to these patterns.

- **Identify the specific pattern of word**

After the process of defining the word belong to SMJ patterns with six letters, then must identify the actual pattern for word, because there are a several patterns have six letters as in Chapter 2 (see Table 2.3). The process to identify the specific pattern need to check additional letters but this check differ from one pattern to another. Researchers will explain in detail about these checking and how to identify specific pattern in next section in this chapter.

Finally, the output of Pattern Identification step is the extraction all of words from query which come on the SMJ patterns.

3.2.3.2 Step two: Word Recognition

This step helps us to identify each word came in one of BP patterns represent plural or not, because not any word come on BP patterns represent plural or not. As stated before any word came on BP pattern represents plural as it may be singular. Researchers used watan-2004 corpus to categorize all input words either they represent plural or singular. This categorization depends on specific patterns. After analysis of those words, researchers deduce some rules that help us to identify the BP words, and these rules will be added to IRS to identify all words represent plurals.

The output of this step is to identify which word in the query represent plural to take it as input to the next step.

3.2.3.3 Step three Singular forms

After pattern identification as in step one, and word identification as in step two, then should get the singular form of word which appear in the query and represent plural. The output of this step is getting all singular of words represent BP. The importance of this step is shown in the next stage which called "*Query Expansion*".

3.2.4 Query Expansion

Query expansion is the process of adding search terms to a user's query terms. Study aims get the singular form of each word represent BP appeared in the query. This to increase document retrieval by adding the singular form to the existed query. The reason for this addition is that indexing algorithms differentiate between singular form and plural form for Broken Plurals words. As showed the problem in Section 1.1. For example, if user typed a query:"تقارير الادارة" then the query must be edited to "تقارير تقرير الادارة" by adding the singular form of word (reports,"تقارير"), if system don't make expansion for a query, it will lose all documents containing the word (report,"تقرير") as a result.

3.2.5 Matching

In this stage must do match between query expansion and exist indexed documents to retrieve all documents that satisfy user query.

In the next section will show patterns that can be used to prove the proposed methodology. Also in the next section will explore how applied those stages on SMJ patterns (تفاعيل ، فعاعيل ، فياعيل) which belong to Broken Plurals patterns.

3.3 Pattern Analysis

As referred in Section 2.3.3.2 Broken Plural have a sets of patterns divided into three categories, our research focuses on SMJ patterns. This Section give an analysis of three patterns of SMJ, and make analysis by applying the methodology stages for each patterns.

3.3.1 Pattern Tfaaeel(تفاعيل)

As explored the (Tfaaeel,تفاعيل) pattern is one of SMJ patterns with have six letters which belong to BP patterns, researcher apply the proposed method steps on this pattern.

3.3.1.1 Pattern Tfaaeel(تفاعيل) identification

To identify the word on this pattern, first must check the number of character if it equal 6 then check the character in position 3 if it equivalent to (alef,"ا") then

check the character on position 5 if it equivalent to (yaa,"ي ") then the word on SMJ with six-party BP patterns. Then we need to know know is the word on Tfaaeel(تفاعيل) pattern by check the character on position 1 is it equivalent to (taa,"ت") then the word in pattern Tfaaeel(تفاعيل). Table 3.2 illustrate how we identify Tfaaeel pattern.

Table 3-2: Tfaaeel(تفاعيل) pattern identification

6	5	4	3	2	1	Position الموقع
	ي		ا		ت	Letter الحرف

Figure 3.3 showed that, to identify is the word belong to (Tfaeel,تفاعيل) pattern, then must do 4 checks (word contains 6 letters, the character in position 1 equivalent to (ت), the character in position 3 equivalent to (ا), the character in position 5 equivalent to (ي)). All words Extracted on Tfaaeel(تفاعيل) pattern shows in Table 3.3

Table 3-3: Words extracted on Tfaaeel(تفاعيل) pattern

تباريز	تراثيا	**تصاميم**	**تفاعيل**	**تلاميذ**
تباشير	**تراحيب**	**تصاوير**	تفانيا	**تماثيل**
تجاريا	**تراخيص**	**تضاريس**	**تفاويض**	تماشيا
تجاعيد	ترافيس	**تعابير**	**تقارير**	تناديك
تجاويف	**تراكيب**	**تعاريف**	**تقاسيم**	تناميا
تحاشيا	**تراكيز**	تعاطيك	**تقاطيع**	**تناهيد**
تحاليل	**ترانيل**	**تعاليم**	**تقاليد**	**توابيت**
تخاريف	**ترانيم**	تفاديا	**تقاليع**	**تواريخ**
تدابير	تساويا	تفاديت	**تكاليف**	**تواقيت**
ترابيا	تشافيز	**تفاسير**	تلافيا	**تواقيع**
تراتيل	**تصاريح**	**تفاصيل**	**تلافيف**	

3.3.1.2 Word Recognition

This step deduce some rules that help to identify is the word represent plural or not.

As in Table 3.3 the total number of words retrieved are 54 words, and noted that there are 39 words represent plurals and 15 words don't. All these words extracted from watan-2004 corpus.

Based on analysis of these words, found that each word in position six equivalent to (alef,"ا") does not represent plurals, and each word in position six equivalent (kaf,"ك") does not represent plurals. Otherwise the word represent BP.

Some of these word represent BP although they not in (*Tfaeel*,تفاعيل) pattern such as (تلاميذ ، توابيت ، تواريخ), but also the study included their singular form.

3.3.1.3 Singular forms

Word which come on this pattern have four singular forms. The first form is (tfeel,"تفعيل"). Most words which are extracted came in this form, its singular are inform (تفعيل).

For example, the word (reports,"تقارير") its (report,"تقرير"). Figure 3.2 illustrate an example of this singular form.

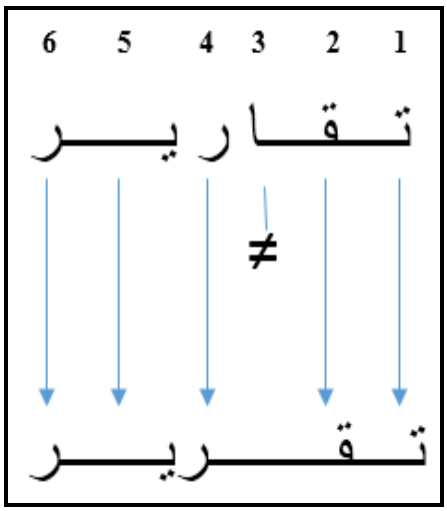

Figure 3-3: Tfeel (تفعيل) singular form example

The second singular form is (tfeaal,"تفعال"). Noted that, one word came in this form such as word (Statues,"تماثيل") which it's singular is (Statue,"تمثال"). Figure 3.3 illustrate an example of this singular form.

