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

العدد 30، **ديسمبر** 2012

التحديات التي تواجه خوارزميات محركات البحث في استرجاع المحتوى العربي على الشبكة العنكبوتية العالمية: دراسة مسحية تحليلية . 2

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.

الاستشهاد المرجعي

النشرتي، مؤمن سيد.التحديات التي تواجه خوارزميات محركات البحث في استرجاع المحتوى العربي على الشبكة العنكبوتية العالمية: دراسة مسحية تحليلية . 2 .- Cybrarians Journal .- ع 30 (ديسمبر 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

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

العدد 30، ديسمبر 2012

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.

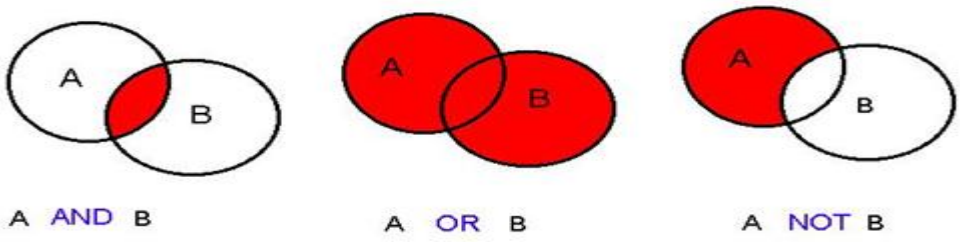
This study tries answering about a set of Queries that related to:

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.

النموذج البوليني The Boolean model .

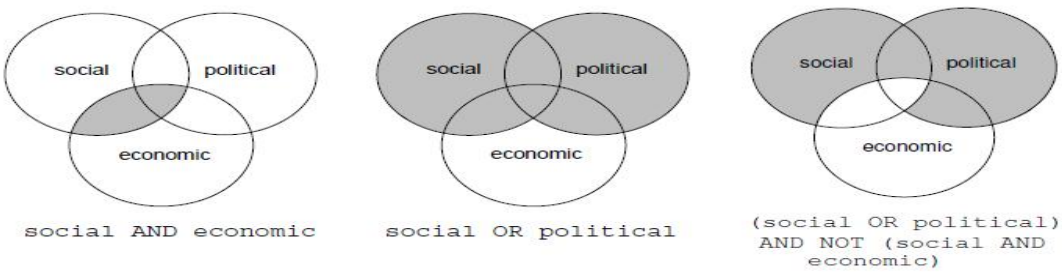
يأتي النموذج البوليني على رأس نماذج استرجاع وترتيب المحتوى في محركات البحث وأقدمها واكثرها خضوعا للدراسة، (ويجب الإشارة إلى امكانية البحث البوليني التي توفرها معظم محركات البحث في وقتنا الراهن تعتمد على النموذج البوليني)، يعتمد النموذج البوليني في أساسه على مفهوم نظرية المجموعات[1] والجبر البوليني، ومنطقية عمله تعتمد على إقران محتوى الوثيقة بمجموعة من الكلمات المفتاحية لتمثيلها، وبالنسبة للاستفسار فيتم أيضا رصد الكلمات المفتاحية به وتحديد سمة الاقتران بين الكلمات المفتاحية من خلال الروابط البولينية مثل (و، أو، لا[2]) كما هو موضح في الشكل رقم (12).



[1] تمثل نظرية المجموعات احد فروع الرياضيات المعنية بنمذجة الكيانات في صورة مجموعات وتحديد درجات الارتباط بين هذه الكيانات في ظل صورتها المجموعية .

[2] وتنسب هذه الروابط إلى عالم الرياضيات والمنطق البريطاني جورج بول Geogre Boole ، الذي يعد أول من استخدم مجموعة من الرموز الرياضية للتعبير عن بعض العمليات المنطقية ، وتتمثل هذه الروابط في الكلمات الثلاث التالية AND , OR NOT.

يعمل النموذج البوليني على توفير نظام يتمتع بالسهولة في ان يدرك من قبل المستفيدين والمتعاملين مع محركات البحث علاوة على ذلك فإن استفسارات المستفيدين تحدد وفقا للروابط البولينية والتي تتسم بالدلالة المحكمة.



شكل رقم (12) يوضح كيفية عمل خوارزمية الربط البوليني

ورغم ما يتمتع به من بساطة واحكام في منطقه **الا ان الكثير من العيوب تعتليه والتي تمثل تحديا في استرجاع وترتيب المحتوى داخل محركات البحث**:

- اولها المعيار ذو الحكم الثنائي على درجة الصلة بين الاستفسار ومحتوى الوثيقة (فمحتوى الوثيقة وفقا للنموذج البوليني إما أن يكون متصل relevant أو غير متصل non-relevant)، فالنموذج البوليني ينظر إلى الكلمات والمفردات الكشفية على كونها موجودة أو غائبة دون تحديد لدرجات الصلة، مما يحول دون استرجاع جيد وفعال لمحتوى الوثائق فالمصطلح الكشفي لايمثل بوزن بين (0،1) كما يحدث في جميع الخوارزميات بل يحصل على أحد الإحتمالين إما (0) وإما (1).

- يتطلب المنطق البوليني استخدام نفس المصطلحات التي كشف بها محتوى الوثائق للتعبير عن استفسار المستفيد، وذلك لضمان نجاح عملية المضاهاة.

- يتطلب المنطلق البوليني تدريباً للمستفيدين على صياغة الاستفسارات، لأنه يختلف عن اللغة الطبيعية في الاستخدام.

- نجد المعامل AND ، يحد من عملية البحث ، فالبحث عن A AND B AND C ، سوف يستبعد محتوى الوثائق التي لا تشتمل على المصطلحات الثلاثة مجتمعة، مع أنه يحتمل أن تكون وثيقة تشتمل على اثنين فقط ذات جدوى للمستفيد. وبالتالي نجد أن المعامل AND غالباً ما يؤدي إلى فشل عملية البحث.

- التعبيرات والروابط البولينية تتسم بأن لها دلالات محددة في كثير من الأحيان، فرغم ما تتسم بها من بساطة إلى انها ليست بسيطة في التعبير عن الحاجات المعلوماتية ففي الواقع نجد أن معظم المستفيدين يجدون صعوبة في التعبير عن مطالبهم وحاجاتهم في ظل إستخدام التعبيرات المنطقية أو البولينية.

Cybrarians Journal
ISSN 1887-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 30، ديسمبر 2012

النموذج ذات الموجهات في الفراغ Vector space model.

يعد نموذج الموجهات أحد وأهم نماذج الترتيب في استرجاع المحتوى استخداما وتوظيفا، وقد كان لبيتر لون الفضل في تطوير منطقية الاعتماد على الموجهات، حيث كان أول من اقترح استخدام النهج الاحصائي في البحث عن المعلومات عام 1957 معتمدا على معيار التماثل بين الاستفسارات والوثائق.

فوفقا للنموذج الاحصائي ينظر إلى محتوى الوثيقة على انه حقيبة كلمات Bag of words بمعنى أن محتوى الوثيقة يشتمل على مصطلحات غير مرتبة وذات ترددية غير منتظمة داخل محتوى الوثيقة.
كانت الفكرة الرئيسية التي بنيت عليها نموذج الموجهات في الفراغ أن استخدام الوزن الثنائي (يتصل أو لا يتصل (0،1)) يحد جدا من عملية الاسترجاع والترتيب الطبقي للنتائج، وبناءا عليه قدم هذا النموذج اطار عمل جديدا يعتمد على ما يعرف بالمطابقة الجزئية (اي ان درجة اتصال أو عدم اتصال الوثيقة بالاستفسار يحدد من خلال أوزان متفاوتة بين قيمتين أدناها يرمز له بالرقم 0 وأعلاها يرمز له بالرقم 1 ويتم ذلك من خلال مجموعة من المعادلات الخاصة بوزن المصطلح بحيث تستخدم هذة الاوزان في نهاية المطاف لحساب درجة التشابه والتماثل بين كل من الوثيقة المختزنة في النظام وبين استفسار المستفيد).

في نموذج فراغ الموجهات تحسب درجة صلة محتوى الوثيقة بالاستفسار من خلال تحديد درجة التشابه بينهما، حيث يمثل كلا من محتوى الوثيقة والاستفسار في صورة موجهات في فراغ متعدد الابعاد كما هو موضح في الشكل رقم (13).

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$
$$q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$

حيث ينطوي كل موجه على اوزان غير ثنائية للمصطلحات الكشفية في كلا من محتوى الوثيقة والاستفسارات والتي يشار اليها بالرمز w_1.
وتحسب درجة الصلة للوثائق من خلال مقارنة انحراف الزوايا بين كل من موجه الوثيقة وموجه الاستفسار كما هو موضح من خلال المعادلة الاتية:

$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|}$$

الأساس الرياضي الذي يعتمد عليه هذا النموذج:

يمكن ان توصف العلاقة بين محتوى الوثيقة D والمصطلح T من خلال المصفوفة **tf -idf** كمعيار كمي يشتمل على محورين اساسين:

المحور الأول: هو تردد المصطلح TF ويشير إلى عدد مرات ظهور المصطلح t في محتوى الوثيقة d وتأتي المعادلة لحساب تردد المصطلح على هذا النحو:

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- حيث تشير $\mathrm{tf}_{i,j}$ إلى حساب تردد المصطلح.
- تشير $n_{i,j}$ إلى عدد مرات ظهور المصطلح t_i في محتوى الوثيقة d_j.
- وتشير $\sum_k \mathrm{n}_{k,j}$ إلى مجموع عدد المصطلحات في اجمالي الوثيقة.

مثال اذا افترضنا ان وثيقة ما تتكون من 100 مصطلح، ويظهر مصطلح المكتبات 4 مرات في الوثيقة فإن المعادلة ستكون (4/100) =0.04

المحور الثاني: هو تردد الوثيقة العكسي، والذي يعمل على حساب نسبة اجمالي عدد الوثائق المختزنة في النظام إلى عدد الوثائق التي تشتمل على المصطلح T وتظهر معادلته على هذا النحو:

$$\mathrm{idf}_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

- تشير idf_i إلى حساب تردد الوثيقة العكسي.
- بينما تشير log إلى حساب لوغارتيم ناتج القسمة.
- وتشير $|D|$ إلى اجمالي عدد الوثائق في النظام.
- وتشير $|\{d : t_i \in d\}|$ إلى عدد الوثائق التي يظهر فيها المصطلح t_i .

وتبعا للمثال السابق، فإذا أفترضنا أن عدد الوثائق المختزنة في النظام تبلغ 1000000 وثيقة ويظهر مصطلح المكتبات في 1000 وثيقة من إجمالي عدد الوثائق وبالتالي يحسب log(1000000/1000)=3.

