بسم الله الرحمن الرحيم

Sudan University of Science and Technology
Collage of Graduate Studies
Collage of Computer Science and Information Technology

Design of Arabic Dialects Information Retrieval Model for Solving Regional Variation Problem

تصميم نموذج لاسترجاع معلومات اللهجات العربية لحل مشكلة التباين الاقليمي

A Thesis Submitted in Partial Fulfillment of the requirements of M.Sc. in computer science

Prepare by:
Rayan Omer Mohamed Ahmed

Supervised by:
Dr. Albaraa Abuobeida Mohamed Ali

November, 2015

بسم الله الرحمن الرحيم

Sudan University of Science & Technology

College of Graduate Studies

جامعة السودان للعلوم والتكنولوجيا

كلية الدراسات العليا

SUST

كلية الدراسات العليا

Approval Page

Name of Candidate: Rayan Omer Mohamed Ahmed

Thesis title: Design Arabic Dilects Information Retrieval Model for Solving Regional Variation Problem

تصميم نموذج استرجاع معلومات اللهجات العربية لحل مشكلة التباين الاقليمي

Approved by:

1. External Examiner

Name: Dr. Adil Ali

Signature: .... Date: 7.12.2015

2. Internal Examiner

Name: Dr. Talaat MohyEldin Wahby

Signature: .... Date: 7-12-2015

3. Supervisor

Name: Dr. Albaraa Abuobeida

Signature: .... Date: 7-12-2015

Sudan University of Science and Technology
College of Graduate Studies

Declaration

I, the signing here-under, declare that I'm the sole author of the (M.Sc.) thesis entitled

Design of Arabic Dialects Information Retrieval Model for Solving Regional Variation Problem

which is an original intellectual work. Willingly, I assign the copy-right of this work to the College of Graduate Studies (CGS), Sudan University of Science & Technology (SUST). Accordingly, SUST has all the rights to publish this work for scientific purposes.

Candidate's name: Rayan Omer Mohamed Ahmed

Candidate's signature: .... Date: 15-2-2016

إقرار

أنا الموقع أدناه أقر بأننى المؤلف الوحيد لرسالة الماجستير المعنونة

تصميم نموذج لاسترجاع معلومات اللهجات العربية لحل مشكلة التباين الاقليمي

وهى منتج فكري أصيل . وباختياري أعطى حقوق طبع ونشر هذا العمل لكلية الدراسات العليا - جامعه السودان للعلوم والتكنولوجيا، عليه يحق للجامعه نشر هذا العمل للأغراض العلمية .

اسم الدارس : ريان عمر محمد احمد

توقيع الدارس : .... التاريخ : ١٥-٢-٢٠١٦م

DEDICATION:

This thesis is dedicated to my mother and my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake and the largest task can be accomplished if it is done one step at a time. It is also dedicated to my brothers and sisters. I am grateful too for the support and advice from my friends especially Ebtihal Mustafa and Rawan Kider. I need to thank the Godfather of this research, Dr. Mohamed Mustafa Ali.

ACKNOWLEDGEMENT:

First and foremost, I would like to thank Allah, then I should extend my deep and sincere gratitude to all of whom directed me, and taught me, and took my hand in order to accomplish this research and particularly Dr. Albaraa Abuobieda has been the ideal thesis supervisor. I would also like to thank Dr. Mohamed Mustafa Ali, whose steadfast support of this research was greatly needed and deeply appreciated.

ABSTRACT:

Information retrieval (IR) is defined as an activity of satisfying the user's information needs from a collection of unstructured data (text, image, and video). One of disadvantage of most IR systems is that the search is based on query terms that entered by users. Then, when Arab user write the query using the term in his dialect or in Modern Stander Arabic (MSA) form, the documents were retrieved contained this query's term only. This problem appears clearly in scientific Arabic's documents, for illustration, the documents that show the compiler concept; it can be found written by the one of the following Arabic words: "الجامع" ,"المفسر" or "المترجم". Thus, our research is focused on the Arabic language, as it is one of the widely spread languages with different dialects.

We propose a pre-retrieval (offline) method to build a statistical based dictionary to expand the query which is based on a statistical methods (co-occurrence technique and Latent Semantic Analysis (LSA) model) which can be defined as a flexible approach because it is based on mathematical foundations to improve the effectiveness of the search result by retrieving the most relevant documents regardless of their dialect was used to formulate the queries.

We designed and evaluated our method and the baseline methods from a small corpus collected manually using Google search engine. The evaluation was done using the average recall (Avg-R), average precision (Avg-P) and average F-measure (Avg-F).

The result of our experiments indicated that the proposed method is a proven to be efficient for improving retrieval via expands the query by regional variation's synonyms, with accuracy 83% in form of Avg-F. Also, statistically our model is significant when it is compared to traditional IR systems by acquired 5.43594E-16 in the t-test.

المستخلص:

استرجاع المعلومات هو عبارة عن عملية ارضاء المستخدمين بتوفير حاجتهم المعلوماتية من مجموعة من البيانات الغير مهيكله (صوت، صورة، فيديو، نص).من التحديات التي تواجه عملية استرجاع المعلومات انه يتم استرجاع الوثائق بتطبيق التطابق الفعلي بين الاستفسار والوثيقة، فقد يقوم المستخدم العربي بالتعبير عن حاجته البحثية بكتابة الاستعلام بلهجته او باللغة العربية الفصحى فيتم استرجاع الوثائق التي تحتوي على الكلمات المكونة للاستعلام التي تمت كتابتها بواسطة المستخدم فقط مما يؤدي الى ضياع الوثائق التي توفر للمستخدم ما يرغب من معلومات بسبباحتوائها على مصطلحات مرادفه لكلمات الاستعلام. هذه المشكلة تظهر بشكل واضح في النصوص العلمية،على سبيل المثال: الوثيقة التي تتناول مفهوم المفسر(In English, Compiler) قد تكتب ايضاً باستخدام مصطلح الجامع او المترجم. في هذا البحث سيتم التعامل مع اللغة العربية لاحتوائها على اختلاف واسع في اللهجات.

تم اقتراح طريقه حل تتم قبل الاسترجاع (خلفيه) تعتمد على طرق احصائية (تقنية الورود ومنهجية التكشيف الدلالي الكامن) التي تعتبر طرق مرنه لاعتمادها على اساس رياضي، وذلك لبناء قاموس يحتوي على المرادفات الخاصة باي كلمة لتوسيع الاستعلام ومن ثم تحسين نتيجة البحث باسترجاع الوثائق الملائمة مع اختلاف لهجة الاستعلام مع لهجة الوثيقة.

تم تصميم وتقييم طريقه الحل المقترحة و طرق الاسترجاع الاخرى باستخدام عدد بسيط من الوثائق التي تم جمعها يدويا باستخدام محرك البحث قوقل. التقييم تم باستخدام متوسط الاستدعاء ومتوسط الدقة و متوسط (F-measure).

النتائج اوضحت ان الحل المقترح فعال جدا في تحسين نتيجة الاسترجاع بتوسيع الاستعلام بالمرادفات الاقليمية المختلفة بدقة 83% باستخدام متوسط F-measure، ايضاً احصائياً طريقتنا لها دلاله مقارنة مع نظام استرجاع المعلومات التقليدي وذلك بالحصول على 5.43594E-16 باختبار الطالب.

Table of Contents

LIST OF TABLES:

LIST OF FIGURES:

LIST OF APPENDIX:

CHAPTER ONE

1. INTRODUCTION:

1.1 Introduction:

In the past, the process of retrieving the required information from a collection of a certain topic was a simple process because of the few amount of information, but with the increasing amount of data such as text, audio, video, and other documents on the internet the process of finding the specified information has become a very difficult process using traditional methods which can be made by the linear search for each document(Sanderson, Croft, 2012).

In 1950 the first Information Retrieval (IR) system was introduced by Calvin Mooer's to solve the issue of searching in huge amount of data (Sanderson, Croft, 2012). Later on, the IR improved as a result of the expansion of the computer systems. With the development of the IR systems, they can process queries and documents in an efficient and effective way (González et al, 2008).

IR is an abbreviation for Information Retrieval, a system that processes unstructured data such as documents, videos and images which consider as the main point of difference from Database " structured data " to reach the point that satisfies the user's need from within large collections (Manning et.al, 2008). In this research we refer to retrieve the relevant text documents only in response to user's information need.

In IR system, users write their needs in the form of a query and authors write their knowledge in the form of a document. To build an IR system which is considered as the main component of search engines must gather a collection of a document to construct which is known as a corpus by using one of gathering methods (manually, crawler, etc.). After that, The IR system applies a set of operations known as preprocessing operations on the documents such as tokenizing documents to words based on white space to extract the terms that are used to build the index which allows us to find the documents that contain a query terms. The same preprocessing operation applied to documents must be applying on queries to make the representation of documents and queries typical. Afterwards, one of IR model is used to retrieve the relevant documents using the index. It then ranks the results using the ranking module. These IR tasks are language independent(Manning et.al., 2008)(Inkpen, 2006).

Over the last year Arabic IR becomes one of the most interesting areas of research due to fastest growth of the Arabic language for the Web. Arabic language is one of the most widely spoken languages in the world. It is a member of Semitic languages. The Arabic Language differs from Indo-European languages in two aspects: morphologically and syntactically (Ali, 2013). The Arabic language is very complex morphological when compared to Indo-European languages because Arabic is root based and very tolerant syntactically, for instance,"اخذت البنت القلم"and "البنت اخذت القلم"(In English, The girl took the pen)has the same meaning despite the order of the words been changed.

The Arabic IR system faces significant challenges to retrieving the Arabic relevant documents due to the ambiguity that is found in it which is caused by the morphology and orthography of the Arabic language which affects the precision of the retrieval system. Regional variation disambiguation is one of the problems facing Arabic information retrieval resulted from the different Arab regions and dialects used in the Arab World (H. AbdAlla,2008). It also plays an important role in the information retrieval because of the increasing amount of Arabic text on the web which can cause a set of documents represented by different words based on a region of authors to carry the same concepts. For instance, The Ministry of Education can be "وزارة التربية والتعليم"and "وزارة المعارف" also mobile phone companies can be "شركات الموبايل" and "شركات الهاتف السيار". Also King can be "الملك"and "الرئيس". The Regional variation problem appears clearly in scientific documents, for example, the documents that show the 'code' concept it can be found written by the one of the following Arabic words:"الشفرة" or "الكود".

The Arab world is divided into six regions based on dialects: Gulf, Morocco, Levantine, Egyptian, Yemen and Iraq. Gulf region includes Saudi Arabia, UAE, Kuwait, Qatar, Bahrain and Oman. Morocco includes Morocco, Algeria, Tunisia and Libya. Levantine cover Lebanon, Jordan, Syria and Palestine. Yemen is in the State of Yemen and Iraq is in the State of Iraq. Within the region can also note the difference.

Two ways to solve the regional variation (Dialect) in the Arabic information retrieval system are: using auxiliary structures like dictionaries or thesauruses. Using this on the web search restricts the synonyms of the word that is found in dictionaries and keeps the search intent is difficult because the words have two sides of meanings: General means in the language and Specific meaning in the context. The other solution is statistical which can be defined as a flexible approach because it is based on mathematical foundations.

This research aims to develop a statistical method that finding the relevant documents to a user's query regardless of the author's dialect and regional variation was used to write the document's contents.

1.2 Problem Statement:

The Arabic language is the most widely spoken languages of the Semitic family and broadly spread because it is the religious language of all Muslims, the language of science in the middle age and part of the curriculum in most of non-Arabic countries such as Iran and Pakistan(Darwish .K, W. Magdy,2014).

The Arabic language is an aggregate of multiple varieties including: Classical Arabic (CA), Modern Standard Arabic (MSA) and Regional or Dialectal Arabic (DA) which are called: Qur'an Arabic, \fuSHa "العربية الفصحى" and\lahja "لَـهْجة"or \ammiyya"عامية", respectively (Darwish .K, W. Magdy,2014). Classical Arabic is the language of the Quran and classical literature. MSA is the universal language of the Arab world which is understood by all Arabic speakers and used in education and official settings.MSA was resulted from adding modern terms to classical Arabic (Quran Arabic). DA is a commonly used, region specific and informal variety, which vary from MSA in many aspects such as vocabulary, morphology and spelling.

The Arab society has a phenomenon known as Diglossia. The term diglossia was introduced from French 'diglossie' by Ferguson (1959). Each Arabic-speaking country has two variations in languages: one of them is used in official communications and what is known as Modern Standard Arabic (MSA). Another variant is non-official language and is used in the everyday between members of the region. It is called local dialects and it differs in between Arabic countries, moreover, different dialects can be found in the same country e.g. The Saudi dialect includes Najdi (Central) dialect, Hejazi (Western) dialect, Southern dialect, etc (Khalid Almeman, Mark Lee, 2013).

Dialects or colloquial can be considered as a new form of synonyms which mean different word to express the same meaning, like the words "موبايل","جوال" ,"سيار" and "محمول" which mean cell phone/portable-phone (Ali, 2013).

On the web, authors write documents to transfer the knowledge that exists on the mind uses his own words. These used words are influenced by the region where authors live which appears in the words that are used by different people from different regions to explain the same concept.

With the huge amount of Arabic data published daily over the Internet, it becomes necessary to develop a method that would help avoid the ambiguity that exists due to the regional semantics overlapping in Arabic words (See Table 1.1). This ambiguity form a great challenge to the Arabic Information Retrieval System because if you don't detect the regional synonyms correctly and accurately it may lead to losing some relevant documents and may cause intent drifting which reduces the precision of Arabic Information retrieval systems (see Figure 1.1, 1.2, 1.3and 1.4) which shows the difference when using two similar words with different result.

Table 1.1: Example of Regional Variations in Arabic Dialect

English	Table	Cat	I_want	Shoes	Baby
MSA	طاولة	قطة	اريد	حذاء	طفل
Moroccan	ميدة	قطة	بغيت	سباط	ذراري
Sudan	طربيزة	كديسه	عاوز	جزمه	شافع
Syrian	طاولة	بسة	بدي	كندره	فصعون
Iraqi	ميز	بزونة	اريد	قندره	زعطوط

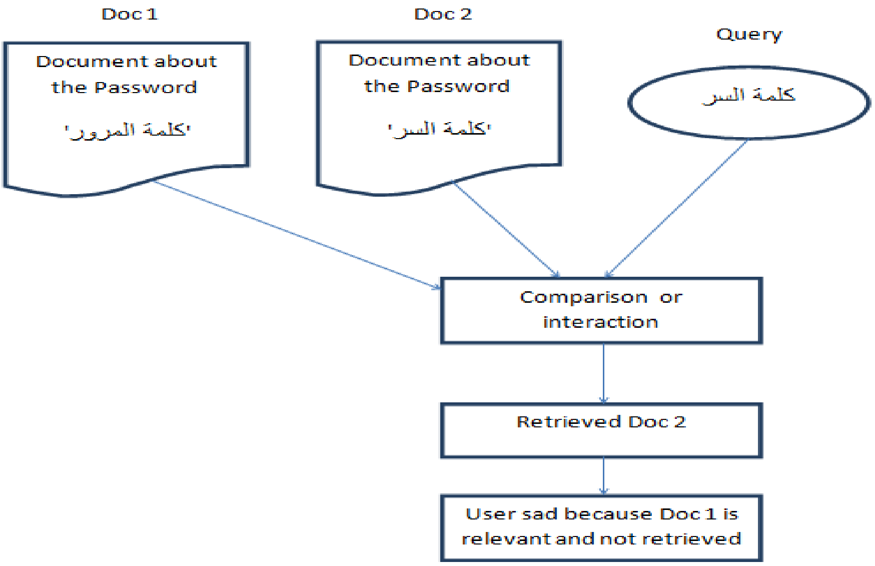

Doc 1
Document about the Password
'كلمة المرور'
Doc 2
Document about the Password
'كلمة السر'
Query
كلمة السر
Comparison or interaction
Retrieved Doc 2
User sad because Doc 1 is relevant and not retrieved

Figure 1.1: Explain when the all Relevant Documents notRetrieved

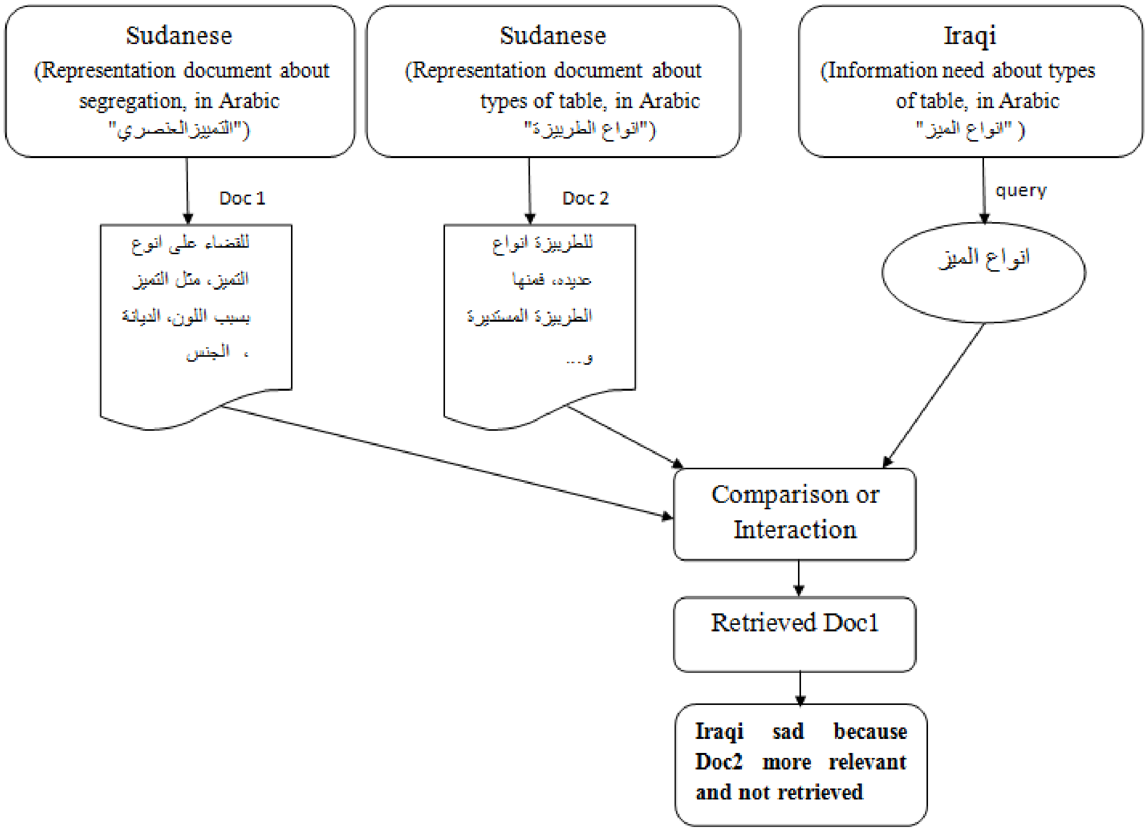
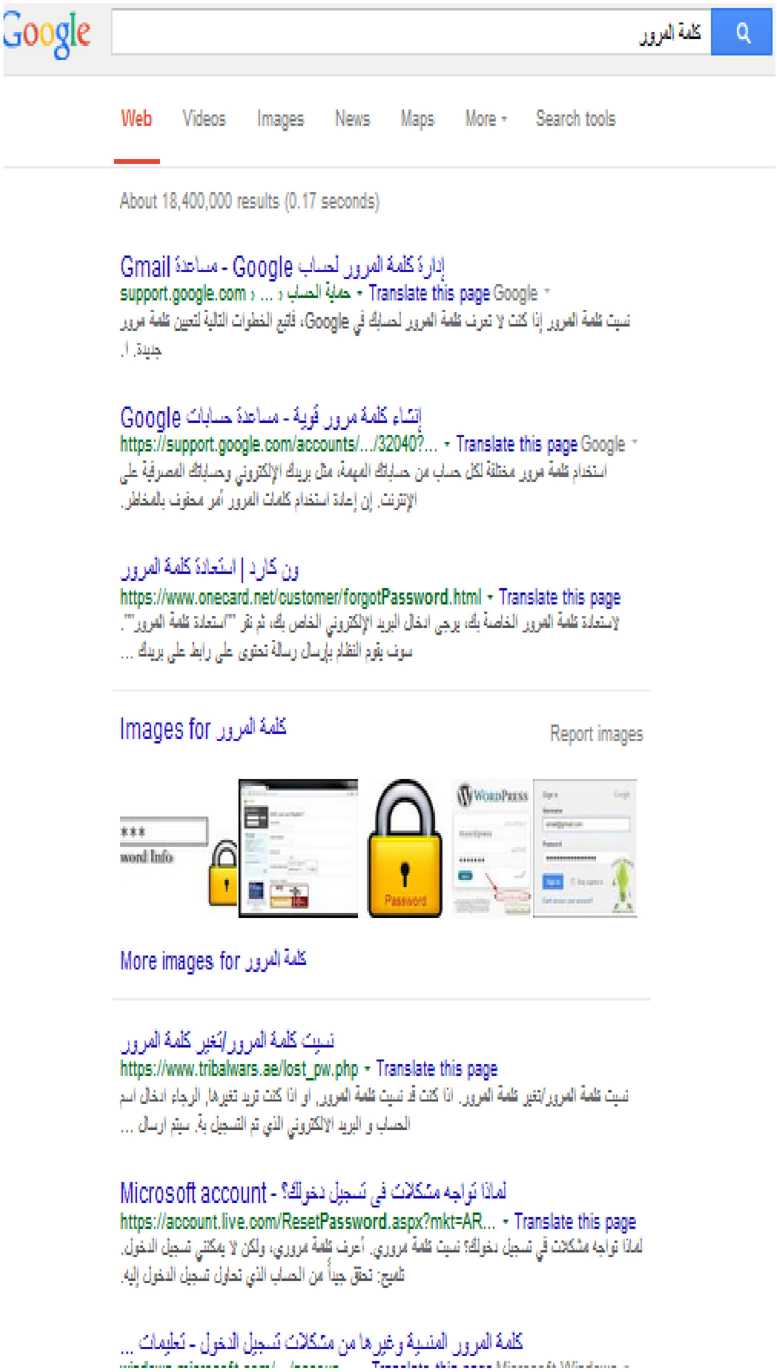

Sudanese
(Representation document about segregation, in Arabic "التميزالعنصري")
Sudanese
(Representation document about types of table, in Arabic "انواع الطريزة")
Iraqi
(Information need about types of table, in Arabic "انواع الميز")
Doc 1
Doc 2
query
للقضاء على انوع
التميز، مثل التميز
بسبب اللون، الديانة
، الجنس
للطريزة انواع
عديده، فمنها
الطريزة المستديرة
و...
انواع الميز
Comparison or Interaction
Retrieved Doc1
Iraqi sad because Doc2 more relevant and not retrieved

Figure 1.2: Explain the Retrieving of Irrelevant Documents



About 18,400,000 results (0.17 seconds)

إدارة كلمة المرور لحساب Google - مساعدة Gmail
support.google.com › ... › حماية الحساب ▾ Translate this page Google ▾
نسيت كلمة المرور إذا كنت لا تعرف كلمة المرور لحسابك في Google، فاتبع الخطوات التالية لتعيين كلمة مرور جديدة !.

