Cybrarians Journal
ISSN 1887-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات

العدد 29، **سبتمبر** 2012

التحديات التي تواجه خوارزميات محركات البحث في استرجاع المحتوى العربي على الشبكة العنكبوتية العالمية: دراسة مسحية تحليلية. 1

إعداد

مؤمن سيد النشرتي

مدرس مساعد، قسم المكتبات والوثائق والمعلومات

جامعة القاهرة، مصر

المستخلص

ترصد هذة الورقة البحثية جانبا مهما في إدارة ومعالجة المحتوى العربي على الانترنت، وهو قضية البحث والاسترجاع لهذا المحتوى، حيث تركز على التحديات التي تواجه خوارزميات محركات البحث الداعمة للغة العربية في استرجاع المحتوى العربي على الانترنت، وذلك في ضوء مجموعة من التساؤلات والتي تحاول الكشف عن:

1- التحديات التي تواجه طرق واليات محركات البحث في الكشف والوصول إلى المحتوى العربي على الانترنت.

2- التحديات التي تواجه منهجيات تكشيف المحتوى العربي داخل محركات البحث.

3- التحديات التي تواجه خوارزميات ونماذج الاسترجاع والترتيب للمحتوى العربي في نتائج محركات البحث.

4- التحديات التي تواجه المستفيدين في صياغة الاستفسارات عن المحتوى العربي على الانترنت.

وفي هذا تعتمد الدراسة على المنهج المسحي لحصر غالبية الخوارزميات والآليات التي تعتمد عليها محركات البحث في استرجاع المحتوى، ثم الاعتماد على النهج التحليلي لدراسة التحديات التي تواجه هذة الخوارزميات.

Cybrarians Journal
ISSN 1887-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 29، سبتمبر 2012

الاستشهاد المرجعي

النشرتي، مؤمن سيد. التحديات التي تواجه خوارزميات محركات البحث في استرجاع المحتوى العربي على الشبكة العنكبوتية العالمية: دراسة مسحية تحليلية .- Cybrarians Journal .- ع 29 (سبتمبر 2012) .- تاريخ الاطلاع <أكتب هنا تاريخ الاطلاع على البحث> .- متاح في: <أكتب هنا تاريخ الاطلاع على البحث>

The Challenges that Facing the Search Engine's Algorithms in Retrieving the Arabic Content on The World Wide Web:
Analytical study

Abstract

This study explores an important side of managing and handling the arabic content on the web, search on arabic content and retrieve issues. this study focuses on the challenges that facing algorithms of search engines that supported arabic language in retrieve the arabic content.

This study tries answering about a set of Queries that related to:

1- What 's the challenges that facing the search engine's crawler?

2- What's the challenges that facing the search engine's indexer?

3- What's the challenges that facing the search engine's ranking?

the study will depend on the analytical methodology to reach final results.

توطيد:

أصبح المسعى الرئيسي لكل دولة أو قومية في وقتنا الراهن أن تحظى بمكانة مرموقة واصيلة لها في الفضاء المعلوماتي، فقد اصبح التنافس الحقيقي بين القوميات هو ما تنفرد به كل دولة عن غيرها من قدرات على انتاج واستثمار ما يتوافر لها من محتوى وتطبيقات وآليات لتعلن بها عن نفسها ولتحظى بهوية فريدة في مجتمعات المعرفة، وعلى هذا نرى أن كثيرا من الآمم قد ادلت بدلوها في مجتمعات المعرفة وقد انحصرت هذه المساهمات بين طابعين الطابع الأول وهو اسهام الدول في إنتاج واستثمار المحتوى ويأتي على رأس هذة الدول دول العالم المتقدم بما توفره من أصالة وحداثة في المحتوى، أما الطابع الثاني فهو قدرة الدول على انتاج واستثمار الاليات والتقنيات لمعالجة هذا المحتوى كدول العالم المتطور كالهند والتشيلي وغيرها، ولكن يظل انتاج المحتوى هو الملك والهوية الفريدة التي تكفل لمنشئه الريادة والأصالة وأن من يتخلف عن هذا الركب بطابعيه سيهوى به في غيابات الفجوة الرقمية، وعلى هذا كانت مبادرات العاهل السعودي خادم الحرمين الشريفين الملك عبد الله بن عبد العزيز – حفظه الله - في التشجيع على اثراء المحتوى العربي على الانترنت بمختلف انواعه وأشكاله سبقا تصطف به الأمة العربية بأسرها في الصفوف الأولى في مجتمعات المعرفة فتحية لهذا الجهد العظيم وسدد الله خطاه نحو رقي هذة الأمة.

مشكلة الدراسة:

تأتي قضية تجهيز ومعالجة المحتوى على الشبكة العنكبوتية العالمية The 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 اكتشافه ونظم ادارته ولكن تبرز قضية هامة على صعيد هذا الأمر وهي قضية استرجاع المحتوى فلا تجود جدوى من وجدو المحتوى ان لم يتم استرجاعه واستثماره من قبل المستفيدين.

وعليه تعد أدوات البحث والاسترجاع وعلى رأسها **محركات البحث** بمثابة حجر الأساس لهذا المحتوى المعلوماتي، وحلقة الوصل بين طرفي النشر والاسترجاع للمحتوى، وتأتي محركات البحث Web Search Engines على رأس أدوات البحث والاسترجاع للمحتوى على العنكبوتية حيث تنفرد بنسبة استخدام تقارب 84% من اجمالي إجراءات البحث عن المحتوى, كما تتصف محركات البحث بإنها أكبر أداة بحث على العنكبوتية حيث تستأصل بأكبر عدد من الإستفسارات، فقد بلغ عدد الاستفسارت الموجه إلى محركات البحث نحو 150 مليون استفسار في اليوم الواحد، فضلا عن كونها أكبر أدوات البحث من حيث حجم تغطيتها للصفحات القابلة للتكشيف حيث تكشف نحو 16% من محتوى العنكبوتية. كما أن 40% ممن يتعاملون مع محتوى العنكبوتية يصلون إلى هذا المحتوى من خلال قوائم نتائج محركات البحث[1] .

ورغم ذلك أيضا لم توفق محركات البحث في تحقيق غايتها فمن حيث حجم التغطية لاتتجاوز كشافات محركات البحث العامة في تغطية المحتوى المتاح على العنكبوتية بنسبة 16%، ليس هذا فحسب بل أن 80% من

[1] Castillo, Carlos. "EffectiveWeb Crawling." Diss. University of Chile, 2004. Web. 12 Oct. 2101. <www.chato.cl/papers/crawling_thesis/effective_web_crawling.pdf>.

المستخدمين لايتجاوزا استخدام أول صفحيتين من نتائج محركات البحث التي تعرض محتوى الشبكة العنكبوتية، ومردود هذا الأمر يعود إلى عدم تحقيق التطابق بين المحتوى المطلوب وبين المحتوى المسترجع من قبل محركات البحث، فضلا عن نسبة التداخل والتكرار في استرجاع المحتوى بين محركات البحث بعضها البعض والتي بلغت نحو 84.9% وبلغت نسبة عدم الرضا من قبل المستفيدين تجاه نتائج محركات البحث التقليدية 44%[2]. وغيرها من الصعوبات والتحديات والتي كانت سببا ودافعا لدراسة التحديات التي تواجه محركات البحث في استرجاع المحتوى المتاخ على العنكبوتية.

أهداف الدراسة:

تسعى الدراسة بشكل مباشر إلى تحقيق الأهداف التالية:

1- التأصيل النظري لبعض التحديات التي تواجه خوارزميات محركات البحث في استرجاع المحتوى العربي.

2- التحليل المنهجي لبعض خوارزميات ونماذج الاسترجاع والترتيب لمحركات البحث في التعامل مع المحتوى العربي.

3- رصد منهجيات واليات محركات البحث في اكتشاف المحتوى العربي المتاح على الانترنت.

4- التعرف على واقع نظم التكشيف الآلية للمحتوى العربي في بنية محركات البحث العالمية.

تساؤلات الدراسة:

تحاول الدراسة الإجابة على مجموعة من التساؤلات المنهجية الآتية:

[2] Asadi, saied & hamied R.jamail."shifts in search engines development: a review of past, present, and future trends in research on search engines"[cited 2010-8-10] available at http://www.webolog.ir

1- ماهي التحديات الخارجية التي تواجه محركات البحث في استرجاع المحتوى العربي على الانترنت.
2- ماهي التحديات الداخلية التي تواجه محركات البحث في استرجاع المحتوى العربي على الانترنت.

منهج الدراسة:

تعتمد هذه الدراسة على المنهج الوصفي التحليلي لوصف ورصد التحديات التي تواجه محركات البحث في استرجاع المحتوى العربي على الانترنت.

الدراسات السابقة:

تناولت العديد من الدراسات قضية المحتوى على العنكبوتية، والتي تباينت فيما بينها حول المنظور او زاوية المعالجة لهذا المحتوى حيث تأتي دراسة *Isil Ozgener and Thomas Dillig*[3] لتركز على قضية نشر المحتوى على العنكبوتية ومدى فاعلية برامج ادارة المحتوى (CMS) في ذلك.

كما قدمت *UNESCO* دراسة حول التنوع اللغوي للمحتوى على الانترنت رصدت من خلالها التوزيع اللغوي للمحتوى على الانترنت ومعدلات الاستخدام وقدمت مجموعة من الإحصائيات حول انتاجية اللغات واسهامها في المحتوى العالمي.

وتأتي دراسة *Tim lee*[4] – مخترع الويب – كأهم الدراسات التي تناولت قضية التحديات التي تواجه استرجاع المحتوى على الانترنت فقد أوضح ان المشكلات والتحديات التي تواجهها محركات البحث في استرجاع المحتوى تكمن في طبيعتها ذاتها، حيث أن هذة المحركات صممت للإجابة على تساؤل واحد " ماهي الوثائق التي تشتمل

[3] Ozgener, Isil. (2005). *Publishing content on the web*. : Stanford university.
[4] Lee, T., & Fischetti, M. (2010). *Weaving the web: the original design and ultimative destiny of the World Wide Web by its inventor* ([Nachdr.] ed.). New York, NY: HarperBusiness.

على الكلمات او الجمل الآتية" دون النظر إلى اعتبارات اخرى كمصداقية وموثوقية المحتوى أو تكامل المعرفي بين المحتوى.

وتاتي دراسة *Ricardo Baeza-Yates*[5] موضحة أن المشكلة الرئيسية في استرجاع المحتوى بشكل في محركات البحث يكمن في قضية التنبوء بتحديد أي من الوثائق تتسم بالصلة لما يمكن أن يقدم من استفسارات وأي منها لايتسم بالصلة.

كذلك اوضح كلا من *PETER BRUSILOVSKY* و*CARLO TASSO*[6] *ان* كافة التحديات التي تواجهها خوارزميات محركات البحث في ادارو واسترجاع المحتوى على الانترنت تدور في فلك عاملين اساسين هما:

- المعالجة اللغوية:

حيث ان غالبية محركات البحث تستند في معالجتها للمحتوى على وجود او غياب الكلمات المفتاحية في النص دون اية محاولة لتحليل المحتوى او تحديد المفاهيم المشار اليها في النص وهو السبب الرئيسي وراء انخفاض الدقة في الاسترجاع فضلا عن الظواهر اللغوية الاخرى كالترادف والتجانس.

- محدودية الآليات والخوارزميات:

وتتجلى هذة المحدودية في التعبير صياغة استفسارات في كلمات قليلة تتراوح في كلمتين ولا تتجاوز الثلاث مما يكفل لمحرك البحث الاجابة السريعة عن الاستفسار (والتي ينظر اليها من قبل البعض على كونها ميزة لها) ولكنها في المقابل تفشل في تحديد وبشكل دقيق ما يريده المستفيد وما لايريده من نتائج.

