جامعة السودان للعلوم والتكنولوجيا
كلية علوم الحاسوب وتقانة المعلومات
قسم الحاسوب ونظم المعلومات

بناء قاموس ثنائي اللغة بإستخدام تقنيات التنقيب على الويب

Building Bilingual Dictionary Using Web Mining Techniques

مشروع مقدم كأحد متطلبات الحصول على بكالوريوس الشرف في الحاسوب ونظم المعلومات

أكتوبر 2015

بسم الله الرحمن الرحيم

جامعة السودان للعلوم والتكنولوجيا
كلية علوم الحاسوب و تقانة المعلومات
قسم الحاسوب ونظم المعلومات

بناء قاموس ثنائي اللغة بإستخدام تقنيات التنقيب على الويب

Building Bilingual Dictionary Using Web Mining Techniques

إعداد الطالبات:
1. **شذى عوض علي**
2. **فرحة أحمد المصطفي إبراهيم**
3. **وفاء حمد قسم الله**

مشروع مقدم كأحد متطلبات الحصول على بكالوريوس الشرف في الحاسوب ونظم المعلومات

إشراف الدكتور: هشام عبدالله
توقيع الدكتورالمشرف:

التاريخ:15/ أكتوبر/2015

الآية

قال تعالي :

﴿اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ (1) خَلَقَ الْإِنْسَانَ مِنْ عَلَقٍ (2) اقْرَأْ وَرَبُّكَ الْأَكْرَمُ (3) الَّذِي عَلَّمَ بِالْقَلَمِ (4) عَلَّمَ الْإِنْسَانَ مَا لَمْ يَعْلَمْ (5) ﴾

صدق الله العظيم

سورة العلق (الأية 1-5).

الحمَّد

الحـمد لله رب العالمين،أعطى اللسان،وعَلَّم البيان، وخلق الإنسان، فبإي ألآء ربكما تكذبان .. لك الحمد يا من هو للحمد أهل، أهل الثناء والمجد، أحقُ ماقال العبد و كلنا لك عبد، لك الحمد مادعوناك إلا حسنُ ظنٍ بك و ما رجوناك إلا ثقة فيك، و ماخفناك إلا تصديقاً بوعدك و وعيدك لك الحمد حمداً كثيراً طيباً مباركاً فيه، وصلى الله على سيدنا محمد خاتم الأنبياء والمرسلين أجمعين بشر وأنذر و وعد و أوعد، أنقذ الله به البشر من الضلالة وهدى الناس الى صراط المستقيم، صراط الله الذي له مافي السموات ومافي الأرض الا الى الله تصير الأمور.

نتقدم ببحثنا هذا الي زملائنا الطلاب والي كل من يجمعنا بهم رباط العلم ,والذي نامل ان ينال القبول ,وان يكون إضافة حقيقية للجهد المبذول في سبيل تنمية البلاد ونسأل الله ان يجعله عملاّ مباركاّ متقبلاّ , وان يكون في ميزان حسنات كل من ساهم في إخراجه في هذة الصورة ونسال الله ان يديم نعمته علينا وان يحفظ وطننا من كل كيد ومن كل شر وان يهدينا سواء السبيل ونسأل الله عز و جل أن يوفقنا و يجعل النجاح حليفنا.

الإهداء

إلي من تقف كل كلماتي وحروفي عاجزة في حضرتها..إلي أغلي ما وهبني الرحمن وأثمن ما جاد به ربي علي..إلي من تمنح للنجاحات معني..وللدورب ضياء..إلي سر بسمتي والسبب وراء سعادتي إلي أطيب وأنقي وأحن قلب..إلي منبع كل ماهو جميل...

أمي

إلي من زرع وسقي ورعي..إلي قدوتي ومصدر عزي وفخري وتاج علي رأسي..إلي من علمني كيف أحب العلم وأحترم من يعلمني..إلي من أطمح أن أكون الشخص الذي أردني أن أكون ووقف بجانبي وساندني..إلي من يكفيني فخراً أنني إبنته...

أبي

إلى من حبهم يجري في عروقي ويلهج بذكراهم فؤادي ..ودعموني شتي أنواع الدعم...

إخوني وأخواتي

إلى من سرنا سوياً ونحن نشق الطريق معاً نحو النجاح والإبداع.. إلى من تكاتفنا يداً بيد ونحن نقطف زهرة تعلمنا...

أصدقائي وزملائي وزميلاتي

إلى من علمونا حروفاً من ذهب وكلمات من درر وعبارات من أسمى وأجلى عبارات في العلم..إلي الشموع التي ظلت تضئ بسخاء طوال الاربعة أعوام..وأبت إلي أن ترقي بنا خطوة بخطوة نحو النجاح...

أساتذتي الأجلاء

وأخيراً وليس آخراً....

إلي من أفاض علينا بعلمه الغزيز ووجه خطواتنا وصححها بكل رحابة صدر..إلي من تعجز ألسنتنا عن شكره علي كل ما قدمه لنا من نصح وإرشاد ومتابعة حتي وضحت الفكرة وأكتملت الصورة ... أستاذي المشرف

د / هشام عبدالله

الشكر و التقدير

الشكر لله من قبل ومن بعد في الظاهر والباطن ذا المنة والفضل....

يقول الرسول عليه أفضل الصلاة وأتم التسليم: (لا يشكر الله مـن لا يـشكر الناس) .

لابد لنا ونحن نخطو خطواتنا الأخيرة في الحياة الجامعية من وقفة تعود إلى أعوام قضيناها في رحاب الجامعة مع أساتذتنا الكرام الذين قدموا لنا الكثير باذلين بذلك جهودا كبيرة في بناء جيل الغد لتبعث الأمة من جديد و قبل أن نمضي نقدم أسمى آيات الشكر والامتنان والتقدير والمحبة إلى الذين حملوا أقدس رسالة في الحياة إلى الذين مهدوا لنا طريق العلم والمعرفة إلي جميع أساتذتنا الأفاضل ونخص بالشكر والتقدير المشرف الدكتور :**هشام عبدالله** لتفضله بالإشراف على هذه الدراسة ومنحها الوقت والجهد والنصح رغم مسئولياته المتعدده فلم يبخل علينا بالتوجيه والإرشاد طوال فترة البحث. و كما نخص بالشكر كل الشكر من ساعدتنا و ساندتنا لإكمال هذه الدراسه الأستاذة الفاضلة: **إبتهال مصطفي**.

المستخلص

الويب هو أكبر مصدر للبيانات في العالم، إنهحقل متعدد التخصصات يتضمن إستخراج البيانات، و التعلم الآلي، ومعالجة اللغات الطبيعية، والإحصائيات، وقواعد البيانات، وإسترجاع المعلومات، والوسائط المتعددة، وغيرها.

المشروع يحل مشكلةعدم تغطية المفردات (out of vocabulary) مثل المصطلحات التي صيغت حديثاً والمصطلحات الفنية وأسماء الأعلاموغيرها التيلا توجد في القواميس ثنائية اللغة (عربي- إنجليزي) لذلك تم إستخدام تقنية تنقيب الويب المستخدمةلحل هذه المشكلة، .يتم البحث عن المصطلح هو مجرد في القاموس ثنائي اللغة إذا تم العثور على معناها يتم إسترجاعه وإذا لم يوجد معناها في القاموس يتم تنقيب الويب للحصول على العديد من المعاني المحتملة للمصطلح المدخل وتم إستخدام إختبار (Chi-Square) لإيجاد أقرب معنى للمصطلح. وتم إدخال 180 مصطلح وكانت نسبة المصطلحات الصحيحة 80% ونسبة المصطلحات الخطأ 20% ونعزي ذلك لمشاكل المعالجة المبدئية للغة العربية.

Abstract

Web is the largest source of data in the world , it is an interdisciplinary field that includes data mining, machine learning, natural language processing, statistics, databases, information retrieval, multimedia, and others. The project solves the problem out of vocabulary like newly coined terms such as technical terms, proper names and others that do not exist in bilingual dictionaries (Arabic-English) .result of that it used the Web mining technology to solve this problem. A search for an abstract term is performed in a bilingual dictionary, if found its meaning it will be retrieved ,if there is no meaning in the dictionary, the system will use the Web mining for finding many of the potential of the term entered meanings, and(Chi-Square test)is used to create a closer sense of the term. will insert 180 term , the percentage of correct terminology 80% , the percentage of error terminology 20% , the reason of that it comes from the problems of the pre-process the Arabic language

شرح المصطلحات

الاختصار	المصطلح	شرح المصطلح
WWW	World wide web	الويب أو الشبكة العنكبوتية العالمية
	web mining	التنقيب في الويب
	hyper link	الوصلات التشعبية
HTML	Hyper Text Markup Language	لغة ترميز النصوص التشعبية
IR	Information retrieval	عملية إسترجاع المعلومات
IE	Information extraction	عملية إستخراج المعلومات
	Web content mining	تنقيب محتوي الويب
	web structure mining	تنقيب بنية الويب
	Web usage mining	تنقيب إستخدام الويب
CLIR	Cross-lingual Information Retrieval	إسترجاع المعلومات ثنائية اللغة
	machine translation	الترجمة الآلية
	Parallel corpora	المجاميع المتوازية
	bilingual dictionaries	القواميس ثنائية اللغة
OOV	out of vocabulary	عدم تغطية المفردات
	Transliteration	النقل الحرفي(التعريب)
	Phonetic mapping	الخرائط اللفظية
	Corpus	المجاميع
	Exploiting Web Corpora	استغلال شبكة المجاميع
	Object Oriented languages	لغات كائنية التوجه

XML	EXtendable Markup Language	لغة الترميز القابلة للتوسيع
KDT	Knowledge Discovery in Texts	إكتشاف المعرفة في النصوص
	Document Structure	هيكلة الوثيقة
	PageRank	تصنيف الصفحات
	HITS	الزيارات
	Named Entity	اسم الكائن
ME	Maximum Entrop	نموذج الحد الاقصى الانتروبي
NP	Noun Phrase	الجملة الأسمية
EM	Expectation Maximization	خوارزمية اقصى توقع
TF-IDF vectors	Term Frequency-Inverse Document frequency	تردد المصطلح العكسي لتكرار الوثيقة
DOM	Document Object Model	نموذج كائن المستند
	anchor-text	مرساة النصوص
	Chi-Square	إختبار إحصائي
	Stopwords	كلمات غير ذات معنى
SAX	Simple API for XML	واجهة برمجة التطبيقات المبسطة XML

فهرس الأشكال

فهرس الجداول

فهرس المحتويات

الباب الأول

المقدمة(الإطار العام)

1.1 المقدمة

التنقيب في الويب هو إستخدام تقنيات إستخراج البيانات للإكتشاف التلقائي وإستخراج المعلومات المفيده من الويب،قد تم الاستفادة من تنقيب فى الويب في حل مشكلة عدم تغطية المفردات (oov) فى القواميس ثنائية اللغة (عربي/إنجليزي) . يتم ذلك بعدد من إجراءات التنقيب عن المصطلح وإجراء إختبارإحصائي للحصول على أقرب ترجمة للمصطلح.

