الآية

قال تعالى :

(قَالُواْ سُبْحَانَكَ لاَ عِلْمَ لَنَا إِلاَّ مَا عَلَّمْتَنَا إِنَّكَ أَنتَ الْعَلِيمُ الْحَكِيم)

سورة البقرة الآيه(32)

الحمـــد لله

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 وكلنا لك عبد ,لك الحمد مادعوناك إلا حسنُ ظنٍ بك ومارجوناك إلا ثقة فيك , وماخفناك إلا تصديقاً بوعدك ووعيدك لك الحمد حمداً كثيراً طيباً مباركاً فيه , وصلى الله على سيدنا محمد خاتم الأنبياء والمرسلين أجمعين بشر وأنذر ووعد وأوعد ,أنقذ الله به البشر من الضلالة وهدى الناس الى صراط المستقيم ,صراط الله الذي له مافي السموات ومافي الأرض الا الى الله تصير الأمور .

الإهداء

بدأنا بأكثر من يد وقاسينا أكثر من هم وعانينا الكثير من الصعوبات وهانحن اليوم والحمد لله نطوي سهر الليالي وتعب الأيام وخلاصة مشوارنا بين دفتي هذا العمل المتواضع إلى منارة العلم والإمام المصطفى إلى الذي علم إلى سيد الخلق إلى رسولنا الكريم خاتم الأنبياء والمرسلين

الي منارة العلم والإمام المصطفى إلى الأُمي الذي علم إلى سيد الخلق إلى رسولنا الكريم سيدنا خاتم الأنبياء والمرسلين محمد صلى الله عليه وسلم

إلى الينبوع الذي لا يمل العطاء إلى من حاكت سعادتي بخيوط منسوجة من قلبها إلى والدتي العزيزة

إلى من سعى وشقى لأنعم بالراحة والهناء الذي لم يبخل بشئ من أجل دفعي في طريق النجاح الذي علمني أن أرتقي سلم الحياة بحكمة وصبر إلى والدي العزيز

إلى من حبهم يجري في عروقي ويلهج بذكراهم فؤادي إلى أخواتي وإخواني إلى من سرنا سوياً ونحن نشق الطريق معاً نحو النجاح والإبداع إلى من تكاتفنا يداً بيد ونحن نقطف زهرة تعلمنا إلى زميلاتي وزملائي إلى من علمونا حروفاً من ذهب وكلمات من درر وعبارات من أسمى وأجلى عبارات في العلم إلى من صاغولنا علمهم حروفاً ومن فكرهم منارة تنير لنا سيرة العلم والنجاح إلى أساتذتنا الكرام

الشكر والتقدير

اللهم لك الحمد كله، ولك الملك كله وبيدك الخير كله وإليك يرجع الأمر كله علانيته وسره ،اللهم صل على سيدنا محمد وعلى آله وصحبه أجمعـين,وعملاً بقول الرسول عليه أفضل الصلاة وأتم التسليم:

((لا يشكر االله مـن لا يـشكر الناس))

أخرجه البخاري

لابد لنا ونحن نخطو خطواتنا الأخيرة في الحياة الجامعية من وقفة نعود 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 وقبل 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 والتقديرالمشرف الدكتور : محمد المصطفى

الذي نقول له بشراك قول رسول الله صلى الله عليه وسلم:

"إن الحوت في البحر والطير في السماء ليصلون على معلم الناس الخير"
لتفضله بالإشراف على هذه الدراسة ومنحها الوقت والجهد والنصح رغم مسئولياته المتعدده فلم يبخل علينا بالتوجيه والإرشاد طوال فترة البحث وكما نخص بالشكر كل الشكر من ساعدتنا وساندتنا لإكمال هذه الدراسه الأخت الفاضلة نسيبة الهادي

المستخلص

نظراً للتطور السريع في الويب وإنتشار المعلومات عليه بصورة واسعة أصبحت هنالك الكثير من التحديات والصعوبات التي تواجة أنظمة إستراجاع المعلومات وخصوصاً عند إنتشار الوثائق علي الويب بلغات مختلفة أو وجود الوثيقة الواحدة مكتوبة بأكثر من لغة(وثيقة متعددة اللغات) فكانت من التحديات كيفية فهرسة هذه الوثائق بالأخص متعددة اللغات بطريقة تؤدي إلى ذيادة كفاءة الإسترجاع في أنظمة إسترجاع المعلومات لذلك تم إستخدام الطريقة المركزية الموزعة التي لها فوائد عدة خاصة في فهرسة الوثائق متعددة اللغات، لبناء فهارس أحادية اللغة للوثائق أحادية اللغة وفهرس للوثائق متعدد اللغات ثم البحث في هذه الفهارس بإستخدام إستعلامات المستخدمين وبهذه الطريقة نكون قد أتِحنا للمستخدم كتابة الإستعلام باللغة التي يراها مناسبة أو التعبير عن مصطلحاته باللغة التي يعرفها, ولكن عند البحث في مجموعة الفهارس تكون هنالك أكثر من قائمة مسترجعة(قائمة لكل فهرس) فكان لابد من إيجاد طريقة لدمج هذه النتائج المسترجعة من مجموعة الفهارس في قائمة واحده تعرض للمستخدم، بحيث تكون فيها الوثائق ذات الأهمية الأكبر للمستخدم والصلة الأعلى بالإستعلام في أعلى القائمة حتى تكون النتيجة مقنعة ومفيدة للمستخدم، ولهذا تم إستخدام ثلاث خوارزميات لدمج النتائج ومقارنة نتائجها بإستخدام معيار DCGلقياس القيمة الخام(كفاءة الإسترجاع، ووجد أن خوارزميةRaw Scoreتعطي أفضل النتائج, وكان لطول الإستعلام (ذو اللغتين(عربي-إنجليزي) تأثير على نتائج خوارزمية CORI .لذلك نوصي بمعالجة الطول لتحسن كفاءة الاسترجاع.