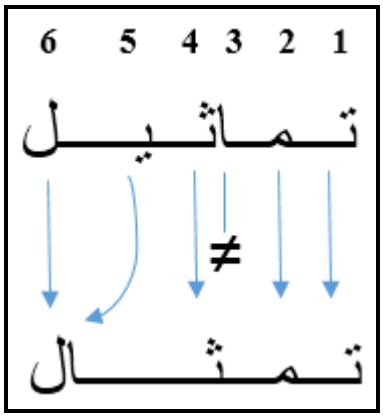

Figure 3-4: Tfaal (تفعال) singular form example

The third form is (tfelh,"تفعيله"), then note that some words came on this formula such as word (orders,"تعاليم") its singular is (order,"تعليمة").

The fourth singular form is (faalool,"فعلول"). Noted that one word came in this form such as word (coffins,"توابيت") which singular form is (coffin/ark,"تابوت").

3.3.2 Pattern fyaeel(فياعيل)

3.3.2.1 Pattern fyaeel(فياعيل) identification

To identify the word on this pattern, first must check is the word in six-party BP patterns by checking the character on position three if it equivalent to (alef,"ا") then check the character on position five if it equivalent to (yaa,"ي ") then the word on six-party BP patterns, after this checking, system must know is the word on fyaeel(فياعيل) pattern, by check the character on position two is it equivalent to (yaa,"ي") is it true then the word on pattern fyaeel(فياعيل). Table 3.4 illustrate Fyaeel patten identification.

Table 3-4: Fyaeel(فياعيل) patten identification

6	5	4	3	2	1	Position الموقع
	ي		ا	ي		Letter الحرف

Figure 3.6 showed that, to identify is the word belong to Fyaeel (فياعيل) pattern, then must do 4 checks (word contains 6 letters, the character in position 2 equivalent to (ي), the character in position 3 equivalent to (ا), the character in

position 5 equivalent to (ي)). Words extracted on fyaeel (فياعيل) pattern shows in Table 3.5

Table 3-5: Words extracted based on fyaaeel (فياعيل) pattern

بياجيو	دياويو	فيافيك
بيانيو	رياديا	فياليت
حياتيا	رياضيا	قياديا
حياديا	سياحيا	قياسيا
خياشيم	سياسيا	لياليا
خياليا	سياسيو	لياليك
دياجير	غيابيا	بيانيس

3.3.2.2 Word Recognition

To identify is the word represent plural or not, system must analyze set of words which obtained from previous step. As in Table 3.5, the total number of words are retrieved 21 words. Noted that, there are only two word represent plurals which is (خياشيم، دياجير) and 19 words don't. Based on analysis of those words found that, each word in position six equivalent to (meem"م", or Raa"ر",) represent plurals, otherwise the word does not represent plurals.

3.3.2.3 Singular form

Words that came on this patterns have one singular form (fyool,"فيعول"). Noted that, there are one word represent plurals came on this pattern it is (Gills,"خياشيم") which singular form of this word is (Gill,"خيشوم"). Figure 3.4 illustrate an example of this singular form.

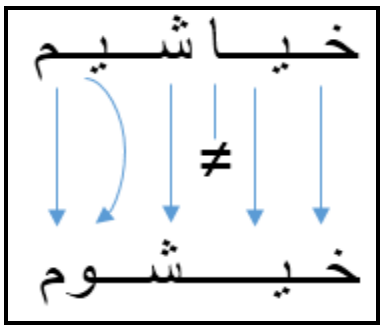

Figure 3-5: Fyool (فيعول) singular form example

3.3.3 Pattern faaeel(فعاعيل)

3.3.3.1 Pattern faaeel(فعاعيل) identification

To identify the word on this pattern first must check is the word in six-party BP patterns by checking the character on position three if it equivalent to (alef,"ا") then check the character on position five if it equivalent to (yaa,"ي ") then the word on six-party BP patterns, after this checking system must know is the word on faaeel(فعاعيل) pattern, must check the character on position two is it equal the character in position number four, is it true then the word on pattern faaeel(فعاعيل). Table 3.6 illustrate Faaeel patten identification.

Table 3-6: Faaeel (فعاعيل) patten identification

6	5	4	3	2	1	Position الموقع
	ي	ع	ا	ع		Letter الحرف

Table 3.6 showed that, to identify is the word belong to Faaeel(فعاعيل) pattern, then must do 3 checks (word contains 6 letters, the character in position 2 equivalent to the character in position 4, the character in position 5 equivalent to (ي)).

Words extracted all on faaeel (فعاعيل) pattern and shows in Table 3.7

Table 3-7: Words extracted on faaeel(فعاعيل) pattern

حنانيك	**شبابيك**	**فقاقيع**
خفافيش	ضبابيا	**قراريط**
دنانير	شبابيا	هلاليا
سلاليم	**عقاقير**	

3.3.3.2 Word Recognition

To identify is the word represent plural or not, must analysis set of words which obtained from previous step. As in Table 3.3 above the total number of words are retrieved 11 words, and noted that, there are 7 words represent plurals and 4 words don't. All these words extracted base on watan-2004 corpus. Based on analysis of those words, and then found that the character in position six for each word

equivalent to (alef" ا "or kaaf "ك") don't represent plurals, otherwise the word is represent BP.

3.3.3.3 Singular forms

Words come on this patterns have three singular forms the first form is (fa-aal,"فعال"). Also noted that, there are some words which extracted came in this form such as (Bats,"خفافيش") which its singular is (Bat,"خفاش"). Also the singular of the word (Windows,"شبابيك") is (window,"شباك "). The second form is (fayaal,"فيعال"). Also noted that, there are two words came in this form such as word (Dinars,"دنانير") which its singular is (Dinar,"دينار"). The third singular form is (foaalh,"فُعاله") .Noted that, one word came in this form such as word (Bubbles,"فقاقيع") which its singular is (Bubble,"فقاعه"). Figure 3.5 illustrate that, to acquire this singular form, must remove some letters from the plural and add a new character to the singular.

