Information Retrieval System for Digital Libraries Using Fuzzy Logic

Ghayda A.A. Al-Talib

Nawar Abdul Ghani Sultan

ghaydabdulaziz@uomosul.edu.iq

nawarabd8@gmail.com

College of Computer Sciences and Mathematics
University of Mosul

Received on:4/10/2010

Accepted on:10/11/2010

ABSTRACT

This research is one of the important steps towards processing the most significant challenge in digital libraries and web, which is computing document's rank, its importance, and its relevance to the user information need. This is achieved through the utilization of fuzzy logic high potential capabilities in dealing with such sort of problems and providing notable flexibility for the user to get his favorite subjects.

The research is concerned with designing and implementing a proposed information retrieval system, called FIRS (Fuzzy Information Retrieval System). This system is developed to deal with huge database, which contains different text file types and sizes. This database is distributed over a collection of server computers, connected with the intranet network that is dedicated for this system.

The system has the ability for mining of data available in this database and retrieving the useful information that corresponds with the user need. This is accomplished through applying the proposed algorithms for indexing process, document's rank computations, generating keywords process, and finally, displaying information retrieval results. The proposed system gives high quality results comparing with other information retrieving algorithms.

Keywords: Information Retrieval, Digital Library, Fuzzy Logic

نظام لاسترجاع المعلومات في المكتبات الالكترونية باستخدام المنطق المضرب

نوار عبد الغني سلطان

غيداء عبد العزيز الطالب

nawarabd8@gmail.com

ghaydabdulaziz@uomosul.edu.iq

كلية علوم الحاسوب والرياضيات

جامعة الموصل

تاريخ قبول البحث: 2010/11/10

تاريخ استلام البحث: 2010/10/4

الملخص

يمثل هذا البحث أحد الخطوات الهادفة لمعالجة واحدة من أهم التحديات على الويب والمكتبات الرقمية، ألا وهي حساب رتبة الوثيقة، أهميتها، ومدى صلتها بطلب المستخدم وتلبية احتياجاته للمعلومات، وذلك من خلال الاستفادة من الإمكانيات الكبيرة للمنطق المضرب في التعامل مع مثل هذا النوع من المشاكل، وتقديم المرونة العالية للمستخدم في توضيح المواضيع والمجالات التي يهتم بها.

اشتمل هذا البحث على تصميم وتنفيذ نظام مقترح لاسترجاع المعلومات، وسمي FIRS (Fuzzy Information Retrieval System). صُمم هذا النظام لكي يتعامل مع قاعدة بيانات ضخمة وموزعة على مجموعة من

الحاسبات (الخوادم servers) المرتبطة بشبكة الانترنت المصممة خصيصاً لعمل النظام، والتي تضم أنواع وأحجام مختلفة من الملفات النصية.

وتتطور إمكانية النظام في التنقيب عن البيانات الموجودة في قاعدة البيانات واسترجاع المعلومات المفيدة منها والتي تلبي حاجة المستخدم بشكل جيد، من خلال تطبيق الخوارزميات المقترحة في عمليات الفهرسة وحساب رتبة الوثائق وفي عمليات توليد الكلمات المفتاحية وعرض نتيجة عملية الاسترجاع، والتي أظهرت جودة عالية عند حساب النتائج مقارنةً مع غيرها من الخوارزميات.

الكلمات المفتاحية: استرجاع المعلومات، المكتبات الالكترونية، المنطق المضرب

1. المقدمة

يلعب الحاسوب الإلكتروني دوراً مهماً في تصميم وبناء نظم المعلومات الحديثة فهو يحقق لنظام المعلومات مزايا السرعة والدقة والثقة والصلاحية، ويترتب عليها جميعاً الكفاءة العالية في الأداء. كما يقوم الحاسوب بأجراء العمليات الحسابية المعقدة والتي يصعب تنفيذها يدوياً بالإضافة إلى القدرة الفائقة على تخزين كم هائل من المعلومات بطريقة منظمة بحيث يسهل استرجاعها في أوقات ضئيلة للغاية كما يستطيع الحاسوب الإلكتروني إنجاز كافة المهام الأخرى التي يقوم بتنفيذها نظام المعلومات ومنها تحقيق أمن وسلامة البيانات والضمان الكامل ضد فقدها أو تلفها من خلال المستفيدين.

فإن التطور المذهل في عدد الأنظمة الآلية المخصصة للمكتبات على مختلف أنواعها يبين أهمية هذه التقنية بالنسبة للمكتبات ولقد أجريت العديد من المسوحات "Surveys" لأعداد الأنظمة الآلية في المكتبات وخاصة في الولايات المتحدة الأمريكية. حيث أجري مسح عام 1964 ثبت فيه أن هناك 25 نظاماً آلياً يعمل في المكتبات في أمريكا. وفي عام 1971 أجري مسح آخر يعرف بمسح لارك (Lark Survey) تم فيه إحصاء (1366) نظام آلي للمكتبات (ما بين أنظمة متكاملة أو أجزاء من أنظمة في (506) مكتبة). [2]