ويحسب معدل التردد العام للوثيقة من خلال حاصل ضرب تردد المصطلح X تردد الوثيقة المعكوس المعادلة الاتية:

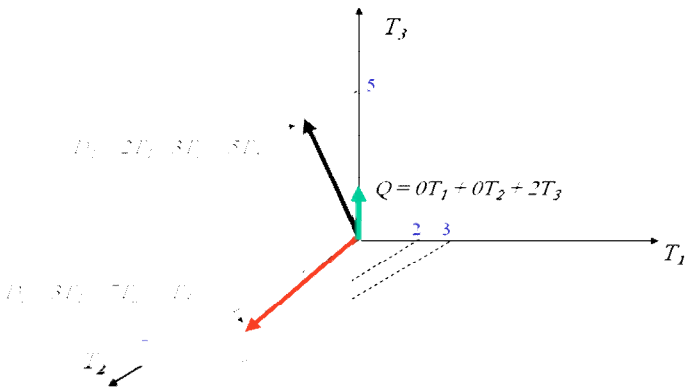
$$d_t = TF(d,t) * IDF(t)$$

ومن خلال المثال السابق تكون المعادلة 0.04X3=0.12 اي ان رتبة الوثيقة يساوي 0.12، ولعل من الملاحظ ان اجمالي القيم ستأتي منحصرة بين رقمي 1 و 0.

$$TF(d,t) = \begin{cases} 0 & \text{if n(d, t) =0} \\ 1 + log(1 + log(n(d,t))) & \text{otherwise} \end{cases}$$

وعليه يحسب جيب الزاوية الخاصة بالتشابه بين الوثيقة والاستفسار من خلال المعادلة الاتية:

$$sim(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \, \|\mathbf{q}\|} = \frac{\sum_{i=1}^{N} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} * \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$



شكل رقم (13) توضح المنهجيات المختلفة لحساب جيب الزاوية بين الأستفسار والوثيقة.

ان عيوب هذا النموذج تمثل تحديا كبيرا في استرجاع المحتوى في محركات البحث والتي تتمثل في اعتماده وبشكل أساسي على المضاهاة المعجمية Lexical، بحيث يقوم باسترجاع محتوى الوثائق التي تستخدم الكلمات الكشفية التي وردت في استفسار المستفيد، واعتباره اكثر محتوى الوثائق صلة بالموضوع، وعليه يكتنف الاسترجاع وفقا للمضاهاة اللغوية على مشكلتين اساسيتن:

- الترادف اللغوي Synonym: تنبع من امكانية استخدام نفس المفهوم ولكن بعبارة او كلمة اخرى فالبسؤال مثلا عن المساجد لاتظهر النتائج التي تشتمل على لفظة الجوامع.

- التجانس اللغوي Polysemy**:** اما التجانس فينبع أن الكلمة الواحدة قد تحظى باكثر من معنى في سياقات مختلفة ومثالا على ذلك قد يسترجع المحتوى عن موضوع القروض ومحتوى اخر يشتمل على اسم صلاح الدين ومحتوى اخر عن الدين الاسلامي نظرا لاستخدام لفظة الدين.

وعليه فإن استخدام المضاهاة اللغوية يؤدي إلى استرجاع محتوى قد لايتصل باستفسار المستفيد، فضلا على تعامله مع الكلمات في صورة فردية دون النظر إلى السياق[3].

نموذج المنهج الإحتمالي Probabilistic model:

قدم المنهج الاحتمالي لإول مرة في مضمار استرجاع المعلومات عام 1960 على يد Maron and Kuhns في مقالتهما المعروفة بأسم On Relevance, Probabilistic Indexing and Information Retrieval كأول عمل علمي يتطرق إلى استخدام المنهج الاحتمالي في استرجاع المعلومات، وعليه ظهر ما يعرف بالتكشيف الاحتمالي Probabilistic Indexing[4]. أما نموذج الاسترجاع الخاص بهذا المنهج فقد قدم في عام 1976 على يد كلا من S. E. Robertson and K. Sparck Jones[5] والذي عرف بأسم نموذج الاسترجاع القائم على الفصل الثنائي The binary independence retrieval model **(BIR)**، وتمثل مفهوم هذا النموذج في أن مجموعة الوثائق المختزنة في نظام استرجاع المعلومات تنقسم إلى مجموعتين ثنائيتين مستقليتين عن بعضهما البعض، المجموعة الاولى تعرف بمجموعة الصلة والتي يتسم محتواها بالصلة بالاستفسار، والمجموعة الاخرى تعرف بمجموعة اللاصلة والتي يتسم محتواها بعدم الصلة بالاستفسار.

تمثل جوهر هذا النموذج في سؤال منطقي وهو **"ماهو احتمالية صلة وثيقة محددة باستفسار محدد؟!"** من خلال هذا السؤال تبلورت رؤية هذا النموذج في قياس وتحديد الوثائق وفقا لإحتمالية صلتها بالاستفسار.

إن الفكرة الاساسية لهذا النموذج تتمثل في فرضية أحتمال أن نظام استرجاع المعلومات يشتمل على وثائق تتصل باستفسار المستفيد تمام الصلة وهناك مجموعة اخرى بمنأى عن هذة الصلة، فوفقا لهذا النموذج تسمى مجموعة الوثائق ذات الصلة بمجموعة الجواب المثالي *ideal* answer set ، وبتوفير توصيف كامل لهذة المجموعة من الوثائق (مجموعة الجواب المثالي) تتضاءل مشاكل استرجاع محتوى الوثائق، ورغم ذلك تظهر عقبة اخرى في صعوبة معرفة ماهية هذة الخصائص والسمات بشكل قاطع.

ما يمكن ان يستشف من معالجة الإستفسار يتمثل في الكلمات المفتاحيةالواردة في محتوى الوثيقة والتي تحظى بدلالات لغوية او اصطلاحية يمكن ان تستخدم في وسم هذة الخصائص والمميزات.

[3] Dominich, S. (2008). The modern algebra of information retrieval . Berlin: Springer.
[4] Maron, M. E., & Kuhns, J. (1960). On relevance, probabilistic indexing and information retrieval. Journal of the Association for Computing Machinery, 7(3), 216–244.
[5] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Sciences,* 27(3):129-146, 1976.

Cybrarians Journal
ISSN 1887-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 30، ديسمبر 2012

رغم ذلك فأن هذة الخصائص والسمات لاتكن معروفة الا في وقت الاستعلام أو الاستفسار.حيث يتمثل الجهد الاساسي في التخمين الاولي لما يمكن أن يكون خصائص وسمات للوثيقة مما يسمح بانشاء وصف أولي احتمالي لمجموعة الجواب المثالية على الأستفسار من الوثائق.

ثم تأتي الخطوة الثانية متمثلة في تفاعل المستفيد مع النتائج بغرض تحسين الوصف الإحتمالي لمجموعة الأجابة المثلى، هذا التفاعل يمكن أن يجري على هذا النحو:

يقوم المستفيد بمراجعة متفحصة للوثائق المسترجعة، ثم تحديد أي منها ذو صلة بالاستفسار وأي منها لا يتصل بها، ثم يقوم النظام باستخدام هذة المعلومات لتنقيح وصقل الوصف الخاص بمجموعة الجواب المثلى، وبتكرار هذا الاجراء لعدة مرات يتوقع ان هذا الوصف سوف يتطور ويصبح اكثر صلة وقرابة إلى الوصف الحقيقي لمجموعة الاجابة المثلى.

تتخذ معادلة النموذج الاحتمالي هذا الشكل:

- q للاشارة إلى استفسار المستفيد.
- d_j للاشارة إلى الوثائق في نظام استرجاع المعلومات.
- R للاشارة إلى مجموعة الاجابة المثلى.
- $\tilde{R}$ للاشارة إلى الوثائق التي لا صلة لها
- $P(R|d_j)$ للاشارة إلى احتمالية ظهور مجموعة الاجابة المثلى من اجمالي وثائق النظام.
- $P(\overline{R}|\vec{d}_j)$ للاشارة إلى احتمالية ظهور الوثائق ذات عدم الصلة ضمن اجمالي وثائق النظام.
- sim للاشارة إلى حساب درجة التشابة والتماثل.

بحيث تتخذ المعادلة هذة الصورة لحساب احتمالية صلة الوثائق او عدم صلتها:

$$sim(d_j, q) = \frac{P(R|\vec{d}_j)}{P(\overline{R}|\vec{d}_j)}$$

ووفقا لقاعدة Bayes [6] تتخذ المعادلة هذة الصورة:

$$sim(d_j, q) = \frac{P(\vec{d}_j|R) \times P(R)}{P(\vec{d}_j|\overline{R}) \times P(\overline{R})}$$

أن ما يعيب هذا النوذج ويمثل تحديا في استرجاع وترتيب المحتوى في محركات البحث يكمن في:

- الحاجة إلى التخمين للفصل الأولي لمحتوى الوثائق في مجموعتين مجموعة ذات صلة ومجموعة لاتتسم بالصلة.

[6] تعد هذة النظرية احد نظريات مجال الاحتمالات في علم الرياضيات والتي تعني بقياس العلاقة بين احتمالين شرطيين والذي عادة ما يعلوهما التناقض فيما بينهم.

- إن هذة الطريقة لاتأخذ في الإعتبار وتيرة تردد المصطلحات الكشفية داخل الوثائق (مما يرجعنا إلى نظرية الحكم الثنائي الخاصة بالنموذج البوليني).[7]

نموذج التكشيف الدلالي الكامن: Latent Semantic Indexing Model :

إن نماذج الاسترجاع - السابق ذكرها - تعتمد وبشكل اساسي على الاسترجاع من خلال اختزال محتوى كلا من الاستفسار والوثيقة في مجموعة من المصطلحات الكشفية أو الكلمات المفتاحية.

ثم يعمد بعد ذلك إلى قياس درجة التشابه بين كلا منهما، ثم الاسترجاع وفقا لهذا الاساس، وتعرف هذة المنهجية بإسم الاسترجاع وفقا للتشابه المعجمي lexical matching method أي من خلال الاعتماد على المضاهاة بين الاحرف المكونة للمصطلحات الكشفية الوارة في كلا من الوثيقة والاستفسار ، ووفقا لهذة المنهجية السالفة يتسم اداء ومعدل الاسترجاع في محركات البحث بالفقر، وعدم الدقة، وهي الأسباب وراء تدني معدلات الاسترجاع والترتيب والمرجعية في ذلك تعود إلى سببين:

- ففي العادة تتوافر العديد من المصطلحات التي يمكن ان يستخدمها المستفيد في التعبير عن حاجته البحثية او المفهوم المجرد لاستفساره، وهذة القضية تعرف بترادف المصطلحات فمن الممكن ان يقوم المستفيد بالتعبير عن مفهوم مجرد بلفظة **الجوامع** مثلا في حين أن مستفيد اخر يعبر عنها بلفظة **المساجد**، في حين أن كلا من اللفظين يشيرا إلى مفهوم واحد وهو دور العبادة لإقامة الصلاة عن المسلمين، ووفقا للمنهجية المضاهاة المعجمية فإن بعض الوثائق ذات الصلة التي لم تكشف وفقا للمصطلحات الكشفية لدى المستفيد قد لاتسترجع.