إنشاء كلمة مرور قوية - مساعدة حسابات Google
https://support.google.com/accounts/.../32040?... ▾ Translate this page Google ▾
استخدام كلمة مرور مختلفة لكل حساب من حساباتك المهمة، مثل بريدك الإلكتروني وحساباتك المصرفية على الإنترنت. إن إعادة استخدام كلمات المرور أمر محفوف بالمخاطر.

ون كارد | استعادة كلمة المرور
https://www.onecard.net/customer/forgotPassword.html ▾ Translate this page
لاستعادة كلمة المرور الخاصة بك، يرجى ادخال البريد الإلكتروني الخاص بك، ثم نقر "استعادة كلمة المرور". سوف يقوم النظام بإرسال رسالة تحتوي على رابط على بريدك ...

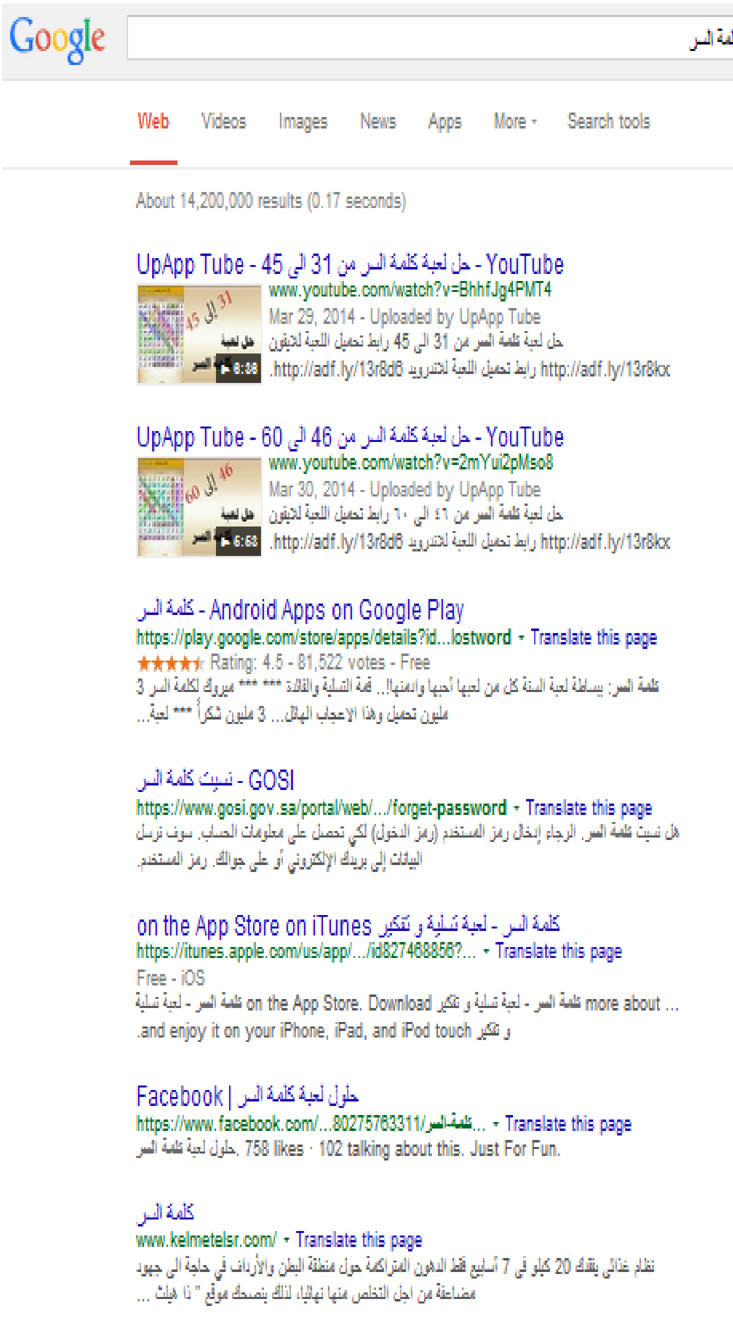
Images for كلمة المرور   Report images

More images for كلمة المرور

نسيت كلمة المرور/تغير كلمة المرور
https://www.tribalwars.ae/lost_pw.php ▾ Translate this page
نسيت كلمة المرور/تغير كلمة المرور. اذا كنت قد نسيت كلمة المرور, او اذا كنت تريد تغيرها, الرجاء ادخال اسم الحساب و البريد الالكتروني التي تم التسجيل بة. سيتم ارسال ...

لماذا تواجه مشكلات في تسجيل دخولك؟ - Microsoft account
https://account.live.com/ResetPassword.aspx?mkt=AR... ▾ Translate this page
لماذا تواجه مشكلات في تسجيل دخولك؟ نسيت كلمة مروري. أعرف كلمة مروري، ولكن لا يمكنني تسجيل الدخول. تلميح: تحقق جيداً من الحساب الذي تحاول تسجيل الدخول إليه.

كلمة المرور المنسية وغيرها من مشكلات تسجيل الدخول - تعليمات ...



About 14,200,000 results (0.17 seconds)

حل لعبة كلمة السر من 31 الى 45 - UpApp Tube - YouTube
www.youtube.com/watch?v=BhhfJg4PMT4
Mar 29, 2014 - Uploaded by UpApp Tube
حل لعبة كلمة السر من 31 الى 45 رابط تحميل اللعبة لاتيفون http://adf.ly/13r8kx رابط تحميل اللعبة لاندرويد http://adf.ly/13r8d6.

حل لعبة كلمة السر من 46 الى 60 - UpApp Tube - YouTube
www.youtube.com/watch?v=2mYui2pMso8
Mar 30, 2014 - Uploaded by UpApp Tube
حل لعبة كلمة السر من ٤٦ الى ٦٠ رابط تحميل اللعبة لاتيفون http://adf.ly/13r8kx رابط تحميل اللعبة لاندرويد http://adf.ly/13r8d6.

كلمة السر - Android Apps on Google Play
https://play.google.com/store/apps/details?id...lostword ▾ Translate this page
★★★★★ Rating: 4.5 - 81,522 votes - Free
كلمة السر: ببساطة لعبة السنة كل من لعبها أحبها وادمنها!.. قمة التسلية والفائدة *** مبروك لكلمة السر 3 مليون تحميل وهذا الاعجاب الهائل... 3 مليون شكراً *** لعبة...

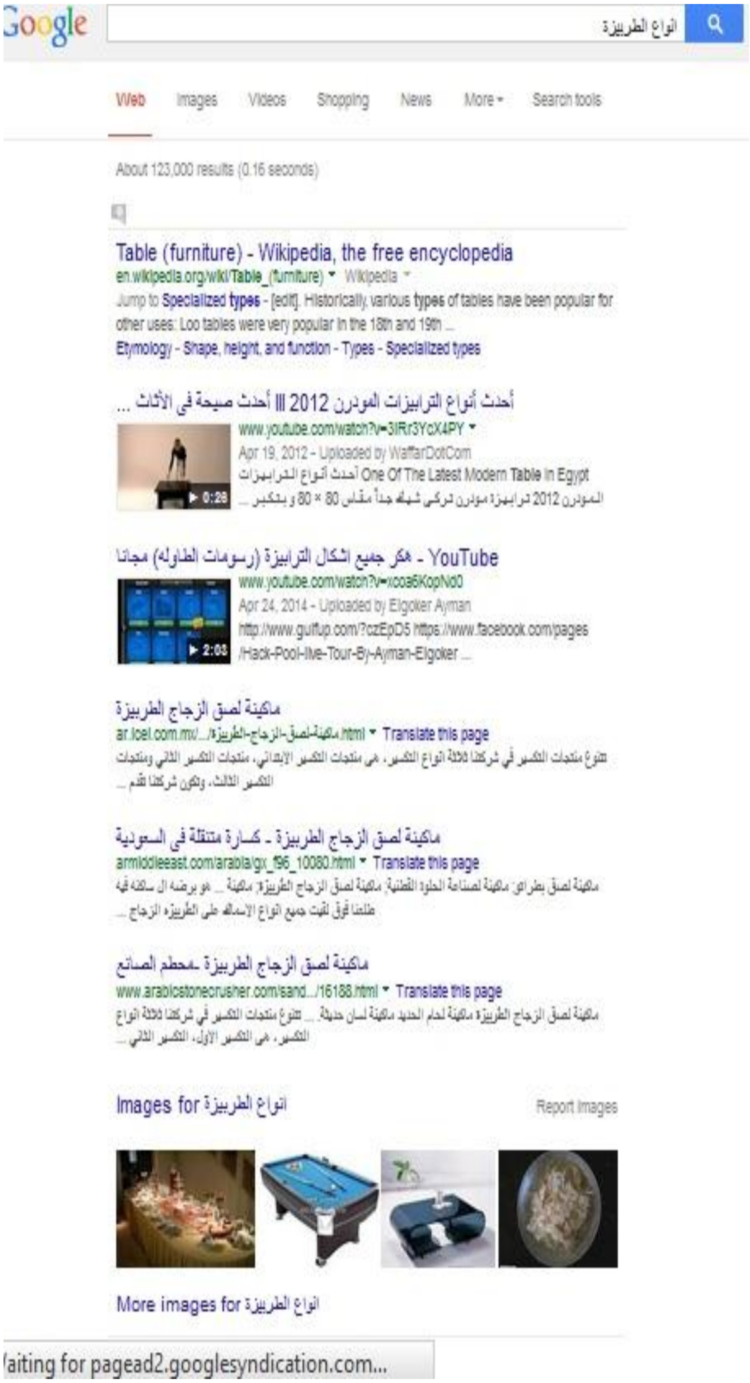
نسيت كلمة السر - GOSI
https://www.gosi.gov.sa/portal/web/.../forget-password ▾ Translate this page
هل نسيت كلمة السر. الرجاء إدخال رمز المستخدم (رمز الدخول) لكي تحصل على معلومات الحساب. سوف ترسل البيانات إلى بريدك الإلكتروني أو على جوالك. رمز المستخدم.

كلمة السر - لعبة تسلية و تفكير on the App Store on iTunes
https://itunes.apple.com/us/app/.../id827468856?... ▾ Translate this page
Free - iOS
... more about كلمة السر - لعبة تسلية و تفكير on the App Store. Download كلمة السر - لعبة تسلية و تفكير and enjoy it on your iPhone, iPad, and iPod touch.

حلول لعبة كلمة السر | Facebook
https://www.facebook.com/...80275783311/كلمة-السر... ▾ Translate this page
حلول لعبة كلمة السر. 758 likes · 102 talking about this. Just For Fun.

كلمة السر
www.kelmetelsr.com/ ▾ Translate this page
نظام غذائي يفقدك 20 كيلو فى 7 أسابيع فقط الدهون المتراكمة حول منطقة البطن والأرداف في حاجة الى جهود مضاعفة من اجل التخلص منها نهائيا، لذلك ينصحك موقع " نا هيلث ...

Figure 1.3: Example of Retrieving documents when write query "كلمة المرور" and " كلمة السر"using Google search engine



About 123,000 results (0.16 seconds)

Table (furniture) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Table_(furniture) ▾ Wikipedia ▾
Jump to Specialized types - [edit]. Historically, various types of tables have been popular for other uses: Loo tables were very popular in the 18th and 19th ...
Etymology - Shape, height, and function - Types - Specialized types

أحدث أنواع الترابيزات المودرن 2012 ||| أحدث صيحة في الأثاث ...
www.youtube.com/watch?v=3IRr3YcX4PY ▾
Apr 19, 2012 - Uploaded by WaffarDotCom
أحدث أنواع الترابيزات One Of The Latest Modern Table In Egypt
المودرن 2012 ترابيزة مودرن تركى شيك جداً مقاس 80 × 80 و بتكبير ...
0:28

هكر جميع اشكال الترابيزة (رسومات الطاوله) مجانا - YouTube
www.youtube.com/watch?v=xcoa6KopNd0
Apr 24, 2014 - Uploaded by Elgoker Ayman
http://www.gulfup.com/?czEpD5 https://www.facebook.com/pages
/Hack-Pool-live-Tour-By-Ayman-Elgoker ...
2:03

ماكينة لصق الزجاج الطربيزة
ar.icel.com.mx/.../ماكينة-لصق-الزجاج-الطربيزة.html ▾ Translate this page
تتنوع منتجات التكسير في شركتنا ثلاثة انواع التكسير، هي منتجات التكسير الإبتدائي، منتجات التكسير الثاني ومنتجات التكسير الثالث، وتكون شركتنا تقدم ...

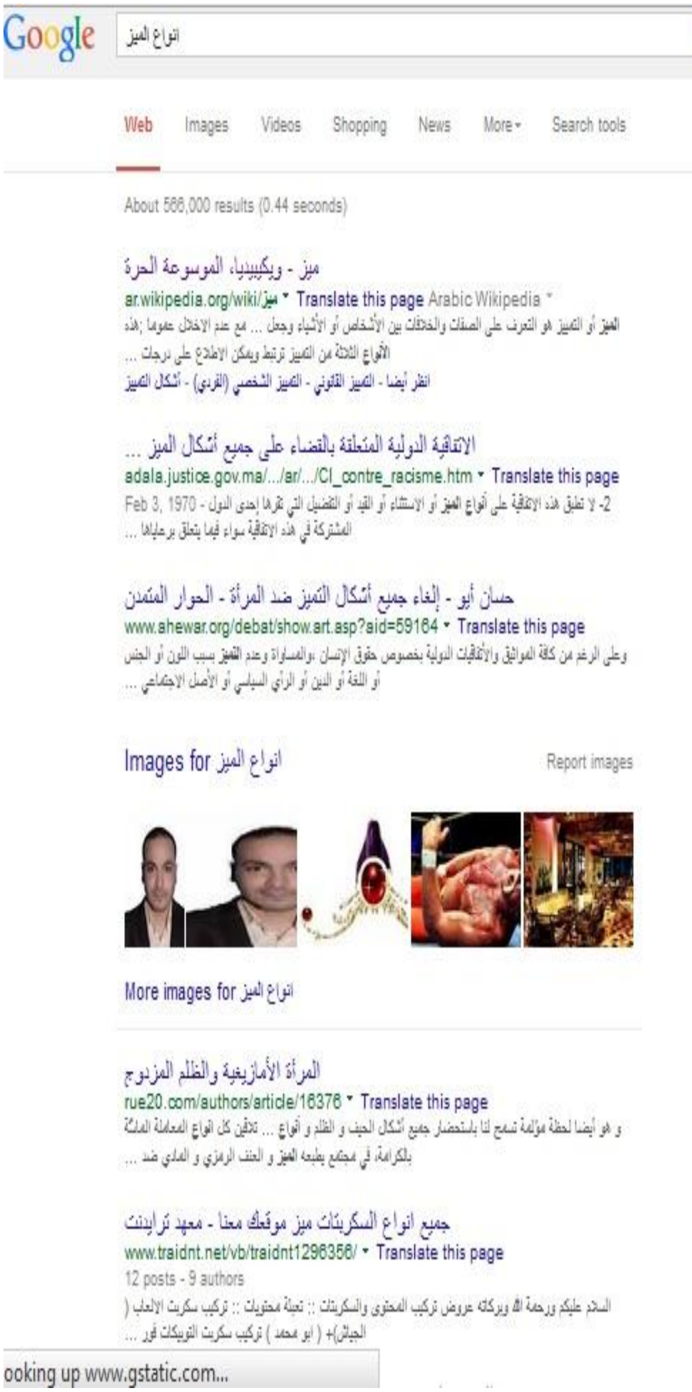
ماكينة لصق الزجاج الطربيزة - كسارة متنقلة في السعودية
armiddleeast.com/arabia/gx_f96_10080.html ▾ Translate this page
ماكينة لصق بطراق: ماكينة لصناعة الحلوى القطنية; ماكينة لصق الزجاج الطربيزة; ماكينة ... هو برضه ال ساكنه فيه طلعنا فوق لقيت جميع انواع الاسماك على الطربيزه الزجاج ...

ماكينة لصق الزجاج الطربيزة -محطم الصانع
www.arabicstonecrusher.com/sand.../16188.html ▾ Translate this page
ماكينة لصق الزجاج الطربيزة ماكينة لحام الحديد ماكينة لسان حديثة ... تتنوع منتجات التكسير في شركتنا ثلاثة انواع التكسير، هي التكسير الأول، التكسير الثاني ...

Images for انواع الطربيزة
Report images

More images for انواع الطربيزة

/aiting for pagead2.googlesyndication.com...



About 566,000 results (0.44 seconds)

ميز - ويكيبيديا، الموسوعة الحرة
ar.wikipedia.org/wiki/ميز ▾ Translate this page Arabic Wikipedia ▾
الميز أو التمييز هو التعرف على الصفات والخلافات بين الأشخاص أو الأشياء وجعل ... مع عدم الاخلال عموما ;هذه الأنواع الثلاثة من التمييز ترتبط ويمكن الاطلاع على درجات ...
انظر أيضا - التمييز القانوني - التمييز الشخصي (الفردي) - أشكال التمييز

الاتفاقية الدولية المتعلقة بالقضاء على جميع أشكال الميز ...
adala.justice.gov.ma/.../ar/.../CI_contre_racisme.htm ▾ Translate this page
Feb 3, 1970 - 2- لا تطبق هذه الاتفاقية على أنواع الميز أو الاستثناء أو القيد أو التفضيل التي تقرها إحدى الدول المشتركة في هذه الاتفاقية سواء فيما يتعلق برعاياها ...

حسان أيو - إلغاء جميع أشكال التميز ضد المرأة - الحوار المتمدن
www.ahewar.org/debat/show.art.asp?aid=59164 ▾ Translate this page
وعلى الرغم من كافة المواثيق والاتفاقيات الدولية بخصوص حقوق الإنسان ،والمساواة وعدم التميز بسبب اللون أو الجنس أو اللغة أو الدين أو الرأي السياسي أو الأصل الاجتماعي ...

Images for انواع الميز
Report images

More images for انواع الميز

المرأة الأمازيغية والظلم المزدوج
rue20.com/authors/article/16376 ▾ Translate this page
و هو أيضا لحظة مؤلمة تسمح لنا باستحضار جميع أشكال الحيف و الظلم و أنواع ... تلاقي كل انواع المعاملة الماسة بالكرامة، في مجتمع يطبعه الميز و العنف الرمزي و المادي ضد ...

جميع انواع السكربتات ميز موقعك معنا - معهد ترايدنت
www.traidnt.net/vb/traidnt1296356/ ▾ Translate this page
12 posts - 9 authors
السلام عليكم ورحمة الله وبركاته عروض تركيب المحتوى والسكربتات :: تعبئة محتويات :: تركيب سكربت الالعاب (الجباش)+ (ابو محمد) تركيب سكربت التويكات فور ...

ooking up www.gstatic.com...

Figure 1.4: Example of Retrieving documents when write query "الطربيزة" and "الميز" using Google search engine

1.3 Research Questions:

The core goal of this research is to develop method to expand queries by Arabic regional variation synonyms to handle missed retrieval for relevant documents using Arabic dialect test dataset. In particular, the research questions are:

- What are the methods that can be used to discover the Regional Variations (Dialects) in the Arabic language?
- How the proposed method can enhance the relevant retrieving?

