Development Search Engine Kernel to Prevent Indexing Hurtful Sites and Evaluation Performance

Sirar Hammoud*

(Received 20 / 2 / 2022. Accepted 23 / 6 / 2022)

□ ABSTRACT □

This research contains web search engine kernel to control hurtful sites and prevent indexing it.

This research contain information about search in web , retrieval system , types of search engines and basic architectures of building search engines . It suggests architecture of search engine kernel to do final planner of search engine architecture ,It build parts of search engine and execute test to get results.

Keywords : Search Engine, Crawler, Indexer, Retrieval System

* **Work Supervisor, Department Computer and Automatic Control Engineering, Faculty of Mechanical and Electrical Engineering, Tishreen University, Lattakia, Syria. engineerengineer2011@hotmail.com**

تطوير نواة محرك بحث لحجب المواقع الالكترونية الضارة وحظر فهرستها وتقييم أدائه

سرار حمود*

(تاريخ الإيداع 20 / 2 / 2022. قُبِل للنشر في 23/ 6 / 2022)

□ ملخّص □

يقدم هذا البحث نواة محرك بحث يمكنه العمل ضمن شبكة الإنترنت، قادر على التحكم بالمواقع الالكترونية الضارة و حجبها وحظر فهرستها. فقد تم دراسة مسألة البحث عن المعلومات عبر الإنترنت ونظم استرجاع المعلومات وأنواع محركات البحث والمعماريات الأساسية لبناء المحركات ومن ثم اقتراح معمارية محرك بحث يصلح نواة لمحرك البحث المرغوب وتحديد المخطط النهائي لمعمارية محرك البحث حيث تم بناء مقاطع محرك البحث وإجراء الاختبارات والنتائج.

الكلمات المفتاحية : محرك البحث – العنكبوت – المفهرس – نظام الاسترجاع .

* **مشرف على الأعمال– قسم هندسة الحاسبات و التحكم الآلي– كلية الهندسة الميكانيكية و الكهربائية– جامعة تشرين– اللاذقية– سورية.**
engineerengineer2011@hotmail.com

مقدمة:

تكمن مشكلة الإنترنت في عدم تنظيم المعلومات ضمنها، ويرجع هذا إلى نمو شبكة الإنترنت بسرعة كبيرة، فقد ظهرت مسألة زيادة حجم المعلومات من جهة، وتنظيم الوصول والبحث عن هذه المعلومات من جهة ثانية، حيث تشكل شبكة الانترنت خزاناً معلوماتياً هائلاً يضاف إليها ملايين الصفحات يومياً، فظهرت الحاجة إلى تصميم وسائل للوصول إلى هذه المعلومات بسرعة وسهولة بهدف توفير الوقت عن طريق استخدام محركات البحث. وبالتالي فإن محرك البحث هو برنامج مصمم للمساعدة في البحث عن المعلومات المطلوبة المتاحة على شبكة الإنترنت بسهولة ويسر حيث يعيد قائمةً بالمراجع التي توافق تلك المعلومات خلال فترة زمنية محددة [1]،[2] .

أهمية البحث و أهدافه:

تنوعت الدراسات المرجعية عن محركات البحث، فمنها اهتم بمقارنة أداء محركات البحث في الحصول على المعلومات الطبية والصحية، وبعضها تناول دراسة تصميم محركات بحث متعددة، وإحدى الدراسات كانت عن المشكلات التي تواجه محرك البحث في الجدولة وتنفيذ الاستعلام، ومنها تناول تحسين طرق استرداد المعلومات لفهرسة مستندات النص العادي، وتحسين تقنيات محرك البحث، واحدى الدراسات اهتمت ببناء محركات بحث متخصصة باللغات المتعددة، وفي هذا البحث تم تطوير نواة محرك بحث، يتميز بقدرته على التحكم بالمواقع الالكترونية الضارة وإمكانية حجبها وحظر فهرستها، بالإضافة لاسترجاع أحدث النسخ عن المعلومات المطلوبة في مجال البحث العلمي، والحصول على إحصائيات لمعرفة وقت وعدد مرات تكرار البحث.[5]،[6]،[7].

طرائق البحث و مواده:

يعتمد بناء محركات البحث على مجموعة متنوعة من أدوات التطوير البرمجية، ومن أهم الأدوات المستخدمة في تصميم محرك البحث المتخصص المقترح، هي نظام إدارة قواعد البيانات MySql التي تتمتع بالعديد من الميزات، وكذلك لغة PHP المتعددة الخصائص التي جعلتها الأساس لتطوير برامج الويب المختلفة[8] ،[9]،[10]،[11]،[12].

يتطلب بناء محرك البحث المقترح بناء كل من العنكبوت، الفهرسة، الاسترجاع، يتم بناء العنكبوت بهدف إيجاد صفحات جديدة على الويب وإضافتها، أما الفهرسة فإن هدفها أن تأخذ المعلومات من العنكبوت و تعمل على تصنيفها ضمن فهارس مرتبة ومقسمة بترتيب معين لتسهيل تنفيذ عملية البحث عليها، أما بناء الاسترجاع فيقوم على تلقي الاستعلام و البحث ضمن الفهرس للمطابقة و تقديم النتائج الموافقة للاستعلام[13] ،[14]،[15]،[16] ،[17]،[18] .

منهجية بناء العنكبوت المتبعة:

انطلاقاً من تحليل عمل العنكبوت المبين في الشكل (1) يمكن تقسيم عملية البناء إلى الخطوات التالية :

1- تخزين رابط صفحة البداية من أجل الفهرسة.

2- قراءة محتويات هذه الصفحة للبحث عن روابط موجودة في هذه الصفحة .

3- عند وجود رابط جديد إضافة هذا الرابط للفهرسة .

4- العودة إلى النقطة الثانية.

5- عند انتهاء الروابط في الصفحة تنتهي عملية البناء .

منهجية بناء المفهرس المتبعة:

إن تنفيذ المفهرس المبين في الشكل (1) يتم بالخطوات التالية:

1- قراءة محتويات الصفحة ووضعها في قاعدة البيانات مع الرابط الذي تنتمي إليه.