2.1 مشكلة البحث

عدم تغطية المفردات (out of vocabulary) يسبب عدة مشاكل خاصة فى مجال إسترجاع المعلومات و الترجمة الآلية، هذه المشكلة تنشأ من حقيقة أن بعض المصطلحات مثل مصطلحات صيغت حديثا والمصطلحات الفنية، والكلمات المركبة، و أسماء الأعلام أثناء الترجمة قد لا توجد في القاموس ثنائي اللغة.

3.1 أهمية البحث

أصبح الويب مصدراًغنياً بالمعلومات ويستمر التوسع في الحجم والتعقيد .أصبح هنالك إلتباس في إيجاد معاني المصطلحات لذا كان لا بد من السعي نحو إيجاد طريقة تُمكن المستخدم من إيجاد معاني المصطلحات حتي التي ظهرت حديثاًولا توجد في القاموس ثنائي اللغة.

4.1 أهداف البحث

- بناء نظام يبحث عن المصطلحات التي صيغت حديثا والمصطلحات الفنية، والكلمات المركبة، أسماء الأعلام، الغير موجودة في القاموس ثنائي اللغة.
- تنقيب الويب و ذلكبعرض مجموعة من نتائج البحث و توضيح أكبر تكرارللمصطلح المدخلة .
- تطبيق إختبار(Chi-Square) للتمكّن من الحصول على أقرب نتيجة محتملة لمعنى المصطلح الذى يراد البحث عنها .

5.1 منهجية البحث

في المشروع قيد الدراسة ستتم معالجة مشكلة عدم تغطية المفردات (out of vocabulary) وذلك بإستخدام تقنية تنقيب الويب (web mining) وتطبيق إختبار(Chi-Square) .

6.1 حدود البحث

هذا المشروع يختص بإيجاد معاني الكلمات العربية، يتم تجريد المصطلح المدخل، تتم عملية البحث عن معناه بتنقيب محتوى الويب، يكون التنقيب في أول صفحة من الويب، لا تتضمن عملية البحث تنقيب بنية الويب و تنقيب إستخدام الويب، نتائج معاني المصطلحات تكون باللغة الإنجليزية .

7.1 هيكلية البحث

يتضمن البحث بالإضافة إلي هذا الباب الابواب التالية :
الباب الثاني : يحتوي على فصلين و همانبذة عامة عن تنقيب الويبوالترجمة .

الباب الثالث: يحتوي على فصلبين هما الدراسات السابقة والتقنيات و الادوات المستخدمة في البحث.
الباب الرابع : يحتوي على تطبيق النظام الذي تم بناءه .
الباب الخامس : النتائج والتوصيات والخاتمة
والمراجع والملاحق.

البـــاب الثانـي

الإطار النظري

الفصل الأول: التنقيب في الويب

الفصل الثاني: الترجمة

الفصل الأول

التنقيب في الويب

1.1.2 مقدمة

يتناول هذا الفصل نبذه عن التنقيب في الويب وتصنيفاته وعلاقته بالمجالات الاخرى.

2.1.2 التنقيب في الويب

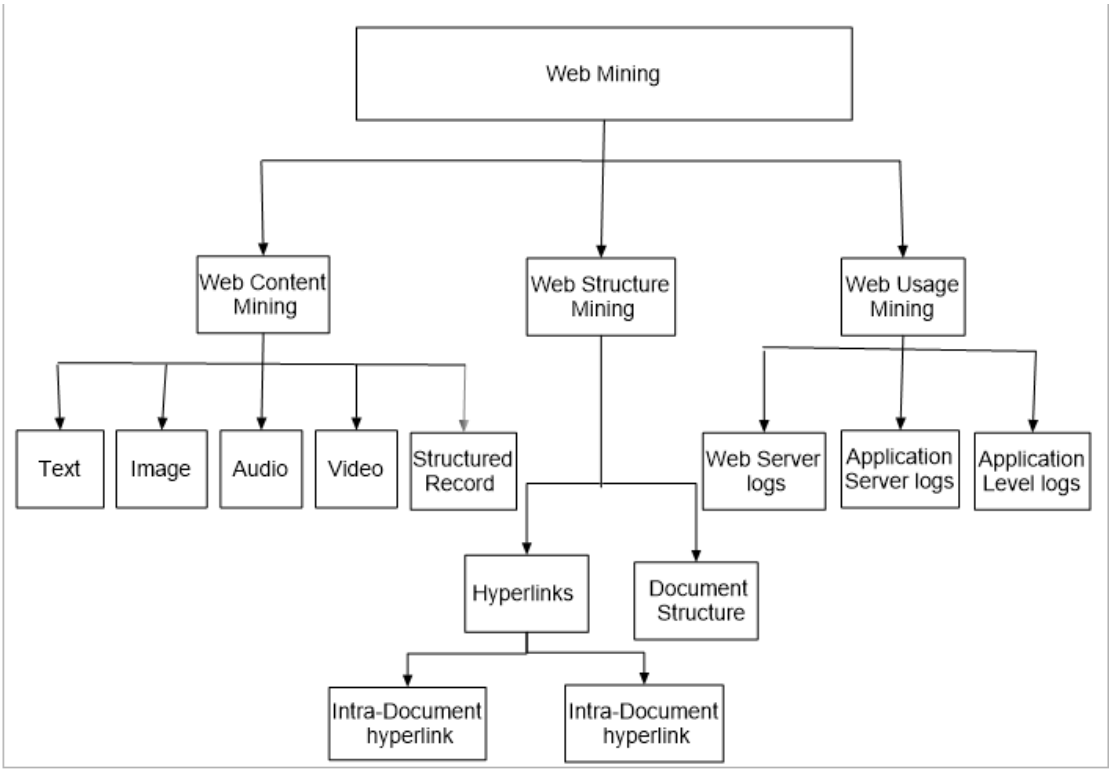
التنقيب في الويب هو إستخدام تقنيات إستخراج البيانات للإكتشاف التلقائي وإستخراج المعلومات المفيده من الويب لجعلها أكثر فائدة وأكثر ربحية وزيادة كفاءة تفاعلنا مع الويب.

يعتبر تنقيب الويب مختلف لأن الشبكة عبارة عن مجموعة ضخمة من الوثائق, الويب ديناميكي جدا لأنه يحدثإنشاء بإستمرار لصفحات جديدة, والتحدي يعني تطوير خوارزميات تنقيب الويب الجديدة والتكيف مع خوارزميات تنقيب البيانات التقليدية لإستغلال الوصلات التشعبية (hyper link) وأنماط الوصول وقواعد المعرفه على الويب. ومن تطبيقاته التجاره الالكترونيه, وإسترجاع المعلومات,وإدارة الشبكه .[1]

يتم تحليل تنقيب الويب لمهام فرعية وتتضمن إيجاد الموارد وهي عملية إسترجاع البيانات على الانترنت بإتصال (online) أو دون إتصال (offline) من المصادر النصية المتاحه على الويب مثل النشرات الإخبارية الإلكترونية ووكالة الأنباء الإلكترونية ومحتويات النص من وثائق الHTML بعد إزالة علامات(HTML Tags) وايضاً من مهام التحليل إختيار المعلومات والمعالجة المسبقة (pre-process) وهو نوع من عملية تحويل إسترجاع البيانات الأصليه إلي عملية إسترجاع المعلومات (IR) والمعالجة المسبقة تهدف للحصول على التمثيل المطلوب مثل العبارات الموجودة في المدونة، وأيضاً من مهام التحليل والتعميم التحقق من الصحة وهي عملية إكتشاف المعلومات أو المعرفة من الويب.[2]

1.2.1.2 تصنيفات التنقيب في الويب

يتم تصنيف التنقيب في الويب إلي ثلاثة أنواع وهي تنقيب محتوىالويب، وتنقيب إستخدام الويب، و تنقيب بنية الويب، ويتم التعرف عليها بصورة أكثر تفصيل في هذا الفصل. الشكل أدناه يوضح هذه التصنيفات:



الشكل 1.2 تصنيفات تنقيب الويب

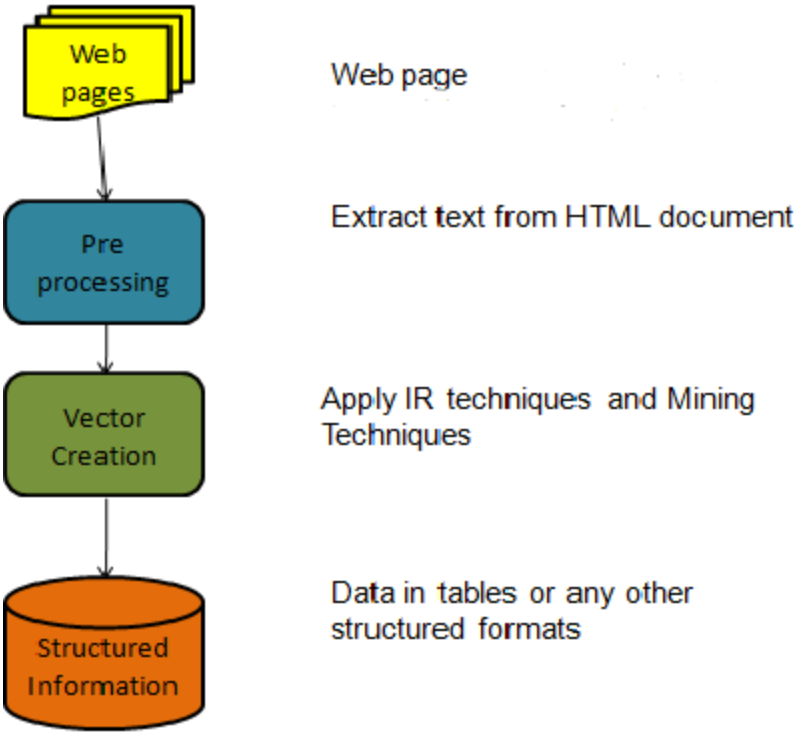
1. تنقيب محتوى الويب (Web Content Mining)

يهدف تنقيب محتوى الويب إلي إكتشاف معلومات مفيدة من محتويات الويب والوثائق، ومحتويات الويب يمكن أن تشمل عدةأنواع من البيانات مثل النصوص والصور والصوت والفيديو وكذلك الإرتباطات التشعبية، ويطلق على البحوث التي أجريت مؤخراً على تنقيب أنواع متعددة من البيانات بمصطلح تنقيب بيانات الوسائط المتعددة, أي أن تنقيب بيانات الوسائط المتعددة مثيل لتنقيب محتوى الويب.[2]