بينما اوضح *Markov* [7] ان تحديات محركات البحث منبعها يعود إلى المحدودية معالجة الشبكة العنكبوتية لدلالة المحتوى، فصفحات العنكبوتية لا تحمل دلالة لمحتواها ولكن تحمل تنسيقا جيد وتمثيل عظيما للبيانات ، اما

[5] Yates, R., & Neto, B. (1999). *Modern information retrieval* . New York: ACM Press ;.
[6] Peter Brusilovsky , Carlo Tasso, Preface to Special Issue on User Modeling for Web Information Retrieval, User Modeling and User-Adapted Interaction, v.14 n.2-3, p.147-157, June 2004.
[7] Zdravko, Markov & Daniel T. Larose. Data-mining the Web : uncovering patterns in Web content, structure, and usage. John Wiley & Sons, Inc.2007

الروابط فتكاد تنعدم دلالاتها على العنكبوتية والدلالة الوحيدة التي تحملها في اطار العنكبوتي هي ان الموقع هذا يرتبط بالموقع ذاك دون اية تحديد لدلالة الربط او نوع الارتباط.

كما اوضح *G.Madhu* ان مشكلات البنية الاسترجاعية للعنكبوتية تمثل التحديات الاساسية لمحركات البحث والمتمثلة في:

- افتقار العنكبوتية للبنية السليمة لتمثيل المحتوى.
- الغموض المعلوماتي الناتج عن ضعف منهجيات الربط بين المحتوى.
- قابلية الاستخدام وما تعنية من قدرتها في التعامل مع العدد الهائل من المستفيدين والحجم الهائل من المحتوى.
- افتقار عنصر المصداقية والثقة فيما تحمله العنكبوتية من معلومات.
- افتقار اليات وخوارزميات العنكبوتية من الفهم المنطقي لما تعرضه من محتوى[8].

ويرى كلا من *Van Harmelen & Stuckenschmidt* أن التحديات التي تواجهها محركات البحث تكمن في افتقار العنكبوتية إلى النماذج المفاهيمية للمحتوى المعروض وعدم وضوح حدود وملامح الشبكة العنكبوتية في ظل ديناميكيتها المفرطة[9].

التمهيد:

أمست الشبكة العنكبوتية في العقود الأخيرة من القرن العشرين قناة الاتصال والنشر الفعالة على الانترنت على مختلف الأصعدة العلمية والاجتماعية والثقافية، ومصدرًا أساسيًا لزخم متراكم من مصادر المعلومات، كما أمست

[8] G. Madhu, A. Govardhan, T. V. Rajinikanth: Intelligent Semantic Web Search Engines: A Brief Survey CoRR abs/1102.0831: (2011).
[9] Stuckenschmidt, Heiner, and Frank Harmelen. *Information sharing on the semantic Web* . Berlin: Springer, 2005.

أيضا أن تكون المضخة الاساسية للمحتوى بتنوع مضامينه وأشكاله ولغاته - فقد قدر حجم محتوى الشبكة العنكبوتية بنحو 7 ملايين صفحة في اغسطس عام 2000 بعدد مستخدمين لها قدر بـــ 500 مليون مستخدم، بينما بلغ حجم الشبكة في اغسطس 2010 نحو 28.9 مليار صفحة بعدد مستخدمين قدر نحو 1.9 بليون مستخدم[10].

وقد وجد هذا المحتوى طريقه للنشر والأتاحة في بيئة ديناميكية اتسمت بفجوة عظيمة في تحقيق التكاملية بين الآلة ومحتواها وبين تحقيق الرضا بين الانسان ومايسترجع من محتوى فضلا عن عجز وصعوبة تشهده تقنياتها في ملاحقة ومعالجة واسترجاع المحتوى ذات التباين والتنوع الموضوعي واللغوي والنوعي والشكلي والجغرافي.

فرضت البنية المعمارية والإسترجاعية للشبكة العنكبوتية مجموعة من التحديات التي تتعلق بفاعلية استرجاع المحتوى على الانترنت، فلم تصمم العنكبوتية على أن تكون قاعدة بيانات Database يخضع فيها المحتوى للهيكلة والتنظيم المطرد - وما تكفله قواعد البيانات من مخططات للتشارك وارجاءات الابحار ونظما في الاسترجاع، بل صممت العنكبوتية لتتيح من خلالها كل شئ عن أي شئ. مما استتبع في ان تكون اقرب لمقولة *George Meghabghab* بأن العنكبوتية تمثل "الحياة البرية للمحتوى"[11].

ما اتسمت به العنكبوتية في إدارتها للمحتوى عن نظائرها من نظم ادارة المحتوى هو عنصر **الازدواجية**، فرغم كونها بيئة استرجاعية توفر مجموعة من ادوات البحث والاسترجاع، الإ انها تعمل في نفس الوقت كبيئة للنشروالتوزيع والاتاحة الحرة للمحتوى، مما أوجد العديد من التحديات غير المسبوقة على مختلف الاصعدة في التعامل مع المحتوى، هذا الأمر جعل من أمر ضبط المحتوى وتنظيمه أمرا يكاد ان يكون مستحيلا في اكتماله.

ويرى *Ricardo Baeza-Yates* أن ابعاد البحث عن المحتوى واسترجاعه على الشبكة العنكبوتية ينطوي على ثلاثة محاور اساسية:

[10] *The size of the world wide web.* Retrieved 8, 2, 2010, from The size of the world wide web: http://www.worldwidewebsize.com/

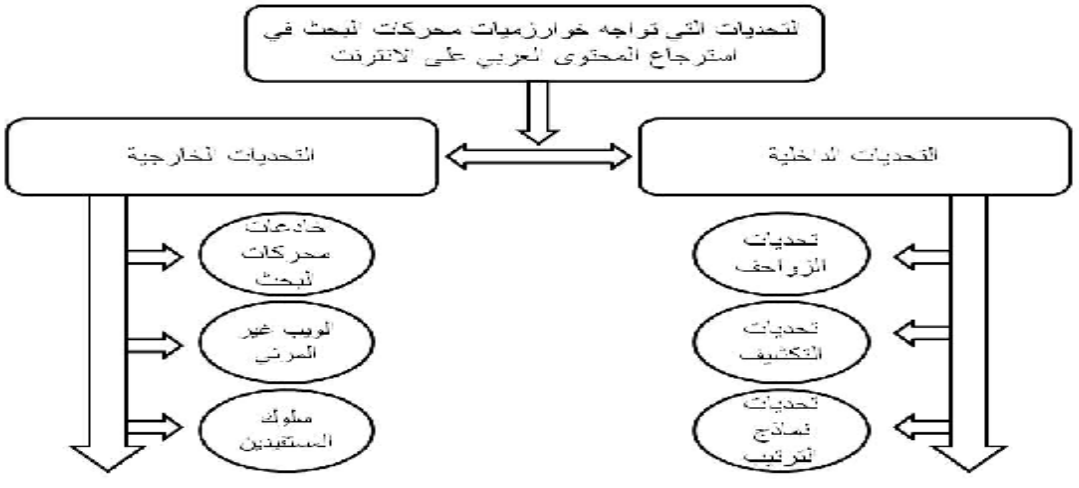
[11] Meghabghab, George, and Abraham Kandel. Search engines, link analysis, and user's web behavior: 74 tables; [a unifying web mining approach]. Berlin: Springer, 2008. Print..

- **المحتوى**: وهو جوهر العنكبوتية ويمثل الهدف الأساس من وجود الاطراف الاحقة.
- **العنكبوتية**: وتمثل بيئة الاسترجاع والبحث والتي تعد اكبر مستودع للبيانات.
- **ادوات البحث**: وهي الاداة التي تمثل حلقة الوصل بين المستفيدين من جانب ومحتوى العنكبوتية من جانب اخر.
- **المستفيدون**: وهم منشئوا المواقع ومستخدموها ويتباينون فيما بينهم في انماط سلوكهم البحثي وفقا لمجموعة من المتغيرات الثقافية والتعليمية وغيرها[12].

وفي هذا، تلقى هذه الورقة البحثية الضوء على التحديات وأوجه القصور التي تواجه خوارزميات محركات البحث في استرجاع المحتوى عامة والمحتوى العربي على وجه الخصوص وعليه تأتي مباحث هذة الدراسة على في النحو الاتي:

[12] Yates, R., & Neto, B. (1999). *Modern information retrieval* . New York: ACM Press ;.

- التحديات الداخلية لمحركات البحث.
  - التحديات التي تواجه الزواحف في اكتشاف المحتوى العربي على الانترنت.
  - التحديات التي تواجه تكشيف المحتوى العربي داخل محركات البحث.
  - التحديات التي تواجه خوارزميات الترتيب والاسترجاع للمحتوى العربي في محركات البحث.
- التحديات الخارجية لمحركات البحث
  - خادعات محركات البحث وتأثيرها على استرجاع المحتوى العربي.
  - العنكبوتية الخفية وما تشمله من محتوى يصعب استرجاعه.
  - سلوك المستفيدين في البحث وتأثيره على استرجاع المحتوى العربي في محركات البحث.



شكل رقم (1) يوضح أنواع التحديات التي تواجه محركات البحث.

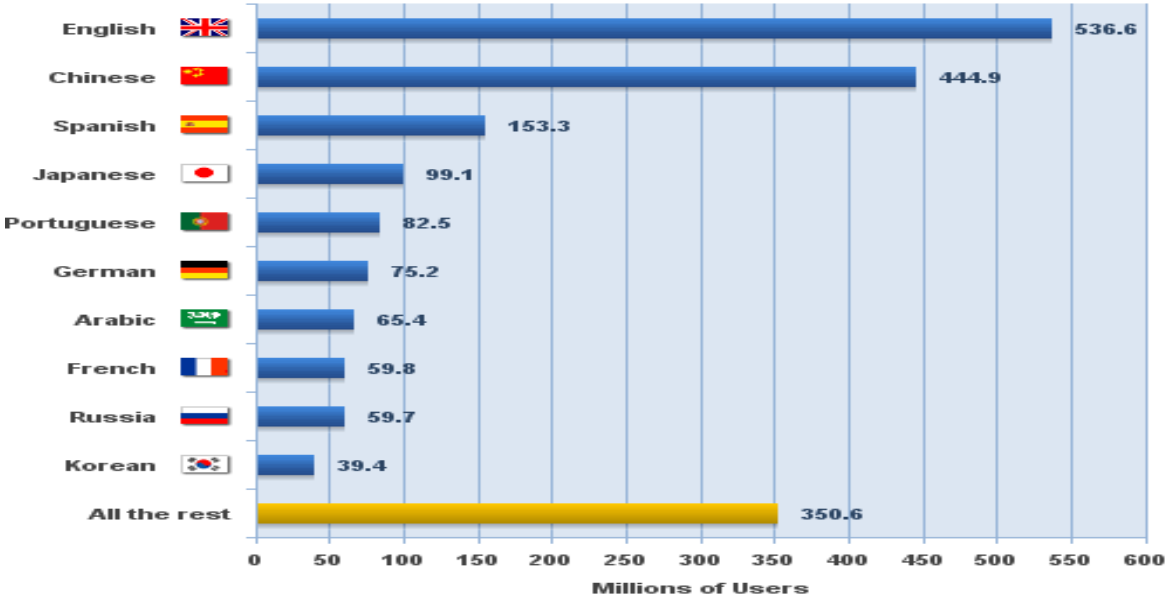
Cybrarians Journal
ISSN 1887-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 29، سبتمبر 2012

أولاً: واقع اللغة العربية في الشبكة العنكبوتية ومحركات البحث:

في عالم الانترنت يعد المحتوى هو الملك The content is a king فبدون محتوى تصبح الشبكة العنكبوتية بتقنياتها وبروتكولاتها وبرامجها كالأنابيب الفارغة دون ماء فيها, وتسهم اللغة بشكل أساسي في تشكيل المحتوى وبشكل خاص هناك مقولة تقول أن من يستطيع أن يسوق لغته يستطيع أن يسوق منتجه في عصر اقتصاد المعرفة، وتعتبر اللغة ومعالجتها آليا أحد أهم عناصر البنى الاساس التي يقوم عليها صناعة المحتوى وتشمل صناعة المحتوى كل ماينتجه النشر الالكتروني من مواقع ووثائق وملفات ذات وسائط متعددة.

