العنوان:	نظم استرجاع المعلومات العربية و اتجاهات البحوث المعاصرة
المصدر:	اعلم -السعودية
المؤلف الرئيسي:	دبور، عبدالرحمن بن غالب
المجلد/العدد:	ع14
محكمة:	نعم
التاريخ الميلادي:	2015
الشهر:	فبراير / ربيع الثاني
الصفحات:	46 - 89
رقم MD:	650437
نوع المحتوى:	بحوث ومقالات
قواعد المعلومات:	ACI, HumanIndex
مواضيع:	اللغة العربية ، تكنولوجيا المعلومات ، استرجاع المعلومات ، البحوث العلمية
رابط:	http://search.mandumah.com/Record/650437

نظم استرجاع المعلومات العربية واتجاهات البحوث المعاصرة

د. عبد الرحمن غالب دبور

أستاذ قسم علم المعلومات المساعد

رئيس قسم المعلومات بجامعة طيبة

ملخص:

تتناول هذه الدراسة خصائص اللغة العربية وكفاءتها في استرجاع المعلومات. ونماذج استرجاعها، وتركز الدراسة كذلك على التطورات الحديثة في تقويم نظم استرجاع المعلومات العربية. وقياسات بحوث استرجاع المعلومات، بدايات بحوث استرجاع المعلومات العربية، وتطور بحوث استرجاع المعلومات العربية في القرن الحادي والعشرين. وتقارن الدراسة بين البحوث العربية والأجنبية في هذا المجال، واتجاهات المستفيدين العرب نحو الويب ونتائج البحوث المستقبلية.

١-١- تمهيد:

أولاً: منهجية الدراسة:

مقدمة:

تبدأ الدراسة بآيات القرآن الكريم التي ذكرت اللغة العربية وحكمة الله في جعلها وعاء للرسالة الخاتمة الخالدة، ومن بين هذه الآيات ما يلي:

{ وإِنَّهُ لَتنزِيلُ رَبِّ العَالَمِينَ (١٩٢) نَزَلَ بِهِ الرُّوحُ الأَمِينُ (١٩٣) عَلَى قَلْبِكَ لِتَكُونَ مِنَ المُنذِرِينَ (١٩٤) بِلِسَانٍ عَرَبِيٍّ مُّبِينٍ } [الشعراء: ١٩٢ – ١٩٥].

{ إِنَّا أَنزَلْنَاهُ قُرْآنًا عَرَبِيًا } [سورة يوسف: الآية ٢].

{ كِتَابٌ فُصِّلَتْ آيَاتُهُ قُرْآنًا عَرَبِيًا لِّقَوْمٍ يَعْلَمُونَ } [سورة فصلت: الآية ٣].

{ وَكَذَلِكَ أَنزَلْنَاهُ قُرْآنًا عَرَبِيًا وصَرَّفْنَا فِيهِ مِنَ الوَعِيدِ لَعَلَّهُمْ يَتَّقُونَ أَوْ يُحْدِثُ لَهُمْ ذِكْرًا } [سورة طه: الآية ١١٣].

هذه الآيات السابقة تتكامل مع آيات كثيرة بالقرآن الكريم، بدأت بسورة العلق وأول آية فيها اقرأ والمقصود بذلك أن يقرأ الإنسان كل ما فيه صلاح له في الدنيا والآخرة، كما أن الآيات تخاطب العقل (تعقلون) وتخاطب العلم (يعلمون)، وتنزلت على قلب رسول الله محمد صلى الله عليه وسلم لتكون من المنذرين بلسان عربي مبين.

وتعد اللغة العربية من اللغات الخمس لمنظمة الأمم المتحدة، وهي اللغة الأم التي يتحدث بها حوالي ثلاثمائة مليون نسمة وهي لغة القرآن الكريم وبالتالي فهي اللغة الثانية للمسلمين حول العالم والذين يبلغ

عددهم أكثر من بليون ونصف نسمة.

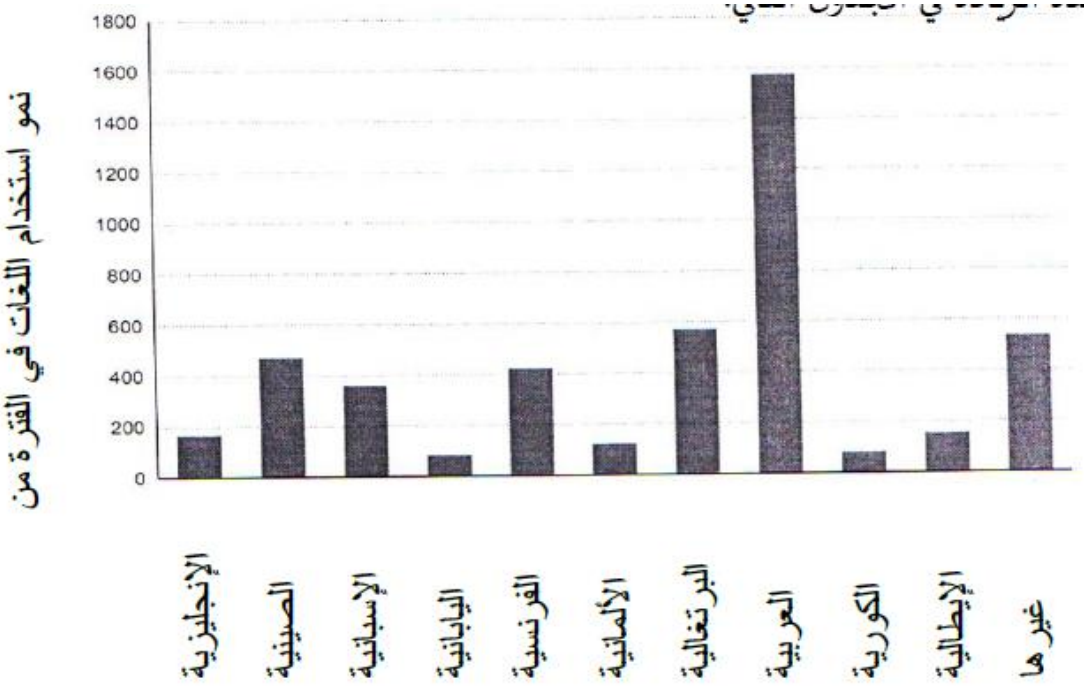
كما يعتبر استرجاع المعلومات التكنولوجيا الأساسي وراء محركات البحث للويب وأصبح الويب مصدرًا رئيسيًّا للمعلومات، حيث يتعامل مع بلايين الوثائق المتاحة للبحث ويتزايد حجمها كل يوم. وطبقًا لإحصائيات أبريل ٢٠٠٨ فهناك (١٦٥,٠٠٠,٠٠٠) مجال واضح للأسماء على الإنترنت.

http://news.netcraft.com/archives/2008/04/14/April2008_web_server_survey.html.

أما بالنسبة للمستخدمين للإنترنت في الشرق الأوسط فقد زاد عددهم بنسبة ٩٢٠% في الفقرة ٢٠٠٠ وحتى ٢٠٠٧ أما عدد الذين يستخدمون الإنترنت باللغة العربية فقد وصل إلى ٤٦,٣٥٩,١٤٠ شخص في نوفمبر ٢٠٠٧، أي أن هذه الزيادة قد بلغت ١٥٧٦% زيادة على عدهم عام ٢٠٠٠.

http://www.internetworldstats.com.stats7.htm.

وتظهر هذه الزيادة في الجدول التالي:



المصدر:

Language growth in the internet between 2000 and 2007, Source: Internet World Stats {Miniwatts International, 2007}. From: Nwesri, A. F. Ph. D Melbourne Australia 2008.

هذا والمعروف أن اللغة العربية، لغة عالية الصرفية (Highly Inflected Language) فكلماتها مشتقة من كلمات الجذور (Roots) وتمتد بواسطة حروف إضافية في أول الكلمة ووسطها وآخرها

(Prefix Infixes and Suffixes) وهذا يؤدي إلى عدد (٢٥٥٢) شكل مختلف للكلمة (الفعل) ويؤدي إلى (٥١٩) شكل للاسم (Attia, 2006) ويلاحظ أن هذه الحروف الإضافية في أول ووسط وآخر الكلمة تلتصق بالكلمة. وليست في شكل منفصل كما هو الحال باللغة الإنجليزية، وبالتالي البتر (Truncation) البسيط يمكن تطبيقه لاستبعاد اللواحق Suffixes من الكلمة الإنجليزية ولا يمكن استبعاده في اللغة العربية وبالتالي فلا بد من إتباع أساليب أخرى في هذا الاستبعاد.

وعلى الرغم من أن اللغة العربية لغة عالية الصرفية .Inflected L وعلى الرغم من الزيادة السريعة الواضحة في أعداد المستخدمين، فإن معظم محركات البحث تحصر بحوثها في أشكال الكلمات غير الصرفية non – inflected ومعظم المستفيدين يعتمدون على محركات مثل جوجل وياهو على الرغم من وجود محركات بحث عربية مثل:

(http://www.araby.com) Araby

(http://www.ayna.com) Ayna.

هذا وبحوث استرجاع المعلومات العربية (AIR) بدأت منذ ١٩٨٩ وحتى نهاية القرن العشرين واستخدام فيها تجارب صغيرة ومجموعات صغيرة للاختبار، ومعظم هذه البحوث قد ركزت على استخلاص الجذور ومقارنة فعاليتها في تكشيف النص العربي باستخدام الجذور والجذوع Stems أو الكلمات، ولكن القرن الحادي والعشرين قد بدا بدراسات Text retrieval Conference TREC حيث ركزت على فاعلية الأساليب في استرجاع اللغة الواحدة Monolingual وعبر اللغات المختلفة Cross Language، وعلى الرغم من الاتجاه القوي باستخدام (TREC) إلا أن بحوث استرجاع المعلومات العربية كانت تختبر عينات صغيرة نسبيًا بالمقارنة بالبحوث الإنجليزية وغيرها.

وإذا كانت هذه الدراسة قد تناولت في الجزء الأول على بعض الأساس النظري لاسترجاع المعلومات بصفة عامة فقد تناول الجزء الثاني التطبيقات التي اشتملت على (١٢) دراسة في البدايات حتى نهاية القرن العشرين وعلى التطبيقات العربية المتطورة (١٢) دراسة أيضًا ومعظمها نشر على هيئة أطروحات لأصحابها أو مقالات نشرت في الدوريات المعلوماتية العالمية مثل (JASIS, ARIST) وغيرها.

١-٢ بعض المفاهيم المرتبطة بموضوع الدراسة:

١-٢-١ تعريفات استرجاع المعلومات (من أحدثها إلى ما قبل ذلك):

تعريف مانينج ٢٠٠٨:

يمكن تعريف استرجاع المعلومات كحقل أكاديمي للدراسة بأنه إيجاد المواد (الوثائق عادة) ذات الطبيعة غير التركيبية unstructured (نص عادة) والتي ترضى أو تستجيب للحاجة المعلوماتية من

داخل مجموعات ضخمة (تكون عادة مخزنة على الحاسبات) (Manning C. D et al, 2008: 1).

تعريف الموسوعة الحرة ويكيبيديا (٢٠٠٨):

استرجاع المعلومات هو علم البحث عن الوثائق وعن المعلومات داخل الوثائق وعن الميتادايا الخاصة بالوثائق. فضلًا عن بحث قواعد البيانات العلاقية (Relational Data Base) والشبكة العنكبوتية العالمية (الويب) وهناك تداخل (Overlap) في استخدام مصطلحات استرجاع البيانات واسترجاع الوثائق واسترجاع المعلومات واسترجاع النص (Text) ولكل واحدة من هذه المصطلحات إنتاجه الفكري الخاص. وله نظريته وله عملياته (Processes) وتكنولوجياته، كما أن علم استرجاع المعلومات هو علم متعدد الارتباطات (Interdisciplinary) يعتمد على علم الحاسب والرياضيات وعلم المكتبات وعلم المعلومات والعمارة المعلوماتية (Architecture Information) وعلم النفس المعرفي، واللغويات (Linguistics) والإحصاء والفيزياء.

أما نظم استرجاع المعلومات الآلية، فهي تستخدم لمعالجة ما يسمى بالمعلومات الزائدة (Information Overload) والعديد من مراكز المعلومات والجامعات والمكتبات. تستخدم استرجاع المعلومات (IR) لتقديم إتاحة للكتب والدوريات وغيرها من الوثائق، وعلى ذلك فإن محركات البحث بالويب تعتبر أكثر الاستخدامات المرئية (Wikipedia: The Free Encyclopedia) http://en.wikipedia.org/wikiInformationRetrieval.

تعريف بوز – بيتس ١٩٩٩:

استرجاع المعلومات هو جزء من علم الحاسب يدرس استرجاع المعلومات (وليس البيانات) من مجموعة من الوثائق المكتوبة، وتهدف الوثائق المسترجعة إلى إرضاء الحاجة المعلوماتية للمستفيد ويكون ذلك عادة باللغة الطبيعية (Baeza-Yates,R.etal,1999:444).

تعريف روبرت كورفهاج (Korfhage, R. 1997):

في كتابه عن اختزان المعلومات واسترجاعها لم يحدد فيه تعريفًا معينًا ولكنه في المقدمة أشار إلى أن مجال استرجاع المعلومات يعتبر واحدًا من تخصصات علم المعلومات. كما أن كتابه قد أعد من منظور علم المعلومات أكثر منه من منظور علم الحاسب الآلي، وأن محور الكتاب هو دراسة مشكلة تقديم استجابات لتلبية احتياجات معلوماتية تعتمد على الحاسبات والاتصالات عن بعد لمصادر المعلومات، كما يعتبر محاولة لتقديم جسر بين المداخل الاعتيادية إلى المداخل الجديدة المصممة حول مفاهيم وتكنولوجيا معاصرة خاصة واجهات التفاعل (Interfaces) واسترجاع النص الكامل واسترجاع الصورة والصوت، وهناك تركيز خاص في الكتاب على تفاعل المستفيد مع النظام وواجهات استرجاع المعلومات المرئية وأشار إلى ضرورة إحاطة المستفيد بأساسيات الرياضيات عن المجموعات (Sets) والمنطق

والاتجاهات (Vectors) والاحتمالات والإحصاء لا سيما وأن النماذج الرياضية تلعب دورًا رئيسيًّا في تطوير نظرية اختزان استرجاع المعلومات.

تعريف لانكستر ووارنر:

عن أساسيات استرجاع المعلومات اليوم الطبعة الإنجليزية ١٩٩٣ جاء فيه: مصطلح استرجاع المعلومات، كما يستعمل في الغالب الأعم من الحالات يعتبر مرادفًا للبحث في الإنتاج الفكري – فهو عملية البحث في مجموعة من الوثائق (وفقًا لأوسع معنى لمصطلح الوثيقة) للتحقق من تلك الوثائق التي تتناول موضوعًا بعينه، ومن ثم فإن أي نظام مصمم لتيسير عملية البحث في الإنتاج الفكري يمكن أن يسمي، وبلا تجاوز، نظاماً لاسترجاع المعلومات، والفهرس الموضوعي بالمكتبة أحد أنواع النظم وكذلك الكشاف الموضوعي المطبوع أو الإلكتروني، ثم أكمل المؤلفان (لانكستر ووارنر) هذا التعريف بقولهما: ومن الواضح أن استرجاع المعلومات (ليس بالمصطلح المرضي للدلالة بوجه خاص على نوعية النشاط الذي يستعمل معه عادة.