وفي عام 1984 أجري مسح آخر ثبت فيه أنه يوجد حوالي (30000) ثلاثين ألف نظام آلي خاص بالمكتبات (إن أغلب هذه الأنظمة تعمل على الحاسوب الشخصي PC)، ويلاحظ المدى الذي وصل إليه الرقم خلال 13 سنة هي الفارق الزمني بين إحصاء لارك 1971 وإحصاء عام 1984، حيث تضاعفت أعداد الأنظمة الآلية للمكتبات بنسبة 59 ضعفاً، والحقيقة أن ذلك يعود إلى سببين رئيسيين هما:

1. الاحتياجات الفعلية للمجتمعات الحديثة المتمثلة بضرورة السيطرة على الكم الهائل والمتنامي في المعلومات المطلوب تداولها من قبل هذه المجتمعات والحصول على المعلومات اللازمة منها بسرعة ودقة وفاعلية. والإفادة منه في التنمية الاقتصادية والاجتماعية والثقافية المنشودة.

2. الإمكانيات الكبيرة التي توفرها الحواسيب الآلية وتكنولوجيا الاتصالات الحديثة المتمثلة في الطاقات التخزينية الكبيرة وسرعة المعالجة التي أصبحت تقاس بأجزاء بسيطة من الثانية وتضاعفت مئات المرات سرعة تبادل البيانات وانخفضت بشكل كبير تكلفة هذه العمليات، إضافةً إلى إمكانية المعالجة عن بعد واستخدام الحواسيب وبنوك المعلومات وغيرها.

إن تاريخ استخدام النظم الآلية في المكتبات يعود إلى عام 1935 وهو العام الذي شهد إدخال أول آلة معالجة بيانات في المكتبات حيث قامت جامعة تكساس باستخدام أجهزة بطاقات مثقبة Punched Card في نظام الإعارة Circulation System، ثم تلتها مكتبة بوسطن العامة في استخدام البطاقات المثقبة لتحليل بعض إحصائيات

التزويد وتوالت النظم الآلية بعد ذلك حيث استخدمت المكتبات الرقمية Digital Libraries، وكان أول من طالب باستخدام هذا النوع من الأجهزة في المكتبات كلا من ملفن ج. فواخت Melvin J. Voight المسؤول عن مكتبة جامعة كاليفورنيا ومعه كلاي ل. بري Clay L. Perry من مركز الحاسوب في نفس الجامعة عام 1962 حيث كان مشروعهما التجريبي هو تحويل تسجيلات عدد (700) سلسلة إلى الشكل المقروء آلياً بالإضافة إلى طباعة قائمة بالأعداد شهرياً مع قائمة كاملة بكل ما تملكه المكتبة.

ويلاحظ أن أغلب الأنظمة الآلية التي ظهرت كانت أجزاء من أنظمة أي أنها لم تكن أنظمة متكاملة Integrated Systems) يمكن أن تضم جميع عمليات المكتبة في آن واحد ولكنها كانت تتعامل مع جزء واحد فقط من عمليات المكتبة مثل الفهارس أو الإعارة ولكن مع عام 1961 قامت المكتبة الطبية الوطنية National Library of Medicine بالعمل مع مشروع مدلاز Medical Literature Analysis and Retrieval System ومراجعة وظائف النظام في محاولة لمكنة كل وظائف المكتبة بالإضافة إلى إجراء عمليات البحث الببليوغرافي وإصدار كشاف Meoicus Index وكذلك عمليات الفهرسة الآلية والاستعارة الآلية والمساعدة في الاقتناء وضبط الدوريات وبالتالي ظهور أول نظام آلي متكامل في المكتبات عام 1966. ولم يقتصر الأمر على ذلك فمن التغيرات والتطورات المثيرة للاهتمام في مجال المكتبات والمعلومات ذلك التطور الذي حدث على الخدمات التي تقدمها المكتبات حيث قامت وكالة الفضاء ناسا Nasa باختبار أول نظام للبحث الانتقائي للمعلومات SDI يعمل على الحاسوب الآلي، حيث يقوم المستفيد بتحديد الموضوعات التي يرغب في الاطلاع عليها ويقوم النظام الآلي بالمقارنة بين موضوعات المقالات ووصفات المستفيد الموضوعية ويقوم بإمداده بقائمة من المقالات تطابق اهتماماته التي قام بتحديدتها سابقاً [2].

إن كانت المكتبات العامة من بين المؤسسات الأولى التي تبنت أنظمة استرجاع المعلومات. وطُورت أنظمة استرجاع المعلومات المستخدمة في المكتبات العامة في البداية من قبل المؤسسات الأكاديمية وبعد ذلك من قبل الباعة التجاريين، وفي السنوات الأخيرة تم انجاز تطورات كبيرة في مجال تقنية المعلومات وخصوصاً بعد ظهور الويب والتحسين المستمر في أداء محركات البحث مما جعل المكتبات الإلكترونية الأكثر رواجاً وفي جميع التخصصات. وتضمن الجيل الأول من أنظمة المكتبات آليات وتقنيات بدائية جعلت البحث يعتمد على اسم المؤلف والعنوان.

