# ما وراء محركات البحث Meta - Search Engines (*)

**د. خالد عبد الفتاح محمد**

مدرس علوم المكتبات والمعلومات

بكلية الآداب جامعة المنيا

**مقدمة :**

تعتبر شبكة المعلومات العالمية «الإنترنت» وخصوصاً الويب ، واحدة من أبرز ملامح مجتمع المعلومات المعاصر ، ويرجع ذلك إلى ضخامة وتنوع مصادر المعلومات المتاحة على هذه الشبكة والعدد الضخم من المستفيدين الذين يقومون باستخدام هذه المصادر . وبالتالي فإنه من الصعب الوصول إلى هذه المصادر دون وجود أدوات تيسر عملية البحث والاسترجاع .

وتوجد ثلاث أدوات لبحث واسترجاع المعلومات المتاحة على شبكة المعلومات العنكبوتية (الويب) وهى : أدلة البحث Directories – ومحركات البحث المستقلة Individual Search Engines – وما وراء محركات البحث Meta - Search Engines . ونظراً لتعدد الدراسات والبحوث باللغة العربية واللغة الإنجليزية التي تناولت أدلة البحث ومحركات البحث المستقلة ، وندرة الدراسات التي تناولت ما وراء المحركات ، فإن هذه الدراسة تركز على ما وراء المحركات .

وتعتبر ما وراء المحركات واحدة من أهم أدوات بحث واسترجاع مصادر المعلومات المتاحة على شبكة المعلومات العالمية (الإنترنت) في الوقت الحالي . وتقوم هذه المحركات بصفة عامة بتلقي استفسارات المستفيدين وإرسالها إلى مجموعة منتقاة من محركات البحث المستقلة ، ثم تتلقى النتائج من هذه المحركات وتقوم بدمجها ومعالجتها ثم فرزها في قائمة مرتبة وفقا لخوارزميات الدمج والترتيب Merging Algorithms – هذا بالإضافة إلى بعض العمليات الأخرى مثل تحليل الاستفسارات وترجمتها لكي تتوافق مع إمكانيات البحث المختلفة للمحركات المشاركة في النظام ،

(*) قدم فى مؤتمر المكتبات والمعلومات فى مجتمع المعرفة : الحاضر والمستقبل ، جامعة عين شمس ، كلية الآداب ، ٣١ مارس – ١ ابريل ٢٠٠٤ م .

ولكي تستفيد أيضاً من القيمة المضافة لعمليات التشغيل المتداخل Interoperability – التي توفرها خوارزميات الدمج والترتيب (Dowork, 2000).

يواجه الباحثون في عمليات بناء وتطوير ما وراء المحركات عدة تحديات أساسية من أهمها :

١ – اختيار محركات البحث المستقلة وتجميعها فى قائمة موحدة وترتيبها وفقاً لأولويات الدمج .

٢ – دمج النتائج المسترجعة .

٣ – ترتيب وفرز النتائج المسترجعة .

وفيما يلي عرض للأسس والمعايير المستخدمة في كل مرحلة من هذه المراحل :

## ١ - اختيار محركات البحث المستقلة وتجميعها فى قائمة موحدة وترتيبها وفقاً لأولويات الدمج .

تعرف هذه العملية في الإنتاج الفكري المتخصص في مجال استرجاع المعلومات بعملية اختيار وفرز قواعد البيانات (Williams, M, Preece, 1979; Williams, M; Maclaury, Rouse, 1979) . ويقوم المبرمجون في هذه المرحلة بتجميع قوائم شاملة بمحركات البحث المستقلة للاختيار من بينها وفقاً لأحد المعايير التالية:

### ١/١ حجم التغطية في قواعد البيانات

**Databases Coverage**

في هذه الحالة يقوم المبرمج بتجميع قائمة شاملة بأشهر محركات البحث المتاحة وأكثرها شمولاً من حيث عدد التسجيلات التي تم تكشيفها والمتاحة فعلياً للبحث والاسترجاع ، ثم يقوم المبرمج بالمقارنة بين قواعد البيانات . وذلك عن طريق تشغيل برنامج للفرز حيث يقوم هذا البرنامج بفرز قواعد البيانات وترتيبها من الأكثر شمولاً إلى الأقل شمولاً . ونظراً لأن محركات البحث المستقلة تنوع فى تغطيتها لمصادر المعلومات المتاحة على شبكة الإنترنت من حيث نوع صفحات المعلومات (مثل صفحات الويب ، صفحات البى دى أف ، صفحات الأوفيس ، أو قواعد البيانات) . فتتم المقارنة بين هذه الأنواع المختلفة لترتيب المحركات وفقاً للاحتياجات الأساسية لما وراء المحركات وليس السياسات المتبعة في المحركات المستقلة . وتجدر الإشارة هنا إلى أنه توجد مصادر متعددة على الشبكة العالمية توفر إحصائيات دقيقة عن معدلات التغطية في محركات البحث المستقلة . ومن أبرز هذه المصادر :