تتكون بيانات محتوى الويب من بيانات غير مهيكلة مثل النصوص الحرة، وبيانات شبه مهيكلة مثل مستندات الHTML، وبيانات أكثر تنظيماً مثل البيانات في الجداول أو قاعدة بيانات إنشاء صفحات الHTML. ومع ذلك، فإن الكثير من بيانات محتوى الويب هو نص البيانات غير المهيكلة . ويطلق على الأبحاث حول تطبيق تقنيات إستخراج بيانات النص غير المهيكل بإكتشاف المعرفة من النصوص (KDT) أو إستخراج بيانات النص أوتنقيب النص، أي أنتنقيب النص مثيل لتنقيب محتوى الويب.[3]

تنقيب محتوىالويب له وجهات نظر مختلفة من حيث إسترجاع المعلومات (IR), وقواعد البيانات (DB)، إن هدف تنقيب محتوى الويب من وجهة نظر ال IR هوالمساعدة في إيجاد المعلومات أو تصفيتها للمستخدمين، أما من وجهة نظر الDB فهي تهدف لنمذجة البيانات على الويب .[4]

الشكل أدناة يوضح سيرعمليات تنقيب محتوىالويب :



الشكل 2.2 سير عمليات تنقيب محتوىالويب

خطوات تنقيب محتوىالويب

عند إستخراج معلومات محتوى الويب باستخدام التنقيب في الويب هناك أربع خطوات أساسية هي :

- الجمع - جلب المحتوى من الويب.
- التحليل - إستخراج البيانات القابلة للإستخدام من البيانات المنسقة (PDF, HTML).
- تحليل الtokenize – تصنيف، تصفية،فرز،... الخ
- الإنتاج - تحويل نتائج التحليل إلي شيء مفيد (تقرير،فهرس البحث،... الخ).[5]

2. تنقيب إستخدام الويب (Web Usage Mining)

تنقيب إستخدام الويب هيالعملية التي من خلالها يمكننا التعرف على أنماط التصفح من خلال تحليل سلوك المستخدم، أنه يركز على التقنيات التي يمكن إستخدامها للتنبؤ بسلوك المستخدم أثناء تفاعلة مع الويب، وهو يستخدم البيانات الثانوية على الويب، يشمل هذا النشاط الإكتشاف التلقائي لأنماط وصول المستخدم من واحد أو أكثر من خوادم الشبكة،ومن خلاله يمكننا التأكد عما يبحث عنه المستخدمين على الويب، الفائدة منه العثور على سلوك المستخدم التي يمكن أن تساعد في إعادة تنظيم مواقع الويب بحيث يمكن تقديم نوع عالي من الخدمة.[6]

مراحل تنقيب إستخدام الويب

1- جمع البيانات : هو إكتشاف البيانات المخفية وإستخدام الأنماط التي يمكن أن تساعد مديري الويب لتحسين الإدارة والأداء والسيطرة على خوادم الويب.
2- تجهيز البيانات : إختيار بيانات مفيدة ومهمة في مرحلة ما قبل المعالجة، قد تم إختيار البيانات بانواعها لتوليد النماذج العنقودية للعثور على وصول المستخدم على الويب، إزالة البيانات غير ذات الصلة هي الخطوة الأولى في هذه المهمه، ثم فهرسة بيانات الوصول.
3- تجميع البيانات : يستخدم أسلوب التجميع على نطاق واسع في المشاريع المختلفة من قبل الباحثين لإيجاد أنماط الاستخدام أو التشكيلات الجانبية للمستخدم، خوارزميات التجميع تصبح أكثر طريقة للتنقيب في المواقع (لوصف إجراءات المستخدم) وصفحات الويب.
4- اكتشاف النمط والتحليل : إستخدام إكتشاف النمط وتحليل المعلومات ذات الصلة والمفيده تُمكّن من التنبؤ على أساس تحليل البيانات والرسم البياني.

وتشمل إستخدامات بيانات الويب بيانات وصول خادم الويب، وتعريف المستخدمين، والإستعلامات، وبيانات التسجيل، والكوكيز، والبيانات المرجعية، وأي بيانات أخرى ناتجه من التفاعل[7] .

أقسام تنقيب إستخدام الويب

1- الوصلات التشعبية (Hyperlinks) :
هي الوحده الهيكلية التي تربط جزءمن صفحة الويب بموقع آخر، سواء داخل نفس صفحة الويب أو على صفحة ويب مختلفة .

2- هيكلة الوثيقة (Document Structure):
هي تنظَّم محتوى صفحة الويب في شكل شجرة تنظيمية، إستناداًعلى مختلف علامات HTML وXML داخل الصفحة .[7]

3. تنقيب بنية الويب (Web Structure Mining)

يستند تنقيب بنية الويب على الهياكل ذات الصلة، مع أو بدون وصف الروابط،نموذج سلسلة (Markov) يمكن إستخدامها لتصنيف صفحات الويب ومفيده لتوليد المعلومات مثل التشابه والعلاقة بين مختلف المواقع، هدفه توليد ملخص منظم حول مواقع وصفحات الويب، أنه يستخدم بنية الشجرة لتحليل ووصف HTMLأو XML.

وقد أُقترحت بعض الخوارزميات لنموذج طوبولوجيا الويب مثل الزيارات (HITS)، وتصنيف الصفحات (PageRank) وتحسين الزيارات عن طريق إضافة محتوى المعلومات إلي هيكل الروابط، وبإستخدام تصفية القيم الشاذة يتم تطبيق هذه النماذج كوسيلة لحساب ملاءمة كل صفحات الويب.[6]

3.1.2 علاقة التنقيب في الويب بالمجالات الأخرى

كثيراً ما يرتبط التنقيب في الويب مع إسترجاع المعلومات وإستخراج المعلومات والتعلم الآلي ولكن يوجد إختلاف في ما بينهم وذلك موضح فيما يلي :

1.3.1.2 التنقيب في الويب وإسترجاع المعلومات

إسترجاع المعلومات هوإسترجاع تلقائي لجميع الوثائق ذات الصلة بينما في الوقت نفسه إسترجاع عدد قليل من غير ذات الصله, هدفه الرئسي البحث عن الوثائق المفيده، يشمل البحث في Information retrieval:(IR) النمذجة، وتصنيف الوثائق، وواجهات المستخدم، والترشيح، أي أن التنقيب في الويب هوجزء من عملية إسترجاع المعلومات(IR).

2.3.1.2 التنقيب في الويب وإستخراج المعلومات

إستخراج المعلومات هو نوع من مرحلة ما قبل المعالجه في عملية تنقيب الويب، وهوخطوة بعد عملية ال (IR) وقبل أن يتم تنفيذ تقنيات إستخراج البيانات, هدفه إستخراج الحقائق ذات الصله من الوثائق, و مهمته هيكلة بنية الوثيقه أي أن Information extraction:(IE) أدق من (IR)، يستخدم تنقيب الويب لتحسين إستخراج معلومات الويب (التنقيب الويب هوجزء من IE).

3.3.1.2 التنقيب في الويب والتعلم الآلي

تنقيب الويب هو ليست شبيه للتعلم من الويب أو تقنية التعلم الآلي المطبقه على الويب، تقنية التعلم الآلي تساعد وتدعم التنقيب في الويب كما يمكن تطبيقها على عمليات تنقيب الويب وأيضاً بعض الاساليب أو الطرق التي تستخدم في تنقيب الويب بجانب طرق التعلم الآلي أي أن تنقيب الويب متداخل مع التعلم الآلي في الويب.[2]

4.1.2 مشاكل اللغة العربية

اللغة العربية تعيش جملـة مشكلات يمكن حصرهـا في مـــا يـأتي:

1.4.1.2 مشكلة المصطلحات

تواجه اللغة العربية في الوقت الحاضر فقراً في المصطلحات العلمية المعبرة عن مختلف مجالات الحضارة العصرية ومخترعاتها الصناعية التي تخرج إلي عالمنا في كل حين بالعشرات والمئات... وفي الحقيقه يوجد عدم تناسب المفردات العربية الأدبية مع مفرداتها العلمية والتكنولوجية.

من أمثلة المصطلحات التي يتجلى في كثير منها تكلُّف وثِقل كالدراجة النارية (وهي في الحقيقة أنواع كثيرة، ولكل نوع إسم خاص في اللغات الأجنبية) كسيارة الشحن' وسيارة النقل، وسيارة الإسعاف كأنه لا يوجد ما يعبر به عن أنواع السيارات سوى بهذه الإضافة المتكررة.[8]

2.4.1.2 كثرة المفردات

تعاني اللغة العربية من مشكلة كثرة مفرداتها الزائدة على اللزوم في بعض المجالات دون الأخرى، وعيوب هذه الكثرة تتمثل في المترادفات (الملاحظة) الموجودة للشيء الواحد، مما يجهد القارئ والمتعلم الإحاطة بكل المترادفات، كأن نجد مثلاً عشرات الأسماء للأسد، ومثلها للسيف والعسل .[8]

3.4.1.2 مشكلة المعجم العربي

1. الكتابة العربية لا تبين نطق الحروف التي ترسمها وتحتاج إلي إشارات مضافة لإبانة ذلك فالألفاظ بغير هذه الإشارات من الممكن أن تقرأ على عدة أوجه.
2. عدم ترتيب المواد ترتيبا داخلياً، ففيها خلط الأسماء بالأفعال والثلاثي بالرباعي والمجرد بالمزيد وخلط المشتقات بعضها ببعض.
3. إن المعجمات العربية أهم لتفي بعض الأحيان النص على ضبط الكلمة وبيان باب الفعل الثلاثي ومن أمثلتة قلبته أي أحببت قلبه وقلبت النخلة أي نزعت قلبها ولم يذكر الباب وقد ذكر غيره أنه من باب يفعل (بفتح فكسر).
4. غموض العبارة وتعريف اللفظ الغامض بلفظ غامض وغيرها .[9]

4.4.1.2 مشكلة النَّحو والصَّرف

قد فكر البعض في حل مشكلة القراءة بواسطة إشراك الحركات مع الحروف، بحيث تكون الحركات جزءآ من الكلمة، غير أن هذه الفكرة لن تحسم المشكلة، فمثلا : عندما ندرس قاعدة «الفاعل» في لغتنا، فإنه لا يكفي أن نعرف أنه الذي يفعل الفعل بل أن نعرف كذلكأنه مرفوع، وأنه يجب أن يقع بعد الفعل «كربح السابق» مثلا. أما إذا قلنا «السابق ربح» فيصير لفظ «السابق» مبتدأ. ويصير الفاعل مستتر في هذه الحالة.فنحن إذاً أمام مشكلة أساسية خطيرة هي مشكلة النحو، وأما مشكلة القراءة، فهي نتيجة لصعوبة النحو نفسه.[10]