أما في الجيل الثاني، فقد اشتملت وظيفة البحث على:

- 1- عنوان الموضوع.
- 2- الكلمات المفتاحية.
- 3- وسائل استفسار كثيرة التعقيد.
- وكان التركيز في الجيل الثالث على:
- 1- الواجهات الرسومية.
- 2- النماذج الإلكترونية.
- 3- خصائص النص التشعبي.
- 4- معمارية النظام المفتوح: هو ذلك النظام الذي له القدرة على التغيير المستمر لعملياته وبنيته الداخلية [10].

2. هدف البحث

تُعتبر عملية استرجاع المعلومات ذات أهمية كبيرة في تقديم الخدمات ونقل المعلومات ومحتويات صفحات الويب العلمية، وكذلك محتويات الوثائق الموجودة في المكتبات الرقمية إلى المستخدمين من خلال الوسائل والأدوات البحثية المنهجية السهلة والسريعة التي تحقق استرجاع للمعلومات للاستفادة منها بالشكل الذي يلبي احتياجاتهم الفعلية والقناعة التامة في مجال دراستهم أو بحوثهم.

وعليه نحاول هنا الإجابة على التساؤلات التالية والتي تشكل المحور الرئيسي لإطار البحث:

- 1- ما هي آلية العمل في تقنية أنظمة استرجاع المعلومات؟ وما هي الأجزاء الرئيسية التي تتكون منها؟
- 2- كيف نستطيع الحصول على نتائج جيدة تلبي حاجة المستخدمين؟
- 3- ما هي أسباب تصاعد نتائج الاسترجاع غير ذات العلاقة مقارنة مع نتائج أخرى ذات علاقة؟
- 4- كيفية تحسين وتطوير عمليات واستراتيجيات الاسترجاع لكي تعطي نتائج أكثر دقة للمستخدمين؟
- 5- ما هي الخوارزميات المستخدمة من أجل حساب رتب الوثائق وأيهما تعطي نتائج ذات نوعية عالية؟

3. أنظمة استرجاع المعلومات

يعتبر علم استرجاع المعلومات أحد فروع المعرفة القديمة والذي يضاهي قدم علم الحاسوب نفسه، إن لم يكن الأقدم. ويعد هذا التعريف في حقل استرجاع المعلومات من أقدم التعاريف، إن لم يكن الأقدم بينها، وهو الذي ذكره مويرس(1950) حيث عرّف استرجاع المعلومات: "هو اسم العملية أو الطريقة التي تمكّن المستخدم من تحويل حاجته للمعلومات إلى قائمة فعلية من الوثائق الموجودة في المخزن (مثل القرص الصلب) والذي يحوي المعلومات المفيدة للمستخدم "[6].

و يشتمل نظام استرجاع المعلومات بشكل عام على جزئين:

1. الأرشيف النصي: الذي هو عبارة عن مجموعة من الوحدات النصية (والتي غالبا ما يطلق عليها بـ "الوثيقة").
 2. محرك الاسترجاع: يقوم مستخدم النظام بتقديم الطلبات التي تصف نوع الوثائق المطلوبة، ويقوم محرك الاسترجاع بمطابقة الطلبات مع الوثائق الموجودة في الأرشيف النصي، ومن ثم يرجع للمستخدم جزء من مجموعة الوثائق الكلية والتي تعطي أفضل تطابق مع حاجة المستخدم [4].
- و إن المتطلبات الرئيسية لهذه الأنظمة هي:

1. يجب أن يكون المستخدم قادر على إدخال أي كلمة من اللغات الطبيعية، أية عبارة، أو جملة إلى النظام بدون الحاجة إلى الدخول في عمليات الفهرسة والبحث، وهذا له دلالة ضمنية على أن أنظمة استرجاع المعلومات للنص الكامل تقوم بفهرسة كل كلمة في الوثيقة بشكل آلي.
2. النظام يجب أن يُرتب الوثائق المسترجعة حسب احتمالية فائدتها لطلب المستخدم.
3. يجب أن يدعم النظام إعادة صياغة آلية لجمل البحث بالاعتماد على المعلومات السابقة للمستخدم (التغذية العكسية)[5].

واستنادا على ما سبق ذكره فيمكن اعتبار نظام استرجاع المعلومات احد حقول علم الحاسوب والذي يهدف إلى تخزين كمية كبيرة من المعلومات والسماح بالوصول السريع إليها، وهذه المعلومات يمكن أن تكون من أي نوع: نصية، سمعية، أو مرئية [9].

4. الحسابات غير العددية واسترجاع المعلومات

عُرف مصطلح الحسابات غير العددية (Soft Computing-SC) من قبل Zadeh [7]، وهو يشير إلى التعاون بين الأساليب المفيدة لحل المشاكل التي تتطلب نوعاً من الذكاء بعيداً عن الحسابات التقليدية. يُقدم SC مجموعة من التقنيات الملائمة لمعالجة الغموض، عدم الموضوعية، وعدم الدقة في العديد من المشاكل القائمة. إنَّ استرجاع المعلومات يهدف إلى نمذجة، تصميم، وتنفيذ نظم قادرة على توفير خدمات سريعة وفعالة في الوصول إلى المحتوى القائم على كميات كبيرة من المعلومات. ويهدف أيضاً إلى تقدير مدى علاقة عناصر المعلومات بحاجة المستخدم الممثلة بعبارة الاستفسار. وهذه مهمة صعبة جداً ومعقدة نظراً لما يتخللها من انعدام في الموضوعية، الغموض، وعدم الدقة.