**www.searchengineswatch.com**

http://www.search-engine-index.co.uk/

**Search Engines.com**

**Show Down Statistics**

### ٢/١ معدلات الاستخدام أو الاستفسار

**Query Load**

في هذه الحالة يتم تحديد عدد الاستفسارات التى توجه إلى كل محرك بحث على حدة وترتيبهم من المحرك الأكثر استفساراً إلى المحرك الأقل استفساراً . كما أن بعض ما وراء المحركات تأخذ في الاعتبار نسبة الاستفسارات الناجحة إلى نسبة الاستفسارات الفاشلة . ويمكن الحصول على هذه

الإحصائيات من خلال تحليل ملف الاستفسارات أو ما يعرف بملف اللوج في كل محرك مستقل على حده . لكن من عيوب هذه الطريقة أنها تتطلب قدر كبير من التعاون من المحركات المستقلة ، وهو أمر غير مرغوب في تلك البيئة ، نظراً للطبيعة التنافسية الشديدة التي تحكم هذا المجال . حيث أن الحصول على هذه الملفات قد يؤدى إلى الكشف عن أساليب تحليل الاستفسارات والخوارزميات المستخدمة في عمليات التكشيف والاسترجاع . هذا وإن كانت هذه الأمور من السهل الكشف عنها من خلال التحليل الدقيق للنتائج المسترجعة والأساليب المفضلة لدى هذه المحركات في بناء إستراتيجيات البحث . ولعل أبرز نماذج التعاون في هذا المجال هو ما قدمته محركات البحث المستقلة (Excite, Alta Vista and Ask Jeeves) – للباحثين من ملفات بغرض التحليل والدراسة ، للتعرف على طبيعة الاستفسارات الموجهة إلى هذه المحركات . ومن أمثلة الدراسات التي تناولت محركات البحث المستقلة بالفحص والتحليل ما يلي :

(Spink, Bateman & Jansen, 1998; Jansen et el, 1998; Saracevic & Kantor, 1998; Spink, Bateman & Jansen, 1999; Spink & Ozmutlu, 2001; Goodrum & Spink, 2001; Spink, 2002) .

### ٣/١ وقت الاستجابة Response Time

يتم قياس متوسط الوقت الذي يستغرقه كل محرك على حده في إجراء البحث واستعراض النتائج ، ثم يتم ترتيب المحركات وفقاً لسرعة الاستجابة من الأكفأ إلى الأقل كفاءة . هذا وإن كان الفارق بين محركات البحث من حيث وقت الاستجابة هو فارق غير محسوس إلا أن مؤشر وقت الاستجابة عامل في غاية الأهمية بالنسبة لمطوري ما وراء المحركات نظراً لما تتطلبه العملية من إجراء البحث في أكثر من محرك مستقل . بالتالي فإن سرعة المحركات المستقلة تؤثر بالتبعية على سرعة ما وراء المحركات . وهذه الطريقة سوف تضمن كفاءة عالية من حيث سرعة الاستجابة ولكنها لا يمكن أن تضمن بأي حال من الأحوال كفاءة وفعالية المواد المسترجعة .

### ٤/١ تقييم النتائج المسترجعة من المحركات المستقلة: Individual Search Engines Results Evaluation

ويشمل التقييم ثلاثة معايير أساسية من مقاييس التقييم في مجال استرجاع المعلومات وهم: الاستدعاء والدقة والترتيب أو الفرز . وتوجد العديد من الدراسات التي قارنت بين محركات البحث من حيث دقة النتائج المسترجعة . ومن أمثلة هذه الدراسات :

(Chu & Rosenthal, 1996; Ding & Marchionini, 1996; Su, 1997; Wishard, 1998; Lighton & Sirvastava, 1999; and Dennis et al., 2002).

وتتسم هذه الدراسات بالمقارنة بين محركات البحث في بيئتها الطبيعية من حيث مقومات البحث والمجموعات وطبيعة الاستفسارات . ويعرف هذا الاتجاه في الأدبيات بالاتجاه العملي Operational Approach .

كما يوجد نوع آخر من الدراسات تولى المقارنة بين محركات البحث المستقلة عن طريق فصل عناصر المقارنة لتجربتها في المعمل . ويعرف هذا الاتجاه بالاتجاه المعملي Laboratory Approach . حيث تتم التجارب على عناصر معينة في محركات البحث دون العناصر الأخرى للتعرف على مدى تأثيرها على كفاءة ودقة الاسترجاع (Rusmussen, 2003) .

هذا ويوجد عدد قليل من الدراسات تناولت الاستدعاء ودقة الترتيب كأساس للمقارنة بين محركات البحث . ويرجع ذلك إلى صعوبة قياس الاستدعاء ، نظراً لاستحالة تحديد عدد الوثائق الصالحة في قاعدة البيانات . بالإضافة إلى أن قياسه غير عملي نظراً لضخامة حجم النتائج المسترجعة لكل استفسار ، والتي قد تصل في بعض الأحيان إلي آلاف بل مئات الآلاف من صفحات المعلومات (Gordon & Pathak, 1999; Clarke & Willett, 1999) .