ونظام استرجاع المعلومات لا يسترجع معلومات، والمعلومات في الواقع ليست بالشيء الذي يمكن لمسه أو رؤيته أو سماعه أو الإحساس به، فالمعلومات هي الشيء الذي يغير الحالة المعرفية للشخص في موضوع ما، وهو أفضل تعريف نقدمه (لانكستر ووارنر، ترجمة حشمت قاسم، ١٩٩٧، ص٢٧).

وتصدق هذه التعريفات على نظم استرجاع المعلومات العربية وإن اختلفت اللغة العربية بخصائصها وتميزها الصرفي (النحوي) أو الدلالي أو الصوتي (Phonetic) وقد تبنى غنيم تعريف شارل ميدوز أن استرجاع المعلومات هو استدعاء الرموز أو الأوعية الكاملة للمعلومات من أماكن اختزانها. استجابة للاستفسارات التي يتقدم بها المستفيدون المحتملون من هذه المعلومات (محمد سالم غنيم ٢٠٠٨: ص٩٠).

١-٢-٢ بعض المصطلحات المرتبطة بنظم استرجاع المعلومات العربية:

- علم الصرف Morphology = فرع من الدراسات اللغوية يتعلق بالتركيب الداخلي للكلمة.
- أجزاء مضافة Affixes = أجزاء مضافة إلى الجذع أو الجذر (في أول أو داخل أو نهاية الكلمة).
- الاشتقاق Derivation = اشتقاق وهو قرع من علم الصرف (مورفولوجي) يهتم باشتقاق كلمة في القاموس أو المعجم Lexicon من كلمة أخرى مثل Keeper من Keep أو Hopeless من hope.
- إضافة قبل الجذر أو الساق Prefixes = الوحدات أو الحروف الصغيرة التي تضاف قبل الجذر

أو قبل الجذع (الساق).

- جذر (Root) = الجذر وهو الذي يمكن تحليله إلى أجزاء أصغر.
- جذع (Stem) = جزء من الجذر يتحد مع أجزاء صغيرة Morphemes. في أول أو آخر أو داخل الجذع هو الجزء من الكلمة المتروك بعد استبعاد الوحدات المضافة Affixes.
- Suffixes = وحدات صغيرة تحدث بعد جذر أو جذع الكلمة.
- التصريف Inflection = هو فرع من المورفولوجي يتصل بفئات الصرف مثل الرقم أو الصوت أو الفعل.
- التجذير Stemming: التجذير هو أسلوب اختصار الكلمات إلى جذورها النحوية Grammatical Roots.

١-٣ مشكلة الدراسة وتساؤلاتها:

تتركز مشكلة الدراسة في محاولة التعرف على تطور البحوث والدراسات في مجال استرجع المعلومات بصفة عامة واسترجاع المعلومات العربية بصفة خاصة فضلًا عن التعرف على خصائص اللغة العربية ومدى دخولها في مجال استرجاع المعلومات العربية ويمكن بلورة هذه المشكلة في التساؤلات التالية:

- ما خصائص اللغة العربية والتي تميزها عن غيرها من اللغات ومدى تطويع الآلات ومحركات البحث للملائمة معها؟
- ما أهم التطورات التي دخلت على بحوث استرجاع المعلومات بصفة عامة؟
- ما البدايات التي تزامنت مع استخدام اللغة العربية في نظم استرجاع المعلومات المعتمدة على الحسابات؟
- ما أهم التطورات العلمية والتقنية التي حدثت في استرجاع المعلومات العربية خلال القرن الحادي والعشرين ودخول تريك (TREC) في البحوث العربية ذات الصلة؟
- ماذا عن إسهام الجامعات الأجنبية والعربية بهذا المجال؟

١-٤- الدراسات السابقة:

جاءت هذه الدراسات ضمن الجزء التطبيقي لاسترجاع المعلومات العربية حيث شملت (١٢) دراسة في القرن العشرين وشملت (١٢) دراسة في القرن الحادي والعشرين مع المقارنة بين بعض الدراسات الأولى وتعديلاتها في الدراسات الأحدث.

١-٥ منهج البحث وأدواته:

تستخدم هذه الدراسة المنهج الوثائقي والمسح للإنتاج الفكري مع استقراء الإنتاج الفكري المتصل

بنظم استرجاع المعلومات، والإفادة من محركات البحث خاصة جوجل والإصدارات الخاصة بالإنتاج الفكري العربي لمحمد فتحي عبد الهادي.

ثانياً: خصائص اللغة العربية وكفاءتها في استرجاع المعلومات:

٢-١ الخصائص (*):

هناك ثلاثة أشكال للغة العربية وهي: الفصحى Classical/ المعيارية الحديثة/ العامية Colloquial ويعتبر الشكلان الأول والثاني كلغة فصحى أو التي تعني أكثر فصاحة أو بلاغة وهي لغة القرآن. واللغة العربية المعيارية الحديثة مشتقة من اللغة الفصحى ومعظم الكتب المطبوعة باللغة العربية والصحف مكتوبة بهذا الشكل. أما اللغة العامية (المنطوقة) فهي تختلف من قطر عربي إلى آخر (Al Dayel, A. 2013).

كما أن اللغة العربية تتميز بالكتابة من اليمين لليسار. كما أن لها ثمانية علامات مميزة diacritic Marks والتي يمكن أن تغير معنى الكلمة بطريقة جذرية اعتمادًا على مكانها في الكلمة وعلى سبيل المثال فكلمة (كَتبَ) تعني he wrote وكلمة كُتُب تعني books وهكذا كما أن الحروف العربية عددها (٢٨) حرف، ولكن كل حرف يمكن أن يكون له أشكالًا مختلفة قد تصل إلى أربعة أشكال في الكتابة: منفرد/ في بداية الكلمة/ في وسطها/ أو في نهايتها، وذلك بناء على مكان هذه الحروف في الكلمة، وكل اسم في الحروف العربية يتم نطقه ككلمة (مثلًا حرف أ ينطق أليف Alif وحرف ل ينطق لام (Lam) أي أن الصوت الفردي لا يمكن استخدمه للدلالة على حروف ألفبائية وبالتالي فإن المختصرات لا تستخدم في اللغة العربية كما هو الحال في اللغة الإنجليزية (Moukdad & Large, 2001)، كما أن اللغة العربية تتميز عن أخواتها السامية بالأمور التالية:

- أنها أكثر أخواتها احتفاظًا بالأصوات السامية.
- أنها أوسع أخواتها جميعًا، وأدقها في قواعد النحو والصرف.
- أنها أوسع أخواتها ثروة في أصول الكلمات والمفردات.

(*) لقد جاءت هذه خصائص اللغة العربية وارتباطها باسترجاع المعلومات في مصادر كثيرة تذكر منها باللغة الإنجليزية: الموسوعة البريطانية (٢٠٠٥) والموكداد وفرعل (Moukdad & Large, 2001)، وإبراهيم أبو الخير (ch, 11, ARIST 2008) وأبو سالم (١٩٩٩) والخراشي (١٩٩٤ (٢٠٠٤) ودرويش (٢٠٠٢) (٢٠٠٣) والحميدي (١٩٩٥) وغيرهم... الخ وباللغة العربية (محمد سالم غنيم ٢٠٠٨) وأحمد بدر ومحمد فتحي عبد الهادي ونارمان متولي (٢٠٠٦) وأسامة لطفي (١٩٩٩) وحشمت قاسم (١٩٩٥) وعلي السليمان الصوينع (١٩٩٤) ومساعد الطيار (١٩٩٨) (٢٠٠٠) ونبيل على (عدة مؤلفات)... إلخ.

هذا ومعظم الكلمات العربية مشتقة صرفيًا من قائمة قصيرة من الجذور المولدة، وهذه الجذور قد تكون ثلاثية أو رباعية أو خماسية الأضلاع، أما الجمع العربي Arabic Plural فهو معقد للغاية صرفيًا Morphologically لأنه يعكس عدم انتظام في الشكل أي أكثر تعقيدًا من الكلمة الإنجليزية المناظرة. اعتمادًا على الجذر فقد يكون الجمع عن طريق إضافة لواصق أو بداية الكلمة Suffixes or Prefixes أو وسط الكلمة Infixes (Moukdad & Large, 2001).

أي أن اللغة العربية تمتلك المقومات والدقة الصارمة وأن ألفاظها هي أوزان موسيقية وهي اللغة الحية الوحيدة في العالم الذي استمرت دون تغيير في كلماتها ونحوها وتراكيبها منذ حوالي خمس عشرة قرنًا فضلًا عن ضبطها للأساليب البلاغية ودلالات المعاني الكثيرة وأخيرًا فمعظم مشتقاتها تقبل التصريف وبالتالي فهي أكثر مرونة بالنسبة لتلبية احتياجات مريديها.

٢-٢ اللغة العربية واسترجاع المعلومات:

على الرغم من الزيادة المستمرة للمستخدمين للإنترنت من الذين يتحدثون بالعربية إلا أن محركات البحث باللغة العربية ما زالت في وضع متدني.

تتأثر عملية استرجاع المعلومات باللغة وكيفية معالجة محركات البحث لخصائص هذه اللغة (Moukdad, 2004) فاللغة العربية تحمل حوالي خمسة مليون كلمة مشتقة من حوالي (١١,٣٥٠) جذر Roots مقارنة باللغة الإنجليزية التي تحتوي على ١,٣ مليون كلمة من بينها ٤٠٠,٠٠٠ كلمة مفتاحية أي أن اللغة العربية تحتوي على كلمات أكثر من اللغة الإنجليزية، لأن اللغة العربية تحتوي على مئات الاشتقاقات من جذر واحد وهناك خاصية مميزة للغة العربية وهي خاصية الكلمات التي تحمل معاني متعددة، وهذه الخاصية الأخيرة تمثل تحديات كثيرة لعملية الاسترجاع، وهناك مشكلة أخرى تتصل باللغة العربية على الخط المباشر في مناطق مختلفة داخل العالم العربي حيث تستخدم كلمات مختلفة لنفس الشيء، فمصطلح Newspaper يترجم في بعض الدول على إنه صحيفة وفي بعضها الآخر جريدة ويؤدي ذلك إلى أن هناك استجابة لسؤال البحث يكون غير صالح Not relevant ومن هنا ظهر الفرض الذي يقول البحث سيكون أكثر دقة وأكثر تحديدًا إذا اعتمد على المعاني وليس الكلمات.

وإذا كانت اللغة الإنجليزية تحتل مكانًا مركزيًا في بحوث استرجاع المعلومات (يصل المحتوى إلى حوالي ٦٧%) في شبكة الإنترنت، فإن المحتوى العربي المنشور إلكترونيًا على الشبكة يصل إلى المرتبة العاشرة (يصل المحتوى إلى حوالي ٠,٠١% على الإنترنت) ويحتل المحتوى باللغات الأخرى إلى حوالي ٣٢% على نفس الشبكة (نبيل عبد الرحمن: ٢٠١١).

ومن الممكن أن نقص هذه البحوث العربية في مجال استرجاع المعلومات يعود إلى خصائص اللغة العربية حيث يقال عادة بأنها لغة اشتقاقية بينما اللغة الإنجليزية لغة لصقية. فضلًا عن نقص الميزانيات

ومصادر التقييم المعيارية للممارسين في حقل استرجاع المعلومات العربية.

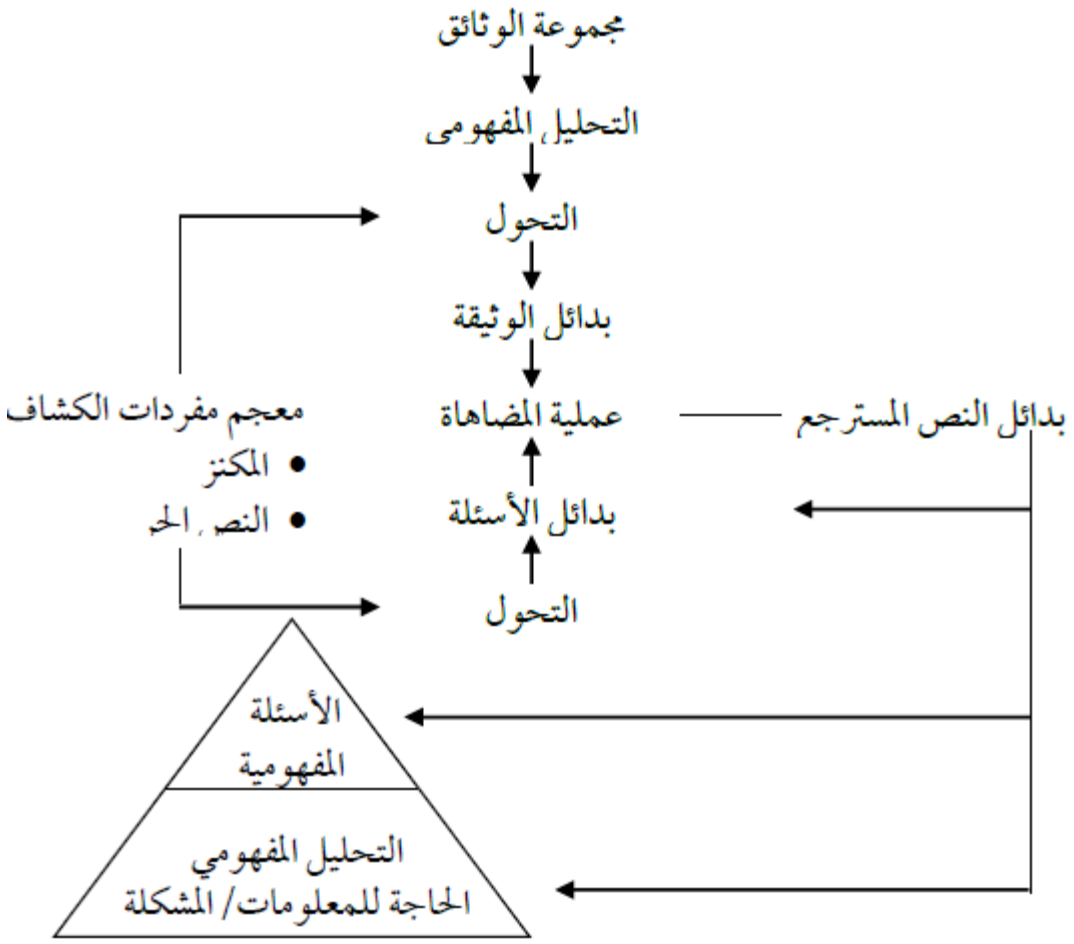
وتتناول هذه الدراسة خصائص اللغة العربية واتجاهات البحوث تطورها خاصة في القرن الحادي والعشرين مع دخول تريك (TREC) في بحوث المجال عام ٢٠٠٠.

