1.1 المقدمة

تأتي قضية تجهيز ومعالجة المحتوى على الشبكة العنكبوتية العالمية(World Wide Web) من أبرز القضايا التي تؤرق مجتمعات المعلومات في الوقت الراهن، وينطوي هذا الأمر على عدد من القضايا الفرعية الأخرى كقضية نشر المحتوى وفاعليته وطرق إكتشافه ونظم إدارته ولكن تبرز قضية هامة على صعيد هذا الأمر وهي قضية إسترجاع المحتوى فلا توجد جدوى من وجود المحتوى إن لم يتم إسترجاعه وإستثماره من قبل المستخدمين.

وعليه تعد أدوات البحث والإسترجاع وعلى رأسها محركات البحث في الويب (Web Search Engines) بمثابة حجرالأساس لهذا المحتوى المعلوماتي، وحلقة الوصل بين طرفي النشر والإسترجاع للمحتوى.

وعلما بأن المعلومات الموجودة على الشبكة العنكبوتية أصبحت متواجدة بلغات متعددة ومختلفة وأن الوثائق الإنجليزية هي المهيمنة علي الشبكة العنكبوتية ، ظهرت أهمية انظمة إسترجاع المعلومات ثنائية اللغة(CLIR) وفيها يقوم المستخدم بكتابة الإستعلام بلغة ويتم إسترجاع النتيجة بلغة أخرى خصوصاً في البلدان التي لا تنطق بالإنجليزية . وبما أن المحتوى المعلوماتي على الويب أصبح متواجد بأكثر من لغة كان لابد من السماح للمستخدم بكتابة الإستعلام بأي لغة وإسترجاع النتائج بأكثر من لغة ومن هنا ظهرت أهمية أنظمة إسترجاع المعلومات متعددة اللغات(MLIR) ، وبما أن النتائج المسترجعة تكون من أكثر من مجوعة كان لابد من دمج هذه النتائج في قائمة واحدة علي حسب أهميتها للإستعلام أو صلتها به .

ومع تطور المحتوي المعلوماتي على الشبكة العنكبوتية أصبحت الوثيقة الواحدة موجودة بأكثر من لغة (mixed documents) فكان لابد من إيجاد طريقة لفهرسة هذه الوثائق حتي نتمكن من البحث فيها بطريقة فعالة ومن هنا كان لابد من تمكين المستخدم من كتابة الإستعلام الواحد بأكثر من لغة(mixed query) وأن تكون النتائج المسترجعة من البحث بهذا الإستعلام تضم الوثائق الأعلى صلة بالإستعلام بغض النظر عن لغة الوثيقة.

2.1 مشكلة البحث

من الصعوبات التي تواجه أنظمة إسترجاع المعلومات هي كيفية فهرسة الوثائق خصوصاً متعددة اللغات(التي تكون مكتوبة بأكثر من لغة) بطريقة مثلى لتحسين كفاءة الإسترجاع ، وأيضاً عندما يقوم المستخدم بكتابة إستعلام بلغتين فإن ترتيب الوثائق المسترجعة من عملية البحث يتم على أساس المطابقة بين الإستعلام والوثيقة وليس على أساس أهميتها(relevant) بالنسبة للمستخدم حيث أن الوثائق ذات اللغات المتعدده تكون ذات وزن أعلى من الوثائق ذات اللغة الواحدة، مع معرفة أنه من المحتمل أن تكون هنالك وثائق ذات وزن أقل ولكنها أعلى صلة بالإستعلام من الوثائق متعددة اللغات، بالتالي تكون النتيجة المسترجعة غير مقنعة للمستخدم،مثال لذلك إذا كان المستخدم يبحث عن (مفهوم ال deadlock) يقوم محرك البحث بتحويل الإستعلام إلى اللغة العربية (monolingual arabic) مرة بحيث يصبح (مفهوم الاقفال) وأخرى إلى الإنجليزية (monolingual English)وتصبح (deadlock concept) وإستعلام ثالث بدمج الإستعلام باللغتين العربية والانجليزية (مفهوم الاقفال deadlock concept)، ويتم البحث بهذا

الإستعلام في مجموعة الوثائق العربية ويتم تجاهل وزن الجزء الإنجليزي من الإستعلام (أي وزنه يساوي صفر)، والبحث مرة اخرى في مجموعة الوثائق الإنجليزية وفي هذه الحالة يتم تجاهل وزن الجزء العربي من الإستعلام (أي وزنه يساوي صفر) ، والبحث اخيراً بالإستعلام في مجموعة الوثائق ذات اللغتين وهنا تكمن المشكلة في كيفية دمج النتائج المسترجعة من مجموعة الوثائق العربية والوثائق الإنجليزية والوثائق متعددة اللغات بناء على أهميتها بالنسبة للمستخدم.

3.1 أهمية البحث

رغم أن إستخدام محركات البحث إنتشر بصورة أكبر من السابق، وأن معظم المستخدمين لايتجاوزوا إستخدام أول صفحتين من نتائج محركات البحث التي تعرض محتوى العنكبوتية ، مردود هذا الأمر يعود إلى عدم تحقيق التطابق بين المحتوى المطلوب وبين المحتوى المسترجع من قبل محركات البحث ، و من الصعوبات والتحديات التي كانت سبباً ودافعا لدراسة التحديات التي تواجه محركات البحث في إسترجاع المحتوى هو أن الوثائق اصبحت متواجدة بأكثر من لغة، وأن الوثيقة نفسها يمكن أن تكون مكتوبه بأكثر من لغة (خصوصاً الوثائق غير الإنجليزية) و كان لابد لمحركات البحث المستقبلية أن تضع في إعتبارها أن المستخدم قد يكون غير قادر على التعبير عن مصطلح ما يريد البحث عنه أو لايعرف معنى التعبير بلغته(خصوصاً في البلدان التي لاتنطق بالإنجليزية) فيقوم بكتابة الإستعلام بأكثر من لغة مثلاً إذا كان يريد البحث عن "مفهوم الnetwork " ولا يعرف معنى كلمة network بالعربية فيقوم بكتابة الإستعلام كالتالي " مفهوم ال network " .

لذا كان لا بد من السعي نحو إيجاد طريقة تُمكن المستخدم من كتابة الإستعلام بأكثر من لغة و إسترجاع النتائج ودمجها علي أساس صلتها بالمستخدم وليست على أساس المطابقة أو الوزن بحيث تكون النتائج النهائية ذات صلة بطلب المستخدم .

4.1 أهداف البحث

1. تطبيق خوارزميات لدمج النتائج وإنتاج قائمة واحدة من الوثائق المرتبة على حسب أهميتها للمستخدم .
2. زيادة كفاءة النظام بزيادة كفاءة عملية الإسترجاع من الفهارس .
3. السماح للمستخدم الذي لايستطيع التعبير عن المصطلحات التي يريد البحث عنها بأن يكتب الاستعلام بأي لغة واسترجاع نتائج تكون مقنعه وذات أهمية بالنسبة للمستخدم.
4. بناء وتطوير معمارية جديدة لفهرسة الوثائق(indexing).

5.1 منهجية البحث

في المشروع قيد الدراسة ستتم فهرسة الوثائق بإستخدام طريقة الفهرسة المركزية الموزعة وبناء عدة فهارس والبحث فيها بإستخدام الإستعلام ومن ثم تطبيق خوارزميات عدة لدمج النتائج المسترجعة من الفهارس المختلفة في قائمة واحدة .

6.1 حدود البحث

سنتناول في هذه الدراسة أنظمة إسترجاع المعلومات وبالأخص أنظمة استرجاع المعلومات ذات اللغات المتعددة بهيكلية فهرسة تدمج بين طريقة الفهرسة المركزية والفهرسة الموزعة وقبل دمج الوثائق الناتجة من عملية البحث يتم تطبيع وزنها على حسب الخوارزمية التي أستخدمت في عملية الدمج .

7.1 هيكلية البحث

يتضمن البحث بالإضافة إلى هذا الباب الأبواب :

الباب الثاني والذي يتضمن نبذة عامة عن إسترجاع المعلومات حيث يحتوي على فصلين هما المقدمة بالإضافة إلى الدراسات السابقة.

الباب الثالث و يتضمن منهجية البحث التقنيات والأدوات المستخدمة في البحث.

الباب الرابع يتحدث عن تطبيق المشروع المقترح و خطوات حل المشكلة.

الباب الخامس يتضمن الخاتمة التي تحتوي على النتائج والتوصيات والمراجع.

1.2استرجاع المعلومات(Information Retrieval)

تعني عملية إسترجاع المعلومات إسترجاع الوثائق (documents) التي تطابق (matching) طلب المستخدم (query) وذلك حسب صلتها به.

وبذلك فإن عملية إسترجاع المعلومات تحوي عدة عمليات يقوم بها نظام إسترجاع المعلومات قبل البدء في مطابقة طلب المستخدم مع الوثائق التي يجب إسترجاعها، وتلك العمليات هي :

1. **التقطيع(Tokenization)**

هي عملية تقطيع المحتوى إلى كلمات ذات معنى تسمى قطع (Tokens) وهي أصغر وحدة لها معنى .

2. **إستبعاد الكلمات غير ذات المعنى(Removing Stopwords)**

هي الكلمات التي ليس لها معنى مفيد .

3. **تطبيع حروف بعض الكلمات (Normalization)**

هي عملية توحيد الكلمات التي تم تقطيعها ووضعها في الصيغة الموحدة وذلك لزيادة مدى المطابقة بين مجموعة الكلمات المقتطعة .

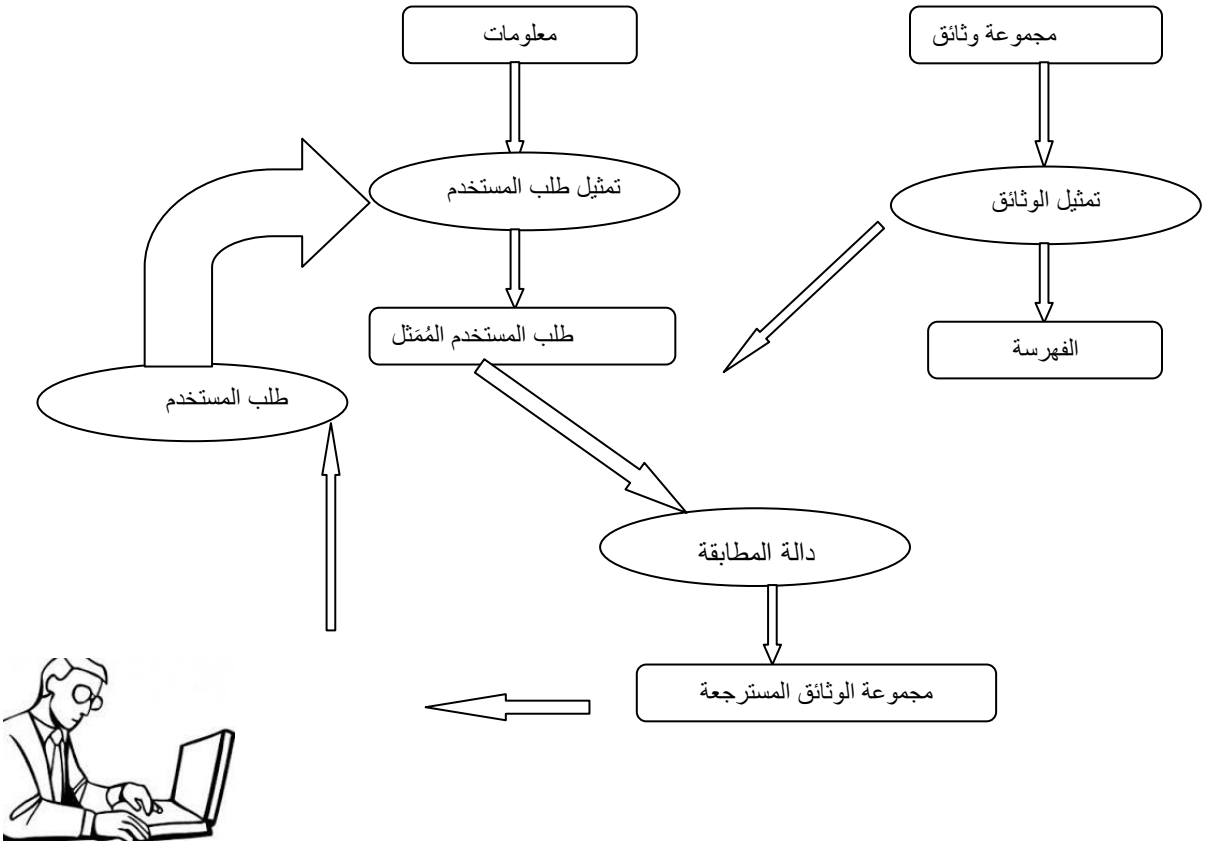
4. **التشذيب أو تحويل الكلمات إلى أصل تشترك فيه (Stemming)**

هو عملية إزالة بوادئ الكلمه ولواحقها وذلك لإنتاج الجذر أو الأصل .

5. **فهرسة الوثائق بناءاً على تشذيب الكلمات (Indexing)**

هي العملية التي يتم فيها تمثيل الوثائق في شكل مجموعة من الكلمات المفتاحية .

وسيتم شرح هذه العمليات لاحقاً في هذا الفصل .



شكل (1.2): العمليات الأساسية في نظام إسترجاع المعلومات

ومن الرسم يتضح أن المشكلة الأساسية هي عملية مطابقة طلب المستخدم مع الوثائق التي يتم إسترجاعها.

1.1.2 العمليات الأساسية في نظام إسترجاع المعلومات

1.1.1.2عملية التمثيل

يتم فيها وضع المحتوى في صيغة أو هيكلية محدده

ويمكن تصنيف عملية التمثيل إلى : .

أ- **تمثيل الوثائق**

بحيث يتم وضع الوثائق المحفوظة في هيكلية تسمح بفهرستها والبحث فيها.

والفهرسة (Indexing) هي التي يتم فيها تمثيل الوثائق في شكل مجموعة من الكلمات المفتاحية(Keyword) والكلمات المفتاحية التي يتم تمثيلها في الفهرس تسمى المصطلح (Term)وهو الوحدة الأساسية المستخدمة في تمثيل الإستفسار ويمكن أن يكون المصطلح عبارة عن كلمة (Word) أو جملة (Phrase) إعتماداً على مانحتاجه من فهرسة الملفات،وعملية إنتاج الفهرس أحيانا تمر بعدد من العمليات وذلك إعتماداً على اللغة مثلاً عملية التشذيب والتقطيع(Tokenizing)، وناتج عملية الفهرسة هو وصف منظم للملفات قابل للبحث في شكل مجموعة من (stemming)

المصطلحات.

ب- **تمثيل طلب المستخدم**

بحيث يتم تمثيل الإستعلام (query) ووضعه في هيكلية محددة لإستخدامه في عملية البحث في الوثائق المفهرسة.