2- حساب وزن كل كلمة مع تجاهل بعض الكلمات الشائعة بكثرة والتي تملك أهمية قليلة جداً في تحديد المستندات المطابقة للاستعلام و تخزينها في قاعدة البيانات و يتم تعيين الوزن بحيث يكون مساوياً لعدد مرات ورود الكلمة في الصفحة.

3- تنتهي مهمة الفهرسة للصفحة الحالية و يتم فهرسة الصفحة التالية التي تم الحصول عليها من قائمة الروابط في بناء العنكبوت.

منهجية بناء الاسترجاع المتبعة:

إن تنفيذ الاسترجاع المبين في الشكل (2) في محرك البحث يتم بالخطوات التالية:

1 – إدخال اسم الموضوع المطلوب البحث عنه.

2- البحث في جداول قاعدة البيانات عن الكلمات التي تطابق اسم موضوع البحث و ترتيب النتائج.

3- في حال لم يطابق اسم الموضوع أحد محتويات قاعدة البيانات لا تظهر نتائج ويعود إلى النقطة الأولى للبحث عن موضوع آخر، حيث يتم حجب المواضيع غير المرغوب إيجادها بالبحث.

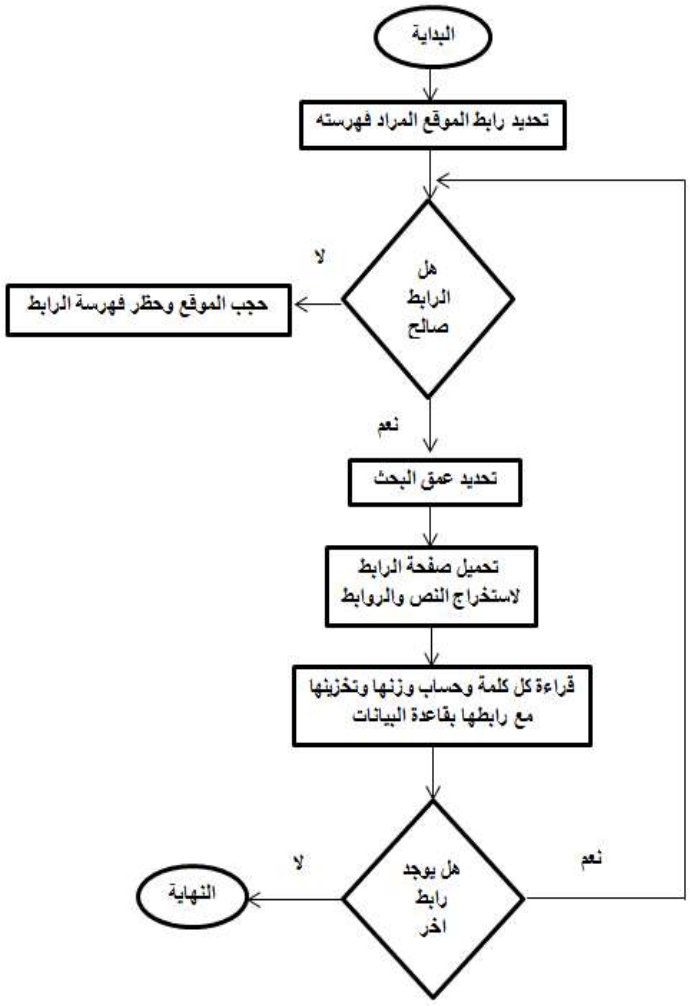
معمارية نواة محرك البحث المقترح:

لقد تم اقتراح محرك بحث مكون من مجموعة من المركبات المرتبطة بعضها ببعض، بنظام منطقي موضح عبر المخطط المنهجي المبسط المبين في الشكل (3).

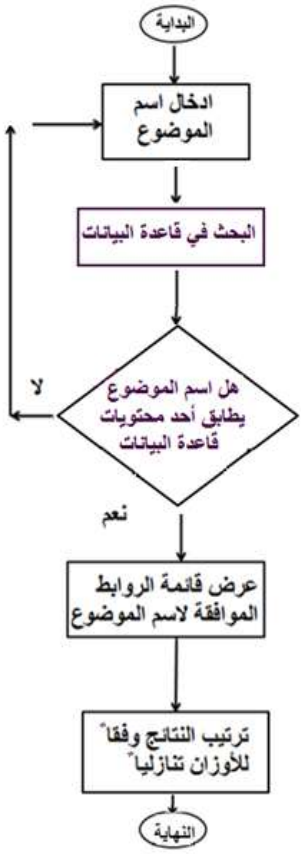
يغطي المخطط المقترح المراحل الأساسية لبناء محرك البحث، حيث يقلع في البداية لتحديد يتم تحديد عمق البحث والتقاط رابط الموقع المراد فهرسته وتحميل صفحة الرابط، الصفحة المحملة تحمل في مكان تخزين مؤقت، حيث تتم عليها مجموعة من العمليات، ومن ثم يتم إعراب الصفحة ويتم استخراج النص والروابط منها وتمريرها لمكون إلغاء التكرار، حيث يتم اختبار في حال كان هناك صفحة بنفس المحتوى محملة مسبقا من رابط آخر يقوم فلتر الروابط (URL Filter) بتحديد فيما إذا سيتم استبعاد الروابط اعتمادا على عدة اختبارات. فمثلا قد نرغب باستبعاد بعض الروابط من بعض المجالات، مثلا كل الروابط المنتهية بـ com.، في هذه الحالة يقوم الاختبار بفلترة الروابط المنتهية بـ com. ثم يتم قراءة كل كلمة مع حساب أوزان الكلمات وتخزينها مع الرابط الموافق في قاعدة البيانات وبهذا تكون قد تمت عملية الفهرسة. ثم يتم التأكد من وجود روابط أخرى، وفي حالة النفي تنتهي عملية تحميل الروابط ،أما في حالة الإيجاب يتم وضع رابط الصفحة (الجديد) في قائمة الروابط وتحميل صفحة الرابط التالي (الجديد) بعد اختبار صلاحية الموقع. وعند البحث عن موضوع معين ضمن مجال تخصص محرك البحث يتم البحث بقاعدة البيانات عن الروابط التي توافق اسم الموضوع و ترتيبها تنازلياً.