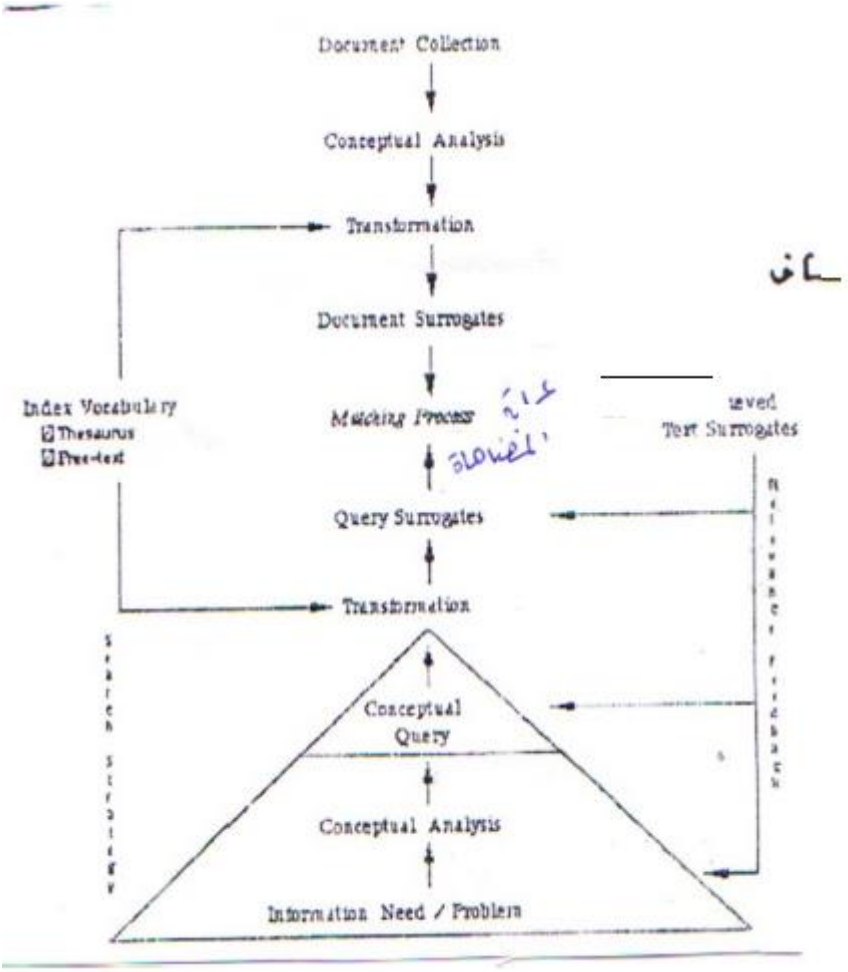
وإذا كان استرجاع المعلومات العربية ما زال في مراحله الأولى، فلا بد أن نعمل على الإفادة من البحوث المتطورة التي تمت في اللغة الإنجليزية، ومن هنا كان الاهتمام في هذه الدراسة ببعض التطورات التي حدثت في تقييم الاستدعاء والدقة Recall & Precision وفي محركات البحث العربية التي ما زالت في وضع ضعيف وغيرها من التطورات، وإذا كانت معظم البحوث العربية قد توقفت عند عام ٢٠٠٤ فتحاول هذه الدراسة التركيز على بعض التطورات التي حدثت حتى عام (٢٠١٣).

ثالثاً: نماذج استرجاع المعلومات:

النماذج الكلاسيكية:

في استرجاع المعلومات هي البوليني والمتجهات والاحتمالي، وفي النموذج البوليني تمثل كل من الوثائق والأسئلة بمجموعة من المصطلحات الكشفية وفي نموذج المتجهات فإن الوثائق والأسئلة تمثل بواسطة المتجهات وهو نموذج يرتكز على رياضيات الجبر، أما النموذج الاحتمالي فهو يعتمد على أن التمثيلات تعتمد على نظرية الاحتمالات، ولما كانت عملية استرجاع الوثائق كاستجابة لطلب وأسئلة الباحث تتم بناء على مصطلحات التكشيف، فإن كثيرًا من المعاني والدلالات Semantics) في الوثائق وأسئلة المستفيد تفقد عندما يتم وضع النص مكان الكلمات، والنتيجة زيادة عدد الوثائق غير الصالحة irrelevant، فضلًا عن أن معظم المستفيدين ليس لديهم تدريب في صياغة الأسئلة ومن هنا فقد اقترح بعض الباحثين الإفادة من البدائل Surrogates لكل من الوثائق والأسئلة كما تظهر في الشكل التالي:


مجموعة الوثائق
التحليل المفهومى
التحول
بدائل الوثيقة
عملية المضاهاة
بدائل النص المسترجع
معجم مفردات الكشاف
• المكنز
• النص الحر
بدائل الأسئلة
التحول
الأسئلة المفهومية
التحليل المفهومي الحاجة للمعلومات/ المشكلة


Document Collection
Conceptual Analysis
Transformation
Document Surrogates
Index Vocabulary
Thesaurus
Free-text
Matching Process
Query Surrogates
Transformation
Conceptual Query
Conceptual Analysis
Information Need / Problem

The figure above: represents a general model of the information retrieval process, where both queues of the user's information need and the document collection have be translated into the form of surrogates to enable the matching process to be performed. The figure has been adapted from Lancaster and warner (1993).

http://www.scils.rutgers.edu/asporrn/infocrystal/ch.2html.

يمثل هذا الرسم نموذج عام لعملية استرجاع المعلومات حيث يتم ترجمة كل من الأسئلة الخاصة باحتياجات المستفيد للمعلومات، مع مجموعة الوثائق وهذه الترجمة للأسئلة والوثائق تكون على شكل بدائل Surrogates حتى يمكن القيام بعملية المضاهاة وقد تم تطويع هذا الرسم من كتاب لانكستر ووارنر (١٩٩٣).

وفيما يلي نبذة عن نماذج استرجاع المعلومات بصفة عامة:

١- نموذج البحث البولييني (Boolean Search Model):

وهو يعتمد على نظرية المجموعة Set وعلى الجبر البوليني حيث يسمح للمستفيد بالتعبير عن استفساره باستخدام ثلاث عوامل للربط بين المصطلحات هي (و and أو or ليس not أو فيما عدا)، ويعتمد هذا النموذج على مضاهاة مصطلحات الاستفسار البولييني مع المصطلحات المستخدمة لتمثيل محتوى الوثائق، ليحدد ما إذا كان مصطلح الاستفسار يظهر أولا يظهر ضمن المصطلحات التي تمثل الوثيقة أو الوثائق في قاعدة البيانات التي تضم التسجيلات الببلوجرافية. ولا بد أن يكون المصطلحات الخاصة بمجموعة الوثائق مطابقة لتلك التي ترد في استفسار البحث حتى يتم استرجاعها، ولا يسمح هذا النموذج بترتيب بيانات الوثائق المسترجعة حسب أهميتها ومدى صلتها بالموضوع، فالمواد المسترجعة تعد على نفس الدرجة من الأهمية.

٢- نموذج الاسترجاع الاحتمالي (Probabilistic Retrieval Model):

وقد عرف فيما بعد بأنه الاسترجاع المستقل الثنائي، حيث يتم في هذا النموذج استخدام نظرية الاحتمالات كأساس لعملية المعالجة، فبدلًا من مطابقة نفس المصطلحات الواردة في استفسار البحث مع المصطلحات الواردة في وصف الوثائق، فإنه يتم وفقًا للنموذج الاحتمالي إحصاء أو تقدير الاحتمالات التي يمكن أن تكون فيها الوثيقة ذات صلة بمستفيد معين، ومن ثم ترتب الوثائق المسترجعة

ترتيبًا تنازليًّا وفقًا لاحتمالات صلتها بالاستفسار وفائدتها بالنسبة للمستفيد؛ فعلى سبيل المثال قد يتم تحديد الوثيقة ذات الصلة عن طريق توظيف المعلومات التاريخية لاستخدام تلك الوثيقة لاحتساب وإحصاء احتمالات صلتها بالاستفسار. والمقصود بذلك أن يتم تتبع عدد المرات التي حكم فيها المستفيدون على الوثيقة من قبل أشخاص آخرين استخدموا نفس مصطلح البحث، فإنه يُمكن الحكم على الوثيقة بأنها ذات صلة بالاستفسار في حالة استخدامهم لنفس مصطلح البحث، وبمعنى آخر إذا استخدم مستفيد مصطلح بحث لاسترجاع وثائق، وحكم على وثيقة من بينها على أنها ذات صلة بالموضوع، وتكرر هذا الحكم على الوثيقة بأنها ذات صلة بالاستفسار، ويذكر أن هذه طريقة واحدة من بين طرق متعددة تستخدم لتحديد احتمالات صلة الوثيقة بالاستفسار وكان هناك اختلاف بين الباحثين حول المقارنة بين الاسترجاع الاحتمالي ونموذج الاتجاهات، وانتهت إلى أن نموذج الاتجاهات يسبق نموذج الاحتمالات بالنسبة للمجموعات العامة (Baeza – Yates p34).

٣- نموذج معالجة حيز المتجهات (The Vector Processing Model):

يعمل على تحديد درجة التشابه بين الوثائق والاستفسارات من خلال قياس أوزان المتجهات للمصطلحات، ويمكن وفقًا لهذا النموذج الحكم على تشابه وثيقتين باحتساب مصطلحات التكشيف المتشابهة فيها، كما يوضع في الحسبان أيضًا المصطلحات المشتركة التي تغيب عنهما معًا. وبذلك فإنه يمكن استخدام مقارنة حيز المتجهات في نصين مع بعضهما لتحديد التشابه بين النصين، كما يمكن احتساب حيز المتجهات في الاستفسار (query vectors) ومقارنته بحيز المتجهات في الوثائق المخزنة لتحديد التشابه بينهما.

The term weights are ultimately used to compute the degree of similarity between each document and user query.

٤- نموذج البحث الأفضل مضاهاة والتغذية الراجعة ذات الصلة (Best Match Searching and Relevance Feedback Model):

يمكن بتطبيق هذا النموذج الحصول على المخرجات مرتبة (ranked) حسب ارتباطها بالاستفسار. ويتطلب ذلك تحديد أوزان مصطلحات البحث؛ حيث يتم وفقًا لنظام وزن المصطلحات تحديد قيمة رقمية لكل مصطلح تكشيف سواء في الاستفسار أو في الوثيقة بحيث تعكس تلك القيمة أهمية المصطلح، ويتم استخدام تلك الأوزان لاحتساب مدى التشابه بين الوثيقة والاستفسار عند إجراء عملية المضاهاة بين مصطلحات الاستفسار في مقابل مصطلحات الوثائق. ويتم بعد ذلك ترتب المواد المسترجعة ترتيبًا تنازليًّا حسب درجة التشابه بينها وبين الاستفسار، ووفقًا لتلك النظرية فإن الوثائق التي تضم مصطلحات أكثر واردة في الاستفسار فإنها تأتي في مقدمة قائمة النتائج المرتبة (ranked list).

ويمكن للمستخدم من خلال هذا النموذج أن يضع استفساره في شكل عبارة باللغة الطبيعية، ويتولى النظام ضبط الاختلافات في الإملاء والمترادفات في حالة تطبيقه بعض الإجراءات الآلية مثل: الإجراءات الحسابية (Computational procedure)، وإجراءات الدمج (Conflation Procedures) الممثلة في استخدام لوغاريتمات التجذير (Stemming Algorithms) التي تعمل على إرجاع المصطلح بأشكاله المختلفة إلى شكل واحد بتجريده من اللواحق والسوابق التصريفية والاشتقاقية.

أما عملية التغذية الراجعة ذات الصلة فتقوم فكرتها على أن تعمل النظم على احتساب أوزان المصطلحات مرة أخرى بعد حصول المستفيد على نتيجة البحث مرتبة (ranked)، ويختار منها ما يجده ملائمًا وذو صلة باستفساره، وبالتالي فإن النظام يحصل على تغذية راجعة نتيجة لاستخدام المستفيدين للنتائج، وذلك على اعتبار أن استخدام المستفيدين لوثائق معينة يدل على وثاقة صلتها بالاستفسار، ومن ثم فإن النظام يعيد احتساب أوزان الوثائق وفقًا للتغذية الراجعة التي يحصل عليها، ويعيد ترتيب الوثائق تبعًا لذلك.

٥- نموذج معالجة اللغة الطبيعية (the natural Language Processing):

بتطبيق هذا النموذج فإن النظام لا يعتمد على مصطلحات الاستفسار والوثيقة فقط، ولكن يعالج الجمل والصيغ ويعمل على مضاهاتها. ويتطلب بناء النظم التي تعمل على معالجة نصوص اللغة الطبيعية على ثلاث مستويات من المعالجة هي:

*** التحليلي النحوي (Syntactic analysis):**

يتطلب فهم بناء الجمل ويعتمد على الكلمات التي تضمنها القواميس (lexicon) والمعلومات المقترنة بها مثل القواعد والعلامات النحوية.

*** التحليل الدلالي (Semantic analysis):**

يتعامل مع معاني الكلمات في الجمل وفقًا لما هو مخزن في قاعدة المعرفة.

*** التحليل الواقعي أو العملي (Pragmatic analysis):**

يأخذ في الاعتبار السياق الذي جاءت فيه المصطلحات (Ghowdhury, G. 2004).

هذا ويتطلب وضع الأسئلة الجيدة good Queries التفكير في الكلمات التي تظهر في الوثيقة ذات الصلة باستخدام هذه الكلمات كأسئلة ومدخل نمذجة اللغة Modeling لاسترجاع المعلومات تتصل مباشرة بنمذجة هذه الفكرة. فالوثيقة هي مضاهاة جيدة good match للسؤال. إذا كان نموذج الوثيقة يمكن أن يولد السؤال، وهذا المدخل بالتالي يقدم لنا إدراكًا مختلفًا لبعض الأفكار الأساسية المتصلة برتبة ranking الوثيقة... ذلك لأن الوثيقة التي تذكر مصطلحات السؤال أكثر من غيرها هي

التي تأخذ درجة أعلى في الترتيب (Manning et al 2008: 218).

رابعاً: التطورات الحديثة في تقييم وقياس نظم استرجاع المعلومات العربية:

٤-١ مقدمة:

قام ثلاثة من الباحثين في المجال وهم أحمد عبد اللالي وجيم كوي وحمدي سليمان (Abdelali, Ahmed; Comie, Jim; Soliman, Hamdy,. April 2004 بدراسة جيدة عن منظورات استرجاع المعلومات العربية، وقد تناول المؤلفون الثلاثة بعض المصادر المتوفرة من أجل اختبار نظم استرجاع المعلومات العربية لاسيما وأن هناك معايير تقييم منذ عام ٢٠٠٠ بواسطة (CLEF/ TREC).

وقد كان للنمو السريع للويب مصحوبًا بالمصادر المتعددة اللغات، وأدوات التكشف والاسترجاع أثره الواضح في تطور استرجاع المعلومات المتعددة اللغات، وتشير الإحصائيات أنه بعد عام ١٩٩٥ أي عند صدور جريدة الشرق الأوسط على الخط المباشر ازدادت مواقع الويب العربية بطريقة تضاعفية حيث وصلت عام ٢٠٠٠ إلى حوالي ٢٠،٠٠٠ موقع عربي (أي حوالي ٧% من المواقع المنشورة على الويب) وإذا كان عدد الحروف العربية هو (٢٨) حرف إلا أنها يمكن أن تتسع إلى تسعين عنصر عن طريق إضافة الأشكال Shapes والعلامات Marks وحروف العلة Vowels ومعظم الكلمات العربية مشتقة من الجذور Roots المكونة من ثلاثة حروف ساكنة Consonants.

وقد تصل الجذور إلى اثنين أو أربعة أو نادرًا خمسة حروف ساكنة.