2.1.1.2 عملية المطابقة

يتم فيها مطابقة طلب المستخدم مع الوثائق المفهرسة وإرجاع الوثائق المطابقة وتدعى هذه الوثائق بالوثائق ذات الصلة، وترتَب تنازلياً حسب درجة صلتها بالموضوع. ومن أهم الصفات التي يجب أن تتحقق في هذه المرحلة هي الفاعلية، صحة النتائج، السرعة، قلة نسبة الخطأ، حسن الأداء في الإسترجاع، صغر حجم الفهرس، وعملية الإسترجاع لهذه الوثائق تسمى الإستدعاء أو الإرجاع (Recall) وتزداد هذه الخاصية بزيادة الوثائق المسترجعة.

ومن خلال الوثائق المسترجعة يمكننا تحديد مدى نجاح نظام إسترجاع المعلومات من فشله وذلك بتحديد مدى الصلة بينها وبين طلب المستخدم وهذه تسمى الدقة (Precision) والدقة هي القدرة على تحليل جميع الكلمات المدخلة وإعطاء تقسيم صحيح واحد على الأقل، ولتحقيقها لابد من الاعتناء بأمرين مهمين هما :

1. تحديد هل الوثيقة المسترجعة ذات صلة بموضوع طلب المستخدم أم لا.
2. تحديد ترتيب الوثائق ذات الصلة على حسب صلتها بالموضوع (Ranking)

وكفاءة عملية المطابقة تعتمد على عدة أمور:

1. حجم مجموعة الوثائق.
2. مواضيع تلك الوثائق.
3. ثقافة المستخدم الذي يصيغ طلبه للمعلومات.

2.1.2 نماذج أنظمة إسترجاع المعلومات

1.2.1.2 النموذج المنطقي(Boolean Model)

تصاغ فيه الإستعلامات من قبل مجموعة من المصطلحات (Term) مع العوامل المنطقية (And ,Or ,Not) وهذه العوامل المنطقية يتم التعامل معها في الإسترجاع بطريقة مشابهة لإستخدامها في الجداول الحقيقة التقليدية فبعد بناء الوثائق يتم تمثيلها في شكل مصطلحات (Term) فإذا كان المدخل مساوياً للواحد(1)يعني أن المصطلح مذكور في الوثيقة المعينة، وإذا كان مساوياً للصفر(0) فهذا يعني أنه لم يذكر في الوثيقة. النموذج المنطقي لا يأخذ في الإعتبار عدد مرات ظهور المصطلح (Term Frequency) في الوثيقة بل على حسب ظهور المصطلح في الوثيقة أو عدم ظهوره.

وللنموذج المنطقي جوانب قصور وهي :

أولاً: لا يوفر وثائق لها ترتيب ووزن(Ranked and Weighted) ويركز حول ما إذا كانت الوثيقة مطابقة للاستعلام أم لا.

ثانياً: الإستعلام في النموذج المنطقي معقد نسبياً.

ونلاحظ أنه يتم ترتيب النتائج زمنياً بدلاً من تقدير دقيق لدرجة أهميتها للإستعلام.

2.2.1.2 نموذج الإسترجاع المرتب(Ranked Retrieval Model)

هو نموذج إسترجاع يقوم على تقدير درجة أهمية الوثائق وتحديد أي منها هو الأفضل مطابقة للإستعلام ،في هذا السياق ثبت أن النماذج المرتبة أكثر فعالية من النموذج المنطقي، ومايلي يصف لنا أمثلة النماذج المرتبة.

1. نموذج الناقلات(Vector Space Model)

هو نموذج يستمد إسمه من حقيقة أنه يمثل كل من الوثائق والإستعلامات في شكل ناقل ويتم تمثيل كل مصطلح في بين الوثيقة (similarity) الناقل في أبسط معانيه، ونجد أن الطريقة المتبعة في هذا النموذج لقياس التشابة والاستعلام هي قياس جيب تمام الزاوية بين متجة الوثيقة ومتجة الإستعلام فإذا كانت هذه الزاوية صغيرة هذا يعني أن متجه الوثائق يعتبر ذا صلة أكبر بالنسبة لمتجة الإستعلام .

ا هوالوثيقة ويتم تمثيلة إستناداً على تردد المصطلحات (j) هو المصطلح و (i)حيث أن (**W**ij) ويرمز للوزن هنا بالرمز

والذي يعرف بأنه عدد مرات ظهور المصطلح في كل وثيقة، أو إستناداً على(terms frequencies)

، الذي يستخدم لتحديد أهمية المصطلح فيIDF: TF.IDF scheme (Inverse Document Frequency

مجموعة الوثائق وعادة ما تسخدم الصيغة التالية :

$$\text{cosine(dj ,q)}=\frac{(dj*q)}{||\text{dj }||*||\text{q }||}=\frac{\sum_{i=1}^{|V|} wij*wiq}{\sum_{i=1}^{|V|} wij^2 \ *\sum_{i=1}^{|V|} wiq^2}$$

حيث:-

jd: متجه الوثيقة.

متجه الاستعلام.q:

.j في الوثيقةiوزن المصطلح wij:

q.وزن المصطلح i في متجه الاستعلامwiq:

هو عدد الابعاد في الناقلات وهذا هوالعدد الاجمالي للكلمات الهامه.|V| :

2- نموذج الإسترجاع الإحتمالي(Probabilistic Retrieval Model) :

تقاس (Similarity)في فضاء الناقلات بطريقة عشوائية نوعاً ما مثلاً: النموذج يفترض أن الوثائق التي تكون قريبة في الزاوية من الإستعلام (أي الوثائق التي تكون الزاوية بينها وبين الإستعلام صغيره) تعتبر ذات صلة أكبر بالنسبة لمتجه الإستعلام. لكن في نموذج الإسترجاع الإحتمالي يتم إستخدام طرق أكثر دقة بحيث يتم ترتيب الوثائق على حسب الإحتمالية المقدره لهذه الوثيقة بأن تكون ذات صلة بالإستعلام وهذا يعتبر أساس مبدأ نموذج الإسترجاع الإحتمالي الذي طور من قبل روبرتسون .

في نموذج الإسترجاع الإحتمالي على نظام إسترجاع المعلومات أن يقرر بأن الوثائق إما تنتمى لمجموعة الوثائق ذات الصلة (elevant setr) أو مجموعة الوثائق التي ليس لها صلة بالإستعلام (non relevant) .

ولإتخاذ هذا القرار يفترض وجود مجموعة وثائق ذات صلة، ومجموعة وثائق ليست ذات صلة لهذا الإستعلام ومهمته هي حساب إحتمالية أن تنتمي الوثيقة المعينة إلي مجموعة الوثائق ذات الصلة أو مقارنتها مع إحتمالية أن تكون الوثيقة تنتمي للوثائق ليست ذات الصلة .

إذا كان:

D:الوثيقة

R:ذات صلة للوثيقة

NR : ليست ذات صلة للوثيقة

يمكن حساب الإحتمالية بإستخدام مايسمى Bayes Rule

P(R|D)=p(D\R)*p(R)\p(D)

P(NR|D)=p(D\NR)*p(NR)\p(D)

P(R) : هي إحتمالية (Probabilty) أن الوثيقة (Document) التي تم إسترجاعها بالنظام تنتمي للوثائق ذات الصلة (Relevant Documetents)

P(D\R) : بإفتراض أن لدينا Relevant Document (R) فإن P(D\R) هي إحتمالية أن الوثيقة المسترجعة تنتمي للوثائق ذات الصلة .

P(R|D):هي إحتمالية المطابقة للوثيقة (D)

P(NR|D):الوثائق ليست ذات الصلة

وتعتبر الوثيقة D على أنها ذات صلة أذا كان :

P(R\D)>p(NR\D)

تنتمي إلي الوثائق ذات الصلة D كوزن لتحديد إمكانية أن تكون الوثيقة P(R\D) وp(D\NR) يتم إستخدام

وتعتبر خوارزمية **Best Match25** هي من أشهر خوارزميات نموذج الإسترجاع الإحتمالي المستخدم على نطاق واسع وهي الخوارزميه التي تم استخدامها في هذا البحث.

3.1.2 المراحل الأساسية في عملية تمثيل المعلومات

1.3.1.2 التقطيع (Tokenization)

تعتبر المرحلة الأولى والأساسية في عملية تمثيل المعلومات ويتم فيها تقطيع المحتوى إلى كلمات ذات معنى تسمى قطع (Tokens) وهي أصغر وحدة لها معنى، ويتم ايضاً حذف بعض الحروف الخاصة كعلامات الترقيم والشرطات وغيرها.
الكلمات الناتجة من عملية التقطيع يمكن الإستفادة منها في عملية الفهرسة.
هناك العديد من الإستراتيجيات التي يمكن أن تستخدم للتقطيع منها :

التقطيع بالفراغات

يتم إعتبار أي كلمة بعدها فراغ قطعة مستقلة، ومن عيوب هذه الطريقة أن الكلمات المقتَطعَة قد تكون إختصار أو علامة ترقيم أو ضمير أو كلمات موصولة مثل (عبدالله) أو مفصولة مثل (علاء الدين – هي كلمة واحدة ولكن مفصولة بفراغ) وهذا الإختلاف في القطع قد يؤدى إلى نتائج خاطئة، ويعتبر الفراغ بين الكلمات في اللغة العربية ميزة جيدة لاتوجد في بقية اللغات مثل اللغات الأسيوية لذلك لا يمكن تطبيق هذا النوع من التقطيع على كل اللغات.

ويتم إعتبار اي كلمة بعدها علامة (-) قطعة مستقلة، ومن عيوب هذه الطريقة أن بعض الكلمات قد تحتوي على هذه العلامة كجزء أساسي من الكلمة. فـ مثلاً كلمة " كريم" سيتم تقطيعها إلى قطعتين " ك " و "ريم "، إذا كان

المستخدم يقصد التشبيه بالغزال ريم، فهي تم تقطيعها بشكل جيد. لكن في حالة أنه كان يقصد صفة الكرم فهذه الطريقة تعتبر غير مجدية وستعطي نتائج خاطئة.

2.3.1.2 الكلمات المراد حذفها(Stopwords)

تعتبر المرحلة الثانية في عملية تمثيل المعلومات ويتم فيها حذف الكلمات التي لا تحمل معنى لذاتها والتى تعتبر كلمات ربط كحروف الجر والضمائر وأسماء الموصول وغيرها . مثال لذلك:حروف الجر(by،ك)، وال(Article) (Stopword) مثل(an)،والضمائر مثل (أنت) فمثل هذه الكلمات لا يتم فهرستها. هناك كلمات لا يجب إعتبارها وقد تبين أنه ليس بالصحيح دائماً إزالة (a) فإنه سيتم حذف حرف(Vitamin a) وبالتالي حذفها مثل إذا قلنا ال(Stopwprds) وخاصة في اللغات الإعرابية مثل اللغة العربية وهي غالباً تكون مقيدة في تحديد أجزاء الكلام، ويتم حذف هذه الكلمات الزائدة قبل الترجمة.

أمثلة لبعض الكلمات الزائدة في اللغة العربية : من، إلى، هو،.....،الكلمات المترجمة من اللغات الأخرى، والضمائر، وحروف الجر، العبارات مثل (السيد، العزيز،...).

3.3.1.2 التطبيع (normalization)

بعد تقسيم النص إلى مجموعة من الكلمات المتقطعة (Tokens) في هذه المرحلة يتم توحيد تلك الكلمات المتقطعة ووضعها في صيغة موحدة (Canonical Form) وذلك لذيادة مدى المطابقة بين مجموعة الكلمات المتقطعة، ومن طرق التطبيع التعديلات الإملائية أو مايسمى ب(الإستبدال) وهذه العملية تعتمد على اللغة فمثلاً في اللغة العربية يتم إستبدال (ي) بحرف(ى)، وإستبدال(أ،إ) بحرف (ا)، وأيضا من العمليات التي تتم في هذه المرحله عملية التعديل في الكلمات كتحويلها إلى حروف صغيره (lower case).

4.3.1.2 التشذيب (stemming)

هو عملية إزالة بوادئ الكلمة ولواحقها وذلك لإنتاج الجذر كما في بعض الحالات أو النابعة (الأصل) كما في حالات اخرى، وهذه الطريقة تجمع كل الكلمات التي لها نفس الأصل وتمتلك بعض العلاقات الدلالية بحيث يتم فيها تخطيط وتحويل كل الصيغ المختلفة للكلمة إلى صيغ مشتركة وشكل موحد لتلك الكلمات التي لها نفس الجذر ومثال لذلك إستخدام التشذيب في اللغة الإنجليزية حيث يتم إسترجاع كل الوثائق(documents)التي تحوي ”play,plays,player,playing “وتلك الصيغ الناتجة يجب أن تكون ملائمة لعملية الفهرسة والبحث. وكمثال من اللغة العربية جذر كلمة "طفيليات " هو "طفل " وجذر كلمة أطفال هي أيضاً "طفل" ولكن الكلمتان تختلفان في المعنى .

تبرز أهمية هذه العملية في إجراءات التصنيف وبناء الفهارس وبحث وإسترجاع المعلومات وذلك لكونها تقلل من تأثير إختلاف الأنماط والأشكال للكلمات العربية، كما أنه يساهم في تقليل الحجم المطلوب لخزن الكلمات في حالة تخزينها في الفهارس بأشكالها المختلفة والموجودة في النصوص الأصلية.

للتشذيب تأثير كبير على الاستدعاء(recall) بمعنى أن عدد الوثائق المسترجعه كبير،وإذا كانت تلك الوثاق المسترجعه بعد عملية التشذيب ذات صله عاليه بطلب المستخدم (query) هذا يعني نجاح ودقة تلك العملية (precision) . وسنقوم بتوضيح الإستدعاء والدقة لاحقاً .

طرق التشذيب

هنالك العديد من طرق التشذيب لكن أغلبها يعتمد اعتماداً كُلياً على اللغة التي يتم التعامل معها.وفيما يلي سيتم توضيح طريقين من طرق التشذيب :

أ-التشذيب الخفيف (Light Stemming)

في هذا النوع يتم تجريد الكلمات من السوابق واللواحق بحيث تستخدم الكلمات المجردة لفهرسة الوثائق وغالباً ما تكون الكلمات التي تجرد من نفس الكلمة لها نفس المعني و بذلك فإن هذه الطريقة تقوم باسترجاع عدد الوثائق ذات الصلة بطلب المستخدم و نتيجةً لذلك فإن هذا النوع من التشذيب يحِد من نطاق البحث بصورة مؤثرة.

خصائصه :

1- غير فعَّال في حالة وجود كلمتين لهما نفس المعنى ويختلفان في الميزان الصرفي مما يقلل من إحتمالية المطابقة و هذه المشكلة تعرف بالـقصور في التشذيب (under-stemming).
2- لم يساعد في التعرف على أجزاء الكلام (part of speech) أي معرفة نوع الكلمة في مجموعة الوثائق هل هي (إسم، فعل، صفه حال (الظروف الزمانية، الظروف المكانية)) فمثلا الإسم والفعل اللذان لهما نفس الصيغة (single form) عند التشذيب يتم إسترجاع الوثائق التي وردت في كليهما .