أما بالنسبة لدقة الترتيب فمن الصعب أيضاً قياسه بطريقة إجرائية نظراً لاختلاف مجموعة الوثائق المسترجعة من محرك لآخر . ولذلك تعددت الدراسات التي تقارن بين محركات البحث من حيث دقة الاسترجاع ، حيث أثبتت التجارب أن قياس الدقة في الاسترجاع هي أهم مقاييس التقييم في بيئة محركات البحث المتاحة على الشبكة العنكبوتية . ويمكن الاعتماد على هذه الدراسات في ترتيب محركات البحث من حيث دقة هذه المحركات في استرجاع النتائج الصالحة .

## ٢ - دمج النتائج المسترجعة

**Fusing or Combining Search Results**

توجد أربعة طرق أساسية لدمج البيانات معروفة ومستخدمة في مجال استرجاع المعلومات . وهذه الطرق هي :

### ١/٢ دمج النتائج المسترجعة وفقا لاستراتيجيات بحث متنوع

**Fusing Different Search Strategies**

وتعتمد هذه الطريقة على التنويع فى طريقة بناء إستراتيجية البحث لنفس الإستفسار . حيث يتم توجيه هذه الإستراتيجيات المتنوعة لنفس المحرك . ثم يتم دمج النتائج المسترجعة بعد استبعاد النتائج المتكررة Overlapped Results .

وقد أثبت كلا من (Saracevic & Kantor, 1998) أنه عند توجيه إستراتيجيات بحث متنوعة لنفس المحرك فإنه يسترجع نتائج مختلفة ، بعض هذه النتائج صالحة وبعضها غير صالحة . وتجدر الإشارة هنا إلى وجود قدر كبير من التداخل والتكرار نظراً لأن الاستفسارات المتنوعة تعالج نفس الموضوع.

### ٢/٢ دمج النتائج المسترجعة وفقا لأساليب متنوعة لوزن المصطلحات

**Fusing According to Term Weighting Schemes**

في هذه الحالة يتم استخدام مجموعة موحدة من الوثائق في بناء عدة قواعد بيانات وفقاً لطرق متنوعة لوزن المصطلحات . ثم يتم توجيه نفس

الاستفسارات لكل قاعدة بيانات على حده ، ثم يتم دمج النتائج المسترجعة من قواعد البيانات بعد استبعاد المكررات . وقد أكد لى (Lee, 1995) أن استخدام أكثر من طريقة لوزن المصطلحات يؤدى إلى تحسين كفاءة الاسترجاع .

### ٣/٢ دمج النتائج وفقا لأجزاء الوثائق المكشفة Data Fusion According to Document Representation

وتعتمد هذه الطريقة على التنويع فى أجزاء الوثائق المكشفة ، حيث يتم إعداد قواعد بيانات مستقلة حسب الجزء المكشف من الوثيقة . فعلى سبيل المثال يتم تكشيف عناوين الوثائق فقط في قاعدة بيانات ويتم تكشيف المستخلصات في قاعدة بيانات أخرى . ويتم إجراء البحث في كل قاعدة بيانات على حدة ، ثم تدمج النتائج المسترجعة بعد استبعاد المكررات ، لتحديد مدى تأثير هذه الأجزاء على فعالية الاسترجاع . وقد اكتشف كاتزر وزملاءه أن إجراء البحث على أجزاء متنوعة من الوثيقة يؤدى إلى استرجاع نتائج بنفس الكفاءة والفعالية ، مما يؤدى إلى زيادة معدلات الدقة والاستدعاء عند دمج هذه النتائج (Katez et al. 1982).

### ٤/٢ دمج النتائج المسترجعة من نظم استرجاع متعددة Data Fusion According to Multiple Retrieval Systems

في الثلاثة نماذج السابقة يمكن استخدام نظام استرجاع موحد مع التنويع في طرق التكشيف أو بناء إستراتيجيات البحث أو أجزاء الوثائق المكشفة. أما في هذا النموذج فيتم التنويع في المصدر بأكمله . حيث يتم الدمج من مصادر متعددة Multiple Sources . وهذا هو النموذج السائد في كل ما وراء المحركات والنظم التي تعتمد على استخدام بروتوكول استرجاع المعلومات Z39.5 (Khaled, 2004) . ومن الفروق الأساسية أيضاً أن الثلاث طرق السابقة تكشف مجموعة موحدة من الوثائق ، بينما يعتمد هذا النموذج على مجموعة مختلفة من الوثائق مع وجود قدر من التداخل والتكرار بين هذه المصادر المتنوعة .

وتجدر الإشارة هنا إلى أنه توجد أربع حالات لمجموعة الوثائق المكشفة تصلح لعملية دمج البيانات. وهذه الحالات هي: (Yang & Zang, 2000) .