ويتضمن الـ SC مجموعة أساليب مختلفة لحل مثل هذه المشاكل، كالمنطق المضطرب، الخوارزميات الجينية، الشبكات العصبية، المجاميع التقريبية (rough sets)، وشبكات Bayesian [8]. حيث مع البدايات الأولى لظهور نظرية المجاميع المضطربة من قبل L. A. Zadeh في عام 1965، تم انجاز تقدم واضح في الاستفادة من مفاهيم المجاميع المضطربة في حقول متعددة من العلوم المرئية والحسية. حيث أُستخدمت المجاميع المضطربة وبشكل مُستفيض في علم الرؤية الحاسوبية (computer vision) والذكاء الاصطناعي. فالمجاميع المضطربة تتعامل مع الغموض والإبهام المطور في نظام الاستيعاب البشري، وتوفر بنية نظام مُمتازة لوصف، تحليل، وتفسير الأحداث الغامضة والمشكوك فيها. حيث يُعد نظام الرؤية البشري أساساً كنظام مضطرب، وذلك لان بإمكاننا فهم وتفسير العالم المرئي الغامض من حولنا [3]. و استرجاع المعلومات يستفيد من المنطق المضطرب في عمليات تكتل المعلومات، استخلاص النصوص، نماذج لغة الاستفسار، وعنقدة الوثائق [8].

5. الكتب الرقمية

يعد الكتاب الرقمي ثورة علمية يمكن أن تغير مسار التاريخ في القرن الحادي والعشرين، أكثر مما غيره اختراع المطبعة، إذ أصبح الكتاب الرقمي (الإلكتروني E.Book) في متناول اليد، سهل النشر، سهل التوزيع والحصول عليه، سهل القراءة والاطلاع، وما الكتاب الرقمي في الواقع إلا مادة علمية، من حروف وصور وتقارير وإحصاءات، محفوظة في نوع ما من أنواع كثيرة من الملفات.

ويختلف الكتاب الرقمي في ميزاته وخصائصه عن الكتاب الورقي في نقاط كثيرة، أهمها [1]:

1. الحصول على الكتاب من غير ثمن أو بئس زهيد جداً.
2. سهولة قراءة الكتاب على الحاسوب، إذ يمكن، في كثير من الحالات، تغيير حجم الحرف، والتحكم في درجة الإضاءة، بما يريح القارئ.
3. الوصول إلى الفكرة المطلوبة في الكتاب، باستعمال أداة "البحث" وبوساطة كلمة واحدة، للوقوف على المواضيع المطلوبة من الكتاب.
4. حفظ ملايين الكتب في حيز محدود جداً في الأقراص الليزرية أو الأقراص الصلبة (Hard) المدمجة داخل الحاسوب أو الأقراص المحمولة المستقلة عنه، مما يتيح حمل ملايين الكتب في السفر والرحلات والانتقال بها من مكان إلى مكان بيسر.
5. سرعة القراءة والتأليف والتوثيق، وتحقيق ثورة علمية ومعرفية.

6. النظام المقترح لاسترجاع المعلومات باستخدام المنطق المضطرب

إن نظام Fuzzy Information Retrieval System (FIRS) مُصمم لكي يتعامل مع قاعدة بيانات ضخمة موزعة على مجموعة من الحاسبات (الخوادم) المرتبطة بشبكة الانترنت، والتي تضم أنواع وأحجام مختلفة من الملفات النصية. وهذا النظام له القدرة على التنقيب عن البيانات الموجودة في قاعدة البيانات واسترجاع المعلومات المفيدة منها والتي تُلبّي حاجة المستخدم بشكل كفوء.

وسوف نتناول بدايةً عملية تحليل الوثائق النصية وطرق التعامل معها.

7. تهيئة الوثيقة النصية

إن عملية بناء الفهرس تعتبر من العمليات الرئيسية للاسترجاع الكفوء للوثائق وهذه العملية تتطلب خطوتين رئيسيتين:

1. تحليل الوثيقة، يتضمن ذلك تحديد الكلمات المهمة فيها، وترشيحها، وهذا يتطلب تحليل كل وثيقة موجودة في قاعدة البيانات (مثال ذلك: وثائق الويب) وتنظيم هذه الوثائق على شكل عناصر أساسية للوثيقة (العنوان، اسم المؤلف، المصدر) وكيفية تمثيل المعلومات فيها، على شكل جداول، معلومات حرجة في النص، مخططات، رسومات، أو صور. والقرار الذي يجب أن يُتخذ هو تحديد أيّ من المعلومات أو أجزاء الوثيقة سوف تفهرس وأياها لا تدخل في عملية الفهرسة.

2. تحليل الرموز أو استخلاص المصطلحات، والقرار الذي يجب أن يُتخذ هو تحديد أي الكلمات (أو العبارات) يمكن أن تُستعمل كمرجع للوثيقة من أجل التمثيل الأفضل للمحتويات المعنوية للوثائق.