ب-التشذيب بإيجاد جذر الكلمة (Root-based Stemming)

في هذا النوع يتم إستخراج جذور الكلمات الموجودة في المستند وذلك بإجراء بعض العمليات اللغوية عليها ومن ثم يتم إستخدام الجذور المستخرجة ككلمات دلالية (Indexing Terms) في الفهرس. التشذيب بإيجاد جذر الكلمة قائم على الجذر حيث يقوم

بعمليات لغوية حتى يستخرج أصل الكلمة (root)، ويهتم هذا النوع من التشذيب بإستخراج جذور الكلمات الموجودة في الوثيقة وإستخدامها ككلمات دلالية (Indexing Terms)، وبذلك فإن جميع الكلمات الواردة في الوثيقة الواحدة والتي لها نفس الجذر ستُفَهرس بنفس الكلمة بالرغم من أنه ليس بالضرورة أن يكون لها نفس المعنى مما يؤدي إلى تقليل حجم الفهرس.

ومن خصائصه :

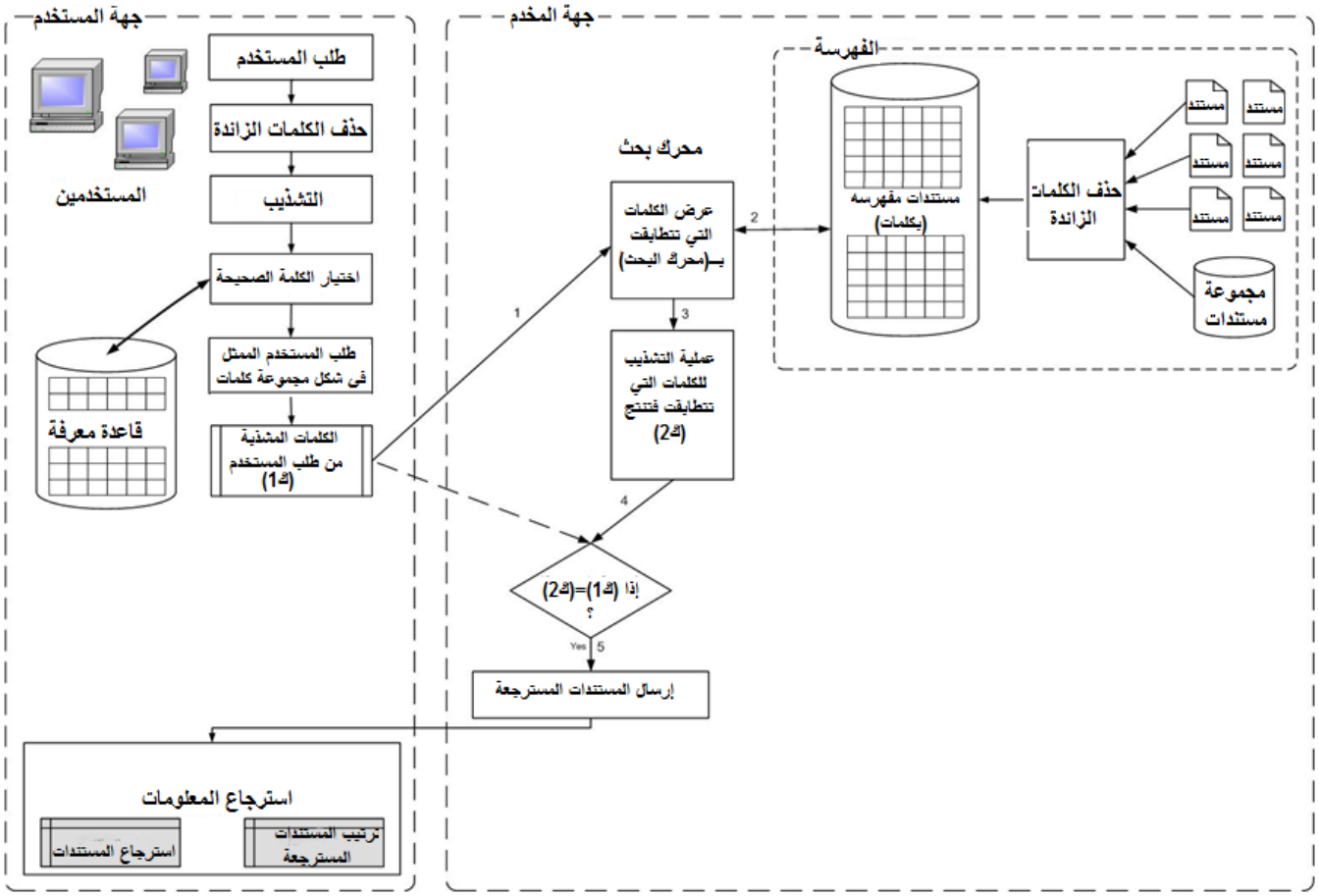
أ- أن هذه الطريقة تسترجع للمستخدم جميع الوثائق التي تحتوي على أي صيغة صرفية للكلمات الواردة في طلب المستخدم مما يزيد من إمكانية عثور المستخدم على المعلومة المطلوبة .

ب- في حالة مجموعات الوثائق الضخمة جداً والمتجددة بإستمرار كصفحات الإنترنت فإن هذه الطريقة غير مجدية لأنها ستوسع كثيراً من نطاق البحث ولن تصل بالمستخدم إلى المعلومة المطلوبة.

ت- يقوم بتقليل حجم الفهرس.

ث- يركز على إستدعاء أكبر كمية من الوثائق التي تشترك كلماتها في الأصل مع كلمات طلب المستخدم ولكن بعض الكلمات لها نفس الأصل وليست لها علاقة ببعضها البعض من ناحية المعنى مما أدى إلى كثرة في الإستدعاء وإسترجاع الوثائق التي يريدها المستخدم والتي لا يريدها مثلاً كلمة "يذهبان" و"ذهب"(gold) لهما نفس الجذر وهو "ذهب" ولكنهما يختلفان في المعنى، فبالتالي تَجِد هذه الطريقة من الدقة، و هذه المشكلة تعرف بالتشذيب أكثر من اللازم (over-stemming).

كل التفاصيل التي ذُكرت في النقاط السابقة يمكن إجمالها في الشكل



الشكل (2.2) يوضح المراحل الاساسية في تمثيل المعلومات :

يمكن قياس أداء أنظمة إسترجاع المعلومات بطرق مختلفة إعتماداً على مهمة الإسترجاع وذلك على حسب الوثائق المسترجعة سواء أنها كانت ذات صلة أو ليست ذات صلة بطلب المستخدم لتقييم تلك الوثائق.ومن هذه الطرق:

1- الإستدعاء(Recall)

هي نسبة تعبر عن عدد الوثائق المسترجعة وذات صلة بالإستعلام من مجموعة الوثائق الكلية

$$Recall = \frac{\text{number of retrieved relevant documents}}{\text{number of relevant documents in the collection}}$$

2- الدقة (Precition)

هي نسبة تعبر عن عدد الوثائق المسترجعَة ذات الصلة من مجموعة الوثائق المسترجعة

$$Precition = \frac{\text{number of retrieved relevant documents}}{\text{number of retrived documents}}$$

ولقياس جودة ترتيب الوثائق نستخدم مقياس DCG (Discounted Cumulative Gain) ، وهذا المقياس غالباً ما يستخدم لقياس فعالية خوارزميات محركات البحث أو التطبيقات المشابهة لها .

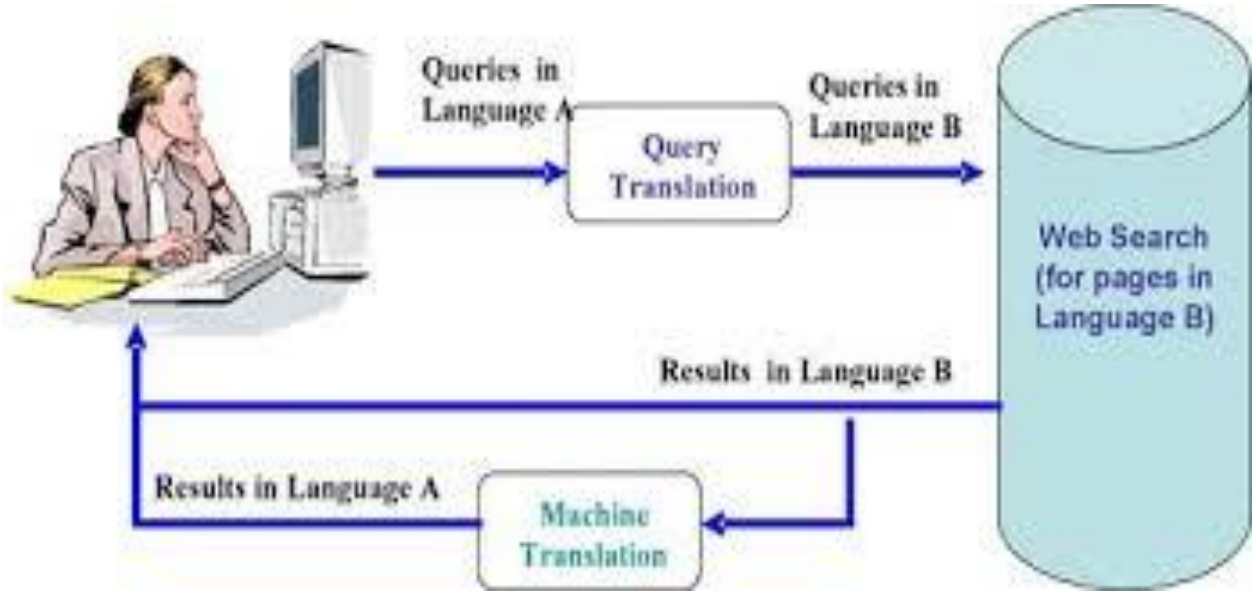
2.2 الدرسات السابقة

في أنظمة إسترجاع المعلومات ذات اللغة الواحدة (Monolingual IR) يتم تمثيل كل من الإستعلامات ومجموعة الوثائق بلغة واحدة، وتكون النتيجة المسترجعة بنفس اللغة ولكن طبيعة المعلومات المتاحة على الشبكة العنكبوتية أنها توجد بلغات مختلفة، فقد يحتاج المستخدم إلى كتابة الإستعلام بلغة (مختلفة عن لغة الوثيقة) ويريد النتيجة بلغة أخرى. أو كتابة الإستعلام نفسه بأكثر من لغة ومن هنا ظهر مفهوم CLIR و MLIR وفيما يلي توضيح لكل منهما .

1.2.2 أنظمة الإسترجاع ثنائية اللغة(CLIR)

فيها يقوم المستخدم بكتابة الإستعلام بلغة واحدة مختلفة عن لغة الوثيقة ربما لأن المستخدم لا يجيد لغة الوثيقة ولكن تكون النتيجة بلغة مختلفة عن لغة الإستعلام .وتسمى اللغة المكتوب بها الإستعلام باللغة المصدر، وتسمى لغة الوثائق المسترجعة باللغة الثانوية (Languge Target).

حيث تتم ترجمة الإستعلام من اللغة المصدر إلى اللغة الثانوية بإستخدام أحد طرق الترجمة والبحث بهذا الإستعلام في مجموعة الوثائق لإسترجاع النتائج.



شكل (3.2) : مفهوم انظمة الأسترجاع ثنائية اللغة

ونظراً لأن الوثيقة الواحدة على الشبكة العنكبوتية قد تكون مكتوبة بأكثر من لغة خصوصاً الوثائق المكتوبة بلغة غير الإنجليزية أو أن المستخدم يريد البحث عن بعض المصطلحات التي لايستطيع التعبير عنها بلغته، ينتج عن ذلك كتابة المستخدم للإستعلام بأكثر من لغة ومن هنا ظهر مفهوم أنظمة إسترجاع المعلومات متعددة اللغات (MLIR).

2.2.2 أنظمة إسترجاع المعلومات متعددة اللغات(Multilingual Information Retrieval)

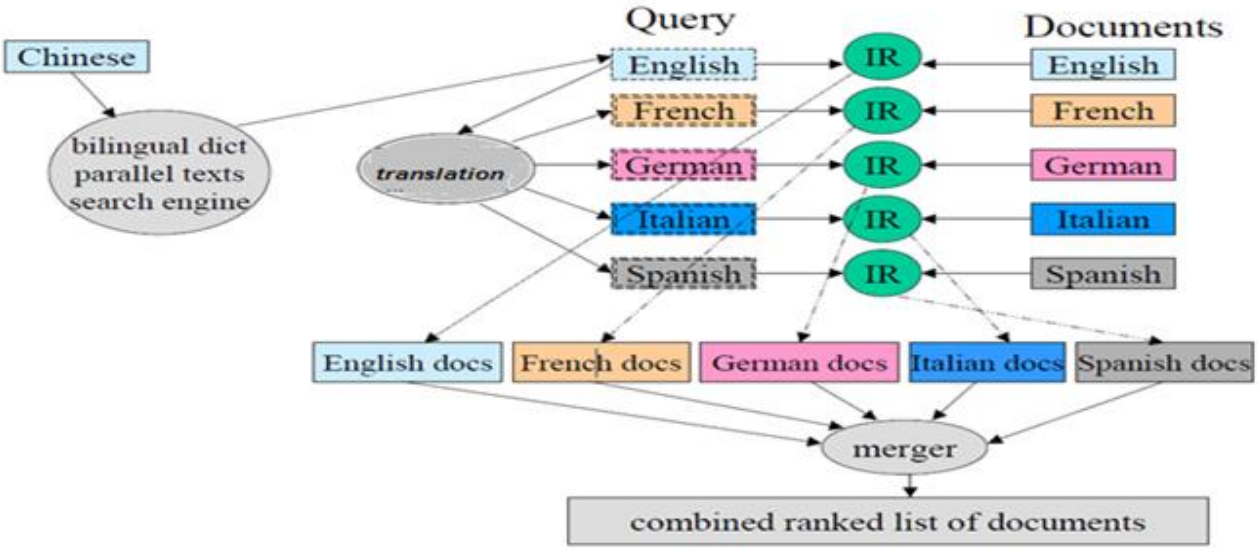
MLIRهو مهمة البحث عن الوثائق ذات الصلة في مجموعة من الوثائق سواء كانت بلغة واحدة أو متعددة اللغات لإستعلام مكتوب بأكثر من لغة، ويتم فيها ترجمة الإستعلام إلى مجموعة من اللغات والبحث بهذه الإستعلامات المترجمة في مجموعات الوثائق(البحث بكل إستعلام مترجم في مجموعة الوثائق التي لها نفس لغته) ثم دمج الوثائق المسترجعة من عملية البحث بكل إستعلام وعرض قائمة وثائق موحدة الترتيب (ranked) للمستخدم بغض النظر عن اللغة.

الإستعلام متعدد اللغات

هو إستعلام مكتوب باكثر من لغة، مثلا "مفهوم ال deadlock"ومعنى هذا الاستعلام بالعربية هو مفهوم لإقفال وهذا الاستعلام مكتوب باللغتين العربية والإنجليزية وغالباً ما تكون الاجزاء الانجليزية في الاستعلام هي عبارة عن مصطلحات .

الوثائق متعددة اللغات

هي وثائق مكتوبة بأكثر من لغة ، وفي هذه الوثيقة تكون هنالك لغة أساسية ولغات أخرى ثانوية، وغالباً ما تكون اللغة الثانوية هي اللغة الإنجليزية .