أما عن واقع اللغة العربية على الانترنت فيتمثل في حجم المحتوى العربي المتاح على الشبكة العنكبوتية والذي يقدر حجمه بنسبة بلغت 1.4% من حجم المحتوى المتاح على الانترنت، حيث احتلت اللغة العربية المرتبة السابعة من حيث اللغات المستخدمة على الانترنت بواقع استخدام بلغ نسبة 3.3% من اجمالي مستخدمي الانترنت في العالم (كما هو موضح في الشكل رقم 2)، اما عن معدل النفاذ إلى الانترنت في الوطن العربي فقد بلغ نحو 17.4%[13] من اجمالي سكان الوطن العربي بواقع 65.4 مليون نسمة من أجمالي عدد سكان الوطن العربي حتى عام 2010، وتظهر الإحصائيات ايضا أن 65% من المستخدمين العرب للانترنت يعتمدون على اللغة العربية في البحث والتصفح كما هو موضح في الشكل رقم (3).

[13] Internet world stat.http://www. Internetworldsta.com



شكل رقم (2) يوضح ترتيب اللغة العربية من حيث عدد مستخدميها على الانترنت[14]

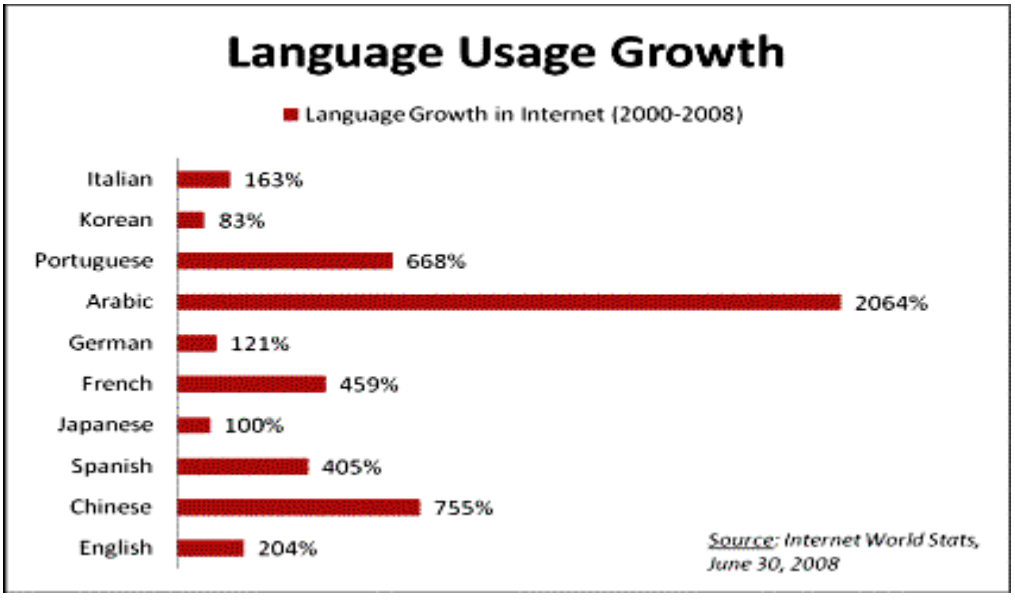
[14] Top Ten Internet Languages - World Internet Statistics. (n.d.). *Internet World Stats - Usage and Population Statistics*. Retrieved July 20, 2011, from http://www.internetworldstats.com/stats7.htm

Arabic Speaking Internet Users					
COUNTRIES	Population (2009 Est.)	Internet Users, Latest Data	Penetration (% Population)	User Growth 2000-2009	% Users in Table
Algeria	34,178,188	4,100,000	12.0 %	8,100.0 %	6.8 %
Bahrain	728,709	402,900	55.3 %	907.3 %	0.7 %
Comoros	752,438	23,000	3.1 %	1,433.3 %	0.0 %
Djibouti	724,622	19,200	2.6 %	1,271.4 %	0.0 %
Egypt	78,866,635	16,636,000	21.1 %	3,596.9 %	27.6 %
Iraq	28,945,569	300,000	1.0 %	2,300.0 %	0.5 %
Jordan	6,269,285	1,595,200	25.4 %	1,153.4 %	2.6 %
Kuwait	2,692,526	1,000,000	37.1 %	566.7 %	1.7 %
Lebanon	4,017,095	945,000	23.5 %	215.0 %	1.6 %
Libya	6,324,357	323,000	5.1 %	3,130.0 %	0.5 %
Mauritania	73,129,486	60,000	1.9 %	1,100.0 %	0.1 %
Morocco	31,285,174	10,442,500	33.4 %	10,342.5 %	17.3 %
Oman	3,418,085	557,000	16.3 %	518.9 %	0.9 %
Qatar	833,285	436,000	52.3 %	1,353.3 %	0.7 %
Saudi Arabia	28,686,633	7,761,800	27.1 %	3,780.3 %	12.9 %
Somalia	9,832,017	102,000	1.0 %	50,900.0 %	0.2 %
Sudan	41,087,825	4,200,000	10.2 %	13,900.0 %	7.0 %
Syria	21,762,978	3,565,000	16.4 %	11,783.3 %	5.9 %
Tunisia	10,486,339	3,500,000	33.4 %	3,400.0 %	5.8 %
United Arab Emirates	4,798,491	3,558,000	74.1 %	384.1 %	5.9 %
Palestine	2,461,267	355,500	14.4 %	915.7 %	0.6 %
Yemen	22,858,238	370,000	1.6 %	2,366.7 %	0.6 %
TOTAL	344,139,242	60,252,100	17.5 %	2,297.7 %	100.0 %

شكل رقم (3) يوضح احصائيات حول استخدام الانترنت في الوطن العربي[15]

[15] http://www.internetworldstats.com/stats19.htm

وفي هذا السياق يجدر الإشارة إلى أن أستخدام اللغة العربية على الانترنت قد تضاعف من عام 2000 – 2008 أكثر من 2000% كما هو موضح في الشكل رقم (4)، ورغم ذلك يوضح اوضح Hermann Havermann انه لا يوجد حتى الان محرك بحث عربي خالص القوام وفقا لمفهوم ورؤية محركات البحث, وماهو متاح في الاسواق من محركات بحث عربية تدعي على انفسها محركات للبحث وهي في الأساس ادلة بحث لا يعتمد البحث فيها على قاعدة بيانات خالصة له[16].



شكل رقم (4) يوضح معدلات نمو استخدام اللغة العربية على الانترنت ما بين عامي 2000 - 2008[17].

تعد اللغة العربية أقدم اللغات على مستوى العالم وسادس لغة معترف بها على مستوى الامم المتحدة وهي أعقد اللغات السامية وأغناها صوتاً وصرفاً ومعجماً وقد أوضح نبيل علي خصائص اللغة العربية من منظور المعالجة الآلية المعلوماتية لها موضحا كينونتها بإنها تمتاز:

1. - التوسط اللغوي.

[16] Andrew Hammond. Arabic search engine may boost content. http://www.abc.net.au
[17] http://www.imakenews.com/lweaver/e_article001189962.cfm?x=bdS7pcv,b6wMC6pH,w

2. - حدة الخاصية الصرفية.
3. - المرونة النحوية.
4. - الانتظام الصوتي.
5. - ظاهرة الإعراب.
6. - الحساسية السياقية.
7. - تعدد طرق الكتابة وغياب عناصر التشكيل.
8. - ثراء المعجم واعتماده على الجذور.
9. - شدة التماسك بين عناصر المنظومة اللغوية.[18]

وأوضح كلا من LARGE ANDREW AND HAIDAR MOUKDAD[19] مجموعة من التحديات التي تواجه خوارزميات محركات البحث في التعامل مع المحتوى باللغة العربية وهي:

1. اشتمال الكلمات العربية على بعض السوابق مثل اداة التعريف ال وعدد اخر من السوابق كحروف الجر والتي لاتاتي بشكل منفصل عن البينية التركيبية للكلمة مما يؤدي إلى ان ترتب هذة الكلمات وفقا لسوابقها في الكشاف.
2. التركيب الصرفي للكلمات في اللغة العربية.
3. احرف العلة في اللغة العربية.
4. رد الكلمة لجذورها او مادتها اللغوية للبحث عنها.
5. مشكلات اختلاف البنية الصرفية للكلمة في حالة التانيث والتذكير والجمع والمثنى.
6. مشكلة الشدة والحرف المزدوج.

[18] نبيل علي. العرب وعصر المعلومات.عالم المعرفة.الكويت:المجلس الوطني للثقافة والفنون والاداب.1994.ص333.
[19] HAIDAR MOUKDAD AND ANDREW. Lost In Cyberspace: How Do Search Engines Handle Arabic Queries?

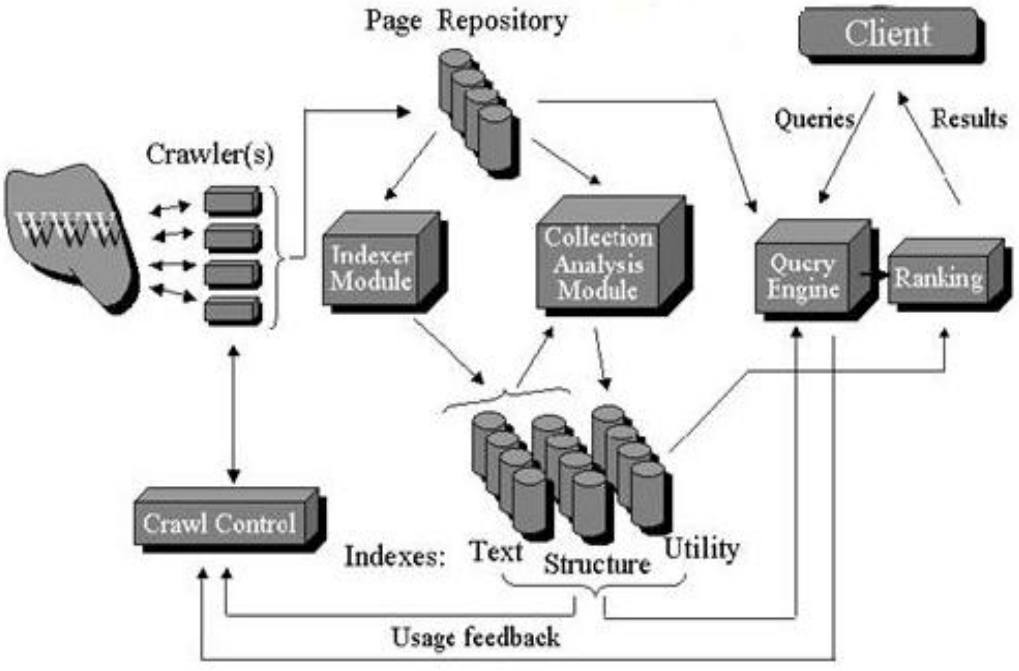
وكذلك أوضحا ايضا[20] أن مشكلات محركات البحث في التعامل مع الاستفسارات الموجهة اليها باللغة العربية يعود إلى قصور خوارزمياتها في التعامل مع اللغة العربية ولهذا تاتي هذة الدراسة في كدراسة حصرية وتحليلية للتحديات التي تواجه خوارزميات محركات البحث في التعامل مع المحتوى باللغة العربية.