5.4.1.2 مشكلة إزالة الزوائد

معظم الانظمه المتوفرة حالياً تقوم بإزالة الأحرف الزائدة عن طريق مقارنة بداية و نهاية الكلمة بمجموعة من اللواحق المحتملة .إذا طابقت بداية ونهاية الكلمة إحدى هذه اللواحق يتم إزالتها بدون النظر إلي الجزء المتبقي من الكلمة بعض الانظمة يشترط لإزالة هذه اللواحق أن يكون الجزء المتبقي من الكلمة ثلاثة احرف أو أكثر، مثلاً الكلمتين" وزير "،" وكيل "مثلاً نجد أن هذا الأسلوب يؤدي إلي قلب معاني الكلمات العربية إلي كلمات أخرى مخالفة وهذا من شأنه أن يجعل من يبحث عن الوثائق التي تحتوي كلمة " وزير" أن يسترجع الوثائق التي تحتوي على كلمة " زير " والعكس صحيح.[11]

6.4.1.2 مشكلـــة التشكيل

تستخدم اللغة العربية أيضا التشكيل لتمثيل الشكل الإعرابي للكلمة ضمن الجملة، لكن عند غياب التشكيل وهذا حال معظم النصوص العربية حالياً يصعب أخذ القرارحول معنى الكلمة كما يصعب تحديد ما إذا كانت الأحرف الأولى أو الأخيرة من الكلمة هي أحرف زائدة .مثل االحرفا لأول من الكلمة "وسادة" قد يكون حرف جرأوعطف إذا اعتبرنا أن الكلمة هي"وسادة" ويكون حرفا اصلياً إذا كانت الكلمة هي" وِسادة".[11]

7.4.1.2ي مشكلـــة جمع الاسماء

جمع الأسماء فهو من أصعب أبواب النحو بينما الجمع في اللغات اللاتينية لا يتوقف سوى على إضافة حرف معين إلي الاسم المفرد. فنحن أولاً أمام ثلاثة أنواع من الجمع : جمع المذكر السالم، وجمع المؤنث السالم، وجمع التكسير. أن جمع التكسير غني بالأوزان التي لا حصر لها، وأغلبها ليست لها قاعدة مطردة. ولعل الكثير منا يحار في جمع هذه الكلمات كلها أو بعضها : «شق، طل : قفا، رئبال، رؤوم» وأمثالها كثير جدا.[10]

الفصل الثاني
الترجمة

1.2.2 المقدمة

يتناول هذا الفصل الترجمة و أنواعها وطرق حل المشاكل التي تواجهها .

2.2.2 الترجمة

معنموالمعلوماتأصبحالإنترنتمتاحلجميعالناس فيأنحاءالعالم و بلغاتمختلفةبواسطةوسائلالإعلام،يحتاج المستخدمون لإسترجاع المعلومات في لغة آخري غيرلغتهم لذلك هم بحاجة إلي ترجمة هذه المعلومات.

1.2.2.2طرق الترجمة

هنالك عدة طرق للترجمة منها ترجمة القاموس، والترجمة الآلية، و ترجمةالمجاميع المتوازية وسنتعرف عليها بالتفصيل فيما ياتي:

1.ترجمة القاموس(Dictionary Translation)

أصبحت قواميس القراءه الآلية متاحة على نحو متزايد وتستخدم في الترجمةعن طريق بحث القاموس البسيط، لذلك هذة الطريقة بسيطة نسبيأًمقارنة مع البدائل ولكنها تعاني من نقطتي ضعف الغموض و عدم تغطية المفردات.

الغموض هو مشكلة رئيسية تؤثر على الأنظمة التي تستخدم ترجمة القاموس، لأن القاموس ثنائي اللغة يوفر ترجمات متعددة لكل مصطلحات الاستعلام، إختيار ترجمة دقيقة من مجموعة من البدائل هي مهمة غير بسيطه، وتناولت الأنظمه السابقه مشكلة الغموض بطريقة بدائية بسيطة عن طريق إختيار الترجمة الأولى التي يمنحها القاموس.إستراتيجية إزالة الغموض لديها عيوب واضحة، وسرعان ما حلت محلها تقنيات أكثر تطوراً إستغلت إحصائيات مشاركة الحدث للمصطلح، هذة الطريقة قادرة على تحديد الترجمة الأكثر إحتمالا للإستفسار المعين.[12]

2.الترجمة الآلية (Machine translation)

الترجمة الآلية هي العملية التي يتم من خلالها إستخدام برامج الكمبيوتر لترجمة النص من اللغة الطبيعية(مثل إنجليزي) إلي لغة آخري (مثل عربي) .ومعنى النص في لغة المصدر يجب أن يتم له إستعادة كاملة في لغة الهدف.ولذلك يجب للترجمة تفسير وتحليل كل عنصر من العناصر في النص ويعرف كيف أن كل كلمة قد تؤثر على الأخرى.

للترجمة الآلية عيوب عديدة منها أن تطويرها يتطلب قدراً كبيراً من الوقت والموارد,وهي تحتاج بيانات تدريب واسعة، وأيضا يحدث بها مشكلة ال OOV . [13,14]

3. المجاميع المتوازية(Parallel Corpora)

المجاميع المتوازيةهي مجموعة كبيرة من الوثائق وترجمتها في واحدة أو أكثرمناللغات الأخرى، تُحلل هذه الوثائق المرتبطة ويمكن أن تستخدم لإستنتاج ترجمة أكثر مُلائَمة من بين اللغات فيCorpus، أوهي الموارد الغنية التي تحتوي على علاقات الترجمة بين النصوص والجمل والعبارات والكلماتويمكن الحصول عليها من مصادر مختلفة (مثل المنظمات الدولية و الشبكة العالمية), أو يمكن أن تترجم من قبل الإنسان يدويا أو بإستخدام نظم الترجمة الآلية.إستخدام المجاميع المتوازية في الترجمة، يجب محاذاة النص الأصلي وترجمته إليجملة او فقره،ثم تستخدم هذه الأزواج المُنحاذه لتدريب نموذج الترجمة الإحصائي.

من عيوب ترجمة المجاميع المتوازية صعوبة الحصول على مجموعات الوثائق المناسبة, وبناء المجاميع المتوازيةيستغرق وقتا طويلاً حتى عندما يقتصر على مجالات معلومات محددة,وكذلك تعاني هذه الترجمة من مشكلة الOOV .[12,13]

الطريقتين الثانية والثالثة يتطلبان عمل مكثف جدا، الترجمه الآليه تتطلب معرفة دقيقة لقواعد اللغتين و برمجتها وهذه العملية مكلفة ومرهقة، وترجمة المجاميع المتوازية تتطلب جهد ووقت لبرمجة كافة قواعد اللغة لبناء المجاميع المتوازية (parallel corpora) .مع تزايد القواميس التي يتم قراءتها آليا، أصبحت الطريقة الأولى أكثر إستخداماً .

3.2.2 حل مشاكل الترجمة

الهدف الرئيسي من الترجمة هو نقل المعني من لغة إلي لغة أُخري، إن تقنيات الترجمة التي إنتشرإستخدامها لها مزايا وعيوب، يمكننا أن نلخص المشاكل الكبرى التي تواجه تقنيات الترجمة في مشكلة الغموض في المعني ومشكلة عدم تغطية المفردات(OOV) ، ولكن مشكلة الغموض لا نتحدث عنها لأنها خارج نطاق المشروع .يقترح الباحثون في المجال حلول متعددة لمعالجة مشكلة عدم تغطية المفردات، نوضح منهم الاتي :

1.3.2.2 التعريب(Transliteration)

هو عملية يتم فيها تمثيل الكلمات من حروف أبجدية إلي حروف أبجدية أُخرى .على سبيل المثال : الإسم الروماني الأبجدي "محمد " تم ترجمتة الي الإسم الإنجليزي Muhammad.

يمكـن أن يـتم التعريـب عـن طريـق تحديـد أوجـه التشـابه فـي الهيكـل الهجـائي للغتـين، وتسـتخدم هـذه التشـابهات بعـد ذلـك لتوليـد قواعـد تحـدد كيفيـة توضـيح الكلمـة بلغـة إلـي أُخـرى. ولكـن هـذا الـنهج يعمـل عنـدما تكونواللغـات تشـترك فـي أبجديـة متماثلـه مثـل الإنجليزيـة والفرنسـية. التعريـب بـين لغـات لا تتشـابهه فـي الابجديـة مثـل العربيـة والإنجليزيـة تتطلـب عمليـه وسـيطه تعـرف بالتحويـل الصـوتي (phoneticmapping).[12,13]

2.3.2.2إستغلال مجاميع الويب (Exploiting Web Corpora)

تتزايد المعلومات على الويب بإستمرار، هناك إمكانيات كبيرة لإستغلال الإنترنت مثل المجاميع التلقائيه للعثورعلى ترجمه فعالة لمصطلحات الإستعلام.إقترح Pu-Jen Chengوآخرون [19]نظام على الويب للتعامل مع ترجمة الإستعلامات المجهولة، الويب يتألف من النصوص الغنية بمزيج من اللغات المتعددة كثيراً منها يحتوي على ترجمة ثنائية اللغة من أسماء الاعلام، أسماء الشركات، وأسماء الأشخاص ويتم البحث عن مصطلحات باللغة الإنجليزية فقط لصفحات بلغة معينة على سبيل المثال الصينية أو اليابانية التي تم إرجاعها عادة من قائمة مرتبة من ملخصات (بما في ذلك العناوين والأوصاف للصفحة)لمساعدة المستخدمين على تحديد وثائق مثيرة للإهتمام .ثم يتم تنقيب ترجمة الإستعلام من خلال الإسترجاع الديناميكي من صفحات نتائج البحث بلغتين لإستخراج معني قريب للترجمة .

البـــاب الثالث

الدراسات السابقة والتقنيات والأدوات المستخدمة

الفصل الأول:الدراسات السابقة

الفصل الثاني:التقنيات والأدوات المستخدمة

الفصل الأول

الدراسات السابقة

1.1.3 مقدمة

أصبح الويب مصدراً غنيا للمعلومات ويستمر التوسع في الحجم والتعقيد. هنالك تحدي في إسترجاع صفحة الويب المطلوبة على الويب بكفاءة وفعالية.الدراسات التي تناولت هذا الموضوع وإن كانت بعض هذه الدراسات ليست لها علاقة وثيقة بالنظام الحالي لكنها تناولت الفكرة والمفهوم العام وسوف يتم التعرض لها من خلال هذا الفصل.