#### حالة التساوي Equivalent Case

وهى الحالة التي تكون فيها الوثائق المكشفة في كل قواعد البيانات واحدة دون أي اختلاف فيما بينها .

#### حالة الاشتمال Inclusion Case

وهى الحالة التي تكون فيها إحدى قواعد البيانات شاملة وقواعد البيانات الأخرى تتضمن جزء من الوثائق المكشفة في قاعدة البيانات الشاملة .

#### حالة الاختلاف Disjoint Case

وهي الحالة التي لا يوجد فيها أي تشابه بين قواعد البيانات من حيث مجموعة الوثائق المكشفة.

#### حالة التداخل والتكرار Overlapping Case

وهي الحالة التي تتداخل فيها قواعد البيانات

من حيث مجموعة الوثائق المكشفة . وهذه هي الحالة السائدة في كل ما وراء المحركات المتاحة على شبكة الإنترنت .

## ٣ - فرز وترتيب النتائج المسترجعة

**Results Merging / Ranking**

تعد هذه الخطوة هي أكثر الخطوات أهمية في عملية دمج النتائج المسترجعة في ما وراء المحركات ، حيث أن معظم هذه المحركات عادة ما تستخدم نفس الوسائل والأساليب في الخطوتين السابقتين ، بينما يعد الأسلوب المستخدم في مرحلة الفرز والترتيب هو العنصر المميز لمحرك عن الآخر . وعموماً ، يوجد أسلوبان أساسيان يستخدمان لتحديد الترتيب الأمثل للنتائج المسترجعة وهما :

- أسلوب التحميل والتحليل

Downloading and Analyzing

- أسلوب الترتيب وفقاً للافتراضات المنطقية

Merging According to Logical Assumptions

وفيما يلي عرض لكل أسلوب مع التركيز على الخوارزميات المستخدمة والأساس الذي بنيت عليه .

### ١/٣ أسلوب التحليل والتحميل :

يعرف هذا الأسلوب في أدبيات استرجاع المعلومات بأسلوب فحص أو تفتيش الوثائق Documents Fetching ويعتمد هذا الأسلوب على تحميل الوثائق المسترجعة بأكملها أو أجزاء منها من خادم محرك البحث المستقل إلى خادم ما وراء المحركات . ثم يتم تحليل هذه الوثائق باستخدام وسائل متعددة لعل أشهرها حساب درجة التشابه Similarity Score باستخدام طرق متنوعة لوزن المصطلحات Term Weighting Schemes (Meng et. el., 2002) .

وتستخدم درجة التشابه في ترتيب الوثائق حسب ارتباطها بموضوع الاستفسار ، وحسب درجة التشابه بين مصطلحات الاستفسار والكلمات المكشفة من الوثيقة . وتوجد العديد من نظم التحميل والتحليل المتاحة حالياً ، ولعل أبرزها gGoiss, (Lawrence & Giles, 1998) CORI, and CVV .

وتجدر الإشارة هنا إلى أن هذه النظم عادة ما تتضمن خوارزميات للاختيار والتحميل والتحليل والدمج في نفس الوقت ، حيث أنها عادة ما تتضمن كل الوظائف اللازمة لما وراء المحركات .

#### ١/١/٣ القاموس العام لخادم الخوادم gGoiss

**Generalized Glossary of Server's Server**

وفي هذا النظام تمثيل قاعدة البيانات بتردد الوثائق Documents Frequency التي تشتمل على المصطلحات الواردة في الاستفسار ومجموع وزن مصطلحات الاستفسار في كل وثيقة . بالتالي يمكن التعرف على ترتيب قواعد البيانات ثم ترتيب الوثائق المسترجعة من كل قاعدة بيانات لإعداد القائمة النهائية (Gravano & Gracia - Molina, 1995) .

### ٢/١/٣ ترتيب صحة الاختلاف

### Cue Validity Variance

يستخدم هذا النظام مزيج من تردد الوثائق ودرجة التشابه لترتيب الوثائق وفقاً لمدى ارتباطها بمصطلحات الاستفسار . وتعمل هذه الحزمة بكفاءة في حالة التعاون بين محركات البحث المستقلة مع ما وراء المحركات من خلال مده بإحصائيات دقيقة ومستمرة عن قواعد البيانات وخوارزميات البحث المستخدمة في التكشيف والتحليل (Yuwono & Lee, 1997) .