1.4 Objective of the Research:

The goal of this research is to develop method able to identify the Arabic regional variation synonyms accurately in monolingual corpora to assist users in finding the information they need regardless of any variation (dialect) was used to formulate the query. The study should meet the following objectives:

- To build small Arabic dialect corpus.
- To device statistical method works with Arabic dialect corpus for extraction Arabic regional variation synonyms.
- To improve the performance of Arabic Information retrieval system by using query expansion techniques.

1.5 Research Scope:

The scope of this research is in the Information Retrieval area. Within the field of information retrieval we focus on synonym discovery in Arabic language from our corpus. These synonyms form the regional variations (Arabic dialect) in vocabulary.

1.6 Research Methodology and Tools:

This thesis introduces the Arabic region variation is a problem for Arabic Information retrieval systems.

To solve the problem of this research we will do the following: Collect a set of documents manually using Google search engine to build a small corpus containing different Arabic documents contains regional variations words to form a test data set and also construct the set of queries and binary relevance judgments. After that, we done some of preprocessing operation and filtered the frequent words, and used the co-occurrence technique and Latent Semantic Analysis (LSA) model.

A Co-occurrence technique used to collect the words that co-occur together in the documents. We used the LSA model to analyze the dataset to extract the high similar word in the test dataset. This analyze assumes that terms occur in the similar context are synonym. Because this approach is based on co-occurrence of words, so maybe gathering words occur together permanently as synonyms. To detraction this issue we set a threshold of revision the semantic space extracted using the LSA model. Afterward, merge the result of Co-occurrence and LSA by using the transitive property concept to build statistical dictionary contains each word and the synonyms.

To browse the result set of Arabic Dialect IR system as search engines we will use Lucene packet for indexing and searching and Java server page language (JSP) with Jakarta tomcat as server to design the web page. This web page allows the user to enter the query and then use the dictionary to expand the queries by terms was gathered as synonym dialects and then retrieves the relevant documents to increase a recall and precision of the IR system.

1.7 Research Organization:

The present research is organized into five chapters entitled: introduction; literature review and related work; research methodology; results and discussion; and conclusion.

Chapter One of the research is mainly an introduction to the research which includes a problem statement and the aims of the research, in addition to the scope of the research, the research methodology and questions, and finally an organization of the chapters.

Chapter Two is deal with the background relating to the research. The background gives an overview of information retrieval(IR) and linguistic issues which have an effect on information retrieval. It is then followed by the related works.

Chapter Three is a detailed description of the proposed solution which describe the method architecture.

Chapter Four (results and discussion): covers the system evaluation. An attempt was made to represent the retrieval performance of our method, in addition to offering a discussion of the results of a method.

Chapter Five is the last chapter of the research. It is a summary of the work which has been carried out in the current research. It also shows the main findings of the system evaluation and attempts to answer the research questions. The chapter presents several recommendations. The chapter ends with some suggestions for future work to be done in this area.

CHAPTER TWO

2. LITRIAL REVIEW

2.1 Introduction:

In this chapter, we describe the basic concepts that are require to conduct this research. We first describe the basic concepts about information retrieval in section 2.2 such as preprocessing operation, indexing, retrieval models and retrieval evaluation measures. Second, we describe brief overview about Arabic language and challenges in section 2.3. Final, section 2.4 for related works.

2.2 Information Retrieval:

There is a huge amount of data such as text, audio, video and other documents available on the internet. Users express their information needs using a query containing a set of keywords to access for this data. Users can use two ways to find this information: search engines, for which the information retrieval system (IR) is considered an essential component (see Figure 2.1).Users can also use browse directories organized by categories (such as Yahoo Directories) (H. AbdAlla,2008).

IR is a process of manipulates the collection of data to achieve the objective of IR which retrieves only relevant documents for a user query with a rapid response. Relevance denotes how well a retrieved document or set of documents meets the information need of the user.

The query search is usually based on so-called terms. These terms can be words, phrases, stems, root and N-grams. To extract these terms from the document collection we apply a set of operations called the preprocessing operation. These extracted terms are used to build what is known by index, used for selecting documents that contain a given query terms(Ruge. G, 1997). Afterwards, the searching model retrieves the relevant documents using the index. It then ranks the results by the ranking module (Inkpen, 2006).We will describe these concepts in details in the next subsections.

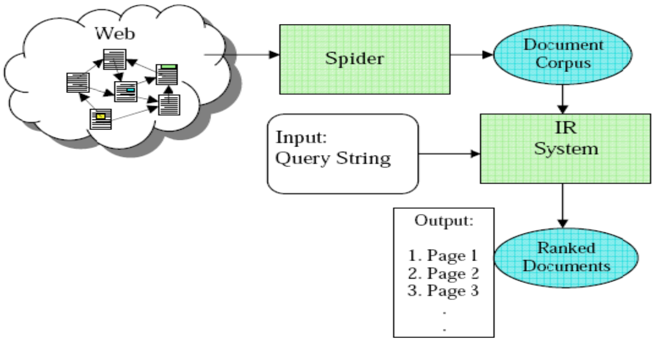


Figure 2.1: Search Engines Architecture

2.2.1 Text Preprocessing in Information Retrieval:

The content of the documents in the IR is used to build the index which helps retrieve the relevant document. But the content of this document it needs to processing to use in IR tasks due to may contain unwanted characters or multiple variation for the same word etc. . Preparing these documents for the IR task goes through several offline preprocessing operations which are language dependent, namely: Tokenization, Stop word removal, Normalization, Lemmatization and Stemming.

2.2.1.1 Tokenization:

In this operation the full text is converted into a list of meaningful pieces called token based on delimiters such as the white space in Arabic and English languages. The task of specifying the delimiter becomes more challenging because it can cause unwanted retrieval results in several cases. One example is when you are dealing with languages (Germany or Korean) that don't have a clear delimiter. Another example is observe if this consequence of words represents one word or more i.e. co-occurrence and in number case (3/20/92, F-12, 123-65-905).(Manning et al, 2008) (Ali, 2013)

2.2.1.2 Stop-Word Removal:

Stop words usually refer to the most common words in a language. In other word, a set of common words which would appear to be of little value in helping select documents matching such as determiners (the, a, an), coordinating conjunctions (for, an, nor, but, or, yet, so) and prepositions (in, under, towards, before).(Manning et al, 2008)

The stop-word removal operation is done by removing these stop words. Stop-words are eliminated from both query and documents

2.2.1.3 Normalization:

Normalization is defined as "a process of canonicalizing tokens so that matches occur despite superficial differences in the character sequences of the tokens (Manning et al, 2008)". It used to handle the redundancy which is caused by morphological variations in the way the text can be represented. This process includes two acts: Case Folding, a process that replaces all letters with lower case letters (Information and inFormAtion convert into information). Another process is eliminating the elements in the document that are not for indexing and unwanted characters (punctuation marks, document tags, diacritics and kasheeda). For example, removing kasheeda, known also as Tatweel, in the word "البيانــــــات" or "البيــــــانات" (in English, data) becomes written "البيانات".

The main advantage of normalizing the words is maximizing matching between a query token and document collection tokens.(Ali, 2013)

2.2.1.4 Lemmatization:

Another process is known as lemmatization, which means use morphological and syntactical rules to obtain dictionary forms of a word, which is known as the lemma, for example, am, are, is and cutting convert to be and cut, respectively.(Manning et al, 2008)

2.2.1.5 Stemming:

Stemming terms is a linguistic process that attempts to determine the base (stem) of each word in a text. in other word, a technique for reducing a word to its root form(Manning et al, 2008). For instance, the English words connected, connection, connections are all reduced to the single stem "connect" and Arabic words like يلعبن ,تلعبون ,يلعبون and يلعبان may all be rendered to لعب (meaning: play), the main advantage of stemming words is reducing the amount of vocabulary and as a consequence the size of index and allowing it to retrieve the same document using various forms of a word. The most popular and fastest English stemmer is Porter's stemmer and Light10 in Arabic (Ali, 2013).

When we build IR System we select the preprocessing operation we want to apply and not require apply all this operation.

The same preprocessing steps that were performed on the documents are also performed on the query to guarantee that a sequence of characters in the text will always match the same sequence typed in a query. The query preprocessing operation is done in the search time.

2.2.2 Indexing:

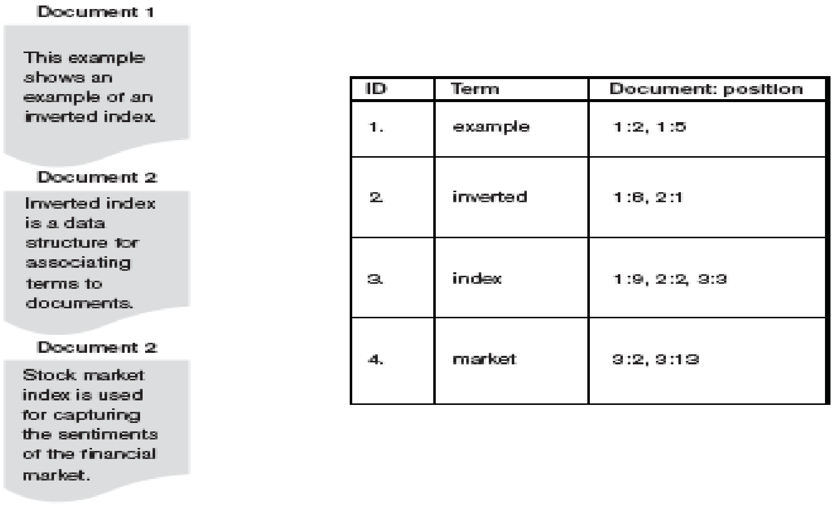
IR systems allow us to search over millions of documents. Finding the documents that contain the search terms from the document collection can be made by the linear search for each document. But this take time and increase the computing processes, it also retrieves the exact matching word only (Manning et al, 2008). To avoid this problem we will use what is known as index.

Index can be defined in general as a list of words or phrases (heading) and associated pointers (locators) to where useful material relating to that heading can be found in documents . Using this concept in the IR leads to improve the speed of searching and relevant retrieving by the assistance of the text preprocessing operations to form the indexing unit which knows the term (Manning et al, 2008).

The indexing unit may be: a word, stem, root, or n-gram. These unit can be obtained by: tokenizing the document base on white spaces or punctuation, use a stemmer to remove the affix, doing morphological operation to provide the basic manning of a word, and enumerating all the sequences of n characters occurring in term, respectively(Manning et al, 2008).

2.2.2.1 Inverted Index:

An inverted index is a data structure that stores a list of distinct terms which are found in the collection, this list is called a dictionary, lexicon or a term index. For each term a list of all documents that contain this term is attached, and it is known as the posting list (Elmasri R., S. Navathe, 2011). see Figure 2.2 below.



Document 1

This example shows an example of an inverted index.

Document 2

Inverted index is a data structure for associating terms to documents.

Document 2

Stock market index is used for capturing the sentiments of the financial market.

ID	Term	Document: position
1.	example	1:2, 1:5
2.	inverted	1:8, 2:1
3.	index	1:9, 2:2, 3:3
4.	market	3:2, 3:13

Figure 2.2: Inverted Index

Inverted index construction is done by collecting the documents that form the corpus. Afterwards the preprocessing operation is done on the documents to obtain the vocabulary terms; this term is used to build the forward index (document-term index) by creating a list of the words that are in each document. Finally, we invert or reverse the document-term matrix into a term-document stream to get the inverted index, this is why we got the word inverted index(Manning et al, 2008).

There are two variants of inverted index: record-level or inverted file index, it tells you which documents contain the term. And the word-level or full inverted index which contains additional information besides the document ID such as positions for each term within the document. This form of inverted index offers more functionality such as phrase searches.(Manning et al, 2008).

Given inverted index to search for documents relevant to the query, our first task is to determine whether each query term exists in the dictionary and then we identify the pointer to corresponding positing to retrieve the documents information and manipulate it based on various forms of query logic (Elmasri R. , S. Navathe, 2011).

2.2.3 Retrieval Models:

The IR model is a process that describes how an IR system represents documents and queries and how it predicts the retrieved documents that are relevant to a certain query.

The following sections will briefly describe the major models of IR that can be applied on any text collection. There are two main models: Boolean model and Ranked retrieval models or Statistical model, which includes the vector space and the probabilistic retrieval model.

2.2.3.1 Boolean Model:

The Boolean model or exact match model is a first IR model. This model is based on set theory and Boolean algebra. Queries are Boolean expression of keyword formalized using the operation of George Boole's mathematical logic, which define three basic operators (AND, OR and NOT) and use the bracket to indicate the scope of operators(Elmasri R., S. Navathe, 2011). Figure 2.3 illustrate how the Boolean model works.

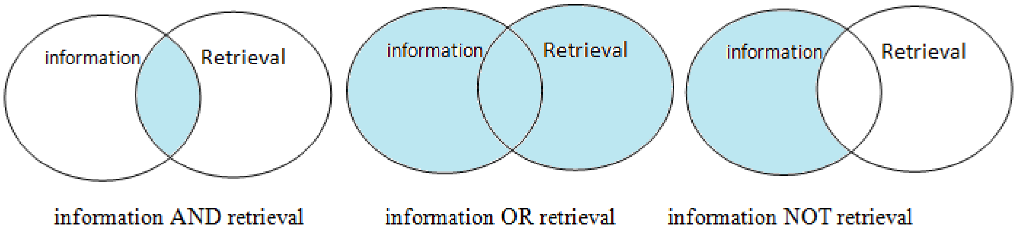


Figure 2.3:Boolean Combinations

Documents are considered as relevant to Boolean query expression if the terms that represent that document match the query expression exactly by tacking the query logic operators into account(Manning et al, 2008).

The main disadvantages of this model are does not provide a ranking for the result set, retrieving only exact match documents to query words and not easy for formalizing complex query.

2.2.3.2 Ranked Retrieval Models:

IR models use statistical information to determine the relevance of document with respect to query and ranked this documents descending according to relevance.

There are two major ranking models in IR: Vector Space Model and Probabilistic Retrieval Model(Ali, 2013).

1. **Vector Space Model:**

Vector Space Model (VSM) is a very successful statistical method proposed by Salton and McQill (Ali, 2013). The model represents the documents and queries as vector in multidimensional space, each dimension was represent term. The degree of multidimensionality is equal to the number of distinct word in corpus, in other word, number of terms that were used to build an index.

The vector component can be binary value represents the absence or presence of a given term in a given document which ignore the number of occurrences. Also, can be numeric value announce the term weight which reflect the degree of relative importance of a term in the corpus (Berry et al, 1999). This numeric value computed by combination of term frequency (**tf**) that can be defined as the number of occurrence of term in document, and the inverse document frequency (**idf**) which mean estimate the rarity of a term in the whole document collection (terms that occurs in all the documents is less important than another term whose appearance in few documents) - see Equation 2.1 and 2.2.TF-IDF weighting introduces extreme weights to words with very low frequencies and down weight for repeated terms. Other weighting methods are raw term frequency and inverted document frequency but these methods are not commonly used (Singhal A., 2001).

Retrieving the relevant documents corresponds to specific query do by computing the similarity between a query vector and the document vectors which deal with it as threshold or cutoff value. Cosine similarity is very commonly used in VSM which formulated as an inner product of two vectors divided by the product of their Euclidean norms - see Equation 2.3. Afterward, the documents ranking by decreasing cosine value that resulted as values between 1 and 0. Other similarity measures are possible such as a Jaccard Coefficient, Dice and

Euclidean distance. Figure 2.4 visualize an example of representing document vector and query vector in three dimension space.

$$\mathbf{W_{t,d} = tf_{t,d} \cdot idf_t} \qquad \mathbf{(2.1)}$$

$$\boldsymbol{idf_t} = \mathbf{\log \frac{|D|}{df_t}} \qquad \mathbf{(2.2)}$$

Where:

- $\mathbf{|D|}$ is the total number of documents in the collection.
- $\mathbf{df_t}$ is the number of documents in which a term appears.

$$\boldsymbol{sim(q,d) = \cos\theta} = \frac{\vec{q}.\vec{d}}{|\vec{q}|.|\vec{d}|}\mathbf{(2.3)}$$

Where:

- $\vec{\mathbf{q}}\,.\,\vec{\mathbf{d}}$ is the inner product of the two vectors.
- $|\vec{\mathbf{q}}|\mathbf{and}|\vec{\mathbf{d}}|$ are the Euclidean length of **q** and **d**, respectively.

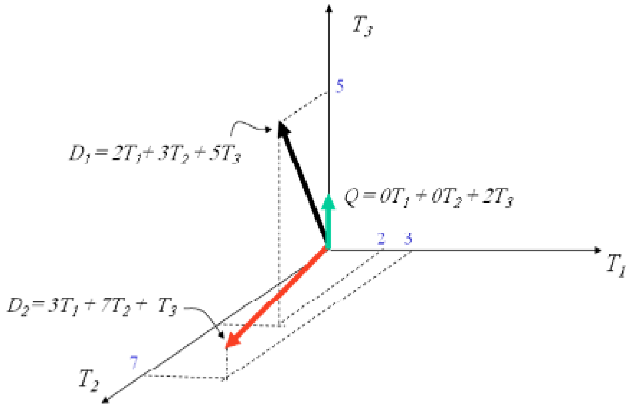


Figure 2.4: Query and Document Representation in VSM

Vector Space Model (VSM) solved Boolean model problem, but it suffers from main problem, namely (Singhal A., 2001): sensitivity to context, which is mean if the document is similar topic to query, but represented by different terms (synonyms) then wont retrieve since each of these term has a different dimension in the vector space. This problem was solved by a new version called latent semantic Analysis (LSA).

2. Probabilistic Retrieval Model:

Users usually write a short query that makes the IR system has an uncertain guess of whether a document is relevant for the query. Probability theory provides a principled foundation for such reasoning under uncertainty.

Probabilistic Retrieval Model is based on the probabilistic ranking principle (PRP), which state that a documents in collection should be ranked decreasing based on their probability of being relevant to the query by represent the document and query as binary term incidence vectors (presence or absence of a term) to predict a weight for that term and merge all weights of the query terms to determine if the document is relevant and amount of it or not relevant, P(R|D)(Singhal A., 2001), With this representation, many possible documents have the same vector representation and recognizes no association between terms(Manning et al, 2008). This concept is the basis of classical probabilistic models which known as Binary Independence Retrieval (BIR) model which is a ratio between the probability that the document belongs to relevant set of documents and the probability that the document belongs to the set of irrelevant documents- see the following formal:

$$\mathrm{Odd}(R|D) = \frac{P(R|D)}{P(\bar{R}|D)} = \frac{P(R|D)}{1-P(R|D)} \qquad (2.4)$$

The Binary Independence Retrieval Model was originally designed for short catalog records of fairly consistent length, and it works reasonably in these contexts. For modern full-text search collections, a model should pay attention to term frequency and document length. BestMatch25 (BM25 or Okapi) is sensitive to these quantities. From 1994 until today, BM25 is one of the most widely used and robust retrieval models (Ali, 2013). The equation used to compute the similarity between a document d and a query q is:

$$\mathrm{BM25}(d,q) = \sum_{t\in q}\left[\log\frac{N-n+0.5}{n+0.5}\right] . \left[\frac{(k_1+1)tf_{tD}}{k_1((1-b)+b\frac{L_D}{\mathrm{avg}\,L_D})+tf_{tD})}\right] . \left[\frac{(k_3+1)tf_{tq}}{k_3+tf_{tq}}\right](2.5)$$

Where:

- **N** is the total number of documents in a collection.

- $\mathbf{n}$ is number of documents containing the term.
- $\mathbf{tf_{tD}}$is the frequency of term t in the document D.
- $\mathbf{L_D}$is the length of document D.
- $\mathbf{avg\ L_D}$is the average document length across the collection.
- $\mathbf{k_1}$is a parameter used to tune term frequency in a way that large values tend to make use of raw term frequency. For example, assigning a zero value to k_1 corresponds to not considering the term frequency component, whereas large values correspond to raw term frequency. k_1 is usually assigned the value 1.2.
- **b** is another free parameter where b [0,1]. The value 1 means to completely normalizing the term weight by the document length. b is usually assigned the value 0.75.
- $\mathbf{k_3}$ is another parameter to tune term frequency in query q

2.2.4 Type of Information Retrieval System:

IR System has been classified into three groups: Monolingual, Cross-lingual and Multilingual. Monolingual IR system mean the corpus contained documents for single language, when the users search; query must be written by the same language of documents. Cross-lingual or Cross Language Information Retrieval (CLIR) system the collection consist document in single language and users written queries using language differ from documents language to retrieve that documents match the translated query. The last group of IR systems is Multilingual system in this case the corpus contained mixed documents and query also written in mixed form.(Ali, 2013)

2.2.5 Query Expansion:

Query expansion is the technique of adding more information (synonyms and related terms) to the input query in order to give more clarity to the original query and improve the performance of IR system. This technique is based on finding the relationships between the terms in the document collection. Figure 2.5 illustrates how the original query "Java" extended by the related term "sun" to retrieve more relevant documents were semantically correlated.