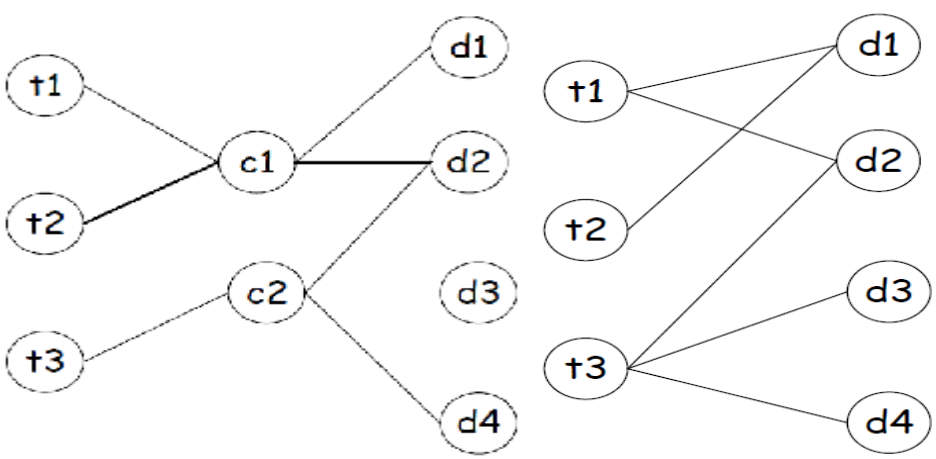
- قد تسترجع العديد من الوثائق التي لاتتصل بالاستفسار داخل مجموعة الوثائق المسترجعة، والمرجعية في ذلك تعود إلى ان العديد من البنى اللفظية للكلمات تحمل في طياتها الكثير من البنى الدلالية أو ما يعرف بالتجانس اللفظي، وعليه فإن المصطلحات الكشفية في استفسار المستفيد سوف تضاهي وفقا للمضاهاة المعجمية مجموعة من الوثائق تشتمل على نفس المصطلحات الكشفية ولكن لاتتصل بموضوع الاستفسارمثل البحث عن لفظة **الدين** فقد ينطوي الأمر على ان تسترجع وثائق موضوعها الديانات، واخرى موضوعها القروض والاقتراض، واخرى تحتوى لفظة الدين كاسم شخص كأسم صلاح الدين او عماد الدين.

فالافكار التي ترد في نصوص الوثائق أاقرب ان توصف من خلال المفهوم بدلا من أن توصف من خلال الالفاظ، ومن ثم فإن تحقيق الاسترجاع وفقا للمفهوم أو المعنى يعد أقرب صلة باستفسار المستفيد، وعلية فإن عملية مضاهاة الوثائق

[7] A probabilistic model of information retrieval: development and comparative experiments. K. Sparck Jones. Information Processing and Management 36 (2000) 779±808

بالاستفسار المقدم يجب أن تعتمد على المضاهاة وفقا للمفهوم بدلا من أن تعتمد على المضاهاة وفقا للفظ مما يسمح باسترجاع الوثائق التي لم تكشف وفقا لمصطلحات الكشفية للاستفسار.

عرفت هذة المنهجية بأسم التكشيف الدلالي الكامن Latent semantic indexing وقد قدم هذا النموذج كمهنجية آلية لمعالجة بعض أوجه القصور المتأصلة في تقنيات التي تعتمد على المضاهاة المعجمية والاسترجاع وفقا للكلمات المفتاحية في الوثيقة، يعتمد النموذج الدلالي الكامن على تحليل درجة العلاقة الدلالية بين محتوى الوثائق من خلال معدلات احصائية، فوفقا لهذة المنهجية فأن محرك البحث يقوم باسترجاع محتوى الوثائق ذات الصلة بمفهوم الاستفسار وليس وفقا لتشابه الكلمات المفتاحية حتى ولو لم تشتمل هذة الوثائق على مصطلحات الاستفسار كما هو موضح في الشكل رقم (14):



شكل رقم(أ). شكل رقم (ب).

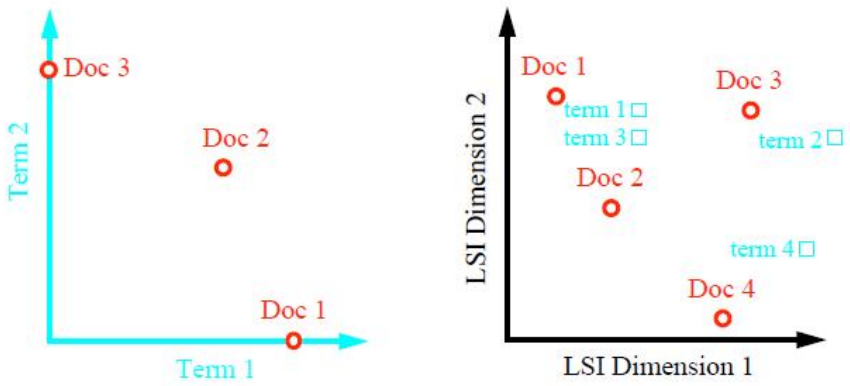
شكل رقم (14) يوضح هذا الشكل منهجية عمل نموذج التكشيف الدلالي الكامن.

يوضح هذا الشكل منهجية عمل التكشيف وفقا للمضاهاة المعجمية والتكشيف وفقا للمفهوم، حيث في الشكل رقم (أ) يتضح ان كل لفظ يقابله مجموعة من الوثائق اما في الشكل (ب) فان الوثائق ترتبط بالمفاهيم التي تدل عليها مجموعة من الالفاظ المختلفة

قدم النموذج التكشيف الدلالي الكامن في عام 1990 ضمن ورقة بحثية لمجموعة من الباحثين في معمل Bell Communications Research ("Bellcore") ضمن ورقة بحثية [8]

يعد نموذج التكشيف الدلالي الكامن نموذج حسابي يعتمد على تقنية الجبر الخطي بهدف تجميع الكيانات (المفاهيم) المتشابه معا ويتم تحقيق ذلك من خلال الاتي:

- تحديد المفاهيم الاساسية المشتركة الموجودة داخل مجموعة من الوثائق.
- تنظيم هذة المفاهيم داخل فراغ متعدد الابعاد .
- التحليل الجبري لهذة المصفوفة.



الشكل رقم (15) مقارنة بين نموذج الاسترجاع فراغ الموجهات وبين نموذج الاسترجاع الدلالي الكامن.

ففي اغلب نماذج الاسترجاع يتم رصد محتوى الوثائق وفقا للمصطلحات المعبرة عنها، أما في النموذج الدلالي فالوثائق ترصد وفقا للمفاهيم ثم تحصر الالفاظ التي تحدد موضوع الوثائق من داخلها كما هو موضح في الشكل رقم (15).

إن عيوب هذا النموذج ووالتي تعتبر تحديا في استرجاع وترتيب المحتوى في محركات البحث: هو قدرته المحدودة في التعامل مع الاعداد الضخمة من الوثائق كما هو الحال في بالنسبة إلى حجم مصادر الشبكة العنكبوتية[9].

نموذج الاسترجاع (الروابط الفائقة الناجمة عن البحث الموضوعي hyperlink-induced topic search) (HITS).

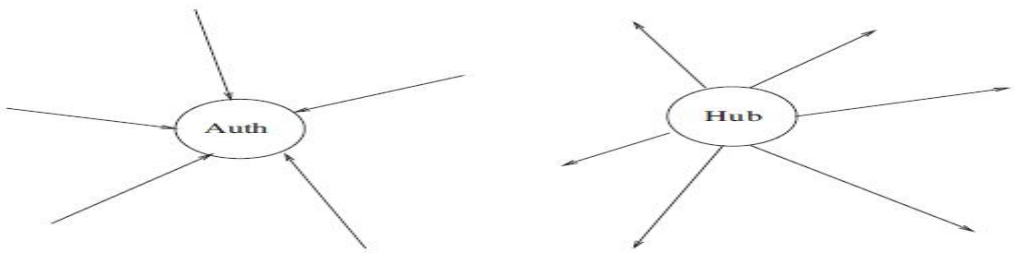
طورت هذه الخوارزمية عام 1997 على يد Jon Kleinberg في نفس الوقت الذي طورت فيه خوارزمية الترتيب الطبقي، تعتمد هذة الخوارزمية على اكتشاف وترتيب محتوى الوثيقة ذات الصلة بموضوع محدد – وتعد هذة الخوارزمية

[8] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (September,1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6): 391-407

[9] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (September, 1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6): 391-407

الآن جزءا اساسي في خوارزميات محرك البحث Ask (www.ask.com) – بمعنى ان هذة الخوارزمية تعتمد على أن يوجه المستفيد اولا الاستفسار لإداة البحث ثم تسترجع النتائج من الكشاف او قاعدة البيانات لتبدء مرحلة الترتيب للنتائج وفقا لعنصرين اساسين هما:

- المواقع الارتكازية Hubs Nodes: وهي المواقع التي تشتمل على محتوى يصدر منه الرابط في اشارة منها لمواقع الاستنادية.
- المواقع الاستنادية او ذات الموثقية Authorities Nodes: ويقصد بها المواقع التي يرد اليها الرابط من قبل المواقع المحورية كما هو موضح في الشكل رقم (16).

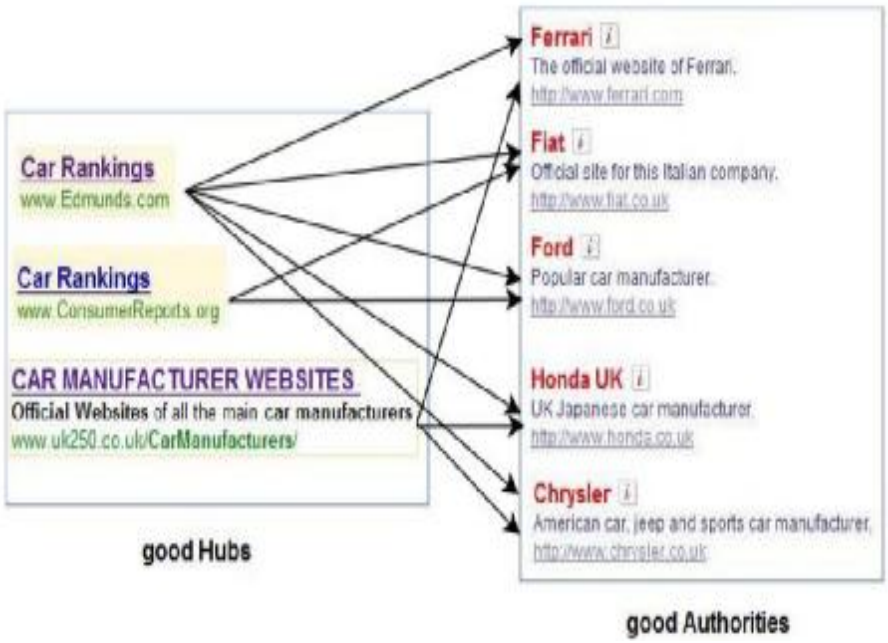


شكل رقم (16) يوضح بنية المواقع الارتكازية والمواقع الاستنادية

فلو افترضنا أن الصفحة I تشتمل على محتوى ذات موثقية authority للاستفسار المقدم لمحرك البحث عن " اشهر صناع المركبات" في ظل أنها تشتمل على محتوى قيم عن الموضوع المراد الاستفسار عنه، حيث تعد الصفحات الرسمية لمنتجي السيارات بمثابة صفحات استنادية ذات موثوقية لهذه العملية البحثية كموقع تويوتا ومرسيدس وفيات وغيرها كما تعد مواقع وصفحات وكلاء المبيعات لهذة السيارات بمثابة صفحات استنادية ايضا للموضوع المشار اليه، على الجانب الاخر كيف يتم تحديد هذة المواقع الموثوقية للاستفسار في ظل عدم اشتمالها على كلمة مركبات وفي ظل عدم اشتمال محرك البحث على مكنز او خريطة انطولوجية لحصر المترادفات والمفردات.