شكل(4.2) : أنظمة إسترجاع المعلومات ذات اللغات المتعددة(MLIR)

3.2.2 الطرق المتبعة في فهرسة الوثائق

كما علمنا سابقاً أن عملية الفهرسة تعني تمثيل الوثائق في شكل مجموعة من الكلمات المفتاحية وتلك الكلمات المفتاحية التي يتم تمثيلها في الفهرس تسمى المصطلح (term)وهو الوحدة الأساسية المستخدمة في تمثيل الإستعلام والوثيقة.

في البداية كانت أنظمة إسترجاع المعلومات تقوم بفهرسة الوثائق بطريقة مركزية وذلك بعمل فهرس واحد لجميع الوثائق سواء كانت أحادية أو متعددة اللغات، وطريقة أخرى للفهرسة هي الفهرسة الموزعة أي لكل لغة فهرس منفرد، وطريقة ثالثة تقوم بالدمج بين الفهرسة المركزية والفهرسة الموزعة لتلافي عيوب كلاً منهما وسيتم ذكر هذه العيوب لاحقاً.

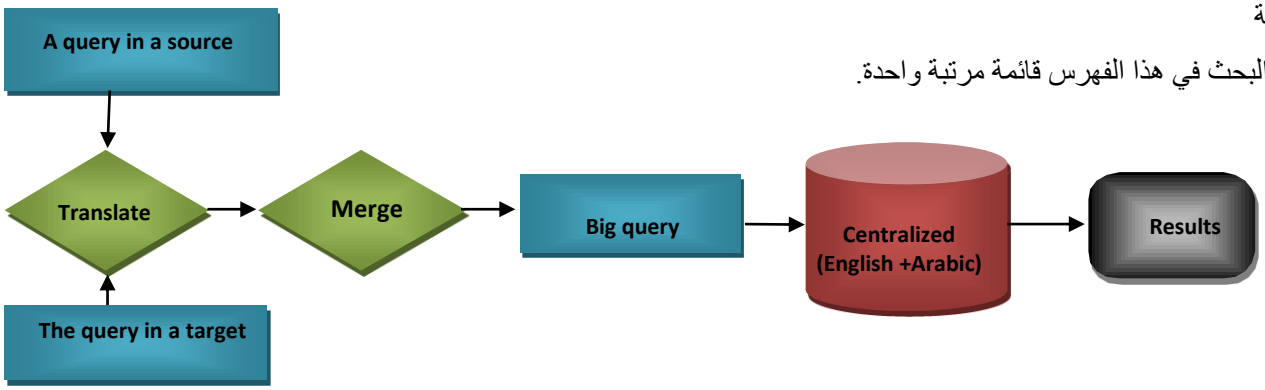
وفيما يلي شرح لطرق الفهرسة كما تم تصنيفها مسبقاً إلى:

1. طريقة الفهرسة المركزية (Centralized Indexing Approach)

في هذه الطريقة يتم بناء فهرس مركزي وحيد لكل الوثائق بإختلاف لغاتها،أي تتم فهرسة الوثائق العربية والإنجليزية وذات اللغات المتعددة في ذات الفهرس. وعندما يقوم المستخدم بكتابة إستعلام معين بغض النظر عن لغته سواء كان بلغة وحيدة أو بأكثر من لغة فإنه يتم البحث بهذا الإستعلام في الفهرس المركزي الذي تم بناءه لكل الوثائق بلغاتها المختلفة، وبعد عملية البحث يتم إسترجاع النتائج المطابقة لطلب المستخدم.

مثال لذلك إذا كان المستخدم يبحث عن (مفهوم ال deadlock) في فهرس مركزي يضم وثائق عربية ووثائق إنجليزية ووثائق ذات لغتين(عربي-إنجليزي)، يتم ترجمة الجزء العربي من الإستعلام إلى اللغة الإنجليزية وترجمة الجزء الإنجليزي إلى اللغة العربية ودمج الأجزاء المترجمة في إستعلام واحد ليصبح (مفهوم الاقفال deadlock concept)، ويتم البحث بهذا الاستعلام في الفهرس المركزي وبالنسبة لمجموعة الوثائق العربية يتم تجاهل وزن الجزء الانجليزي من الإستعلام (أي وزنه يساوي صفر)، وبالنسبة لمجموعة الوثائق الإنجليزية وفي هذه الحالة يتم تجاهل وزن الجزء العربي من الإستعلام (أي وزنه يساوي صفر)، وبالنسبة للوثائق ذات اللغتين لن يتم تجاهل وزن أي من الجزء العربي والجزء الإنجليزي من الإستعلام، وتكون نتيجة

البحث في هذا الفهرس قائمة مرتبة واحدة.

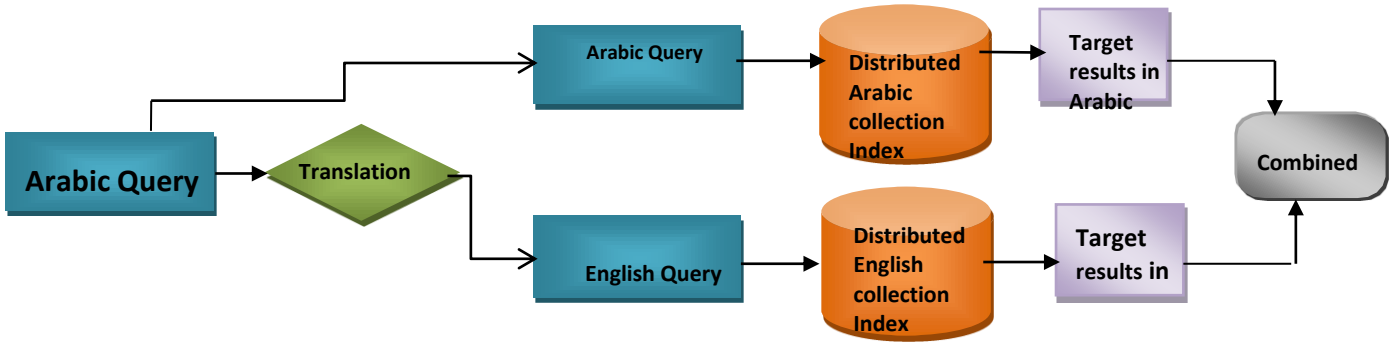


شكل(5.2) طريقة الفهرسة المركزية

2. طريقة الفهرسة الموزعة(*Distributed* Indexing Approach)

في هذه الطريقة تتم فهرسة الوثائق بطريقة موزعة أي لكل لغة فهرس منفرد، ويعتمد هذا المنهج على توزيع الوثائق وفقاً للغاتها، تتم فهرسة الوثائق العربية في فهرس عربي أحادي (monolingual arabic) وفهرسة الوثائق الإنجليزية في فهرس أحادي أخر(monolingual English) والوثائق متعددة اللغات يتم فهرسة محتوياتها على حسب لغاتها.

مثال لذلك إذا قام المستخدم بكتابة إستعلام "مفهوم العولمة" يتم ترجمة هذا الإستعلام إلى اللغة الإنجليزية "**globalization concept**" ومن ثم يتم البحث بالإستعلام العربي("مفهوم العولمة") في الفهرس العربي وإسترجاع النتيجة، ومن ثم البحث بالإستعلام الإنجليزي (**globalization concept**)في الفهرس الإنجليزي ودمج تلك النتائج.

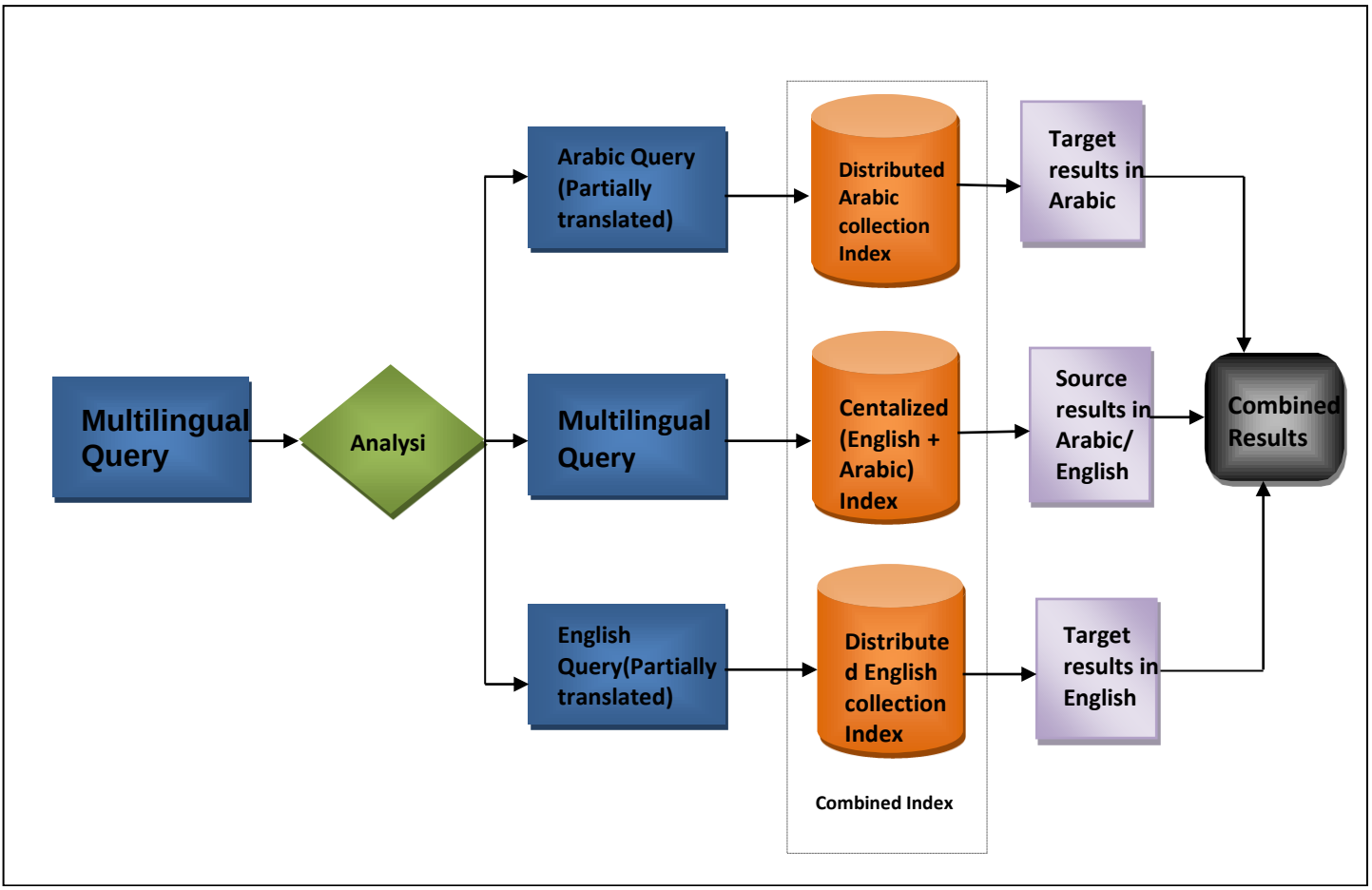


شكل(6.2) يوضح الفهرسة الموزعة

3. الطريقة المركزية الموزعة(Hybrid approach of Indexing)

تدمج هذه الطريقة بين الفهرسة المركزية والفهرسة الموزعة للإستفاده من مميزات كل منهما وفي هذه الطريقة تتم فهرسة الوثائق أحادية اللغة كما تتم فهرستها في الطريقة الموزعة وذلك ببناء فهرس منفرد لكل لغة، وفهرسة الوثائق متعددة اللغات بالطريقة المركزية وذلك ببناء فهرس مركزي لكل من الوثائق متعددة اللغات(عربي-إنجليزي).

مثال لذلك إذا كان المستخدم يبحث عن (مفهوم ال inheritance) يقوم محرك البحث بتحويل الإستعلام الي اللغة العربية (monolingual arabic) مرة بحيث يصبح (مفهوم الوراثة) ويتم البحث به في الفهرس العربي، ومرة أخرى الي الانجليزية (monolingual English)ويصبح (inheritance concept) بحيث يتم البحث به في الفهرس الإنجليزي.وإ ستعلام ثالث بدمج الإستعلام باللغتين العربية والانجليزية(مفهوم الوراثة inheritance concept)، ويتم البحث بهذا الإستعلام في الفهرس المركزي متعدد اللغات بحيث يتم تطبيع وزنه والبحث به في ذلك الفهرس ومن ثم دمج النتائج المسترجعة من كلا من الفهارس.



شكل(5.2) طريقة الفهرسة المركزية

4.2.2 خوارزميات دمج النتائج (Results Merging (Algorithms

1. خوارزمية القيمه الخام (Raw Score)

في هذه الخوارزمية يتم دمج الوثائق من قائمة النتائج المسترجعة من مجموعة الوثائق العربية مع قائمة النتائج المسترجعة من مجموعة الوثائق الإنجليزية وقائمة الوثائق المسترجعة من مجموعة الوثائق متعددة اللغات، وذلك بأخذ وثيقة من العربية ووثيقة من الإنجليزية وأخرى من متعددة اللغات ومن ثم تقوم بعرضها للمستخدم.

2. خوارزمية تطبيع الوزن(Normalized score merging)

تعتمد هذه الخوارزمية على أكبر وزن في قائمة الوثائق المسترجعة وأيضاً على أقل وزن في تلك القائمة بحيث يتم تطبيع وزن(score) كل وثيقة بطرح وزنها من أكبر وزن في القائمة مقسوماً على المدى(أكبر وزن - أقل وزن) ومن ثم دمج النتائج المسترجعة من كل فهرس بناءً على هذا الوزن المطبع، فمن عملية البحث يتم إسترجاع ثلاث قوائم الأولى تحتوي على وثائق عربية والثانية تحتوي على وثائق إنجليزية والثالثة تحتوي على وثائق باللغتين(عربي-إنجليزي) وتقوم بدمج تلك الوثائق بأخذ الوثيقة الأولى من كل قائمة من القوائم المطبعة ومقارنة أوزانها ووضع الوثيقة ذات الوزن الأعلى في بداية القائمة التي سيتم عرضها للمستخدم.

3. خوارزمية CORI (Collection Retrieval Inference Network)

لكي نقوم بتطبيع الوزن قمنا بإستخدام هذه الخوارزمية لأنها تعتبر اكثر دقة من سابقاتها وفيها يتم تطبيع الوزن بناءً على طول الوثائق المسترجعة من كل مجموعة. بحيث يكون الوزن النهائي المطبع لاي وثيقه عبارة عن وزن الوثيقة مضروباً في وزن المجموعة التي تنتمي له، ونستخدم ذلك الوزن المطبع لدمج تلك النتائج .