2.1.3الدراسة الأولي

Named Entity Translation with Web Mining and Transliteration

تقدم هذه الدراسة نهجاجديداً لتحسين ترجمة أسم الكائن (Named Entity)، والذي يشير إلي مجموعة من المفاهيم مثل أسماء الناس، وأسماء الأماكن, وأسماء المنتجات وغيرها، تركزهذه الدراسة على ترجمة أسماء الأشخاص من الإنجليزية إلي الصينية من خلال الجمع بين نهج التحريف مع تنقيب الويب، لذلك إستخدمت معلومات الويب كمصدر لإكمال التحريف لتحسين دقة الترجمة من 18.5% إلي 47.5%, أستخدمت معلومات التحريف لتوجيه تنقيب الويب و لتحسين تغطية الخيارات المرشحة للترجمة من 54.5%إلي .574 ويعمل نموذجMaximum Entrop(ME) لتصنيف المرشحين للترجمة عن طريق الجمع بين تشابه النطق والسياق باللغتين.في هذه الدراسة يوصى بتطبيق النظام بين أزواج لغات مختلفة.[16]

3.1.3 الدراسة الثانية

Base Noun Phrase Translation Using Web Data and the EM Algorithm

تقدم هذهالدراسة حل لمشكلة ترجمة الجملة الاسمية (NP) من اللغة الانجليزية إلي الصينية، يتم إدخال الجملة الاسمية في الويب وتحدد مجموعة من الخيارات المرشحة لترجمة الجملة وبعد ذلك تحدد الترجمة المحتملة من بين المرشحين للترجمة بواحدة من الطريقتين وهما مُصَنِّف (Naïve Bayesian) مع خوارزميةExpectation Maximization (EM) وسميت (EM-NBC-Ensemble)والطريقة الثانية استخدمت TF-IDF vectors مع خوازمية EM وسميت (EM-TF-IDF) وكانت النتائج على التوالي 91.4% و 79.8%.كما موضح ادناه:

Data	Accuracy(%)		Coverage (%)
	Top 1	Top 2	
Web (EM-NBC-Ensemble)	62.9	79.7	91.4
Non-web (EM-NBC-Ensemble)	56.9	74.7	79.3
Web (EM-IF-IDF)	62.2	79.8	91.4
Non-web (EM-TF-IDF)	51.5	71.4	78.5

الشكل 1.3 يوضح نتائج هذه الدراسة

في هذه الدراسة يقترحوا تطبيق النظام بين أزواج لغات مختلفة.[17]

4.1.3 الدراسة الثالثة

Translating out-of-vocabulary (OOV)

أستخدمت هذه الدراسة نظام STRAND الذي يقوم بالبحث في الإنترنت عن النص المتوازي و لوحظ من قبل Zhang و Vines،عندما تحدث المصطلحات الإنجليزية في صفحات الويب الصينية،خاصة عندما تحدث داخل قوسين، فهيمن المحتمل جدا أن تكون ترجمة للمصطلح الصيني السابق له مباشره .Cheng وآخرون لاحظوا أن في حالة حدوث المدى الصيني في صفحة الإنترنت الإنجليزية أن ترجمته تكون موجودة عادة في نفس الصفحة.

في هذه الدراسة تم إدخال 310 مصطلح صيني من فئات مختلفة وحقق معدل الإدخال 95٪. ودقة الترجمة الاجمالية 90٪. في هذه الدراسة يوصى بتطبيق النظام بين أزواج لغات مختلفة.[18]

5.1.3 الدراسة الرابعة

Term Translation Extraction Using Web Mining Techniques

في هذه الدراسة تم تطوير نظام CLWS،يدعى(LiveTrans) في الوقت الحالي، لترجمة المصطلحات من اللغة الإنجليزية إلي الصينية، يمكن للنظام أن يولداقتراحات ترجمة مختلفة لإستعلامات الويب الغير معروفة، النظام يستخدم على نحو فعال نوعين من موارد بيانات الويب :

مرساة النصوص(anchor-text) وصفحات نتائج البحث (search-result pages) إستخدمإختبار (chi-square)لتحديد الترجمة الأكثر إحتمالاً وهي فعالة في تقليل عدد أخطاء الربط غير المباشر.

في هذه الدراسة تم جمع أنواع مختلفة من مصطلحات الإختبار مثل سجلات الإستعلام الكبيرة و المصطلحات الفنية، النهج حقق دقة بمعدل 67٪ و 44٪ على التوالي تم إختيارها عشوئياَ .في هذه الدراسة يوصى بتطبيق النظام بين أزواج لغات مختلفة.[19]

الفصل الثاني

التقنيات والأدوات المستخدمة

1.2.3المقدمة

هذا الفصل يحتوي على وصف التقنيات المستخدمة والأدوات التي يمكن أن تساعد على عمل هذا النظاموهي :

2.2.3 لغة الجافا (Java)

هي عبارة عن لغة برمجة إبتكرها جيمس جوسلينج (James Gosling) في عام 1992 أثناء عمله في مختبرات شركة صن (Sun Microsystems) وذلك لإستخدامها بمثابة العقل المفكر المستخدم لتشغيل الأجهزة التطبيقية الذكية مثل التلفاز التفاعلى،من خصائصها إنها إحدى اللغات الكائنية التوجه (languages Oriented(Object،آمنة ، وجيدة الأداء، وإضافة الحركة والصوت إلي صفحات الويب، وكتابة الألعاب والبرامج المساعدة، وإنشاء برامج ذات واجهة مستخدم رسومية، تتميز بتصميم برمجيات تستفيد من كل مميزات الأنترنت، توفر لغة الجافا بيئة تفاعلية عبر الشبكة العنكبوتية وبالتالي تستعمل لكتابة برامج تعليمية للإنترنت عبر برمجيات المحاكاة الحاسوبية للتجارب العلمية وبرمجيات الفصول الإفتراضية للتعليم الإلكتروني والتعليم عن بعد. لا تنحصر فاعلية الجافا في الشبكة العنكبوتية فقط بل تمكننا من إنشاء برامج للإستعمال الشخصي والمهني حيث يوجد العديد من البرمجيات التي تسهل عملية كتابة الأوامر كـ netbeans وEclipse.[20]

في المشروع محل الدراسة قمنا بإستخدام هذه اللغة لأنها تدعم الكثير من المكتبات (API) التي تساعدنا في تطبيق النظام بصورة أكثر كفاءة.

(eXtendable MarkupLanguage)3.2.3XML

في البدايه تم وضعها من قبل W3C Generic SGML تحت رعاية w3 في عام 1996 ويرأسها جون بوزاك من(Sun Microsystems) بمشاركة فعالة مع فريق العمل.

XML هي اختصار لeXtendable Markup Language أي لغة الترميز القابلة للتوسّع, وهي ليست لغة برمجيَّة بل إنَّها تنتمي لعائلة لغات الترميز "Markup Languages" الَّتي تنتمي إليها لغة HTML.

وهي لا تحتوي على أوامر أو عمليَّات، تم تصُمّيمها لنقل وتخزين البيانات، وظيفتها تنحصر على ترميز النصوص برموز معيَّنة تفهمها جميع المتصفحات,تتميز بأنها ليس لديها وسومٌ محددة يتم تصميم الوسوم بواسطة المستخدم, حساسة لحالة الأحرف ويجب أن يكون لكل وسم فتح وسم إغلاق.[21]

NetBeans4.2.3

هو يرنامج مفتوح المصدر مكرّس لتقديم منتجات تطوير البرمجيات القوية التي تلبي إحتياجات المطورين والمستخدمين والشركات، يُمكّن المؤسسات التي تعتمد عليه كأساس لمنتجاتها البرمجية من تطوير منتجاتها بسرعة وكفاءة وسهولة من خلال الإستفادة من نقاط القوة في منصة جافا (Java platform) ومعايير الصناعة الأخرى ذات الصلة .

في يونيو من عام 2000، تم جَعله مفتوح المصدر من قِبل شركة (Sun Microsystems) ، حَيث ظلت هذه الشركة هي الراعية لبرنامج (لبرنامج) حتى شهر يناير من عام 2010 .

المنتجان الأساسيان لبرنامج(NetBeans) هما (لبرنامجIDE) و (NetBeans Platform) و هما مُتاحان مجاناً للإستخدامات التجارية و غير التجارية؛ حيث أن شفرة المصدر (source code) لكليهما متاحة لأي جهة من أجل إعادة إستخدامها بالطريقة التي تراها مناسبة .

يحتوي القسم القانوني (legal section) من هذا يرنامج على المعلومات المتعلقة بالترخيص، وقضايا حقوق التأليف والنشر، وسياسة الخصوصية وشروط الإستخدام .

من مميزات انه مفتوح المصدر (Open Source) و يعمل على العديد من المنصات بما فيها (Windows, Linux, Solaris, and the MacOS) ومعظم المطورين يعتبرون أن بيئة العمل (NetBeans) هي البيئة الأساسية لإنشاء مشاريع بلغة الجافا، لأنها توفر الكثير من المزايا للعمل و إنشاء البرمجيات بإستخدام لغة الجافا، كما أنه يوفر بيئة تطور متكاملة لدعم عدة لغات أخرى مثل: (PHP, JavaFX, C/C++ and JavaScript) . [22]

5.2.3إختبارChi-Square

هو إختبار إحصائي يطبق على مجموعة من المتغيرات لتقييم الإحتمال الأفضل،

(Pearson's Chi-Square test) قد تم التحقق من خصائصها أول مرة من قبل(Karl Pearson)في عام 1900، أستخدمت مبكراً من قبل (Church and Gale) في عام 1991 في إحصائيات معالجة اللغة الطبيعية لتحديد أزواج الترجمة في المجاميع، أيضا طبقت كمقياس للتشابهه في المجاميع من قبل

(Kilgariff and Rose)في عام 1998.

إخترنا إختبارالChi-Square لأنه أفضل دقة في الترجمة وأيضاً parameters(المدخلات) المطلوبة لإختبار chi-square يمكن الحصول عليها على نحو فعال من محركات البحث في العالم الحقيقي.

خوارزمية Chi-Square

$$S_{\chi 2}(s,t) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)},$$

s: الكلمة العربية المدخلة على سبيل المثال كلمة (كتاب).

t: الكلمة الانجليزية المحتملة (book).

a: عدد نتائج البحث التي تحتوي على كل من t,s (4).

b: عدد نتائج البحث التي تحتوي على s و لا تحتوي على t (3).

c: عدد نتائج البحث التي تحتوي على t و لا تحتوي على s (2).

d: عدد نتائج البحث التي لا تحتوي على كل من t,s (2).

N : كل نتائج البحث . N= a+b+c+d (11).

الباب الرابع

التطبيق

1.4 مقدمة

يتناول هذا الباب خطوات النظام, ووصفه, ومعماريته,و مخططاته.