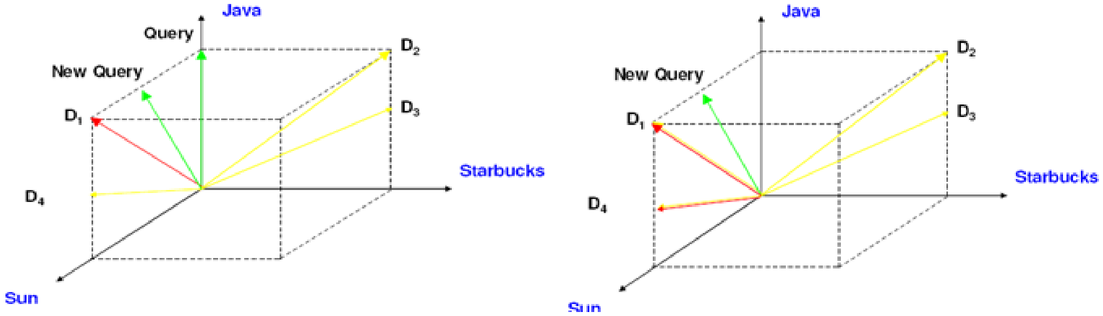


Figure 2.5: Extended the Query "java" by the Related Term "sun"

Query expansion can be done by one of two ways: automatically using resources such as WordNet or thesaurus which each term in the query will expand with words that listed as similarity related in it, these resources can be generated manually by editors (e.g.., PubMed) or via the co-occurrence statistics.The advantage of this approach is not requiring any user input to select the expansion terms, however it's very expensive to create a thesaurus and maintain it over time.

Another way to expand the queries will do semi-automatically based on relevance feedback when the search engine shows a set of documents (Shaalan K., 2012). Relevance feedback approach made by two manners (Manning et al, 2008): The first one which was proposed by Rocchio in 1965, users mark some documents as relevant and the other documents as irrelevant. Use the marked documents to form the new query and run it to return the new result list. We can iterate it several times. The second one was developed in the early 1990s (Du S., 2012), automate the part of selecting the relevant documents in the prior method by assuming the top K documents are relevant, after that do as the previous approach. These approaches suffer from query drift due to several iterations and made long queries that expensive to process.

Query expansion handles the issue of term mismatch between a query and relevant documents. Get an appropriate way to expand the query without hurting the performance nor allow search intent drift is crucial issue due to success or failure is often determined by a single expansion term (Abdelali, 2006).

2.2.6 Retrieval Evaluation Measures:

In order to measure the IR system’s performance, the test collections, which is consisted of a set of documents, queries, and relevance judgments that specify which documents are relevant to each query, and an evaluation techniques are used. These evaluation measures depend on type of assessing documents if it unranked (binary relevance judgments) or ranked set.

Two basic measures can be used in the binary relevance assumption (document is relevant or irrelevant to the query) is precision and recall. Precision is defined as the ratio of relevant documents correctly retrieved by the system with respect to all documents retrieved by the system(see Equation 2.6).Recall is defined as the ratio of relevant documents were retrieved from all relevant documents in the collection(see Equation 2.7).For a certain query the documents can be categorized into four sets. Figure 2.6 is a pictorial representation of these concepts. When the recall increases by returning all relevant documents in the collection for all queries the precision typically goes down and vice versa. In all IR systems we should tune the system for high precision and high recall. This can be made by trades off precision versus recall, this concept called an F-measure. The F-measure or F-score is the harmonic mean of precision and recall (see Equation 2.8). The main benefit from the harmonic mean is automatically biased toward the smaller values. Thus a high F-score mean high precision and recall.

	Relevant	Irrelevant
Retrieved	A	C
Not retrieved	B	D

Figure 2.6: Retrieved vs. Relevant documents

$$precision = \frac{A}{(A \cup C)} \tag{2.6}$$

$$Recall = \frac{A}{(A \cup B)} \tag{2.7}$$

$$F = \frac{2pr}{p+r} \tag{2.8}$$

When considering the relevance ranking, we can use the precision to evaluate the effectiveness of the IR System as the same way of Boolean retrieval by treating all documents above the given rank as an unordered result set, and calculate precision at cutoff k. This is called precision at K measure. This measure focuses on retrieving the most relevant documents at a given rank and ignores the ranking within the given rank. The main objection of this approach, it does not take the overall recall in the account.(Ali, 2013) (Webber, 2010)

Recall and precision can also be combined to evaluate the ranked retrieval results by plotting the precision and recall values to give which is known as a precision-recall curve (Manning et al, 2008).There are two ways of computing the precision: Interpolate a precision or Mean Average Precision (MAP). The interpolated precision at the i-th standard recall level is the largest known precision at any recall level between the i-th and (i + 1)-th level.MAP is the average precision at each standard recall level across all queries, this measure is widely used in the evaluation of IR systems(Manning et al, 2008)(Ali, 2013) (Elmasri R. , S. Navathe, 2011) (Webber, 2010).

To evaluate the effectiveness of our graded relevance we use the Discounted Cumulative Gain measure (DCG), a commonly used metric for measuring the web search relevance (Weiet al, 2010). DCG is an expansion of Cumulative Gain (CG) which sum of the graded relevance values of a result set without taking into account the position of the document in the result-see equation 2.9 (Ali, 2013).

$$CG_p = \sum_{i=1}^{p} gradedrel_i \qquad (2.9)$$

The DCG is based on two assumptions: the highly relevant documents are more useful than lesser relevant documents and more valuable when appear with a top rank in the result list. Stand on these assumptions we note the DCG measures the total gain of a document which accumulate from the top to the bottom based on its position and relevance in the provided list-see Equation 2.10. The principle of DCG is the graded relevance value of the document is a discount logarithmically by the position of it in the result.

$$DCG_p = gradedrel_1 + \sum_{i=2}^{p} \frac{gradedrel_i}{\log_2 i} \qquad (2.10)$$

Evaluate a search engine's performance can't make using DCG alone, for the reason that result lists vary in length depending on the query. Normalized Discounted Cumulative Gain (NDCG)-see Equation 2.11- measure was used to solve this issue by normalizing the DCG value by the use of the Idle DCG (IDCG) value that is obtained from the perfect ranking of documents using the same query(Ali, 2013)..

$$NDCG_p = \frac{DCG_p}{IDCG_p} \qquad (2.11)$$

No single measure is the correct one for any application choose measures appropriate for task.

2.2.7 Statistical Significance Test:

Statistical significance tests help us to compare between the performances of systems to know if an improvement of one system over another has significant mean or just occurred by pure chance (CD Manning, H Schütze,1999). Suppose we would like to know whether the average precision of a system that expands queries by words that used in the other Arab society (method A) is significantly better than the same system with non-expansion(method B). The evaluation well done in the same environment, in the context of IR that is mean the same set of queries(CD Manning, H Schütze,1999).

The most commonly used statistical tests in IR experiments are the Student's t-test (Abdelali, 2006). Tests of significance are typically to a 95% confidence level and the remaining 5% of performance is considered as an acceptable error level that is meant if a significance test is reliable, then at 95% of choices of 'A' will go above that of 'B' and the 5% is the probability of being a false positive. In further words, since the significance value represents the probability of error in accepting that the result is correct, the value 0.05 is considered as an acceptable error level(p-value< 0.05)(Ali, 2013)(Abdelali, 2006).

Student‘s t-test is hypothesis testing .Hypothesis testing involves making a decision concerning some hypothesis or question to decide whether this question, given the observed data, can safely assume that a certain hypothesis is true, or that we have to reject this hypothesis. T-test use sample data to test hypotheses about an unknown data mean, and the only available information about the data comes from the sample to evaluate the differences in means between two groups. The test looks at the difference between the observed and expected means, scaled by the variance of the data (see Equation 2.12).(CD Manning, H Schütze,1999)

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{N}}} \qquad (2.12)$$

where:

- $\bar{X}$ is the sample mean
- μ is the mean of the distribution
- S^2 is the sample variance
- N is the sample size

2.3 Arabic Language:

The Arabic language is the most widely spoken language of the Semitic family which also includes Hebrew(spoken in Israel), Tigre(spoken in Eritrea) ,Aramaic(spoken in Iraq) and Amharic(spoken in Ethiopia)(Ali, 2013).Arabic is broadly spread because it is the religious language of all Muslims, language of science in the middle age and part of the curriculum in most of non-Arabic countries such as Iran and Pakistan. Arabic is the only language of Semitic languages, which preserved the universality while most Semitic languages have abolished.

The Arabic alphabet consists of 28 basic characters which are called '*hurofalheaja*' (حروف الهجاء) which are written and read from right to left and numbers from left to right (see Figure 2.7). In the past these characters were written without dots and diacritical marks. In the seventh century, dots and diacritical marks were added to the language to reduce ambiguity (Ali, 2013) (Abdelali, 2006).Arabic language doesn't have letters dotted by more than three dots (see Figure 2.8). The typographical form of these characters depending on whether they appear at the beginning, middle or end of a word, or on their own (see Table 2.1) and the diacritical marks for each character are set according to the meaning we want to obtain from the word. Arabic words are divided into three types: noun, verb, and particle. Noun can be singular, dual, or plural and masculine or feminine (Darwish .K, W. Magdy,2014) (Musaid, 2000).

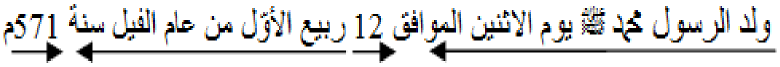

ولد الرسول محمد ﷺ يوم الاثنين الموافق 12 ربيع الأوّل من عام الفيل سنة 571م

Figure 2.7: Arabic language writing direction

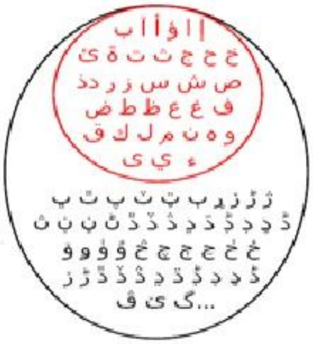


Figure 2.8: Difference between Arabic and Non-Arabic letter

Table 2.1: Typographical Form of /ba/ Letter

/ba/ letter (حرف الباء)			
Beginning	Middle	end of a word	their own
بدر	مبادئ	تعجب	ب

The Arabic language is an aggregate of multiple varieties including: Classical Arabic (CA), Modern Standard Arabic (MSA) and Regional or Dialectal Arabic (DA) which are called: Qur'an Arabic, FUSHA"العربية الفصحى" and LAHJA"لـَهْجة" or AMMIYYA"عامية", respectively. Classical Arabic is the language of the Quran and classical literature.MSA is the universal language of the Arab world which is understood by all Arabic speakers and used in education and official settings. Dialectal Arabic is a commonly used, region specific and informal variety, which have no standard orthographies but have an increasing presence on the web.(Ali, 2013)(Darwish .K, W. Magdy,2014) (Mona Diab,2014)

The Arabic Language varies from European and Asian languages in two aspects: morphologically and syntactically (Ghassan Kanaan et.al,2005). The Arabic language is very complex morphologically when compared to Indo-European languages because Arabic is root based while English for example is stem based and highly derivational(Abdelali, 2006). The words are derived from a root (which is usually a sequence of three consonants) by applying patterns which involve adding infix or replacing or deleting a letter or more from the root using derivational morphology (srf, علم الصرف) which define as the process of creating a new word out of an old word, usually by adding affixes and then adding prefixes and suffixes if needed(Ghassan Kanaan et.al, 2005). Adding prefix and suffix to the words gives them some characteristics such as the type of verb (past, present or امر) and gender, number, respectively. Although, Arabic has very complex morphology it is very flexible syntactically as it tolerates modifying the order of the words in the sentence e.g. "كتب الولد القصيدة" has the same meaning of "الولد كتب القصيدة"(Ali, 2013)(Abdelali, 2006).

The Arabic language is categorized as the seventh top language on the web (see Figure 2.9) which shows how Arabic is the fastest growing language on the web among all other languages (Darwish .K, W. Magdy,2014). As there are few search engines interested in Arabic language they don't handle the levels of ambiguity in Arabic which will be mentioned below. This leads researchers to focus on Arabic language information retrieval and natural language processing systems.

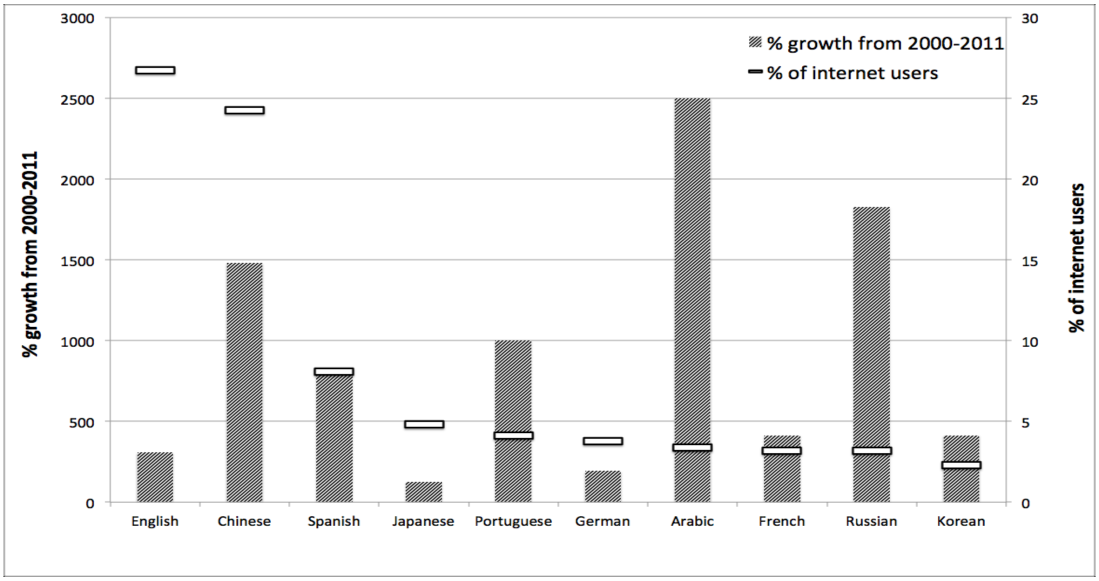


Figure 2.9: Growth of Top 10 languages in the Internet by 31 Dec 2011 (Darwish .K, W. Magdy,2014).

2.3.1 Level of Ambiguity in Arabic Language:

The Arabic language poses many challenges for retrieval due to ambiguity that is found in it which is caused by one or more of the Arabic features. We expound these levels of ambiguity in details and describe their effects on retrieval in the following subsections.

2.3.1.1 Orthography Level:

Orthographic variations in Arabic occur due to various reasons. The different typographical forms for one letter such as ALEF (أ,إ ,آ and ا), YAA with dots or without dots (ي and ى) and HAA (ة and ه) play a role in variations. Substituting one of these forms with another will sometimes changes the meaning of the words. For instances, "قران" (meaning: Quran) it change to "قرآن" (meaning: marriage contract) also "ذُره" (meaning: Corn) it change to "ذَره" (meaning: Jot). Occasionally, some letters when replaced with other letters can cause misspelling but do not change the meaning and phonetic of the words e.g. "بهاء" and "بهائية" (meaning: his glory). These variations must be handled before using the words in document retrieving by normalizing the letter (Ali, 2013) (Darwish .K, W. Magdy,2014). This has been done for four letters:

1. أ,إ ,آ and ا normalized to ا.
2. ي and ى normalized to ى.
3. ة and ه normalized to ه.
4. ؤ, ئ and ء normalized to ء

An additional factor that can cause orthographic variation is the presence and absence of diacritical mark. Diacritical mark refers to symbol or short vowel that come above or below Arabic character to define the sense of the words and how it will be pronounced which helps us to minimize the ambiguity. For instance, "حَب"(meaning: seed) it change to "حُب"(meaning: love). Every Arabic letter can take any one of these marks: KASRA, FATHA, DAMA and SUKUN. The first mark is written below the letters and the rest are written only above the letters. FATHA, KASRA and DAMA called the short vowel. Extra diacritics mark which is used to implicit repetition of a letter is SHADDA that appears above the character. Nunation or TANWEEN is a short vowel in double form which is unlike other diacritical marks does not change the meaning of words but just the sound. These diacritics mark can be combined (Ali, 2013) (Darwish .K, W. Magdy,2014)(Abdelali, 2006). Table2.2 illustrated how diacritical marks change the pronunciation of letter.

Table 2.2: Effect of diacritical mark in letter pronunciation

Letter	Diacritics mark	Sound	Letter	Diacritics mark	Sound
بً	Nunation	/ban/	بَ	FATHA	/ba/
ب	Nunation	/bin/	ب	KASRA	/bi/
بٌ	Nunation	/bun/	بُ	DAMA	/bu/
بّ	SHADDA	/bb/	بْ	SUKUN	/b/
بُّ	Combination	/bbu/	بًّ	Combination	/bban/

Although the diacritical marks remove ambiguity, most of the text in a web page is printed without these diacritical marks. This issue can be solved by performing diacritic recovery but this is very computationally expensive, large index and facing problem when dealing with unseen words. The commonly adopted approach is removing all diacritical marks this increases the ambiguity but computationally efficient (Darwish .K, W. Magdy,2014).

Orthographic variations can also occur with transliteration of non-Arabic words to Arabic (Darwish .K, W. Magdy,2014). For example, England transliteration to"انجلترا" and "انكلترا" also bachelor it gives different forms like "بكالوريوس" and "بكلوريوس".This problem causes mismatching between the documents and queries if the systems depend on literal matches between terms in queries and documents.

2.3.1.2 Morphological Level:

Arabic language is derivational system based on a set of around 10000 roots (Darwish .K, W. Magdy,2014). We can build up multiple words from one root which made the Arabic has complex morphology which can increases the likelihood of mismatch between words used in queries and words in documents. For instance, creating words like /***kitāb*** "book", ***/kutub*** "books", /***kātib*** "writer", ***/kuttāb*** "writers", ***/kataba*** "he wrote", */ya**ktu**b**u*** "they write", from the root (ktb) "write". The root is a past verb and singular composed of three consonants (tri-literals), four consonants (quad-literals) or five consonants (pet-literals), which always represents lexical and semantic unit. Words derived by using a pattern which refer to standard frame which we can apply on roots by adding infix, deleting character or replacing a letter by another letter. Subsequently, attaching the prefix and suffix for adding the characteristics which mentioned earlier section if needed. The main pattern in Arabic is "فعَل" (transliterated as f-à-l) and other patterns derived from it by affix letter at the start "يفعل" (transliterated as y-fà-l), medially"فعال" (transliterated as f-à-a-l), finally "فعلن"(transliterated as f-à-l-n) or mixture of them "يفعلون"(transliterated as y-f-à-l-o-n). The new pattern words may have the same meaning of roots or different meanings. Table 2.3 show derivational morphology of "كتب" KTB (in English: writing).(Ali, 2013) (Darwish .K, W. Magdy,2014) (Musaid, 2000)

Table 2.3: Derivational Morphology of "كتب" /KTB "writing"

Word	Pattern	Meaning	Word	Pattern	Meaning
كتاب /kitāb/	فعال	Book	مكتبة/maktaba/	مفعلة	Library
كتب /kutub/	فعل	Write	مكتب /maktab/	مفعل	Office
مكتوب /maktūb/	مفعول	Letter	كاتب /kātib/	فاعل	writer

The Arabic language attach many particles include suffix like (وا,نا,هم ..etc) and prefix like (ت,ن,س ..etc) to words which it make it so difficult to known if these particles are attached particles or a part of roots. This issue is one of the IR ambiguities.

There are many solutions to handle the morphology issues to reduce the ambiguity: one of them is by using the morphological analyzer technique to recover the unit of meaning (root). This solution is facing ambiguity in indexing and searching because all fended analyses has the same degree of likeness. Another solution made by finding all possible prefix and suffix for the word and then compares the remaining root with a list of all potential roots. This approach has the same weakness of the previous solution. The most common solution is so-called light stemming which improves both recall and precision (Darwish .K, W. Magdy,2014)

Light stemming is affix removal stemming which chop out the suffixes and prefixes of the word without trying to find the linguistic root. Light stemming like light10 is stem-based which outperforms root-based approaches like Khoja that chopping off prefixes, infixes and suffixes (Ali, 2013).

The light10 stemmer removes the prefix (ال، وال،لل، بال، كال، فال، و) and the suffixes (ها، ان، ات، ون، ين، يه، ية، هـ، ة، ي) from the words (Ali, 2013). But Khoja use the lists of valid Arabic roots and patterns. After every prefix or suffix removal, the algorithm compares the remaining stem with the patterns. When a pattern matches a stem, the root is extracted and checked against the list of valid roots. If no root is found, the original word is returned (KHOJA S., GARSIDE R., 1999).

2.3.1.3 Semantic Level:

Documents are constructed for communication of knowledge. The knowledge exists in the author’s mind; the author uses his own words to transfer this knowledge. Arabic has a very rich vocabulary, many of these words describes different forms of a particular word or object. This phenomenon is known as synonyms, that is, two or more different words have similar meaning which can used by different authors to deliver the same concept. This phenomenon causes a greater challenge in finding the semantically related documents.