تحليل النظام :

تعدّ عملية تحليل النظام مرحلة هامة في بناء محرك البحث وتتم باستخدام لغة UML وهي لغة نمذجة رسومية تقدم لنا صيغة لوصف العناصر الرئيسة للنظم البرمجية. ويتم تحليل النظام اعتماداً على الخطوات المبينة في الجداول(1) و(2) و(3) و(4) و(5) و(6) و(7) و(8) و(9) و(10) ، والأشكال (4) و(5) .



الشكل (1) بناء العنكبوت و المفهرس

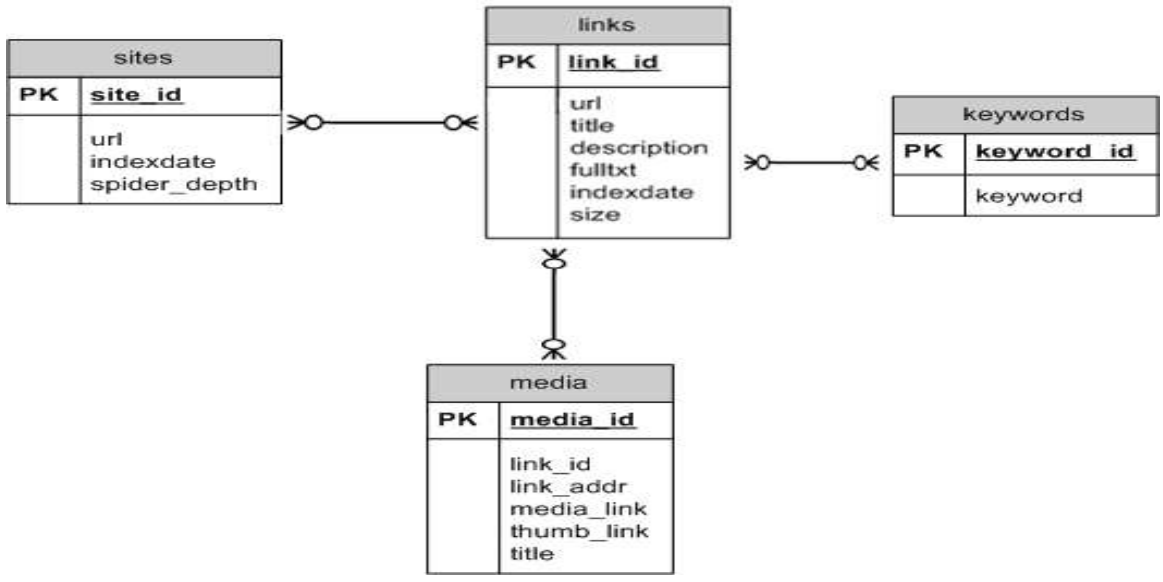


الشكل (2) نظام الاسترجاع

الشكل (3) المخطط المقترح لمعمارية برنامج محرك البحث

الجدول (1) جدول كيانات النظام		
اسم الكيان	اسم جدول الكيان	خصائص الجدول
كيان المواقع الصالحة	good sites	site_id_g: تمثل رقم الموقع الصالح الذي تم فهرسته. url_g: رابط الموقع الصالح الذي تم فهرسته. indexdate: تاريخ فهرسة الموقع. spider_depth: عمق البحث.
كيان المواقع الضارة	hurtful sites	site_id_h: تمثل رقم الموقع الضار الذي تم حجب فهرسته. url_h: رابط الموقع الضار الذي تم حجب فهرسته.

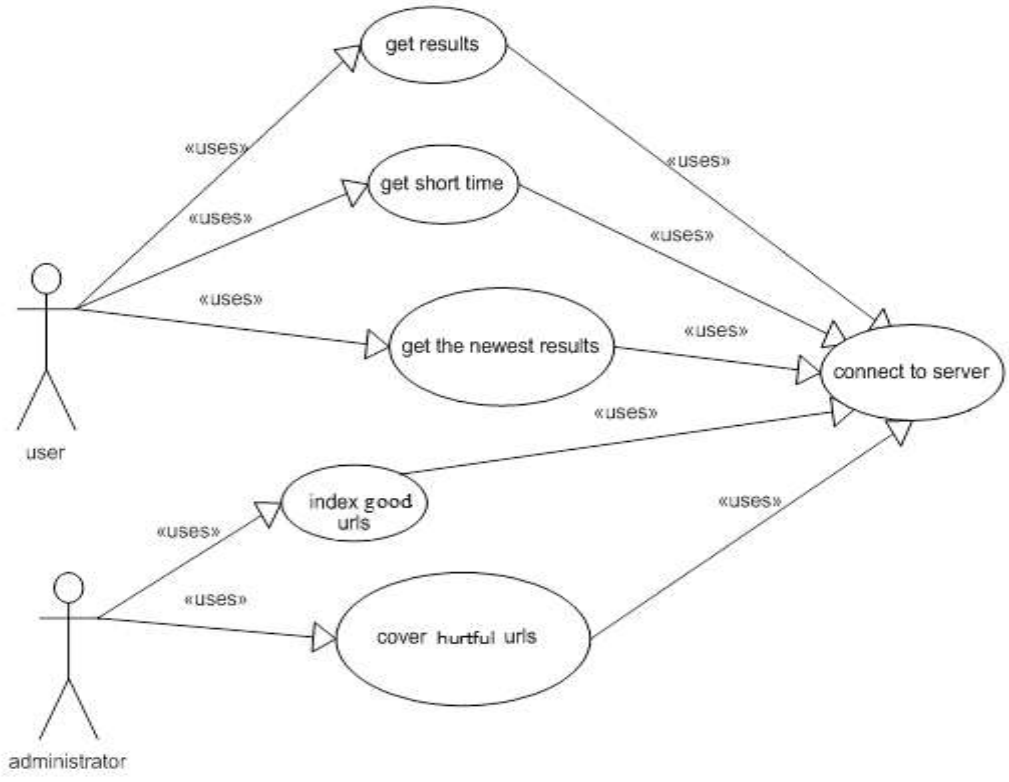
كيان الكلمات	Keywords	keyword_id: تمثل رقم الكلمة التي تمت قراءتها و تخزينها بقاعدة البيانات. keyword: الكلمة التي تمت قراءتها و تخزينها بقاعدة البيانات.
كيان الروابط	Links	link_id: رقم الرابط الذي تم فهرسته. site_id: رقم الموقع الذي يتبع له الرابط . url: الموقع الذي تمت فهرسته. title: عنوان الرابط الذي تمت فهرسته. description: الوصف المختصر للرابط الذي تمت فهرسته. fulltxt: النص الكلي للرابط الذي تمت فهرسته. indexdate: تاريخ فهرسة الرابط. size: حجم محتوى الرابط الذي تمت فهرسته.
كيان كلمة الرابط	link_keyword	link_id: ويمثل رقم الرابط الذي تم فهرسته. keyword_id: ويمثل الكلمة التي تمت قراءتها و تخزينها بقاعدة البيانات. weight: وزن الكلمة حسب عدد مرات تكرارها في محتوى الرابط. hits: عدد مرات تكرار وجود الكلمة في محتوى الرابط. indexdate: تاريخ فهرسة الرابط.
كيان الصور	Media	media_id: تمثل رقم الصورة. link_id: رقم الرابط الذي يتضمن الصورة. link_addr: عنوان الرابط الذي يتضمن الصورة. media_link: رابط الصورة. thumb_link: موقع الصورة المصغرة التي تم تخزينها بقاعدة البيانات. title: عنوان الصورة التي تم تخزينها بقاعدة البيانات.