٤-٢ أدوات استرجاع المعلومات العربية:

يمكن تقسيم هذه الأدوات إلى فئتين هما:

(أ) الشكل الكامل المعتمد على استرجاع المعلومات Full Form based – IR:

معظم محركات البحث التجارية هي من هذا النوع مثل محرك بحث الويب صخر www.alidrisi.com، www.ayna.com وغيرها من المحركات متعددة اللغات ذات الكود الواحد UNI code مثل: www.google.com أو www.alltheweb.com.

(ب) المورفولوجي المعتمد على استرجاع المعلومات Morphology based – IR:

تشير الجهود المبذولة في البيئة الأكاديمية لتقييم نظم أكثر تعقيدًا إلى الجيل القادم من محركات البحث العربية، حيث توجد عدة مداخل شاملة علم أحرف اللغة Morphology الجذوع والجذور المعتمد على التجذير الخفيف Light Stem (larkey et al 2002) وهناك أيضًا أجهزة التجذير stemmer غير الإحصائية أو نماذج n – gram وبصفة عامة يمكن القول أن استخدام أجهزة

التجذير stemmers فإنها تحسن كلًّا من الاستدعاء recall والتحقيق Precision كما أظهرت تجارب لاركي (Larkey, Connel, 2002) أن أجهزة التجذير الخفيفة light stemmers التجذير العادية.

٤-٣ المجموعات الضخمة Coropa:

وجدت هذه المجموعات الضخمة اللازمة في اختبارات استرجاع المعلومات مع دخول TREC فهناك مجموعة LCD (869 megabytes) من المقالات الإخبارية العربية والتي تضم أكثر من ٣٨٣,٨٧٢ وثيقة من وكالة الصحافة الفرنسية AFP والتي استخدمت في تقييمات TREC المعاجم Lexicons. ومن بين المصادر الأخرى هناك قواميس أحادية اللغة Monolingual وثنائية اللغة bi lingual – ومنها على الخط المباشر on-line والتي استخدمها لاركي (Larkey et al 2002/ 2003) في بعض تجاربه، وهناك جهود فردية أخرى استخدمت في تطبيقات مختلفة مثل (Zajac, et al 2001).

٤-٤ مقاييس وأدوات التقييم Evaluation Measures and Tools:

تتمتع اللغة العربية بدرجة عالية من التصريف اللغوي Inflection ذلك لأن مورفولوجيا اللغة العربية هي علم فيه كثير من التحدي بالنسبة لاسترجاع المعلومات، وأن هذا التعقيد يأتي لعدم وجود مسافات Spaces بين الكلمات فضلًا عن الالتباس بالنسبة لبعض الحروف والكلمات وعلى سبيل المثال لا الحصر فالألف (أ) يمكن أن تكون أ أو إ، كما أن اللواحق واللواصق Prefixes and Suffixes يمكن أن تكون العلامات النحوية على هيئة اثنين أو ثلاثة أو حتى أربعة مثل (ليعلمانكما) (تعلمونهن) أو سيعلمانها they will teach it to you… you teach them ولحل هذه المشكلة هناك اتجاهان رئيسيان مستخدمان لبناء المحللات المورفولوجية morophological analysers أو الجدول المعتمد على المحللات المورفولوجية والقواعد أو التحليل المورفولوجي للباحث بيزلي (Beesley, Ken 1998) وبتحليل نظم استرجاع المعلومات المعاصرة بتحليل تريك TREC لتقييم اللغة العربية يمكن أن يقال بأن أداء نظم استرجاع الوثائق من المجموعة العربية يشبه الأداء في النظم المعاصرة في اللغات الأخرى، فتقييم تريك TREC يستخدم LDC لجسد من مقالات AFP من عام ١٩٩٤، وحتى عام ٢٠٠٠ وكانت اللغة المستخدمة هي اللغة المعيارية الحديثة MSA وقد أثبت قانون زيف Ziph's أن هذه المجموعة كافية وكاملة وممثلة، وإن كانت المجموعة صغيرة لعلم النحو Syntex والمورفولوجيا فضلًا عن الهجاء Spelling للجمل المترجمة والتي كتبت بحروف أخرى فهي غير منتظمة باللغة العربية كما هو واضح في الجدول التالي والنتيجة هي أن النظام سوف لا يقدم نتائج كاملة:

Word AFP	English	Occurrences	Word AFP	English	Occurrence
لوس أنجليس	Los Angeles	21	نيوهامبشير	New Hampshire	15
لوس أنجلوس	Los Angeles	23	نيو هامبشر	New Hampshire	٩
لوس أجيلس	Los Angeles	2	جوهانس	Johannes	4
لوس أنجيلس	Los Angeles	34	يوهانس	Johannes	74
إنجلترا	England	2	يوهانيس	Johannes	8
إنجلترا	England	1	جوهانز	Johannes	1
إنجلترا	England	1	جوهانسبورغ	Johannesburg	173
كارولاينا	Carolina	26	جوهانسبرغ	Johannesburg	15
كارولينا	Carolina	14	جوهانسوبورغ	Johannesburg	1
ويسكونسين	Wisconsin	8	جوهانسورغ	Johannesburg	1
ويسكنسن	Wisconsin	2	فايمار	Weimar	3
ويسكونسن	Wisconsin	16	فيمار	Weimar	10

مصادر تريك ٢٠٠١ المستخدمة بواسطة المشاركين.

ومما سبق نجد أن الجيل التالي لنظم استرجاع المعلومات العربية سيحتاج إلى العمل ببيانات أكثر ثراء، والبيانات الحالية على الإنترنت يمكن أن تكون نقطة بداية، وإذا كان هذا التطور قد تم عام ٢٠٠٤، فقد فضلت الدراسة الإشارة لبعض المقاييس الجديدة للبحث الذي قدمه توماس ماندل (mandl, Thomas 2006) حيث تقدم لنا هذه الدراسة للمسارات الحالية وتطورها داخل TREC, NTCIR, CLEF وإذا كانت الدراسة قد ذكرت بعض جهود TREC في الأمثلة السابقة المتصلة

باسترجاع المعلومات العربية، فإن النماذج التالية تعكس الوضع الأحدث حتى عام ٢٠٠٨ ولكن في مجال استرجاع المعلومات بصفة عامة:

(أ) مؤتمر استرجاع النص TREC Text Retrieval conference:

يعتبر هذا المؤتمر أول المبادرات التقييمية على نطاق واسع عام ٢٠٠٨ في السنة السابعة عشرة وهو ممول من المعهد الوطني للمعايير والتكنولوجيا NIST في ولاية ميريلاند بأمريكا Maryland, U. S. A ليبدأ عهدًا جديدًا في بحوث استرجاع المعلومات (Voorhees, E: 2006) لأول مرة في علم المعلومات.

هذا وقد قام تريك بتنظيم تقييمه في عدد ٢٦ مسار ومن بينها خلال السنوات الأخيرة:

- مسار الإجابة على الأسئلة QA ويتطلب نظمًا تركز على الإجابات القصيرة إلى الأسئلة المعتمدة على الحقائق في مختلف التخصصات.
- أما المسار الأكثر مشاركة من الباحثين فكان عام ٢٠٠٥ في مجال الوراثة Genomics والذي يجمع ين الوثائق العلمية النصية مع البيانات والحقائق (Hersh et al 2006) وضع هذا المؤتمر مجتمع الأنفورماتيكا البيولوجية (Bio - informatics) إلى جانب المتخصصين في استرجاع المعلومات وانتهى المؤتمر عام ٢٠٠٧.
- في مسار الاسترجاع القوي (Robust Retrieval track) استخدم قياسات تقييميه جديدة تركز على الأداء الراسخ على جميع الموضوعات بدلًا من مكافأة النظم التي قدمت متوسط جيد للدقة precision.
- مسار المدونات The Blog Track والذي بدأ عام ٢٠٠٦ لاكتشاف السلوك ألمعلوماتي في المجموعات الضخمة من البيانات الاجتماعية المحسبة.

(ب) ندوة تقييم عبر اللغات Cross Language Evaluation Forum:

في عام ٢٠٠٠ كان تقييم نظم استرجاع المعلومات المتعددة اللغات قد انتقل إلى أوروبا وعقد المؤتمر الأول هناك، فكل لغة لها مورفولوجيات مختلفة عن الأخرى في إنشاء الكلمات والمترادفات وغيرها وبالتالي فلا بد من احتياج المصادر اللغوية للوغاريتم لاسترجاع كل لغة على حدة ويهدف CLEF إلى تدعيم ذلك الاتجاه وركز CLEF على مجموعات وثائق اللغات: الإنجليزية والفرنسية والأسبانية والإيطالية والألمانية والهولندية والتشيكية السويدية والروسية والفنلندية والبرتغالية والبلغارية والهنجارية وتشير نتائج بحوث CLEF أنها أدت إلى تقدم علمي فضلًا عن تحسن في النظام، وعلى سبيل المثال فقد كانت الأرقام n – grams يمكن أن تمثل النص بدلًا من المصطلحات التجذيرية Mc Namee)

.and Mayfeld, 2004) Stemmed Terms

(ج) المبادرة اللغوية الآسيوية Asian Language initiative (NICIR):

هذه المبادرة تركز على تكنولوجيات اللغات المحددة اللازمة للغات الآسيوية وعبر البحوث اللغوية ومن بينها اللغة الإنجليزية (Kando and Evans, 2007).

وشملت عبر اللغات Cross languages اللغات الآسيوية الثلاث وهي الصينية واليابانية والكورية، لاسيما وأن هذه اللغات تعتبر أكثر تعقيدًا من اللغات في الدول الغربية.

أما بالنسبة لمقاييس التقييم:

فقد ظهر الاختلاف Variance بين الأسئلة Queries أكبر من الاختلاف بين النظم، فهناك أسئلة صعبة للغاية، وهناك نظم قليلة تحلها، ولكنها تؤدي إلى نتائج سيئة جدًّا في معظم النظم (Harman and et al, 2004) وما زالت مشكلة ربط تقييم النتائج مع إرضاء احتياجات المستفيدين غير واضحة (Turpin, A, et al., 2006) وما زالت المقاييس التقليدية هي السائدة حول الاستدعاء Recall والدقة Precision كما أظهرت النتائج العامة أن تقييم نظم استرجاع المعلومات في مجال معين لا يمكن نقلها إلى مجالات أخرى وهناك تطورات رياضية بالنسبة لقياس الاستدعاء والدقة وردت في كتاب Baeza – Yates كما يلي:

٤-٥ مدى ملائمة الاستدعاء والدقة للتقييم الحديث لكل من استرجاع المعلومات واسترجاع المعلومات العربية:

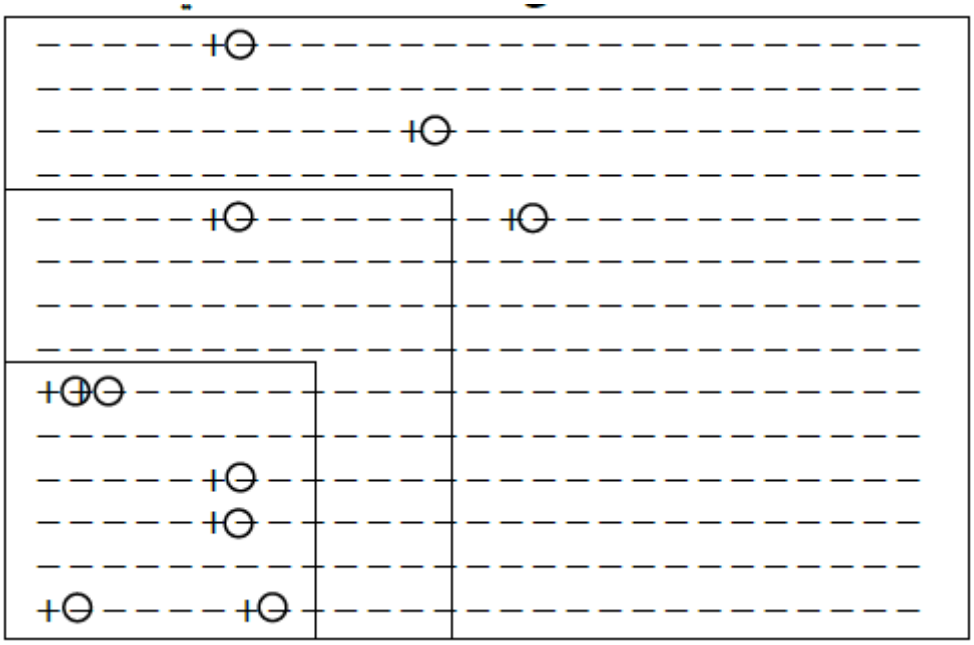
المقصود بالاستدعاء والدقة Recall and Precision:

في أحد البحوث في مجال استرجاع المعلومات كانت النتائج الخاصة بالقياس كما يلي:

(٥٧) وثيقة استرجعت في المربع الأصغر ومن بينها كان عدد الوثائق ذات الصلة (٦) فقط والباقي (٥١) غير صالحة irrelevant وبالتالي فتحسب الدقة:

$$\frac{٦}{٥٧} = ٠,١ \text{ والاستدعاء } \frac{٦}{١١} = ٠,٥$$

والعدد (١١) يدلنا على مجموع الوثائق ذات الصلة أي الصالحة (٣ + ٢ + ٦).



لقد استخدم الاستدعاء Recall والدقة Precision بشكل واسع لتقييم أداء لوغاريتم الاسترجاع ومع ذلك فهناك مشكلات مع هذين القياسين (Tague – Sutcliff, 1992/ Raghaven et al 1989).

أولًا: أن التقدير الصحيح لأعلى درجة في الاستدعاء بالنسبة لسؤال معين يطلب معرفة بكل الوثائق في المجموعة، وفي المجموعات الضخمة مثل هذه المعرفة غير متوفرة وبالتالي فإن الاستدعاء لا يمكن تقديره بدقة.

ثانيًا: أن الاستدعاء والدقة هي مقاييس مرتبطة ببعضها وفي مواقف عديدة فإن استخدم قياس واحد يضم كلًّا من الاستدعاء والدقة قد يكون هو المناسب.

ثالثها: أن مقياس الاستدعاء والدقة يقيس تأثير مجموعة من الأسئلة والتي يتم تجهيزها بطريقة الدفعة Batch Processing على عكس النظم الحديثة التي تعتبر الجانب ألمفتاحي هو التفاعلية Interactivity (وليس التجهيز على دفعات) بالنسبة لعملية الاسترجاع، ونتيجة لذلك فإن القياسات التي تقوم بالتقدير الكمي للمعلوماتية Quantify Informativenes تكون أكثر صلاحية لعملية الاسترجاع وأخيرًا فإن تعريف الاستدعاء والدقة يكون يسيرًا عندما يكون الترتيب الخطي Linear Ordering إجبارياً، أما بالنسبة للنظم التي تتطلب ترتيبًا ضعيفًا فإن الاستدعاء والدقة ستكون غير كافية (Baeza, Yates, 1999: 81) ثم قام المؤلف بتفصيل البدائل المتصلة بالقياسات للاستدعاء والدقة ومن بينها:

١- **المتوسط الهارموني (F) Harmonic mean**: للاستدعاء والدقة والذي يجمع بين القياسية كما جاء في البند ثانيًا أعلاه ومعادلته كما يلي:

$$F(J) = \frac{2}{\frac{1}{R(J)} + \frac{1}{P(J)}}$$

حيث يعتبر (J) هي الاستدعاء Recall بالنسبة للوثيقة (J) وترتيبها Ranking هو p(J) هو الدقة Precision بالنسبة للوثيقة (J) في الترتيب أما F (J) فهو المتوسط الهارموني لكل من r (J)، P (J) أما الدالة F (Function) فتدلنا على القيم Values في المسافة (٠,٠١)، والقيمية (صفر) عندما لا يسترجع النظام أي وثيقة صالحة والرقم (١) يدلنا على أن جميع الوثائق المرتبة تعتبر صالحة، كما يفترض في المتوسط الهارموني (F) قيمة عالية فقط عندما يكون كلًّا من الاستدعاء Recall والدقة Precision عاليًا وبالتالي فإن تحديدنا للقيمة الأعلى للدالة (F) يمكن تفسيرها كمحاولة للتعرف على الحل الوسط Compromise بين الاستدعاء والدقة.