يتم تحديد المواقع الموثوقية من خلال تتبع الروابط من نقاط ارتكازية محددة تكون بمثابة دليل موثوق به لمحرك البحث ثم يتم تتبع الروابط التي يشير فيها لمواقع اخرى، فلو افترضنا ان الصفحة I بمثابة صفحة ذات موثقية authority للاستفسار المقدم لمحرك البحث عن " اشهر صناع المركبات" في ظل أنها تشتمل على معلومات قيمة عن الموضوع المراد الاستفسار عنه، حيث تعد الصفحات الرسمية لمنتجي السيارات بمثابة صفحات استنادية ذات موثوقية لهذه العملية البحثية كموقع تويوتا ومرسيدس وفيات وغيرها كما تعد مواقع وصفحات وكلاء المبيعات لهذة السيارات بمثابة صفحات استنادية ايضا للموضوع المشار اليه، على الجانب الاخر كيف يتم تحديد هذة المواقع الموثوقية للاستفسار في ظل عدم اشتمالها على كلمة مركبات وفي ظل عدم اشتمال محرك البحث على مكنز او اخريطة انطولوجية لحصر المترادفات والمفردات؟ يتم تحديدها

من خلال تتبع الروابط من نقاط ارتكازية محددة تكون بمثابة دليل موثوق به لمحرك البحث كما هو موضح في الشكل رقم 17:



شكل رقم (17) يوضح المثال السابق

وتتمثل العلاقة بينهما في علاقة تبادلية تعزز بعضهم البعض، بمعنى أن المواقع الارتكازية ذات الجودة العالية تشير إلى المواقع الاستنادية الجيدة والعكس ايضا. وعليه تعمل هذة الخوارزمية على تحديد درجة خاصة لإرتكازية الوثيقة ودرجة اخرى لإستنادية نفس الوثيقة.

وعلى هذا فان كل محتوى يحظى بدرجتين واحدة للمواقع التي تشير اليها - وهي في هذة الحالة نقطة ارتكازية - واخرى للمواقع التي تشير إلى هذة الصفحة، ويحدد على اساسهما رتبة الموقع في قائمة النتائج:

وعلية تحسب درجة الموثوقية والنقطة الارتكازية للوثيقة p على النحو الاتي:

$$\sum_{i=1}^{n} hub(i)$$

لحساب قيمة موثوقية الصفحة

$$\sum_{i=1}^{n} auth(i)$$

لحساب قيمة النقاط الارتكازية للصفحة

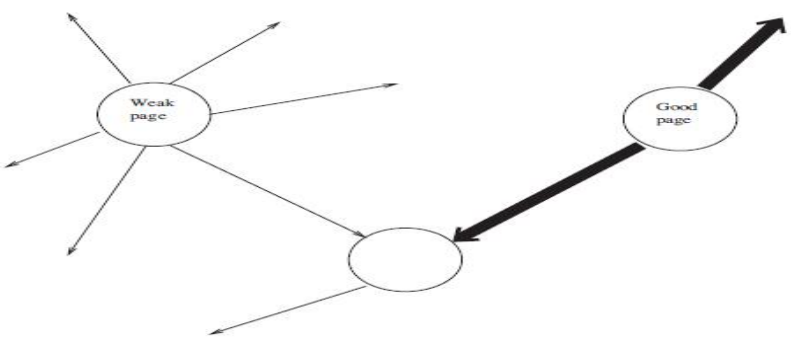
حيث تشير n إلى مجموع عدد المواقع التي ترتبط بالصفحة p , أما i فتشير إلى الصفحة المرتبطة ب p بشكل مباشر[10].

نموذج الاسترجاع (ترتيب الطبقي للصفحات PageRank).

طور هذا النموذج على يد كلا من Sergey Brin و Lawrence Page عام 1997، وقد عرف هذا النموذج ب" بانها المنهجية التي تعني بحساب رتبة محتوى كل صفحة على العنكبوتية اعتمادا على نمذجة العنكبوتية في مخطط بياني قائم على الروابط والمواقع". ولقياس جدوى هذة الخوارزمية قاما كلا من Brin و Page بتصميم محرك البحث الشهير Google.

أن الجانب الذي التفتت اليه هذة الخوارزمية هو النظر إلى الكيف دون الكم، بمعنى الاخذ في الاعتبار جودة الروابط بدلا من النظر إلى عدد الروابط ، فتستند هذة الخوارزمية على مبدئين اساسين هما:

- تمثل الروابط مؤشرات جيدة لتحديد اهمية محتوى الوثيقة التي تشير اليها.
- الروابط الصادرة من وثائق تحظى باهمية في موضوعها تعد مؤشرا جيدا لجودة الوثيقة التي تشير اليها، عن الوثيقة التي يشار اليها من قبل وثائق اقل في الاهمية والجودة كما هو موضح في الشكل رقم (18).



شكل رقم (18) يوضح منهجية PageRank ويوضح دلالة ان الرابط الفائق يكتسب قوة من قوة الصفحة ومحتواها

الخوارزمية:

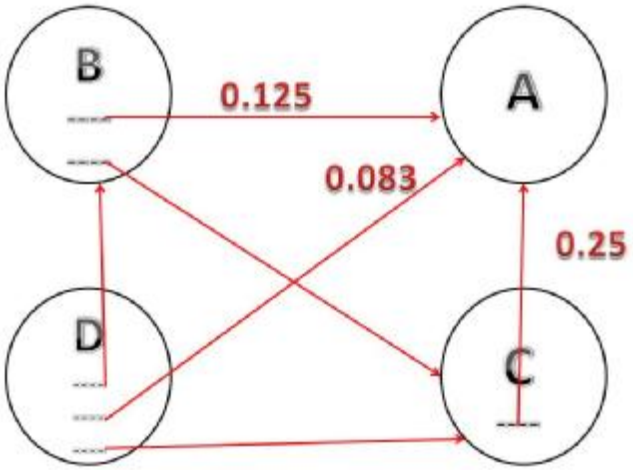
بداية يجدر الاشارة إلى ان خوارزمية الترتيب الطبقي (the pagerank) صدرت في اكثر من صيغة واكثر من معادلة متتالية، وسيتعرض الدراسة في هذا المقام إلى الصيغة البسيطة من هذة المعادلات.

[10] Levene, M. (2010). An introduction to search engines and web navigation (2nd ed.). Hoboken: Wiley.

Cybrarians Journal
ISSN 1887-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 30، ديسمبر 2012

تعتمد خوارزمية الترتيب الطبقي على نظرية احصائية تعرف بنظرية التوزيع الاحتمالي والتي تعمل على احتمالية تحديد قيمة لمتغير ما (كصفحة عنكبوتية اوالرابط الفائق) تم اختياره عشوائيا، هذة القيمة في هذا المقام هي الاهمية والتي يمكن ان تتروح ما بين قيمتين اساسيتين هما (0 & 1).

ولنفترض ان لدينا بنية بيانية لشبكة عنكبوتية تتكون من 4 نقاط (NODES) (أربع صفحات) A,B,C and D وأن أهمية هذة الصفحات تتوزع بالتساوي بينهم – أي تقسيم رقم 1 الذي يشير إلى وجود اهمية للبنية البيانية للشبكة بالتساوي – فيكون نصيب كل صفحة هو 0.25، ولنفترض أن بين هذة الصفحات مجموعة من الروابط والتي سيتم الاعتماد عليها لحساب رتبة الصفحة A، هذة الروابط تتمثل في الشكل رقم (19).



شكل رقم (19) يوضح كيفية حساب رتبة الصفحة من خلال الرابط

حيث تشير كلا من الروابط الموجودة في الصفحات B,C,D إلى الصفحة A كما هو موضح في الشكل، مع الآخذ في الاعتبار أن الصفحة A تحسب رتبتها من خلال قيمة الرابط الذي يشير اليها فإذا كانت الصفحة B تحظى بقيمة مقدارها 0,25 موزعة هذة القيمة على رابطين فإن قيمة الصفحة B بالنسبة إلى الصفحة A هي: 0.25/2=0.125 وتحسب قيمة الرتبة A من خلال المعادلة الاتية:

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}.$$

$$PR(A) = \frac{0.25}{2} + \frac{0.25}{1} + \frac{0.25}{3} = 0.458$$

ومن ثم تكون رتبة الوثيقة A في موضوع تخصصها 0.458 على صعيد الشبكة العنكبوتية.

التحديات التي تواجه خوارزميات page rank & HITS في ترتيب النتائج المسترجعة.

ورغم ذلك تمثل التحدي الذي واجه محركات البحث في تطبيق هذة الخوارزمية في:

1. هذة الخوارزمية في الاساس تعتمد على نمذجة وثائق العنكبوتية وروابطها في صورة تعرف بالمخطط البياني للعنكبوتية والذي ينطوي على انه قد يتخذ اكثر من صورة وهيئة وشكل مما قد ينطوي على تبديل الادوار في النقاط الارتكازية والمواقع ذات الموثوقية.

2. تعتمد هذة الخوارزمية على فرضية اساسية مفادها ان الصفحات المرتبطة ببعضها عبر الروابط الفائقة تنتمي إلى نفس الموضوع او بدرجة شبيهة منه، مع الاخذ بالاعتبار ان بيئة العنكبوتية لاتتسم بالاستقرار الكامل في ما تحمله وما تطرحه المواقع من محتوى وموضوعات، فضلا عن التغذيات المرتدة التي قد تنأى بالموضوع الاساسي إلى موضوعات اخرى مما ينطوي ايضا على زيادة في عدد الروابط الفائقة او قد تتواجد بعض الروابط التي تشير إلى مواقع تتناول موضوعات اخرى وهو ما يعرف بقضية توليد الموضوعات topic generalized ومن ثم فان فرضية دلالة الربط وفقا لتشابه المواقع ذات الموضوع الواحد تتضاءل.

3. ان الاعتماد على تحليل الروابط الفائقة ينطوي على ما يعرف بظاهرة الانحراف الموضوعي *topic drift* في ظل ما توفره الروابط من قدرة على الابحار[11].

ويعد التحدي الاكبر الذي يواجه الترتيب وفقا للروابط الفائقة هو خادعات الروابط والتي ستتناول في قسم التحديات الخارجية لمحركات البحث في استرجاع المحتوى على العنكبوتية.

التحديات الخارجية لمحركات البحث في أسترجاع المحتوى من على العنكبوتية:

4/1- خادعات المحتوى لمحركات البحث:

يمكن القول أن بعض التحديات والمشكلات التي تواجهها محركات البحث في أسترجاع المحتوى ارتبط ظهورها بوجود الشبكة العنكبوتية والبعض الاخر كان بمثابة تغيرات مستحدثة لمشكلات حظيت من قبل بالدراسة والاهتمام العلمي في ادبيات مجال استرجاع المعلومات أي قبل ظهور محركات البحث.

[11] Langville, A. N., & Meyer, C. D. (2006). Google's PageRank and beyond: the science of search engine rankings. Princeton, N.J.: Princeton University Press.

اتسم نمط سلوك المستفيدين من محركات البحث إلى اللجوء والاعتماد بصورة قصوى على النتائج الاولى داخل الصفحات الاولى من نتائج محركات البحث وقصر التصفح عليها دون النظر إلى باقي النتائج في الصفحات الاخرى فقط أوضحت Silverstein ان 85% من عمليات البحث يقتصر تصفح نتائجها على الصفحة الاولى فقط ولعل المرجعية في ذلك إلى قناعة ذاتية من جانب المستفيدين في أن النتائج ذات الصلة باستفسارة لابد ان تظهر اولا، وانه كلما توغل في النتائج الاخرى ابتعدت به عن مجال استفساره[12].