[20] HAIDAR MOUKDAD AND ANDREW. Lost In Cyberspace: How Do Search Engines Handle Arabic Queries?

ثانياُ:التحديات الداخلية لمحركات البحث في التعامل مع المحتوى العربي (التحديات النابعة من مشكلات في خوارزمياتها):

1- معمارية محركات البحث:

تلعب معمارية محركات البحث *The Search Engines Architecture* دورا رئيسا في اكتشاف وتكشيف واسترجاع المحتوى على الانترنت ورغم أن محركات البحث قد تختلف فيما بينها من حيث نطاق الشمولية والتغطية والحداثة والتركيز النوعي، الإ أنها تتفق فيما بينها من حيث المعمارية الاساس للتكوين والتشغيل والتي تتمثل في الشكل رقم (5).



شكل رقم (5) يوضح البنية المعمارية الأساس لمحركات البحث.

تشتمل بنية محركات البحث العديد من المكونات والنظم الفرعية والتي تتكامل فيما بينها بهدف تحقيق نسبة استرجاع مرتفعة، وتتمثل الوحدات الفرعية المكونة لبنية محركات البحث في:

- **الزواحف**:*The crawler*:

وهو برنامج يعمل على توفير المحتوى لمحرك البحث حيث يقوم بتصفح العنكبوتية من خلال تتبع الروابط الفائقة بين المواقع ومصادر المعلومات للوصول إلى الصفحات التي تشتمل على المحتوى. ثم استخراج URLs واعطائها إلى وحدة التحكم للزاحف.

- **وحدة التحكم للزاحف** *The Crawler Control Unit*:

تقوم هذه الوحدة بتحديد أي الروابط الفائقة التي سيتم زيارتها مستقبلا وتغذية الزاحف بالخوارزميات الخاصة بعملية الزحف، وبمجرد أن تكتمل عملية الزحف تقوم وحدة التحكم للزاحف بإعلام الكشافات التي تم بنائها مسبقا. حيث يحدد كشاف البنائي *The Structure Index* للزاحف أي الروابط التي يجب أن تستكشف وأي منها يجب أن يتم تجاهلها، كما تعتمد أيضا هذة الوحدة على استخدام التغذية المرتدة *feedback* من انماط الاستخدام للاسترشاد بها فيما بعد في عملية الزحف.

- **وحدة تحليل المجموعات** *The Collection analysis module* :

هي الوحدة المسئولة عن انشاء الكشافات من واقع تحليل الوثائق وتحديد طبيعة البني التكوينية للوثائق.

- **المكشف** *The Indexer* :

تشتمل هذة الوحدة على ثلاثة أنماط من الكشافات.

1- **كشافات النص** *The text index*: والذي يشتمل على الكلمات المفتاحية والعناوين والجمل الدلالية الواردة في محتوى الوثيقة المكشفة، وتحفظ في ملف يعرف بالملف المقلوب *Inverted file*. حيث يعمل على استخراج كافة الكلمات من كافة الصفحات، وتسجيل محددات الفريدة للمواقع ومكان ظهور كل كلمة. ويعد الناتج عن هذا الأمرهو مجموعة جداول ضخمة قابلة للبحث، والتي توفر كافة محددات الموقع التي تشير إلى الصفحات التي تظهر فيها الكلمات والعبارات.

2- **كشافات البناء** *The Structure index*: والتي تعكس الروابط بين الصفحات، وتشتمل على المعلومات التي تتعلق ببنية الروابط الفائقة للصفحات المكشفة وتحفظ في ملف يعرف بالكشاف الاساسي وغالبا ما يمثل العمود الفقري للزواحف حيث تعتمد عليه الثانية من خلال الروابط الفائقة في تتبع الصفحات لسحبها.

3- **كشافات الاغراض الخاصة** *The Utility index*: ككشافات الكيانات الاخرى غير الكيانات المكودة بالنصوص الفائقة مثل كشافات الملفات التي كتبت بصيغة *PDF* وكشافات الصور كشافات الزواحف وتعمل بشكل مستقل عن استفسارت المستفيدين.

- **مستودع الوثائق** *The pages repository*:

خلال عملية الزحف والتكشيف تقوم محركات البحث بتخزين وحفظ الصفحات الملتقطة من العنكبوتية في مستودع يعرف بمستودع الصفحات، بعض محركات البحث تقوم تخزين الصفحات التي تم زيارتها خلال فترة بناء الكشاف بشكل مؤقت، هذا الحفظ المؤقت يعمل على استرجاع صفحات النتائج بسرعة كبيرة، بالاضافة إلى تسهيلات بحثية من الممكن أن توفرها.

- **محرك الاستفسار** *The query engine*:

وهو المسئول عن استلام طلبات البحث والاستفسارات من المستفيدين.

- **وحدة الترتيب** *The Rank Module*:

وهي الوحدة المسئولة عن ترتيب وفرز النتائج ذات الصلة باستفسارات المستفيدين[21].

2- نشأة تحديات محركات البحث في استرجاع المحتوى:

ظهرت العديد من الدراسات العربية والأجنبية التي تتناول التأريخ لمحركات البحث ونشأتها وقد أفضت فيها بالتحليل والدراسة منها دراسة *History of Search Engines: From 1945 to Google Today* [22] والتي تعد من أوفي الدراسات عن تاريخ محركات البحث ودراسة محمد عبد المولى[23] وعلى ذلك أثر الباحث أن يتطرق إلى تاريخ ونشأة التحديات التي تواجه محركات البحث في استرجاع المحتوى على الانترنت بدلا من التعرض لتاريخ محركات البحث.

يعود تاريخ أول تحدي يواجه محركات البحث إلى تاريخ أولى الخدمات البحثية عن المحتوى التي وفرتها الشبكة العنكبوتية، والتي تمثلت فيما توفره بعض الخوادم من إمكانيات بحثية للمستفيدين في الصفحات المحلية لديها (فيما عرف فيما بعد بالبحث الداخلي في المواقع *Web site Internal Search*). أما عن محركات البحث فقد كانت رؤية مطوريها لها تتمثل في الاعتماد على تقنيات نظم استرجاع المعلومات التقليدية للبحث على العنكبوتية، فقاموا ببناء قواعد بيانات ضخمة في محاولة لنسخ متماثل لمحتوى العنكبوتية المصاغ بلغة النص الفائق HTML داخل هذة القواعد، ليكون البحث والاسترجاع من خلالها بدلا من البحث والاسترجاع من العنكبوتية بصورة مباشرة، وعليه أقتصرت هذة المحركات البحثية على لغة النص الفائق دون النظر إلى غيرها من الصيغ التي كانت متاحة في ذلك الوقت كصيغ (*BiblTex*) متجاهلة اياها، وقد أفضى هذا التجاهل إلى انشاء اول تحد يواجه محركات البحث في استرجاع المحتوى وهو العنكبوتية غير المرئية.

[21] Baeza-Yates, R., & Castillo, C. (n.d.). Web Search. *Waterloo Univesity*. Retrieved July 20, 2011, from softbase.uwaterloo.ca/~tozsu/courses/cs856/W05/.../Ricardo-WebSearch.pdf.

[22] History of Search Engines: From 1945 to Google Today. *Search Engine History.com*. Retrieved July 20, 2011, from http://www.searchenginehistory.c

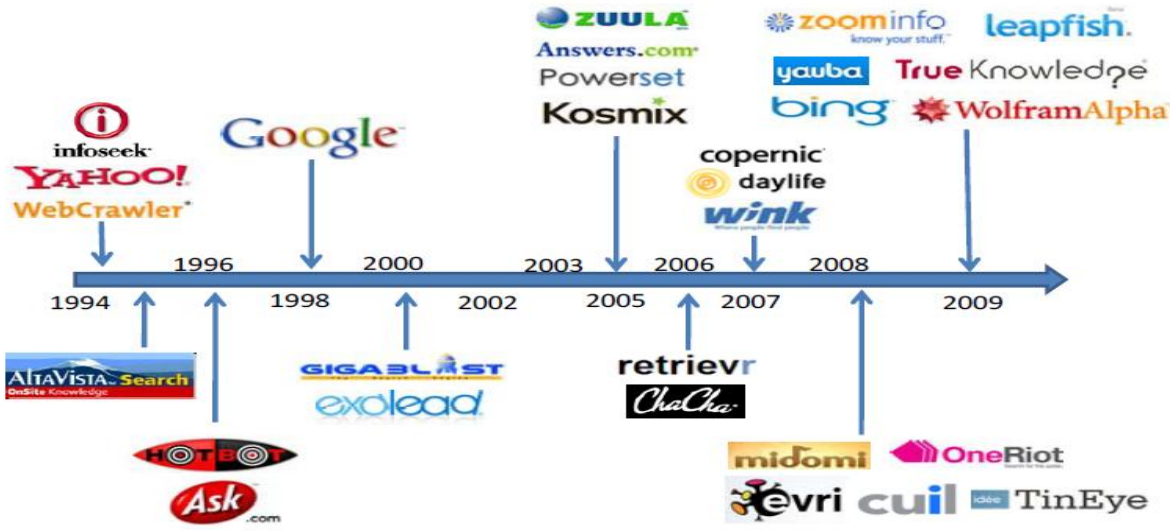
[23] محمد عبد المولى محمود .محركات البحث:من اين بدأت وإلى اين انتهت:بنيتها واساليب الاسترجاع. العربية 3000 متاح في : http://www.arabcin.net/arabiaall/index.html

يعود تاريخ التحد الثاني الذي واجه محركات البحث في بواكيرها، إلى معضلة افتقار الشبكة العنكبوتية (منصة النشر على الانترنت) إلى البنية المركزية ذات منطقية البحث والبناء المتسق الأمر الذي استتبع عدم القدرة على العثور على كافة انواع واشكال المحتوى المتاح على الانترنت مما أجهد محركات البحث في أكتشاف المحتوى المتاح على الانترنت.

وعليه دعت الحاجة إلى توفير برنامج حاسبي يعمل على تصفح وتجميع المحتوى من على العنكبوتية، وتمثل هذا البرنامج في الزواحف او الروبوتات (*Spiders*)، ولكن سمة تحد آخر ظهرت ملامحه، مفاده أن الزواحف تقوم بزيارات متكررة ولكن للصفحات التي تشتمل على محتوى يتمتع بشعبية عالية –(يقصد بالشعبية هنا انها تحظى بعدد من الروابط والإشارات الراجعة اليها من محتوى اخر) على حساب المواقع الاخرى التي قد تشتمل على محتوى مناظر في الأهمية، كما أن هذة الزواحف اتسمت في بعدم اشتمالها على قدرة تجنب التكرار في التجميع لمحتوى سبق لها أن قامت بتجميعه في نفس الفترة مع تجاهل تام للصفحات الجديدة الاخرى والتي قد تشتمل على محتوى جديد بدورها.

الفترة	اسم محرك البحث
1993	ALIWEB (Archie Linking), WWWWander, JumpStation, WWWWorm
1994	EINet Galaxy, WebCrawler, Lycos, Yahoo!
1995	Infoseek, SavvySearch, AltaVista, MetCrawler, Excite
1996	HotBot, LookSmart
1997	NorthernLight
1998	Google, InvisibleWeb.com
1999	FAST
2000+	Hundreds of search tools

شكل رقم (6): جدول زمني يوضح نشأة محركات البحث.