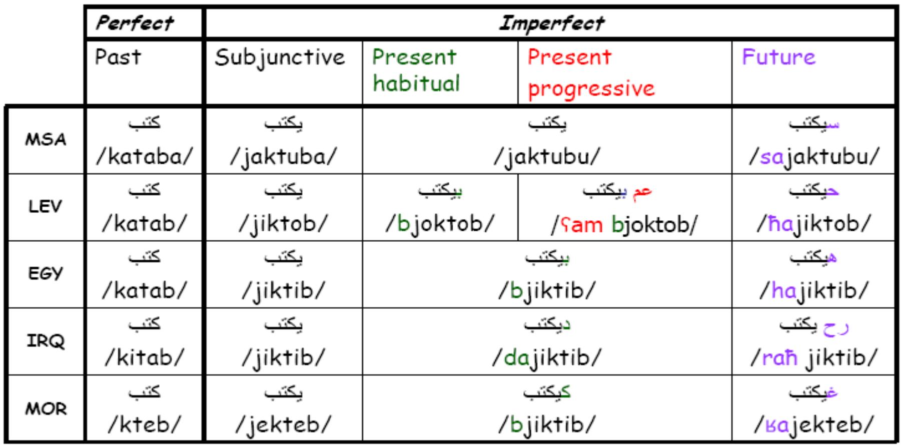
In the past, synonym in Arabic has two forms(H. AbdAlla,2008): different words to express the same meaning, e.g. "المسند,الدهر,السمير,العتك,السبت" (meaning: year) or resulting from applying morphological operation to derive different words from the same root, e.g. "عرض" (meaning: display) and "يعرض " (meaning: displaying). At the present time, regional variations or dialects in vocabulary considered as a new form of synonym, like the words (الصحيه,السبيطار,اسبتاليه and دختر) which mean hospital.

Dialects or colloquial is the number of spoken vernaculars in Arab world. Arabic speakers generally use the dialects in daily interactions. There are four main dialects, namely: North Africa (Maghreb), Egyptian Arabic (Egypt and the Sudan), Levantine Arabic (Lebanon, Syria, Jordan and Palestine/Palestinians in Israel), and Iraqi/Gulf Arabic (Abdelali, 2006). Dialectical differences within the same region can be observed. Dialects Arabic (DAs) differ lexically (see Table 2.4), morphologically (see Figure 2.10) and lesser degree syntactically(see Table 2.5)from MSA and also from one another, and does not have standard spelling because pronunciations of letters often differ from one dialect to another. Changes of pronunciations can occur in stems. For example, the letter "ق" /q/ is typically pronounced in MSA as an unvoiced uvular stop (as the /q/in 'quote'), but as a glottal stop in Egyptian and Levantine (like /A/ in 'Alpine') and a voiced velar stop in the Gulf (like /g/ in 'gavel').Some changes also occur in phonetics of prefixes and suffixes, for example in the Egyptian dialect the prefix "س" /s/ meaning 'will' is converted to "ح" /H/ in North Africa(Khalid Almeman, Mark Lee,2013) (Abdelali, 2006) (Hassan Sajjad et al, 2013).

In Arabic, such differences we mentioned above have a direct impact on Arabic processing tools. Dialect electronic resources like corpora and dictionaries and tools are very few, but a lot of resources exist for MSA(Wael, Nizar, 2012). There are two approaches for dealing with region variation: the first one is dialect-to-MSA translations which can be done by auxiliary structures like dictionaries or thesauruses and the second is mathematically and statistically model.

Table 2.4: Lexically Variations in Arabic Language

English	MSA	Iraq	Sudanese	Libya	Morocco	Gulf	Philistine
Shoes	حذاء	قندره	نعال– جزمه	مداس	سباط	جوتي	كندره
Pharmacy	صيدلية	ازة خانة	شفخانه– اجزخانه	–	فرماسيان	–	–
Carpet	سجاده	اورطه– سوباط	سجاده	–	سداجة	زوليه	–
Hospital	المستشفى	–	اسبتاليه	السبيطار	–	الدختر- سبيتار	–



	Perfect	Imperfect			
	Past	Subjunctive	Present habitual	Present progressive	Future
MSA	كتب /kataba/	يكتب /jaktuba/	يكتب /jaktubu/		سيكتب /sajaktubu/
LEV	كتب /katab/	يكتب /jiktob/	بيكتب /bjoktob/	عم بيكتب /ʕam bjoktob/	حيكتب /ħajiktob/
EGY	كتب /katab/	يكتب /jiktib/	بيكتب /bjiktib/		هيكتب /hajiktib/
IRQ	كتب /kitab/	يكتب /jiktib/	ديكتب /dajiktib/		رح يكتب /raħ jiktib/
MOR	كتب /kteb/	يكتب /jekteb/	كيكتب /bjiktib/		غيكتب /ʁajekteb/

Figure 2.10: Morphological Variations in Arabic Language

Table 2.5: Syntactically Variations in Arabic Language

Dialect/Language	Example
English	Because you are a personality that I cannot describe.
Modern Standard Arabic	لانك شخصية لا استطيع وصفها
Egyptian Arabic	لانك شخصية بجد مش هعرفاوصفها
Syrian Arabic	لانك شخصية عنجد مارح اعرف اوصفها
Jordanian Arabic	انت جد شخصية مستحيل اقدر اوصفها
Palestinian Arabic	عن جد شخصية ما بتنوصف
Tunisian Arabic	على خاطرك شخصية بلحق منجمشنوصفها

2.3.2 Region Variation Approaches:

2.3.2.1 Dialect-to-MSA Translation Approach:

Translation in general, is a process of translate word from language (e.g. Arabic) to another (e.g. English). IR used this idea to translate query form one language to another in order to help a user to find relevant information written in a different language to a query this concept known as cross-language information retrieval (CLIR).

To manipulate with Arabic dialects in IR, researchers have used different translation approaches same as CLIR approaches to map DA words to their MSA equivalents rather than mapping a words to unlike language. The translation approaches are machine translation, parallel corpora and machine readable dictionaries (Ali, 2013) (Nie, 2010).

1. Machine Translation Approach:

In general, we can classify Machine Translation (MT) systems into two categories: the rule-based MT system and the statistical MT system. The rule-based MT system using rules and resources constructed manually. Rules and resources can be of different types: lexical, phrasal, syntactic, semantic, and so on. Statistical Machine Translation (SMT) is built on statistical language and translation models, which are extracted automatically from large set of data and their translations (parallel texts). The extracted elements can concern words, word *n*-grams, phrases, etc. in both languages as well as the translations between them (Nie, 2010).

2. Parallel Corpora Approach:

Parallel Corpora are texts with their translations in another language are often created by humans as a manual translation process (Nie, 2010). Finding the translation of the word in other language do with aligned the text. To get the relevant document for specific query regard less of user's region using this approach we need to multidialectal Arabic parallel corpus.

3. Dictionary Translation Approach:

Dictionary is a list of word or phrase in the source language and the corresponding translation in the target language. There are many bilingual dictionaries available in electronic forms. The IR researchers extended this idea to build monolingual dictionaries to solve the dialect issue.

2.3.2.2 Statistically Model Approach:

A Statistical model can be defined as a flexible approach because it is based on mathematical foundations. The main idea of this approach relies on the assumption that terms occur in similar context are synonyms. The remain of this section contains illustration of the commonly statistical model which known as Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI).

Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI) (DuS., 2012)is an extension of the vector space retrieval model to deal with language issue of ignoring the semantic relations (synonymy) between terms in VSM to retrieve the relevant documents regardless of exact matching between a query terms and documents by finding the hidden meaning of terms(Inkpen, 2006).The difference between LSI and LSA are LSI using for indexing and LSA using for everything.LSA is a mathematical and statistical approach, claiming that semantic information can be derived from a word-document co-occurrence matrix. LSA also used in automated documents categorization (clustering) and polysemy Phenomenon which refers to the case that a term has multiple meanings e.g. "عامل" (EAMIL) which mean 'worker' and 'factor'. LSA basing on assumption that words that are used in the same contexts are close in meaning and then represents it in similar ways, in other word, in the same semantic space.(DuS., 2012)

LSA uses the mathematical technique to reduce the dimension of a term-document matrix to group those terms that occur in similar contexts (synonyms) in one dimension (latent semantic space) rather than dimension for each terms as VSM (Du S., 2012). The dimension reduction technique was use here called singular value decomposition (SVD) which can applied in any matrix that vary from the principal component analysis (PCA)which manipulate with rectangular matrices only (Kraaij, 2004).

Singular value decomposition (SVD) is a reduction technique that project semantically related terms onto same dimension and independent terms onto different dimension, based on this concept the recall of query will be improved(Kraaij, 2004).SVD decompose the term-document matrix$A_{m\times n}$ into the product of three matrices(see Equation 2.13 and Figure 2.11) to obtain low rank approximation matrixA'.The first component in the equation describes the term matrix and the second one is square diagonal matrix which contain non-zero entries called singular values of matrix A that sorting descending to reflect the important of dimension to assist in omitted all unimportant dimensions from U and V. The third is a document vectors. The choice of rank, latent features or concepts (*r*) is critical to the performance of LSA. Smaller (*r*) values generally run faster and use less memory, but are less accurate. Larger *r* values are more true to the original matrix, but require longer time to compute. Experiments prove choosing values of *r* ranged between 100 and 300 lead to more effective IR system (Berry et al, 1999) (Abdelali, 2006).

$$A = U\sum V^T = (\mathrm{orthonormal})_{\mathrm{m\times r}}(\mathrm{diagonal})_{\mathrm{r\times r}}(\mathrm{orthonormal})_{\mathrm{r\times n}} \qquad (2.13)$$

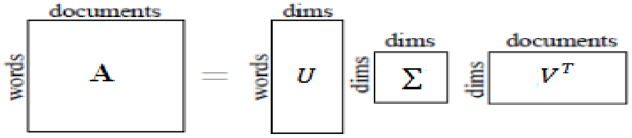


Figure 2.11: SVD Matrices

where:

- Orthonormal matrix means vectors have unit length and each two vectors are orthogonal.
- Diagonal mean matrix all elements are zero expect the diagonal

In order to retrieve the relevant documents for the user, a user's query adapt using SVD to r-dimensional space(see Equation 2.14). Once the query and documents represent in LSI space, now we can use any similarity measure such as cosine similarity in VSM to return the relevant documents(Manning et al, 2008).

$$\hat{q} = q^T U_r \sum_r^{-1} \qquad (2.14)$$

Advantage of LSI:

- Mathematical approach: this makes it strong and can be applied in any text collection language.
- Handling synonyms and polysemy Phenomenon, Formally, polysemy (words having multiple meanings) and synonymy (multiple words having the same meaning) are two major obstacles to retrieving relevant information (Du S., 2012).

Disadvantage of LSI:

- Calculation of LSI is expensive (Inkpen, 2006).
- Cannot be used an inverted index due to cannot locate documents by index keywords (Inkpen, 2006).
- Derivational of words casus camouflage these can be solve using stemmer.
- Require re-computation for LSI representation when new documents added (Manning et al, 2008).

2.4 Related works:

Some work has been proposed to deal with Arabic Dialect in IR. these work classify to two approaches: the first one is dialect-to-MSA translations which can be done by auxiliary structures like dictionaries or thesauruses and the second is mathematically and statistically model (Distributional approaches) is based on the distributional hypothesis that words that occur in similar contexts also tend to have similar meanings/functions.

To manipulate with Arabic dialects in IR, researchers have used different translation approaches was mentioned above to map DA word to their MSA equivalents.

(Wael, Nizar,2012) they describe the implementation of MT system known as ELISSA. ELISSA is a machine translation (MT) system from DA to MSA. ELISSA uses a rule-based approach that relies on the existence of DA morphological analyzers, a list of hand-written transfer rules, and DA-MSA dictionaries to create a mapping of DA to MSA words and construct a lattice of possible sentences. ELISSA uses a language model to rank and select the generated sentences. ELISSA currently handles Levantine, Egyptian, Iraqi, and to a lesser degree Gulf Arabic.

(Houda et al, 2014)present the first multidialectal Arabic parallel corpus, a collection of 2,000 sentences in Standard Arabic, Egyptian, Tunisian, Jordanian, Palestinian and Syrian Arabic which makes this corpus a very valuable resource that has many potential applications such as Arabic dialect identification and machine translation.

Another approach to deal with Arabic Dialect by building monolingual dictionaries to solve the dialect issue, (Mona Diab et.al, 2014) build an electronic three-way lexicon, Tharwa. Tharwa is the first resource of its kind bridging two variants of Arabic (Egyptian Arabic, MSA) with English besides it is a wide coverage lexical resource containing over 73,000 Egyptian entries and provides rich linguistic information for each entry such as part of speech (POS), number, gender, rationality, and morphological root and pattern forms. The design of Tharwa relied on various preexisting heterogeneous resources such as Hinds-Badawi Dictionary (BADAWI) which provides Egyptian (EGY) word entries with their corresponding English translations and definitions, Egyptian Colloquial Arabic Lexicon (ECAL) is a machine readable monolingual lexicon which contain only EGY entries with a phonological form, an undiacritized Arabic script orthography form, a lemma, and morphological features for each word, Columbia Egyptian Colloquial Arabic Dictionary (CECAD) is a three-way (EGY-MSA-ENG) small lexicon consists of 1,752 entries extracted from the top most frequent entries in ECAL, CALIMA Lexicon (CALIMA-LEX) is an EGY morphological analyzer relies on the ECAL and SAMA Lexicon is a morphological analyzer for MSA.

Some related works deal with Arabic Dialect in IR systems are based on Latent Semantic Analysis (LSA) which is a Statistical model which consider as a flexible approach because it is based on mathematical foundations. The assumption behind the proposed LSA method is that it is nearly always possible to determine the synonyms of a word by referring to its context.

(Abdelali, 2006) discussed ways of improving search results by avoiding the ambiguity of regional variations in Arabic-speaking countries through restricting the semantics of the words used within a variation using language modeling (LM) techniques. Colloquial Arabic that were covered by Abdelali categorize to Levantine Arabic, Gulf Arabic, Egyptian Arabic and North-African Arabic. The proposed solutions, Abdelali alleviate some of the ambiguity inherited from variations by clustering the documents based on variant (region) using the *k*-means clustering algorithm and built up index corresponding to each cluster to facilitating a direct query access to a more precise class of documents (see Figure 2.12). Once the documents are successfully clustered, the clusters will be merged to build the language model (LM).Semantic proximity is represented by semantic vectors, based on vector space models. The semantic vectors form from term-by-term matrix show the co-occurrence between the terms within specific size of window. The size of the matrix reduces by Singular Value Decomposition (SVD) method to construct which is Known Latent Semantic Analysis (LSA). The results proved significant improvement in recall and precision compared to the baseline system by applying query expansion techniques.

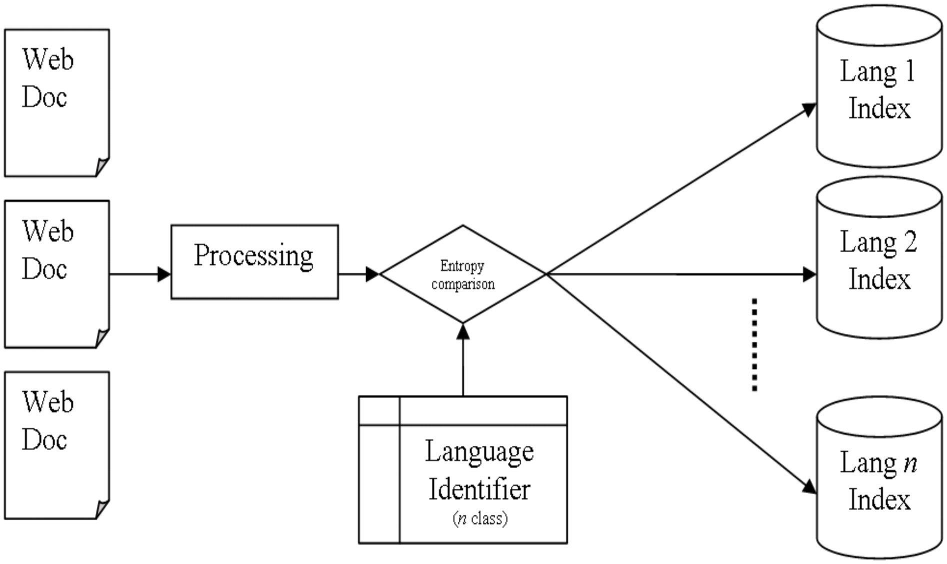


Figure 2.12: Process of searching on multi-variant indices engine

(Mladen Karan et.al., 2012) proposed a method for identifying synonyms in Croatian language using two basic models of distributional semantic models (DSM) on the larger Croatian Web as Corpus (hrWaC corpus) and evaluated the models on a dictionary-based similarity test. Theses DSMs approaches, namely: latent semantic analysis (LSA) and random indexing (RI).

In order to reduce the noise in the corpus we filtered out all words with a frequency below 50. This left us with a corpus containing 5,647,652 documents, 1.37G tokens, 3.89M word-form types, and 215,499 lemmas. To remove the morphological variations which scatter vectors over inflectional forms, we use the semi-automatically acquired morphological lexicon for Croatian language to employed lemmatization and consider all possible lemmas when building DSMs.

Evaluation was done based on 10 models: six random indexing models and four LSA models. The differences between models come from the way of how the large size of the hrWaC corpus is reflected in the dimensions in term-context co-occurrence matrices. LSA uses documents and paragraphs, and RI uses documents, paragraphs, and neighboring words as contexts. Results indicate that LSA models outperform RI models on this task. The best accuracy was obtained using LSA (500 dimensions, paragraph context): 68.7%, 68.2%, and 61.6% on nouns, adjectives and verbs, respectively. These results suggest that LSA may be better suited for the task of synonym detection in Croatian language and the smaller context (a window and especially a paragraph) gives better performance for LSA, while RI benefits more from a larger context (the entire document) which a reduced amount of noise into the distributions.

(G.Bharathi , D.Venkatesan, 2012) proposed an approach increases the performance of IR system by increasing the number of relevant documents retrieved. The proposed solutions done by apply set of preprocessing operation on the documents and then compute the term weight for each term in the document using term frequency-inverse document frequency model (tf-idf). It is utilized the term weight to preparing the document summary using the distinct terms whose frequencies are high after preprocessing of the documents. After that, the approach extract the semantic synonyms for the terms in the documents summary using Conservapedia thesauri and then clusters the document set by applying the K-means partitioning algorithm based on the semantically correlated. Retrieving the relevant documents are made by finding query and cluster similarity. The experiment showed that his method is promising and resulted in a significant increase in the number of relevant documents retrieved than the traditional tf-idf model alone used for document clustering by K-means.

CHAPTER THREE

3. RESEARCH METHODOLOGY:

3.1 Introduction:

The classic IR problem is to locate desired text documents using a search query consisting of a keyword express user's information need. Typically, the main interface of the IR system provides the user with an input field for the query. Then, all matching documents that have the query’s term are found and displayed back to the user. In our approach, we focus on query manipulation by using the query expansion technique to expand it by set of regional variation synonyms to retrieve all documents meet user's information need irrespective of user's dialect. Our method could be described as a pre-retrieval system that manipulates the query in a manner that guarantees a better performance.

This chapter divided to two sections. First, we explain the problem of the previous methods in section 3.2. Second, we describe in detail the proposed method to show how we could able to fill this research gab and reach the goal of research in section 3.3.

3.2 Previous Methods:

As we referred before in section 2.4, the early solutions addressed the problem of regional variations in IR systems. These solutions was classified to two methods based on the concept was used: Translation approaches or Distributional approaches.

(Wael,Nizar, 2012)(Houda et.al, 2014) (Mona et.al, 2014) were used the translation approaches concept to solve the dialect problem in IR. These methods, however, are suffers from a common problem known as out-of-vocabulary (OOV) which mean many words may not be listed in their entries and also deal with MSA corpus only, and any method has unique defect, the first way needs large training data and rule to translate DA-to-MSA. These requirements are considered obstacle to it due to less of available Arabic dialects resource. A more important drawback of the second approach, huge amounts of parallel text are required to infer translation relations for complex lemmas like idioms or domain specific terminology. And the drawback of the last method is lack of coverage to dialects because still no one machine readable dictionary cover all Arabic dialects, most of available dictionary deal with Egyptian because Arabic Egyptian media industry has traditionally played a dominant role in the Arab world.

Other solutions used the second approach.(Abdelali,2006)improve search results by combine clustering technique to build up index corresponded to each cluster, language model to restricting the semantics of the words used within a variation and use the LSA to find the Semantic proximity. (G.Bharathi , D.Venkatesan, 2012) extracts the semantic synonyms for a term in the documents by abstract the documents using the term frequency - inverse document frequency (tf-idf) to extract the height terms weight and then use the Conservapedia thesauri to find the synonyms for this terms, then clusters the document summary. Finding the relevant documents is made by compute the similarity between query and cluster.

The obvious shortcomings for the first solution, building index for each region and then make the query's access to appropriate index based on dialect was used to write a query and then find the Semantic proximity to retrieve a relevant documents is huge the IR performance. And the main limitation of the second method is using thesauri structure to summarize the documents then they inherited the drawbacks of auxiliary approaches (OOV) and also huge the IR performance due to finding query and cluster similarity at runtime.