وعليه ادرك مديرون المواقع والقائمين على إدارة صفحات العنكبوتية تبعية منطقية مفادها في أن تضمين المواقع داخل النتائج العشر الاولى يؤدي إلى ما يعرف بإرتفاع معدل المرور إلى محتوى الموقع traffic to web site[13] ، وعلى النقيض فإن استثناء أو استبعاد المواقع من الشاشة الاولى أو النتائج الاولية للبحث يسمح لعدد محدود من المستفيدين من رؤية محتوى الموقع أو تصفحه.

ويعد هذا الامر هو بؤرة أهتمام المواقع ذات التوجه التجاري – والتي تعتمد في دخلها على أرتفاع معدلات المرور اليها – حيث تعمل على ان تظهر في أول عشر نتائج للاستفسار، مما يساعد على تعزيز من سمعة الموقع والتي بدورها يساعد على جعل المزيد من المواقع تشير اليه وبالتالي تساعد على توليد مستوى أعلى في النتائج ومن ثم الحصول على القدر الاوفر من تضمين الاعلانات والدعاية داخلها.

وبالتالي يعمل أغلب موفري المحتوى على العنكبوتية – خاصة المواقع التجارية – على جعل محتوى صفحات مواقعهم تحظى بترتيب طبقي عالي في نتائج محركات البحث داخل محركات البحث العامة والاكثر استخداما . ويتم تحقيق ذلك عن طريق منهجية واضحة ومباشرة وهي تحسين جودة صفحات الموقع سواء كان في المحتوى أو الشكل ولكن تكتنف هذة الطريقة الكثير من المال والوقت والجهد والموارد الأخرى، وعوضا عن كل ذلك يلجأ بعض موفري المحتوى إلى طرق اخرى مختصرة تتمثل في التلاعب في نتائج محركات البحث من خلال استخدام اساليب غير اخلاقية عند بناء محتوى المواقع او تصميم الصفحات، فيما تعرف هذة المحاولة بتضليل خوارزمية الترتيب الطبقي لنتائج محركات البحث بخداع محركات البحث search engine spam.

[12] Craig Silverstein ,Hannes Marais ,Monika Henzinger,Michael Moricz. Analysis of a very large web search engine query log. ACM SIGIR Forum, 1999 - portal.acm.org

[*] ظاهرة المرور إلى المواقع Web traffic تعرف بانها حجم البيانات التي ترسل وتستقبل من قبل زائري المواقع والتي تقدر وفقا للمعادلة بحساب عدد الزائرين وعدد الصفحات التي تم تصفحها من خلالهم ويعد هذا المقياس مؤشر مهما للقائمين على ادارة المواقع في تحديد جدوى صفحاتهم.

Cybrarians Journal
ISSN 1887-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 30، ديسمبر 2012

المفهوم والتعريف:

يعود تاريخ اللفظة في الانتاج الفكري إلى عام 1996 حينما صك Eric Convey هذا المصطلح في مقالته [14] Porn sneaks way back on Web ،موضحا ان سمة مشكلة تظهر في نتائج محركات البحث، وان مرجعية هذة المشكلة تعود إلى ان مديرو المواقع الذين يلجأون إلى اضافة مئات المصطلحات في بنية محتوى وثائقهم للحصول على رتبة اعلى وقد اطلق على هذة العملية مسمى Spamdexing في اشارة منه إلى خداع كشافات محركات البحث.

ويمكن أن يعرف خداع محركات البحث ب "المنهجية او اسلوب الذي يعمد إلى استخدام بعض الاليات المصممة عمدا لرفع ترتيب المواقع او الصفحات في نتائج محركات البحث"[15].

إن الهدف الاساسي لمحركات البحث هو توفير نتائج عالية الدقة والجودة عبرالتحديد الكامل والدقيق لكافة الوثائق التي تتصل موضوعيا باستفسار البحث.[16] والعمل على ترتيب النتائج التي تتسم بالاهمية عن مثيلاتها وفقا لدرجات الصلة الموضوعية.

فمفهوم الصلة الموضوعية يشير إلى التشابه النصي بين الاستفسار المقدم والوثيقة المسترجعة. اما عن مفهوم الاهمية فهو يشير إلى قوة الوثيقة وعالميتها ودرجة تمتعها برواج وفقا لمجالها كما يستدل ايضا على اهميتها من واقع الروابط الفائقة (فعلى سبيل المثال الوثيقة التي تشتمل على الكثير من الروابط الداخلية تحظى باهمية عن مثيلاتها).

وعمليا تقوم محركات البحث بالدمج بين معيار الصلة الموضوعية ومعيار الاهمية كمؤشر لترتيب النتائج المتعلقة بالاستفسار في صورة خوارزمية خاصة بمحرك البحث. ولتحقيق هذة الاهداف يلجأ مخادعوا محركات البحث spammers إلى دراسة خوارزميات الترتيب الطبقي لمحرك البحث بعناية لايجاد سبل الافادة منها وبشكل اكثر دقة يعمل المخادعون إلى تحديد المؤشرات والعوامل والمقاييس المستخدمة في نظم الاسترجاع والترتيب الطبقي لمحركات البحث بصورة اقرب مما يقومون به قراصنة الحاسب في التعامل مع فرائسهم من الشبكات والنظم والحواسب مما يجعل خداع محركات البحث احد اهم التحديات التي تواجه محركات البحث فنتائج محركات البحث تتاثر بشكل كبير وخطير مما قد يستتبع ان يفقد محرك البحث مصداقية نتائجه لدى المستفيدين وبالتالي عدم ثقة المستفيد في هذا المحرك مرة اخرى.

اوجة الضرر التي قد تلحقه الخادعات بمحركات البحث فتتمثل في :

- تصعيب مهمة المستفيدين في ارضاء حاجتهم المعلوماتية information needs.
- ارساء خلفية سيئة تجاه محرك البحث ونتائجه وعدم مصداقيته.

---

[14] Eric Convey.Porn sneaks way back on Web.The Boston Herald, May 22, 1996

[15] Gyongyi, Z., & Garcia-Molina, H. (n.d.). Web Spam Taxonomy. Web Spam Taxonomy. Retrieved July 21, 2011, from air-web.cse.lehigh.edu/2005/gyongyi.pdf

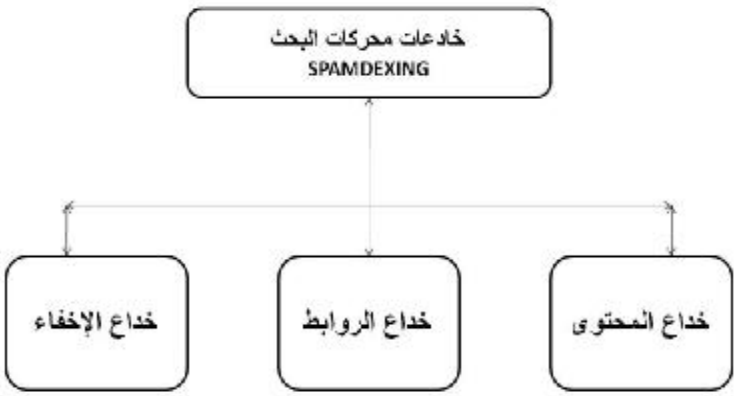
[16] Gyongyi, Z., & Garcia-Molina, H. (n.d.). Web Spam Taxonomy. *Web Spam Taxonomy*. Retrieved July 21, 2011, from air-web.cse.lehigh.edu/2005/gyongyi.pdf

- تلويث كشاف محركات البحث بمواقع زائفة.
- حرق النطاق الترددي للزواحف او ما يعرف بBandwidth.
- تشوية نتائج محركات البحث.
- قدرة الخادعات على توليد العديد من شبيهاتها.
- وقد يمتد الضرر إلى اصحاب المواقع، فقد يدفع مدراء المواقع إلى التخلي عن جهودهم في توفير صفحات ومحتوى ذو جودة مرتفعة، اخذين على عاتقهم مهمة انشاء محتوى يستطيع ان يتعامل مع محركات البحث مما يستتبع ان يجعل محتوى الشبكة العنكبوتية موجه إلى محركات البحث بدلا من ان يوجه إلى مستفيدي الشبكة العنكبوتية الامر الذي يشكل كارثة في نهاية المطاف على الشبكة العنكبوتية[17].

فئات خادعات محركات البحث:

يمكن ان تصنف خادعات محركات البحث من خلال تقنياتها وأهدافها إلى:

- خداع محتوى المواقع.
- خداع روابط الصفحات.
- الخداع معتمد على اخفاء او تضمين الصفحات.



خداع المحتوى:

تعتمد هذة الفئة على تغيير محتوى صفحات العنكبوتية للحصول على رتبة أعلى في نتائج محركات البحث وتعد اشهر تقنيات الخداع المعتمد على المحتوى هي تكرار الكلمات المفتاحية داخل صفحات العنكبوتية، واستخدام الكلمات غير المتصلة بسياق او موضوع الصفحة او اضافة قاموس كامل من المصطلحات في نهاية الصفحة سعيا به للأن يكشف في

[17] Detecting Spam Web Pages. Marc Najork. Microsoft Research, Silicon Valley

محركات البحث ومن ثم تسكينها وفقا للمصطلحات والكلمات الواردة في القاموس داخل قاعدة بيانات محرك البحث ومن ثم استرجاع الصفحة في كافة عمليات البحث في ظل الاعتماد على المماثلة او المشابهة في الاسترجاع.

دورالمصطلحات والالفاظ في خداع المحتوى:

يجدر في البداية إلى اشارة ان كافة اشكال المحتوى في صفحات ومواقع العنكبوتية يتم تمثيلها وفقا للفظ والمصطلح فالصورة والملفات الفيديو والملفات ثلاثية الابعاد وغيرها يتم تكويدها وفقا للغة النص التشعبي الفائق HTML تكود في صورة نصية أي في صورة الفاظ ومصطلحات طبيعية او برمجية ولايقتصر الامر على هذا فحسب بل ان تكشيف الوثائق (مهما كان محتواها نصي او صوري او ملف فيديو) داخل محركات البحث يتم وفقا للنص.

إن محركات البحث تلجأ في تقييم درجة الصلة الوثائق بالاستفسار إلى تحديد المصطلحات الواردة في الاستفسار داخل صفحات المواقع، او مايعرف بالمضاهاة، وغالبا ما يكون ظهور هذة المصطلحات في اماكن محددة داخل محتوى الوثيقة ، تعرف هذة الاماكن وفقا لخوارزميات محركات البحث بأسم الحقول Fields وتعد اشهر الحقول داخل الوثيقة هي حقل جسم الوثيقة والعنوان وواصفات البيانات والمحدد الفريد للموقع وتعد هذة الحقول بمثابة مفاتيح محتوى الوثيقة ومحدداتها، وبالتالي تلجأ محركات البحث إلى تحديد درجة الصلة بين الاستفسار والوثيقة من خلال محتوى هذة الحقول، ومن ثم تعد هذة الحقول ايضا بمثابة الصيد الثمين لخادعات محركات البحث بل تمثل النقاط الحيوية للتلاعب.