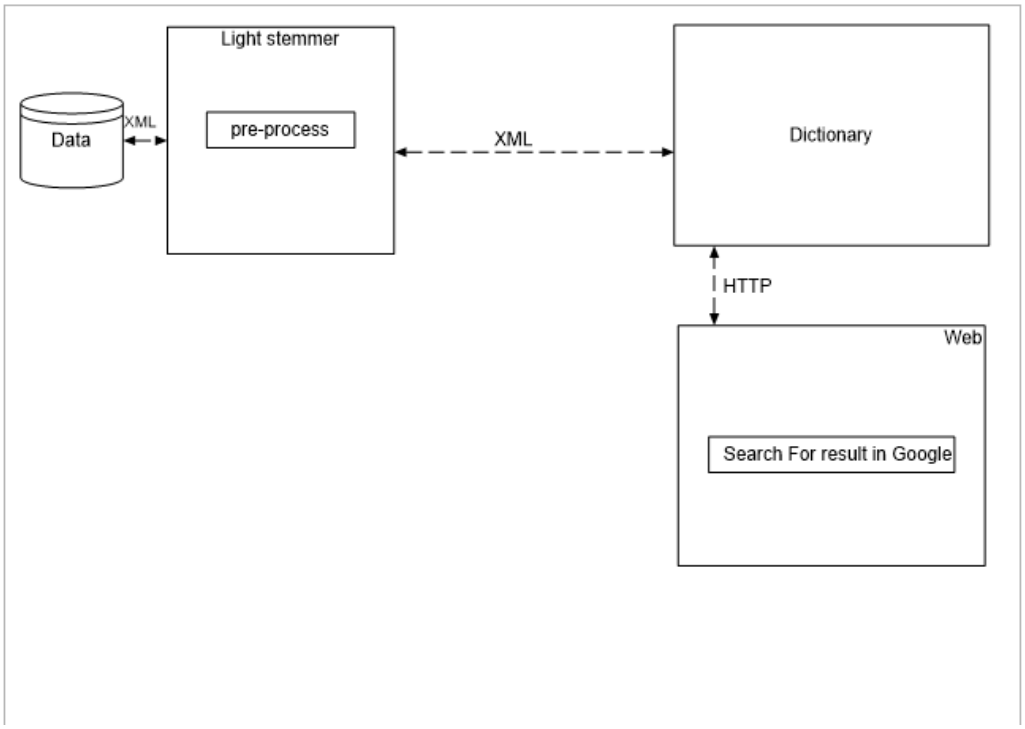
2.4خطوات النظام

1. إدخال المصطلح في الواجهة.
2. تجريد المصطلح المدخل من السوابق واللواحق مثل حروف العطف, وحروف التعريف وغيرها، إستخدمنا (light stemmer version 10)
3. يتم البحث عن المصطلح في القاموس الثنائي اللغة إذا وجدت يتم إسترجاع المعنى .
4. إذا لم يوجد في القاموس يتم تنقيب الويب للبحث عن المعنى .

3.4 معمارية النظام (Architecture)

1- بيانات إختبار النظام عبارة عن مستندات XML.
2- المعالجة المبدئية(pre-process) :يتم تجريد المصطلح المراد البحث عنهمن السوابق واللواحق.
3- البحث في القاموس(dictionary) :عبارة عن مستند XML فيه بعض المصطلحات ومعانيها .
4- الويب (web): فيه يتم البحث عن المصطلحالذي لا يوجد في القاموس وإستخدمنا في البحث نتائج بحث قوقل.

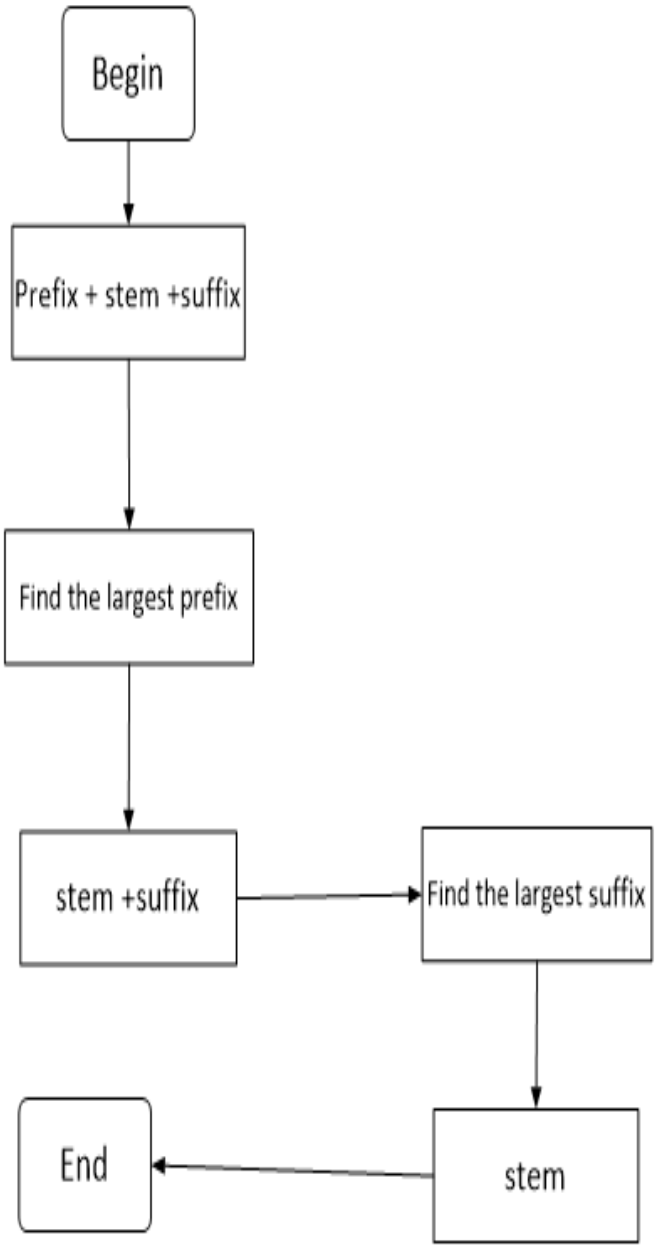
الشكل أدناه يوضح معمارية النظام:



الشكل 1.4 معمارية النظام

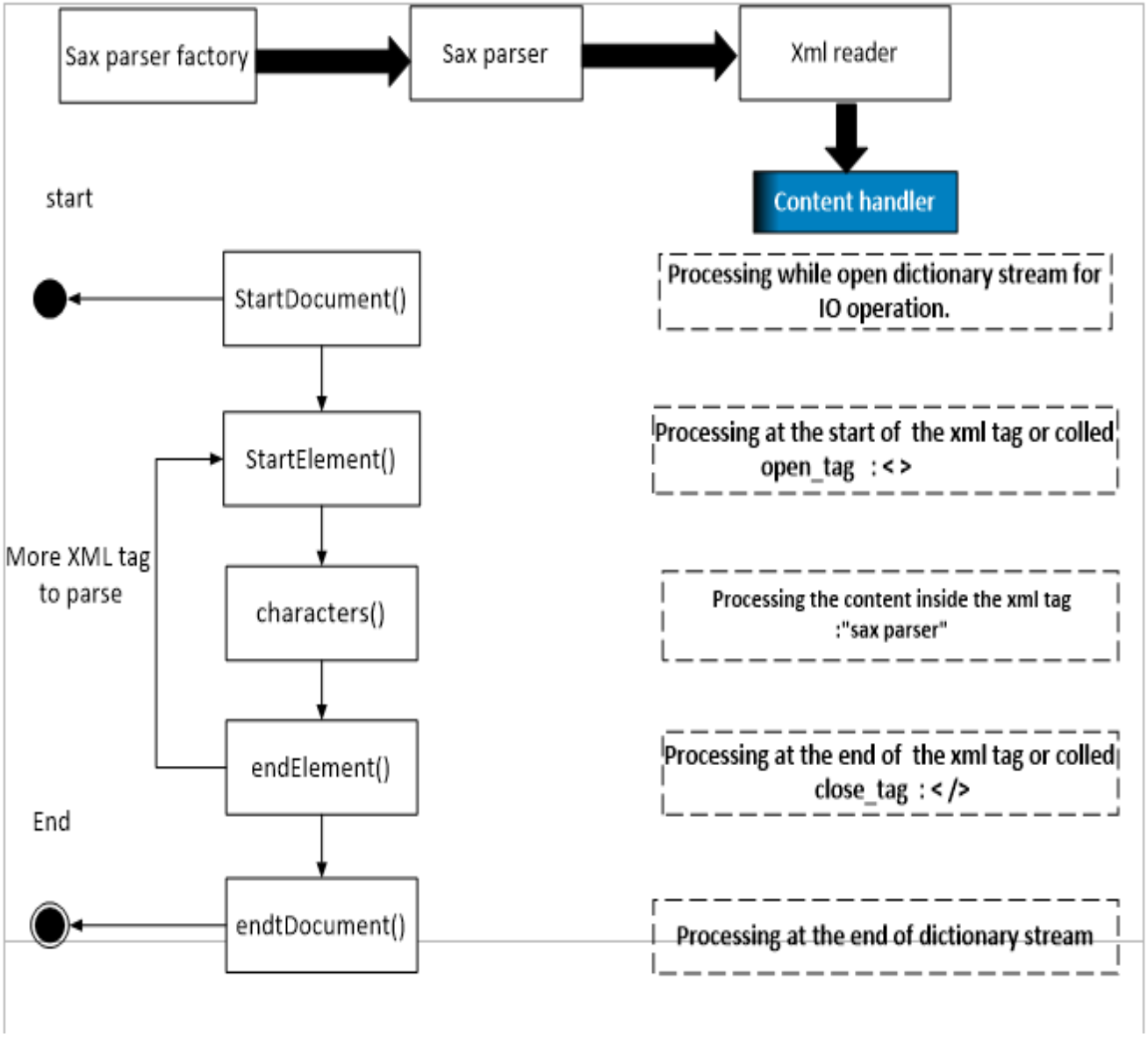
4.4 مخططات النظام

المخطط أدناه يوضح عملية تجريد المصطلح من السوابق واللواحق:



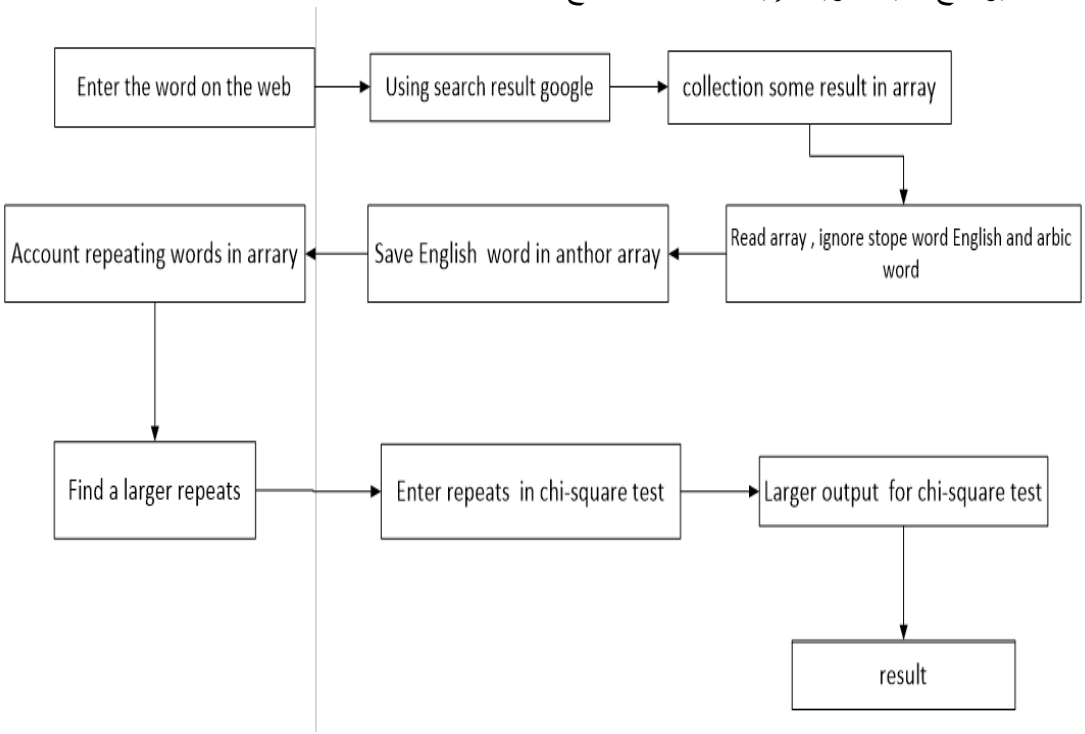
الشكل 2.4 تجريد المصطلح

الشكل أدناه يوضح عملية القراءة من القاموس الذي هو عبارة عن مستندXML



الشكل 3.4 طريقة قراءة XML

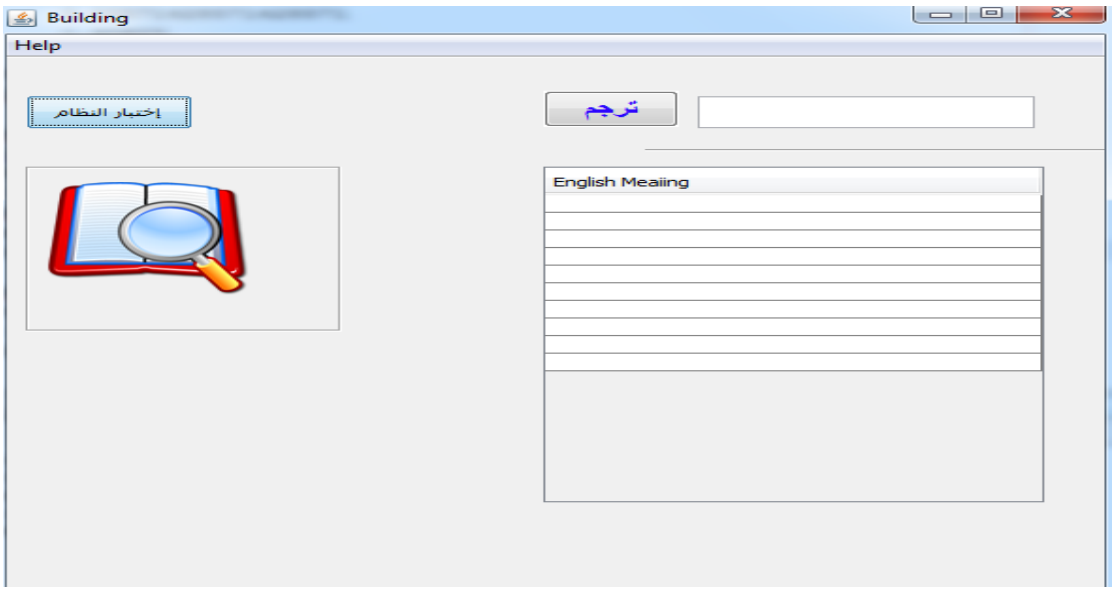
الشكل أدناه يوضح تنقيب الويب لإيجاد معنىالمصطلح المدخل :



الشكل 4.4 تنقيب الويب لإيجاد معنىالمصطلح

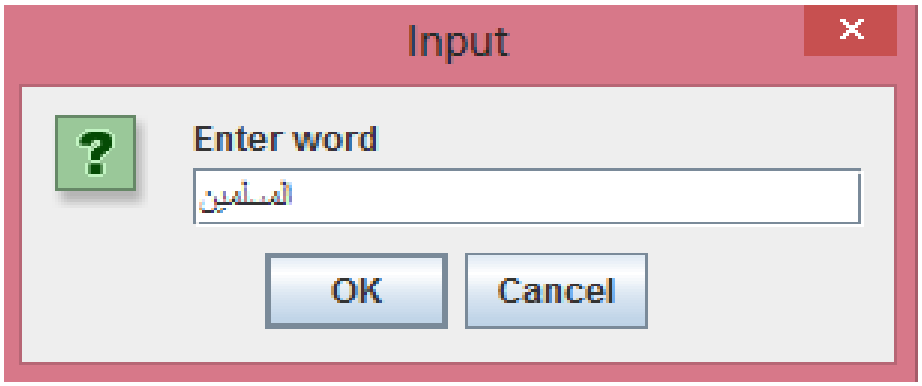
5.4 وصف النظام

يبدا النظام بإدخال المصطلح في الواجهه كما موضحة في الشكل (5.4) وعند الضغط علي زر (ترجم) يجرد المصطلح قبل البحث عنه من السوابق مثل (ال ، وال ، بال ، كال ، فال ، لل ، و) واللواحق مثل (ها ،ان ، ات ، ون ، ين ، يه ، ية ، ه ، ة ، ي) وبعد ذلك يتم البحث عنهمجردلكي يكون البحث بصورة فعالة .



الشكل 5.4 واجهة التنفيذ

وعملية التجريد لا تكون من نواتج التنفيذ و الشكل أدناه يوضح كيفية التجريد عندإدخال كلمة (المسلمين)



الشكل 6.4 ادخال الكلمة للتجريد

تجريدها من (ال) و (ين) وناتج التجريد موضح أدناه:

الشكل 7.4 ناتج التجريد

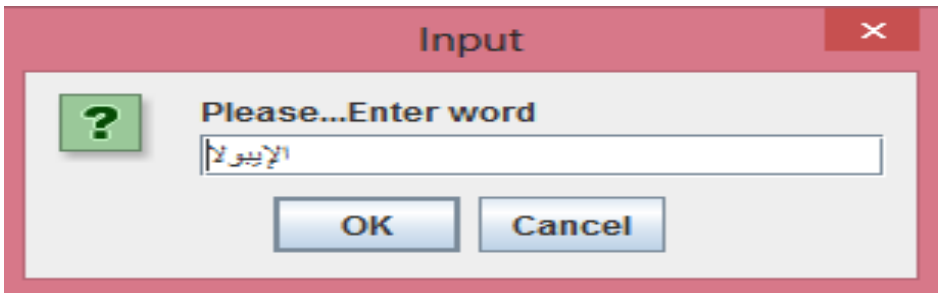
يتم البحث عن المصطلح في القاموس وهو عباره عن مستند xml(كما موضح في الشكل 7.4) و يستخدم parser (SAX) و(DOM parse) لقراءة مستندات الXML وهنا إستخدمنا للقراءة الSAX لأن DOM يُحمِّل كل مستندات الXML في الذاكرة، القاموس حجمه كبير لذلك يحتاج الي ذاكرة كبيرة ويحتاج أيضا إلي زمن ليتم فيه التحميل، بينما الSAX بتحمل جزء من مستنداXML لذلك لا يستنزف ذاكرة ولا زمن كبير في البيانات الضخمة .

```
<dictionary>
  - <entry>
        <word>انتكص</word>
        <pos>R060</pos>
        <eng_word>fall back</eng_word>
    </entry>
  - <entry>
        <word>انتكس</word>
        <pos>R060</pos>
        <eng_word>relapse</eng_word>
    </entry>
  - <entry>
        <word>انتكس</word>
        <pos>R060</pos>
        <eng_word>be reversed</eng_word>
    </entry>
  - <entry>
        <word>انتكس</word>
        <pos>R060</pos>
        <eng_word>be inverted</eng_word>
    </entry>
  - <entry>
        <word>انتكث</word>
        <pos>R060</pos>
        <eng_word>be violated</eng_word>
    </entry>
```

الشكل 8.4جزء من القاموس

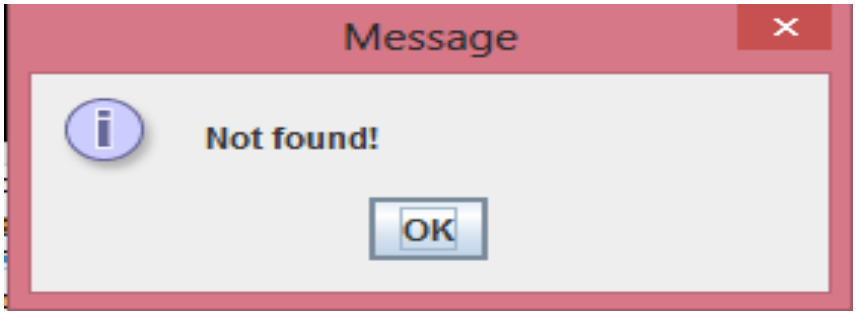
تتم القراءه من مستندال XML بفتح القاموس وذلك عن طريق "tag"<dictionary> نقارن كل بداية ونهاية ال(tags) التي بالملف مع ال(tags) التي حددناها في ال(code) وذلك بالدالة startElement وendElementإذا كانت متطابقة نقرأ محتوى الtags بالدالة charactersوبذلك عندما ندخِل المصطلحونقارنهمع محتوى<word></word>إذا تطابقا يعني أنهموجودو يتم إسترجاع محتوى< /eng_word ><eng_word> الذي يدل على المعنى،أما إذا لم يتطابقا تعني أنه لم يتم العثور عليه في القاموس لذلك يتم إستخدام تنقيب الويب (web mining).