والفكرة الأساسية لهذه الخوارزمية هي زيادة وزن الوثائق التى لها وزن أعلى من الوزن المتوسط، ويقل وزن الوثائق التي لها وزن أقل من الوزن المتوسط . ومن دواعي استخدامنا لهذه الخوارزمية انها بسيطة لأن مدخلاتها هي وزن الوثائق وطول النتائج المسترجعة . وايضاً هذه الخوارزمية لا تحتاج الى معلومات عن مجموعة الوثائق وبالتالي فإن الوسيط (Broker) الذي يتعامل مع هذه الخوارزمية لا يحتاج الى تخزين معلومات عن مجموعة الوثائق . ولكن اذا كانت التنبؤات عن مجموعة الوثائق مطلوبة في بيئة متغيرة او حركية (مثل الويب) فإن هذه المعلومات تحتاج الى التحديث بصورة دورية وهذا غير ممكن الا بوجود تعاون بين الوسيط وخوادم المجموعات (Collection Server).

في هذا الباب سيتم توضيح الطرق المتبعة في عملية فهرسة الوثائق والخوارزميات المتبعة في دمج النتائج المسترجعة من عملية البحث في مجموعة الفهارس والتقنيات التي تم إستخدامها لتكوين بيئة التجربة .

1.3 الطرق المتبعة في فهرسة الوثائق

1.1.3 الطريقة المركزية

في البداية كانت أنظمة إسترجاع المعلومات تقوم بفهرسة الوثائق بطريقة مركزية وذلك بعمل فهرس واحد لجميع الوثائق سواء كانت أحادية أو متعددة اللغات . ويتم حساب الوزن للمصطلح الوارد في الإستعلام والوارد في الوثيقة كالآتي :

TF * log (N/DF)

بحيث أن :

TF : عدد مرات ظهور المصطلح في الوثيقة المعينة

DF : عدد الوثائق التي ورد فيها المصطلح

N: عدد كل الوثائق في المجموعة المعينة

ولكن لهذه الطريقة عدة عيوب وعيبها الأساسي هو زيادة الوزن (Overweighting) ويعني أن أوزان الوثائق التي تنتمي إلى المجموعة التي تحتوي على وثائق بسيطة سيكون أعلى من أوزان الوثائق التي تنتمي إلى المجموعة التي تحتوي على وثائق أكثر.وذلك لأنه سيزيد عدد الوثائق(لأنه يتم وضعها في مجموعة واحدة أي أن عدد الوثائق N سيزيد) ولكن عدد تكرارت المصطلح او عدد مرات ظهوره (TF) و عدد الوثائق التي ورد فيها المصطلح(DF) ستظل كما هي من دون تغيير مما يؤدي إلى زيادة أوزان المصطلحات التي تظهر في المجموعة التي تحتوي علي وثائق بسيطة فمثلاً إذا كان عدد وثائق اللغة العربية (2,000 وثيقة) وعدد وثائق اللغة الإنجليزية (18,000 وثيقة) وتم دمج هذه الوثائق في مجموعة واحدة (مركزية) سيصل عدد الوثائق الكلية إلى (20,000 وثيقة)، ولذلك ستكون هناك زيادة في وزن الوثائق العربية لأن المصطلحات الواردة في الوثائق العربية تجعلها مميزة أكثر من نظيراتها في اللغة الإنجليزية [1] .

وعند التحدث عن الإستعلامات و الوثائق متعددة اللغات سيكون هنالك عيب آخر لهذه الطريقة وهو أن المصطلحات المتشابهة عبر اللغات تحسب أوزانها بطريقة منفصلة على أساس أنها مصطلحات مختلفة، فمثلاً إذا كان لدينا الإستعلام الآتي ("الوراثة Inheritance" حيث أن كل مصطلح هو ترجمه للآخر)سيكون وزن المصطلح "وراثة" في الفهرس المركزي مختلف عن وزن المصطلح "Inheritance" وسيتم جمع الوزنين للمصطلحين أعلاه للبحث في الفهرس وعند البحث فإن الوثائق العربية سيوجد فيها المصطلح"وراثة" فقط وفي الوثائق الانجليزية سيوجد المصطلح"Inheritance" فقط ولكن في الوثائق متعددة اللغات(عربي-انجليزي)سيوجد المصطلحان مما يزيد من وزن الوثائق متعددة اللغات وهذا هو السبب الذي يؤدي إلى هيمنة الوثائق متعددة اللغات على بداية القائمة المسترجعة .احياناً يكون المصطلح مصحوباً بترجمته في الوثائق خاصة في الوثائق ذات اللغات غير الانجليزية مما ينتج عنه تكرار المصطلح "Co-occurring Terms " فمثلاً مصطلح "الإقفال " في الوثائق ذات اللغة العربية قد يكون مصحوباً بترجمته "deadlock" مما ينتج عنه تكرار للمصطلح لكن بلغات مختلفة وهذه الخاصية موجودة على عدد معقول من الوثائق غير الإنجليزية على الشبكة العنكبوتية. وعندما يأتي الحديث عن الإستعلام ات متعددة اللغات في الفهارس المركزية فإن هذه المشكلة "Co-occurring Terms " تؤدي إلى زيادة أوزان الوثائق متعددة اللغات مما يجعلها تكسب اوزاناً إضافية هي ليست جزءاً حقيقياً من وزنها الأصلي فمثلاً إذا كان لدينا وثيقتين هما

و1، و2 . وكانت الوثيقة **و1** وثيقة متعددة اللغات حيث فيها اللغة العربية هي اللغة الرئيسية بالاضافة الى اللغة الإنجليزية كلغة ثانوية وكانت الوثيقة **و2** هي وثيقة ذات لغة احادية وهي الانجليزية

و1:" تؤدي عملية التطبيع normalization لإنشاء مجموعة جدوال tables ذات..."

و2:" The process of normalization leads to the creation of tables, whose…"

ولكن الوثيقة و2 هي بالضبط ترجمة للوثيقة **و1** فإذا كان الإستعلام كالتالي "التطبيع normalization"هذا الإستعلام يجعل الوثيقة متعددة اللغات **و1** ذات صلة أعلى لذلك تكون في أعلى قائمة الوثائق المسترجعة لكن يمكن ان توجد وثائق اخري ذات صلة اكبر بالإستعلام من هذه الوثيقة لكن تكرار المصطلحات في الوثيقة متعددة اللغات أدَى إلى زيادة وزنها مما جعلها تعتلي قائمة الوثائق المسترجعة وهذه من المشاكل الرئيسية في الوثائق والإستعلام ات متعددة اللغات في الطريقة المركزية[6] .

ولكن من مميزات الفهرسة المركزية أنها تقوم بعملية البحث والإسترجاع من فهرس واحد بغض النظر عن لغة الوثيقة وبالتالي لا تتطلب هذه الطريقة عملية دمج لتلك الوثائق وهذه الميزة مفيدة جداً للوثائق ذات اللغات المتعددة وذلك لأن عملية الإسترجاع من الفهرس الواحد ذات كفاءة أعلى من عملية الإسترجاع من الفهارس المتعددة [1] .

2.1.3 الطريقة الموزعة

ولحل مشكلة زيادة الوزن في طريقة الفهرسة المركزية تمت فهرسة الوثائق بطريقة موزعة أي لكل لغة فهرس منفرد ، ويعتمد هذا المنهج على توزيع الوثائق وفقاً للغاتها[2].

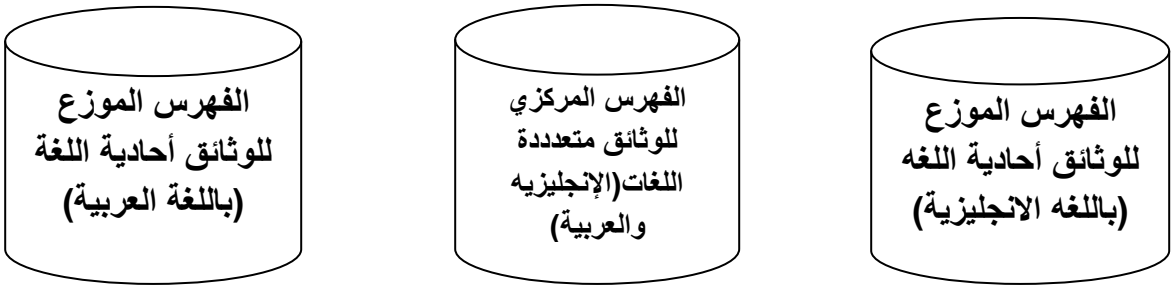
ولكن عند التحدث عن هذه الطريقة لن تكون مشكلة الوثائق المتعددة فقط محصورة على الوزن وإنما تمتد الى كيف ستتم فهرسة الوثائق متعددة اللغات ؟ فكما ذكر سابقاً أن هذه الطريقة تعتمد على تقسيم الوثائق و فهرسة محتوياتها على حسب اللغة [8]. ولكن عند إجراء عملية التقسيم هذه على الوثائق متعددة اللغات (على حسب اللغات الواردة فيها) هذا يجعل المعلومات والمصطلحات الموجودة في الوثيقة متعددة اللغات تفقد قيمتها ومعناها ،بالإضافة إلى ذلك إذا تم تقسيم الوثائق متعددة اللغات وفهرستها على عدد من المجموعات الفرعية (sub-collections) إعتماداً على اللغات المكونة منها تلك الوثيقة ، فإن ذلك سيؤدي الى نقصان وزن الوثيقة متعددة اللغات "underweighting" في كل مجموعة فرعية لأن الوزن يتم حسابه فقط من جزء الوثيقة الموجود في المجموعة الفرعية وليست من الوثيقة كاملة وهذا يجعل الوثائق متعددة اللغات غير قابلة للمقارنة مع الوثائق احادية اللغات(لأن الوثيقة احادية اللغة لن يتم تقسيمها إلى وثائق فرعية فبالتالي لا تفقد معناها ولا ينقص وزنها)مما يجعل الوثائق متعددة اللغات حتى وإن كانت ذات صلة أعلى بالإستعلام ان تكون في نهاية القائمة المسترجعة ، وهذا هو الفرق بين الطريقة الموزعة والطريقة المركزية حيث أنه في الطريقة المركزية تتم فهرسة الوثيقة متعددة اللغات بإفتراض أنها لغة أحادية مما يؤدي إلى زيادة اوزان الوثائق متعددة اللغات وفي الطريقة الموزعة يتم فهرسة الوثيقة متعددة اللغات بتقسيمها إلى وثائق فرعيه حسب اللغات المكونة لها مما يؤدي إلى نقصان اوزانها [2].

هنالك فائده كبيره في الفهارس الموزعة أحادية اللغه وهي أن تكون عملية الإسترجاع منها ذات كفاءة عالية وذلك نتيجه للمطابقه بين الوثيقة والإستعلام الذي تم ترجمته،خاصة مع الجزء الواحد المترجم للإستعلام ذو اللغات المتعددة.

3.1.3 الطريقة المركزية الموزعة

بعد التعرف علي الطريقتين أعلاه و معرفة عيوب كل منهما كان لابد من استخدام طريقة للحد من سلبيات كل منهما وتمكين المستخدم من البحث باستخدام الإستعلام ات متعددة اللغات في الوثائق متعددة اللغات وان تكون نتيجه البحث تضم الوثائق ذات الصلة الأعلى بالإستعلام بغض النظر عن اللغه المستخدمه في كتابته او امكانيه المستخدم في التعبير عن احتياجاته . لذلك تم إستخدام هذه الطريقة التي تدمج بين الفهرسة المركزية والفهرسة الموزعة لبناء الفهرس في هذا البحث للحد من سلبيات كلا من الطريقتين السابقتين والإستفاده من فوائدهما.

و هذا النهج الهجين من كلتا الآليتين يساعد في التعامل مع الوثائق أحادية اللغة و متعددة اللغات بصورة فعالة ووفقا لذلك من المحتمل أن يكون ذلك النهج الهجين أكثر ملائمه من بقية الطرق وذلك ببناء فهرس أحادي اللغه للوثائق أحادية اللغه (monolingual) وهذا يكون مثل الطريقة الموزعة وبناء فهرس اخر للوثائق ذات اللغات المتعددة (الطريقة المركزية) كما هو مبين في الشكل التالي(باعتبار اللغتين الإنجليزية و العربية):



شكل (1.3) يوضح إستخدام الفهرسة التي تدمج بين الموزعة والمركزية

وهذه الطريقة المقترحة تقلل من مشكلة زيادة الوزن(Overweighting) الموجودة في الفهرسة المركزية إلى أدنى مستوى وذلك لأن الوثائق أحادية اللغة لن تتم فهرستها في الفهرس المركزي(أي لن يتم دمجها في فهرس واحد مع الوثائق متعددة اللغات) وبالتالي لن يزيد عدد الوثائق الكلي الموجود في المجموعة وهذا يحد من مشكله زيادة الوزن . إلى جانب حل مشكلة زيادة الوزن في الطريقة المركزية من ناحية أخرى هذه الطريقة تضم فوائد الفهرسة المركزية التي تقوم بعملية البحث والإسترجاع من فهرس واحد بغض النظر عن لغة الوثيقة كما ذكر سابقاً بالتالي يمكن بناء فهرس مركزي وحيد يحوي الوثائق ذات اللغات المتعددة أما الوثائق ذات اللغه الواحده يتم فهرستها بالطريقة الموزعة.

أيضاً هذه الطريقة المقترحه تحد من مشكلة نقصان الوزن(underweighting)بالنسبة للوثائق متعددة اللغات في الطريقة الموزعة لانه لايتم تقسيم الوثيقة متعددة اللغات إلى مجموعة من الوثائق الفرعية أي لايتم فصل محتواها على حسب اللغة لذلك لن تضيع المعلومات الموجودة في تلك الوثائق ولن تفقد قيمتها وعلى عكس الطريقة الموزعة هذا النوع من الفهرسة يجعل أوزان الوثائق ذات اللغات المتعددة مكافئة لغيرها من الوثائق الأحادية(لانه لا يتم تقسيم الوثيقة متعددة اللغات).

فمثلاً إذا كان لدينا ثلاث فهارس أحدهم للغة العربية يحتوي على مجموعة الوثائق ذات اللغة العربية فقط وفهرس للغة الانجليزية يحتوي على مجموعة الوثائق ذات اللغة الانجليزية فقط و الفهرس ذو اللغتين للوثائق التي تحتوي على اللغتين معاً فإذا كان المستخدم يبحث عن (مفهوم ال deadlock) فإن عملية البحث ستتم كالتالي: يقوم محرك البحث

بتحويل الإستعلام الي اللغه العربية (monolingual arabic) مرة وذلك بترجمة الأجزاء الإنجليزية في الإستعلام بحيث يصبح (مفهوم الإقفال) ويتم به البحث في فهرس الوثائق العربية (ويتم تجاهل وزن الجزء الانجليزي من الإستعلام أي أن وزنه يساوي صفر) ويتم تحويل الإستعلام الي اللغة الانجليزية (monolingual English) وذلك بترجمة الاجزاء العربية في الإستعلام بحيث يصبح (deadlock concept) ويتم البحث به في فهرس الوثائق الإنجليزية (ويتم تجاهل وزن الجزء العربي من الإستعلام أي أن وزنه يساوي صفر) ' ويتم دمج الإستعلامين أعلاه(العربي،الانجليزي) ليصبح الإستعلام (مفهوم الاقفالdeadlock concept، ويستخدم هذا الإستعلام للبحث في مجموعة الوثائق ذات اللغتين بحيث تنتج من عملية البحث في الثلاثة فهارس ثلاث قوائم تحتوي على الوثائق التي تعتبر ذات صلة أعلى بالإستعلام بحيث تكون هذه القوائم مرتبة ترتيباً تنازلياً إعتماداً على أوزان الوثائق في كل قائمة وهنا تكمن المشكلة في كيفيه دمج النتائج المسترجعة من الفهارس الثلاثة في قائمة واحدة تظهر للمستخدم بحيث

تكون فيها الوثائق الاكثر صلة بالإستعلام في أعلى القائمه المسترجعة .