خوارزميات خداع محتوى المواقع:

إن اغلب خوارزميات محركات البحث المستخدمة في ترتيب النتائج تعتمد وبشكل اساسي على استخدام انماط مختلفة من المعادلات والحسابات المختلفة ولكن تعد خوارزمية tf-idf القاسم المشترك بين اغلب هذة الخوارزميات وتحسب العلاقة بين الوثيقة والاستفسار من خلال المعادلة الاتية والتي سبق الاشارة اليها:

$$tf\text{-}idf(p,q) = \sum_{t \in p, t \in q} tf(t) \cdot idf(t)$$

وفي هذا يعتمد الخادعون spammers على هذة المعادلة لتحقيق هدفين اساسين اولهما: جعل صفحاتهم تتصل بعدد كبير من الاستفسارات وذلك من خلال جعل صفحاتهم لا تحصل على الرقم صفر كناتج لهذة المعادلة ويتم تحقيق ذلك من خلال تضمين مجموعة كبيرة وضخمة من المصطلحات بنية الوثيقة.

اما الهدف الثاني فهو فيتمثل في جعل الوثيقة تكثر صلة باستفسار محدد وبالتالي تحظى برتبة عالية في نتائج محركات البحث من خلال ان تحصل على درجة عالية في المعادلة السابقة تكاد تصل إلى رقم 1 ويتم تحقيق ذلك من خلال تكرار مصطلحات محددة في بنية الوثيقة او في الحقول التي سبق الاشارة اليها.

تقنيات خداع المحتوى:

- خداع حقل جسم الوثيقة document body:

في هذا الحقل يتم تكرار المصطلحات الخادعة في المضمنة في جسد الوثيقة.

- خداع حقل العنوان title:

مؤخرا، تعمل خوارزميات محركات البحث على اعطاء اعلى وزن للمصطلحات التي ترد في عنوان الوثيقة وبالتالي يعمل الخادعون على تضمين المصطلحات الخادعة في عنوان الوثيقة.

- خداع حقول واصفات البينات meta tags:

تظهر حقول واصفات البيانات في راس الوثيقة التي تم تكويدها بلغة النص الفائق التشعبي شكل رقم 2 والتي تعد بمثابة هدف اساي ورئيسي للخادعات ونظرا لذلك استغنت محركات البحث في الوقت الراهن عن اعطاء وزن مرتفع للكلمات التي تظهر في حقول واصفات البيانات بل وصلت إلى انها تتجاهلها تماما

<meta name= “keywords” content=”money, bank, finale”>

- خداع حقل العنوان الموحد للموقع URL:

يعد محدد الموقع بمثابه محدد فريد لدرجة صلة الوثيقة بالاستفسار ورغم ذلك تخلت محركات البحث في الاعتماد عليه نظرا لمتداد يد الخادعات لتشملها، حيث يعمل الخادعون على انشاء محدد مواقع طويل يشتمل على مصطلحات خادعة وبالتالي حينما يتم تكشيفه تدرج هذة المصطلحات في كشافات محرك البحث ومن ثم ظهور نتائج لاتتصل باستفسار المستفيد.

- تكرار الكلمات المفتاحية Repeated keywords:

حيث يتم في هذة التقنية تكرار الكلمات المفتاحية عدة مرات في الوثيقة، وتعد هذة التقنية فعالة جدا في ظل اعتماد محركات البحث على خوارزمية تعطي وزن اكبر للوثيقة باحتساب عدد مرات ظهور الكلمات المفتاحية وفي ظل اشتمال الوثيقة على نفس الكلمات مفتاحية التي وردت في استفسار المستفيد.

- اخفاء المحتوى Content Hiding:

تعتمد هذة التقنية على اخفاء بعض الكلمات المفتاحية بحيث لا تعرض حينما يقوم المتصفح بعرض الصفحة، ويتم ذلك من خلال عملية تمويه للكلمات والمصطلحات حيث يتم جعل لون هذة الكلمات بنفس لون خلفية الوثيقة، او من خلال كتابتها بصورة ضئيلة ومتناهية في الصغر.

- إغراق محتوى الوثائق بالمصطلحات الاستفسارية:

حيث تعتمد هذة التقنية على اغراق محتوى الوثيقة بالمصطلحات الخادعة والكلمات المفتاحية التي لا علاقة لها بالموضوع ويتم ذلك من خلال انشاء ما يعرف بقاموس المصطلحات الخادعة داخل الوثيقة بحيث يشتمل هذا القاموس على مصطلحات مختلفة تضاهي اي كلمات مفتاحية قد ترد في اي استفسار يوجه إلى محرك البحث مما يستتبع ان تظهر الوثيقة كنتيجة للعديد من الاستفسارات، ويتم اخفاء القاموس في نهاية الوثيقة داخل كود النص الفائق للوثيقة.

- الخداع المعتمد على واصفات الكيانات:

سبق وقد اشرنا ان محركات البحث في تكشيفها للصور وملفات الفيديو وغيرها من ملفات الوسائط المتعددة - والتي عرفت باسم الكيانات objects للتمييز بينها وبين نص الوثيقة- تعتمد النصوص الملحقة بالكيان كالنص الموجود في اسفل الصورة كتعليق عليها وتوضيح لها وهو ما يعرف في لغة النص الفائق التشعبي بحقل بديل النص Alt text وعادة لايرى المستخدم هذا الحقل الإ عندما يقف عليه بالفأرة، وتلجأ محركات البحث لهذا النص في تكشيف محتوى الصورة، مما جعل الخادعون يعملون على اضافة مصطلحات خادعة داخل بينة هذا الحقل بحيث تكشف المصطلحات الخادعة مع المصطلحات الاساسية مما يوفر فرصة في ظهور الوثيقة عند البحث عن صورة لموضوع معين.

- ناسخات المواقع:

تعمل هذة التقنية على استنساخ او كشط المحتوى الكامل للموقع بهدف انشاء مثل هذا الموقع لجمع عائدات مادية والتلاعب في رتب نتائج محركات البحث.

خداع الروابط:

تعتمد الكثير من محركات البحث على خوارزميات تحليل الروابط بين الوثائق وذلك في ظل ما تتمتع به بنية الشبكة العنكبوتية من روابط بين وثائقها، مما جعل محركات البحث تعتمد على الروابط كمؤشر في تحديد اهمية الوثيقة ودرجة الصلة بينها وبين الموضوعات المستفسر عنها من قبل المستفيد.

تقوم تقنيات خداع الروابط على تغيير بنية الروابط لصفحات المواقع لمهاجمة محركات البحث المعتمدة على استخدام خوارزميات الروابط للترتيب الطبقي لنتائجها مثل خوارزمية رتبة الصفحة pagerank ونموذج HITS والتي سبق الاشارة اليها وتعد اشهر تقنيات الخداع المعتمد على الروابط هي بناء ما يعرف بمزرعة المواقع وتبادل الروابط وهو ما سنتعرض له الان.

- مزارع الروابط Link farm:

تعتمد هذة التقنية على اضافة عدد من الروابط التي تشير إلى مواقع معروفة وتتمتع بشعبية على امل أن تكون الصفحة مصدرا للصفحات الموثوقة بها، ويتم تجميع هذة الروابط من خلال نسخ المواقع المرتبة في الادلة الموضوعية Directory

- والتي حظيت بفهرسة دقيقة وموثوقية – ثم تسكينها في الصفحة الخادعة وبالتالي حينما تقوم محركات البحث بترتيب هذة الصفحة نجد انها تحظى برتبة عالية في كونها مصدرا للروابط او المواقع ذات الموثوقية.

- خداع معتمد على اخفاء او تضمين الروابط Create a honey pot:

تعتمد هذة التقنية على الروابط التي تشير إلى صفحة ما، فمن الممكن ان تشتمل بعض الصفحات على معلومات مفيدة تجعل المستفيد يستخدمها، ولكن تشتمل هذة الصفحات على روابط خفية للصفحة الخادعة من خلال هذة الروابط تحظى هذة الصفحة برتبة عالية في نتائج محركات البحث.

- تبادل الروابط Link exchange:

تعتمد هذة التقنية على اتفاق بين اصحاب المواقع او مديريها والتي تتمثل في اضافة الرابط الخاص بك في صفحتهم على ان تضيف رابطهم في صفحتك، بغض النظر عما اذا كان رابطهم يتصل بموضوع صفحتك او لا مما يسمح بظهور هذة الصفحة في نتائج محركات البحث.

- شراء الروابط Link purchase:

وتعتمد هذة التقنية على ان تقوم بعض المواقع بأن تدفع مبالغ لبعض المواقع لتضمين روابط صفحاتهم ضمن صفحات الموقع الاخر.

- شراء النطاقات المنتهية Expired domains:

وتعتمد هذة التقنية على ان يقوم المخادعون بشراء نطاقات منتهية او قد غيرت ووضع محتوى غير مفيد، في ظل تمتع هذا النطاق برتبة جيدة داخل نتائج محركات البحث.

- تقنية صفحات الابواب الخلفية Doorway pages:

تعتمد هذة التقنية على انشاء صفحة خادعة من اجل خداع كشاف محركات البحث تظهر في نتائج محرك البحث وعندما يقوم المستفيد بالنقر عليها يتم اعادة توجيهه إلى صفحة اخرى مما يسمح بارتفاع رتبتها داخل محرك البحث في ظل تكرار الزيارات لها من قبل المستفيدين، إن فكرة هذة التقنية تعتمد على كود يعرف بأسم Meta refresh والذي يعمل على اعادة تنشيط الصفحة خلال فترة زمنية وفقا لمحدد اخر وتوضح في المثال الآتي:

```
<meta http-equiv="refresh" content="5;url=http://example.com/" >
```

يعد هذا الكود من اكواد الميتا الخاص بعمل بعمل تنشيط للصفحة خلال فترة زمنية معينة وعلى ان يحدد URL الخاص بالموقع كما هو مبين في الاعلى ، ما يقوم به الخادعون Spammers ان يلصقوا URL الخاص بهم مما يؤدي إلى انه اثناء

تحميل الصفحة الرئيسية يتم عمل تنشيط لها خلال خمس دقائق وتوجيه الصفحة للURL الآخر مما يوفر فرصة لوثيقة الخادعة ان تحظى برتبة عالية نتائج محركات البحث في ظل ارتفاع معدلات الزيارة لها.
ولا تعد هذة الطريقة هي الطريقة الوحيدة لعمل الابواب الخلفية بل يمكن انشاؤها من خلال ما يعرف بنصوص الجافا JavaScript .

- الروابط المنفجرة Link bombing spam:

حيث ان الروابط في البنية العنكبوتية بمثابة روابط حرة يمكن ان تطعم بالكثير من العبارات والكلمات المفتاحية، ومن ثم يكون تحظى الصفحة المشار اليها برتبة علاية في نتائج محركات البحث.

- خادعات التعليقات والمدونات Comment or blog spam:

تعتمد هذة التقنية على ان يقوم الخادعون بلإضافة روابط خادعة لبعض المواقع والمدونات التي تحظى بسمعة طيبة في ترتيب نتائج محركات البحث.

خادعات اخفاء الصفحات:

من الطبيعي ان يقوم الخادعون Spammers بإخفاء الادلة او ما يثبت انشطتهم الخادعة، ويتم ذلك من خلال استخدام العديد من التقنيات لتخفى عن المستفيدين الذين يقومون بزيارة المواقع او الصفحات الخادعة، او اخفائها عن محرري شركات محركات البحث الذين يسعون إلى تحديد حالات الخداع.