الشكل (4) مخطط ERD

الجدول (2) متطلبـات النظام
تخصيص البحث عن معلومات بمجالات محددة . get results
اختصار زمن البحث عن المعلومات (حسب الإمكانات المتاحة). get short time
استرجاع أحدث النسخ عن المعلومات المطلوبة.get the newest results
إضافة المواقع الصالحة وفهرستها.index good urls
حجب المواقع الضارة أثناء البحث أو الفهرسة. cover hurtful urls

الجدول (3) مستخدمي النظام	
اسم المستخدم	الوصف
الباحث	يقوم بالبحث عن أحدث المعلومات المتخصصة بمجالات محددة بزمن قصير
المدير	إضافة المواقع وفهرستها و إمكانية حجب فهرسة رابط الموقع الضار



الشكل (5) مخطط حالات الاستخدام (Use case diagram)

الجدول (4) حالة الاستخدام الحصول على النتائج			
اسم حالة الاستخدام	وصف موجز	الفاعل	الشروط السابقة
الحصول على النتائج	تسمح هذه الخدمة للباحث تخصيص البحث عن معلومات محددة	الباحث	لا توجد شروط مسبقة
التدفق الرئيس			التدفقات البديلة
المستخدم	النظام		يقوم الباحث بإدخال كلمة بحث غير موجودة بقاعدة البيانات فيعرض النظام رسالة بعدم وجودها
في حالة البحث عن نصوص يقوم الباحث بإدخال كلمات البحث المرادة ومن ثم يضغط زر البحث في حالة البحث عن صور يقوم الباحث بإدخال كلمات البحث المرادة ومن ثم يضغط على البحث عن طريق الصور و من ثم يضغط زر البحث			
	يقوم النظام بعرض النتائج التي حصل عليها من المخدم بأجزاء من الثانية و يظهر الزمن الصغير المستغرق بعملية البحث		

الجدول (5) حالة الاستخدام اختصار زمن البحث			
اسم حالة الاستخدام	وصف موجز	الفاعل	الشروط السابقة
اختصار زمن البحث	تسمح هذه الخدمة للباحث بالحصول على زمن بحث قصير عن المعلومات (حسب الإمكانات المتاحة)	الباحث	لا توجد شروط مسبقة
التدفق الرئيس			التدفقات البديلة
المستخدم	النظام		يقوم الباحث بإدخال كلمة بحث غير موجودة بقاعدة البيانات فيعرض النظام رسالة بعدم وجودها
في حالة البحث عن نصوص يقوم الباحث بإدخال كلمات البحث المرادة ومن ثم يضغط زر البحث في حالة البحث عن صور يقوم الباحث بإدخال كلمات البحث المرادة ومن ثم يضغط على البحث عن طريق الصور و من ثم يضغط زر البحث			
	يقوم النظام بعرض النتائج التي حصل عليها من المخدم بأجزاء من الثانية و يظهر الزمن الصغير المستغرق بعملية البحث		

الجدول (6) حالة الاستخدام استرجاع أحدث النسخ			
اسم حالة الاستخدام	وصف موجز	الفاعل	الشروط السابقة
استرجاع أحدث النسخ	تسمح هذه الخدمة للباحث باسترجاع أحدث النسخ عن المعلومات المطلوبة.	الباحث	لا توجد شروط مسبقة
التدفق الرئيس			التدفقات البديلة
المستخدم		النظام	يقوم الباحث بإدخال كلمة بحث غير موجودة بقاعدة البيانات فيعرض النظام رسالة بعدم وجودها
في حالة البحث عن نصوص يقوم الباحث بإدخال كلمات البحث المرادة ومن ثم يضغط زر البحث في حالة البحث عن صور يقوم الباحث بإدخال كلمات البحث المرادة ومن ثم يضغط على البحث عن طريق الصور و من ثم يضغط زر البحث			
		يقوم النظام بعرض أحدث النتائج عن المعلومات المطلوبة التي حصل عليها من المخدم	

الجدول (7) حالة الاستخدام فهرسة المواقع الصالحة			
اسم حالة الاستخدام	وصف موجز	الفاعل	الشروط السابقة
فهرسة المواقع الصالحة	تسمح هذه الخدمة للمدير بإمكانية إضافة المواقع الصالحة وفهرستها مع إمكانية حجب المواقع الضارة ومنع فهرستها	المدير	لا توجد شروط مسبقة
التدفق الرئيس			التدفقات البديلة
المستخدم		النظام	يقوم المدير بإدخال موقع غير موجود أو خاطئ فيتم عرض رسالة خطأ
يقوم المدير بالدخول إلى إعدادات البرنامج ثم إلى تبويب الفهرسة و من ثم يقوم بإدخال رابط الموقع المراد فهرسته و يحدد عمق الفهرسة و يختار فيما إذا أراد أن تتم فهرسة الموقع بفهرسة جميع روابط الموقع التابعة لنفس مجال الموقع أو فهرسة جميع روابط الموقع بغض النظر عن المجال ،			
	يقوم النظام بعرض نتائج الفهرسة التي حصل عليها من المخدم		