In our proposed method we used distributional approaches to build auxiliary structure (see Figure 3.1). This is done by applied set of preprocessing operations and then combined terms-pair co-occurrence with LSA to extract synonyms of words from monolingual corpus to build a statistical dictionary to expand user's query. This to improve the relevant retrieving performance. The next sections illustrate the proposed method in details.

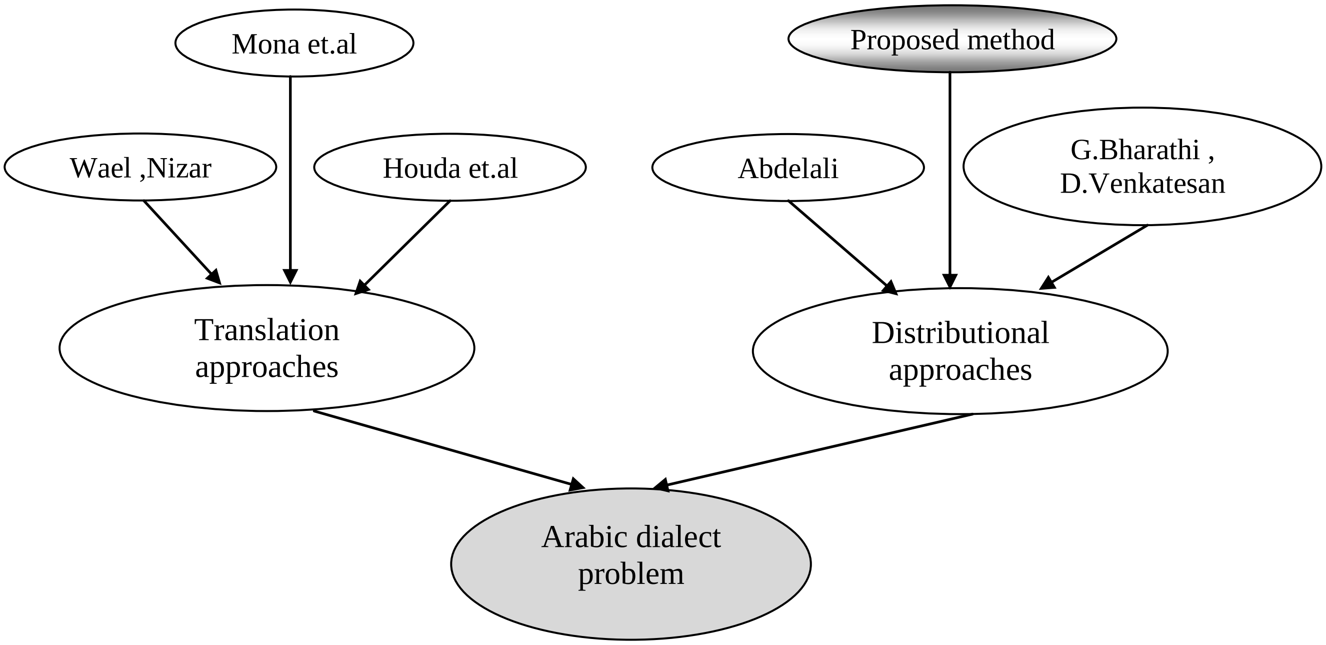


Figure 3.1: Research gab approaches

3.3 Proposed Method:

We proposed a method for building a statistical based dictionary from a monolingual corpus to expand the query using synonyms (regional variations) of the word in the other Arab world. This statistical based dictionary aim to improve the performance of Arabic IR system to assist users in finding the information they need regardless of their nationality. The proposed method is decomposed into three phases (see Figure 3.2) as follows:

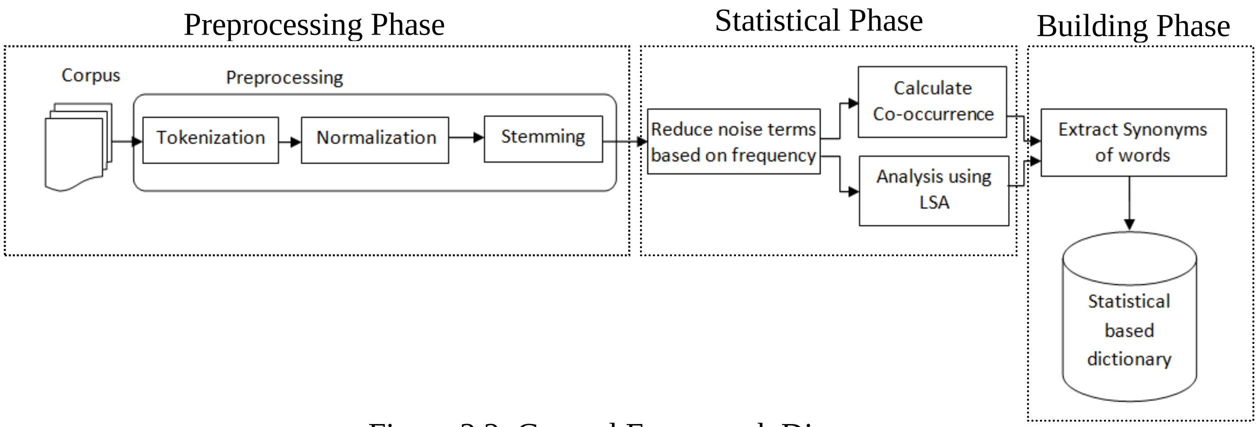


Figure 3.2: General Framework Diagram

Preprocessing Phase:

This phase contains two steps to prepare the data. The output of this phase will be directed as input to the next phase.

1. Collect a collection of documents manually to build a monolingual corpus contain different Arabic dialects to form a test data set and also construct the set of queries and relevance judgments.
2. Apply some of the preprocessing operations as follows:

   2.1. Tokenize the corpus into words.

   2.2. Normalize the words as follow:

   i. Remove honorific sign
   ii. Remove koranic annotation
   iii. Remove tatweel
   iv. Remove tashkeel
   v. Remove punctuation marks
   vi. Converteأ, إ ,آ to ا
   vii. Converteة to ه
   viii. Converteي, ئ to ى
   ix. Converteؤ to و

   2.3. Stem the words as follow:

   - For each word has more than 2 character remove the 'و' from beginning if found, for instance, والاقدام becomes الاقدام (In English, Foot) and check if the picked token is not stop words.
   - Remove 'ء' from end of all words to make شئ, شيء and شي same.
   - Remove the stop words.
   - If the length of the word`s is equal to four characters, then we don’t apply stemming and just remove the 'ال' and 'لل' from the beginning of the words if there are any. For example, الفل and للفل becomes فل (In English, Jasmine)
   - If the length of the word`s is more than four characters, then remove the 'ال' , 'لل' , 'بال', 'فال' and 'ل' from the beginning of the words if there are any.

- If the length of the word`s is more than five characters after apply the previous step then we should stem the word by remove the 'كم' ,'ها' ,'يا' , 'يه' ,'ان' ,'ون' , 'ين' and 'ات' from the end of the words.

Table3.1: Effect of Light10 Stemmer

Before Stemming	After Stemming	Meaning of the words before stemming	Meaning of the words after stemming
الدرج	درج	Stairs	Stairs
درجة	درج	Degree	
القصة	قص	Store	Cut
القص	قص	Cutting	
الآلة	ال	Machine	No meaning

The main goal from these levels of stemming is to maintain the meaning of the words as much as possible so as to prevent the meshing of words which affect their meaning.

According to the Table 3.1, we noticed that the first two words 'الدرج' and 'درجة' and the other set of words 'القصة' and 'القص' both with different meanings end up having the same meaning after applying light10 stemming. However some words will carry no meaning at all after being stemmed, such as 'الآلة' which will turn out to be 'ال'. 'ال' in Arabic is simply an article.

For this reason we assumed that all words with characters between 3 and 5 are representational lexical and semantic units (root) because the Arabic language is a derivational system based on a unit called the root (see in section 2.3.1.2).

Flow of stemming preprocessing operation was shown in Figure 3.3.

Statistical phase:

In this phase we done some of statistical operations as follow:

1. Reduce the noise in the corpus by filter out all words with height document frequency and re-write the corpus.
2. Calculate the co-occurrence between each terms-pair in the new corpus, this co-occurrence used as a link between documents.

3. Analyze the new corpus to extract the semantic similarity of the words of each other in the Arab world. This will do by using Latent Semantic Analysis (LSA) model (see in section 2.3.1.3.4) and apply the cosine similarity (see Equation 3.1)to find similarity between the word vectors.

$$sim(q,d) = \cos\theta = \frac{\vec{q}.\vec{d}}{|\vec{q}|.|\vec{d}|} \qquad (3.1)$$

Where:

- $\vec{\mathbf{q}}\,.\,\vec{\mathbf{d}}$is the inner product of the two vectors.
- $|\vec{\boldsymbol{q}}|$and$|\vec{\boldsymbol{d}}|$are the Euclidean length of **q** and **d**, respectively.

Because this approach is based on co-occurrence of the words, so maybe gathering words occur together permanently as synonyms and destroy some synonymous because not occur in the same context. To detract the first issue, we set a threshold to revise the semantic space extracted using the LSA model. And the second issue solved by the next phase.

Building phase:

In this phase we used the outcome of phase two to build the statistical dictionary by use the subsequent steps:

1. For each term "A" get co-occurrence words "{B_1, B_2, B_3 …}", if "A" has high weight.
2. Select "B_i" as related word to "A" if this term-pair co-occurrence has high similarity in LSA semantic space.
3. For each related word "B_i" to term "A" gets all word that co-occurs with it "{C_1, C_2, C_3 …}".
4. From term-pair co-occurrence "B-C" get the high similar term-pair "B-C" using the LSA space.
5. Select the words "C_i" as synonyms to "A" if it get by more than or equals to half of related terms and has high weight.

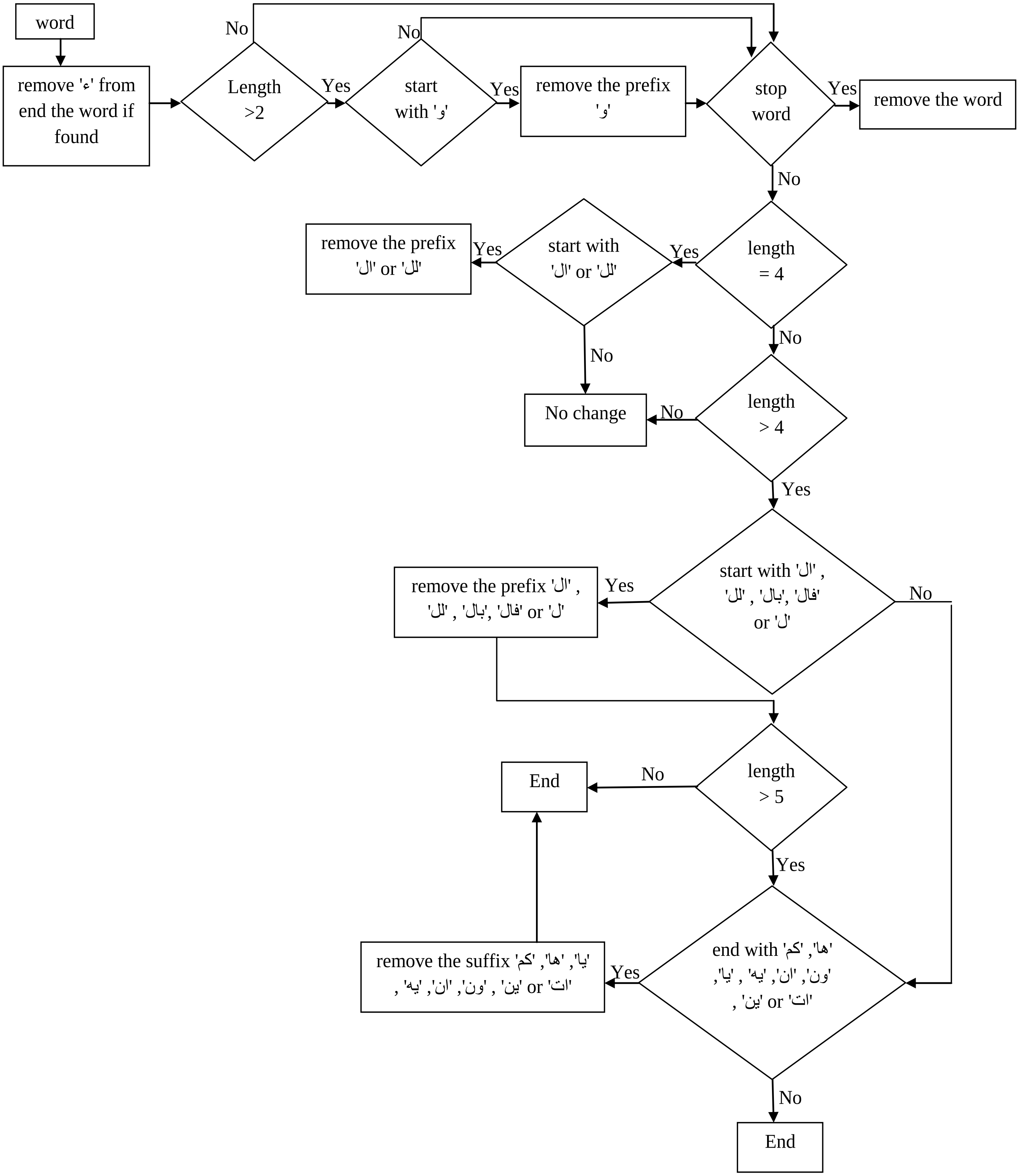

word
remove 'ء' from end the word if found
Length >2
No
Yes
start with 'و'
No
Yes
remove the prefix 'و'
stop word
Yes
remove the word
No
length = 4
Yes
start with 'ال' or 'لل'
Yes
remove the prefix 'ال' or 'لل'
No
No change
No
length > 4
No
Yes
start with 'ال' , 'لل' , 'بال', 'فال' or 'ل'
Yes
remove the prefix 'ال' , 'لل' , 'بال', 'فال' or 'ل'
No
length > 5
No
End
Yes
end with 'كم' , 'ها' ,'يا' , 'يه', 'ان', 'ون' , 'ين' or 'ات'
Yes
remove the suffix 'كم', 'ها' , 'يا' , 'يه', 'ان', 'ون' , 'ين' or 'ات'
No
End

Figure 3.3: Levels of Stemming

When the statistical dictionary is built, we will build the index. When a user enters a query's term in the search field, we apply the same preprocessing operation that was applied to build the statistical dictionary. After that, the resulting term is searched of in the statistical dictionary along with its synonyms which will be found with the resulting term in the dictionary to expand the query – see Figure 3.4.

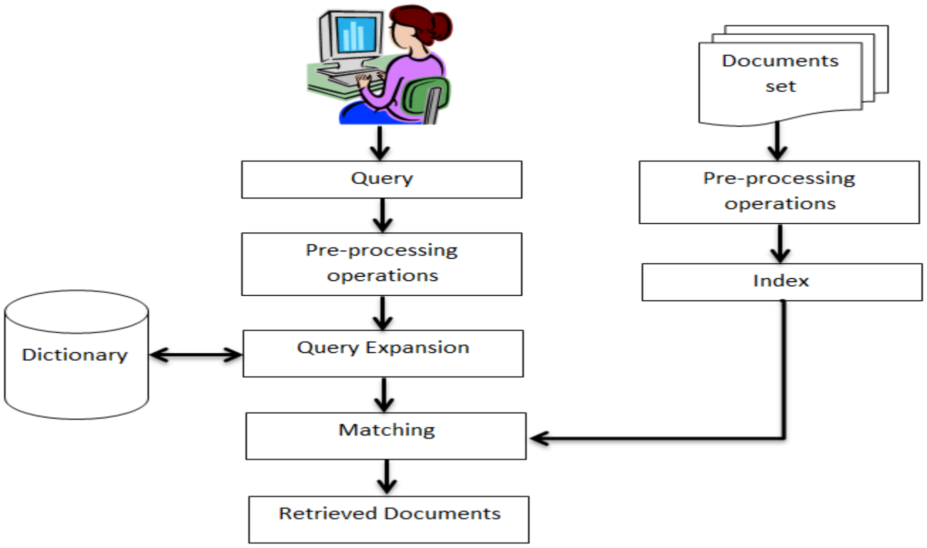


Figure 3.4: Proposed Method Retrieval Tasks

Now to understand this method we will look at the following example. Suppose the user wants to find information about "eye glasses", and he searched for his query using the Moroccan dialect which calls it "نواظر". In the corpus there are many documents that contain this user's information need - see Appendix B -but they cannot be retrieved because the query term would not be found in the relevant documents. To solve this issue, our method concerns that the documents which talk about the same subject contain the same keywords. Taking this assumption into account, we get all the words that co-occur with the term "نواظر" and select from it those words that have high similarity with it in the semantic space - see Table 3.2. For each word that co-occurs with the term "نواظر", we applied the same previous step to extract the highly similar words that co-occur with it - see Table 3.3, 3.4, 3.5, 3.6and 3.7 below.

Table 3.2: high similar words that co-occur with "نواظر" term

Term	Related term
نواظر	عدسا
	رويه
	عدسه
	طبيب
	نظر

Table 3.3: high similar words that co-occur with "عدسا"

Term	Related term
عدسا	طرق
	كشمه
	رويه
	عدسه
	طبيب
	نظر
	نواظر
	بصر
	نظار
	نضار
	جلاكوم
	مبصر

Table 3.4: high similar words that co-occur with "عدسه"

Term	Related term
عدسه	عدسا
	طرق
	كشمه
	رويه
	طبيب
	نظر
	نواظر
	بصر
	نظار
	نضار
	جلاكوم
	مبصر

Table 3.5: high similar words that co-occur with "رويه"

Term	Related term
رويه	طرق
	قطه
	سنوري
	عدسا
	كشمه
	عدسه
	طبيب
	نظر
	بزون
	ثدي
	بسين
	نواظر
	هر
	بصر
	نظار
	كديس
	نضار
	جلاكوم
	قطوف
	مبصر

Table 3.4: high similar words that co-occur with "طبيب"

Term	Related term
طبيب	عدسا
	رويه
	عدسه
	اطبا
	دختر
	نظر
	خسته
	سبيطار
	نواظر
	بصر
	نظار
	مستشفى
	باطن
	سبيتار
	عياد
	اسبتال

Table 3.5: high similar words that co-occur with "نظر"

Term	Related term
نظر	عدسا
	رويه
	عدسه
	طبيب
	عدلى
	بارك
	توثيق
	بسمله
	شاهد
	مشهود
	عرف
	قبض
	اصفا
	مرمز
	برمجي
	نواظر
	بصر
	نضار
	جلاكوم
	عقد
	قاضى
	قانون
	شخصى

Then, from these words related to the term "نواظر" we will see that there is a term, "نظارة" for instance, that is related to more than half the terms related to "نواظر", and therefore we ensure that "نظارة" is a synonym for "نواظر" but only if it has a high weight in the corpus. From the words in the tables above, we will find that only the following terms: "كشمه" ,"قطه","بزون","بسين","نواظر","هر","نظار","كديس","نضار", "جلاكوم", "قطوف", "مبصر", "دختر", "خسته", "سبيطار", "مستشفى", "سبيتار", "اسبتال", and "اصفا" have a high weight based on our corpus, and others have a low weight because they are repeated in many documents. Now since we ensured that the following words meet the first condition (to have a high weight) we will move to the second condition (being related to more than half the related words). According to Table 3.8 below, which shows the number of times for each word is retrieved by the related terms, we notice that the words "كشمه","نظار", "نضار", "جلاكوم" and "مبصر"

meet the second condition. We now know that these words meet both the necessary conditions, therefore, we add them as synonyms of the word "نواظر" to the dictionary to expand the query.

Table 3.6: Number of Times that Word Retrieved by the Related Terms

Term	Times
كشمه	3
قطه	1
بزونه	1
بسينه	1
الهر	1
النظاره	4
كديسه	1
نضاره	4
جلاكوما	4
قطوف	1
مبصره	3
الدختر	1
الخسته	1
السبيطار	1
مستشفى	1
سبيتار	1
الاسبتاليه	1
واصفات	1

CHAPTER FOUR

4. EXPERIMENT AND EVALUATION:

4.1 Introduction

This thesis challenges to improve the performance of Arabic IR system by developing a method able to identify the Arabic regional variation synonyms accurately in monolingual corpora. This method aims to assist users in finding the information they need apart from any dialect that was used to query formulation.

In particular, the chapter will evaluate our approach, which was shown in the previous chapter. This evaluation aims to show the significant impact of using these proposed approaches on Arabic IR effectiveness and determine if they provide a significant improvement over some well-established baseline systems.

This chapter as follows: Section 4.2 define the test collection, section 4.3 explain the tool, Section 4.4 define the baseline methods, Section 4.5 give explanation about the experiments procedures and section 4.6 is devoted to experiments and results.

4.2 Test Collection:

Test collection is used to evaluate the IR systems in laboratory-based evaluation experimentation. To measure the IR effectiveness in the standard way, we need a test collection consisting of three things: a document collection (data set) which contains textual data only, a test suite of information needs; expressible as queries (query set) and a set of relevance judgments. In the next subsection we discuss these components that are used in this research.