- الحجب Cloaking:

تعتمد هذة التقنية على أن تقوم بعض خادمات المواقع الخادعة Sam web searver بتحديد زواحف محركات البحث وذلك من خلال عناوينها IP الخاصة بالزاحف، ومن ثم تقوم هذة الخادعات بتوفير محتوى خاص للزاحف غير الذي تقدمه للمستفيدين ومن ثم الحصول على الظهور في بعض النتائج غير المتصلة.

- الخداع المعتمد على الطبقات Layer based spam:

تعتمد هذة التقنية على الخداع من خلال ما يعرف بطبقات CSS (Cascading Style Sheets) تنسيق طراز الصفحة، حيث تعد تقنية CSS أحد التقنيات الخاصة بتنسيق الصفحات من خلال طبقات، يقوم المخادعون بتزويد الموقع بطبقة غير مرئية للمستفيد حين العرض حيث تشتمل على المحتوى الخادع لمحركات البحث.

- الخداع الاستفساري Query spam:

حيث تعمد هذة التقنية على توجيه اسئلة لوحدة الاستفسار الموجودة داخل بنية الزاحف بهدف تلويث ملف الاستفسار لديه.

يمكن القول بان هناك ثلاثة عوامل رئيسية خارجية تقف **كعقبة في طريق محركات البحث في القضاء على الخادعات**.

1- التطور السريع لاساليب وخوارزميات خداع محركات البحث لملاحقة تطور خوارزميات التكشيف في محركات البحث.

فقد عملت الخادعات spams على تضليل محركات البحث منذ الوهلة الأولى لظهور الثانية، فبداية عمدت الخادعات إلى تقنية حشو الكلمات المفتاحية keyword stuffing داخل كشافات محركات البحث في ظل اعتماد محركات البحث على منهجيات وخوارزميات نظم المعلومات التقليدية في التكشيف. مثل تقنية TF-IDF.

وقد كان هذا الأمر دافعا للباحثين والعاملين في حقل محركات البحث على توفير خوارزميات تكشيف اخرى لمحركات البحث بمنأى عن خادعات البحث، فظهرت خوارزمية الترتيب الطبقي المعتمد على الروابط link based ranking algorithm في ظل تضاؤل فاعلية الخوارزميات السابقة، ولكن سرعان ما أوجد الخادعون اسلوبا جديدا للتعامل مع هذة الخوزارزمية الجديدة عرف هذا الاسلوب بإسم link bombing or link farms الأمر الذي جعل الباحثين إلى تصوير هذا الوضع بالحرب بين الخادعين ومحركات البحث.

2- التنوع الواسع في تقنيات خادعات محركات البحث:

يعمل منشؤا خادعات محركات البحث على انشاء العديد من خوارزميات وتقنيات الخداع المختلفة مما يجعل الأمر صعبا على محركات البحث في ايجاد منهجية او خوارزمية موحدة للكشف عن خوارزميات الخداع ويعد احد اهم الأمثلة في هذا المقام اسلوب الخداع المعتمد على نمطية الحجب .cloaking behavior (أحد التقنيات الخاصة بخداع محركات البحث يعتمد على ان يرسل إلى زاحف محرك البحث محتوى يختلف عن المحتوى الذي يراه المستفيدين من خلال اخفاء بعض النصوص في صفحات التكويد الخاصة بالموقع). بالاضافة إلى هذا يصعب على محركات البحث التعرف المسبق على طبيعة ونوع الخداع المحتمل ان يتعرض له بحيث يمكن ان يشبه هذا الأمر بمعركة مع خصم غير مرئي.

3- استمرارية النمو الهائل للشبكة العنكبوتية:

حيث تعد استمرارية نمو الهائل للشبكة العنكبوتية سببا تتضاءل فرضية الاعتماد على البشر في ملاحقة والتحقق من صفحات المواقع بدلا من الآلة (الزواحف)، على الرغم مما يكفله طبيعة العمل البشري من القدرة المنطقية في الحكم على المحتوى مما يجعل فرصة خداع محركات البحث تتضاءل امام الخادعين، ولكن اغلب ما تتخذه محركات البحث من اجراءات تأخذ الطابع الآلي وبالتالي يجدر على محركات البحث في ظل عدم اعتمادها على القدرة البشرية واستبدالها بالطاقة الالية في الحكم على المحتوى ان تعمل على اختبار خوارزمياتها الخاصة بمكافحة الخادعات على نطاق واسع ومتنوع من البيانات المتاحة على الويب، وفي هذا المقام يجب توضيح نقطة هامة ان عالبية الخوارزميات التي تتخذها

محركات البحث في مكافحة الخادعات تعمل بشكل جيد على العينات من المواقع والصفحات ولكن الأمر يختلف جذريا في التعامل مع المحتوى المتدفق من العنكبوتية[18].

العنكبوتية غير المرئية كتحد لخوارزميات محركات البحث في أسترجاع المحتوى:

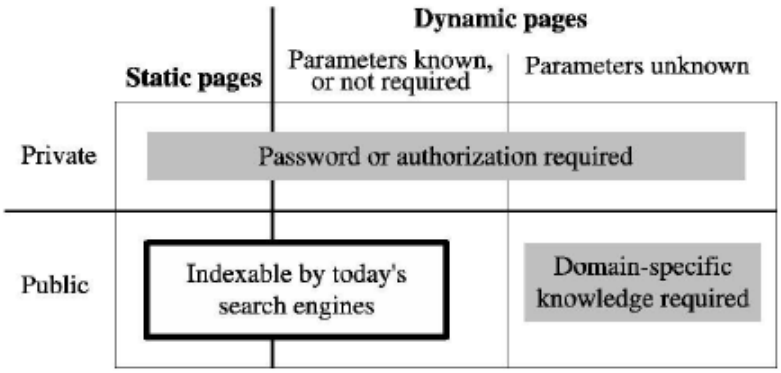
يمكن النظر إلى العنكبوتية العالمية وفقا لمنظور البحث والاسترجاع على انها تنقسم إلى قسمين:

القسم الأول: العنكبوتية السطحية Surface web أو العنكبوتية القابلة للتكشيف Publicaly Indexable web أو العنكبوتية المفتوحة Open web :وهي تلك الصفحات والمواقع التي تخضع للتجميع والتكشيف والاسترجاع من قبل محركات البحث من خلال تتبع الروابط.
القسم الثاني: العنكبوتية غير المرئية Invisible web: وتعرف ايضا بالعنكبوتية العميقة the deep web، ، أو العنكبوتية المظلمة Darknet او ما تحت العنكبوتية Undernet: وتشتمل على مصارد المعلومات التي لايمكن لمحركات البحث ان تخضعها للتجميع او التكشيف او التسكين في فهارسها. كما هو موضح في الشكل رقم (20).

إن محركات البحث التقليدية تعتمد في انشاء فهارسها وكشفاتها على تصفح وتتبع محتوى المواقع بغية اكتشاف الصفحة وتتبع ما بها من روابط لتجميع صفحات اخرى، وعليه يسقط من حسبانها البيانات والمصادر ذات الطبيعة المستقلة عن الروابط الفائقة هذة المصادر تشكل ما يعرف بالعنكبوتية الخفية، كما أن عجز محركات البحث لايتمثل في عدم قدرته على التعامل مع العنكبوتية الخفية وحسب بل يمتد ليتضح أن تغطية محركات البحث لما هو متاح على العنكبوتية المرئية لا يتجاوز من 20% - 50% ويعود ذلك إلى عدة اسباب:

- اعتماد محركات البحث على تقنيات تتسم بالمحدودية في قدراتها حيث لايمكن لزواحف محركات البحث مواكبة ومتابعة وتيرة الديناميكية التي تتخذها العنكبوتية في ظل التزام هذة الزواحف بجداول زمنية قد تتفاقم التغييرات في المواقع خلال تكرار الزيارة لها من قبل الزاحف.
- ارتفاع التكاليف التي تنطوي على تشغيل محرك البحث بصورة شاملة فمن المعروف ان تحديد مواقع مصادر المعلومات والحفاظ على حداثة كشافات محرك البحث بصورة دورية يعد امرا مكلفا للغاية.
- محاولة محركات البحث تجنب الخادعات ذات التأثير الضار لفهارسها وكشفاتها، ومن خلال تطويرها لمنهجيات صارمة تكفل تجنب مع هذة الخادعات، تحمل هذة المنهجيات عيوبا تستقضي استبعاد محتوى اخر.
- اختلاف محركات البحث فيما بينها من حيث خوارزميات التجميع والتكشيف والاسترجاع لكل منهم، وقد ادى ذلك لما يعرف بالتداخل والتكرار في نتائج محركات البحث.

[18]Wu, B. (n.d.). FINDING AND FIGHTING SEARCH ENGINE SPAM. *CiteSeerX*. Retrieved July 21, 2011, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.6961



شكل رقم (20) رسم توضيحي يوضح فئات الوثائق من خلال انواع العنكبوتية

تعرف العنكبوتية غير المرئية "بانها ذلك الجانب من العنكبوتية الذي يشتمل على محتوى لايمكن لمحركات البحث التعامل معه بالتجميع أو التتبع أو الاكتشاف نظرا للقيود والمحدودية التي تتمتع بها تقنياتها من زواحف وبرامج تصفح أو نظرا لسياسات التوجيهية لمحركات البحث في الاختيار المتعمد لمصادر ذات سمات وطبيعة محددة"[19].

ويمكن تحديد خصائص العنكبوتية غير المرئية على النحو الاتي:

- يقدر حجم العنكبوتية غير المرئية من 400 - 500 مرة من العنكبوتية المرئية.
- تشتمل العنكبوتية على ما يقرب من 70 الف تيرابايت من المعلومات مقارنة ب9 الاف تيرابايت متاحة على العنكبويتة القابلة للتكشيف.
- تشتمل العنكبوتية غير المرئية على ما يقرب من 550 بليون وثيقة منفردة مقارنة ببليون وثيقة متاحة على نظيرتها.
- يشتمل العنكبوتية غير المرئية على اكثر المصادر تتمتع بمصداقية وموثوقية عما هو متاح على العنكبويتة السطحية.
- تشتمل غير المرئية على اغلب المصادر التي يمكن ان تحقق نسبة مرتفعة من تلبية الحاجات المعلوماتية.
- 95% من معلومات العنكبوتية غير المرئية متاح مجانا[20].

إن التحديات التي تواجهها محركات البحث في التعامل مع محتوى العنكبوتية غير المرئية تكمن في تعاملها مع هذة المفردات:

1. مستودعات قواعد البيانات information stored in databases:

[19] Sherman, C., & Price, G. (2001). *The Invisible Web: uncovering information sources search engines can't see*. Medford, N.J.: CyberAge Books.
[20] MICHAEL K. BERGMAN. The Deep Web: Surfacing Hidden Value.avaliable: http://brightplanet.com/images/uploads/12550176481-deepwebwhitepaper.pdf

فهناك الالاف او ربما الملايين من قواعد البيانات التي تشتمل على معلومات عالية الجودة والموثوقية والمصداقية، ما يتمثل في تحد لمحركات البحث يكمن في ان كل قاعدةبيانات تعتبر نظاما فريدا من حيث تصميم هياكل البيانات ومن حيث قدرات البحث والاسترجاع، ومن ثم فزواحف محركات البحث يمكن لها ان تصل إلى واجهات الاستخدام ولكن لاتتمكن من الانتقال إلى المرحلة التالية وهي مستودعات قواعد البيانات نظرا لاختلاف بنيتها عما هو مبرمج عليه الزاحف في التعامل معه.

