# WOJOOD Arabic Web Search Engine

## وجود - محرك بحث يفهم اللغة العربية - مشروع تخرج عام ٢٠٠٨

Prepared by:
Ali Salhi
Anwar Hithnawi

# **WOJOOD** Arabic Web **Search Engine**

**WOJOOD** is an intelligent multilingual web search engine that focuses on Arabic. This project is concerned with building a well structured information retrieval system to help people find and manage Arabic web information they need in a minimal time. Through a period of six months this project has focused on working on topics related to retrieving Arabic web information such as Arabic language processing, document indexing, focused crawling (crawling Arabic web pages), search in Arabic PDF files(with special attention to newspapers content).

## Search...

**WOJOOD** depends on a group of powerful natural language processing methods that guarantee the best search results in Arabic documents.
**WOJOOD** provides the following web search services:
Searching the web, and searching the social bookmark system.
Incremental Search, one of the features our system will provide is to enable users to perform their searches incrementally. This means giving them a method to keep them updated, so when the user searches for query A for the first time the search engine will provide the user with available results. Next time the user searches for A, the search engine will return the new results introduced since his last search for A.

## Social Bookmark...

**WOJOOD** Social bookmark is a social bookmarking web service for storing, sharing, and discovering web bookmarks. With time this bookmark system will have a large number of web pages in almost all the topics that concern the users and that people find worth saving, and we will give the users the chance to search this database. This will limit the returned results for queries by restricting it to those results that other people found valuable. On the other side we believe this will enable us to have good repository of the best on the Arabic web.
Beside this, the social bookmark system will play a role in ranking the web pages so by time the number of people saving a particular web page will influence its ranking among other pages. That is human factor will play part in the page ranking process.

## Arabi...

Arabi is a package of natural language processing methods that employs Artificial intelligence, statistics and the language characteristics so as to help search engines better understand what users want.
This package includes:

- Arabic Language detector: This tool will allow us to determine if the language of the document is Arabic not Persian or Urdu or any other languages that use Arabic alphabetical characters, in order to crawl and index the Arabic web.

- Automatic spell corrector: This tool will enable us to give the user a list of few words that is probably the best replacement of the user's entered query if it is misspelled.

- Automatic categorizer: This tool will allow the system to categorize the Arabic web content into ten different categories that we believe are the most appropriate to categorize the content of the Arabic web.

- Automatic root extractor: This tool will automatically predict the roots of the words. We need this tool so as to index the web based on the roots of the words, instead of the words, which is more efficient. Moreover, we will need it to expand the user's queries.

- Arabic Query Expansion: If Arabic retrieval system restricts its search to the exact query without looking for its relevant words or derivatives, returned results will be poor, so to overcome such a problem, expansion techniques are used in search engines,

we will rely on the stemmer, root extractor and small repository of Arabic synonyms we have in order to expand user's queries so as it include relatives and derivatives of the query elements in the search.

### Arabic web database-AWD…

We started building the Arabic web database, a database that contains all the Arabic words on the web with their frequencies. This database will be the base, most of the Arabic NLP methods will depend on. So far the database contains around (568,106) words.

### Arabic PDF files…

One of the topics that concerns us through this project is Arabic PDF files; how we can index them so as to make them searchable by our search engine, as we believe that the content of the Arabic PDF files is very valuable specially the newspapers contents. Unfortunately none of the search engines currently index the content of the newspapers and most of them even have problems indexing the normal Arabic PDF files. So far we were able to fix the problem of indexing the normal PDF files; by normal we mean the PDF files that have been created under windows platform. Since the problem is due to the fact that Arabic is written from the right to the left, unlike most of the languages, the PDF to text converters convert the PDF in reverse order. We fixed this problem through reversing the text once again by using a function that reverses the text. As for the newspapers contents, we faced a different challenge and that's due to the fact that most of the newspapers are prepared using InDesign or Page Maker under the Mac OS platform which uses different representation of the non-ASCII characters. So far, we were able to develop our own algorithm that enables us to read the text of the PDF files but with some errors. Currently we are working on making it function in a better way.