2.3 خوارزميات دمج النتائج(Results Merging Algorithms)

1.2.3 خوارزميه القيمة الخام (Raw Score)

في هذه الخوارزمية لا يتم تطبيع وزن الوثائق او تعديله وإنما يتم دمج القوائم المسترجعة من الفهارس الثلاثة (عربي ،انجليزي،متعدد اللغات) في قائمة واحدة كما ذكر في الباب السابق إعتماداً على أوزانها الأصلية دون تعديل[3] فمثلاً إذا أجرينا عملية البحث باستخدام الإستعلام الآتي "مفهوم الbinary tree"،وكانت الوثائق المسترجعة من الفهرس ذو اللغة العربية هي :

Da1 :"في علم الحاسوب شجرة البحث الثنائية أوأحياناً يطبق عليها شجرة البحث المرتبة ..."

Da2 :"هي بنية معلوماتية حيث ان كل رأس فيها اثنين من الابناء على الاكثر غالباً مميزين ب اب ايسر واب ايمن ..."

Da3: " هي الشجرة التي يكون لكل عقدة فيها ابن وحيد فقط . يكون ارتفاع هذه الشجرة عادةً..."

بالاوزان الآتيه على التوالي: 0.67558967 ، 0.590278973 ، 0.318662405

والوثائق المسترجعة من الفهرس ذو اللغة الانجليزية هي:

De1: the basic concepts of binary trees, and "then ..."

De2: ” The binary tree is a fundamental data structure used in computer science...“.

De3: In computer science, a binary tree is a "tree data structure in which each node has ..."

بالاوزان الآتيه على التوالي : 0.7761199 ، *0.7066437* ، *0.63523394*

والوثائق المسترجعة من الفهرس متعدد اللغات هي :

Dm1: ”مفهوم الشجرة الثنائية *binary tree* هي عبارة عن هياكل بيانات...“

Dm2: "... *nodes* الشجرة الثنائية تتكون من مجموعة من العقد“

Dm3: "مفهوم الشجرة الثنائية (*binary tree concept*) : هي جزء اساسي من هياكل البيانات..."

بالأوزان الآتيه على التوالي :

0.7046437 ، 0.6044762 ، *0.0364784*

فإنه في هذه الخوارزمية سيتم دمج الوثائق المسترجعة في القوائم الثلاث بمقارنة وزن الوثيقة الأولى من كل قائمة ووضع الوثيقة ذات الوزن الأعلى في بداية القائمة التي سيتم عرضها للمستخدم ،أي أنه سيتم مقارنة اوزان الوثائق

الثلاث، وزن الوثيقة Da1 في القائمه المسترجعة من الفهرس ذو اللغه العربية مع وزن الوثيقة De1 في القائمة المسترجعة من الفهرس ذو اللغة الإنجليزية ومع وزن الوثيقة Dm1 في القائمة المسترجعة من الفهرس ذو اللغتين(عربي-انجليزي)ويتم وضع الوثيقة التي تحتوي على وزن أكبر(وهي *De1*) في أعلى القائمة التي سيتم عرضها للمستخدم والإنتقال إلى الوثيقة التي تليها في القائمة ومقارنتها مع الوثيقتين من الفهرسين الاخرين إلى ان نتوصل إلى نهاية القوائم الثلاث .وستكون القائمة النهائية عند دمج القوائم الثلاث هي

De1,De2,Dm1,Da2,De3 ,Dm2 ,Da2,Da3,Dm3

ولكن من عيوب هذه الخوارزمية انه يمكن أن تكون أوزان الوثائق في القوائم المختلفة غير قابلة للمقارنة مع بعضها البعض فمثلاً يمكن أن تكون أوزان الوثائق في القائمه المسترجعة من الفهرس العربي كبيرة جداً و أوزان الوثائق في القائمة المسترجعة من الفهرس الإنجليزي صغيرة جداً بالتالي ستظهر الوثائق العربية أولاً مع العلم بأنه يمكن أن تكون الوثائق الإنجليزية ذات صلة أكبر بالإستعلام من الوثائق العربية.

2.2.3 خوارزمية تطبيع الوزن(Min Max)

بما أنه يمكن أن تكون أوزان الوثائق في القوائم المختلفة غير قابلة للمقارنة عندها يكون من المعقول أن يتم تطبيع الوزن أو تعديله لكل قائمة من القوائم المسترجعة من الفهارس الثلاثة كلاً على حده قبل أن يتم دمجها في قائمة واحدة ومن إحدي الطرق المستخدمة هي قسمة اوزان الوثائق في كل قائمة على أكبر وزن (max score)وهو وزن أول وثيقة في تلك القائمه المسترجعة المرتبة مما يجعل أوزان تلك الوثائق محصورة بين 0 و 1 الوثيقة ذات أعلى وزن سيطبع وزنها للواحد وآخر وثيقة في تلك القائمة سيطبع وزنها ويكون قريباً من الصفر.ومن ثم يتم إعاده ترتيب القوائم بعد تعديلها ومن ثم دمجها في قائمة واحدة تظهر للمستخدم [4].

وأيضاً في هذه الخوازميه يتم استخدام المعادله التاليه لتطبيع الوزن:

$$\mathbf{normalized\_score} = \frac{\text{original\_score} - \text{min\_score}}{\text{max\_score} - \text{min\_score}}$$

حيث أن:

Original_score:هي الوزن الأصلي للوثيقة

min_score:هي أصغر وزن في القائمه المسترجعة (وزن آخر وثيقة في القائمة)

Max_score:هو أكبر وزن في القائمه المسترجعة (وزن أول وثيقة في القائمة)

فباعتيار القائمة المسترجعة من الفهرس العربي min_score و Max_score هي 0.67558967 ، 0.318662405 وعند تطبيق المعادلة أعلاه سيكون الوزن الجديد المطبع لهذه القائمة لكل من الوثائق الثلاث *Da1,Da2,Da3*هو 0،0.761،1 على التوالي

وعند تطبيق المعادلة على القوائم الثلاث وتطبيع أوزان الوثائق فيها ودمجها في قائمه واحدة كما ذكر في الباب السابق فإن القائمه النهائية التي سيتم إظهارها للمستخدم هي :

De1,Da1,Dm1,Dm2,Da2 ,Da2 ,De3,Da3,Dm3

3.2.3 خوارزمية CORI

كما ذكر سابقا هذه طريقة أخرى لتطبيع الوزن وذلك باستخدام أوزان الوثائق بالإضافة إلى الوزن القائمة المطبع(weighted score)الذي يتم استخراجه من القائمة التي تنتمي لها الوثيقة[5].

وهذه الطريقة تعتبر أكثر دقة من سابقاتها ومن أفضل خوارزميات دمج النتائج وفيها يتم تطبيع الوزن بناء على طول الوثائق المسترجعة من كل مجموعة.بحيث يكون الوزن النهائي المطبع لأي وثيقة عبارة عن وزن الوثيقة مضروباً في وزن المجموعة التي تنتمي لها، ونستخدم ذلك الوزن المطبع لدمج تلك النتائج [7].

والفكرة الأساسية لهذه الخوارزمية هي زيادة وزن الوثائق التى لها وزن أعلى من الوزن المتوسط ، ويقل وزن الوثائق التي لها وزن اقل من الوزن المتوسط . ومن دواعي إستخدامنا لهذه الخوارزمية أنها بسيطة لأن مدخلاتها هي وزن الوثائق وطول النتائج المسترجعة . وأيضاً هذه الخوارزمية لا تحتاج الى معلومات عن مجموعة الوثائق وبالتالي فإن الوسيط (Broker) الذي يتعامل مع هذه الخوارزمية لا يحتاج الى تخزين معلومات عن مجموعة الوثائق . ولكن إذا كانت التنبؤات عن مجموعة الوثائق مطلوبة في بيئة متغيرة او حركية (مثل الويب) فإن هذه المعلومات تحتاج الى التحديث بصورة دورية وهذا غير ممكن إلا بوجود تعاون بين الوسيط وخوادم المجموعات (Collection Server).

وهذه الخوارزمية تقوم بحساب وزن الوثائق بناءً على طول النتائج المسترجعة (عدد الوثائق المسترجعة) لكل مجموعة ، والوزن لكل مجموعة يقوم على إفتراض أن كل الوثائق المسترجعة هي أعلى صلة بالإستعلام وتسخدم هذه الطريقة لتطبيع الوزن.

$$\mathbf{normalized_{score}} = \text{original}_{\text{score}} * [1 + \frac{\text{Si} - \text{avg\_S}}{\text{avg\_S}}]$$

حيث:

avg_S :متوسط وزن الوثائق في المجموعة الواحدة

si : هو وزن مجموعة الوثائق ذات اللغة المراد تطبيع وثائقها.

وبعد أن تقوم خوارزمية ال *CORI* بتعديل أوزان الوثائق عند تطبيق معادلة تطبيع الوزن أعلاه يتم دمج القوائم الثلاث في قائمة واحدة بمقارنة وزن الوثيقة الاولى من كل قائمة ووضع الوثيقة ذات الوزن الأعلى في بدايه القائمة التي سيتم عرضها للمستخدم ويكون ترتيب القائمة الناتجه كالتالي :

Dm1,Dm2,Da1,Da2,De1,De2,De3, Da3,dm3

3.3 التقنيات المستخدمة

1.3.3 لغة الجافا (Java)

هي عبارة عن لغة برمجة ابتكرها جيمس جوسلينج (James Gosling) في عام 1992 أثناء عمله في مختبرات شركة صن (Sun Microsystems) وذلك لاستخدامها بمثابة العقل المفكر المستخدم لتشغيل الأجهزة التطبيقية الذكية مثل التلفاز التفاعلي.

مميزاتها :

- ❖ إحدى لغات الكائنية التوجه (Object Oriented)، آمنة وجيدة الأداء.
- ❖ إضافة الحركة والصوت إلى صفحات الويب، كتابة الألعاب والبرامج المساعدة، وإنشاء برامج ذات واجهة مستخدم رسومية.
- ❖ تصميم برمجيات تستفيد من كل مميزات الأنترنت، توفر لغة الجافا بيئة تفاعلية عبر الشبكة العنكبوتية وبالتالي تستعمل لكتابة برامج تعليمية للإنترنت عبر برمجيات المحاكاة الحاسوبية للتجارب العلمية وبرمجيات الفصول الإفتراضية للتعليم الإلكتروني والتعليم عن بعد. لا تنحصر فاعلية الجافا في الشبكة العنكبوتية فقط بل تمكننا من إنشاء برامج للاستعمال الشخصي والمهني حيث يوجد العديد من البرمجيات التي تسهل عملية كتابة الأوامر كـ netbeans وEclipse [9].

في المشروع محل الدراسة قمنا بإستخدام هذه اللغة لأنها تدعم الكثير من المكتبات (API) التي تساعدنا في تطبيق النظام بصورة أكثر كفاءة.

2.3.3 تقنية JSP (Java Server Page)

هي تقنية صممت لتساعد مطوري البرمجيات في إنشاء صفحات ويب متغيرة المحتوى (ديناميكية). تم إصداراها عام 1999 بواسطة شركة (Sun Microsystems) وهي تشبة لغة PHP إلا أنها تستخدم لغة البرمجة جافا. لتنفيذ ونشر محتوى الصفحات المكتوب بهذه اللغة لابد أن يكون خادم الويب (web server) متوافق مع لغة سيرفلت (servlet) مثل Apache Tomcat أو Jetty.

مميزاتها :

- ❖ لا تعتمد على منصة بعينها (Multi Platform).
- ❖ يوجد العديد من المكونات التي يمكن إعادة استخدامها داخل صفحة JSP ولعل أشهرها (JavaBeans).
- ❖ نسبة لإعتمادها على لغة الجافا فهي تمكن من الإستفادة من قوة ومزايا الجافا [10].

في المشروع محل الدراسة تم إستخدام هذه التقنية لأن النظام وكما ذكرنا سابقاً تم تطويره بلغة الجافا ولأن لغة JSP تُمكن من التعامل مع برامج مكتوبة بلغة الجافا بكل سهولة ، قمنا بإستخدامها لبناء واجهة ويب للنظام.

3.3.3 خادم ويب تومكات (Apache Tomcat Server)

هو خادم ويب وحاوية سيرفلت (Servlet) مفتوح المصدر أصدرته مؤسسة أباشي للبرمجيات (Apache software Foundation)، كما يوفر أيضاً هذا الخادم العديد من المميزات التي تهيئ منصة عمل لتطوير تطبيقات و خدمات الويب.

مميزاته : المرونة (flexibility)، الإستقرار (Stability) [11].

في المشروع محل الدراسة قمنا بإستخدام هذا الخادم نسبة لأن لغة الويب المستخدمة في صفحاتنا هي JSP وهو يتيح لنا إمكانية ترجمة هذه اللغة بصورة مجانية.

4.3.3 محلل جيريكو لوثائق HTML (Jericho HTML Parser)

هو مكتبة جافا تتيح التحليل والتلاعب في أجزاء من وثيقة HTML، و أيضاً الأجزاء التي بجانب الخادم (server-side tags)، كما أنه يمكن عمل إستنساخ حرفي لأي جزء في وثيقة HTML غير معترف به أو غير صالح.

مميزاته :

- التعامل مع الوثائق الكبيرة في شكل أجزاء متسلسلة مما يحافظ على مساحة الذاكرة.
- متطلباته من الموارد ذات نسبة معقولة مقارنة مع المحللات الأخرى.
- يسمح بإستخلاص كل النص الموجود بالوثيقة المحددة.
- يسمح بحذف المسافات غير المرغوب فيها بين الكلمات.
- يسمح بإستخلاص النص الموجود في الوثيقة المحددة بصورة مناسبة لتغذية محركات البحث مثل: Apache Lucene [12] .

في المشروع محل الدراسة تم إستخدام هذا المحلل في إستخلاص النص العربي(النص فقط) من صفحات HTML المخزنة كوثائق لفهرستها و البحث فيها.

5.3.3 محرك البحث لوسين (Lucene)

هومكتبة مكتوبة بلغة جافا بواسطة مؤسسة أباشي للبرمجيات (Apache software foundation) تُستخدم للبحث عن المعلومات بداخل الوثائق صممت لإضافة إمكانات البحث للتطبيقات الخاصة بالمستخدم.

مميزاته :

- ❖ الأداء العالي و إمكانية التوسعة.
- ❖ مجاني مفتوح المصدر.
- ❖ بالإضافة لإمكانيات البحث توفر هذه المكتبة أيضاً بعض الإضافات أو الظائف جديدة مثل: تظليل النص و التدقيق الإملائي.