فمثلا مصطلح (الإيبولا) تم إدخاله في القاموس كما موضح أدناه :



الشكل 9.4إدخال المصطلح في القاموس

وكان ناتج البحث عنه في القاموس (not found) كما موضح أدناه اي أنه غير موجود

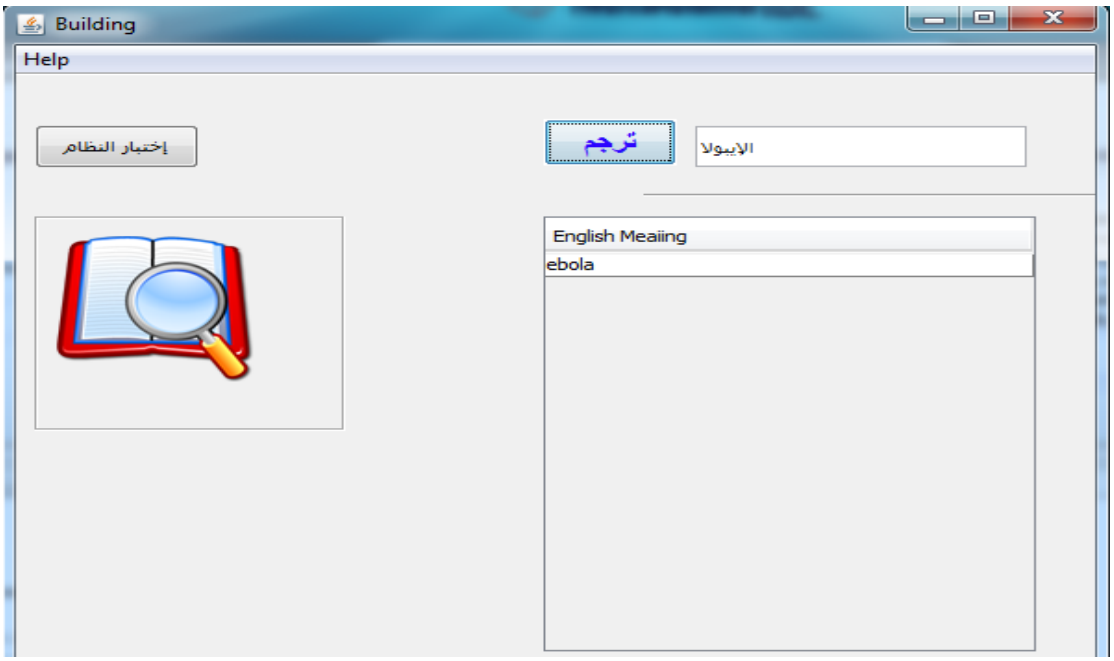


الشكل 10.4 ناتج الإدخال

نقوم بإدخال المصطلحفي الويب،يتم إسترجاعمجموعة من نتائج البحث عن طريق محرك بحث قوقل إستجابة للإستعلام عن المصطلح, فأخذنا بعض نتائج البحث و حفظناها في مصفوفة، بعد ذلك تمت قراءة النتائج منالمصفوفة وجردناها من *English stop* words*s*ومن الكلمات العربية وحفظنا المصطلحات الإنجليزية في مصفوفة أخرى وحسبنا تكرار أي مصطلح في هذه المصفوفة، من هذا التكرار حددنا أكبر تكرارات محتملة، أدخلنا في إختبارChi-Square المصطلح المدخل مع أكبر تكرار محتمل له ونتائج البحث، لإيجاد عدد نتائج البحث التي بها كل مصطلح مع أكبر إحتمال معنى له، وعدد نتائج البحث التي بها المصطلح فقط، و عدد نتائج البحث التي بها إحتمال المعنى فقط، وعدد نتائج البحث التي لا توجد بها كل من الكلمتين.بذلك تحسب نسبة المعنى المحتملوهكذا مع بقية أكبر تكرارات محتملة، أكبر نتيجة للإختبار من التكرارات المدخله تعتبر معني المصطلح. كما في الشكل(10.4) أدناه تم إدخال مصطلح (الإيبولا) في الويب كانأكبر تكرار لمعنىالمصطلح هو (Ebola).

الشكل 11.4 نتيجة إدخال المصطلح في الويب

الشكل أدناة يوضح ناتج تنفيذ النظام بعد إدخال مصطلح (الإيبولا)



الشكل 12.4 ناتج تنفيذ النظام

الباب الخامس

النتائج والتوصيات

1.5 النتائج

لإختبار النظام تم إدخال 180 مصطلح باللغة العربية(بصورة عشوائية من مختلف الأشخاص)، كانت نسبة المصطلحات الصحيحة 80% و نسبة الخطأ20%ونعزي ذلك لمشاكل المعالجة المبدئية للغة العربية– خارج نطاق الدراسه- فمثلا مصطلح (السيلكون) عند إدخالة تم حذف(ال)و (الواو) و (النون) لذلك كانت ترجمتة خطأ.

الكلمات التي تأتي معانيها بكلمات متكرره في صفحات الويب مثل (ago) و(cached)أعتبرت مع (stopwords) للحصول على نتائج افضل.

الكلمات	النسبة
معاني المصطلحات الصحيحة	%80
معاني المصطلحات الخطأ	%20

الجدول1.5نتائج الدراسة

2.5 التوصيات

تجربة النظام ببيانات أكبر و زيادة صفحات نتائج البحث لتحسين النتائج. إستخدام تقنيات أخرى فى التنقيب ومقارنة النتائج مع النتائج التى تحصلنا عليها. إستخدام تقنيات لمعالجة اللغة العربية فى مرحلة المعالجه المبدئيه التي ربما تحسن من النتائج.

3.5 الخاتمة

عدم تغطية المفردات(oov) تنشأ من حقيقة بعض المصطلحات التي لا توجد في القواميس ثنائية اللغة بسبب انها صيغت حديثا لذا فكرة الدراسة كانت تركز لإيجاد طريقة لحل هذه المشكلة، وبحمد الله وتوفيقة قمنا بتطوير نظام لحل هذه المشكله بين اللغتين العربية والإنجليزية وذلك عن طريق تنقيب محتوى الويب وإجراء اختبار إحصائي، وتم إختبار النظام بإدخال بعض مصطلحات حيث ظهرت بعض المشاكل في المعالجة المبدئية للغة العربية والتي أثرت علي بعض نتائج النظام . نتمنى من الدارسين والباحثين والمهتمين بالمجال إيجاد طرق لحلول هذه المشاكل لجعل النظام أكثر كفاءة.

المراجع

[1]الرابط يحتوي على معلومات عن تنقيب في الويب

www.upet.ro/annals/economics/pdf/2012/part1/Dinuca-Ciobanu.pd

التاريخ: 2015/6/10 الزمن:3:46PM

[2]Raymond Kosala and Hendrik Blockeel.Department of Computer Science Katholieke ,Universiteit Leuven Celestijnenlaan .Web Mining Research: A Survey.Belgium (July 2000).

[3] Johnson, F., Gupta, S.K.Web Content Minings Techniques: A Survey, International Journal of Computer Application.

[4]الرابط يحتوي على معلومات عن تنقيب محتوىالويب:

http://www.claes.sci.eg/coe_wm/WebContentMining.pdf

التاريخ: 2015/6/12 الزمن:3:46PM

[5]الرابط يحتوي علي خطوات تنقيب محتوىالويب :

www.scaleunlimited.com/about/web-mining/

التاريخ: 2015/6/12 الزمن:1:46PM

[6]Shaily G.Langhnoja 1, Mehul P. Barot2, Darshak B. Mehta. Computer Department, M.E.(Pursing) Gujarat Technical University web sage Mining Using Association Rule Mining on Clustered Data forPattern Discovery.(June 2013)

[7]R.Malarvizhi, K.Saraswathi. Department of Computer Science, Government Arts College (Autonomous). Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study.(August 2013)

[8]الرابط يحتوي علي مشكلة المصطلحات وكثرة المفردات في اللغة العربية :

http://uqu.edu.sa/page/ar/148320.

التاريخ: 2015/4/8 الزمن:3:20PM

[9]الرابط يحتوي علي مشكلة المعجم العربي:

www.iasj.net/iasj?func=fulltext&aId=51275

التاريخ: 2015/4/8 الزمن:5:52PM

[10]الرابط يحتوي علي مشكلة جمع الاسماء و مشكلة النحو والصرف :

www.habous.gov.ma/daouat-alhaq/item/332-مشاكل-اللغة-العربية

التاريخ: 2015/4/8 الزمن:6:14PM

[11]Abdusalam F.A. Nwesri .School of Computer Science and Information Technology, RMIT University , Melbourne 3001, AustraliaArabic Text Processing for Indexing and Retrieval(معالجة النصوص العربية لأجل فهرسة و الاسترجاع)

[12] Zhou, D., M. Truran, T. Brailsford, V. Wade and H. Ashman (2012). "Translation techniques in cross-language information retrieval." ACM Computing Surveys (CSUR) 45(1):

[13] Mohammed Mustafa Ali. (2013). Mixed-Language Arabic- English Information Retrieval.University of Cape Town. Cape town.

[14] Ballesteros, L. A. (2001). Resolving ambiguity for cross-language information retrieval: A dictionary approach, University of Massachusetts Amherst.

[15] Cheng, P.-J., J.-W. Teng, R.-C.Chen, J.-H.Wang, W.-H.Lu and L.-F.Chien (2004).Translating unknown queries with web corpora for cross-language information retrieval.Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM.

[16] Zhichun Road, Haidian District. Institute of Information Science, Academia Sinica , Beijing 100080, Taiwan .Named Entity Translation with Web Mining and Transliteration

[17]] Yunbo Cao, Hang Li. Base Noun Phrase Translation Using Web Data and the EM Algorithm.

[18]Ying Zhang, Fei Huang and Stephan Vogel.Language Technologies Institute, Carnegie Mellon University.Mining Translations of OOV Terms from the Web through Crosslingual Query Expansion 5000 Forbes Ave. Pittsburgh, PA 15217, U.S.A.

[19] Wen-Hsiang Lu. College of Electrical Engineering and Computer Science , National Chiao Tung University. Term Translation Extraction Using Web Mining Techniques.In Taiwan, November 2003.

[20]رابط يحتوي على معلومات عن لغة جافا:

http://ar.wikipedia.org/wiki/%D8%AC%D8%A7%D9%81%D8%A7_%28%D9%84%D8%BA%D8%A9_%D8%A8%D8%B1%D9%85%D8%AC%D8%A9%29

التاريخ: 2015/8/4 الزمن:10:46AM

[21]الرابط يحتوي علي معلومات عن XML :

http://www.tech-wd.com/wd/2010/01/02/xml-first-lesson/

[22] هذا الرابط يحتوي علي معلومات عن netbeans
https://netbeans.org/about/

الملاحق

ملحق : يوضح توجيهات لمستخدم النظام