### ٣/١/٣ شبكة مدلول استرجاع المجموعات Core Net

### Collection Retrieval Inference Network

تعتمد هذه الشبكة على تمثيل قواعد البيانات بعدد الوثائق المسترجعة لكل استفسار ، وتردد الكلمات في كل وثيقة ثم يتم إعطاء كل وثيقة درجة تشابه معينة وفقاً لتردد المصطلحات ضمن مجموعة الوثائق المسترجعة ، بالتالي يمكن ترتيب قواعد البيانات ومجموعة الوثائق المسترجعة وفقاً لارتباطها بمصطلحات الاستفسار . ويتطلب هذا النظام درجة كبيرة من التعاون بين ما وراء المحركات والمحركات المستقلة (Callan & Connel, 2001). وتوجد نماذج أخرى تستخدم أسلوب التحميل والتحليل ومنها الاسترجاع الفائق للوثائق (OptDoc Ret.) Optimal Document Retrieval . ويستند هذا النظام أيضاً على نفس الأسس المستخدمة في النماذج السابقة ، والتي تتمثل في تحميل الوثائق المسترجعة من خادم المحرك المستقل إلى خادم ما وراء المحركات ، ثم تحليلها وتكشيفها وترتيبها .

ولعل أبرز مميزات أسلوب التحميل والتحليل هو الاعتماد على أسلوب موحد في التحليل والترتيب بصرف النظر عن الخوارزميات التي تستخدمها المحركات المستقلة في الترتيب . ولهذا النموذج عدة عيوب ، لعل أبرزها :

١ – أنه يحتاج إلى وقت طويل لتحميل وتحليل الوثائق وهو ما لا يتناسب مع طبيعة مستخدمي الويب .

٢ – أنه يتطلب مساحات تخزين كبيرة ، حيث يتم تحميل الوثائق المسترجعة على خادم ما وراء المحركات ، هذا بالإضافة إلى خوارزميات التكشيف والتحليل والفرز .

٣ – يحتاج هذا النموذج إلى أنظمة استرجاع ذات كفاءة عالية لكي تقوم بعمليات التحليل والترتيب بفاعلية وكفاءة . حيث أن عمليات البحث في المحركات المستقلة والتحميل والتحليل وبناء ملفات الوثائق واستبعاد المكررات وبناء القوائم الموحدة ، ثم في النهاية استخدام أسلوب موحد لعرض النتائج المسترجعة ، لابد أن تتم كل هذه العمليات على الهواء On the Fly وهي عملية معقدة ودقيقة إلي درجة بعيدة . ويصلح هذا النموذج ويعمل بكفاءة عالية في نظم التجميع على الخط المباشر Aggregator Online Systems . وهي النظم التي يقوم فيها المورد بتجميع أكبر عدد ممكن من قواعد البيانات ، ويتيحها للاسترجاع على الخط المباشر . بالتالي فإن هذه البيئة

تسمح بقدر كبير من التعاون بين قواعد البيانات المستقلة ونظام التجميع . وهو ما لا يتوافر في بيئة الويب التي تقوم على التنافس الشديد بين محركات البحث .

### ٢/٣ اسلوب الترتيب وفقاً للافتراضات المنطقية :

**Merging Upon Logical Assumptions**

يعتمد هذا الأسلوب على استخدام الترتيب الأصلي الوثائق المسترجعة من المحركات المستقلة في إنتاج قائمة موحدة من خلال بناء خوارزميات فرز وترتيب تعتمد على الافتراضات المنطقية . ومن أبرز الخوارزميات المستخدمة في هذا النموذج .

#### ١/٢/٣ الحشو والإدراج Interleave

وتعتمد هذه الطريقة على ترتيب قواعد البيانات ترتيباً تنازلياً وفقاً لمقاييس متعددة ، مثل شمول التغطية ، دقة الاسترجاع ، أو وقت الاستجابة . ثم يتم ترتيب الوثائق وفقاً لترتيب قواعد البيانات ، حيث تأتى الوثيقة رقم ١ من قاعدة البيانات رقم ١ في الترتيب رقم ١ في القائمة الموحدة ، تليها الوثيقة رقم ١ من قاعدة البيانات رقم ٢ ، ثم الوثيقة رقم ١ من قاعدة البيانات رقم ٣ ، ثم الوثيقة رقم ٢ من قاعدة البيانات رقم ١ ، وهكذا إلى أن يتم الحصول على العدد المرغوب من الوثائق في القائمة الموحدة (Meng et el., 2002) .

كما قام فورهيز وزملاءه (Voorhees et el., 1995) باستخدام قواعد الاختيار العشوائي في ترتيب قواعد البيانات ، بالتالي يكون لكل محرك نفس الفرصة في أن يسبق المحركات الأخرى في الترتيب في القائمة النهائية ويستند نموذج الحشو والإدراج على افتراض أن الوثيقة المسترجعة من محرك بحث أكثر أهمية ربما تكون أفضل من وثيقة أخرى لها نفس الترتيب استرجعت من محرك آخر أقل أهمية . ومصطلح أهمية هنا يشير إلى موقع محرك البحث في قائمة المحركات المستقلة .