شكل رقم (7): الخط الزمني لنشاة أشهر محركات البحث على صعيد العالم منذ 1994 حتى عام 2009[24]

الزواحف والتحديات التي تواجهها في تجميع المحتوى لمحركات البحث:

تعرف زواحف العنكبوتية بأنها برنامج مصمم لتجميع المحتوى من على العنكبوتية، جدير بالذكر أن هذة البرامج لايقتصر استخدامها على محركات البحث فحسب بل تستخدم لإغراض متعددة قام بتحديدها كلا من Marc Najork &Christopher Olston على النحو الاتي:

- أحد المكونات الاساسية لمحركات البحث لتجميع صفحات العنكبويتة.
- ارشفة العنكبوتية.
- التنقيب عن البيانات على العنكبوتية.
- لرصد نمو العنكبوتية للخروج بدلالات احصائية[25].

[24] The Search Engine Industry. *Tommaso Buganza and Emanuele.springer .2010.*
[25]Christopher Olston and Marc Najork. Web Crawling. Foundations and Trends in Information Retrieval. Vol. 4, No. 3 (2010).

إن الضحامة التي تتمع بها الشبكة العنكبوتية من حيث حجم محتواها المتاح لا يجعل الزاحف قادرًا الإ على تجميع جزء ضئيل جدا من محتوى العنكبويتة خلال فترة زمنية محددة، ولذلك يجب على الزاحف ان يحدد اولوياته من عملية تجميع المحتوى وفقا للغة او المكان أو الموضوعات.

لايقتصر الامر على ضخامة العنكبوتية فحسب بل يزداد الامر سوءا في ظل ارتفاع معدلات التغيير والتحديث للمواقع خلال فترة زمنية متلاحقة وعليه قد ينقضي الامر بأن الصفحة التي جمعها الزاحف تخضع لاحتمال انها قد حذفت او عدلت او حدثت.

أولا: سياسات الزاحف:

إن سلوك الزاحف على العنكبويتة العالمية هو نتيجة لمجموعة من السياسات، حيث تحتاج الزواحف إلى سياسة لجدولة عملية التجميع ويجب أن تتسم هذه السياسة بالذكاء في التجميع وهو الأمر الذي يمثل **تحديا للزاحف في تجميع والتقاط المحتوى**، ففي ظل ارتفاع معدلات نمو محتوى العنكبوتية، تتسم الزواحف بمحدودية في الطاقة الاستيعابية للتجميع المحتوى.

إن نمطية عمل الزواحف تحدد وفقا لمجموعة من السياسات والتي تركزعلى مجموعة من الاهداف وتشمل:

- **سياسة الاختيار** *Selection policy*: حيث يحدد فيها طبيعة المحتوى الذي يجب أن يجمع.
- **سياسة تكرار الزيارة** *Re-visit policy*: فتتمثل في تحديد أوقات اعادة التجميع والجدول الزمني لها.
- **السياسة الأخلاقية للزاحف** *Politeness policy*: وتتمثل في سياسة التهذيب في عدم الاثقال على المواقع في تحميلها.

وسوف يتم تناول هذة السياسات بشئ من التفصيل:

1- سياسة الاختيار:

أوضحت دراسة *Lawrence and Giles*[26] ان تغطية محركات البحث لاتتجاوز نسبة 16% لما هو متاح على العنكبوتية من محتوى قابل للتكشيف، كما أن زواحف محركات البحث لا تجمع الا جزء ضئيل جدا من هذا المحتوى.

ومن ثم كان لزاما على محركات البحث أن تضع سياسات لإختيار المحتوى هذة السياسات تعد بمثابة محددات لهوية محركات البحث، فوفقا لسياسات الاختيار توجد محركات البحث العمودية *Vertical search engine* والتي ينصب اهتمام زاحفها على تجميع المحتوى من دولة واحدة أو منطقة واحدة كالوطن العربي مثلا، ويعد هذا النوع من أقل أنواع الزواحف مواجهة لتحديات الشبكة العنكبوتية حيث تتمثل وجهته في المواقع التي حددت له من واقع نطاقات اسمائها.

أما النوع الاخر فيتمثل في الزواحف ذات سياسة الاختيار للمحتوى العالمي *Global Web Search Engine* والتي تصطبغ محركات البحث خاصتها بالصبغة العالمية، ويعد **التحدي الرئيسي الذي يواجه هذة الفئة من الزواحف** في تحديد متى تتوقف عن التجميع والزحف في ظل فضاء يتسم بلا نهاية فيه، وفي هذا الصدد تستخدم الزواحف معيار تقليدي يعرف بعمق الرابط *link depth* ومفاده أن نقطة البداية تتمثل في الصفحة الرئيسة، وعلى الزاحف أن يبتبع الروابط بداخلها وفقا لمستوى محدد، يحدده محرك البحث، ويمثل هذا الأمر تحديا **أخر فما هو المستوى الذي حدد للزاحف بالتوقف عنده وماهي الشروط التي يستوفيها هذا المستوى؟**

أما النوع الثالث من محركات البحث وفقا لسياسات الاختيار يعرف بالزواحف الموضوعية المركزة *topical Focused crawlers* والتي ينصب اهتمامها على تخصصات محددة ويتمثل **التحدي الذي يواجه الزواحف المركزة** أن الحاجة إلى توفير القدرة على التنبوء بدرجة صلة الصفحات التي يجمعها قبل عملية التجميع، ثانيا صعوبة تحديد أماكن الوثائق الصالحة على العنكبوتية، وثالثا كيف يمكن للزاحف استبعاد وترشيح المحتوى أو الوثائق غير الصالحة قبل أن يجمعها ورابعا أن معظم هذة الزواحف يعتمد على محركات البحث العامة كنقطة بداية له ومن ثم تنسحب عليه التحديات التي تواجهها محركات البحث العامة.

[26] Lawrence, S., & Giles, C. L. (1998, March 4). Searching the World Wide Web. SCIENCE. Retrieved July 20, 2011, from clgiles.ist.psu.edu/papers/Science-98.pdf

2- سياسة إعادة التجميع او الزيارة *Re-visit policy* :

تعد الشبكة العنكبوتية بيئة ديناميكية متغيرة لنشر المحتوى، وعليه نجد أن عملية الزحف تستغرق وقتاً طويلاً عادة قد تصل إلى اسابيع أو شهور في تجميع الصفحات، وحينما يفرغ الزاحف من تجميعه للمحتوى أو للوثائق تكون الكثير من التحديثات والإضافات قد حدثت لما قام بتجميعه من محتوى.

إن رؤية محركات البحث في الزحف تتمثل في أنها لاتعمل على أكتشاف ما يستجد من تعديلات في المواقع المجمعة لديها، بل تجميع ماهو جديد من الصفحات، في المقابل نجد أنها توفر خوارزميات أخرى تعمل من خلالها على اكتشاف التعديلات في المواقع المخزنة لديها، تتمثل هذة الخوارزميات في خوارزمية الحداثة *Freshness* وخوارزمية عمر الوثيقة أو الصفحة *Age*.

- حداثة الصفحة *Freshness*: يعد مقياسا ثنائي يعمل على اكتشاف ما إذا كانت النسخة المجمعة حديثة أم لا. فحداثة الصفحة p في قاعدة بيانات محرك البحث في الوقت t تعرف من خلال هذة الخوارزمية الاتية:

$$F_p(t) = \begin{cases} 1 & \text{if } p \text{ is equal to the local copy at time } t \\ 0 & \text{otherwise} \end{cases}$$

- عمر الصفحة *Age*: تعد خوارزمية لقياس ما إذا كانت النسخة المحفوظة قد عفا عليها الزمن أو لا، فعمر الصفحة P في قاعدة بيانات محرك البحث يحسب من واقع الزمنT على هذا النحو:

$$A_p(t) = \begin{cases} 0 & \text{if } p \text{ is not modified at time } t \\ t - \text{modification time of } p & \text{otherwise} \end{cases}$$

قد تسعى الزواحف احيانا بجانب هدفها الأساسي إلى الحفاظ على متوسط أو معدل حداثة الصفحات المحفوظة لديها بشكل مرتفع، او الحفاظ على متوسط او معدل عمر الصفحات عند ادنى مستوى ممكن.وهذة ليست علاقة مطردة او متساوية ففي الحالة الاولى يركز الزاحف على كم عدد الصفحات القديمة المهملة أما في الحالة الثانية فيركز على عمر النسخ التي يحتفظ بها.

3- السياسة الاخلاقية للزواحف *Politeness policy*:

ان الزواحف تعد من التقنيات المفيدة في التجميع ولكنها في ذات الوقت ترهق الكثير من خوادم الشبكات في ظل تحميل هذة الخوادم طلبات أكثر مما يتحملون خلال زيارتها لهم.

وفي هذة الحال تقدم العديد من المواقع والشبكات إلى استخدام مايعرف ببروتكول إستقصاء أو استبعاد الزواحف *robots exclusion protocol* وهو معيار يمكن لمديروا المواقع أو منشئوا المحتوى من تحديد أي الاجزاء من خوادهم لاينبغي أن تكون متاحة للزواحف، اما فيما يتعلق بالفاصل الزمني لإتاحته وتصفحه فيها تتروح ما بين 20 ثانية إلى 3- 4 دقائق[27].

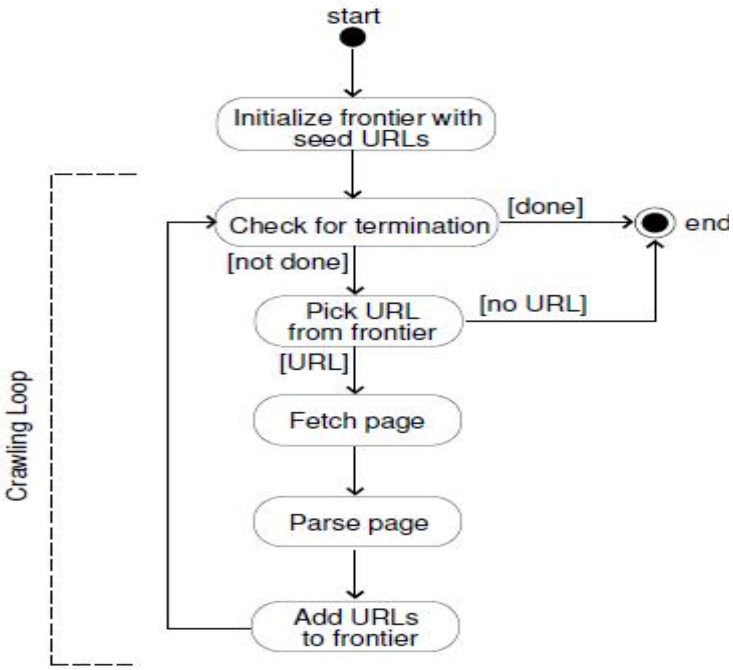
ثانيا : البنية المعمارية للزواحف ومنهجية عمله:

[27] Baeza-Yates, Ricardo, and Carlos Castillo. "Web Search." Waterloo Univesity. University of Chile, n.d. Web. 20 July 2011. <softbase.uwaterloo.ca/~tozsu/courses/cs856/W05/.../Ricardo-WebSearch.pdf>.

يعد من الضروري أن يتمتع الزاحف باستراتيجيات وسياسات زحف كما أوردنا سابقا، ومن ثم يتطلب ذلك بنية معمارية ذات أداء قوي ومرتفع، ومع ذلك فإن بناء زاحف قوي يرتبط بالعديد من التحديات المتعلقة بكفاءة الشبكة المكشفة والفاعليتها.