طريقة عمله:

هنالك بعض العمليات تتم على الوثائق المراد البحث فيها و عمليات أيضاً تتم على طلب المستخدم.

العمليات التي تتم على مجموعة الوثائق هي:

أ- **إستخلاص المحتوى (Acquire content)**

يتم إستخلاص محتوى الوثائق المراد فهرستها للبحث فيها.

ب- **بناء وثيقة جديدة (Build document)**

يتم بناء وثيقة جديدة مبنية في صورة مفتاح و قيمة لتكون مهيكلة بصورة تسمح بالبحث فيها.

ت- **تحليل الوثيقة (Analyze document)**

يتم تطبيع و تشذيب الكلمات الموجودة بالوثيقة لرفع درجة كفاءة البحث.

ث- **فهرسة الوثيقة (Index document)**

تتم إضافة الوثيق المحللة للفهرس.

العمليات التي تتم على طلب المستخدم هي:

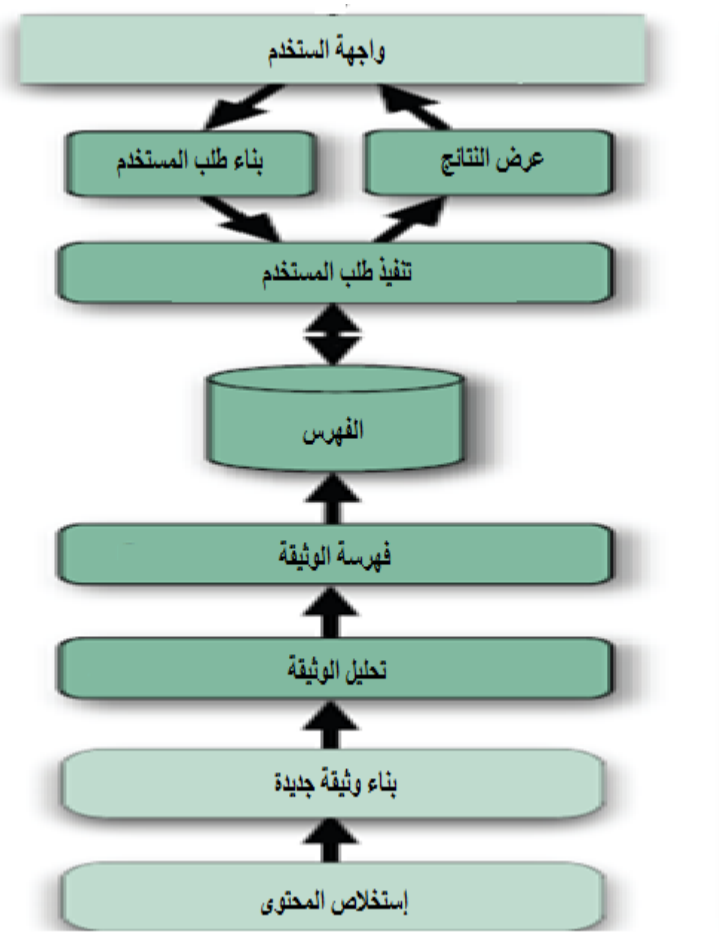
i. **بناء طلب المستخدم (Build query)**

يتم تكوين طلب المستخدم بإستخدام الحزمة QueryParser التي تقوم بتحويل النص الذي أدخله المستخدم إلى طلب ذو صيغة محددة متوافقة مع صيغة البحث.

ii. **تنفيذ طلب المستخدم (Run query)**

يتم البحث في الفهرس عن طلب المستخدم لإسترجاع الوثائق الأكثر ملاءمة.

الشكل (1.3) يوضح مجمل الخطوات والعمليات التي يقوم بها لوسين[13].



شكل (2.3) العمليات التي تتم في محرك بحث لوسين

في النظام محل الدراسة تم إستخدام محرك بحث لوسين في إضافة إمكانية البحث في الوثائق التي تمت فهرستها بالطريقة قيد الدراسة.

1.4 خطوات تطبيق المـشروع

تم تكوين بيئة الإختبار لهذا المشروع بإستخدام محرك البحث لوسين (Lucene) النسخه (4.6.0) والذي يستخدم للبحث عن المعلومات بداخل الوثائق، ولإستخلاص النصوص من الصفحات تم إستخدام محلل الجريكو (Jericho) وحيث تم في هذه التجربة إستخدام عدد 20الف وثيقة مختلفة المواضيع مثل وثائق تتحدث عن الذكاء الإصطناعي وأخرى عن حالة الجمود(deadlock) وغيرها من المواضيع. ثم عملية تطبيع النصوص العربية والإنجليزية ومن ثم تشذيبها وتكوين الوثائق الجديدة وفهرستها ببناء فهرس بالطريقة المقترحة وهي التي تدمج بين الطريقة المركزية والموزعة كما تم توضيحها سابقاً،وترجمة الأجزاء العربية والإنجليزية والبحث بها في ذلك الفهرس،وأخيراً يتم دمج النتائج المسترجعة من عملية البحث بخوارزميات الدمج التي تم توضيحها سابقاً.وفيما يلي شرح لهذه الخطوات:

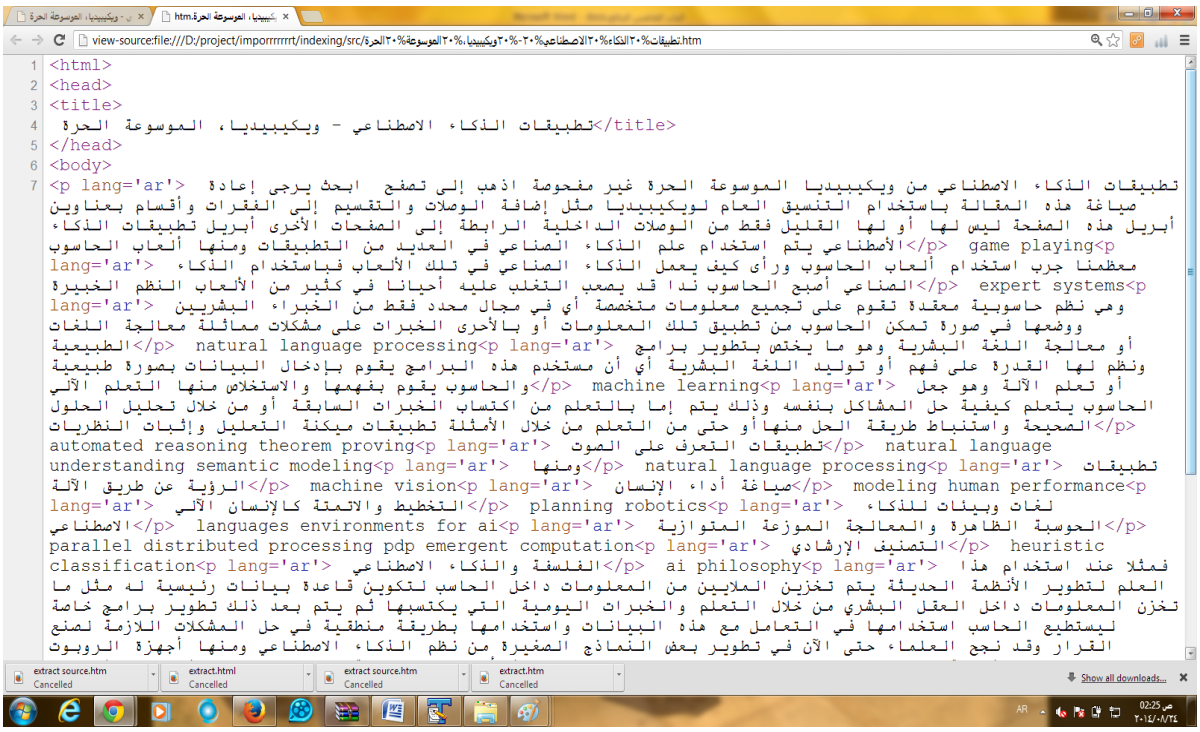
أولاً: تحديد مجلد أو أكثر لفهرسة الملفات الموجودة بداخله

تتم قراءة الملفات الموجودة بداخل المجلد ملف،ملف وإستخلاص النص فقط من الملفات ذات الإمتداد(html)أو(htm).بإستخدام محلل جريكو بحيث يتم تجاهل الصور و النصوص غير العربية الموجودة بداخل ملفات (HTML). يتم وضع النص المستخلص في شكل سلسلة من الحروف (string) وإعتبارها مدخل لعملية التطبيع.

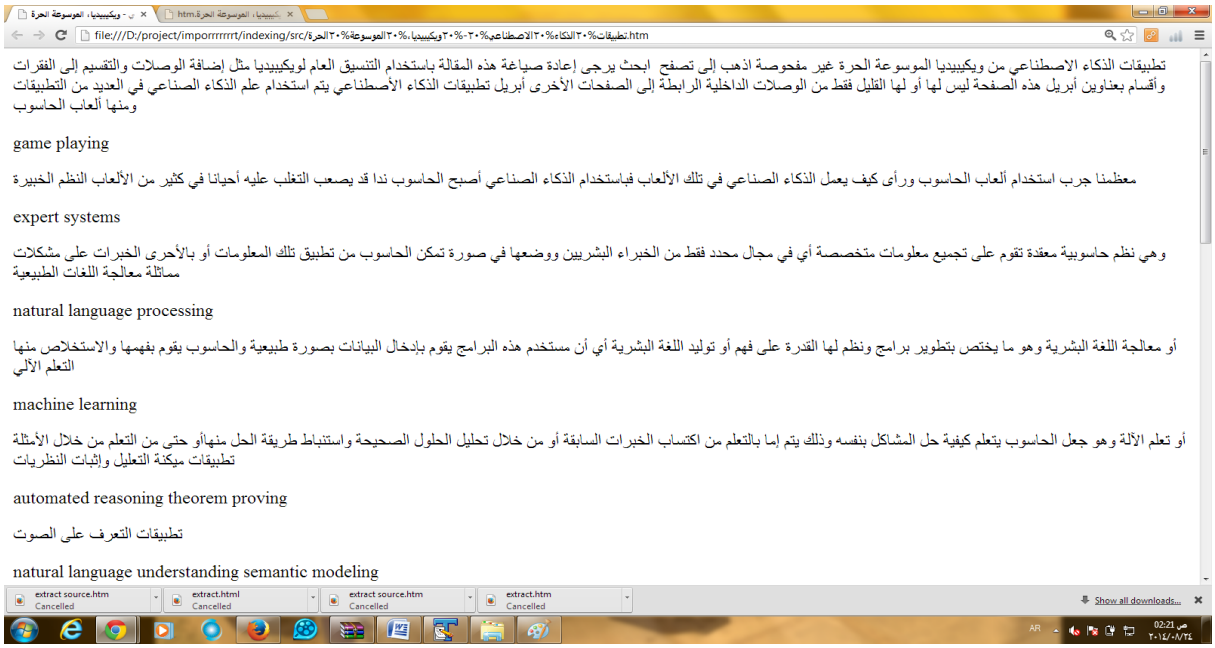
وفيما يلي توضيح لصفحة قبل وبعد إستخدام محلل الجريكو(Jericho)

الشكل(1.4): الصفحة قبل إستخدام محلل الجريكو

ولتنقية مثل هذه الصفحة قمنا بإستخلاص النص فقط (العربي والإنجليزي) وإزالة كل الصور وعلامات الترقيم والأحرف بأي لغة خلاف اللغتين الإنجليزية والعربية وبما أن الصفحة كما موضح في النموذج أعلاه خليط بين اللغتين العربية والانجليزية فعندما تكون الفقره باللغة العربية نقوم بتغير ال<tag> لهذه الفقرة الى العربية ليصبح <tag lang =”ar”> وفيما يلي توضيح لذلك :



الشكل(2.4) : عملية تعريف الtag



الشكل (3.4) : الصفحة بعد إستخدام الجريكو

ثانياً: تطبيع محتوى ملفات (HTML)

يتم تطبيع النص المستخلص و ذلك بتوحيد الحروف – مثلا: قلب كل (أ)، (إ) و (آ) الى (ا)- وإزالة التشكيل ما عدا الشدة بإعتبارها حرف من حروف الكلمة الأصلية ليكون النص بعد ذلك جاهز للتشذيب.

ثالثاً: تشذيب النص العربي والإنجليزي (Stemming)

بعد تطبيع النص يتم تشذيب النص لرفع كفاءة البحث والذي يقوم بإزالة بوادئ ولواحق الكلمات وإرجاعها الى أصلها وله أنواع فلتشذيب الكلمات العربية تم إستخدام (light10 stemmer)وهذا المشذب مستخدم على نطاق واسع في إسترجاع المعلومات العربية.

نوع المشذب	السوابق	اللواحق
Light 10	ال، وال، بال، كال، فال، لل، و	ها، ان، ات، ون، ين، يه، ية، هـ، ة، ي

جدول(1.4): السوابق والواحق في Light10

و بذلك نكون بحاجة للتعرف على الأسماء و الأفعال في سلسلة الحروف المراد تشذيبها.

رابعاً: يتم تكوين الوثيقة الجديدة المراد فهرستها وإضافتها الى الفهرس

بعد تطبيع و تشذيب النص المكتوب في الوثيقة لابد من تكوين وثيقة جديدة يتم حفظ المحتوى فيها في شكل مفتاح و قيمة لتصبح الوثيقة قابلة للبحث كما هو موضح في الجدول:

الحـــقل	القـيمة
المسار	مسار الوثيقة الاصلي
رقم الوثيقة	رقم الوثيقة (لكل وثيقة رقم)
التعديل	تاريخ آخر تعديل تم على الوثيقة
تعريف المستخدم	مسار الوثيقة مضاف الى التاريخ الذي تمت فيه إضافة الوثيقة
عنوان الوثيقة	العنوان الموجود في الإشارة<title>
النص	النص الموجود في الإشارة <body>

جدول (2.4) : مكونات وثيقة لوسين

ومن ثم تتم إضافة الوثيقة الجديدة في الفهرس الذي تم تحديد موقعه من قِبل المستخدم.
بذلك تكون كل الملفات التي تحتوي على نصوص عربية وإنجليزية في المجلد الذي حدده المستخدم فُهرست في فهارس تم حفظها في المكان الذي حدده المستخدم أيضاً.