#### ٢/٢/٣ تحويل أرقام الوثائق إلى رقم تشابه عام :

**Convert Document Rank to Global Similarity Scores**

قام لى (Lee, 1997) بتصميم نموذج لترتيب القوائم النهائية يعرف باتجاه درجة التشابه . يستخدم هذا النموذج ترتيب الوثيقة الأصلي الذي تنتجه المحركات المستقلة من أجل ترتيب قوائمها في إنتاج القائمة الموحدة . ويعتمد هذا النموذج على المعادلة التالية :

$$\text{درجة التشابه} = ١ - \frac{\text{ترتيب الوثيقة} - ١}{\text{عدد الوثائق المسترجعة من المحرك المستقلة}}$$

والافتراض الأساسى هنا أن الوثيقة المسترجعة ضمن عدد أكبر من الوثائق أفضل من وثيقة أخرى لها نفس الترتيب ومسترجعة ضمن عدد أقل من الوثائق . على سبيل المثال ، فالوثيقة رقم ١ المسترجعة ضمن ألف وثيقة تعتبر أفضل من وثيقة رقم ١ ومسترجعة ضمن خمسمائة وثيقة .

كما قام كلا من يونو ولى (Yuwono & Lee, 1997) بإعداد معادلة لتحويل رقم الوثيقة المحلى Local Rank Score إلى رقم تشابه عام Global Similarity Score من خلال تطبيق المعادلة التالية : نفترض أن لكل استفسار أن ترتيب

محرك البحث $D_i$ هو $r_i$ وأن $r_{min}$ هو ترتيب آخر قاعدة بيانات في القائمة ، r هو الترتيب المحلى للوثيقة المسترجعة ، g هي درجة التشابه العام . والمعادلة المستخدمة في ترتيب القائمة النهائية :

$$g = 1 - ( r - 1 ) * Fi$$

حيث أن F هي :

$$Fi = (r_{min}) / (m * ri)$$

حيث أن m تمثل العدد المرغوب من الوثائق .

على سبيل المثال نفترض وجود قاعدتين بيانات D1 و D2 ونفترض أن ترتيبهم r1 = 0.2 و r2 = 0.5 ونفترض أن العدد الكلى المطلوب من الوثائق هو أربعة وثائق ، بالتالي فإن :

rmin = 0.2, F1 = 0.25, F2 = 1,m = 4

ووفقاً للمعادلة فإن الوثائق الثلاثة الأولى في D1 يحصلون على درجات تشابه ١ ، ٠,٧٥ ، ٠,٥ على التوالى . والوثائق الثلاثة الأولى من D2 يحصلون على درجة تشابه ١ ، ٠,٩ ، ٠,٨ على التوالي . من ثم فإن القائمة النهائية سوف تتضمن ثلاثة وثائق من D2 ووثيقة واحدة من D1 .

### ٣/٣ نماذج لما وراء المحركات المتاحة على شبكة الإنترنت :

لقد ظهرت العديد من أدوات البحث التي تستخدم تقنية ما وراء المحركات خلال الأعوام القليلة الماضية . ويمكن الوصول إلى قوائم شاملة بتجارب بناء ما وراء المحركات من خلال صفحات المعلومات التالية :

Search Engines Watch.com

Big Search Engines Index

Search Engines.com

Show Down.com

وسوف نستعرض فيما يلي نماذج لأفضل التجارب لبناء ما وراء المحركات .

اشتملت صفحة المعلومات[1] Big Search Engines Index في فبراير ٢٠٠٤ على ٤٦ أداة بحث تستخدم تقنية ما وراء المحركات . بعض هذه الأدوات تعرض قائمة شاملة بمحركات البحث المستقلة المرشحة للبحث مثل Ixquick, VIVISMO, Opsearch والبعض الآخر لا يعرض المحركات المستقلة المشاركة في ما وراء المحركات مثل Dogpile, Profusion حيث تستخدم هذه المحركات قالب عام للبحث . ومع ذلك يمكن الوصول إلى القائمة المستخدمة في البحث من خلال خيارات البحث المتقدم Advanced or Customized Search Options . وبمراجعة أبرز النماذج المتاحة لما وراء المحركات اتضح أن المحرك Dogpile لا يقوم بدمج النتائج المسترجعة ، إنما يستعرض نتائج كل محرك مستقل على حدة . بينما يقوم كلا من Inquick, Mamma بدمج النتائج من خلال استخدام المكررات في ترتيب القائمة النهائية ، حيث يتم الدفع بالوثائق التي تظهر في أكثر من محرك بحث مستقل إلى قمة القائمة . بالتالي فإن الوثيقة التي تظهر في ثلاث محركات تسبق وثيقة أخرى ظهرت في محركين فقط (Tsikrik, 2001) . كما تقوم أداة البحث