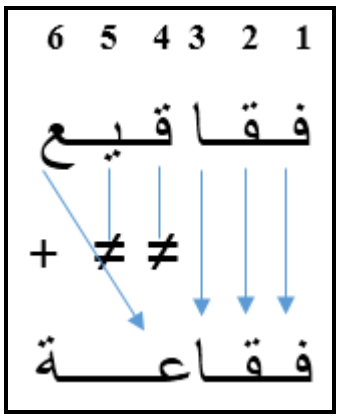

Figure 3-6: Foalh (فعاله) singular form example

To explore the proposed method work, Figure 3.7 shows the input and output of each stage of proposed method using actual query.

Previous sections conducted a study of some SMJ patterns (Tfaeel, Faaeel, Fyaeel, Faaleen) and apply the proposed methodology (Pattern Identification, Word Recognition, Singular Forms).

In the next section will give an information about the evaluation measures and their equation, which will use to evaluate the results before and after applying methodology.

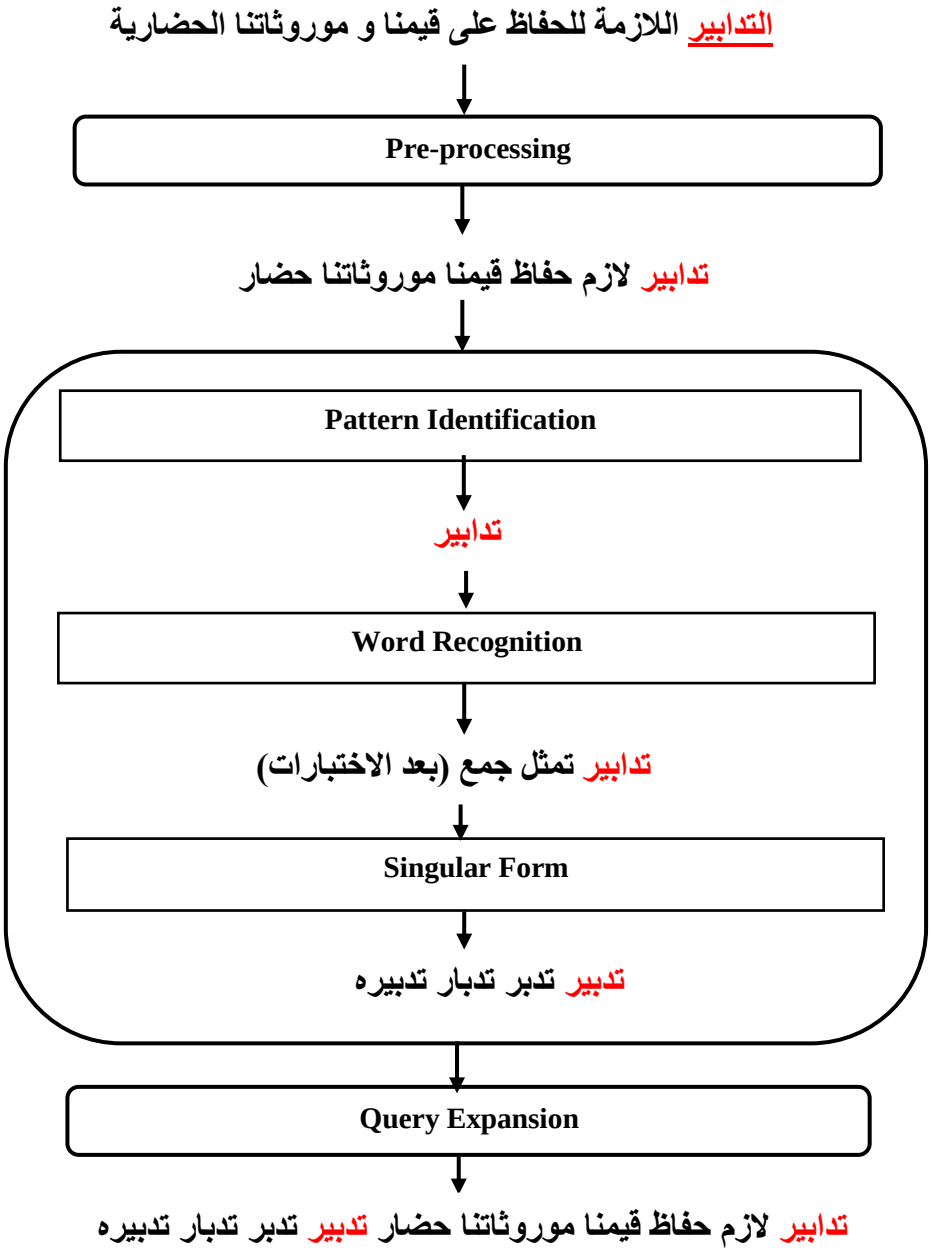

Figure 3-7: Proposed method stages for actual query.

3.4 Evaluation Measures

To evaluate the research findings, need to evaluate the result. There are some measure used for evaluation. Precision, Recall, F-measure is most measure used for this purpose as explored in literature review.

The performance of an IR system can be measured in different ways, depending on retrieval task and used relevance judgment. If the binary relevance judgments are employed for assessing Documents, then precision and recall measures can be used. In the next section, will show how they are computed.

Precision:

Define as the ratio of the number of retrieved relevant documents over the total number of documents retrieved.[6]

Equation (3.1) below illustrates how Precision is calculated.

$$Precision = \frac{number\ of\ retrieved\ relevant\ documents}{number\ of\ retrieved\ documents} \qquad (3.1)$$

Recall: The recall is defined as the fraction of relevant documents that are retrieved. Equation (3.2) below illustrates how recall is calculated.

$$Recall = \frac{number\ of\ retrieved\ relevant\ documents}{number\ of\ relevalt\ documents\ in\ the\ collection} \qquad (3.2)$$

The F-score measure:

Is used to balance system performance on both "Precision" and "Recall" measures. Equation (3.3) below illustrates how F-score measure is calculated.