٢- قياس E Measure E:

ويطلق عليه مقياس تقييم (E) وفكرته تدور حول السماح للمستفيد بتحديد أولوياته: هل يركز على الاستدعاء أو على الدقة، وله معادلاته المختلفة عن السابق وهناك كذلك المقاييس التي تدور حول المستفيد User Oriented Measures بالإضافة إلى قياسات المجموعات المرجعية مثل (/Tipster TREC) ومؤتمرات تريك السنوية التي يشرف عليها NIST وهو المعهد الوطني للمعايير والتكنولوجيا بالولايات المتحدة الأمريكية وهناك أربعة أشكال أساسية لقياسات التقييم وهي:

(a) **إحصائيات جدول التلخيص**: ويحتوي الجدول على إحصائيات بعدد الموضوعات Topics = information requests وعدد الوثائق المسترجعة لهذه الموضوعات وعدد الوثائق الصالحة وأخيرًا الوثائق الصالحة التي كان من الممكن أن تسترجع لكل الموضوعات.

(b) **متوسطات الاستدعاء والدقة**: وهذه تحتوي على جدول أو رسم بياني graph بمتوسط الدقة Precision لكل الموضوعات بالنسبة لعدد (١١) من مستويات الاستدعاء المعيارية.

(c) **متوسطات مستوى الوثائق**:وتشمل متوسط الدقة Precision لجميع الموضوعات عند قيم محددة Specified document cull of value بدلًا من المستويات المعيارية للاستدعاء.

(d) **Average Precision Histogram**: ويحتوي على رسم قياس لكل موضوع على حدة، ويتم ذلك شاملًا لجميع نتائج نظم الاسترجاع المشاركة هذا إلى جانب مجموعات الاختبار وهذه كلها قد ازدهرت مع دخول الويب Web والإنترنت.

وختامًا لهذا كله فيجب الاعتراف بأن هناك مشكلات عديدة تنتظر الحل ومنها ما هي المعايير التي يتبناها المستفيد عندما يحكم على مدى صلاحية الوثائق للسؤال الذي وضعه Judge Reference لأن هذه المعايير ربما تؤثر على إستراتيجية الرتب Ranking وعلى واجهات التفاعل بين المستفيد والنظام Interfaces والمشكلة مع الويب Web واضحة بل وفريدة ذلك لأن مستخدم الويب لا يعرف في أحيان كثيرة ماذا يريده بالفعل وعنده صعوبة في صياغة أسئلته وكيفية تأثير ذلك على الترتيب Ranking مع ضآلة حجم الكشافات التي تحتفظ فيها محركات البحث وفي تقدير الباحث – Baeza Yates تصل هذه الكشافات إلى ٢% فقط من العدد الكلي للصفحات المكشفة ويقترح حلًا هو محركات البحث عن محركات البحث Meta Search engines.

وأخيرًا فتشير (ناريمان متولي ٢٠٠٤) عن مشكلات القياس والصلاحية:

«لقد كان الاستخدام الأول لمصطلح الصلاحية Relevance كمعيار أداء لنظام استرجاع المعلومات، وكان معيار الفاعلية Effectiveness المستخدم هو صلاحية الوثائق لأسئلة حقيقية. ولكن هذا الاختبار فشل بالنسبة لقضية الحكم على الصلاحية، ومن بين التجارب التي تمت على اختبارات كرانفليد تبين وجود عدد ضخم من العوامل التي أثرت على انتظام أحكام الصلاحية، وهذه العوامل تشمل نوع الوثيقة وموضوعاتها ومستوى الصعوبة وأسلوب الوثيقة ودرجة التركيز، والزمن المتاح للقائم بالحكم ونظام ترتيب المواد... الخ مما يؤدي إلى انخفاض أو ارتفاع درجات الصلاحية (ناريمان متولي، ٢٠٠٤).

ومع ذلك وبعد دراسات وتجارب عديدة فقد تبين أن عدم مقدرة الإنسان على تحديد احتياجاته الفعلية بشكل مبسط، سيؤدي إلى سحب البساط من تحت أقام العديد من دراسات التقييم لنظم استرجاع المعلومات.

على الرغم من إمكانية وضع أفكار موضوعية للصلاحية، فالعديد من الباحثين يعتقدون بأن الصلاحية هي بالضرورة مجال ذاتي Subjective أي أن السؤال الواحد الموضوع بواسطة مستفيدين مختلفين، سيكون وثائق مختلفة توصف بالصلاحية.

خامساً – بدايات البحوث في مجال استرجاع المعلومات العربية:

تتناول هذه البدايات نماذج من دراسات المعلومات ونماذج الدراسات اللغوية الإحصائية حتى نهاية القرن العشرين، ولا تضم هذه الدراسة هنا الدراسات حسب تسلسلها التاريخي فحسب وإنما تذكر الدراسة ما طرأ عليها من نقد إيجابي أو سلبي من باحثين آخرين.

١/٥ فمن الدراسات المبكرة عن المجال اللغوي والإحصائي قائمة داود عطية ١٩٧٩ عن المفردات الشائعة في اللغة العربية، حيث احتوت دراسته على عدد (٣٠٢٥).

كلمة اختارها من قوائم أخرى للمفردات الشائعة وتضم في مجملها على (٧٠٠,٠٠٠) كلمة، ورتبها هجائياً في القسم الأول ثم تنازلياً حسب عدد مرات تكرارها ويصف بعض المتخصصين هذه القائمة بأنها تعد أشمل محاولة تمت في اللغة العربية لتحديد الكلمات الشائعة (محمد سالم غنيم: ٢٠٠٨).

٢/٥ أما بالنسبة لدراسات المعلومات فتعتبر دراسة حشمت قاسم ١٩٧٨ عن اللغة العربية في نظم المعلومات المتخصصة، واحدة من الدراسات المبكرة في المجال حيث حاول مقارنة كفاءة أساليب الاسترجاع العربية بنظائرها باللغة الإنجليزية، واختار حشمت قاسم في مجال اللغويات (٢٧) سؤالًا، كما قام بوضع نظام الترجمة الحرفية Transliteration لتجنب المعالجة بالحروف العربية واستخدام أسلوب المضاهاة وكانت النتيجة أنه ليس هناك اختلاف بين الاسترجاع باللغتين العربية والإنجليزية باستخدام قياسات الاستدعاء والدقة، وواضح أن التحدي الذي واجهه هو تمثيل اللغة العربية بالآلات (وهذا ليس مشكلة في الوقت الحاضر نظرًا لأن الآلات قد يسرت كثيرًا من مشكلات اللغة العربية في نظم الاسترجاع (Abu El – Khair, 2003).

٣/٥ قدم حشمت قاسم بعد ذلك عدة كتب ودراسات احتوت على مجال استرجاع المعلومات من بينها دراسته بعنوان (كشاف الكلمات المفتاحية في السياق واحتمالاته في اللغة العربية).

وأكد على أنه ليس هناك في طبيعة اللغة العربية والخصائص البنائية والدلالية لمفرداتها، ما يحول دون استخدام هذا النوع من الكشافات بكفاءة لا تقل عما هو الحال في اللغات الأخرى فضلًا عن إمكانية استخدام الحاسب الآلي لإصدار مثل هذه الكشافات (حشمت قاسم: ١٩٨٥).

٤/٥ وفي نفس الاتجاه كانت دراسة فتحي عثمان أبو النجا ١٩٨٤ الذي نال بها درجة الدكتوراه من جامعة القاهرة بعنوان (وضع نظام عربي لاسترجاع المعلومات في قطاع المعلومات الزراعية، وقد وضع نظم لاختزان المعلومات واسترجاعها يجمع بين الربط المسبق والربط اللاحق، يمكن أن يحقق أعلى درجة استدعاء مع نسبة عالية من الدقة وبذلك تتوافر مقومات نظام المعلومات الجيد.

٥/٥ أما بالنسبة لمحمد عبد الله الأطرم ١٩٩٠: فقد كان مشرفًا على مشروع دعمته مدينة الملك عبد العزيز للعلوم والتقنية بالمملكة العربية السعودية، واستغرق المشروع عامين (١٩٨٩ – ١٩٨٧) ونشر عام ١٩٩٠ في مجلدين بعنوان (كفاءة اللغة العربية في تكشيف واسترجاع الوثائق العربية) وقد اتضح في نهاية المشروع أن للتكشيف الآلي للنصوص العربية وفرزها يتسم بدرجة عالية من الدقة، بشرط تصميم برنامج للتكشيف يعالج قضايا اللغة العربية مبني على فهم كاف لطبيعة اللغة العربية وخصائصها المتميزة.

هذا وقد تأثرت البحوث الأولى بالطريقة التقليدية لتكشيف النص العربي باستخدام كلمات الجذر Root ووضعت نظم تعتمد على التحليل الصرفي واستخراج الجذور Root Extraction والتعليق على ذلك هو أن معظم هذه النظم لم يتم اختبارها باستخدام أساليب تقييم استرجاع المعلومات المعيارية والتي

وضعت فيما بعد مثل أعمال لاركي وزملائه (Larkey, et al 2007).

٦/٥ قام الباحثان الفيداغي والعنزي (Fedaghi and Al – Anzi, 1989) بوضع نظام للعثور على الجذور ذات الحروف الثلاثة (Trilateral) في الكلمة العربية، وللنظام قائمتان: قائمة النماذج (Patterns) وأخرى للجذور العربية الواضحة بالأحرف الثلاثة، وقائمة النماذج لا تحتوي فقط على النماذج العربية الأساسية، ولكنها تضم أيضًا نماذج اللواصق (أول وآخر الكلمة) الصحيحة (Affixes) وحسب ما يذهب إليه الفيداغي والعنزي، فإن Khoja & Garside استطاعا بنجاح استخراج الجذور من حوالي ٨٠% من الكلمات والتعليق على ذلك هو أن الدراسات التالية لم تجد أرقامًا دقيقة بالنسبة لهذه المحاولة الرائدة (Khoja and Garside, 1999).

٧/٥ أما الباحثان الشلبي وإيفانز (Shalabi and Evans, 1998) قاما بتوسيع لوغاريتم الفيداغي والعنزي (١٩٩٨) حيث تبين لهم بعد هذه التجربة وجود جذور ذات الأضلاع الأربعة (Quadri Lateral) للكلمة العربية واستخدما اللوغاريتم الجديد للعثور على كل من الجذور ثلاثية ورباعية الأضلاع، وتم التعرف على دقة وكفاءة اللوغاريتم والتعليق على ذلك هو دون إجراء تجارب على استرجاع المعلومات.

٨/٥ وفي عام (١٩٩٤) قام الباحث علي سليمان الصوينع بدراسة تحت عنوان استرجاع المعلومات في اللغة العربية عالج فيها المشكلات الدلالية والتركيبية للغات الطبيعة ومشكلاتها كالتصريف والاشتقاق والمترادفان والألفاظ المشتركة وصيغة الفعل والرمز والمجاز وكلمات التوقف (Stop Words) كما ركزت الدراسة على كشافات التباديل وكيفية الاسترجاع باستخدامها.

٩/٥ وفي العام نفسه (١٩٩٤) قام الخراشي وإيفنز بدراسة «لمقارنة الكلمات والجذوع والجذور كمصطلحات كشفية في نظام استرجاع المعلومات العربي» استخدم فيه نظام (Micro - Airs) (نظام حاسب لاسترجاع المعلومات العربية) تم تصميمه كنظام تجريبي لفحص عمليات التكشيف والاسترجاع للبيانات الببليوجرافية العربية، واستخدم (٢٩) سؤالًا على قاعدة من (٢٥٥) تسجيله ببلوجرافية تغطي علوم الحاسب من بنك معلومات ببليوجرافي في مدينة الملك عبد العزيز للعلوم والتقنية، وأوضحت تلك التجارب أن استخدام جذور الكلمات كمصطلحات تكشيف يعطي أفضل نتائج من استخدام الكلمات كاملة.

١٠/٥ وفي دراسة مشابهة قام أبو سالم (١٩٩٢) ثم أبو سالم وزميلاه ١٩٩٩) بإعادة تجريب عينة الخراشي، ولكنه استخدم في دراسته (١٢٠) عنوانًا في مجال الحاسب الآلي مع مستخلصاتها وبنفس نظام الاسترجاع للخراشي وهو (Micro Computer Based Arabic Information Retrieval system) وأضاف مقارنًا خاصة بالكلمة والساق والجذر باعتبارها مصطلحات تكشيف وأثبتت هذه

التجارب أن الاسترجاع المبني على الجذع (Stem) متفوق على الاسترجاع المبني على الكلمات، كما أن الاسترجاع المبني على الجذر (Root) أفضل من الاسترجاع المبني على الكلمات، وهو أفضل أيضًا من الاسترجاع المبني على الجذع عند المستويات العليا للاسترجاع (Higher Recall level).

وعلى العكس من تجارب كل من الخراشي وأبو سالم فقد أثبتت البحوث التي جاءت في عصر التريك (TREC 10) حيث جاء فيها ما يلي باللغة الإنجليزية:

TREC10, the Stem – Based Approach Yielded better results (Abu El – Khair (Aljlayl, 2002 chwodhury, 2004: 2008, p. 514) بتطوير ووضع اثنان من نظام التجذير (Two New Stemmers).

١١/٥ أما الباحث حميدي وزملائه (Hmeidi et al 1997) فقد قاموا بمقارنة التكشيف الآلي بالتكشيف اليدوي باستخدام الكلمات والجذور والجذوع (Words, Roots and Stems) واتخذوا في تلك المقارنة مجموعة اختبار مكونة من عدد (٢٤٢) مستخلص وعدد (٦٠) سؤال وانتهى البحث إلى أن التكشيف الآلي يؤدي إلى نتيجة أفضل من التكشيف اليدوي عند استخدام الكلمات كمصطلحات الكشاف، وعند استخدام الجذوع والجذور كمصطلحات تكشيف فهي أفضل فقط في التكشيف اليدوي عند المستويات الحالية للاستدعاء (Recall) أي أعلى من (٠,٣)، (٠,٥) بالتتابع، ومعنى ذلك أن التكشيف اليدوي باستخدام الجذور (Roots) كمصطلحات تكشيف يؤدي إلى نتائج أفضل من استخدام الكلمات والجذوع (Stems).

١٢ /٥ وفي نهاية سنوات القرن العشرين قام كل من خوجا وجارسيد (Khoja and Garside, 1999) بإدخال لوغاريتم جديد يستخلص الجذور من الكلمات العربية ويستخدم هذا اللوغاريتم قوائم النماذج والجذور العربية وبعد إزالة اللواصق واللواحق في بداية الكلمة وفي نهايتها يتم مقارنة اللوغاريتم بالنسبة للجذوع (Stems) والنماذج الباقية.