وجود

## وجود محرك بحث عربي

وجود محرك بحث و مفضلة إجتماعية عربية, من خلال وجود نحن نسعى لتصميم محرك بحث ذكي متعدد اللغات مع التركيز على اللغة العربية. هذا المشروع يعنى ببناء هيكل جديد لنظام إسترجاع المعلومات لمساعدة الناس على العثور على المعلومات التي يحتاجونها بطريقة تتسم بالكفاءة.
خلال ستة أشهر من العمل على هذا المشروع تم التركيز على مواضيع ذات صلة في إسترجاع المعلومات العربية من الويب, مثل معالجة اللغة الطبيعية, فهرسة الوثائق, عملية العنكبة وكيفية تطبيق عمليات زحف مختصة باللغة العربية, البحث في ملفات PDF العربية ( خاصة محتوى الصحف).

### البحث ...

يعتمد وجود على تقنيات لغوية متقدمة تضمن الحصول على أفضل النتائج لكلمات البحث.
يقدموجود خدمات بحثية متعددة منها :
البحث في الشبكة العنكبوتية , البحث في مفضلة وجود الاجتماعية ومفضلتك الخاصة .

### المفضلة ...

مفضلة وجود هي مفضلة اجتماعية توفر للمستخدمين خدمة التخزين و المشاركة و البحث لصفحات الويب المتميزة. نهدف من خلال المفضلة الإجتماعية الى تجميع كل ما هو مفيد و متميز في الويب العربية ذلك لنضمن افضل النتائج لعملية البحث.

مع مرور الوقت ستمكنا المفضلة الإجتماعية من الحصول عدد كبير من صفحات الويب العربية في مختلف الموضوعات التي تهم المستخدم العربي. بناء على ذلك سوف نوفر للمستخدم العربي إمكانية البحث في هذا المخزون الغني من صفحات الويب العربية, وبذلك نحصر نتائج البحث بصفحات الويب التي وجدها الناس قيمة. من جهة أخرى ستمكنا المفضلة الاجتماعية من الحصول على مستودع جيد لأفضل ما في الويب العربية.

الى جانب ذلك , فان المفضلة الإجتماعية سيكون لها دور في ترتيب الصفحة (Page Ranking),إذ مع الوقت فإن عدد المرات التي تم فيها حفظ صفحة معينة من قبل المستخدمين سيؤثر على ترتيب تلك الصفحة مقارنة بغيرها من الصفحات. وبذلك يكون للعامل الإنساني ألأثر في ترتيب صفحات الويب.

### عربي ...

عربي هي حزمة من التقنيات الخاصة باللغة الطبيعية التي تسخر كل من الذكاء الاصطناعي, وخصائص اللغة, والإحصائيات, بما يسمح لمحرك البحث فهم الجوانب المعقدة في اللغة العربية.

حزمة عربي تشمل:

- مصحح املائي آلي: يقوم المصحح ألإملائي من عربي باكتشاف وتصحيح الأخطاء الإملائية العربية. و يقترح ايضا مجموعة من الكلمات التي قد تكون مناسبة لاستبدال الكلمات المغلوطة.
- مصنف آلي: يقوم المصنف الالي من عربي بتنظيم المحتوى العربي لصفحات الويب و تصنيفه إلى شجرة موضوعات منطقية حتى الان تحتوي شجرة التصنيف من عربي على ثمانية موضوعات رئيسية هي: سياسة , أديان, تكنولوجيا, علوم طبيعية, رياضة, طب و صحة, فلك , تجارة و إقتصاد.
- تحليل صرفي لغرض استرجاع المعلومات: يقوم المحلل الصرفي من عربي بتحليل الكلمة العربية لاستخراج مشتقاتها, بحيث لايقتصر البحث على الكلمة ذاتها بل ومشتقاتها ايضا و ذلك للحصول على افضل النتائج الممكنة من عملية البحث.
- محدد لغة: تقوم هذه التقنية من عربي بتحديد ما إذا كان النص المدخل عربي او غيرذلك, الغرض من محدد اللغة هو تمكين محرك البحث وجود من تجميع محتوى الويب باللغة العربية, فكان محدد اللغة وسيلتنا للتمييز بين محتوى الصفحات العربية وتلك غير العربية من اللغات الاجنبية و اللغات التي تستخدم الاحرف العربية كالفارسية وغيرها.
- استخلاص جذور الكلمات: تقوم هذه التقنية من عربي بإعادة الكلمات العربية الى جذورها اللغوية.