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (3.3)$$

Researchers will use these metrics to evaluate the results obtained after the implementation in the next chapter "Result and discussion".

3.5 Summary

This chapter reviewed the stages of proposed method that used to solve the research problem (Pre-processing, Broken Plurals Identification, Query Expansion). This chapter gave a study for some BP patterns, which reviewed how to identify pattern of word and check is that word represent BP or not and also how acquire the singular of word.

The next point chapter “Results and discussion” which will show compare result between Base Line System (Lucene) before and after Appling our proposed method.

Chapter 4
Results and Discussion

4.1 Introduction

As mentioned in the second chapter, there are some previous studies that used to identify the Broken Plurals (BP). Proposed methodology as explained in the third chapter identify other BP patterns(SMJ patterns), which depends on the *Restricted Broken Plural matching Method* which proposed by *Goweder, et al. 2004 [44]*.

Researchers applied this methodology to three patterns of Syntax Montahaa Jemoa (صيغ منتهى الجموع) which belong to BP patterns. Researcher selected a sample documents from watan-2004 corpus (100 documents). Some of these documents contain the words which represent a plural and other documents contain a singular for the same word. Some queries are formulated and determined the documents relevant for each query, and then calculated the results using measures Precision and Recall and F-measure.

Researchers calculated the results before applying the methodology Based Line System (Lucene), and also calculated the results after applying of the methodology (after identification of the Arabic BP). Results calculated only for two patterns (تفاعيل ، فعاعيل) because there are enough documents in the watan-2004 corpus.

Next section shows the result based on sample data set which selected from watan-2004 corpus. Some queries were formulated which contains words belong to SMJ patterns and then calculated the compare between results before and after applying our proposed method.

4.2 The Result Calculated For Patterns

The next section shows the results calculated for two patterns by write some queries and retrieved documents based on these query before and after applying proposed method.

4.2.1 The Result for Faaeel (فعاعيل) pattern

Sample collection contain 30 documents, some documents containing the word (windows,"شبابيك") which represent BP on (Faaeel,"فعاعيل") pattern, and other documents contain the word (window,"شباك"), which represent the singular of word "شبابيك". Some query was formulated to calculate the results.

Query (1):"عوائد شبابيك أماكن العروض"

This query contains BP word (شبابيك). This query formulated because there are some documents related of this query which contains information about the income of place which show films, and then results calculated for Baseline System and after applying the proposed method. Tables 4.1, 4.2 shows the results that were acquired.

Table 4-1: Query (1) Baseline result for (Faaeel,"فعاعيل") pattern

No. Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Precision	Recall	F-measure
All	**11**	**18**	**1**	**0.0555**	**0.0909**	**0.0689**
Top 15		**15**	**1**	**0.0666**	**0.0909**	**0.0768**
Top 10		**10**	**1**	**0.1**	**0.0909**	**0.0952**
Top 5		**5**	**1**	**0.2**	**0.0909**	**0.1249**

Table 4.1 showed that, the number of documents retrieved are (18). The relevant documents in the collection equal (11) and the documents retrieved from BaseLine system equal (1) document while there are (10) relevant documents system fail to include it as result. Figure 4.1 illustrate comparison of results between documents retrieved for Baseline system and after applied proposed method for Query (1). Then the precision and recall and F-measure become low. While the precision for all documents retrieved equal to (0.0555), and also, the value of Recall equal to (0.0909), and F-measure equal to (0.0689). Based on Table 4.1, the value which calculated for these measures explore the weakness of the Base Line system when compared with result acquired by proposed method. Figure 4.2 Illustrate comparison of results between measures for Baseline system and after applied proposed method for Query (1).

Table 4-2: Query (1) Proposed method result for (Faaeel,"فعاعيل") pattern

No. Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Precision	Recall	F-measure
All	**11**	**37**	**11**	**0.2972**	**1**	**0.4582**
Top 15		**15**	**6**	**0.4**	**0.5454**	**0.4615**
Top 10		**10**	**2**	**0.2**	**0.1818**	**0.1904**
Top 5		**5**	**1**	**0.2**	**0.0909**	**0.1249**

From Table 4.2 noted that, the total documents retrieved are (37). The relevant documents in the collection equal (11) and the relevant documents retrieved after applied the proposed method equal (11). Then the measures was increased, and the recall measure equal (1) which mean all relevant documents in the collection was retrieved, and enhanced the precision become (0.2972), and also noted that, the F-measure was increase and equal (0.4582). Figure 4.2 Illustrate comparison of results between measures for Baseline system and after applied proposed method for Query (1).

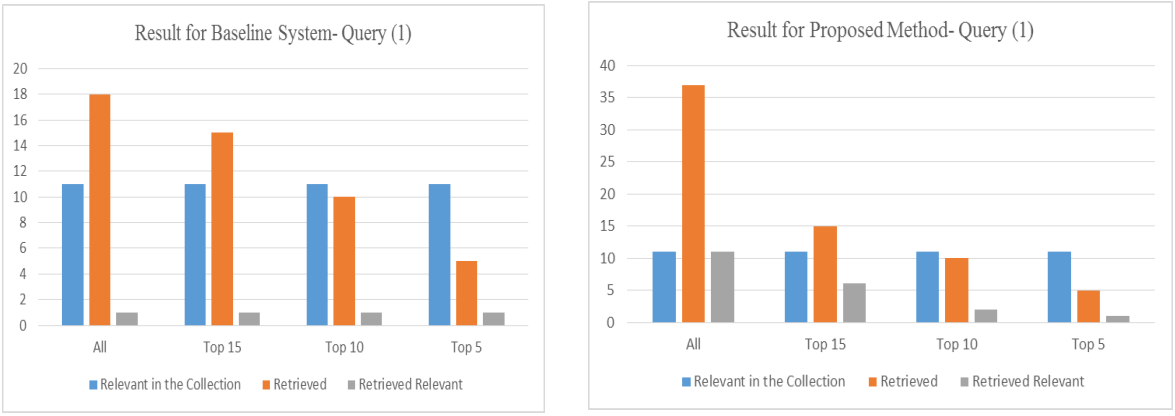

Figure 4-1: Query (1) comparison for retrieved documents.