إذن قبل الشروع بعملية الفهرسة نحتاج إلى مجموعة من التحضيرات والإعدادات التي تتم على الوثيقة من أجل تهيأتها لعملية الفهرسة والتي تختلف حسب نوعية الوثيقة النصية المراد فهرستها لاحظ الشكل (1).

الشكل (1): تهيئة الوثيقة

ولغرض تحليل الوثائق النصية تم اقتراح هيكلية بيانات خاصة بالخوارزمية المقترحة، ونحتفظ في هذه الهيكلية بالمعلومات المهمة المتوفرة في الوثيقة والتي سوف تُسهّل عملية الفهرسة وحساب رتبة الوثيقة، وهذه الهيكلية تضمنت جزئين: جزء خاص بالكلمات ويضم معلومات عن كل كلمة موجودة في الوثائق والجزء الآخر يخص الوثائق التي توجد فيها الكلمة. ويوضح الشكل (2) هيكلية البيانات المقترحة والتي سوف يتم التطرق إليها لاحقاً.

الشكل(2): هيكلية البيانات الخاصة بالكلمات

8 . معمارية النظام

معمارية النظام تتكون من الهيكل الموضح في الشكل(3)، حيث نلاحظ في هذه المعمارية وجود خمسة حواسيب تمثل الخوادم التي تحوي قاعدة البيانات، حيث أن كل خادم يحوي قاعدة البيانات الخاصة به والتي سوف تقهرس لاحقاً. نلاحظ أيضاً أن أحد هذه الخوادم يمثل الخادم الرئيسي (Main Server) والذي سوف يضم في النهاية جميع ملفات الفهرسة التابعة لكل الخوادم. وهناك عدد غير محدد من حاسبات المستخدمين (Clients) مرتبطة بالشبكة مع الخوادم عن طريق HUB، وذلك لان العمل يتم انجازه على شبكة انترانيت، كما نحتاج إلى جهاز Access Point مرتبط بالشبكة لغرض الاتصالات اللاسلكية لحاسبات مستفيدين أخرى.

الشكل (3) معمارية نظام FIRS

وفيما يخص حاسبات المستفيدين فإنها لا تحتوي على قاعدة البيانات، ولكنها تضم الواجهات Interfaces التي تمكن المستخدمين من الاتصال مع الخادم الرئيسي والذي عن طريقه سوف يتم استرجاع المعلومات التي تلي احتياجاتهم.

9. الخوارزميات المقترحة

بما أن نظام FIRS المصمم في هذا البحث يتكون من برنامجين اثنين، أحدهما يعمل على حاسبات الخوادم (servers) والآخر على حاسبات المستفيدين (clients)، كان لابد من عمل خوارزميات خاصة بكل برنامج اعتماداً على الوظيفة التي يقوم بها كل منهما، وفيما يأتي عرض موجز لأساسيات عمل كلا من البرنامجين والخوارزميات المعتمدة.

9.1 الخوارزمية المقترحة لعملية الفهرسة

1. يتم في البداية تحديد مسار قاعدة البيانات والتي نحدد من خلالها عدد الملفات الموجودة فيها.
2. تحديد نوع الملف حيث إذا كان الملف من نوع HTML فسوف يتوجب إزالة الأوسمة tags الموجودة ضمن المحتوى النصي للملف، وتتم هذه العملية باستخدام تعبير نظامي regular expression. أما في حالة الأنواع الأخرى من الملفات فلا نحتاج إلى القيام بهذه العملية.
3. استخلاص النص من الملفات، مع استخدام تعبير نظامي يقوم بحذف الرموز الخاصة special characters من النص.
4. بعدها يتم استقطاع الكلمات من النص حسب تعبير نظامي خاص بأحرف اللغة الانكليزية، وكذلك تفحص كل كلمة باستخدام تعبير نظامي آخر خاص بالكلمات المستبعدة التي لا تدخل ضمن عملية الفهرسة.
5. بعد الحصول على الكلمة الصحيحة تخزن معلوماتها في مصفوفة خاصة بمعلومات الكلمات والموضحة هيكلتها في الشكل (2).
6. وقبل خزن المعلومات يتم التأكد من أن الكلمة لم تخزن مسبقاً، فإذا كان مخزونة مسبقاً يتم إضافة معلوماتها إلى القيد الموجود.
7. يتم حساب عدد مرات تكرار الكلمة في الملف باستخدام خوارزمية Boyer-Moore.
8. هذه العملية تتكرر إلى أن تتم فهرسة جميع الملفات وما تحويه من كلمات.

9.2 الخوارزمية المقترحة لعملية حساب الرتبة

تتم عملية حساب الرتبة، وذلك بالإعتماد على عمليات المنطق المضرب، وبالشكل الآتي:

❖ تضبيب المعلومات Fuzzification:

يتم في هذه العملية حساب دوال العضوية التابعة لكل كلمة موجودة في مصفوفة الكلمات، ويوضح الشكل (4) دوال العضوية المستخدمة في هذه العملية، وحسب الصيغ الآتية:

$$MF_count_low = 1 - (1 / (1 + \exp(-((count - (max_count / 2)))))) \quad \dots(1)$$

$$MF_count_high = 1 / (1 + \exp(-((count - (max_count / 2)))))) \quad \dots(2)$$

$$MF_loc_near = 1 - (loc / end_loc) \quad \dots(3)$$

$$MF_loc_far = loc / end_loc \quad \dots(4)$$

$$MF_URL_low = 1 - (1 / (1 + \exp(-((fileURLcount - (\max_no_of_URLs / 2)))))) \dots(5)$$

$$MF_URL_high = 1 / (1 + \exp(-((fileURLcount - (\max_no_of_URLs / 2)))))) \dots(6)$$

MF_count_low: دالة العضوية لعدد التكرار القليل للكلمة في الملف النصي.