4.2.1 Document Set:

In this experiment we use an Arabic monolingual dataset collected manually from different online sites using Google search engine.

Table 4.1: Statistics for the data set. computed without stemming

Description	Numbers
Number of documents	245
Number of words	102603
Number of distinct words	13170

4.2.2 Query Set:

We are choice a set of 45 queries from different topics (see Appendix C). There are a number of the query was written in Dialects Arabic language and the other in MSA Arabic language. Table 4.2 below show the some sample from the query set.

Table 4.2: Example queries from the created query set

#	Query	Region	Equivalent in English
Q01	الشفرة	MSA	Code
Q02	المستورة	Algeria	Corn
Q03	المزنبلة او البزبور	Gulf and Yemian	Faucet
Q04	اجزخانة	Sudan and Egypt	Pharmacy
Q05	الاورطة	Iraq	Carpet
Q06	الشنطة	Sudan, Libya and Libnan	Bag
Q07	النواظر	Jazzier and Morocco	Glasses
Q08	البندورة	Levant and Tunisia	Tomato
Q09	بطاقة الاحوال المدنية	-	Identity Card
Q10	الانسالة	-	Robot

4.2.3 Relevance Judgments:

In our experiments, we used the binary relevance judgment to evaluate the system performance. That is, a document is assumed to be either relevant (i.e., useful) or non-relevant (i.e., not useful) for each query-document pair. We used the binary relevance due to one aim of this research as mentioned in chapter one which is improving the performance of the Arabic IR system by improving the recall of IR system and not discard the precision. In this case, it is not recommending to use the multi-grade relevance.

4.3 Retrieval System:

For the retrieval system, we used the Lucene IR system (version) to processing, indexing and retrieve the documents and Apache Tomcat Software which allow to browse the result as a search engine. The Lucene IR system is a free, open source IR software library, originally written in Java. Lucene is suitable for any application that requires full text indexing and searching capability, Lucene has been widely recognized for its utility in the implementation of Internet search engines and local, single-site searching. As an example, Twitter is using Lucene for its real time search (https://en..org/wiki/Lucene).

4.4 Baseline Methods:

In this section, we show two baseline methods which was used to evaluate the proposed solution.

1. A baseline method (b) done by applying the preprocessing operations on the words in the documents and locate all documents into index and search for them using the Lucene IR system.
2. A baseline method (bLSA), all extracted word from the documents was manipulated using the preprocessing operations and then analyze the data set by the latent semantic analysis model (LSA) to extract the candidate's synonyms for each word. The environment setup by set the LSA dimension=50 and revise the candidates by use threshold similarity greater than 0.6. Afterward, write the word with candidates' synonyms that meet the threshold condition and write it as dictionary form. After that, index the documents and search for it using the Lucene IR system. When the user writes his query the system finds the synonym(s) of each word in the dictionary and expand the query.

4.5 Experiment Procedures:

As previously described in this research, the study seeks to assess if we using the proposed method in the Arabic IR system can have a significant effect on the retrieval performance. To reach this objective, we did three experiments based on six methods. These methods come from applied two type of stemmer: Light10 and proposed stemmer (see preprocessing phase in section 3.3) on the baseline methods (see in section 4.4) and the proposed method. Table 4.3 show the Abbreviation of the methods which was used in the experiments.

The aim from applied different stemmer to notice how the proposed stemmer aid in improve the performance of IR system behind the proposed solution(see statistical and building phase in section 3.3).

Table 4.3: Abbreviation of Baseline Methods and Proposed Method

Method	Abbreviation	Method by Light10 Stemmer	Method by Proposed Stemmer
1th baseline method	B	b_{light10}	$b_{\text{prostemmer}}$
2th baseline method	bLSA	$\text{bLSA}_{\text{light10}}$	$\text{bLSA}_{\text{prostemmer}}$
Proposed method	Co-LSA	$\text{Co-LSA}_{\text{light10}}$	$\text{Co-LSA}_{\text{prostemmer}}$

4.6 Experiments and results:

In this section, we present some experiments to evaluate the effectiveness of the proposed expansion method. These methods are evaluated in the average recall (Avg-R),average precision (Avg-P) and average F-measure (Avg-F).

There are three experiments was done to evaluate our method. The first experiment is an evaluation of proposed method and baseline methods with the counterpart after applying the two type of stemmer. The second experiment compares the two baseline methods. Afterward, the third experiment is an evaluation of the proposed method with the1th baseline method (b).

Experiment 1:

This experiment tries to find if we are using the proposed stemmer in Arabic IR can improve the retrieval performance. This was done by compared the proposed method and the baseline methods($\text{Co-LSA}_{\text{Prostemmer}}$, $b_{\text{Prostemmer}}$, $\text{bLSA}_{\text{Prostemmer}}$) with the counterpart(Co-

$LSA_{Light10}$, $b_{Light10}$, $bLSA_{Light10}$)when we use the proposed stemmer in the previous chapter and light10 stemmer, respectively.

Results:

The following tables, Table 4.4, Table 4.5 and Table 4.6compare the result of $b_{Light10}$ method with $b_{Prostemmer}$ method, $bLSA_{Light10}$method with $bLSA_{Prostemmer}$ method and Co-$LSA_{Light10}$ method with Co-$LSA_{Prostemmer}$ method, respectively. Figure 4.1, Figure 4.2 and Figure 4.3 Visualize the same results obtained.

Table 4.4: Shows the results of $b_{Light10}$ compared to the $b_{Prostemmer}$

Method	avg-R	avg-P	avg-F
$b_{Light10}$	0.32	0.78	0.36
$b_{Prostemmer}$	0.33	0.93	0.39

Table 4.5: Shows the results of $bLSA_{Light10}$compared to the $bLSA_{Prostemmer}$

Method	avg-R	avg-P	avg-F
$bLSA_{Light10}$	0.87	0.60	0.64
$bLSA_{Prostemmer}$	0.93	0.65	0.71

Table 4.6: Shows the results of Co-$LSA_{Light10}$ compared to the Co-$LSA_{Prostemmer}$

Method	avg-R	avg-P	avg-F
Co-$LSA_{Light10}$	0.74	0.68	0.65
Co-$LSA_{Prostemmer}$	0.89	0.86	0.83

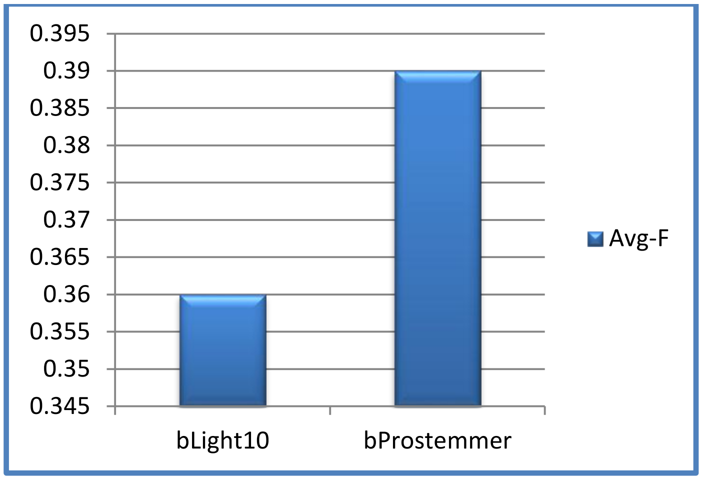


Figure 4.1: Retrieval effectiveness of $b_{Light10}$compared to the $b_{Prostemmer}$, in terms of average F-measure

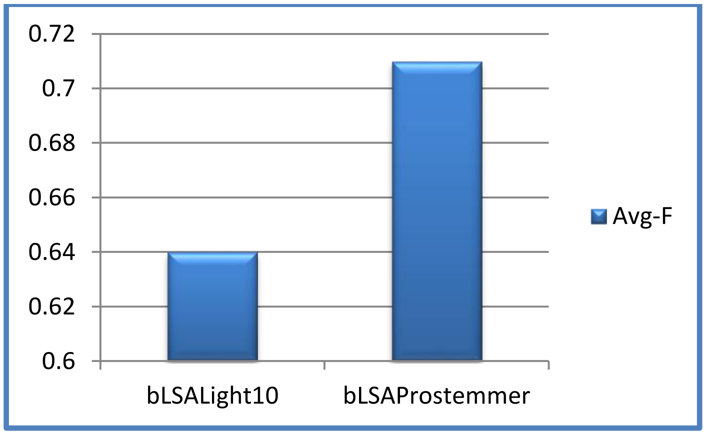


Figure 4.2: Retrieval effectiveness of $bLSA_{Light10}$compared to the $bLSA_{Prostemmer}$

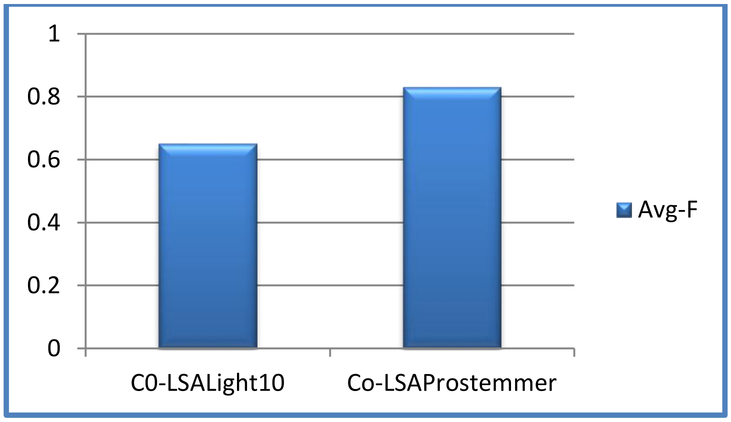


Figure 4.3: Retrieval effectiveness of Co-$LSA_{Light10}$compared to the Co-$Lsa_{Prostemmer}$

Discussion:

In the Figures 4.1, 4.2 and 4.3 above, we noted a very substantial benefit from using the proposed stemmer, with statistically significant differences between $b_{light10}$ and $b_{Prostemmer}$, $bLSA_{light10}$ and $bLSA_{Prostemmer}$, and between Co-$LSA_{light10}$ and Co-$LSA_{Prostemmer}$ (all at p-value<0.01).

Experiment2:

The main objective of this experiment to decide if the latent semantic analysis is able to find synonyms and improve the effectiveness of the IR system (b). And determine if this improves in the effectiveness of bLSA method can have a significant effect on retrieval performance.

This experiment contains two result sections: The first, result after stemmed the data by light10 and the second, the result after stemmed the data set by the proposed stemmer.

Results of Light10 Stemmer:

Experimental results for $b_{Light10}$ and $bLSA_{Light10}$ are shown in Table 4.7 and Figure 4.4.

Table 4.7: Shows the results of $b_{Light10}$compared to the $bLSA_{light10}$

Method	avg-R	avg-P	avg-F
$b_{Light10}$	0.32	0.78	0.36
$bLSA_{Light10}$	0.87	0.60	0.64

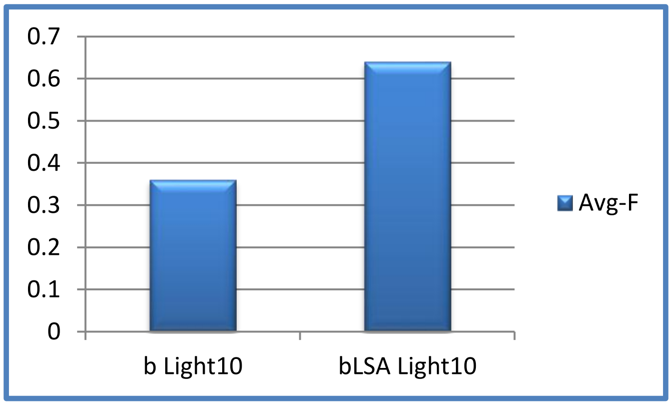


Figure 4.4: Retrieval Effectiveness of $b_{Light10}$compared to the $bLSA_{light10}$

Results of Proposed Stemmer:

The result of the experiment is shown in Table 4.8 and Figure 4.5.

Table 4.8: Shows the results of $b_{Prostemmer}$, compared to the $bLSA_{Prostemmer}$

Method	avg-R	avg-P	avg-F
$b_{Prostemmer}$	0.33	0.93	0.39
$bLSA_{Prostemmer}$	0.93	0.65	0.71

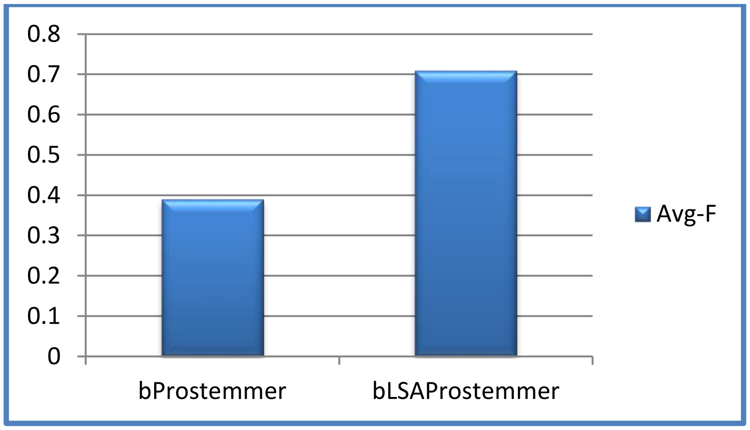


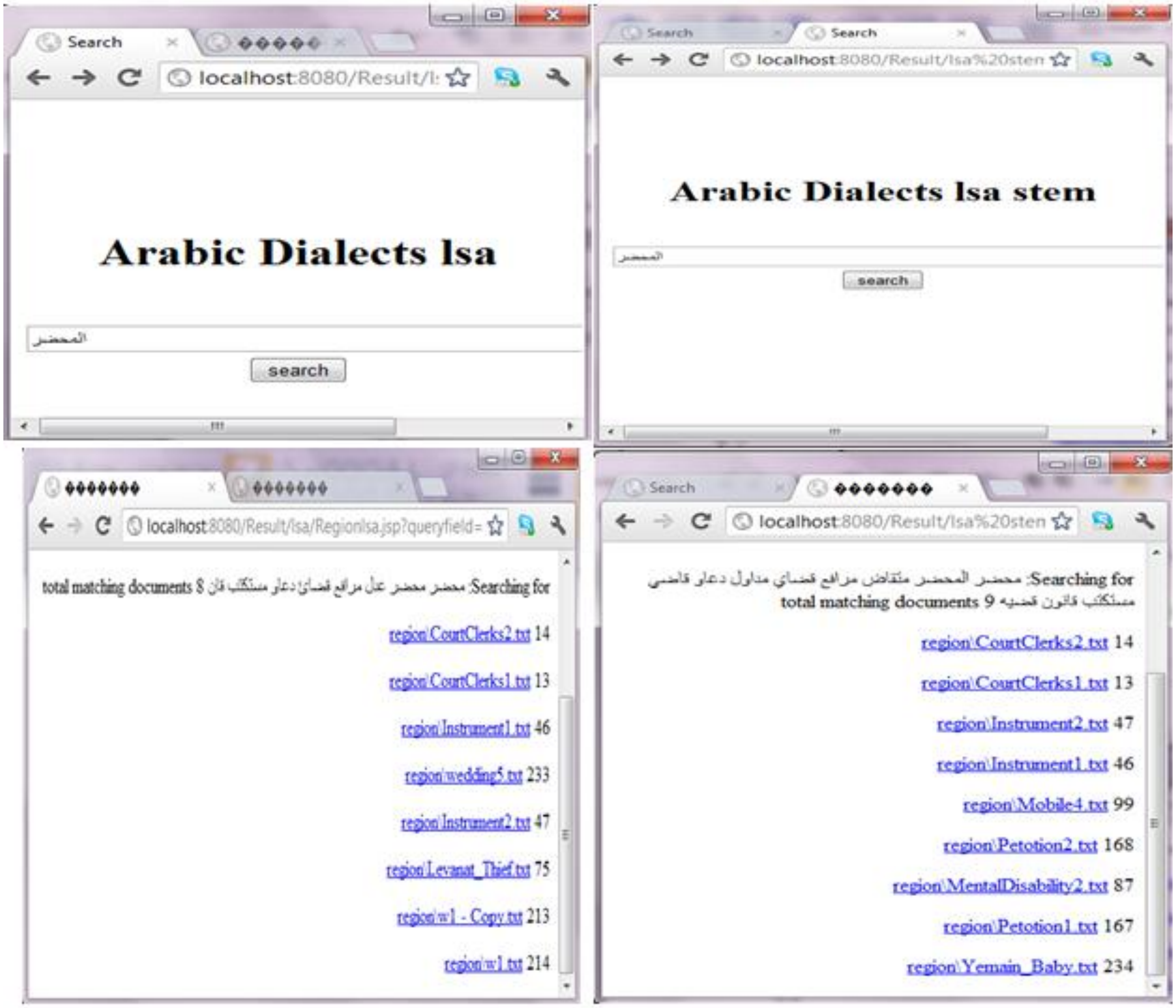
Figure 4.5: Retrieval Effectiveness of $b_{Prostemmer}$compared to the $bLSA_{Prostemmer}$

Discussion:

We noticed the bLSA method improve the Arabic IR retrieval markedly. This improvement occurs as a result of the expansion of the query by the candidate synonyms and then executes the expanded query rather than execute of that entrance query by the user directly. The $bLSA_{Light10}$ and $bLSA_{Prostemmer}$ produce results that are statistically significantly better than $b_{Light10}$and $b_{Prostemmer}$ (t-test, p-value <1.68667E-06) and (t-test, p-value <1.4843E-07).

In spite of the results presented in Figure4.4 and Figure 4.5 indicate the retrieval effectiveness of bLSA method outperforms the b method . We found that improvement was not able to achieve the research challenge. The thesis aims to improve the performance of Arabic IR system by expanding the query by Arabic regional variation synonyms.

The bLSA method based mainly on the LSA model which gathering words occur together permanently as synonyms due to being based on co-occurrence of the words. This method increases the recall of IR system which was appearing in Table 4.7 and Table 4.8through expanding the query by high similar related terms in the semantic space. But this may cause to retrieve irrelevant documents containing these related terms and which leads to lower precision (see Table 4.7 and Table 4.8) and it also leads to intent drifting– see Figure 4.6 to notice that.


Search
localhost:8080/Result/l:
Search
localhost:8080/Result/lsa%20sten
localhost:8080/Result/lsa/Regionlsa.jsp?queryfield=
Search
localhost:8080/Result/lsa%20sten
search
search

Arabic Dialects lsa

المحضر

Arabic Dialects lsa stem

المحضر

Searching for: محضر محضر عل مرافع قضاي دعاو مستكتب قان total matching documents 8

region\CourtClerks2.txt 14

region\CourtClerks1.txt 13

region\Instrument1.txt 46

region\wedding5.txt 233

region\Instrument2.txt 47

region\Levanat_Thief.txt 75

region\w1 - Copy.txt 213

region\w1.txt 214

Searching for: محضر المحضر متقاض مرافع قضاي مداول دعاو قاضي مستكتب قانون قضيه total matching documents 9

region\CourtClerks2.txt 14

region\CourtClerks1.txt 13

region\Instrument2.txt 47

region\Instrument1.txt 46

region\Mobile4.txt 99

region\Petotion2.txt 168

region\MentalDisability2.txt 87

region\Petotion1.txt 167

region\Yemain_Baby.txt 234

Figure 4.6: Result of Submitted "المحضر" query (in English, Court Clerk) in bLSA. the left colum show $\text{bLSA}_{\text{Light10}}$ and the right show $\text{bLSA}_{\text{ProStemmer}}$

Experiment 3:

This experiment aimed to test the impact of the proposed method (Co-LSA) in the effectiveness of the Arabic IR system. It also showed how the proposed method outperforms the baseline. And then, determine if this improves in the effectiveness of the proposed method (Co-LSA) can have a significant effect on retrieval performance.

This experiment contains two results section: The first, result after stemmed the data by light10,the second, the result after stemmed the data set by the proposed stemmer.

Results of Light10 Stemmer:

The result of this experiment is shown in Table 4.9 and Figure 4.7.

Table 4.9: Shows the results of $b_{Light10}$ compared to the Co-$LSA_{Light10}$

Method	avg-R	avg-P	avg-F
$b_{Light10}$	0.32	0.78	0.36
Co-$LSA_{Light10}$	0.74	0.68	0.65

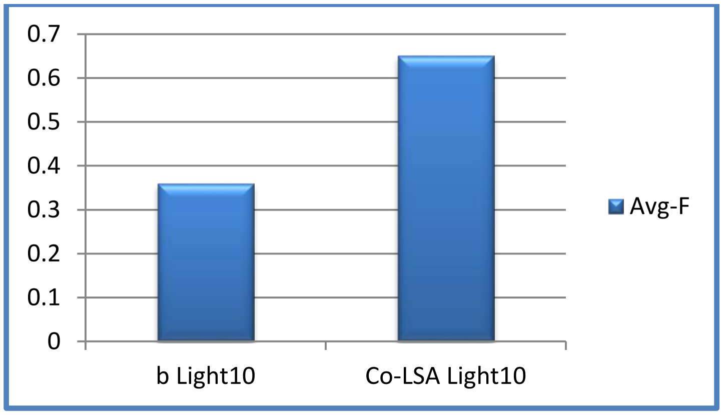


Figure 4.7: Retrieval Effectiveness of $b_{Light10}$ compared to the Co-$LSA_{Light10}$

Results of Proposed Stemmer:

Table 4.10 compares the baseline with our proposed method. Figure 4.8 illustrates this comparison using the F-measure.