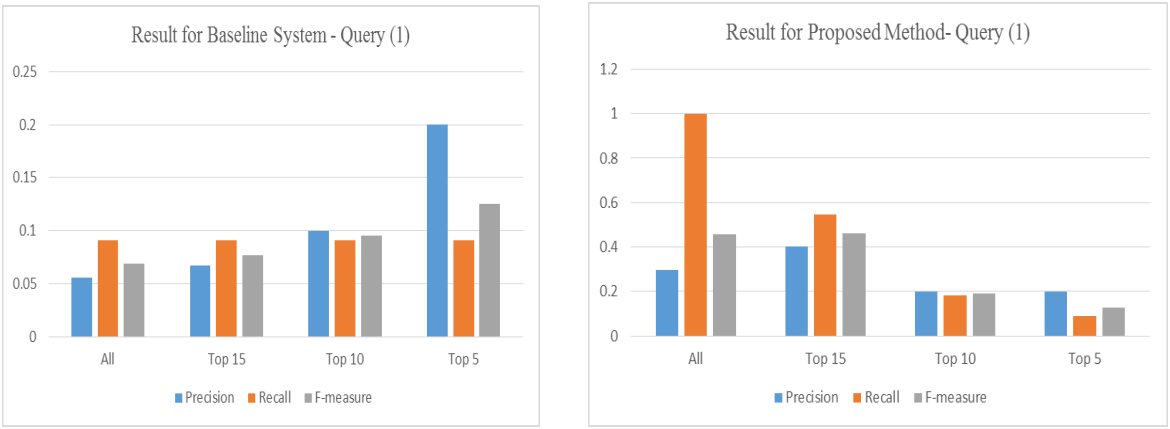

Figure 4-2: Query (1) comparison for measures (Precision,Recall and F-measure).

4.2.2 The Result for Tfaeel (تفاعيل) pattern

Two query was formulated on this pattern the Query (2) contains the word (licenses,"تراخيص") and second query contains the word (handling,"تدابير"). Sample collection contain 30 documents, 15 documents containing the word (licenses,"تراخيص") which represent BP based on (TFaeel,"تفاعيل") pattern and belong to SMJ patterns, and 15 documents contain the word (license,"ترخيص"), which represent the singular of word "تراخيص".

Query 2 :"تراخيص عقودات شبكات السيار"

This query contain BP word (تراخيص) and a query wrote because there are some documents related to this query which contains information about the communication company indentures. Results calculated for Base Line system. As in Tables 4.3 which shows the results that was acquired.

Table 4-3: Query (2) Baseline result for (Tfaaeel,"تفاعيل") pattern

Evaluate Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Precision	Recall	F-measure
All	**8**	**32**	**5**	**0.1562**	**0.625**	**0.2499**
Top 15		**15**	**3**	**0.2**	**0.375**	**0.2608**
Top 10		**10**	**3**	**0.3**	**0.375**	**0.3333**
Top 5		**5**	**1**	**0.2**	**0.125**	**0.1538**

Table 4.3 showed that, retrieved documents by BaseLine System equal (32), and the relevant documents in the collection equal (8) and the documents retrieved from BaseLine system equal (5) documents. Then there are (3) relevant documents was lost. And also Table 4.3 show the measure precision for all document retrieved equal (0.1562) and recall equal (0.625) and F-measure equal (0.2499). Figure 4.3 illustrate comparison of results between documents retrieved for Baseline system and proposed method for Query (2).

Table 4-4: Query (2) Proposed method result for (Tfaaeel,"تفاعيل") pattern

Evaluate Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Precision	Recall	F-measure
All	**8**	**41**	**8**	**0.1951**	**1**	**0.3264**
Top 15		**15**	**6**	**0.4**	**0.75**	**0.5217**
Top 10		**10**	**5**	**0.5**	**0.625**	**0.5555**
Top 5		**5**	**2**	**0.4**	**0.25**	**0.3076**

Table 4.4 showed that, the number of documents retrieved are (41). The relevant documents in the collection equal (8) and the relevant documents retrieved after applied the proposed method equal (8). Then the measures was increased, and the recall measure equal (1), then all relevant documents in the collection was retrieved, and enhanced the precision become (0.1951), and also noted that, the F-measure was increase and equal (0.3264). Figure 4.3 illustrate comparison of results between measures for Baseline system and after applied the proposed method for Query (2).

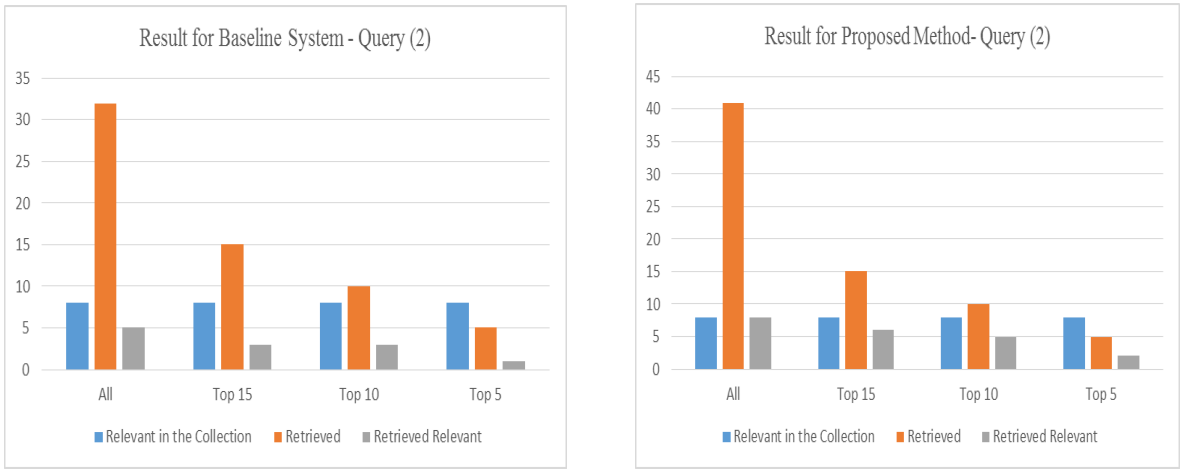

Figure 4-3: Query (2) comparison for retrieved documents.