ABSTRACT

Because of the rapid development and the spread of information widely in the web , there are a lot of challenges and difficulties faced by systems of information retrieval, especially when the spread of the documents on the web in different languages and the presence of a single document written in more than one language (Mixed Documents) was challenging how the indexing of these documents Especially Mixed Documents and then search in these indexes using queries of users, therefore users are allowed to write their quiries in different languages and with their own way , but there will be more than a retrieved list (List each index) therefore it was necessary to find a way to merge these results which are retrieved from the indexes in a single list which will be presented to the user so that the most relevant documents to the user and to the query should occupies top of the list to make the result compelling and useful to the user, and therefore the need to use merging algorithms to merge the results and to compare the results using a standard to measure the efficiency of retrieval Discounted Cumulative Gain(DCG)

The length of a query languages influence the outcome of CORI so we recommend addressing the length which will increase the recovery efficiency.

شرح المصطلحات

شرح المصطلح	المصطلح	الاختصار
الويب أو الشبكة العنكبوتية العالمية	World wide web	WWW
البحث عن الوثائق، وعن داخل الوثائق، وعن المعلومات المتعلقة المعلومات الوصفية بالوثائق، بالإضافة إلى البحث في شبكة الإنترنت وقواعد البيانات	Information retrieval	IR
وثيقة واحدة موجودة بأكثر من لغة	mixed documents	
لغة النص المتشعب	**H**yper **T**ext **M**arkup **L**anguage	HTML
الإستعلام الواحد بأكثر من لغة	mixed query	
أهمية الوثائق الوثائق بالنسبة لإستعلام المستخدم	Relevant	
كتابة الإستعلام باللغة العربية	monolingual arabic	
كتابة الإستعلام باللغة الإنجليزية	monolingual English	
جمع، تخزين البيانات لتسهيل سرعة ودقة استرجاع المعلومات.	Indexing	
مطابقة الوثائق المستجرعة مع	Matching	

إستعلام المستخدم		
عملية تقطيع تيار من نص للوصول إلى كلمات، أو رموز، أو العناصر ذات معنى تسمى القطع(tokens)	Tokenization	
كلمة، أو رمز، أو عنصر ذو معني ناتج من عملية تقطيع نص معين.	Token	
كلمات تُستَبعد قبل، أو بعد، تجهيز البيانات باللغة الطبيعية	Stopwords	
عملية جعل الكلمات موحدة و متسقة إملائياً	Normalization	
الحد من تصريف الكلمة و لإيجاد أصلها أو جذرها	Stemming	
جمع، تخزين البيانات لتسهيل سرعة ودقة استرجاع المعلومات.	Indexing	
يتم تمثيل الوثائق المفهرسة في شكل مجموعة من الكلمات المفتاحية	Keyword	
الكلمات المفتاحية التي يتم تمثيلها في الفهرس	Term	
نسبة تعبر عن عدد الوثائق المسترجعَة من مجموعة الوثائق الكلية	Recall	
نسبة تعبر عن عدد الوثائق المسترجعَة ذات الصلة من مجموعة الوثائق المسترجعة	Precision	

ترتيب الوثائق المسترجعة تنازلياً بناءاً على مدى إرتباطها بطلب المستخدم	Ranking	
عدد مرات ظهور المصطلح في الوثيقة	Term Frequency	
وزن الوثيقة	Weighted	
إسترجاع يقوم على تقدير درجة أهمية الوثائق وتحديد أي منها هوالمطابق للإستعلام	Ranked Retrieval Model	
هو نموذج يمثل كل من الوثائق والإستعلامات في شكل ناقل	Vector Space Model	
المشابهه بين الإستعلام والوثائق	Similarity	
يستخدم لتحديد أهمية المصطلح في مجموعة الوثائق	Inverse Document Frequency	IDF
هو ترتيب الوثائق على حسب الإحتمالية المقدره لهذه الوثيقة بأن تكون ذات صلة بالإستعلام	Probabilistic Retrieval Model	
إحتمالية أن الوثائق التي تم إسترجاعها تنتمي للوثائق ذات الصلة	Probabilty	
أشهر خوارزميات نموذج الإسترجاع الإحتمالي	Best Matches25	
معيار لكفاءة إسترجاع المعلومات في	Discounted Cumulative Gain	DCG

محركات البحث		
تقنية تستخدم للترجمة	Google translator API	

فهرس الأشكال

فهرس الجداول