سادساً: بحوث استرجاع المعلومات العربية منذ بداية القرن الحادي والعشرين وحتى عام ٢٠١٣:

تتميز هذه الفترة باستخدام التحليل الصرفي مع الطرق الدلالية (Semantic).

١/٦ فقد قام مساعد صالح الطيار(*) في أطروحته للدكتوراه من جامعة يمونفورت (٢٠٠٠) بإدخال

(*) الباحث مساعد صالح الطيار تخرج من جامعة الإمام محمد بن سعود بالرياض قسم المكتبات والمعلومات وحصل على درجة الماجستير من جامعة لافبرا في بريطانيا والدكتوراه من جامعة دي مونفورت (De Monfort) والأطروحة عنوانها (Arabic Information

لوغاريتم التجذير (Stemming) واستخدم طريقة الصرف الدلالية (Morpho-Semantic Method) أي المعتمدة على الروابط الدلالية للأشكال الصرفية (Semantic Links of the Morphological Form AIRSMA) وقد جاء في خلاصة (Abstract) بحث مساعد الطيار بعض المعلومات الإضافية كما يلي:

تعتبر نظم استرجاع النص العربي (Arabic text Retrieval Systems) ظاهرة جديدة فقد أدخلت هذه النظم موضوعات كالبتر (Truncation) ولوغاريتم الجذوع (Stemming).

وهذا وتستخدم نظم استرجاع المعلومات العربية ثلاث طرق بحثية هي الكلمة والساق أو الجذع (Stem) والجذر (Root) وتستخدم الكلمة (Word) في مضاهاتها كمصطلح أما المصطلحين الآخرين فيدخلان ضمن التحليل الصرفي، ويلاحظ أن لكل طريقة نواقصها فالكلمة (Word) والجذع (Stem) قد يفوتها تسجيلات صالحة (Relevant Records) (للتعبيرات الصرفية) كما يمكن أن يسترجع الجذر (Root) تسجيلات غير صالحة (Irrelevant)، ومن هنا فقد لجأ الطيار إلى طريقة جديدة وهي طريقة الصرف الدلالية (Morpho Semantic Method) المعتمدة على الروابط الدلالية للأشكال الصرفية (Semantic Links of the Morphological Forms) حيث يتم اختيار عينة ممثلة للأشكال الصرفية العربية (الأسماء والصفات) والت تشكل معظم أشكال الكلمات وذلك على أمل تحسين فاعلية الكلمات والجذوع (Words & Stems) في الاسترجاع فضلًا عن تحسين طريقة الجذر (Root) بعدم استبعاد التسجيلات التي يُمكن استرجاعها، وقد استخدم عدد (٥٩٠) تسجيله عربية كقاعدة بيانات واستخدم كل من الاستدعاء (Recall) والدقة (Precision) كمقياس وتقييم مدى تأثير كل من الكلمة والجذع والجذر على الطريقة الدلالية الصرفية، وكانت النتيجة تفوق الطريقة الدلالية أي تحسين أداء الاسترجاع للكلمة والجذع بالنسبة للاستدعاء (Recall) أي استرجاع تسجيلات أكثر، وتحسين الدقة (Precision) أيضًا (أي تسجيلات أقل أي غير صالحة للاسترجاع).

وللتحقق من الأداء ألاسترجاعي، قام الطيار بتجميع عينة من (٥٩٠) تسجيله عربية ثم إدخالها على قاعدة بيانات تجريبية تم إنشاؤها باستخدام لغة البرولوج (Prolog) وكان من نتائج التجربة كما سبقت الإشارة تحقيق الأسلوب الصرفي الدلالي أعلى نسبة استدعاء (٩١%) بينما جاء أسلوب الجذر في المرتبة الثانية (٨٨%) وأسلوب الجذع (٧٩%) وأسلوب الكلمة (٥٤%).

Retrieval System Based on Morphological Analysis 2000) كما يلاحظ أن الباحث نشر بحثًا عام (١٩٩٨) بمجلة مكتبة الملك فهد الوطنية عن كفاءة التحليل الصرفي في استرجاع النصوص العربية.

٦/ ٢ أما دراسة أحمد المالكي (٢٠٠١) لنيل درجة الدكتوراه من جامعة نيومكسيكو فقد تناولت تجربة تحسين استرجاع المعلومات في بيئة أحادية اللغة وذلك بتطبيق التحليل الصرفي للغة العربية فضلًا عن تجربة أخرى لاسترجاع المعلومات اعتمادًا على معاجم للترجمة بين اللغتين العربية والإنجليزية، وقد اختبر النظام بنجاح المقالات الإخبارية في صحيفتي الراية القطرية والوطن الإماراتية عن طريق عشرين سؤال، واستخدم فيها كلمات كاملة (أي بدون المحلل الصرفي) ومجموعة أخرى استخدم فيها المحلل الصرفي ثم أوجز النتائج حول تقييم النظام.

٦/ ٣ دراسة نزار محمد علي قاسم (٢٠٠١) لنيل درجة الدكتوراه من جامعة المستنصرية بالعراق وتناولت إمكانية رفع كفاءة لغة استرجاع المعلومات باستخدام خصائص المفردات العربية ذات العلاقة باختزان المعلومات واسترجاعها، ومقارنتها بخصائص هذه المفردات باللغة الإنجليزية، وقام بفحص عينتين من المفردات المأخوذة من قوائم رؤوس الموضوعات إحداهما عامة والأخرى متخصصة في علم المكتبات والمعلومات، ومن أهم نتائج الدراسة أن اللغة العربية تأتي بصيغ مختلفة للمذكر والمؤنث وكذلك للمفرد والمثنى والجمع في حالات الرفع والنصب والجر مما يجعل كلًّا منها يحمل صفات دلالية متميزة فضلًا عن وجود إمكانيات جيدة للغة العربية يمكن توظيفها لرفع كفاءة الاسترجاع.

٦/ ٤ في عام (٢٠٠٢) قام كل من درويش وأورد (Darwish and Oard) بتطوير محلل صرفي قاما بتسميته سيباويه (Sebawai) واستخدم هذا النظام معجم (ALPNET) لتقدير احتمالات حدوث النماذج وحروف البداية للكلمة اللواصق (Prefixes) وحروف الكلمة في النهاية (Suffixes) والهدف من هذا النظام هو زيادة التغطية عن طريق معجم البناء الآلي (Automatically Constructing Lexicons) ويستخدم النظام قائمة (بالكلمات والجذور) الزوجية والتي يتم استخلاصها آليًّا باستخدام محلل الصرف (ALPNEt)، وقد تم تمرير قائمتين من الكلمات الزوجية على محلل الصرف (ALPNET) وتم بنجاح الحصول على عدد (٢٨٠٠٧٤) من الكلمات الزوجية، واستخدمت هذه الكلمات الزوجية في تقدير احتمالية حدوث اللواصق واللواحق والنماذج، ذلك لأن محلل الصرف يكشف الجذور عن طريق فحص الكلمات وتحديد إمكانية وجود التركيبات التي تحتوي على اللواصق والجذوع واللواحق ثم يتم مقارنة هذه الجذوع بقائمة الجذوع التي تتكون من (١٠,٠٠٠) جذع للتأكد من الجذع الصحيح، وكان سيباويه (Sebawai) ناجحًا في تحليل حوالي (٩٣%) من الكلمات التي قام محلل الصرف (ALPNET) بتحليلها.

ومما سبق يتضح أنه نظرًا لطبيعة اللغة العربية فإن معظم الأعمال المنشورة عن استرجاع المعلومات العربية (AIR) تتعامل مع التحليل الصرفي للغة العربية، إلى أن دخل المسلك العربي (Track) لنظام (TREC 2001) وكانت الأهداف الرئيسية لمختلف النظم السابقة تتركز في استخراج الجذر (Root)

الخاص بالكلمة العربية (Arabic Word)، وقد أثبتت التجارب أنه عند استخدام مجموعات صغيرة فإنها تشير على أن تكشيف الجذور يكون أكثر تأثيرًا من كل من تكشيف الجذوع وتكشيف الكلمات.

بناء جهاز تجذير عربي لاسترجاع المعلومات (Arabic Stemmer):

٥/٦ قام اثنان من الباحثين وهما (**Chen, Aitao and Gey, F. 2002**) بجامعة كاليفورنيا بيركلي وهما من جماعة بيركلي (Tree 2002) بالمشاركة في مسار (CLIR) للاسترجاع عبر اللغة الإنجليزية – العربية، حيث تمت التجربة على مسار عربي وحيد للغة العربية وثلاثة مسارات عبر اللغتين الإنجليزية والعربية، وكان الهدف من الاسترجاع عبر اللغة هو ترجمة الموضوعات الإنجليزية إلى اللغة العربية باستخدام النظم الآلية للترجمة على الخط المباشر من الإنجليزية للعربية، وكنت أسماء الدورات Runs هي (BKY Mon, BKY CL1, BKY CL2, and BKY CL3) والدورة الأولى (BKY Mon) هي للمسار العربي وحيد اللغة والدورات الثلاث الباقية هي عبر اللغتين الإنجليزية – العربية، ويتناول هذا البحث بناء كلمات الاستبعاد العربية (Arabic Stoplist) ونظامين لتجذير العربية وتكون التجارب على الاسترجاع أحادي اللغة العربية وعلى الاسترجاع اللغوي من الإنجليزية إلى العربية، ويلاحظ هنا أن الباحثين قد أخذا الاتجاه الذي يجمع بين ترجمة الأسئلة (Queries) ولغة الوثيقة باستخدام نظم الترجمة بواسطة آلتين (Two Machines) وقد حققت مسيرتهما عن الاسترجاع عبر اللغة نسبة (٨٧,٩٤%) من الأداء ألاسترجاعي الوحيد للغية، ثم قاما بوضع نظامين أحدهما نظام تجذير عربي يعتمد على (M. T Based Arabic Stemmer) وقد حققت مسيرتهما عن الاسترجاع عبر اللغة نسبة (٨٧,٩٤%) من الأداء الاسترجاعي الوحيد للغة، ثم قاما بوضع نظامين أحدهما نظام تجذير عربي يعتمد على(M. T Based Arabic Stemmer)، والنظام الآخر هو التجذير الخفيف (Light Stemmeq) الخاص بمجموعة بيركلي (Berkeley) وكان أداء وعمل الأخير أفضل من مسار التجذير العربي المعتمد على (M. T) أي أن نتائج التجربة تشير إلى أن امتداد السؤال يؤدي بوضوح إلى أداء أفضل للاسترجاع، وأخيرًا فيجب التنويه إلى أن هذا البحث كان برعاية وكالة الدفاع لمشروعات البحوث المتطورة (DARPA) بأمريكا كما تم البحث داخل مكتب تكنولوجيا المعلومات بالوكالة نفسها.

٦/ ٦ دراسة الشهري استخدمت الطرق الإحصائية في اختزال جذوع الكلمات العربية، وتتضمن هذه المداخل استخدام ما يسمى بـ(n-grams)، والذي بموجبه يتم تقسيم الكلمة لعدد من القطع المتداخلة متساوية الحجم في النص لها (n) من الصفات، أي أن هناك قياسات متماثلة تستخدم لتجميع الكلمات المتشابهة اعتمادًا على التشابه في عدد الجرامات (n-grams).

لقد قام الباحث الشهري (Alshehri, 2002) بدراسة الخصائص الإحصائية للكلمات العربية

وتداخلها (n-grams) باستخدام ستة أجسام عربية، وقد أوصى بأن الحجم المثالي (n-grams) للتكشيف واسترجاع النص العربي هو رقم (٣)، ثم قام بمقارنة تأثير استخدام (tri- grams) وخليط من (٣، ٤، ٥) grams كمصطلحات كشفية لمدخل تكشيف الكلمة، وذكر عن تجارب استخدم فيها مجموعتان للاختبار، إحداهما تحتوي على عدد (٢٤٢) مستخلص علمي عربي وكذلك عدد (٦٠) سؤال، والثانية تحتوي على (١٨٧) مقال صحفي كامل من جريدة الراية وعدد (٣٠) سؤال، وأظهر الباحث أن طريقتي التكشيف بواسطة (n-grams) تفوق طريقة تكشيف الكلمات بالنسبة للمجموعة الأولى ولكنها لا تتفوق في المجموعة الثانية.

وكذلك استخدم الباحث لاركي (Larkey et al 2002) المدخل الإحصائي على اختزال الجذع العربي (Arabic Stemming) دون أن يدخل في عمله أي (n-grams)، أي أن هذا المدخل يعتمد على تحليل الوجود المشترك (co-occurrence) للمصطلحات في النص العربي، واستخدم في البداية النص العربي عملية الاختزال بواسطة (Light Stemmer) وانتهى من التجربة إلى أن هذا المدخل الإحصائي يضيف تحسنًا واضحًا على (Light Stemmer) شاملًا (Light 8, Light2) ولكن ذلك التحسن لم يحدث بالنسبة لـ(Khoja Stemmer) كما انتهى إلى التوصية بعدم استخدام (n-grams) في استرجاع النص العربي.

٦/٧ قام قدري وزميله ني (Kadri and Nie, 2006) بمقارنة نظام اختزال الجذوع المعتمد على اللغة (Linguistic Based Stemming) بنظام اختزال الجذع الخفيف (Light Stemming) واعتمد في البداية على القواعد الإحصائية لحل الغموض المتصل بتتابع اللواصق أو اللواحق كما استخدم الباحثان جسد نظام تريك (TREC 2001) لبناء جميع الجذوع الممكنة ومدى تكرار حدوثها في الجسد (Corpus)، وبالنسبة لجذع الكلمة فقد قام الباحثان بتفكيك (De-composed) لجميع الجذوع الممكنة ثم اختيار أكثرها ملائمة اعتمادًا على إحصائياتها في الجسد. وبالنسبة لنظام الجذع الخفيف فقد قاما ببناء جهاز ستيمر (Stemmer) يقوم ببتر (Truncate) أكثر اللواصق واللواحق تكرارًا في نفس الجسد. وقاما ببناء قائمة من (٤١٣) كلمات الوقف (Stop words) مع معايرة النص باستخدام مدخل مشابه لمدخل ألجلالي (Aljlayl, 2002) وعن طريق المقارنة للنظامين باستخدام اختبار تريك ٢٠٠١ وتريك ٢٠٠٢ للمجموعات، فقد تبين أن استخدم نظام اختزال الجذوع المعتمد على اللغة يؤدي إلى نتائج أفضل من نظام الجذوع الخفيف (Light Stemming) وأن النظام الأخير لا يعتبر أفضل المداخل لاسترجاع المعلومات العربية (TREC = Test Retrieval Conference 2001).