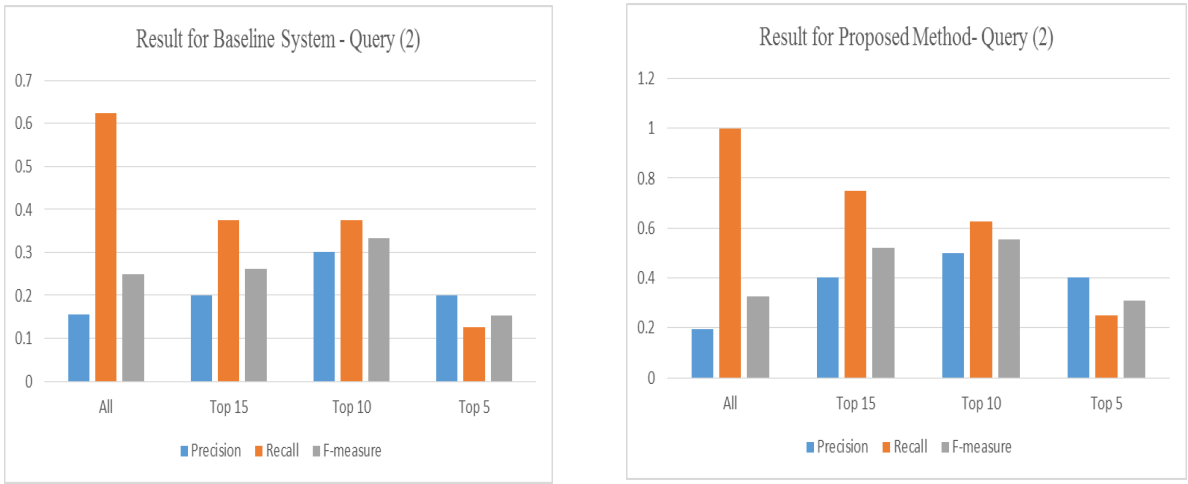

Figure 4-4: Query (2) comparison for measures (Precision,Recall and F-measure).

Other query was formulated and calculated the result which obtaind for baseline system and after apply the proposed method.

Query (3):"**التدابير اللازمة للحفاظ على قيمنا و موروثاتنا الحضارية**"

Table 4-5: Query (3) Baseline result for (Tfaaeel,"تفاعيل") pattern

Evaluate Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Precision	Recall	F-measure
All	**11**	**21**	**7**	**0.3333**	**0.6363**	**0.4373**
Top 15		**15**	**6**	**0.4**	**0.5454**	**0.4615**
Top 10		**10**	**5**	**0.5**	**0.4545**	**0.4761**
Top 5		**5**	**4**	**0.8**	**0.3636**	**0.4999**

Table 4.5 showed that, retrieved documents by Baseline System equal (21), and the relevant documents in the collection equal (11) and the documents retrieved by Baseline system equal (7) documents. Then there are (4) relevant documents was lost. And also Table 4.3 show the precision for all document retrieved equal (0.3333) and recall equal (0.6363) and F-measure equal (0.4373). Figure 4.3 illustrate comparison of results between documents retrieved for Baseline system and proposed method for Query (3).

Table 4-6: Query (3) Proposed method result for (Tfaaeel,"تفاعيل") pattern

Evaluate Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Precision	Recall	F-measure
All	**11**	**24**	**9**	**0.375**	**0.8181**	**0.5142**
Top 15		**15**	**9**	**0.6**	**0.8181**	**0.6922**
Top 10		**10**	**8**	**0.8**	**0.7272**	**0.7618**
Top 5		**5**	**5**	**1**	**0.4545**	**0.6249**

Tables 4.6 showed improved the result measures after applied the proposed method. Documents retrieved are (24), and the relevant documents in the collection equal (11) and the relevant documents retrieved after applied the proposed method equal (9). Then the measures was increased, and the recall measure equal (0.8181), and enhanced the precision which become (0.375), and also noted that, the F-

measure was increase and equal (0.5142). Figure 4.6 illustrate comparison of results between measures for Baseline system and after applied the proposed method for Query (3). The precision for top five document equal (1) that means all top five documents are relevant, this good result.

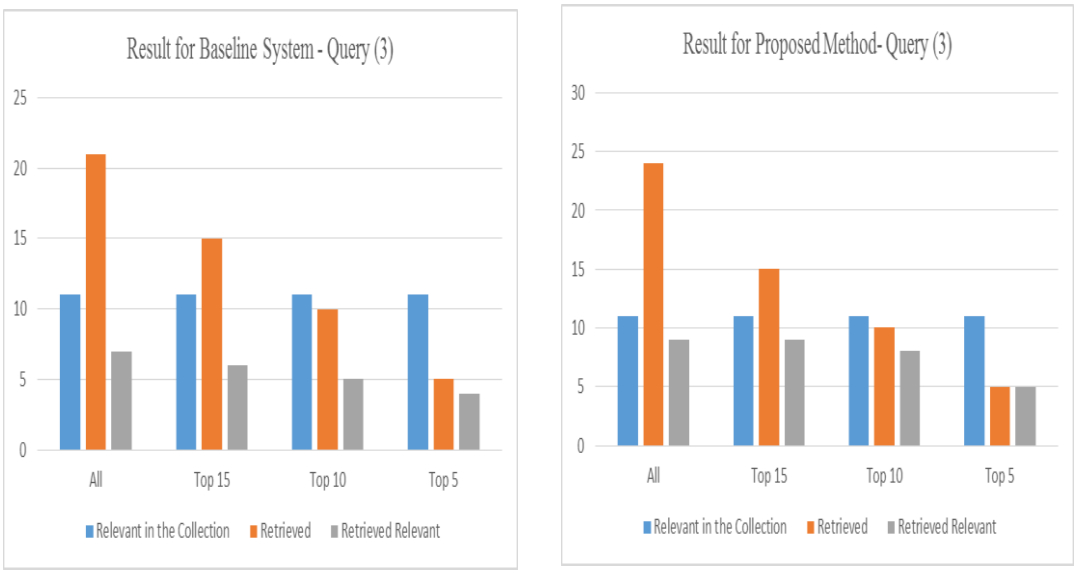

Figure 4-5: Query (3) comparison for retrieved documents.

Figure 4-6: Query (3) comparison for measures (Precision,Recall and F-measure).