يمثل الزاحف قلب محرك البحث، ورغم ذلك فإن خوارزميات الزاحف وتفاصيل أداءه تحاط بسرية خاصة من قبل محركات البحث ذاتها، حيث لا تعلن الشركات عن خوارزميات زحفها، حتى أنه عنما يتم نشر تصاميم الزاحف فكثير التفاصيل الهامة في بنيته لايتم نشرها أو ذكرها، مما يصعب على الباحثين أدراك البنية المعمارية الكاملة لمحرك البحث، ولعل المرجعية في ذلك ترجع إلى أن سمة تخوف لدى محركات البحث من أعلان خوارزميات زواحفها فتصبح فريسة سهلة لخادعي محركات البحث *Search engine Spammer*.

أما البنية النموذجية لزواحف محركات البحث فتتمثل في الشكل رقم (8):



شكل رقم (8) خريطة تدفق توضح بنية الزاحف منهجيته في الزحف[28]

تبدأ عملية الزحف من قبل محركات البحث من خلال تزويد الزاحف بمجموعة من عناوين المواقع والتي لم تتم زيارتها list of unvisited urls والتي تعرف بأسم حدود أو جبهة الزاحف the frontier وهذة القائمة تهيئ كمحددات بذرية seed points حيث يتم توفيرها يدويا أو من خلال برنامج أخر كإدلة البحث yahoo، حيث أن كل عملية زحف تنطوي على أختيار المحدد التالي من جبهة الزاحف، ثم يتم بعد ذلك جلب fetching الصفحات الموافقة لـــ URLs من خلال بروتكول النص الفائق HTTP، ثم يتم بعد ذلك وضع هذة المحددات في قائمة انتظار، بعد ذلك تتم عملية تحليل Parsing لتلك الصفحات لاستخراج URLs (الروابط الخارجة) واضافتها لقائمة المحددات بعد تعيين درجة تقديرية تمثل الجدوى منها، وتكرر هذة العملية بالنسبة للصفحات الجديدة، ويمكن انهاء عملية الزحف عند عدد معين من الصفحات، أم اذا كان الزاحف لديه القدرة فيمكن أن يكمل عملية الزحف، ولكن تتسم جبهته حين إذاً بالفراغ، وذلك يؤدي إلى الوصول إلى طريق مسدود للزحف dead-end، ووهو ما يشير إلى أن هناك عمقا لعملية التجميع Crawling Depth وقد وجد ان عمق التجميع المثالي يتراوح مابين 3 إلى 5 مستويات انطلاقا من صفحات البداية وذلك بهدف الوصول إلى نسبة مرتفعة من الصفحات الهامة التي يتم زيارتها بالفعل من قبل المستفيد.

ويمكن اجمال خطوات الزحف على النحو الاتي:

- البداية من مجموعة بذرية من الصفحات.
- تحديد الصفحات الجديدة التي أضيفت للمجموعة البذرية من خلال التحميل الهابط لها.
- استخراج الروابط الفائقة التي بداخلها.
- حفظ هذة الروابط في قائمة الجلب للاسترجاع.

[28] Pant, G., Srinivasan, P., & Menczer, F. (n.d.). Crawling the Web. *University of Iowa*. Retrieved July 21, 2011, from http://dollar.biz.uiowa.edu/~pant/Papers/crawling.pdf

- الاسترمار في عملية الزحف حتى التوقف عند عدد معين من الصفحات محدد سلفا او فراغ قائمة الجلب.

يمكن أجمال التحديات التي تواجه زواحف محركات البحث في تجميع المحتوى العربي في:

1. ***ماهي الصفحات أو طبيعة المحتوى التي يجب على الزاحف ان يجمعها؟***

ففي معظم الحالات لايستطيع الزاحف تحميل وتجميع كافة الصفحات المتاحة على العنكبوتية، وفي ظل ذلك يجدر على الزاحف ان يحدد الصفحات التي يجب زيارتها وذلك وفقا لأهميتها.

2. ***كيف يحافظ على حداثة الصفحات؟***

فبمجرد أن ينتهي الزاحف من التجميع يجب أن يقوم بإعادة الزيارة خلال فترات منتظمة لما قام بتجميعة من الصفحات حفاظا على حداثتها.

3. ***كيف يتم تحديد الحد الادنى من التحميل والتجميع للمحتوى؟***

فمن المعروف أن زيارة الزاحف للمواقع يستهلك الكثير من الموارد العائدة لمنشئ هذة المواقع، فعند زيارة محرك البحث لتجميع الصفحة p من الموقع S ينطوي ذلك على أن يقوم الموقع باستدعاء الصفحة من نظام الملفات لديه مما يؤدي إلى استهلاك الطاقة التشغيلية للCPU الخاصة بالحاسب أو وحدة التخزين الاساسية، ثم يتم بعد ذلك نقلها من خلال الشبكة مما يؤدي إلى أستهلاك الموارد المادية للموقع.

4. ***كيفية الزحف المتوازي:***

فبسبب حجم العنكبوتية العملاق، يجب على الزواحف أن تقوم بعملية التجميع بشكل متوازي على أغلب الحاسبات فالتوازي يعد أمرا ضروريا في ظل الحاجة إلى تجميع أكبر عدد من الصفحات وفي ذات الوقت.

5. **الكشف عن وجود مكررات على العنكبوتية على صعيد المحتوى.**
6. **تحديد المجموعات البذرية اي ماهي الصفحات التي يجب ان تزار اولا.**

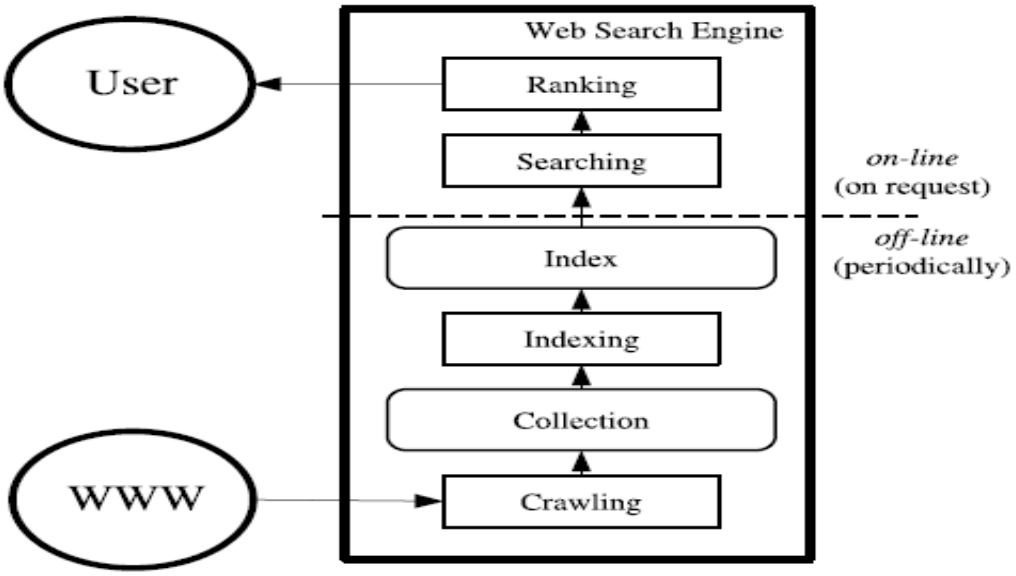
7. **ديناميكية النص الفائق: فقد يشتمل موقع ما على محتوى مصاغ بلغات ديناميكية تتغير كلما قام الزاحف بزيارتها**[29].

تكشيف المحتوى والتحديات التي تواجه في محركات البحث:

إن البحث عن المحتوى على العنكبوتية ينطوي على قسمين كما هو موضح في الشكل رقم (9):

- القسم الأول: البحث غير المباشر *off line*: والذي ينفذ من قبل محركات البحث بشكل دوري وتركز فيه على التحميل الهابط *Downloading* لمحتوى مواقع العنكبوتية لبناء مجموعة الوثائق التي ترحل بعد ذلك إلى الكشاف لتكشف به.
- القسم الثاني: البحث المباشر *online*: والذي يجرى وينفذ من جانب المستفيد دون التقيد بوقت أو زمن عبر استفساراتهم، ويتم استخدام فيه الكشاف لاختيار بعض الوثائق والتي تفرز وفقا لتقدير صلة محتواها باحتياج المستفيدين المعبر عنها داخل الاستفسار.

[29] Castillo, C., (2005) "*Effective web crawling*", SIGIR Forum, ACM Press,. Volume 39, Number 1, N, pp.55-56.



شكل رقم (9) رسم توضيحي لأقسام البحث داخل محركات البحث[30]

إن وثائق الشبكة العنكبوتية تأتي في العديد من التنسيقات المختلفة من صيغ pdf , html وغيرها من تنسيقات الصفحات، وتتمثل المرحلة الاولى في عملية التكشيف هو ايجاد رؤية منطقية قياسية للوثائق، ويعد "نموذج حقيبة الكلمات *Bag of Words*" أكثر الرؤى المنطقية استخداما في محركات البحث، فوفقا لهذا النموذج ينظر إلى المحتوى على أنه مجموعة غير منتظمة من الكلمات والتي ينبغي نظمها في ملف التكشيف والذي يبنى وفقا لرؤية واضعي خوارزميات محرك البحث.

وتمتد هذة الرؤية لتركز على تردد الكلمات داخل محتوى الوثيقة، وسمات تنسيق النص، والبيانات الوصفية للصفحات كالكلمات المفتاحية المدرجة في بنية الوثيقة الخلفية *Meta information*.

يتم استخراج الكلمات الدالة keywords من بنية محتوى الوثيقة من خلال العديد من عمليات التطويع (تطبيع النصوص للمعالجة) *Text Normalization Operations* ، فالخطوة الأولى لعملية التكشيف تنطوي على تطبيع

[30] Baeza-Yates, R., & Castillo, C. (n.d.). Web Search. *Waterloo Univesity*. Retrieved July 20, 2011, from softbase.uwaterloo.ca/~tozsu/courses/cs856/W05/.../Ricardo-WebSearch.pdf

النصوص وفقا لشكل موحد، حيث تكفل هذة العملية أعادة هيكلة النصوص بصورة منطقية توفر من خلالها القدرة على البحث فيها، وذلك في ظل التعامل مع البيئة الرقمية والتي تتسم بعدم اتساق تنسيقاتها وأشكال بياناتها ومن هذة العمليات:

- **التأخيذ** Tokenization :

وتشمل هذة العملية على تفتيت النصوص الكاملة لكلمات وتحديدها، وهنا يتضح أول تحد في عملية التكشيف ففي بعض اللغات مثل اللغة الانجليزية تعد هذة الطريقة مجدية في ظل الكيان المستقل للكلمة الذي يتحقق من خلال المسافات وعلامات الترقيم داخل النص وغيرها من العناصر التي تعمل على تحديد ملامح الكلمة، ولكن لايجدي الامر نفعا في لغات اخرى خاصة اللغة العربية حيث تتشابك بنيتها دون فواصل او محددات مثل اللغة الصينية.

- **قائمة الاستبعاد** stopwords:

وتتمثل في استبعاد الكلمات التي تحمل دلالات معلوماتية ولغوية ضئيلة في الوثيقة، وفي نظم استرجاع المعلومات عادة ما يتم التخلص من هذة الكلمات لاسباب تتعلق بالكفاءة. ولكن سمة ملمح لتحد اخر وهو انها قد يشتمل محتوى الوثيقة على بعض الكلمات والمفردات الفنية والمحورية التي تحمل ايضا دلالات ضئيلة في المعنى ولكن تؤثر بشكل كامل على دلالات المحتوى الكامل للوثيقة، **ثانيا** تشغل هذة الكلمات مساحة كبيرة من حجم الكشاف نظرا لارتفاع تكرار وتردد وتيرة هذة الكلمات في بنية محتوى الوثيقة.