(1) Big Search Engines Index--- http://www.search-engine-index.co.uk/

Meta Crawler بجمع درجة تشابه الوثائق المكررة بالتالي تحصل الوثائق المكررة على درجة أعلى من الوثائق الفريدة Unique Documents . وتعتمد أداة البحث Profusion على وزن المصطلحات ، حيث يتم استخدام كلا من درجة التشابه المسترجعة من المحركات المستقلة والدرجة التي حصل عليها محرك البحث المستقل في مرحلة ترتيب المحركات المستقلة . ولكن المشكلة الأساسية في هذه الطريقة أنه ليست كل المحركات المستقلة تسترجع الوثائق مصحوبة بدرجة التشابه ، ولكنها تسترجع الوثائق مرتبة فقط دون أي معلومات إضافية عن الدرجة التي حصلت عليها كل وثيقة . بالتالي يتطلب استخدام هذه المعادلة تعاون المحركات المستقلة مع ما وراء المحركات (Callan, Lu & Croft, 1995) . أما أداة البحث ميتاجير Meta Ger فتعتمد على نظام التحليل والتحميل لترتيب القائمة النهائية . حيث تستخدم الترتيب الأصلي للوثائق المسترجعة من المحركات المستقلة إلى جانب تردد المصطلحات في عناوين تلك الوثائق أو ما وراء البيانات Metadata أو ملخص الوثيقة . كما تعتمد أداة البحث Inquiries على نظام التحليل والتحميل، حيث يتم تحميل الوثائق بالكامل على خادم ما وراء المحركات ثم تحليلها وبناء الكشافات. وتجدر الإشارة هنا أن أداة البحث Inquiries تعتمد على تردد المصطلحات بالإضافة إلى تقارب المصطلحات Term Proximity من أجل ترتيب القوائم النهائية .

## ٤ - خاتمة .

تتناول تلك الخاتمة الخطوات التي يجب تنفيذها عند بناء محرك بحث يعمل وفقاً لتقنية ما وراء المحركات . وتتطلب تلك العملية تنفيذ الخطوات التالية :

١ - إعداد قائمة شاملة بمحركات البحث المستقلة المتاحة على شبكة الإنترنت .

٢ - تجميع بيانات وإحصائيات دقيقة عن المحركات المستقلة ، تشمل حجم قواعد البيانات ، معدلات الإضافة والتعديل ، ومعدلات الزيادة السنوية ، ونظم التحليل والتكشيف ، ونظم البحث وتحليل الاستفسارات .

٣ - المقارنة بين محركات البحث المستقلة وفقاً لكفاءة الاسترجاع Retrieval Effectiveness من أجل ترتيبها تنازليا .

٤ - اختيار عدد مناسب من المحركات المستقلة من خلال تحليل الفارق بين إسترجاعية Retrievability المحركات العشر الأولى في القائمة والمحركات العشرين الأولى ، فإذا كان الفارق كبير يتم تحليل إسترجاعية الخمسة عشر الأولى مع العشرين ، وهكذا إلى أن يتم تحديد العدد المناسب من المحركات .

٥ - إعداد قائمة بقدرات البحث Search Capabilities للمحركات المستقلة من أجل ترجمة الاستفسارات Query Translation or Query Mapping إلى قدرات البحث الخاصة بكل محرك مستقل .

٦ - اختيار الطريقة المثلى لترتيب القائمة النهائية . هذا وإن كانت طريقة الافتراضات المنطقية القائمة على أساس ترتيب القوائم النهائية بناء

على دفع الوثائق المكررة إلى قمة القائمة ثم ترتيب الوثائق المنفردة بناء على كفاءة المحركات المستقلة هي أفضل الوسائل المتاحة حالياً ، نظراً لأنها تتناسب مع طبيعة العمل في بيئة الويب .

٧ – تحديد النموذج المثالي لوصف النتائج المسترجعة Document Description وهو أمر يحتاج إلى بحث دقيق .

٨ – حفظ استفسارات المستفيدين في ملف خاص يعرف بملف المستفيدين User Profile حيث يتم استخدام هذا الملف في إعادة استفسار Re-Query ما وراء المحركات وضخ النتائج الجديدة للمستفيدين وهو ما يعرف ببوابات الويب Web Portals .

## المراجع


1 - Callan. J & Connel, M. (2001). Query - based sampling of text databases. *ACM Transaction on information systems*, 19 (2) : pp. 97 - 130 .

2 - Chu, H., & Rosenthal, M. (1996). Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. *In S. Hardin (Ed), Proceedings of the 59th Annual Meeting of the American Society for Information Science.* (pp. 1127 - 135), Medford. NJ: American Society for Information Science.

3 - Clarke, S. J., & Willett, P. (1997). Estimating the recall performance of Web search engines. *Aslib Proceedings*, 49 (7), 184 - 189 .

4 - Ding W., & Marchionini, G. (1996) A Comparative Study of Web Service Performance. *In S. Hardin (Ed), Proceedings of the 59th Annual Meeting of the American Society for Information Science* (pp. 136 - 142), Medford. NJ: American Society for Information Science.

5 - Dennis, S., Bruza, P., & McArthur, R. (2002). Web Searching : A Process Oriented Experimental Study of Three Interactive Search Paradigms. *Journal of the American Society of Information Science.* 53 (2) : 120 - 133 .