٨/٦ في عام ٢٠٠٦ صدرت مقلة إبراهيم أبو الخير عن تأثير استبعاد كلمات التوقيف على نظام

استرجاع المعلومات العربي: دراسة مقارنة والتي نشرت في المجلة الدولية لعلوم الحاسبات والمعلومات. تمثل كلمات التوقيف (Stop Words) حوالي (٨٠%) من الوثائق في المجموعة وهي بلا فائدة لأغراض الاسترجاع (Boaza-yates, p. 167) ويطلق عليها البعض كلمات التشويش (Noise Words) والتي تشتمل على حروف الجر وأدوات التعريف والتنكير وما إليها أي أنها عديمة الدلالة (محمد غنيم ص٢٠٦) ويبدأ الباحث إبراهيم أبو الخير دراسته بأن معظم البحوث في حقل استرجاع المعلومات قد تركزت في اللغة الإنجليزية، ومنذ وقت قريب كانت هناك جهود وأعمال لتطوير نظم استرجاع المعلومات في لغات أخرى غير الإنجليزية فالبحوث والتجارب العربية في هذا المجال ما زالت جديدة بل ومحدودة في جوانب دراستها وتسعى هذه الدراسة لتحسين فاعلية الاسترجاع، حيث يتم فحص تأثيرات وزن المصطلح، وكلمات التوقيف على الاسترجاع العربي ومقارنته باستخدام (Lemur Toolkit) وكان أفضل لوغاريتم للمضاهاة (BM25) مع قائمة كلمات التوقيف العامة، كما أن وضع لوغاريتم للمضاهاة (BM25) يمكن أن يؤدي إلى تحسينات أكثر، لقد جاء (Lemur) بتطبيقات عديدة وقد استخدمت هذه الدراسة نسبة صغيرة فقط لتحديد أدائها باللغة العربية.

٩/٦ كما قدم الباحث درويش وزميله أورد (Darwish and Oard, 2007) دراسة تشمل تكشيف مجموعة التريك (TREC 2001) باستخدام الجذور التي أفرزها سيباويه وأنها يمكن أن تقارن بالتشكيف للكلمات ولكنها أقل رتبة بالنسبة لتكشيف الجذوع وتم ترجيح ذلك بأنه قد يعود لحجم المجموعة الصغير أو عدم الدقة الكافية للمحلل analyser.

وهناك بعض نماذج فقط من محاولات تحسين أداء نظم استرجاع المعلومات العربية عن طريق التحليل الصرفي ولكن التعليق النهائي الذي وضعه الباحث عبد السلام النويسري (Abdusalam F. Ahmed Nwesri, 2008) كان كما يلي (ص ٦٨) يساعد التحليل الصرفي في التمييز بين اللواصق واللواحق في الكلمات العربية ولكن التحليل المعمق للكلمات العربية قد أظهر أن هذه الجهود غير مساعدة بالنسبة لاسترجاع المعلومات العربية كما أن التحليل الصرفي يحتاج إلى قوائم (تكون غير مكتملة غالبًا) من الجذور والجذوع والقواعد التي توضع قبل التجربة لمعظم المحللات من اللواحق واللواصق وفي غياب التشخيص يتم الفشل في استخلاص الجذور، وبالتالي فقد استخدمنا القواعد الصرفية في دعم عملية التجذير واستبعاد اللواصق واللواحق (Stemming).

١٠/٦ هذه أطروحة دكتوراه للباحث عبد السلام أحمد نويسري من جامعة (RMIT) في ملبورن، فيكتوريا، استراليا وأجيزت في مايو ٢٠٠٨ بعنوان (أساليب الاسترجاع المؤثرة للنص العربي مع بعض التعليقات).

لقد اقترح الباحث إنشاء معجم (Lexicon) يعتمد على تحسينات على (Light Stemming)

والتي تميز الحروف المحورية أي القلب (Core) من الحروف (المضافة إلى أول أو آخر الكلمة) واقترح قواعد لاختصار معظم الإضافات (affixes) وأظهر تأثير كل قاعدة على حدة على فعالية الاسترجاع، فضلًا عن استخدام جميع القواعد مرة واحدة، وأثبت الباحث أنه باستخدام مجموعة اختبار (TREC 2001) وباستخدام التغذية المرتدة المتعلقة بالقواعد التي وضعها فإن ذلك يؤدي بشكل مؤكد إلى نتائج أفضل من (Light Stemming).

هذا ويؤكد الباحث نويسري باستخدامه مجموعة ضخمة (١,٥٠٠,٠٠٠) وثيقة (بدلًا من المجموعات الصغيرة ودلالتها خاضعة للشك) كما اختار مجموعة ذات (٩٠) موضوع، وقام باختبار فعالية هذه المجموعة الكبيرة، وأثبت أن مداخل التجذير (Stemming) المعتمدة على المعجم تقدم أداء أفضل من مدخل (Light Stemming) وحده.

هذا ويحتوي النص العربي عادة على الكلمات الأجنبية والتي يتم نقلها من الحروف الأجنبية (Transliterated) إلى الحروف العربية، وقام الباحث بمحاولة التعرف على قضيتين هما إثبات الهوية (Identification) والاسترجاع، وبالتالي فقد تم اختبار تأثير المعاجم (Lexicon) والنماذج العربية و(n-grams) في تمييز الكلمات الأجنبية عن الكلمات العربية الوطنية.

وقام الباحث بإدخال قواعد تساعد على ترشيح (Filter) الكلمات الأجنبية وتحسين مدخل (n-grams) المستخدم في التعريف على اللغة واختار أفضل (n-grams) لبناء الكلمة وسمات اللغة. والنتيجة هي أن الجمع بين (n-grams) والمدخل المعجمي (Lexicon) قد أدى إلى عملية نجاح للتعرف على (٨٠%) من الكلمات الأجنبية وبدقة تصل إلى (٩٣%).

خلاصة ما سبق:

أن المداخل التي شملتها هذه الأطروحة تمثل خطوة هامة نحو تحقيق نص عربي عالي التأثير بالنسبة للاسترجاع، أي أن الباحث قام بمقارنة أداء نظم استرجاع المعلومات العربية الجارية (Current) وأظهر أن (Light 10 stemmer) يقوم بالأداء الأفضل من النظم الأخرى باستثناء (Buck walter stemmer) عند استخدام التغذية المرتدة ذات العلاقة (Relevant Feedback)، وتشير النتائج العملية إلى أن الأساليب الفنية المتبعة تستبعد اللواصق (Prefixes) وبالتالي تساعد على فاعلية الاسترجاع، كما أ، الباحث استخدام ثلاثة أشكال من (Light 10 Stemmer) وهي (Light 13, Light 12, Light 11) وذلك للتأكد من استبعاد اللواحق (في آخر الكلمات) (Suffixes).

١١/٦ اتجاهات الباحثين العرب نحو البحث في الويب باستخدام آلية الشرح Paraphrasing ملخص الدراسة (الدايل، عبير ومراد) (٢٠١٣) وتتناول هذه الدراسة سلوك المستفيدين العرب للويب واستخدامهم محركات البحث والحاجة إلى شرح الأسئلة وتحديدها على اعتبار أن هذه أداة هامة

لاسترجاع المعلومات العربية، ونحن في حاجة ماسة للتعرف على كيفية تعامل الباحثين العرب مع محركات البحث وما هي احتياجاتهم من المعلومات لتحسين استرجاع المعلومات العربية، لقد استجاب المئات من المستفيدين العرب لهذا المسح على الخط المباشر وقد عبر هؤلاء عن تعامل محركات البحث المستخدمة (سواء العربية أو المتعددة اللغات) والعوامل التي تؤثر على استخدامهم لمحركات البحث، وأظهرت النتائج إنه عند استخدام الويب للبحث عن معلومات باللغة العربية، فيميل المستفيدون إلى استخدام الأسئلة الشارحة Use Query Paraphrasing أثناء البحث، وإن معظمهم يستخدم محركات البحث المتعددة اللغات وذلك بدلًا من استخدام محركات البحث المخصصة للعربية وأن هذه الدراسة هي الدراسة –كما يقول صاحبها– الأولى عن سلوك الباحثين العرب للويب واحتياجاتهم البحثية، ويذهب القائمون على هذه الدراسة بأن نتائجها ستقدم للباحثين في المجال بعض الضوء لمتابعة البحوث العربية للويب وتحسينها.

لقد وصل عدد استجابات المستفيدين العرب (٤٦٣) وتم تجميع الإجابات على الاستبيان على الخط المباشر، وقد تبين أن المستفيدين ما زالوا يعانون من بعض الصعوبات في وضع الأسئلة السليمة أي شرح هذه الأسئلة والتعبير الأفضل عن احتياجاتهم المعلوماتية، وثبت أن معظم المستفيدين يستخدمون محركات البحث مثل جوجل وياهو أكثر من استخدامهم لمحركات البحث العربية لاسترجاع المحتوى العربي، وإن استخدامهم الأكثر هو لمحركات البحث المتعددة اللغات... وتعكس نتائج هذه الدراسة الحاجة لتحسين إمكانيات استرجاع المعلومات من الوثائق العربية على الإنترنت.

ملخص نتائج الاستبيان:

تظهر النتائج أن ٧٣% من المستفيدين للويب العربي قد استخدموا محركات البحث يومياً وهذا يثبت أهمية محركات البحث كأداة لاسترجاع المعلومات ولكن معظم المستخدمين للويب العربي لم يكونوا راضيين عن محركات البحث العربية وبالتالي فإن ٨٨% من المستفيدين يميلون لاستخدام محركات البحث متعددة اللغات للبحث عن المحتوى العربي، حيث بلغت نسبة المستخدمين للمحركات العربية ١٢% فقط.

١٢/٦ وختامًا فقد شهد عام ٢٠١٣ تطورًا واضحًا في مجال استرجاع المعلومات عن طريق بحوث أكثر تعقيدًا ومن بينها دراسة محمد أوطير وغسان كنعان ورائد كنعان (Qtair, M. 2013) بعنوان التحول المثالي للسؤال العربي باستخدام أساليب توسيع السؤال الشامل: والثلاثة حاصلين على درجة الدكتوراه ويعملون بجامعة عمان العربية بالأردن.

فموضوع توسيع السؤال Query Expansion هو موضوع معقد إذ يعرفه بوزاييتس (BaezaYates et al 1999) بأنه عملية إضافة مصطلحات جديدة لسؤال المستفيد كمحاولة

لتقديم متن أكثر إفادة للمستفيد (ص٤٤٩) وبالتالي فهو توسيع للسؤال اعتمادًا على التماثل ألمكنزي Similarity thesaurus وهو توسيع للسؤال اعتمادًا على المكنز الإحصائي (ص١٣٤) وهو توسيع للسؤال وإعادة وزن المصطلح لخدمة نموذج الاتجاهات (ص١٢٨) Vector Model وهو توسيع للسؤال عن طريق تحليل المتن المحلي (ص١٢٩) ويذهب الباحثون الثلاثة في دراستهم إلى أن متوسط عدد المصطلحات التي يضيفها المستفيد تكون عادة مصطلحين أو ثلاثة، وهذه تؤدي غالبًا إلى مشكلات متعددة، وللتغلب على هذه المشكلات تستخدم أساليب متعددة لتوسيع السؤال (كما سبقت الإشارة) ومع ذلك لا نستطيع أن نجزم بأن واحدًا من هذه الأسئلة هو الحل المثالي خاصة في اللغة العربية نظرًا لتعقد تركيبها المورفولوجي، وبالتالي فقد لجأت هذه الدراسة إلى استخدام مزيج Combination شامل من أساليب التوسيع والتي يمكن استخدامها لتحسين عملية التوسيع والحصول على أكثر الوثائق صلاحية لسؤال المستفيد العربي، وتشير الدراسة في نتائجها إلى أن هذا النظام المستخدم قد حسن من كل من الاستدعاء والدقة، وبالتالي فإن هذا المدخل في البحث قد أفاد من كل من أساليب توسيع السؤال الآلية والتفاعلية نظرًا لأن السؤال قد تم توسيعه آليًا وتم إدخال المستفيد ضمنيًا في توسيع السؤال (والملاحظ أن مراجع هذه الدراسة قد وصلت إلى ٤٧ مرجعًا معظمها حديثة ومعاصرة).

النتائج والبحوث المستقبلية:

١- خصائص اللغة العربية ومدى تطويع الآلات ومحركات البحث للملاءمة معها:

أشكالها الفصحى والعامية – الكناية من اليمين لليسار – لها خواص مميزة بالأصوات السامية والقواعد النحوية والقواعد الدلالية – صعوبة معالجتها في محركات البحث تحمل حوالي خمسة مليون كلمة مشتقة من ١١,٣٥٠ جذر وهذه تمثل للاسترجاع بالمقارنة باللغة الإنجليزية التي تحتوي على حوالي ١,٣ مليون كلمة من بينها ٤٠٠,٠٠٠ كلمة مفتاحية، في الوقت الذي تحتل اللغة الإنجليزية حوالي ٦٧% من شبكة الإنترنت ولكن المحتوى العربي حوالي ١%، ومحركات البحث العربية ما زالت في وضع ضعيف ولا بد أن يشهد المستقبل محركات بحث عربية متطورة مثل جوجل.

٢- بعض التطورات التي دخلت على بحوث الاسترجاع بصفة عامة:

هناك محاولات عديدة للوصول إلى مرحلة المضاهاة Matching بين أسئلة المستفيد والوثائق وكيفية الملائمة بينهما للوصول إلى أعلى استدعاء (Recall) مع تقليل الوثائق غير الصالحة، ودخلت الرياضيات وعمليات التجذير Stemming والترشيح Filtering وتجهيز الوثائق Processing وامتدادات الأسئلة Expansion of Queries وفي اللغة العربية الملائمة في استخدام الكلمة والجذر والجذع للاسترجاع الأفضل.

وقبل دخول تريك (TREC) عام ٢٠٠٠ كانت الجذور تحقق أفضل استرجاع عن الكلمة والجذع (Stem) ويمكن أن يكون ذلك بسبب صغر حجم مجموعة البحث في مختلف البحوث العربية، ولكن بعد دخول تريك (TREC) أمكن للجذوع أن تحقق أفضل النتائج على الجذور والكلمات فضلًا عن ضرورة أن تشهد بحوث المستقبل لوغاريتمات وزن المصطلح لا سيما مع نماذج اللغة والاحتمالات في اللغة العربية فضلًا عن ضرورة الوصول إلى معايير لوغاريتم التجذير Stemming، بالإضافة إلى كلمات التوقيف Stop Words وقوائمها إلى جانب التحول من المجموعات الصغيرة إلى المجموعات الضخمة للوصول إلى نتائج مرضية.

٣- الدراسات المبكرة في المجال اللغوي والإحصائي عن المفردات الشائعة في اللغة العربية:

حيث تمت تجارب باستخدام الحاسب لبناء كشافات الكلمات المفتاحية في السياق وثبتت صلاحية اللغة العربية مع الحاسبات الآلية بل ثبتت كفاءة اللغة العربية في تكشيف واسترجاع الوثائق العربية وهو مشروع تبنته مدينة الملك عبد العزيز للعلوم والتقنية بالسعودية وغيرها من الدراسات الآلية أيضًا على الكلمات والجذور والجذوع والتعرف على أهم المشكلات الدلالية والتركيبية للغة العربية والتصريف والاشتقاق والمترادفات والألفاظ المشتركة... إلخ.

٤- دخلت بحوث القرن الحادي والعشرين في مجالات جديدة مهمة مثل استخدم طريقة:

الصرف الدلالية وتجربة استرجاع المعلومات اعتمادًا على معاجم اللغتين العربية والإنجليزية، فضلًا عن تطوير المحلل الصرفي حتى دخول المسلك العربي لنظام تريك Arabic Track into TREC2001 والمقارنة مع المسارات الإنجليزية، بالإضافة إلى استخدام الطرق الإحصائية في اختزال جذوع الكلمة العربية وتداخلها مع n-grams وأخيرًا بحوث (أساليب الاسترجاع المؤثرة للنص العربي وإنشاء معجم Lexicon الذي يعتمد على تحسينات Light Stemming.