الجدول (8) حالة الاستخدام حجب المواقع الضارة			
اسم حالة الاستخدام	وصف موجز	الفاعل	الشروط السابقة
حجب الكلمات غير المرغوبة	إمكانية حجب البحث عن مواضيع غير المرغوبة و منع فهرسة رابط الموقع الضار و فهرسة الرابط الصالح.	المدير	لا توجد شروط مسبقة
التدفق الرئيس			التدفقات البديلة
المستخدم		النظام	يقوم المدير بإدخال موقع

يختار المدير فيما إذا أراد حجب فهرسة المجالات غير المرغوبة و يختار أيضا فيما إذا أراد حجب فهرسة رابط الموقع الضار أو فهرسة الرابط الصالح و إذا أراد حجب البحث عن كلمات غير مرغوبة		غير موجود أو خاطئ فيتم عرض رسالة خطأ
	يقوم النظام بعرض نتائج الفهرسة التي حصل عليها من المخدم	

الجدول (9) حالة الاستخدام الاتصال مع المخدم

اسم حالة الاستخدام	وصف موجز	الفاعل	الشروط السابقة
الاتصال مع المخدم	تسمح هذه الخدمة للبرنامج بالاتصال مع المخدم للحصول على المعلومات	جميع حالات الاستخدام التالية الحصول على النتائج ، اختصار زمن البحث ، استرجاع أحدث النسخ ، فهرسة المواقع الصالحة وحجب فهرسة المواقع الضارة	أن يتم طلب الاتصال من إحدى الخدمات السابقة
التدفق الرئيس			التدفقات البديلة
المستخدم		النظام	في حال فشل الاتصال يعرض النظام رسالة خطأ توضح المشكلة
يقوم المدير بالدخول إلى إعدادات البرنامج		يقوم النظام بفتح اتصال بالانترنت	
		يرسل النظام المعلومات وينتظر النتائج	
		يعيد النتائج التي تم طلبها	

بعض المقاطع البرمجية المستخدمة ببناء محرك البحث المتخصص المقترح :

1- المقطع البرمجي المستخدم من أجل الاتصال مع قاعدة البيانات :

```
$db_con     = db1_connect() ;
$success    = mysql_select_db ($database1, $db_con);
```

2 - المقطع البرمجي المستخدم من أجل الحصول على النتائج النصية :

```
 if ($media_only==0 ) {
 $text_results = get_text_results($query, $start, $results);
$ip = $_SERVER['REMOTE_ADDR']; $results=$text_results[total_results];
saveToLog($query, $time, $results, $ip, 0);  }
```

3 - المقطع البرمجي المستخدم من أجل الحصول على الصور :

```
if (   $media_only == '1') { media_only($query ,$start, $media_only ); }
```

4 - المقطع البرمجي المستخدم للتحقق من كون الرابط صالح :

```
foreach ($blacklist as $value) {
$met = stripos($data[fulltext], $value);
 if($met)  $found = '1'; }
if ($found == '1') {  printStandardReport('matchBlacklist'); }
```

آلية تقييم محرك البحث المقترح:

يهدف نظام استرجاع المعلومات إلى تلبية احتياجات المستفيدين من المعلومات. وبذلك فإن مقياس نجاح تلك النظم في تحقيق أهدافها يتمثل في مدى رضا المستفيدين عن النتائج التي حصلوا عليها من النظام، ومدى تطابق تلك النتائج مع استفسارات المستفيدين .ويتم عادة تقييم نجاح النظام في أداء وظائفه في اتجاهين هما: [6] ،[8] ،[10]

- ❖ فحص نتائج البحث وتطبيق مقاييس التقييم عليها.
- ❖ سؤال المستفيدين عن درجة رضائهم .

يستطيع القائم على عملية التقييم التحقق بسهولة من مطابقة بعض العناصر لاحتياجات المستفيدين واستفساراتهم كاللغة والتاريخ والتكلفة، إلا أن هناك عناصر أخرى يصعب التعرف على مدى ملاءمتها لاحتياجات المستخدمين كمدى ارتباط نتائج البحث باستعلام المستخدم، كذلك مدى الإضافة التي تحققها تلك النتائج إلى معلومات المستخدم. ويمكن التعرف على تلك الجوانب وغيرها بتطبيق بعض المقاييس على نتائج البحث، ومن بينها الاستدعاء والدقة والجدّة والكلفة.

النتائج والمناقشة:

تم تقييم أداء محرك البحث المقترح الذي تم تصميمه في هذه الدراسة ليكون خاص بالبحث العلمي، والمتخصص في إحدى المجالات مثل : الطب، الاقتصاد ،الفنون ، الرياضيات ، الفيزياء ، الكيمياء ، الحاسوب .. اعتماداً على مقاييس التقييم السابقة وبالتالي يكون لدينا أربع فئات من النتائج موضحة في الجدول (11) ينبغي وضعها في الاعتبار وهي: [6] ،[8] ،[10]

مواد من رابط صالح تم استرجاعها، مواد من رابط صالح لم يتم استرجاعها، مواد من رابط ضار تم استرجاعها، مواد من رابط ضار ولم يتم استرجاعها.

وتعتمد تلك النتائج على بعدين هما: الاسترجاع ، الصلة بالموضوع.