6 - Dowork et al. (2002) Rank aggregation revisited. Citeseer. NEC Research. Available online (08/20/ 2002) http://citeseer.nj.nec.com/1478775.html

7 - Goodrum, A., & Spink, A. (2001). Image Searching on the Excite Web Search Engine. *Information Processing and Management.* 37 (2), 295 - 312 .

8 - Gordon, M & Pathak, P. (1999). Finding Information on the World Wide Web: The Retrieval

*Annual International Conference on Research and Development in Information Retrieval.* pp. 267 - 276.

14 - Leighton, H. V., Srivastava, J. (1999). First 20 Precision among World Wide Web Search Services (Search Engines). *Journal of the American Society for Library and Information Science,* 50, 870 - 881.

15 - Meng, W., Yu, C., & Liu, K. (2002). Building Efficient and Effective Metasearch Engines. *In ACM Computing Survey, 34 (1):* pp. 48 - 49.

16 - Rusmussen, E. (2003). Indexing and Retrieval for the Web. *Annual Review of Information Science and Technology.* Vol. 37. pp. 91 - 123.

17 - Saracevic, T. & Kantor, P. (1998). A Study of Information Seeking and Retrieving. III. Searchers, Searches, Overlap." *Journal of the American Society for Information Science.* 39 (3): pp. 197 - 216 .

18 - Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines : an exploratory study. *Information Processing and Management.* 38 (3): 401 - 426 .

19 - Spink, A., Bateman, J., Jansen. B. J. (1998). Searching Heterogeneous Collection on the Web : Behavior of Effectiveness of Search Engines. *Information Processing and Management.* 35 (2) : 141 - 80.

9 - Jansen, J., Spink, A., Batean, J., & Saracevic, T. (1998). Real Life Information Retrieval: A Study of the User Queries on the Web. *SIGIR Forum,* 32 (1): 5 - 17 .

10 - Katzer, J., McGill, M., Tessier, J., Frakes, W., & Dasgupta, P. (1982). A Study of Overlap among Document Representations. *Information Technology Research and Development.* 1 (2): pp. 261 - 274.

11 - Khaled A. (2004). Merging Multiple Search Results for Meta-Search Engines. Ph.D Dissertation. University of Pittsburgh, USA, 200p.

12 - Lee, J. (1995) Combining Multiple Evidence from Different Properties of Weighting Schemes. Annual ACM Conference on Research and Development in Information Retrieval. *Proceeding of the 18th Annual International Conference on Research and Development in Information Retrieval.* pp. 180 - 188.

13 - Lee, J. (1997). Analyses of Multiple evidence Combination. Annual ACM Conference on Research and Development in Information Retrieval. *Proceeding of the 23rd*

Excite Users. Information Research: An Electronic Journal, 5 (2) : http://www.shef.ac.uk/~is/publications/infers.

20 - Spink, A., Bateman, J., & Jansen, B. J. (1999). Searching the Web: Survey of Excite users. *Internet Research: Electronic Networking Applications and Policy,* 9 (2) : 117 - 128 .

21 - Spink, A., & Ozmutlu, H. C. (2001) What do people ask for on the Web and how do they ask it: Ask Jeeves query analysis. *Information Today,* Inc. 545 - 554 .

22 - Su, L. (1997). Developing a comprehensive and systemic model of user evaluation of Web-based search engines. *Proceeding of the 18th National Online Meeting* (pp. 335 - 344) . Medford, NJ: Information Today.

23 - Tzitzikas, Y. (2001). Democratic data fusion for information retrieval mediators. Citeseer. NEC Research. http://citeseer.nj.nec.com/tziztizkas01democratic. html

24 - Voorhees, E., Gupta, N., Larid, B. (1995). Learning Collection Fusion Strategies. *In the Proceedings of the ACM SIGIR Conference (Seattle, WA):* pp. 172 - 179 .

25 - Williams, M; Preece, S. (1977) Data Base Selector for network use: a feasibility study. Information management in the 1980s: *proceedings of the 40th ASIS Annual Meeting, volume 14,* edited by B.M. Fry. White Plains, New York, American Society for Information Science, Chicago, Illinois, September 26 October 1, 1977 (Abstract).

26 - Williams, M.; Maclaury, K; Preece, S; & Rouse, S. (1979) Data base mapping model and search scheme to facilitate resource sharing. *volume 1. mapping of chemical data bases and mapping of data base elements using a relational data base structure.* Coordinated Science Laboratory, University Of Illinois At Urban-Champaign. 342 P (Abstract).

27 - Wishard, L (1998). Precision among Internet Search Engines: and Earth Science Case Study. *Issues in Science and Technology Librarianship.*

28 - Yang, X. & Zhang, M. (2000) Necessary Constraints for Fusion Algorithms in Meta Search Engine Systems. *In Proceedings of International Conference on Intelligent Technologies, Bangkok. (Accessed through Citeseer Search Engines)* : p. 409 - 416 .