### قاعدة كلمات ...

هي قاعده بيانات تحتوي على اغلب الكلمات العربية في الويب و تكرارها, حتى الان لدينا ما يقارب نصف مليون كلمة عربية مختلفة, تشكل هذه الكلمات القاعدة التي تعتمد عليها معظم تقنيات المعالجة الطبيعية للغة العربية. حتى الآن تحتوي هذه القاعدة على (568,106) كلمة عربية.

### ملفات "PDF" العربية ...

كان لنا وقفةخلال هذا المشروع مع ملفات PDF العربية , كيف بإمكاننا معالجتها و من ثم فهرسة محتواها لجعلها قابلة للبحث من قبل محركات البحث, ذلك لأدراكنا لأهمية محتوى ملفات PDF العربية .
تم معالجة كل من ملفات PDF العادية (ملفات أنشئت في بيئة ويندوز) إذ ان أصل المشكلة في هذه الملفات يكمن في أن اللغة العربية تكتب من اليمين الى اليسار بعكس معظم لغات العالم. حيث أن تحويل ملفات PDF العربية الى ملفات نصية باستخدام المحولات النصية ينتج عنه ملفات نصية معكوسة, تم معالجة هذه المشكلة بعكس النص مرة أخرى للحصول على نص عربي صحيح جاهز للفهرسة. أما فيما يتعلق في محتوى الصحف فالتحدي هنا هو من نوع آخر . يكمن في أن معظم ملفات PDF تنشأ باستخدام برامج مثل InDesign أو PageMaker في بيئة ماك , حيث يتم تمثيل الأحرف العربية في بيئة ماك بشكل مغاير عن تلك التي تمثل بها في بيئة ويندوز, نتيجة لهذا ألاختلاف فإن تحويل هذه الملفات الى ملفات نصية ينتج عنه ملفات نصية غير مقروءة.
حتى الان , قد تمكنا من تطوير خوارزميه تمكنا من قراءة النص من ملفات PDF ولكن مع بعض الاخطاء. نعمل حاليا على تحسين الناتج للأفضل.

# المشروع حصل على المركز الأل في مسابقة أعمال طلبة تكنولوجيا المعلومات في الوطن العربي للعام ٢٠٠٨

## جامعة بيرزيت تحصل على المركز الاول في مسابقة أعمال طلبة تكنولوجيا المعلومات في الوطن العربي

بيرزيت- حصلت جامعة بيرزيت على المرتبة الاولى في مسابقة معرض أعمال طلبة تكنولوجيا المعلومات في الوطن العربي، والذي عقد في جامعة البتراء الأردنية بالتعاون مع اتحاد الجامعات العربية والامانة العامة لجمعية كليات الحاسبات والمعلومات، بمشاركة ١٨ جامعة عربية.

وأكد عميد كلية تكنولوجيا المعلومات د. عدنان يحيى على أهمية مشاركة الطلبة في هذه المسابقة خاصة وأنه أتيحت الفرصة لطلبة كلية تكنولوجيا المعلومات لعرض مهاراتهم امام العاملين والمعنيين من الشركات العاملة في هذا القطاع. وأوضح يحيى ان مشروع «وجود محرك بحث عربي» والذي نفذه الطلبة علي الصالحي وأنوار حثناوي وميرنا فواضلة من خريجي دائرة هندسة أنظمة الحاسوب للفصل الدراسي الثاني ٢٠٠٧-٢٠٠٨، حصل على المركز الاول من بين ٥٠ مشروعاً تم تقديمها، وذلك بعد تقييمه من لجنة متخصصة. مشيراً الى ان هذا المشروع هو مشروع واعد يحمل معنى التطور والحداثة.