Based on those obtained result, researcher can judged our proposed methodology could enhance retrieval by improve the precision and recall and F-measure for Syntax Montaha Jemoaa (SMJ) which belong to Broken Plurals.

4.3 Summary

This chapter gave more details about sample collection that were used to acquire results. Which extracted from watan-2004 corpus.

As explained calculation results of some SMJ patterns based on some of the queries that had prepared. To evaluation the results were compare between the results obtained before(Baseline system) and after applying the proposed method. The matrix that used to evaluate and compare results are (Precision, Recall, F-measure).

Based on those obtained result, researcher can judged our proposed methodology could enhance retrieval by improve the precision and recall and F-measure for Syntax Montaha Jemoaa (SMJ) which belong to Broken Plurals patterns.

Chapter 5
Conclusion and Recommendations

5.1 Conclusion

Identify Arabic Broken Plurals BP represents one of the important challenges for IRS. Our proposed methodology depends on ***Restricted Broken Plural matching Method*** to identify Arabic Broken Plurals which proposed by Goweder, et al2002 as mentioned in related work in Section 2.6. The study focuses on three BP patterns which are (تفاعيل فعاعيل، and فياعيل). To identify BP followed three steps. Step one is ***Pattern Identification;*** the output of this step is the extraction all words from a query which come on the BP patterns, and step two is ***Word Recognition;*** the output of this step is to identify which word in the query represent plural based on some rule. The output of this step is an input to the next step. Step three is ***Singular Form;*** the output of this step is getting all singular of words from a query which represent BP.

After applying these steps proposed method make a query expansion which added the singular form to the existed query to retrieve all document contain singular of BP word. Researchers calculated the results before and after applying the methodology using sample documents and compared with Baseline system (*Lucene*). Based on the result researchers can consider that, our proposed method is successful to identify Broken Plural words and enhance retrieval by referring back to this research hypothesis, found that the study could improve information retrieval for Arabic language especially for query which contain Broken Plurals words.

5.2 Recommendations

- This study was covered only three patterns of Broken Plurals Syntax Montaha Jemoaa (SMJ), we recommend to study remaining patterns.
- The study gave a solution to identify Arabic Broken Plurals and acquire a singular form but there are cost for searching process because there are some singular forms of words which come on same pattern.
- Preprocess may remove some letters from word as suffix, but it original letters, then it made difficult to identify because the word become vague.
- Some BP patterns have three letters, this property made it is difficult to identify, because the root of words in Arabic has 3 letters (فعل).

APPENDIX A

List of Arabic Stop words

ب	حين	فى	تم	ما	كما	بن	لدى	وقالت
ا	الى	في	ضد	مع	لها	به	نحو	وكانت
أ	انه	كل	بعد	هذا	منذ	ثم	هذه	فيه
،	اول	لم	بعض	واحد	وقد	اف	وان	لكن
عن	انها	لن	حتى	واضاف	ولا	ان	واكد	وفي
عند	ف	له	اذا	واضافت	لقاء	او	كانت	ولم
عندما	و	من	احد	فان	مقابل	اي	واوضح	ومن
على	و6	هو	بان	قبل	هناك	بها	يوم	وهو
عليه	قد	هي	اجل	قال	وقال	منها	فيها	وهي
عليها	لا	قوة	غير	كان	وكان	يمكن	يكون	

APPENDIX B

Example of watan-2004 corpus document

صنعاء من جمال نعمان:هددت منظمة اليونسكو السلطات اليمنية بإسقاط مدينة زبيد من قائمة التراث الإنساني خلال 3 أشهر ما لم تتخذ إجراءات سريعة وعاجلة لإنقاذ ما يمكن من معالم وآثار المدينة المسجلة ضمن المواقع الأثرية العالمية المهددة بالخطر .وقالت مصادر مطلعة إن وزارة الثقافة والسياحة والهيئة العامة للحفاظ على المدن التاريخية التي تلقت تحذيرا شديد اللهجة لاحظت جدية غير مسبوقة من قبل مركز التراث العالمي التابع لمنظمة اليونسكو هذه المرة حيث أكدت بأن المؤتمر العالمي للآثار الذي سيعقد في إحدى بلدان القارة الأفريقية بعد نحو 3 أشهر قد يتخذ قرارا بإبعاد مدينة زبيد نهائيا من قائمة التراث العالمي. وباشرت وزارة الثقافة والسياحة بالتعاون مع السلطات المحلية والجهات ذات العلاقة التنسيق لاتخاذ إجراءات عملية للحفاظ على معالم زبيد والتي شملت إقامة ندوة حول المدينة التاريخية كما بدأت في تنفيذ مشروع المجاري وكذا بدء الترميم لبعض المعالم الأثرية الآيلة للسقوط. وكان مختصون في مركز التدريب العالمي قد أوصوا في دراساتهم لواقع مدينة زبيد بضرورة أن تقوم الحكومة اليمنية باعتماد نحو نصف مليار ريال كخطوة أولى في أعمال الترميم والصيانة للمنازل والمواقع الأكثر تضررا. من جهتهم يتحدث سكان مدينة زبيد عن هذه الخطوات الحفاظية ورغم أنها بدأت متأخرة بعض الشيء إلا أنها بطيئة وربما لن يسعفها الوقت لإدراك الكثير من المباني والمنازل والتي باتت آيلة للسقوط في أي لحظة. وكانت توصيات الندوة الوطنية لإنقاذ مدينة زبيد المقامة بجامعة الحديدة للفترة 1316 ديسمبر أكدت على ضرورة الحفاظ على تراث وثقافة هذه المدينة التاريخية وتشكل لجنة للمحافظة على مدينة زبيد وإنقاذها من التدهور والاندثار. ودعا المشاركون في الندوة إلى سرعة اتخاذ الإجراءات والتدابير لإخراج المدينة من دائرة الخطر ووقف كل الأعمال والممارسات التي تشوه الطابع المعماري لها وتهدد موروثها الحضاري والتاريخي. وأكد المشاركون على أهمية دور وسائل الإعلام في توعية المواطنين للحفاظ على تراثهم التاريخي والحضاري وموروثهم الثقافي في كل أنحاء الوطن وعلى وجه الخصوص مدينة زبيد التاريخية كونها الآن تصنف في قائمة التراث العالمي المعرض للخطر.