- **جدذور الكلمات** stemming:

تعمل هذة المنهجية على استخراج الجذور الصرفية لكل كلمات الوثيقة، ويتضح هنا تحد اخر يواجه محركات البحث مفاده في ظل عالمية محركات البحث مثل جوجل عليه ان يتعامل مع لغات تتسم جذورها اللغوية بالتعقيد وعدم المرونة مثل اللغة العربية التي من الممكن ان يكون الجذر اللغوي لكلمة ما لا علاقة له في البناء اللغوي بالمشتق منها.

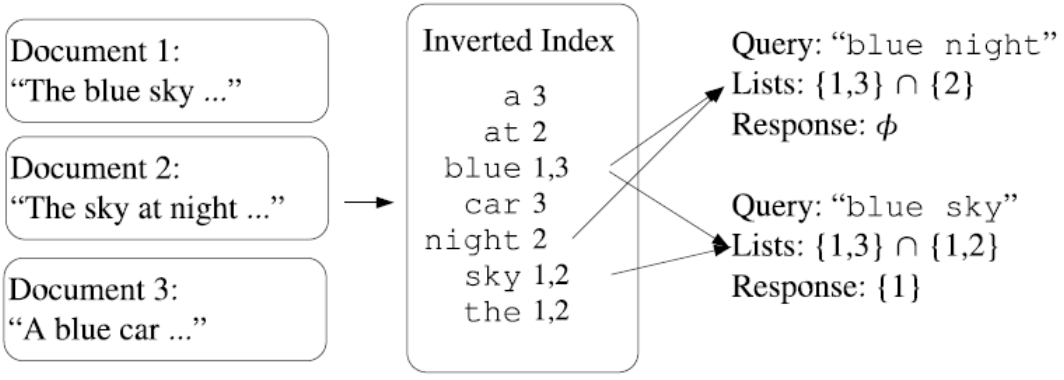
الكشاف المقلوب Inverted Index:

وهو ذلك الملف الذي يعمل على توفير سبل للوصول إلى محتوى الوثائق الذي يشتمل على المصطلحات الكشفية بشكل يضمن الفاعلية في الاسترجاع، يوفر الكشاف المقلوب طريقة مختصرة في عملية البحث، بدلا من البحث قاعدة بيانات الوثائق بأكملها لتحديد المصطلحات الواردة في الاستفسار كما هو موضح في الشكل رقم (10)، فالكشاف المقلوب يعمل على تنظيم المعلومات في قائمة مختصرة من المصطلحات ومن ثم الاعتماد على المصطلح في تحديد مجموعة الوثائق الملائمة وهو بمثابة الكشافات التي تأتي في نهاية المؤلفات والذي يسهل من خلاله تحديد هدف الباحث.

يتألف الكشاف المقلوب من ثلاثة اجزاء رئيسية :

- ملف الوثائق *Document file*: ويعمل على اكساب كل وثيقة معرف رقمي فريد، وتحديد كافة المصطلحات الموجودة داخل الوثيقة. فهو قائمة رتبت بداخلها كافة الكلمات المفتاحية التي وردت في الوثائق.
- القاموس Dictionary: يمثل قائمة مرتبة بظهور وتردد المصطلحات مع مؤشرات لما يعرف بالقائمة المقلوبة وترتب المصطلحات هجائيا بداخله.
- القائمة المقلوبة inversion list: وتشمل المؤشرات من المصطلحات إلى الوثائق التي تشتمل على تلك المصطلحات[31].

[31] Berry, M. W., & Browne, M. (1999). *Understanding search engines: mathematical modeling and text retrieval*. Philadelphia, PA: Society for Industrial and Applied Mathematics.



شكل (10) يوضح كيفية عمل الكشاف المقلوب في الاستجابة على الاستفسارات[32]

أما عن حجم الكشاف فإن اختيار المصطلحات لتسكن في قائمة الظهور يحدد حجم الكشاف، فمن الممكن ان يتسم حجم الكشاف بالصغر من خلال ترتيب محددات الوثيقة فقط للوثائق المطابقة، أما اذا قام محرك البحث بحفظ وتخزين أماكن ظهور المصطلحات في كل صفحة فيؤدي ذلك إلى ضخامة حجم الكشاف، مما يوفر قدرة على الاجابة على الاستفسارات المعقدة أو التي تعتمد على التجاور. **ومن هنا يظهر تحد اخر لمحركات البحث في تكشيف المحتوى** وهو المفاضلة بين صغر حجم الكشاف وبين القدرة على الاجابة على الأستفسارات المعقدة للمستفيدين.

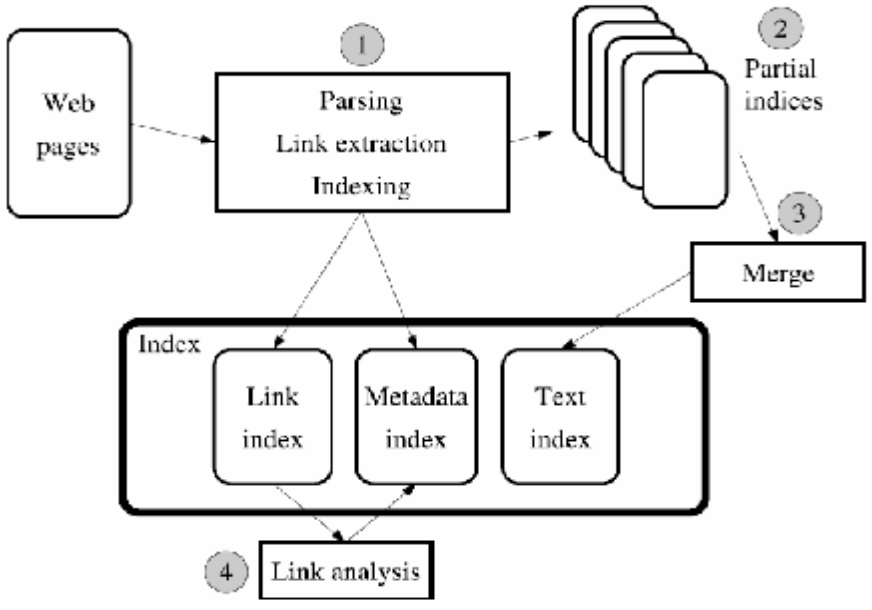
تحد اخر يتمثل في ضخامة حجم الملف المقلوب- والذي يتكون من ملف المصطلحات وملف اماكن ورودها – والمرجعية تعود إلى العلاقة الطردية بين حجم المصطلحات المكشفة وبين حجم اماكن ظهورها، ففي ظل نمو المصطلحات الكشفية في شكل خطي يتضاعف حجم ملف اماكن ظهورها، جدير بالذكر أن ملف الكشاف المقلوب يشغل مساحة من 10 إلى 20 % من حجم الملف الكامل، وبالتالي لايسمح لملف التكشيف بأن يوجد في الذاكرة الرئيسية وعلى هذا أوجد ملف التكشيف العديد من المؤشرات الكشفية التي يقوم ببنائها، بحيث يمثل كل مؤشر

[32] Levene, M. (2010). *An introduction to search engines and web navigation* (2nd ed.). Hoboken, N.J.: John Wiley.

كشفي مجموعة فرعية من اجمالي حجم الملف الكشفي المقلوب، ثم تدمج فيما بينها فيما بعد داخل الملف الكشفي الكامل.

كيف يجيب الكشاف على الاستفسارات:

ان استجابة الكشافات على استفسارات المستفيدين تتم من خلال الشكل رقم (11):



شكل رقم (11) يوضح مراحل عملية التكشيف في محركات البحث[33]

فالتكشيف في العنكبوتية يتضح كما في الشكل السابق على النحو الاتي:

1- يتم تحليل الصفحات واستخراج الروابط الفائقة لبناء ما يعرف بالشكل البياني للعنكبوتية Web graph حيث تحلل الروابط فيما بعد لتوليد درجات لها يمكن ان تحفظ فيما بعد مع البيانات الوصفية لتحديد درجة اهمية الوثيقة.

[33] Berry, M. W., & Browne, M. (2005). *Understanding search engines: mathematical modeling and text retrieval* (2nd ed.). Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics.

2- يبدء في انشاء الكشافات الفرعية في ظل استنفاذ مساحة التخزين الرئيسية يوجه اليها الاستفسار مباشرة.

3- الدمج بين نتائج الكشفات الفرعية داخل كشاف النص الكامل.

4- تحليل الروابط يمكن ان يستغل في تقدير العشرات من الروابط الاخرى.

وينبغي أن يتسم زمن الاستجابة للاستفسار المقدم من قبل المستفيدين لمحركات البحث بالسرعة والفاعلية، ويتم ذلك الأمر بطريقة متوازية عبر العديد من الآليات، فوفقا لمنهجية التوازي، يوزع الكشاف المقلوب بين العديد من الحاسبات ولايتم هذا التوزيع بشكل اعتباطي ولكن يخضع لتقنيتين يعرفا بالملف المقلوب الواسع global inverted file والملف المقلوب المحلي local inverted file.

عند استخدام الملف المقلوب الواسع global inverted file تنقسم قائمة المصطلحات والمفردات بداخل الملف الكشفي إلى العديد من الاجزاء بحيث توزع هذة الاجزاء بعد ذلك على العديد من الحاسبات بحيث يشمل كل حاسب ملف من المفردات يصاحب معه ملف باماكن ظهور هذة المفردات في الوثائق، وعند استقبال الاستفسار ترسل وحدة تعرف بالوسيط broker هذا الاستفسار إلى الحاسب الذي يقتني المفردات المضمنة في الاستفسار ثم تدمج النتائج بعد ذلك.

ما يمثل تحديا في هذا الصدد هو ان بناء هذا النمط من الملفات وصيانته وتحديثه امر مرهق ومكلف نظرا لمرور مايعرف بالوسيط Broker على الحاسبات للكشف عن المفردات التي تطابق الاستفسار.

الطريقة الثانية وتتمثل في استخدام الكشاف المقلوب المحلي local inverted file حيث توزع وتنقسم محددات ومعرفات الوثائق document identifiers على الحاسبات ولكن في هذة الحالة ينبغي على كل حاسب من هذة الحاسبات ان يشتمل على المصطلحات والمفردات الكشفية كاملة ومن ثم تلغى المرحلة الثالثة من الشكل السابق، وعندما يوجه الاستفسار من قبل المستفيدين يبث على كافة الحاسبات مما يوفر سرعة وتوازن في التحميل وهذة المعمارية هي التي تستخدم في اغب محركات البحث في وقتنا الحالي.

ان معالجة الاستفسار في الكشافات ينطوي على ما يعرف بالوسيط المركزي centeral broker، والذي يعين لمهمة توزيع الاستفسارات الواردة على الحاسبات ودمج نتائجها معا، جدير بالذكر ان النتائج تعرض في مجموعات تتكون كل مجموعة من 10 إلى 20 وثيقة من كل حاسب.

مايمثل تحديا أمام محركات البحث يتمثل في ان هذا الوسيط لا يقوم بطلب كافة القوائم او المجموعات التي تضاهي الاستفسار او حتى دمجها معا من كافة الحاسبات (مما يستتبع ذلك ما يعرف بمشكلة التداخل والتكرار في نتائج محركات البحث الواحدة) بل يقوم باخذ اعلى نتائج كل حاسب أو كشاف فرعي وحسب دون الكشف عن باقي النتائج الاخرى.