ويتم النظر إلى الفئات الأربعة وتطبيق معيارين رئيسين عليها لتقييم النتائج وهما :الاستدعاء والدقة ويضاف إلى هذه المقاييس أيضا مقياسي الجدة و الكلفة. [6]،[8] ،[10]

الجدول (11) فئات النتائج

	مواد من رابط صالح	مواد من رابط ضار
تم استرجاعها	a	c
لم يتم استرجاعها	b	d

أولاً :الاستدعاء recall : هو مقياس لدرجة اكتمال الاسترجاع، حيث يقيس نسبة المواد لرابط صالح التي تم استرجاعها فعلياً من النظام، وذلك بتطبيق المعادلة : $R = a / (a+b)$

ثانياً :الدقة precision : هو مقياس لمدى نقاء الاسترجاع، حيث يقيس نسبة المواد المسترجعة لرابط صالح وذلك بتطبيق المعادلة : $P=a/(a+c)$

ثالثاً :الجدة Novelty : يقصد بالجدة نسبة النتائج الجديدة على المستخدم، أي تلك التي لم يتعرف عليها من قبل، و تأخذ الصيغة التالية :n =n/(n+m) حيث n تمثل عدد النتائج الجديدة التي لم يسبق للمستخدم الاطلاع عليها، في حين تمثل m عدد النتائج المعروفة بالنسبة للمستخدم.

رابعاً :التكلفة cost : في الحالات التي يكون فيها المستخدم مطالباً بدفع مقابل مادي عن إجراء البحث، فإنه يمكن حساب التكلفة المباشرة مقابل كل نتيجة لرابط صالح تم استرجاعها ضمن نتيجة البحث، فعلى سبيل المثال لو أن البحث استرجع 20 نتيجة من بينها 18 نتيجة لرابط صالح وكانت تكلفة البحث 1000 ليرة ، فإن تكلفة كل نتيجة تكون 55.5 = 18 / 1000 ليرة.

أما لو استرجع البحث 200 نتيجة من بينها 18 نتيجة لرابط صالح وكانت تكلفة البحث 2000 ليرة ، فإن تكلفة استرجاع النتيجة الواحدة تكون 111.1 = 18 / 2000 ليرة. أي أنها ستة أضعاف تكلفة النتيجة في البحث الأول، وبذلك فإن نسبة الدقة والاستدعاء في البحث لها تأثير على التكلفة، فكلما كان البحث أفضل من حيث الاستدعاء والدقة أدى ذلك إلى احتمال نقص تكلفة كل نتيجة لرابط صالح تم استرجاعها.

ووجد أنه عند قيام العنكبوت بتعقب روابط أربعة مواقع صالحة وقيام المفهرس بفهرسة صفحات تلك الروابط والبحث عن موضوع معين ، ظهرت لدينا النتائج التالية:

تضم قاعدة البيانات 20 مستنداً، واستُرجع 20 منها عند إجراء البحث، وبذلك فإن نسبة الاستدعاء في هذه النتيجة هي: R=20/(20+0) =1 . وبفحص النتائج السابقة تبين أن النتائج الـ 20 تنتمي لروابط صالحة، فإنه يمكن حساب نسبة الدقة على النحو الآتي: P=20/(20+0)=1. وبفحص النتائج السابقة تبين أن 9 نتائج منها سبق لنا الاطلاع عليها من قبل، في حين أن 11 نتائج تعد جديدة بالنسبة لنا، وبذلك يمكن احتساب الجِدّة على النحو الآتي: N= 11/(11+9) = 0.6. والبحث السابق استرجع 20 نتيجة لرابط صالح وإذا كانت تكلفة البحث 100 ليرة، فإن تكلفة كل نتيجة تكون 5 = 20 / 100 ليرة سورية.

وعند قيام العنكبوت بتعقب روابط ثمانية مواقع صالحة و قيام المفهرس بفهرسة صفحات تلك الروابط والبحث عن الموضوع السابق ظهرت لدينا النتائج التالية:

تضم قاعدة البيانات 35 مستنداً، واستُرجع 35 منها عند إجراء البحث، وبذلك فإن نسبة الاستدعاء في هذه النتيجة هي: R=35/(35+0) =1 . وبفحص النتائج السابقة تبين أن 31 منها فقط تنتمي لرابط صالح، في حين 4 نتائج تم استرجاعها تنتمي لرابط ضار، فإنه يمكن حساب نسبة الدقة على النحو الآتي: P=31/(31+4)=0.9، وبفحص النتائج السابقة تبين أن 17 نتيجة منها سبق لنا الاطلاع عليها من قبل، في حين أن 18 نتيجة تعد جديدة بالنسبة لنا، وبذلك يمكن احتساب الجِدّة على النحو الآتي: N= 18/(18+17) = 0.5، والبحث السابق استرجع 31 نتيجة لرابط صالح وإذا كانت تكلفة البحث 100 ليرة، فإن تكلفة كل نتيجة تكون 100/31 = 3.22 ليرة سورية.

وعند البحث عن الموضوع السابق مع قيام العنكبوت بتعقب روابط عشر مواقع صالحة وقيام المفهرس بفهرسة صفحات تلك الروابط ظهرت لدينا النتائج التالية :

تضم قاعدة البيانات 44 مستنداً في موضوع الحواسب، واستُرجع 44 منها عند إجراء البحث، وبذلك فإن نسبة الاستدعاء في هذه النتيجة هي: R=44/(44+0)= 1. وبفحص النتائج السابقة تبين أن 40 منها فقط لها صلة فعلية بموضوع البحث، في حين 4 نتائج تم استرجاعها لا صلة لها بالموضوع، فإنه يمكن حساب نسبة الدقة على النحو الآتي: P=40/(40+4)=0.9، وبفحص النتائج السابقة تبين أن 34 نتيجة منها سبق لنا الاطلاع عليها من قبل، في

حين أن 10 نتائج تعد جديدة، وبذلك يمكن احتساب الجِدّة على النحو الآتي: 0.2 = (34+10)/10 =N، والبحث السابق استرجع 40نتيجة ذات صلة فعلية بالموضوع وإذا كانت تكلفة البحث 100 ليرة، فإن تكلفة كل نتيجة تكون2.5 = 100/40 ليرة سورية.