APPENDIX C

Example of identify Arabic BP using Uni-gram, Bi-gram and Dice coefficient similarity

No.	word	Uni-gram	similarity	Bi-gram	similarity
	أشقياء	ا-ش-ق-ي-ا-ء		اش-شق-قي-يا-اء	
1	ابتداء	ا-ب-ت-د-ا-ء	0.5	اب-بت-تد-دا-اء	0.2
2	إمتطاء	ا-م-ت-ط-ا-ء	0.5	ام-مت-تط-طا-اء	0.2
3	اختشاء	ا-خ-ت-ش-ا-ء	0.66	اخ-خت-تش-شا-اء	0.2
4	اختباء	ا-خ-ت-ب-ا-ء	0.5	اخ-خت-تب-با-اء	0.2
5	أنبياء	ا-ن-ب-ي-ا-ء	0.66	ان-نب-بي-يا-اء	0.4
6	أغبياء	ا-غ-ب-ي-ا-ء	0.66	اغ-غب-بي-يا-اء	0.4
7	أغنياء	ا-غ-ن-ي-ا-ء	0.66	اغ-غن-ني-يا-اء	0.4
8	أقرباء	ا-ق-ر-ب-ا-ء	0.66	اق-قر-رب-با-اء	0.2
9	أصدقاء	ا-ص-د-ق-ا-ء	0.66	اص-صد-دق-قا-اء	0.2

APPENDIX D

Code we added to *Lucene* search code to apply the methodology

```
String word="";
String new_word;
String new_word1;
String new_word2;
String new_word3;
char a='ا';
char h='ه';
char w='و';
char y='ي';
String name=line;
String[]splited = name.split(" ");
for(int i=0;i<splited.length;i++){
if(splited[i].length()==6)
{
if((splited[i].charAt(0)=='ت')&&(splited[i].charAt(2)=='ا')&&(splited[i].charAt(4)=='ي')){
if((splited[i].charAt(5)!='ا')||(splited[i].charAt(5)!='ك')||(splited[i].charAt(5)!='ت')){
System.out.println("kkk");
new_word=splited[i].charAt(0)+""+splited[i].charAt(1)+""+splited[i].charAt(3)+""+splited
[i].charAt(4)+""+splited[i].charAt(5);
new_word1=splited[i].charAt(0)+""+splited[i].charAt(1)+""+splited[i].charAt(3)+""+splite
d[i].charAt(5)+""+h;
new_word2=splited[i].charAt(0)+""+splited[i].charAt(1)+""+splited[i].charAt(3)+""+a+""+
splited[i].charAt(5);
new_word3=splited[i].charAt(0)+""+splited[i].charAt(1)+""+splited[i].charAt(3)+""+y+""+
splited[i].charAt(5)+""+h;
word=name+" "+new_word+" "+new_word1+" "+new_word2+" "+new_word3;
}}
/**************************************************************************/
if((splited[i].charAt(1)==splited[i].charAt(3))&&(splited[i].charAt(2)=='ا')&&(splited[i].ch
arAt(4)=='ي')){
if((splited[i].charAt(5)!='ك')||(splited[i].charAt(5)!='م')){
new_word=splited[i].charAt(0)+""+splited[i].charAt(1)+""+splited[i].charAt(2)+""+splited
[i].charAt(5);
new_word1=splited[i].charAt(0)+""+y+""+splited[i].charAt(3)+""+a+splited[i].charAt(5);
new_word2=splited[i].charAt(0)+""+splited[i].charAt(1)+""+splited[i].charAt(2)+""+splite
d[i].charAt(5)+""+h;
word=name+" "+new_word+" "+new_word1+" "+new_word2;
}}}}
```

References

1. Schütze, C.D.M.P.R.H., *An Introduction to Information Retrieval.* Draft of April 1, 2009.
2. Boualem, M. and R. Abbes, *Information Retrieval in Arabic Language*, 2008.
3. Goweder, e.a., *Identifying Broken Plural in Unvowelised Arabic Text.* 2004(a).
4. Goweder, A., et al. *Identifying Broken Plurals in Unvowelised Arabic Tex*. in *EMNLP*. 2004.
5. *http://ifnlp.org/ir*.
6. Ali, M.M., *Mixed-Language Arabic- English Information Retrieval.* 2013.
7. Larkey, L.S., L. Ballesteros, and M.E. Connell, *Light stemming for Arabic information retrieval*, in *Arabic computational morphology*2007, Springer. p. 221-243.
8. Nie, J.-Y., *Cross-Language Information Retrieval.* 2010.
9. Goweder, A.M., I.A. Almerhag, and A.A. Ennakoa, *Arabic Broken Plural Recognition Using a Machine Translation Technique*, 2008, Citeseer.
10. Abduelbaset M. Goweder*, i.a.a., and anes a. ennakoa***, *Arabic Broken Plural Recognition using aMachine Translation Technique.*
11. ABDUELBASET, I., *ARABIC BROKEN PLURAL RECOGNITION USING A MACHINE TRANSLATION TECHNIQUE.* 2222.
12. Goweder, A., Poesio, M., and De Roeck, A.,, *Broken Plural Detection for Arabic Information Retrieval'.* 2004(b).
13. شحاده, خ.م.ع.: 'جموع التكسير في صحيح البخاري', 2008 /5/13 ، الجامعة الهاشمية
14. صوافطه, س.م.ح.: 'صيغ منتهى الجموع في لسان العرب (دراسة صرفية دلالية)', 2010/10/6م ، جامعة النجاح الوطنية
15. M. Abbas, K.S., D. Berkani, *Evaluation of Topic Identification Methods on Arabic Corpora.* JOURNAL OF DIGITAL INFORMATION MANAGEMENT, 2011.
16. Goweder, A., M. Poesio, and A. De Roeck. *Broken plural detection for arabic information retrieval*. in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 2004. ACM.
17. Mahmoud, R., S. Majed, and Z. Khaldoun, *Improving Arabic information retrieval system using N-gram method.* WSEAS Transactions on Computers, 2009. **10**(2011): p. 125-133.
18. Xu, J., A. Fraser, and R. Weischedel. *Empirical studies in strategies for Arabic retrieval*. in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 2002. ACM.