أحد أوجه القصور التي تكتنف تكشيف المحتوى تتمثل في استغلال ندرة أن يقوم المستفيد بتصفح النتائج التي ترد بعد الصفحة الاولى والثانية او بمعنى اخر قصور التصفح على الصفحتين الاولى والثانية من نتائج محركات البحث الامر الذي دعى محركات البحث بان تلجأ إلى توفير مجموعة نتائج تقريبية وحسب في ظل عدم إكتمال دمج كافة قوائم نتائج الكشافات الفرعية، وبالتالي فإن عدد الوثائق المسترجعة يمكن أن يحسب بسهولة من قبل محرك البحث ثم يسترجع ولهذا السبب عندما يتوجه المستفيد إلى الصفحة الثانية او الثالثة بالنقر عليها فان من الطبيعي ان يقوم محرك البحث باعادة تنفيذ الاجراءات السابقة لأخذ نتائج اخرى من الكشافات الجزئية أو الفرعية والتي لم تحسب في الحلقة الاولى مما يوفر مقدارا هائلا من التداخل والتكرار.

ويمكن إجمال التحديات التي تواجه تكشيف المحتوى في محركات البحث على النحو الآتي:

1- ما تقوم به محركات البحث من عمليات التأخيذ Tokenization والتي تعتمد على أن يقوم محرك البحث بتفتيت المحتوى الكامل إلى كلمات مستقلة وهو لايتناسب مع طبيعة بعض المحتويات الخاصة بالوثائق ففرضا إذا تم تفتيت معادلة حسابية فوفقا لهذا المبدء لايمكن لمحركات البحث ان تسترجع المعادلات الحسابية أو الرياضية.

2- ما تقوم به محركات البحث من أستبعاد لبعض الكلمات في المحتوى والتي قد تحمل دلالة ضئيلة ولكنها في ذات الوقت تلعب دورا محوريا.

3- منهجية التكشيف في محركات البحث التي تقضي باستخراج الجذور الصرفية للمصطلحات الواردة في المحتوى ودون أن تأخذ في اعتبارها أن اللغة العربية تتسم جذورها الصرفية بالتعقيد وعدم المرونة.

4- مفاضلة محركات البحث بين صغر حجم الكشاف وبين القدرة على إجابة الاستفسارات المعقدة من محتوى الوثائق.

5- بناء الكشافات الفرعية ضمن الكشاف المقلوب مما يسمح بوجود تداخل وتكرار في نتائج المحتوى.

التحديات التي تواجه خوارزميات ترتيب نتائج المحتوى في محركات البحث:

منذ البداية الأولى للنظم الآلية لاسترجاع المعلومات في منتصف القرن العشرين، عُمل على توظيف الخوارزميات واللوغريتمات الرياضية والمنطقية في الترتيب الطبقي لنتائج نظم استرجاع المعلومات، فتوالت الاسهامات في تطوير نماذج استرجاع المعلومات، كالنموذج البوليني Boolean retrieval، ونموذج فراغ الموجهات vector space model، ونموذج الاحتمالات probabilistic model، إلى أن سرعان ما تعطلت هذة الخوارزميات بسبب بيئة عمل جديدة تمثلت في الشبكة العنكبوتية، فأقدم المعلوماتيون على تطوير مجموعة من نماذج الاسترجاع وخورازميات الترتيب التي تلائم طبيعة الاسترجاع المحتوى في الشبكة العنكبوتية، كخوارزميات الترتيب المعتمدة على الروابط مثل نموذج ترتيب الصفحات PageRank ونموذج ترتيب المعتمد على الموضوع (Hypertext Induced Topic Search) HITS، إلى ان بعض محركات البحث قد نحنى منحنى اخر، حيث اعتمدت بعض محركات البحث على دمج خوارزميات الترتيب التقليدية مع خوارزميات الترتيب الخاصة بالعنكبوتية واستخدامها لاسترجاع وترتيب المحتوى.

أهمية نماذج استرجاع المحتوى:

أن جوهر التحديات التي تواجهها محركات البحث التقليدية في استرجاع المحتوى يكمن في العيوب وأوجه النقص في نماذج الاسترجاع اعتمادا على رؤية Ricardo baeza Yates حيث أوضح "أن المشكلة الرئيسية في محركات البحث بشكل خاص، تكمن في قضية التنبوء بتحديد أي من محتوى الوثائق يتسم بالصلة لما يمكن أن يقدم من استفسارات، وأي منها لايتسم بالصلة.[34]"، ومثل هذا القرار لايخضع إلى الحدس أو التخمين بل يستند وبشكل أساسي على خوارزميات لترتيب والاسترجاع تعمل على إنشاء قائمة مرتبة بمحتوى الوثائق المسترجعة، ويكون مدلول هذة القائمة مفاده أن الوثائق التي تظهر في أعلى القائمة تحمل محتوى أكثر دلالة بموثوقية الصلة بالاستفسار المقدم وعليه تمثل خوارزميات الترتيب الطبقي ranking algorithms نواة وقلب نظم استرجاع المعلومات بما فيها محركات البحث.

فخوارزميات الترتيب الطبقي هي مجموعة من الفرضيات الرياضية والمبادئ المنطقية ألاساسية التي تسفر عن توفير ما يعرف بنماذج استرجاع المعلومات Information Retrieval Models لتحديد درجة صلة الوثائق بالاستفسار. وعليه تعمل نماذج استرجاع المعلومات بصورة عامة تعمل على تحديد التوقعات والتقديرات المتعلقة بتميز أي من الوثائق تتسم بالصلة واي منها لايتسم بالصلة الاستفسارية.

تعرف نماذج استرجاع المحتوى على أنها مجموعة من الفرضيات والخوارزميات التي تعمل على توفير الترتيب الطبقي لمحتوى الوثائق المتعلقة بإستفسار المستفيد, وبشكل أكثر تفصيلا تعمل نماذج استرجاع وترتيب المحتوى وفقا لمعادلة رباعية تتمثل اطراف هذة المعادلة في العناصر الاتية $[D, Q, F, R(q_i, d_j)]$ حيث يمثل:

- حيث يمثل حرف D مجموعة الوثائق Document داخل محرك البحث.
- يمثل حرف Q مجموعة استفسارات Query المستفيد.
- يمثل حرف F الاطار الخاص Framework بنمذجة المحتوى والاستفسارت معا.

[34] Yates, R., & Neto, B. (1999). *Modern information retrieval* . New York: ACM Press.

- يمثل حرف R دالة الترتيب الطبقي المرتبطة بالترتيب الرقمي Rank لكل من الاستفسار qi والوثيقة dj .[35]

فئات نماذج استرجاع المعلومات:

تعتمد منطقية ترتيب الوثائق المسترجعة على حساب درجة التشابه بين الاستفسار والوثائق المكشفة، وعليه كلما ارتفعت درجة التشابه ارتفعت رتبة الوثائق المشابه للاستفسار. تاتي خوارزميات ترتيب المحتوى في محركات البحث على صعيد بسيط ومعقد على النحو الآتي:

اولا: النماذج البسيطة: وتشمل:

1. التحليل من خلال النقر:

تعتمد هذة المنهجية على استخدام البيانات المتعلقة بتردد اختيار المستفيد لمحتوى صفحة بعينها استجابة للاستفساره كوسيلة للترتيب الطبقي او بمعنى اخر انها تعتمد على تسجيل استفسارات المستفيدين ومحددات المصادر الخاصة بمحتوى الوثائق المسترجعة، والتي قام بالنقر عليها للدلالة على مطابقتها للاستفسار المدخل، ومن ثم يحتل محتوى هذة الوثائق طبقة عليا عن غيرها في حالة إدخال نفس الاستفسار إلى محرك البحث.

2. تحليل الروابط:

تعد مرجعية هذة المنهجية إلى علم المعلومات والمكتبات حيث عُني بدراسة وتحليل الاستشهادات المرجعية، فمنهجية الروابط تعتمد على فحص الروابط التي تشير إلى محتوى الوثيقة، فمن خلالها تحدد درجة أهمية محتوى الوثيقة في موضوع محدد اعتماد على كم ونوعية الروابط التي تشير إلى هذا المحتوى.

3. تردد المصطلح:

[35] Yates, R., & Neto, B. (1999). Modern information retrieval . New York: ACM Press

وهي منهجية حسابية تعتمد على تقييم حساب تردد ظهور المصطلح في محتوى الوثيقة، فبشكل عام يدل التردد المرتفع لظهور الكلمات في محتوى الوثيقة على إحتمالية ان هذا المحتوى أشد ارتباطا بالاستفسار ويصحب هذة التقنية وجود ما يعرف بقائمة الاستبعاد.

4. موقع المصطلح:

في كثير من الاحيان يشير موقع المصطلح إلى أهميته في محتوى الوثيقة، ومن ثم أهمية محتوى الوثيقة في المجال الموضوعي التي تنتمي اليها، علاوة على ذلك تعتمد معظم محركات البحث إلى إعطاء وزن أكثر للمصطلحات التي تظهر بشكل جلي في اجزاء معينة من محتوى الوثيقة، مثل العنوان والفقرة الرئيسية والمستخلصات وتعليقات الصور عن نظيرتها التي تظهر في جسد الوثيقة أو في الهوامش السفلية.

5. تقارب المصطلح:

منطقية هذا العامل تتمثل في أن قرب المسافة بين مفردتين أو أكثر في الاستفسار يحقق نسبة مرتفعة في استرجاع محتوى الوثائق الذي يتشابه في قرب مفرداته مع الاستفسار، وهذا النمط يفضل في البحث عن أسماء الاشخاص والكيانات المعرفة.

6. تنسيق النص:

وهو نمط شكلي معني بتنسيق محتوى الوثيقة حيث أن الكلمات التي تتسم ببنط مختلف في كتابتها عن غيرها كجعلها بخط سميك تمثل أهمية في حمل محتوى الوثيقة إلى الطبقة الاعلى في النتائج المسترجعة.

7. حقل العنوان:

فبتدوين حقل العنوان في تكويد الوثيقة يوفر لها درجة مرتفعة في الترتيب الطبقي .<Title></title>

ثانيا: النماذج المعقدة وتشمل:

- النموذج البوليني The Boolean model.
- نموذج فراغ الموجهات Vector Space Model ونموذج التكشيف الدلالي الكامن latent semantic indexing.
- النموذج الاحتمالي Probabilistic Model.
- نموذج رتبة الصفحة The PageRank.
- نموذج تحديد الرتبة وفقا لروابط الفائقة HITS (Hyperlink Induced Topic Search).

ورغم ما قد يعتلي هذة النماذج من إختلاف في الأساس الخوارزمي والتكويني لهم، إلى ان هناك سمة من القواسم المشتركة بينهم والتي تتمثل في:

- إن غالبية هذة النماذج صممت للتعامل مع النص فقط، فهي تعتمد في خوارزميتها على تكشيف النص بالنص، بمعنى أنه يتم تحديد المصطلحات الكشفية الخاصة بمحتوى الوثيقة وعدد مرات ظهورها لتمثل بعد ذلك في معادلة ما للتعبير عن أهمية الوثيقة.
- يعامل أيضا استفسار المستفيد على غرار محتوى الوثيقة في كونه جزء من النص المكشف ومن ثم تمثيله على غرار الوثيقة.
- بصرف النظر عن الأسس الخاصة بهذة النماذج، فإن تعاملها مع محتوى الشبكة العنكبوتية أكسبها بعدا أخر، وهو الربط بين محتوى الصفحات بعضها البعض، مما كان له عظيم الأثر في استغلال هذا البعد والإفادة منه في ترتيب محتوى الوثائق وتحديد صلتها بالاستفسار.
- إن تحديد درجة الاهمية او التشابه بين الاستعلام وبين محتوى الوثيقة (أو صفحة العنكبوتية) يتم عن طريق الحساب الرقمي والعددي لإهمية الأوزان والروابط.

للبحث بقية العدد القادم