وعند البحث عن الموضوع السابق مع قيام العنكبوت بتعقب روابط ثلاثة عشر موقع صالح وقيام المفهرس بفهرسة صفحات تلك الروابط ظهرت لدينا النتائج التالية:

تضم قاعدة البيانات 47 مستنداً في موضوع الحواسب، واستُرجع 47 منها عند إجراء البحث، وبذلك فإن نسبة الاستدعاء في هذه النتيجة هي: 1= (47+0)/47=R، وبفحص النتائج السابقة تبين أن 42 منها فقط لها صلة فعلية بموضوع البحث، في حين 5 نتائج تم استرجاعها لا صلة لها بالموضوع ، فإنه يمكن حساب نسبة الدقة على النحو الآتي: 0.9=(42+5)/42=P، وبفحص النتائج السابقة تبين أن 44 نتيجة منها سبق لنا الاطلاع عليها من قبل، في حين أن 3 نتائج تعد جديدة، وبذلك يمكن احتساب الجِدّة على النحو الآتي: 0.1 = (44+3)/3 =N، والبحث السابق استرجع 42 نتيجة ذات صلة فعلية بالموضوع وإذا كانت تكلفة البحث100 ليرة، فإن تكلفة كل نتيجة تكون = 100/42 2.4ليرة سورية.

وعند البحث عن الموضوع السابق مع قيام العنكبوت بتعقب روابط عشرين موقع صالح وقيام المفهرس بفهرسة صفحات تلك الروابط ظهرت لدينا النتائج التالية :

تضم قاعدة البيانات48 مستنداً في موضوع الحواسب، واستُرجع 48 منها عند إجراء البحث، وبذلك فإن نسبة الاستدعاء في هذه النتيجة هي: 1= (48+0)/48=R وبفحص النتائج السابقة تبين أن 42 منها فقط لها صلة فعلية بموضوع البحث، في حين أن 6 نتائج تم استرجاعها لا صلة لها بالموضوع، فإنه يمكن حساب نسبة الدقة على النحو الآتي: 0.9=(42+6)/42=P، وبفحص النتائج السابقة تبين أن 45 نتيجة منها سبق لنا الاطلاع عليها من قبل، في حين أن 3نتائج فقط تعد جديدة، وبذلك يمكن احتساب الجِدّة على النحو الآتي: 0.1 = (45+3)/3 =N، والبحث السابق استرجع 42 نتيجة ذات صلة فعلية بالموضوع وإذا كانت تكلفة البحث 100 ليرة، فإن تكلفة كل نتيجة تكون 100/42 2.3=ليرة سورية.

وعند البحث عن الموضوع السابق مع قيام العنكبوت بتعقب روابط سبع وثلاثين موقع صالح وقيام المفهرس بفهرسة صفحات تلك الروابط ظهرت لدينا النتائج التالية :

تضم قاعدة البيانات 54 مستنداً في موضوع الحواسب، واستُرجع 54 منها عند إجراء البحث، وبذلك فإن نسبة الاستدعاء في هذه النتيجة هي:1= (54+0)/54=R ، وبفحص النتائج السابقة تبين أن 46 منها فقط لها صلة فعلية بموضوع البحث، في حين أن 8 نتائج تم استرجاعها لا صلة لها بالموضوع ، فإنه يمكن حساب نسبة الدقة على النحو الآتي: 0.9=(46+8)/46=P، وبفحص النتائج السابقة تبين أن 47 نتيجة منها سبق لنا الاطلاع عليها من قبل، في حين أن 7 نتائج فقط تعد جديدة، وبذلك يمكن احتساب الجِدّة على النحو الآتي 0.1 = (47+7)/7 =N، والبحث السابق استرجع 46 نتيجة ذات صلة فعلية بالموضوع وإذا كانت تكلفة البحث100 ليرة، فإن تكلفة كل نتيجة تكون 2.1= 100/46ليرة سورية.

وعند البحث عن الموضوع السابق مع قيام العنكبوت بتعقب روابط خمسين موقع صالح وقيام المفهرس بفهرسة صفحات تلك الروابط ظهرت لدينا النتائج التالية :

تضم قاعدة البيانات66 مستنداً في موضوع الحواسب، واستُرجع66 منها عند إجراء البحث، وبذلك فإن نسبة الاستدعاء في هذه النتيجة هي:1= (0+66)/66=R . وبفحص النتائج السابقة تبين أن 52 منها فقط لها صلة فعلية بموضوع البحث، في حين 14 نتيجة تم استرجاعها لا صلة لها بالموضوع، فإنه يمكن حساب نسبة الدقة على النحو الآتي: 0.8=(52+14)/52=P، وبفحص النتائج السابقة تبين أن 53 نتيجة منها سبق لنا الاطلاع عليها من قبل، في حين أن 13 نتائج تعد جديدة، وبذلك يمكن احتساب الجِدّة على النحو:

0.2 =(13+53)/13= N، والبحث السابق استرجع 52 نتيجة ذات صلة فعلية بالموضوع وإذا كانت تكلفة البحث 100ليرة، فإن تكلفة كل نتيجة تكون 1.9 = 52 / 100 ليرة سورية.

وعند البحث عن الموضوع السابق مع قيام العنكبوت بتعقب روابط خمس و سبعين موقع صالح و قيام المفهرس بفهرسة صفحات تلك الروابط ظهرت لدينا النتائج التالية :

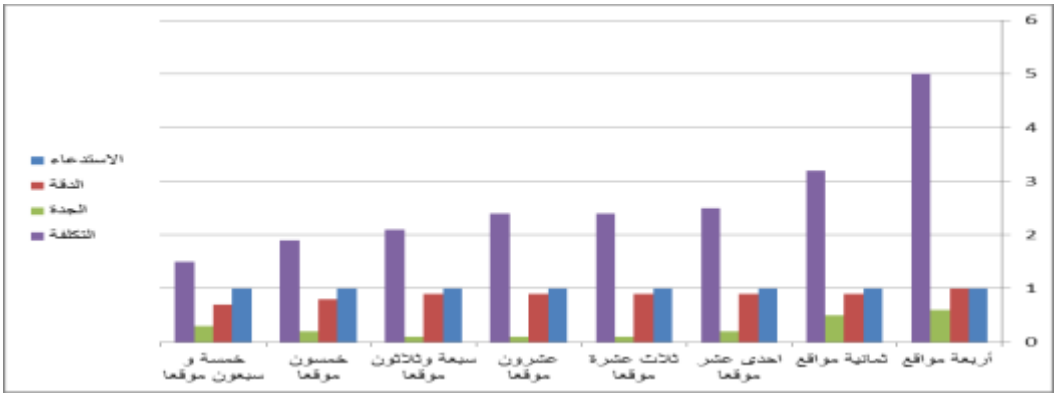
تضم قاعدة البيانات 94 مستنداً في موضوع الحواسب، واستُرجع 94 منها عند إجراء البحث، وبذلك فإن نسبة الاستدعاء في هذه النتيجة هي: 1= (0+94)/94=R. وبفحص النتائج السابقة تبين أن 69 منها لها صلة فعلية بموضوع البحث، في حين أن 25 نتيجة تم استرجاعها لا صلة لها بالموضوع ، فإنه يمكن حساب نسبة الدقة على النحو الآتي: 0.7=(69+25)/69=P، وبفحص النتائج السابقة تبين أن 63 نتيجة منها سبق لنا الاطلاع عليها من قبل، في حين أن 31 نتيجة تعد جديدة، وبذلك يمكن احتساب الجِدّة على النحو: 0.3 = (31+63)/31 =N، والبحث السابق استرجع 69 نتيجة ذات صلة فعلية بالموضوع وإذا كانت تكلفة البحث100 ليرة، فإن تكلفة كل نتيجة تكون 1.5 =100/69ليرة سورية.