MF_count_high: دالة العضوية لعدد التكرار الكثير للكلمة في الملف النصي.

MF_loc_near: دالة العضوية لموقع الكلمة القريب من بداية الملف النصي.

MF_loc_far: دالة العضوية لموقع الكلمة البعيد من بداية الملف النصي.

MF_URL_low: دالة العضوية لعدد التكرار القليل للروابط الموجودة في الملف النصي.

MF_URL_high: دالة العضوية لعدد التكرار الكبير للروابط الموجودة في الملف النصي.

الشكل (4): دوال العضوية

❖ قواعد المعرفة Knowledgebase Rules:

بعد الإنتهاء من عملية التضبيب يتم تكوين قاعدة المعرفة المكونة من المجاميع المضببة (poor, good, excellent) التي تمثل ناتج تنفيذ القواعد وقواعد المنطق المضبب، حيث تُستخدم ثمانية قواعد مع الملفات من نوع HTML أما في حالة الأنواع الأخرى فهناك أربعة قواعد مُستخدمة. وهذه القواعد موضحة في الشكل (5).

1. If (Count is Low) and (Location is Near) then (RANK is Goog) (1)
 2. If (Count is Low) or (Location is Far) then (RANK is Poor) (1)
 3. If (Count is High) and (Location is Near) then (RANK is Excellent) (1)
 4. If (Count is High) and (Location is Far) then (RANK is Goog) (1)
-
1. If (Count is Low) and (Location is Near) and (URLs is Low) then (RANK is Poor) (1)
 2. If (Count is Low) or (Location is Near) or (URLs is High) then (RANK is Goog) (1)
 3. If (Count is Low) or (Location is Far) or (URLs is Low) then (RANK is Poor) (1)
 4. If (Count is Low) or (Location is Far) or (URLs is High) then (RANK is Poor) (1)
 5. If (Count is High) or (Location is Near) or (URLs is Low) then (RANK is Goog) (1)
 6. If (Count is High) or (Location is Near) or (URLs is High) then (RANK is Excellent) (1)
 7. If (Count is High) or (Location is Far) or (URLs is Low) then (RANK is Poor) (1)
 8. If (Count is High) or (Location is Far) or (URLs is High) then (RANK is Goog) (1)

الشكل(5): قواعد المنطق المضبيب

❖ التجميع Aggregation:

بعد تنفيذ القواعد السابقة تتم عملية تجميع النتائج المُستحصلة باستخدام تجميع OR-MAX كما موضح في الشكل(6).

حساب قيمة الرتبة باستخدام طريقة centroid

الشكل(6): تنفيذ القواعد وعملية التجميع وحساب قيمة الرتبة

❖ Defuzzification:

يتم حساب قيمة الرتبة Rank لكل ملف تابع للكلمة بالإعتماد على طريقة النقطة الوسطى centroid كما في الشكل(6)، وحسب الصيغة الآتية:

$$RANK = \frac{\sum Xi \cdot \mu(X)}{\sum \mu(X)} \quad \dots(7)$$

إذا كانت الكلمة موجودة في العنوان وفي محتوى الملف فإن قيمة الرتبة سوف تصبح مضاعفة.

9.3 الخوارزمية المقترحة لعملية البحث والاسترجاع

عندما يقوم المستخدم بإدخال عبارة الإستفسار ويختار الزر Search تبدأ عملية فحص عبارة الإستفسار للتأكد من وجودها. هذه العبارة يتم إدخالها عن طريق SmartTextBox وهو صنف يساعد المستخدم في عملية تصحيح الأخطاء المطبعية التي قد تتواجد في عبارة الإستفسار، والشكل (7) يوضح المخطط الانسيابي لعملية البحث.

الشكل (7): المخطط الانسيابي لعملية البحث

بعد ذلك تُستدعى دالة مخصصة لعمل تهيئة لمصفوفة معلومات الكلمات المفتاحية، وتشمل هذه المعلومات الكلمة المفتاحية، عدد الملفات التي تضم هذه الكلمة، عدد الخوادم التي تضم معلومات عن هذه الكلمة، ومعلومات الملفات التي تظهر فيها الكلمة. كما تهيئ مصفوفة أخرى لحفظ معلومات الملفات المسترجعة، كما هو موضح في الشكل (8).

الشكل (8): هيكلية مصفوفة الكلمات المفتاحية

يتم إستخلاص الكلمات المفتاحية من عبارة الإستفسار وذلك باستخدام أكثر من تعبير نظامي لحذف الأوسمة tags والرموز الخاصة، ومن ثم الحصول على العبارة بالإعتماد على تعبير نظامي خاص بأحرف الأبجدية الإنكليزية. بعدها يتم تقطيع العبارة إلى الكلمات المفتاحية وتخزينها في المصفوفة الخاصة بها.