٥- **لقد تبين باستقراء مراجع هذه الدراسة أن هناك ثلاث عشرة أطروحة للدكتوراه حصل عليها أصحابها العرب** من جامعات القاهرة (٢) والمستنصرية (١) ثم الجامعات الأمريكية: بتسبرج (٤) إلينوى (٣) ونيومكسيكو (١) فضلًا عن لوفبرا بإنجلترا (١) واستراليا (١) والرؤية المستقبلية أن يشهد المجال مزيدًا من دراسات الدكتوراه الماجستير في الجامعات العربية في مجال نظم استرجاع المعلومات العربية.

المراجع العربية:

١- أحمد أنور بدر (يوليو ٢٠٠٩) تطور الإنتاج الفكري لمجال استرجاع المعلومات وواقع تدريسه في بعض أقسام المكتبات والمعلومات المصرية والسعودية: دراسة توثيقية ميدانية الاتجاهات الحديثة في المكتبات والمعلومات، مج ١٦، ع ٣٢ ص١١ – ٥٤.

٢- أحمد أنور بدر (٢٠٠٠) أساسيات استرجاع المعلومات: ص٥٧ – ٧٨ في كتابه تكنولوجيا المعلومات وأساسيات استرجاع المعلومات – الإسكندرية: دار الثقافة العلمية، ١٠٠ ص.

٣- أحمد أنور بدر ومحمد فتحي عبد الهادي ونايمان متولي (٢٠٠١) التكشيف والاستخلاص: دراسات في التحليل الموضوعي – القاهرة، دار قباء، ٤١٦ ص.

٤- أحمد أنور بدر، زهير سليمان مالكي ونايمان إسماعيل متولي (٢٠١٣) مقدمة في علم المعلومات: النشأة والتطور حتى العصر الرقمي – القاهرة: المكتبة الأكاديمية، ٢٩٦ ص.

٥- حشمت قاسم (١٩٨٥) كشافات الكلمات المفتاحية في السياق واحتمالاته في اللغة العربية عالم الكتب، مج ٥، ع٤، ص٦٣٨ – ٦٥٠.

٦- داود عطية عبده (١٩٧٩) المفردات الشائعة في اللغة العربية. الرياض: جامعة الرياض، معهد اللغة العربية، ٣٤٣ص.

٧- علي سليمان الصوينع (١٩٩٤) استرجاع المعلومات في اللغة العربية. الرياض: مكتبة الملك فهد الوطنية، ١٧٦ ص.

٨- فتحي عثمان أبو النجا (١٩٨٥) وضع نظام عربي لاسترجاع المعلومات في قطاع المعلومات الزراعية (دكتوراه جامعة القاهرة).

٩- كنت، ألن (١٩٧١) ثورة المعلومات: استخدام الحاسبات الإلكترونية في اختزان المعلومات واسترجاعها: ترجمة حشمت قاسم وشوقي سالم، مراجعة أحمد بدر، الكويت، وكالة المطبوعات، ٩٨٤ ص.

١٠- محمد سالم غنيم (٢٠٠٨) نظم استرجاع المعلومات العربية: مظاهر الغموض وآفاق الحلول: الرياض: مكتبة الملك فهد الوطنية، ٤٧٤ ص (أطروحة دكتوراه) القاهرة.

١١- محمد بن عبد الله الأطرم (١٩٩٠) كفاءة اللغة العربية في تكشيف واسترجاع الوثائق العربية التقرير النهائي الرياض: مدينة الملك عبد العزيز للعلوم والتقنية.

١٢- مساعد بن صالح الطيار (١٩٩٨) كفاءة التحليل الصرفي في استرجاع النصوص العربية – مجلة مكتبة الملك فهد الوطنية – مج٤، ع١، ص٧ – ٢٣.

١٣- ناريمان إسماعيل متولي (يناير ٢٠٠٤) الإنترنت والأطر البحثية في استرجاع المعلومات. الاتجاهات الحديثة في المكتبات والمعلومات – ع٢١ – ص٥٥ – ٧٧.

١٤- نبيل عبد الرحمن المعثم (نوفمبر ٢٠١١) البحث باللغة العربية على محرك البحث جوجل – مجلة مكتبة الملك فهد الوطنية – مج٧، ع٢.

١٥- نزار محمد علي قاسم (٢٠٠١) إمكانية رفع كفاية استرجاع المعلومات باستخدام خصائص المفردات العربية (أطروحة دكتوراه) الجامعة المستنصرية، قسم المكتبات والمعلومات ص١٨٠.

١٦- نبيل علي (١٩٨٨) اللغة العربية والحاسوب: دراسة بحثية تقديم أسامة الخولي – القاهرة تعريب للنشر، ٥٩١ ص.

المراجع الأجنبية:

- Abdelali, Ahmed; Soliman, Hand}/ (2004) Arab Information Retrieval, Arab Language Processing, v.19.
- Abu El Khair (2006) Effects of stop Words Elimination for AIR: Comparative stud)/ international). of Computing and Information Science v.4 (3) On-line 119-132.
- Abu El Khair, I. (2003). Effectiveness of document processing techniques for Arabic informationretrieval. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh.
- Abu El Khair (2007) Information Retrieval ARIST, v.41 (1) ch.11 p505-531.
- Abu Salem, H. (1992). A microcomputer based Arabic bibliographic information retrieval system with relational thesauri (Arabic-IRS). Unpublished doctoral dissertation, Illinois institute of Technology, Chicago.
- Abu Salem, H,, Al-Omari, M,, & Evens, M. (1999). Stemming methodologies over individual query words for an Arabic information retrieval system, journal of the American Society for information Science, 50,524-529.
- Al Dayal, Abeer and Mourad jykh/ef (2013). Arabic user's attitudes toward web searching using paraphrasing mechanisms journal of Compute Science and information systems, v.2(2), p 34-39.
- Al-Fedaghi, S. and Al-Anzi, F.(1989). A New Algorithm to Generate Arabic Root- Pattern forms. In Proceedings of the 11th National Computer Conference and Exhibition, pp391-400, Dhahran, Saudi Arabia.

- Aljlayl, Mohamed and Frieden, Ophir (2002) On Arabic Search: improving the Retrieval effectiveness via a light stemming Approach CIKM 02 November 4-9, Mclean, Virginia, U.S.A.
- Al-kharashi, l,, & Evens, M. (1994). Comparing words, stems, and roots as index terms in an Arabic information retrieval system, journal of the American Society for information Science, 45(8), 548-560.
- Al-Shalabi, R. and Evans, M. (1998). A Computational Morphology System for Arabic. In Proceedings of the Workshop on Computational Approaches to Semitic Languages. (COLING-ACL '98), pp66-72, Montreal, Quebec, Canada.
- Al-Shehri, Abd-Allah M. (2002). Optimization and effectiveness of N-grams approach for indexing and retrieval in Arabic information retrieval systems. Thesis

(ph.D.) university of Pittsburgh, A. Al Shehri, 2002173p. Pittsburgh.

- Al Tayyar, Musaid Saleh (july 2000). Arabic information Retrieval System based on Morphological Analysis: A comparative study of word, stem, root and morph semantic Methods. Ph.D. in computer science in the department of computer and information science, De Mont fort university, U.S.A.
- Attia, M. (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modeling Finite State Networks. In Proceedings of the Challenge of Arabic for NLP/MT Conference. The British Computer Society, London.
- Baeza-Yates and Ribeiro-Neto (1999) Modern information retrieval pearson: Addison Wesley.

- Beesley, Ken. (1998). Arabic morphological analysis on the internet. In proceedings of the 6th international conference and exhibition on Multi-lingual Computing, Cambridge..
- Chen, Atitao and Gey, F. (2002) Building an Arabic stemmer for information Retrieval In: Proceeding of TREC 2000.
- Chowdhury, G. S. (2004). Introduction to Modern Retrieval. Facet Publishing, London.
- Darwish, K. (2002a). Al-stem: A light Arabic Stemmer. Retrieval june 20, 2005, from www.glue.umd.edu/~kareem/research.
- Darwish, K. (2002b, July). Building a shallow Arabic morphological analyzer in one day. Proceeding of the ACL-02 Workshop on Computational Approaches to Semitic Languages, University of Pennsylvania, Pennsylvania, PA, USA. Retrieval January 26, 2006, from the http://acl.ldc.upenn.edu/w/w02-0506.pdf..
- Darwish, K, & Oard, D,, (2002). CLIR experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval. The Eleventh Text Retrieval Conference (TREC 2002), 703-710. Retrieved January 25, 2006, from http://trec.nist.gov/pubs/trec11/papers/umd.darwish.pdf.
- Harman, Donna et al (2004). The NRRC reliable information access (RLA) workshop (SIGIR) pp528-529.
- Hemeidi, L, Kanaan, G., & Evens, M. (1997). Design and implementation of automatic indexing for information retrieval with Arabic documents, lournal of the American Society for Information Science, 48,867-881.

- Hersh, William et al (2006) TREC: Genomics Track Overview In: Woorkees & Buchlard (2006.
- Ibrahim, Farid M.S. (1988). A syntactically based preprocessor for a limited experimental Arabic document retrieval system. Loughborough: F. Ibrahim, 1988. Thesis (Ph.D.) Loughborough Univ. pfTechnology.
- Kadri, Y. and Nie,). Y. (2006). Effective Stemming for Arabic Information Retrieval, the Challenge of Arabic for NLP/MT. In Proceedings of the International Conference ar the British Computer Scociety, pp68-74, London, UK.
- Kando, Noriko and Evans, David. (2007). Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access. National Institute of Informatics, Tokoyo,)apan.
- Kasem, H. (1978). Arabic in specialist information systems, Unpublished doctoral dissertation, University of London.
- Kent, Allem (1963). Information Analysis and Retrieval Cleveland: Case-Western Reserve University.
- Khoja, S. (2001). Khoja's Arabic stemmer (version 1.0). London: Khoja.
- Khoja, S. & Garside, R. (1999). Stemming Arabic text. Retrieval from www.comp.lancs.ac.uk/computing/users/khoia/stemmer.ps
- Korfhage, R. R. (1997). Information storage and retrieval. New York: Wiley.

- Lancaster, Wilfind and Warner, Awg (1993). Information Retrieval Today:

revised, retitled and expanded edition of information retrieval systems, Arlington, Virginia. Information Resources Press. ترجمة حشمت قاسم إلى العربية وصدر في كتاب عام ١٩٩٧ عن مكتبة الملك فهد الوطنية في الرياض السعودية

- Larkey, L. et al. (2007). Light-Stemming for Arabic Information Retrieval. V38 of Text, Speech and Language Technology, pp221-243.Springer, Netherlands.

- Larkey, L. S., Allan,)., Connel, M. E., Bolivar, A,, & Wade, C. (2003). UMass at

TREC2002: Cross language and novelty tracks. The Eleventh Text Retrieval Conference (TREC2002), 721-732. Retrieved from

http://trec.nist.gov/pubs/trec11/papers/umass.wade.pdf

- Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. SIGIR

2002: proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 275-282.

- Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002). Arabic Information

Retrieval at UMass in TREC-10. The Tenth Text Retrieval Conference (TREC2001), 562-570. Retrieved January 25, 2006, from

http://trec.nist.gov/pubs/trec10/papers/UMass TREC10 Final.pdf

- Malki, Ahmed. (2001). Arabic Interactive Cross-Language Information Retrieval via Natural Language Processing. New Mexico. Thesis (PhD)- New Mexico State University, 168p.
- Mandl, Thomas. (2006). Recent Developments in the Evaluation of Information Retrieval Systems. The European journal for the Informatics Professional, v9, n5, pp44-

55.

- Manning, Christopher, D.; prebhaker Regharan and Hinrich Schutze (2008). Introduction to Information Retrieval. Cambridge: University Press.
- McNamee, Paul; Mayfeld,)ames. (2004). Character N-Gram Tokenization for European Language Text Retrieval. In: Information Retrieval, v7, n1,2, pp73-98.
- Moukdad, Haider and Large, A. (2001). Information Retrieval from Full-Text Arabic Databases: Can Search Engines Designed for English do the Job?. Libri, v51, pp63-74.
- Moukdad, Haider. (2004). Stemming and Root-based approaches to the retrieval of Arabic Documents on the Web. Webology, v3, n1.
- Nwesri, Abdusalam F. Ahmed (May 2008) Effective Retrieval Techniques for Arabic Text. Doctor of Philosophy, School of Computer Science and Information Technology. RMIT University, Melbourne, Victoria, Australia
- Otair, Mohammed et al (2013). Optimizing an Arabic Query using Comprehensive Query Expansion Techniques. International journal of Computer Applications, v. 71 (17), 42-49.

- Raghaven, V. V. et al. (1989). A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. ACM Transactions on Office and Information Systems, v7, n3, pp205-229.
- Tague-Sutcliffe,). (1992). Measuring the Informativness of Retrieval Process. In Proc. Of the 15th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval, pp23-36, Copenhagen, Denmark.
- Turpin, Andew et al. (2006). User Performance versus Precision Measures for Simple Search Tasks. SIGIR Conference on Research and Development, Washington, USA, pp11 -18.
- Voorhees, Ellen and Herman, D. (2005). TREC (Experiment and Evaluation in information retrieval) MIT Press.
- Voorhees, Ellen (2006). Overview of TREC 2006 In: Voohees and Bucklard (2006).
- Zajac, Remi, Malki, Ahmed, Abdelali, Ahmed, Cowie, James, Ogden William C. (2001). Arabic-English NLP at CRL, Proceedings of the Arabic NLP Workshop ACL/EACL 2001.

Dr. abderahmen dabour

Abstract:

This study aims to explore the characteristics of the Arabic language and its efficiency in information retrieval and information retrieval models. Also, It explores recent developments in the evaluation of retrieval systems of Arabic information, measurements of information retrieval research, and the beginnings of the Arabic information retrieval research, and the development of information retrieval research in twenty one century. The study compares between Arab and foreign research, and attitudes of Arabic beneficiaries toward using web in information retrieval and future research results.

As the progress in knowledge became the most accurate method to judge the development in societies, which relies adequately on the progress of the various institutions involved in knowledge, including scientific research institutions, educational institutions and information in various patterns, and involved in the production of information, which has become one of the most prominent elements of knowledge and society's economy, as suggested by the experienced countries such as China, Japan, India, Singapore, Malaysia and Korea, who were able to catch up by implementing foreign knowledge as foundation for their societal development, this research found imperative to introduce the following concepts:

- Information policy, objectives and values that needed to be achieved.
- Information society and Information societal sectors.
- The characteristics of knowledge societies, shifting factors, and their differences with information societies
- knowledge society's requirements and formational stages.
- Information policy of the Chinese development as strategic model that can be applied in Egypt and in the Arab countries (due to the similarity of the two countries in most human development obstacles such as inflation and economic problems, etc.].

•The views and the perceptions of Egyptian experts regarding the anticipating level of success when implementing the Chinese IT policy in Egypt.

- The barriers that blocking the transformation of the Egyptian society to become a knowledge society.
- The recommendations of the Arab experts regarding the requirements that needed to transform the information societies to knowledge societies in the Arab World.