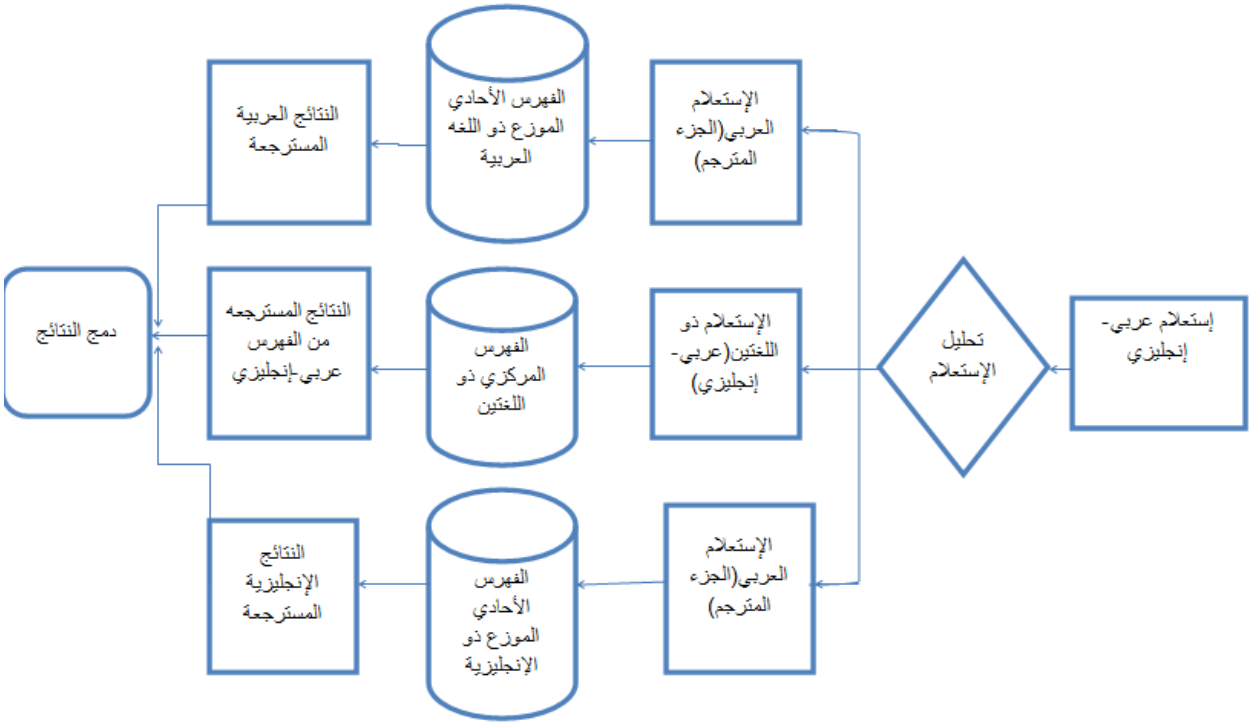
خامساً : ترجمة الأجزاء العربية والانجليزية الواردة في طلب المستخدم

يتم إدخال الإستعلام عربي-انجليزي وترجمة الجزء العربي من الإستعلام عربي-انجليزي الى اللغة الإنجليزية ودمج الجزء المترجم الى الجزء الإنجليزي من الاستعلام الأصلي سينتج إستعلام عربي والبحث بهذا الإستعلام في الفهرس العربي أحادي اللغة وايضاً ترجمة الجزء الإنجليزي من الإستعلام عربي-انجليزي الى اللغة العربية ودمج الجزء المترجم إلى الجزء العربي من الإستعلام الأصلي لينتج الإستعلام الإنجليزي والبحث بهذا الإستعلام في الفهرس الإنجليزي أحادي اللغة وأخيراً دمج الإستعلام العربي أحادي اللغة مع الإستعلام الإنجليزي

أحادي اللغة لتكوين إستعلام عربي-إنجليزي جديد للبحث به في الفهرس ذو اللغتين العربية والإنجليزية،ثم إسترجاع النتائج في ثلاث قوائم ، القائمة الأولى تحوي وثائق عربية والثانية وثائق إنجليزية والثالثة وثائق عربي-إنجليزي .

سادساً : تطبيق خوارزميات الدمج على النتائج المسترجعة

بعد عملية البحث يتم إسترجاع ثلاث قوائم ،القائمة الأولى يتم إسترجاع نتائجها من الفهرس الأحادي ذو اللغة العربية والقائمة الثانية من الفهرس ذو اللغة الإنجليزية وقائمة ثالثة مسترجعة من الفهرس عربي-إنجليزي .ومن ثم يتم دمج تلك النتائج المسترجعة بإستخدام خوارزميات الدمج التي تم توضيحها سابقا.



الشكل(4.4):عملية البحث و دمج النتائج في الفهارس المركزية الموزعة

1.1.5 التصميم التجريبي

في هذا الإختبار سيتم قياس كفاءة طريقة الفهرسة التي تم بناءها وتطبيق خوارزميات لدمج النتائج ومقارنة نتائج تلك الخوارزميات لمعرفة أيهم أحسن كفاءه في عملية الإسترجاع،سيتم قياس الكفاءة عن طريق معيار DCG (Discounted Cumulative Gain) الذي يتم حسابه بناءاً على مدى إرتباط الوثيقة المسترجعة بطلب المستخدم، بالإضافة إلى ترتيب الوثيقة ضمن مجموعة الوثائق المسترجعة .

سيتم تطبيق التجربة على مجموعة وثائق HTML تحتوى على نصوص عربية وأخرى تحتوي على نصوص إنجليزية ووثائق تحتوي على نصوص متعددة اللغات(عربية-إنجليزية) مفهرسة بالطريقة المركزية الموزعة ، ليتم البحث فيها وإسترجاع قوائم نتائج من تلك الفهارس و دمج تلك النتائج في قائمة واحدة مرتبة للمستخدم بإستخدام الخوارزميات الثلاث التي تم توضيحها سابقاً و مقارنتها مع بعضها البعض لتحديد أيهما أحسن كفاءة.

2.1.5 الهدف من التجربة

معرفة مدى تأثير الطريقة الجديدة المستخدمة في الفهرسة على كفاءة إسترجاع المعلومات، و مقارنة كفاءة الخوار زميات الثلاث التي تم تطبيقها لدمج النتائج المسترجعة.

3.1.5 منهجية التجربة

تم إستخدام محرك بحث لوسين في فهرسة الوثائق بطريقة تسمح بالبحث فيها، فأولاً تمت تنقية هذه الوثائق من الصور، و الرموز، و كل العناصر ما عدا النصوص العربية والإنجليزية و بعد ذلك تم تطبيع تلك النصوص التي تكوِّن محتويات كل وثيقة لتصبح جاهزةً لعملية التشذيب. ثالثاً تم تشذيب محتويات هذه الوثائق بإستخدام لايت10،ومن ثم عملية الفهرسة.

تم بناء فهرس مركزي موزع للبحث فيه،ومن ثم تم جمع ثمانية إستعلامات (quires) من مستخدمين مختلفين، وفي هذه الإستعلامات تمت ترجمة الأجزاء العربية والإنجليزية وتم تكوين الإستعلام العربي والإستعلام الانجليزي والإستعلام متعدد اللغات (عربي-انجليزي)،وإستخدام تلك الإستعلامات للبحث في ذلك الفهرس الذي تم بناءه.

من ضمن الوثائق المسترجعة تم إختيار أول عشر وثائق، و تم تقييم كل وثيقة برقم في المدى (0ـــ5) على حسب مدى إرتباطها بطلب المستخدم ، ليتم بعد ذلك حساب معيار DCG (Discounted Cumulative Gain)

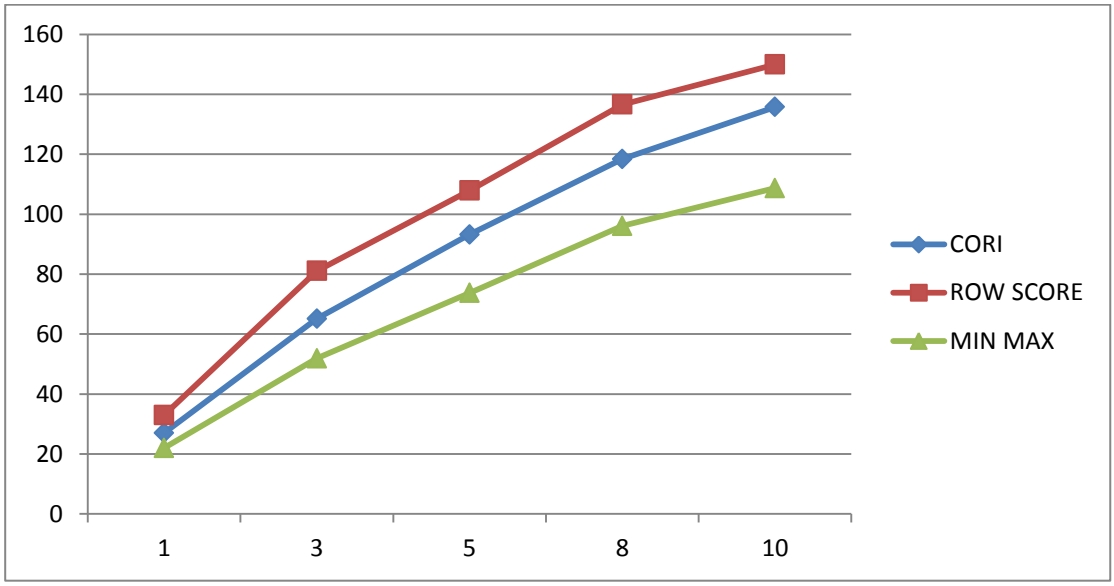
وهو معيار لكفاءة إسترجاع المعلومات في محركات البحث،حيث يتم حسابه لكل طلب في كل خوارزمية من خوارزميات الدمج الثلاثة، لتتم من خلاله مقارنة الكفاءة.

4.1.5 النتائج والمناقشة

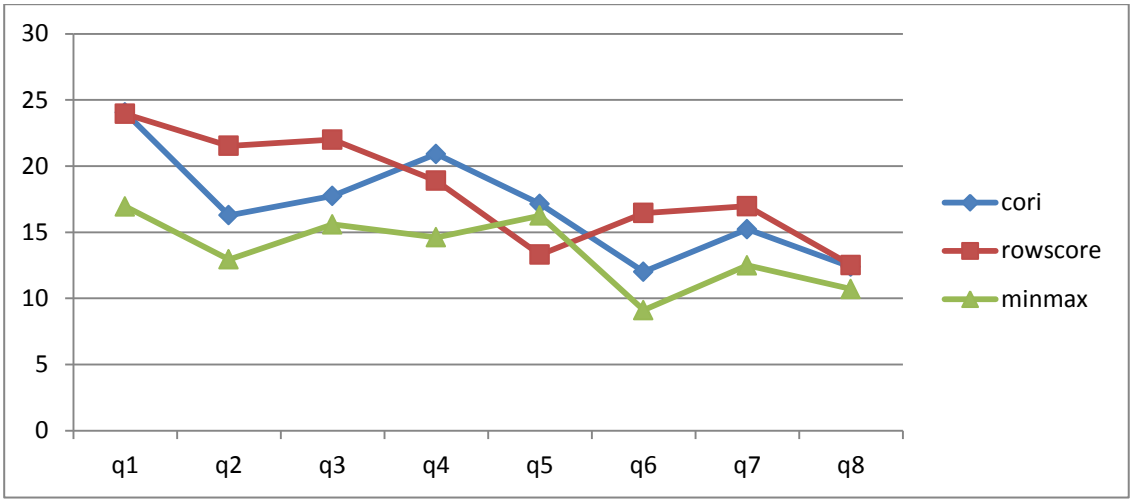
وبعد إجراء عملية البحث بإستخدام الثمانية إستعلامات في الفهرس المركزي الموزع وإسترجاع قائمة نتائج مرتبه من كل فهرس ومن ثم تطبيق خوارزميات المج الثلاثة عليها فكان متوسط معيار DCG لكل من الخوارزميات الثلاثة كالأتي:

- متوسط DCG بإستخدام خوارزمية CORI= 16.97
- متوسط DCG بإستخدام خوارزمية RawScore = 18.74
- متوسط DCG بإستخدام خوارزمية NormalizedScore = 13.59

بالنظر للشكل (1.5) يمكننا مقارنة متوسطات DCG بصورة أوضح، و يتجلى لنا أن خوارزمية Raw Score تفوقت على الأخريات بحصولها على أكبر متوسط من DCG على مستوى عدد الوثائق في قائمة النتائج المسترجعة وذلك على مستوى وثيقه واحده وعلى مستوى ثلاث وثائق وعلى مستوى خمس وثائق وعلى مستوى ثماني وثائق وعلى مستوى عشر وثائق.



الشكل(1.5):متوسطات معيار DCG للخوارزميات الثلاث



الشكل(2.5):نتائج الخوارزميات للثلاث استعلامات

كما ذكر سابقا أن خوارزمية CORI هي الأفضل وكان ذلك بناء على إستخدام إستعلام أحادي اللغة ولكن بالنظر للشكلين (2.5)، (1.5) أعلاه وجد أن خوارزمية Raw Score تعطي نتائج أفضل وذلك لأن عملية البحث هنا تتم بإستخدام إستعلام ذو لغتين(عربي_إنجليزي) مما أدى إلى زيادة طول الإستعلام الذي يتم البحث به في الفهرس متعدد اللغات والذي بدوره أثر على وزن الوثائق وبالتالي أثر علي ترتيب الوثائق حسب الأهمية.

5.1.5 جوانب القصور في نتيجة التجربة

- تم إجراء التجربة على عدد غيركاف من الوثائق العربيه.
- من المفترض أن يتم نوع من إختبارات الأهمية على نتائج طلبات المستخدمين، لكن إختبارات الأهمية يجب أن تتم على الأقل على 31 طلب مستخدم، و في تجربتنا هذه إستخدمنا 8 طلبات فقط، لذلك لم يكن من الممكن إجراء إختبار الأهمية.

6.1.5 الصعوبات التي واجهت البحث

عدم القدره علي الحصول على Google translator API حيث تم اغلاقها من قبل الشركه مؤخرا لذا تم بناء قاموس يحتوي علي اكثر الكلمات استخداما في عملية البحث لاستخدامه في عملية ترجمة الاستعلام.

2.5 التوصيات

- إيجاد طريقة لتعديل معادلة تطبيع الوزن في خوارزمية cori وذلك بإدخال معامل طول طلب المستخدم لإعطاء نتائج أفضل
- تطبيق الفهرسة و البحث على مجموعة أكبر من الوثائق للحصول على نتائج أفضل عند مقارنة خوارزميات دمج النتائج.
- إجراء عملية البحث عن طريق طلبات مستخدمين تتجاوز الثلاثينيات، للتمكن من إجراء إختبار الأهمية.

الخاتمة

بعد توفيق من الله سبحانه وتعالى تم استكمال هذا البحث الذي يستخدم الطريقة المركزية الموزعة لفهرسة الوثائق و التي تمكننا من فهرسة الوثائق متعددة اللغات أو أحادية اللغة بطريقة فعالة تزيد من كفاءة الإسترجاع في انظمة إسترجاع المعلومات،وايضا تمكين المستخدم من كتابة الاستعلام باللغه التي يراها مناسبه أو التعبير عن مصطلحاتة باللغة التي يعرفها.ومن ثم تم تطبيق خوارزميات دمج النتائج على النتائج المسترجعه من عملية البحث بهذا الاستعلام وتكوبن قائمة واحدة تكون فيها النتائج مرتبه علي حسب صلتها بالإستعلام وأهميتها بالنسبه للمستخدم بحيث تكون الوثائق ذات الصله الاعلى بالاستعلام و ذات الاهمية الاكبر بالنسبه للمستخدم في بداية القائمة الناتجة.