وكانت قد شاركت في هذه المسابقة العديد من الجامعات الفلسطينية والأردنية والمصرية واليمنية والكويتية، وقد حظي باهتمام المؤسسات والشركات العاملة في قطاع تكنولوجيا المعلومات المحلية والأقليمية والعالمية.

**From:** Adnan YAHYA
**Sent:** Monday, July 28, 2008 7:38 PM
**To:** AbdLatif ABU HIJLEH
**Cc:** DEAN.IT
**Subject:** الفوز بالمرتبة الأولى لأعمال طلبة تكنولوجيا المعلومات في الوطن العربي

عزيزي د. عبداللطيف،

يسعدني أن أعلمكم أن خريجي الفصل الثاني 2007-2008: علي الصالحي وأنوار حثناوي قد مثّلا جامعة بيرزيت في مسابقة معرض أعمال طلبة تكنولوجيا المعلومات في الوطن العربي والتي عقدت في جامعة البتراء في الأردن بمشاركة 18 جامعة عربية قدمت حوالي 50 مشروعا، وقد احتل مشروع تخرج طلبة كلية تكنولوجيا المعلومات في جامعة بيرزيت: "**وجود:محرك بحث عربي" المرتبة الأولى مع جائزة مالية مقدارها 2000 دينار أردني مع فرص للتواصل مع الشركات العاملة في حقل التخصص.**

العمل الفائز هو عبارة عن مشروع التخرج للطلاب علي الصالحي، أنوار حثناوي وميرنا فواضلة من دائرة هندسة أنظمة الحاسوب.

د. عدنان يحيى
عميد كلية تكنولوجيا المعلومات

## من خلال مشروع محرك بحث عربي أعده ثلاثة من طلبتها

## «بيرزيت» تفوز بالمرتبة الأولى في معرض أعمال طلبة تكنولوجيا المعلومات في الجامعات العربية

رام الله- خاص بـ«القدس الرقمي»- فاز مشروع أعده ثلاثة طلبة من جامعة بيرزيت بالمرتبة الأولى في معرض أعمال طلبة تكنولوجيا المعلومات في الجامعات العربية الذي نظمته جامعة البتراء الأردنية مؤخرا في العاصمة الأردنية عمان بالتعاون مع اتحاد الجامعات العربية والأمانة العامة لجمعية كليات الحاسبات والمعلومات.

وشارك في المعرض الرابع من نوعه الذي تنظمه جامعة البتراء الأردنية والأول من نوعه على المستوى العربي بعد أن كان محصورا في الجامعات الأردنية ١٨ جامعة عربية مثلت دول :فلسطين،الأردن،مصر ،اليمن،والكويت حيث قدمت ٥٠ مشروعا من تصميم طلبة تكنولوجيا معلومات في هذه الجامعات.

والمشروع الفائز بالمرتبة الأولى على الصعيد العربي هو مشروع تخرج لثلاثة طلبة من دائرة هندسة أنظمة الحاسوب في كلية تكنولوجيا المعلومات التابعة لجامعة بيرزيت وهم: علي الصالحي، أنوار حثناوي، وميرنا فواضلة وبإشراف د. عدنان يحيى أستاذ هندسة أنظمة الحاسوب وعميد كلية تكنولوجيا المعلومات في الجامعة.

وأطلق على المشروع الفائز اسم « وجود» وهو عبارة عن محرك بحث و مفضلة إجتماعية عربية،من خلاله يتم الحصول على محرك بحث ذكي متعدد اللغات مع التركيز على اللغة العربية. ويعنى هذا المشروع ببناء هيكل جديد لنظام استرجاع المعلومات لمساعدة الناس على العثور على المعلومات التي يحتاجونها بطريقة تتسم بالكفاءة.