Table 4.10: Shows the results of $b_{Prostemmer}$ compared to the Co-$LSA_{Prostemmer}$

Method	avg-R	avg-P	avg-F
$b_{Prostemmer}$	0.33	0.93	0.39
Co-$LSA_{Prostemmer}$	0.89	0.86	0.83

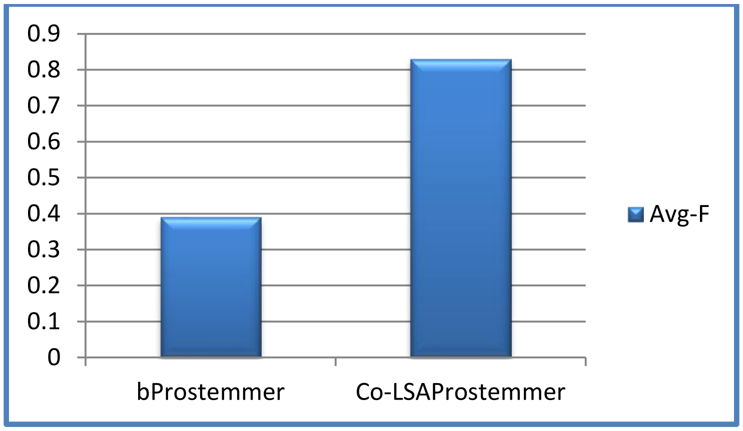


Figure 4.8: Retrieval Effectiveness of $b_{Prostemmer}$ compared to the Co-$LSA_{Prostemmer}$

Discussion:

As we observed in Table 4.9 and 4.10, they found a loss in average precision in Co-LSA method compared to the b method due to the obvious improvement in the recall caused by the proposed method. But also as can be seen in Figure 4.7 and 4.8, Comparing b method with the proposed method shows that our method is considerably more effective in Arabic IR. This difference is statistically significant (p<5.25706E-09) in light10 case and (p<5.43594E-16)in the case of proposed stemmer, using the Student t-test significance measure.

On the test data set, the results presented in this research show that proposed method (Co-$LSA_{Prostemmer}$) is able to solve successfully the research problem and it achieves it in high performance level.

CHAPTER FIVE

5. CONCLUSION AND FUTURE WORK:

5.1 Conclusion:

In this research, we developed synonyms discovery approach for the dialect problem in Arabic IR based on LSA and co-occurrence statistics. We built and evaluated the method through the corpus that gathered manually using Google search engine. The results indicated that the proposed solution could outperform the traditional IR system (1st baseline method) by improving search relevance significantly.

5.2 Limitation:

Although the proposed solution increases the effectiveness of the results significantly, but it suffer from limitations. The shortcomings appeared when dealing with phrases such as "قاعدة البيانات" (in English, Database)which represents one meaning, in spite of that any word has its own meaning carried when it shows up individually. In this situation there are two problems:

1. If the constituent words of the phrases are common and frequent in the dataset it will be given a low weight and thus cleared and will not be finding the synonyms.
2. If given high weight as a result of rarity, we need to find synonyms for any word consisting the phrase separately. This leads to a turn down in the precision which is subsequently decrease the effectiveness of IR systems.

5.3 Future Work:

For future work we intend to address the following:

1. Building standard test collection for evaluating Arabic IR system that dealing with regional variations.
2. Find a way to determine the phrases and manipulate (consider) them as a single word.
3. Handling the Homonymous.

References:

Abdelali, A., *Improving Arabic Information Retrieval Using Local Variations in Modern Standard Arabic*. 2006, New Mexico Institute of Mining and Technology.

Ali, M.M., *Mixed-Language Arabic-English Information Retrieval*. 2013

Berry, M.W., Z. Drmac, and E.R. Jessup, *Matrices, vector spaces, and information retrieval.* SIAM review, 1999. **41**(2): p. 335-362.

CD Manning, H Schütze, Foundations of statistical natural language processing, 1999.

Darwish, K. and W. Magdy, *Arabic Information Retrieval.* Foundations and Trends in Information Retrieval, 2014. 7(4): p. 239-342.

Du, S., *A Linear Algebraic Approach to Information Retrieval.* 2012.

Elmasri, R. and S. Navathe, *Fundamentals of Database Systems sixth Edition Pearson Education.* 2011.

G.BHARATHI and D.VENKATESAN, *Improving information retrieval using document clusters and semantic synonym extraction*,Journal of Theoretical and Applied wikipedia Information Technology, February 2012. Vol. 36 No.2.

Ghassan Kanaan, Riyad al-Shalabi and Majdi Sawalha, *Improving Arabic Information Retrieval Systems Using Part of Speech Tagging*, information technology journal, 2005.4(1): p. 32-37

González, R.B., et al., *Index Compression for Information Retrieval Systems*. 2008.

Hassan Sajjad, Kareem Darwish and Yonatan Belinkov, *Translating Dialectal Arabic to English*,Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1–6,Sofia, Bulgaria, August 4-9 2013. c2013 Association for Computational Linguistics

Houda Bouamor, Nizar Habash and Kemal Oflazer, *A Multidialectal Parallel Corpus of Arabic*, ELRA, May-2014, pages 1240--1245

https://en..org/wiki/Lucene

Inkpen, D., *Information Retrieval on the Internet*. 2006.

Khalid Almeman and Mark Lee, *Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words*. 2013 IEEE

KHOJA, S. & GARSIDE, R. Stemming arabic text. Lancaster, UK, Computing Department, Lancaster University,1999.

Kraaij, W., *Variations on language modeling for information retrieval.* 2004.

Manning, C.D., P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Vol. 1. 2008: Cambridge university press Cambridge.

Mladen Karan, Jan Snajder and Bojana Dalbelo, *Distributional Semantics Approach to Detecting Synonyms in Croatian Language*,2012.

Mona Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Pradeep Dasigi, Heba Elfardyy, Ramy Eskandery, Nizar Habashy, Abdelati Hawwari and Wael Salloum, *Tharwa: A Large Scale Dialectal Arabic - Standard Arabic - English Lexicon*,2014.

Musaid Saleh Al Tayyar,*Arabic Information Retrieval System based on Morphological Analysis*, PHD thesis , July 2000

Mustafa, M., H. AbdAlla, and H. Suleman, *Current Approaches in Arabic IR: A Survey, in Digital Libraries: Universal and Ubiquitous Access to Information*. 2008, Springer. p. 406-407.

Nie, J. Y.,*Cross-language information retrieval*, Synthesis Lectures on Human Language Technologies ,2010.

Ruge, G. *Automatic detection of thesaurus relations for information retrieval applications*. in Foundations of Computer Science. 1997. Springer.

Sanderson, M. and W.B. Croft, *The history of information retrieval research.* Proceedings of the IEEE, 2012. 100(Special Centennial Issue): p. 1444-1451

Shaalan, K., S. Al-Sheikh, and F. Oroumchian, *Query expansion based-on similarity of terms for improving Arabic information retrieval*, in *Intelligent Information Processing VI*. 2012, Springer. p. 167-176.

Singhal, A., *Modern information retrieval: A brief overview.* IEEE Data Eng. Bull., 2001. **24**(4): p. 35-43.

Wael Salloum and Nizar Habash, *A Dialectal to Standard Arabic Machine Translation System*,Proceedings of COLING 2012: Demonstration Papers, pages 385–392, COLING 2012, Mumbai, December 2012.

Webber, W.E., *Measurement in Information Retrieval Evaluation.* 2010.

Wei, X., et al. *Search with synonyms: problems and solutions*. in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. 2010. Association for Computational Linguistics.

Appendix A:

System Design:

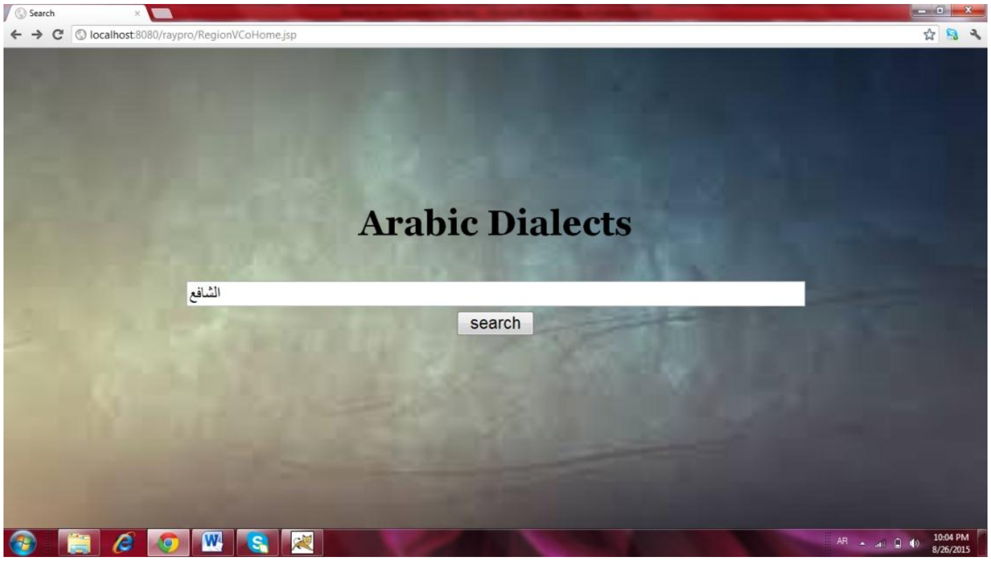
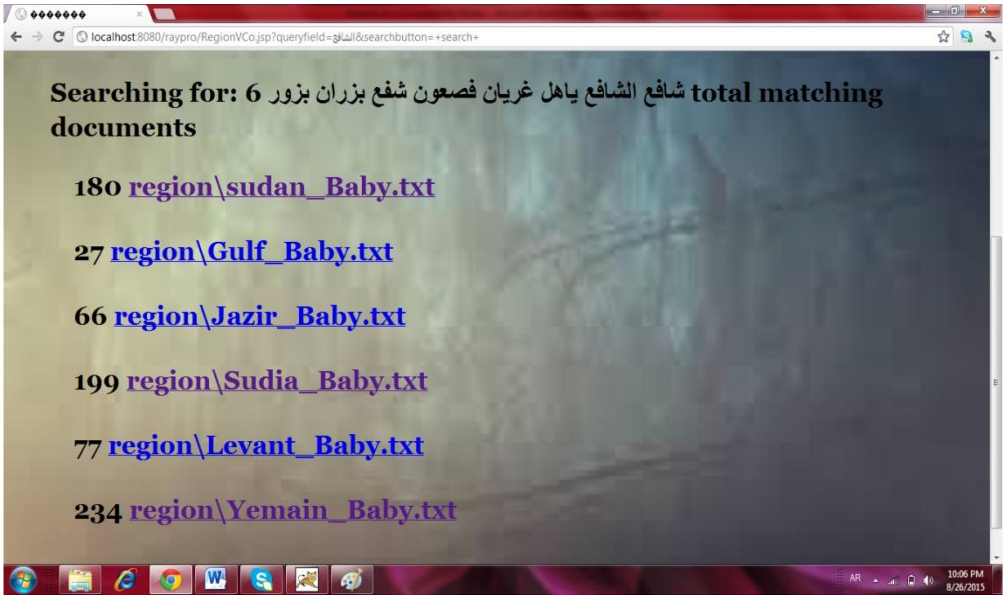


Figure 5.1: Main Interface


◆◆◆◆◆◆◆
localhost:8080/raypro/RegionVCo.jsp?queryfield=الشافع&searchbutton=+search+
AR
10:06 PM
8/26/2015

Searching for: شافع الشافع ياهل غريان فصعون شفع بزران بزور 6 total matching documents

180 region\sudan_Baby.txt

27 region\Gulf_Baby.txt

66 region\Jazir_Baby.txt

199 region\Sudia_Baby.txt

77 region\Levant_Baby.txt

234 region\Yemain_Baby.txt

Figure 5.2: Output Interface

Appendix B:

Document 1:

ما أنواع عدسات الكشمة المتوفرة و ما مميزات كل منها؟

يوجد الان أنواع كثيرة من عدسات الكشمة المتوفرة مع تقدم التكنولوجيا . في الماضي ،كانت عدسات الكشمة تصنع بشكل حصري من الزجاج . اليوم يتم صناعة الكشمة من عدسات مصنوعة من البلاستيك المتطور بشكل عالي . تتميز هذه العدسات الجديدة بخفة الوزن ، غير قابلة للكسر بسهولة مثل العدسات الزجاجية ، وأكثر مقاومة للخدش من العدسات الزجاجية ، اضافة إلى ذلك تحتوي على طبقة اضافية للحماية من الأشعة فوق البنفسجية الضارة لتحسين الرؤية .

عدسات متعددة الكربونات

عدسات تري فكس

عدسات لا كروية

عدسة متلونة بالضوء

Document 2:

التحرر من النواظر :

سواء كنت تستخدمين النواظر منذ سنوات أو اكتشفت حاجتك إلى تصحيح النظر ، فإن العدسات اللاصقة خيار مثالي للتمتع برؤية واضحة ودقيقة.

لا بأس في وضع العدسات في عينيك طوال اليوم ، طالما وافق طبيب العيون على ذلك . أو ربما تفضلين التبديل بين العدسات اللاصقة و النواظر ؟ أروع مزايا العدسات اللاصقة هي كونها ملائمة مهما كان أسلوب حياتك.

لماذا تستخدم العدسات اللاصقة بدلاً من النواظر ؟

تمنحك العدسات اللاصقة الحرية لتري وتعيشي الحياة كما تريدين ، دون أن تعيقك في أنشطتك.

فيما يلي بعض الأسباب التي تجعل من العدسة اللاصقة خيار أفضل من النواظر :

تتميز العدسات بخفة الوزن .

على عكس النواظر فإنها لا ترتفع أو تنخفض أثناء الحركة ولا تسقط أو تنزلق .

ليس عليك القلق من الكسر.

تتحرك العدسات مع عينيك لتمنحك مجالاً كاملاً للرؤية ، مما يعني إمكانية رؤية كل شي من ركن عينك .

لا تسبب انعكاس الضوء ولا تجمع الرذاذ أو تكون بخار – مهما كانت حالة الطقس.

وجهك بدون النواظر يبدو طبيعي أكثر.

من الصعب فقدانها أو كسرها ، ويمكن استبدالها بسهولة أكبر وتكلفة أقل.

يمكنك استعمال النواظر الشمسية على الموضة ودون وصفة طبية.

يمكنك استعمالها في جميع الأنشطة والمغامرات، مثل التزلج على المنحدرات الثلجية. كما أنها لا تعيق ارتداء الخوذات الواقية.

Document 3:

يعاني كثير من الناس من مشاكل في العيون و البصر ، فيكون إحدى الحلول ارتداء النظارات و ذلك لتصحيح الرؤية أو لحماية العين . و هي ضرورية للحفاظ على صحة العين و خاصة إذا أقرها طبيب العيون. كما أن هناك النظارات الشمسية و هي إحدى أنواع النظارات التي تسمح برؤية أفضل في ضوء النهار الساطع، ويمكن أن تحمي من الضرر الناتج من المستويات العالية من الأشعة .

كما اصبح الناس الذين يرتدون النظارات الطبية و الشمسية يهتمون بها كجزء من الموضة ، فهناك اختيارات متعددة لتختار نوع الاطار و العدسات الملائمة لك و التي تواكب آخر صيحات الموضة.

كما يمكن ان ترتدي العدسة اللاصقة بدلا من النظارات و لكن إذا كانت تسبب لك تهيج في العيون فاختر النظارات الطبية المناسبة لك التي تعطي وجهك منظرا جديد و جميل .

Document 4:

كيف تقوم بتنظيف عدسات المبصرة بشكل صحيح ؟

المبصرة الطبية عرضة لتراكم الاوساخ الناتجة من يديك و الوجه و الرموش و تخلق طبقة لزجة من الأتربة و الدهون . و هذا يؤثر علي الرؤيه من المبصرة و يجعل زجاج المبصرة ضبابي قد تكون أنسب و أسرع طريقة لكي تحسن الرؤيه هي مسح عدسة المبصرة بطرف التي شيرت . و لكنك لا تنتبه إلي أن طرفه محمل بجزئيات الغبار التي يمكن أن تؤثر علي عدسة المبصرة ، تحتاج إلي ايجاد طرق جيدة لتنظيف عدسة المبصرة . و الخبر السار ، هنا الذي نعرضه عليك ، يمكنك تنظيف المبصرة بدون الحاجة إلي شراء منظف مكلف ، فقط كمية صغيرة من الصابون السائل كافية للقيام بهذا الغرض .

بالإضافة إلي ذلك فإن جمعية المبصرات الأمريكية توصي بتنظيف المبصرة يومياً و يفضل في الصباح .

لذلك يجب عليك تنظيف المبصرة بصورة منتظمة ، لتحسين الرؤية من خلالها بالإضافة إلي أنها تجعل مظهرك يبدو أنيق

خطوات التنظيف :

يمكنك شطف مبصرتك الطبية تحت الماء الجاري الدافئ .

ثم وضع قطرة من الصابون السائل علي كل عدسة .

البدء بفرك زجاج كل عدسة بأصابعك ، حتي يحدث الصابون رغوة ثم شطفها بالماء .

Document 5:

النضارة هي أداة توضع فوق العينين لكي تساعد الأشخاص ضعيفي البصر على القراءة والرؤية بوضوح أكثر .

تتكون النضارة من إطار لاحتواء العدسات التي يمكن أن تكون مصنوعه من الزجاج أو البلاستيك و العدسة قد تكون عدسة مقعرة أو عدسة محدبة

النضارة الطبية تعتبر وسيلة لإصلاح مشاكل البصر في العين مثل مد البصر أو الحسر (قصر النظر) أو اللابؤرية وتستخدم أيضا لعلاج بعض حالات الحول أو الجلاكوما

العدسات المفضلة للنضارة الطبية هي العدسة الشفافة ، ولكن قد ينصح باستخدام العدسات الملونة في حالات حساسية العين

أفضل طريقة للعناية بها هي غسل النضارة بالماء الدافئ والصابون أو أى سائل منظف ثم شطفها بالماء ثم التنشيف برفق بمادة قطنية ، وذلك لأن مسحها في حالة الجفاف يشوش عمل العدسات . كما أن العرق قد يضر أكثر من الماء لاحتوائه على أملاح تسبب التآكل

Appendix C:

#	Query	Region	Equivalent in English
Q01	الشيك	MSA	Check
Q02	الشفرة	MSA	Code
Q03	المترجم	MSA	Compiler
Q04	المحضر	MSA	Court Clerks
Q05	الشافع	Sudan	Baby
Q06	المشه	Morocco	Cat
Q07	الترب	Egypt	Cemetery
Q08	المستورة	Jazzier	Corn
Q09	المزنبلة او البزبور	Gulf and Yemian	Faucet
Q10	اجزخانة	Sudan and Egypt	Pharmacy
Q11	الاورطة	Iraq	Carpet
Q12	الشنطة	Sudan, Libya and Libnan	Bag
Q13	حوائج	Morocco and Libya	Clothes
Q14	الكرهبة	Libya and Tunisia	Car
Q15	القرلو	Jazzier and Libya	Cockroach
Q16	النواظر	Jazzier and Morocco	Glasses
Q17	المناقش	Jazzier	Earring
Q18	البنكة	Gulf and Iraq	Fan
Q19	الكندرة	Palestine and Jordan	Shoes
Q20	البسكليت	Hejaz	Bicycle
Q21	الكوفيرته	Jazzier	Blanket
Q22	البندورة	Levant and Tunisia	Tomato
Q23	الخسته خان	Iraq	Hospital
Q24	كوجينه	Tunisia and Libya	Kitchen
Q25	بطاقة الاحوال المدنية	-	Identity Card
Q26	الوثيقة العدلية	-	Instrument
Q27	القاش	sudan	Belt
Q28	مطب	MSA	Bump

Q29	الغارو	Morocco	Cigarette
Q30	معطف	MSA	Coat
Q31	الايسكريم	MSA	Ice cream
Q32	فستق العبيد	Iraq	Peanut
Q33	الخدوش	Jordan	Cheeks
Q34	السيمافرو	Libya	Traffic Light
Q35	الرقد	Yemain	Stairs
Q36	الصغيو	Oman	Chick
Q37	الجوال	Gulf	Mobile
Q38	البرمجة كائنية المنحنى	-	Object Oriented Programming
Q39	التخلف العقلي	-	Mental Disability
Q40	واصفات البيانات	-	Metadata
Q41	اللص	MSA	Thief
Q42	الكحته	Syria	Scrooge
Q43	العريضة	-	Petitions
Q44	الانسالة	-	Robot
Q45	النكاح	-	Wedding