إذا كان صندوق الاختيار CheckBox لمتعلقات الكلمة Allrelevant مؤشر فسوف يتم توسيع الكلمات المفتاحية الموجودة ضمن عبارة الإستفسار، وذلك باستخدام طريقة بحث خاصة بالمنطق المضرب، والتي تمثل خوارزمية بحث مبسطة تستند في مبدئها على Levenshtein-Distance وتستخدم للحصول على تقريب للتطابق بين السلاسل الحرفية. وتعطي Levenshtein-Distance مدى التطابق بين سلسلتين حرفية.

تستمر عملية البحث إلى أن يكتمل فحص جميع الكلمات المفتاحية، كما تتم عملية مقارنة الكلمات المضافة إلى القائمة الجديدة لأجل التأكد من أنها لم تظهر مسبقاً في القائمة.

ويتم تعيين القيم إلى المصفوفة الخاصة بالملفات ذات العلاقة بالكلمات المفتاحية وإكمال ملء مصفوفة الكلمات المفتاحية، ويُنجز هذا الأمر سواءً تمت عملية توسعة الكلمات أم لم تتم.

10. تنفيذ نظام FIRS و مناقشة النتائج

إن نظام FIRS له أهمية كبيرة، لما يتمتع به من سرعة ودقة في عمليات الفهرسة والبحث والاسترجاع، وسهولة التعامل معه من قبل المستخدمين. فقد تم تنفيذ و برمجة النظام باستخدام اللغة البرمجية C# ضمن مجموعة لغات البرمجة Visual Studio 2008 بإطار عمل ذي إصدار .NET Framework Ver. 3.5. وتم الإعتماد على الصفوف التي تدعم فكرة المقابض في برمجة الشبكات ضمن نظام التشغيل ويندوز windows sockets. وتدعم لغة C# كتابة برامج ذات خطوط تنفيذية متعددة multithreading وهي من متطلبات تصميم وبرمجة وحدات النظام المصمم، ويعتمد النظام على نموذج العميل/الخادم.

لتطبيق النظام المقترح يجب تحقيق ما يأتي:

- 1- توفير حاسبة ذات مواصفات خاصة حيث نحتاج حاسبة تكون فيها الذاكرة الرئيسية RAM بحجم 8GB أو أكثر ونظام تشغيل 64bit لكي يكون قادر على التعامل مع هذه الذاكرة، ومعالج 2.9GHz، والقرص الصلب لا يقل عن 320GB. وتُسمى هذه الحاسبة بالخادم الرئيسي.
- 2- تهيئة قاعدة البيانات المكونة من الوثائق النصية على القرص الصلب الخاص بحاسبة الخادم الرئيسي.

- 3- تحميل NET Framework Ver. 3.5 على حاسبة الخادم الرئيسي.
- 4- تحميل برنامج الخادم الرئيسي التابع لنظام FIRS الخاص بعملية الزحف والفهرسة على حاسبة الخادم الرئيسي.
- 5- ربط حاسبة الخادم الرئيسي مع جهاز wireless access point لغرض تهيئة شبكة الانترنت.
- 6- بعد ذلك يتم تنفيذ برنامج الخادم الرئيسي لغرض الحصول على ملفات الفهرسة التي نحتفظ بها على القرص الصلب الخاص بحاسبة الخادم الرئيسي.
- بعدها نحمل برنامج المستخدم (الجزء الثاني من النظام المقترح) على حاسبات المستخدمين المرتبطة بشبكة الانترنت من اجل توفير خدمة البحث والاسترجاع. يوضح الشكل(9) واجهة برنامج الخادم الرئيسي، ويوضح الشكل(10) واجهة برنامج المستخدم.

الشكل (9): واجهة برنامج الخادم الرئيسي

الشكل (10): واجهة برنامج المستخدم.

تم إختبار النظام المدعوم بالمنطق المضطرب على مجموعة من عبارات الإستفسار عددها عشرة، وتم استخدام العبارات نفسها في نظام لا يدعم المنطق المضطرب، ونظام آخر يعتمد البحث في واجهة سطح المكتب وهو برنامج

يقوم بعملية الفهرسة والبحث بالاعتماد على المُفهرس lucene المكتوب بلغة C#, وتم حساب معدل مقاييس الاستدعاء والدقة لجميع عبارات الاستفسار وكانت النتائج كما في الجدول (1) والمخططات البيانية في الشكل (11). ثم جمعت نتائج الخوارزميات الثلاث في مخطط بياني واحد موضح في الشكل (12) من أجل توضيح جودة وكفاءة الخوارزمية المقترحة باستخدام المنطق المضرب. وتم تمثيل نتائج الخوارزميات الثلاث بواسطة المدرج التكراري في الشكل (13)، والخوارزمية المقترحة وبرنامج Lucene المُعتمد عالمياً في المدرج التكراري الموضح في الشكل (14).