ويوضح الشكل (6) المخطط البياني لتقييم أداء محرك البحث المقترح حيث يبدو أن نسبة الاستدعاء ثابتة و تساوي 1، كون محرك البحث يطابق الطلب كحروف لغوية و هذا يؤدي لاستدعاء كامل النتائج التي تطابق كلمة البحث و بالتالي نسبة الاستدعاء لن تتغير و تبقى ثابتة مساوية 1.

كما يبدو أن نسبة الدقة تنخفض بشكل بسيط مع زيادة حجم قاعدة البيانات حيث مع أنه مع ازدياد حجم قاعدة البيانات تزداد عدد النتائج التي لا صلة لها بموضوع البحث و بالتالي تنخفض نسبة الدقة .

ويبدو أيضاً أن نسبة الجِدّة تنخفض بشكل صغير مع ازدياد حجم قاعدة البيانات حيث إنه مع ازدياد حجم قاعدة البيانات ينخفض عدد النتائج الجديدة التي لم يسبق للمستخدم الاطلاع عليها و بالتالي تنخفض نسبة الجِدّة .

كما يبدو أن التكلفة تنخفض بشكل كبير مع زيادة حجم قاعدة البيانات حيث إنه مع زيادة حجم قاعدة البيانات يزداد عدد النتائج التي يتم استرجاعها و بالتالي تنخفض تكلفة كل نتيجة يتم استرجاعها .



الشكل (6) المخطط البياني لتقييم أداء محرك البحث المقترح

الاستنتاجات والتوصيات:

بُنيت نواة محرك بحث ويب قادر على التحكم بالمواقع الالكترونية الضارة وامكانية حجبها وحظر فهرستها، ويمكن أن يكون منصة باتجاه بناء محرك بحث منافس لمحركات البحث العالمية الموجودة. و قد تركزت الجهود للاهتمام بجميع التفاصيل و الخيارات لبناء قواعد البيانات و مكونات المحرك الأخرى، وتبين من نتائج تطبيق مقاييس التقييم على نتائج البحث أن شمولية الاسترجاع ثابتة مهما تغير حجم قاعدة البيانات ولكن مع زيادة حجم قاعدة البيانات ينخفض نقاء الاسترجاع قليلاً و تنخفض عدد النتائج الجديدة التي لم يسبق للمستخدم الاطلاع عليها بشكل بسيط كما تنخفض تكلفة كل نتيجة تم استرجاعها بشكل كبير .

أقترح في المرحلة اللاحقة الاهتمام بالجانب الأمني لقواعد البيانات لما له من أهمية في استمرارية عمل محرك البحث والوصول الى الأداء الأمثل.

References:

[1] GREHAN,N**.** *Function of Search Engines***,** Addison-Wesley,2020, 300.

[2] KENT,C. *Search Engine improvement*.5th.ed., Cambridge UP , England, ,2020.

[3] MENG,H .*How Metasearch Engines* work, Springer, German ,2016.

[4] DING,W.H. ; BUYYA,C. *Guided Google: A Meta Search Engine*, Wiley, London,2017, 355.

[5] OLSTON,P.; NAJORK,M. *Web Crawling*, Vol. 4, No. 3 ,2018,175–246.

[6] CROFT,R.; METZLER,S. ;STROHMAN,C. *Search Engines: Information Retrieval*, Incisive Media , New York,2015,500.

[7] HERSH,P. *Information retrieval*.3th.ed., , Binghamton University, New York, 2021.

[8] MANNING,P.; RAGHAVAN,W.; SCHÜTZE,C., An *Introduction to Information Retrieval*, John Wiley & Sons , London ,2017.

[9] LEVENE,N**.** *Search Engines and Web* Navigation, University of Melbourne, Australia,2018.

[10] BAMEFLIH,M. , *Improvement Information retrieval systems*.1st.ed.,KFNL , Riyadh, 2013.

[11] LARRY,A., *PHP 6 and MySQL 5 for Dynamic Web Sites*, Peachpit Press,2018.

[12] Mark L., "*Creating specialized search engines in multiple languages*," IEEE Transactions on Knowledge and Data Engineering, vol. 18,no. 1, pp. 50–75, 2019.

[13] Gautam ,P.; Padmini ,S.; Filippo,M., *Crawling the Web*,2017

[14]. INAMDA,A., *An Agent Based Intelligent Search Engine*, Swami Ramanand Teerth Marathwada University,2018

[15] OLSTON,C., *Crawl ordering by search impact*, in Proceedings of the 1st International Conference on Web Search and Data Mining, 2011

[16] GANAPATHY, A., *Google's deep-web crawl*, in Proceedings of the 34th International Conference on Very Large Data Bases, 2016.

[17] G. Pant and P. Srinivasan, "*Link contexts in classifier-guided topical crawlers*," IEEE Transactions on Knowledge and Data Engineering, vol. 18,no. 1, pp. 107–122, 2006.

[18] J. Dorn, "*Improving search engine technology*," IEEE Transactions on Knowledge and Data Engineering, vol. 18,no. 3, pp. 117–132, 2020.