جدول (1): نتائج عبارات الاستفسار العشر

Recall	Precision بدون استخدام المنطق المضرب	Precision باستخدام برنامج Lucene	Precision باستخدام المنطق المضرب
10%	76.25%	90%	100%
20%	75.13%	79.98%	100%
30%	76.4%	79.5%	90%
40%	68.4%	54.37%	83.7%
50%	64.8%	50.38%	76.4%
60%	65.9%	47.82%	73.9%
70%	62.7%	44.57%	72.9%
80%	59.6%	37.96%	64.03%
90%	56.2%	33.73%	61.1%
100%	53.8%	21.82%	53.4%

الشكل (11): معدل مقياسي الاستدعاء والدقة لعبارات الإستفسار العشر أ- باستخدام المنطق المضرب، ب- بدون استخدام المنطق المضرب، ج- باستخدام برنامج Lucene

الشكل (12): المخطط البياني لنتائج الخوارزميات الثلاثة والذي توضح جودة وكفاءة الخوارزمية المقترحة في استرجاع المعلومات

50

الشكل(13): المدرج التكراري لنتائج الخوارزميات الثلاثة

الشكل(14): المدرج التكراري لنتائج الخوارزمية المقترحة وبرنامج Lucene

11. الاستنتاجات

لقد توصلنا في هذا البحث إلى الاستنتاجات الآتية:

- 1- إن استخدام نظرية المجموعات المضببة في استرجاع المعلومات يمتلك على الأقل واحدة من الخصائص الآتية مقارنةً مع الطرق الكلاسيكية:
 - تكون علاقات الصلة المضببة أكثر تعبيراً من القيم الحقيقية الواضحة، كما تكون عملية بنائها أكثر واقعية.
 - في حالة كون عدد الوثائق كبير جداً فإن المنطق المضبب يوجه المستخدم إلى الوثائق ذات الأهمية الأكثر لحاجته.
 - الإستفسار المضبب يوفر للمستخدم الفرصة الكبيرة للوصول إلى المواضيع التي يهتم بها.
- 2- الخوارزمية المقترحة التي طبقت لغرض حساب رتبة الوثيقة أعطت دقة ومرونة عالية.
- 3- وفرت عملية استخدام التعابير النظامية (Regular Expression) سهولة في عملية تحليل الوثيقة وفي تحديد الرموز الخاصة والكلمات المستبعدة.
- 4- استخدام خوارزمية Boyer-Moore ساعد في حساب تكرار الكلمات في الوثائق بشكل سريع ودقيق.
- 5- أدت الخوارزمية المقترحة لعملية الفهرسة، وإن كانت تتطلب وقتاً للتنفيذ، إلى جعل عملية البحث والاسترجاع تتم بشكل أسرع.
- 6- عندما يُدخل المستخدم استفساراً للنظام فإن النظام يقوم بفحص الاستفسار وتصحيح عبارة الاستفسار أثناء عملية الإدخال وبالتالي يضمن الحصول على نتائج دقيقة دون إشغال المستخدم بعمليات التصحيح والتعديل.
- 7- أدى استخدام الخوارزمية المقترحة لتوليد الكلمات المفتاحية إلى زيادة دقة النتائج.
- 8- أدت عملية توسيع الكلمات المفتاحية باستخدام المنطق المضبب وخوارزمية Levenshtein-Distance إلى تحسين نوعية الاسترجاع وذلك بتوسيع عدد الوثائق ذات الصلة.

المصادر

- [1]. أحمد زياد محبك, 2010, "اللغة العربية والكتاب الرقمي", أستاذ الأدب العربي الحديث بجامعة حلب, عضو المجلس العالمي للغة العربية.
- [2]. جاسم محمد جرجيس, صباح محمد كلو, 2002, مقدمة في علم المكتبات والمعلومات, جامعة صنعاء 146-143.
- [3]. Acharya, T. and Ray, A. K., 2005,"Image Processing: Principles and Application", John Wiley & Sons, Inc, Hoboken, New Jersey.
- [4]. Bhaskar Karn , 1998 ,” Information Retrieval System Using Fuzzy Set Theory – The Basic Concept”, Assistant Professor, Department Of MIS(Management Information Systems), Birla Institute Of Technology, Mesra, Ranchi.
- [5]. D. Hiemstra , 2001,"Using Language Models for Information Retrieval", Universiteit Twente. ISBN 90-75296-05-3,ISSN 1381-3617; No. 01-32 (CTIT Ph.D. Thesis Series) Subject headings: information retrieval, natural language processing.
- [6]. Fabio Crestani and Gabriella Pasi, 1997, "Soft Information Retrieval: Applications of Fuzzy Set Theory and Neural Networks" , Department of Computing Science, University of Glasgow.
- [7]. L.A. Zadeh, 1994, “Fuzzy logic, neural networks and soft computing”, Communication of ACM.
- [8]. Ricardo Baeza-Yates,2003, " Information retrieval in the Web: beyond current search engines " , Center for Web Research, Department of Computer Science, University of Chile, Blanco Encalada, 2120 Santiago, Chile , Received 1 January 2003; accepted 1 July 2003.
- [9]. Van Rijsbergen, C.J., 1979, “Information Retrieval Butterworths”, Department of Computing Science, University of Glasgow, second edition.
- [10]. Yates, R. A. and B. Ribeiro-Neto, 1999, “Modern Information Retrieval”, Addison-Wesley.