2. صيغ التمثيل النصي Formats of text:

حيث ان غالبية محركات البحث صممت لإتقاط وتكشيف النصوص التي صيغت وفقا لصيغ لغة النص الفائق دون النظر إلى صيغ التمثيل النصية الاخرى كpdf وdej vu و Epub وغيرها نظرا لحاجتها لمعالجة حاسبية معينة مما يوفر فجوة في التعامل مع بعض هذة الصيغ.

3. الوثائق المنقطعة Disconnected:

وهي الصفحات المتاحة على العنكبوتية والتي لا يشار اليها عبر الروابط الفائقة من قبل صفحات اخرى وهي تمثل غالبية العنكبوتية غير المرئية، ما يمثل تحديا لمحركات البحث في هذا الصدد ان زواحف محركات البحث تعتمد في زحفها على تتبع الروابط الفائقة للإلتقاط الصفحات ومن ثم اغفال الوثائق والصفحات التي لم يشر اليها من جانب الصفحات الاخرى.

4. المحتوى غير النصي Untextual content:

فالكثير من محركات البحث صممت زواحفها للتعامل مع المحتوى النصي او التركيز بشكل كامل على طبيعة هذا المحتوى دون النظر إلى الكيانات الاخرى ومن ثم اغفال قدر ليس بالصغير من المحتوى الممثل في صور اخرى، ومع ذلك قد نجد الكثير من محركات البحث تضمن امكانية البحث من خلال الكيانات الاخرى كالبحث عن الصور والصوت والفيديو ولكن يجب الاشارة إلى ان محركات البحث لم تقم بتجميع هذة الكيانات بصورة مستقلة بل ان هذة الكيانات مدرجة في وثائق نصية ووتكتسب قابلية البحث من خلال عمليات الفرز التي تتم داخل كشافات محركات البحث كما ان هذة الكيانات تكشف من واقع التعليقات نصية سابقة او لاحقة او من خلال ما يعرف ب" البدائل alternative (ALT) text والذي ينشئ من قبل مصمم الوثيقة.

5. المحتوى ذات الطبيعة الانية Real-time content:

حيث تشتمل العنكبوتية على محتوى يتسم بانها ديناميكة الوقت في وجودها وفي تعطل محتواها، كجداول الطيران واسعار الاسهم وغيرها، الامر الذي لايوفر لمحركات البحث فرصة في التقاطه نظرا للجداول الزمنية لانطلاق الزواحف وتكرار الزيارة.

6. تقييد زواحف محركات البحث من خلال بروتكول استبعاد الروبوتات Robots Exclusion Protocol:

وهو عبارة عن مجموعة من القواعد التي تمكن مديروا المواقع من تحديد الاجزاء المفتوحة من الخوادم ليتعامل معها الزاحف واي منها مغلق خارج نطاق التعامل. ويتم ذلك من خلال ان يقوم مديروا المواقع بانشاء قائمة من الملفات او الادلة التي لا يجب لها ان تجمع او تكشف ثم حفظها في الخادم تحت مسمى robots.txt.

7. اقصاء الزواحف:

يتم ذلك من خلال ان يقوم مديروا المواقع بتضمين عبارة NoIndex في حقول الميتاداتا في راس الوثيقة.

8. كلمات السر:

وتعد التقنية الاكثر قوة في تجنب تكشيف المحتوى للوثائق حيث تعتمد على حاجز تقني لابعاد الزاحف عن التجميع.

سلوك وانماط تفاعل المستفيدون مع محركات البحث:

يمثل المستفيد المحور الثاني بعد محور النظام في منظومة محركات البحث User-oriented وهو ذلك المنحنى من دراسات محركات البحث الذي يهتم بدراسة طبيعة المستفيد البحثية، ونمط سلوكه البحثي، ومقاومات لغته، والمتغيرات العلمية والثقافية والاجتماعية التي تؤثر في صياغة استفساره.

الاستفسارات واسترجاع المحتوى على العنكبوتية:

ذهب Broder إلى ان العنكبوتية قد حملت معها انماطا جديدة لحاجات المستفيدين في التعامل مع المعلومات سواء كان على صعيد البحث أوالاسترجاع، وقد امتدت هذة الانماط لتشكل فئات جديدة من الاستفسارات المعلوماتية لدى المستفيدين، ففي سياق العنكبوتية، لاتقتصر الحاجة the need من وراء الاستفسار على ان تكون ذو طبيعة معلوماتية فحسب بل تمتد لتشمل 3 اقسام اخرى الحاجة:

- الحاجة الملاحية Navigational need.
- الحاجة المعلوماتية Informational need.
- الحاجة الاجرائية Transactional need .

1- الحاجة الملاحية:

يعد نمطا حديثا من الاستفسارات فالغرض منه هو الوصول إلى موقع أو مصدر معلومات معين لدى المستفيد وأن الحاجة من وراء هذا الاستفسار تتمثل في أن المستفيد ربما قد زار أو تعرض لهذا الموقع من قبل او على فرضية ان مثل هذا الموقع يمكن ان يكون موجودا على العنكبوتية. إن هذا النمط من الحاجات او البحث كان يشار اليه سالفا بالبحث عن

"المفردة المعروفة" بمعنى ان المستفيد يكون على دراية وعلم بأن هذة المفردة (سواء كانت هذة المفردة وثيقة او ملف صوتي او موقع ..الخ) موجودة على العنكبوتية.

2- الحاجة المعلوماتية:

يعد هذا النمط تقليديا إلى حد ما فالغرض منه لايختلف كثيرا عن الغرض من الحاجات المعلوماتية من نظم الاسترجاع التقليدية وهو ايجاد المعلومات التي يفترض ان تكون متاحة على العنكبوتية في صورة ثابتة ولكن سمة امتياز امتازت به العنكبوتية تمثل في قدرتها على المزج بين البيانات المتاحة من خلال المصادر المختلفة في صورة اقرب لتحقيق التكاملية بين مصادر العنكبوتية فمثلا من خلال البحث عن مدينة الرياض في بعض محركات البحث تكفل خوارزميات هذة المحركات القدرة على توفير مصادر المعلومات التي تتناول الرياض ثم تستطرد لتسترجع اسماء فنادق الرياض والمعالم الاثرية بها... الخ .

3- الحاجة الاجرائية:

والغرض من طبيعة هذة الحاجة هو الوصول إلى مواقع تتعالا فيها سمة التفاعل سواء كان بين المستفيدين بعضهم البعض (كالشبكات الاجتماعية) او تفاعل المستفيدون مع الالة (كتحميل الملفات والتنزيل الهابط وكالمعاملات التجارية مع البنوك).

اما عن فئات الاستفسارات فتشمل الاقسام الآتية:

1. الاستفسارات البولينية Boolean Queries: وهي تلك الاستفسارات التي تعتمد على الروابط المنطقية (and – or – not) في صياغته.
2. الاستفسارات باللغات الطبيعية Natural Language Queries: وهي تلك الاستفسارات التي تصاغ في صورة سؤال او جملة خبرية.
3. الاستفسار المكنزي Thesaurus Queries: وهو النمط الذي يعتمد على ان يقوم المستفيد باختيار مصطلح من قائمة بالمصطلحات معدة سلفا من جانب نظام الاسترجاع او محرك البحث.
4. الاستفسار المبهم Fuzzy Queries: والذي يعكس انماط الاسئلة التي تشتمل على اخطاء املائية او مشتقات من الجذور اللغوية.
5. الاستفسار بالتجاور Term Searches: وهو نمط الاستفسارات الذي يعتمد على صياغة السؤال في جملة واحدة دون تفتيتها.

استفسارات المستفيدين كتحد يواجه محركات البحث في استرجاع المحتوى:

Cybrarians Journal
ISSN 1887-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 30، ديسمبر 2012

حمل ظهور العنكبوتية العالمية في بدايتها مؤشرات شتى كان أولى هذة المؤشرات نهاية حقبة ما يعرف بالبحث من خلال وسيط المعلومات، فقد كفلت العنكبوتية وللمرة الاولى للمستخدم القدرة على ان يقوم بالبحث بنفسه دون الحاجة إلى وسيط للبحث تتفاقم من خلاله عملية البحث او تتبسط وفقا لعوامل متغيرة.

ويرى Ricardo baeza ان التحديات التي تواجه محركات البحث في تفاعل المستفيدين معها تكمن في محورين اساسيين:

- التعبير عن الاستفسار وصياغته:

ان التعبير عن الحاجات المعلوماتية لدى البشر في صورة استفسارات لايعد امر سهلا في كثير من الاحيان ، حتى وإن كان هذا التعبير باللغة الطبيعية، فالاستفسار في افضل صورة ماهو الا انعكاس لاحتياجات المعلومات وبالتالي قد يفتقر هذا الانعكاس إلى الدقة او الوضوح فضلا عن العوامل الاجتماعية والثقافية والتعليمية التي تلعب دورا في التعبير عن الحاجة المعلوماتية.

- تفسير الاجابات والنتائج:

حتى ان المستفيد قادرا على التعبير بشكل مثالي عن حاجته من خلال استفساره، فمن الممكن للاجابة ان تنطوي على الاف او ملايين الوثائق التي تتفاقم معاها بعض القضايا التي تتعلق بمطابقة حاجة المستفيد أم لا، او كيفية ترتيبها في نسق منطقي واسباب هذا الترتيب، وكيف يمكن ان تستخدم هذة الوثائق بفاعلية من جانب المستفيد.

ويمكن اجمال التحديات التي تتعلق بتفاعل المستفيدين مع محركات البحث على النحو الاتي:

1-حجم الاستفسارات:

إن حجم استفسارات المقدمة إلى العنكبوتية في نمو مستمر فقد بلغت نحو 10 مليار استفسار عام 2008 موجه إلى 5 محركات بحث (Google, Yahoo!, MSN, AOL, Ask Jeeves). كما أن عدد الاستفسارات الموجهه لمحرك البحث جوجل اكثر من 200 مليون استفسار يوميا عام 2003، لتصل هذة النسبة في عام 2010 إلى 3,5 مليار استفسار يوميا بمعدل يصل إلى 40,000 استفسار في الثانية، وبالتالي تعد هذة المؤشرات بمثابة تحد كبير يواجه محركات البحث.

2-صياغة الاستفسار:

عادة مايميل المستفيدون إلى تقديم استفسارت أعم بكثير من الحاجة الفعلية إلى المحتوى، والمرجعية وراء ذلك تكمن في ظن المستفيد بأن محرك البحث يمكن ان يعمل على مستوى أعرض وليس على مستوى التخصيص، هذا من ناحية ومن ناحية اخرى كشفت الدراسات المقارنة – القائمة على تحليل ملفات النظام لمحركات البحث – أن سلوك المستفيدين في البحث على المعلومات ينطوي على أن معدلات الاستفسارت المقدمة لمحرك البحث تتسم بالقصر فمعظم المستفيدين يدخل

مابين مصطلحين إلى ثلاثة مصطلحات في الاستفسار الواحد، ويدخل مابين استفسارين إلى ثلاثة استفسارات في عملية البحث الواحدة.[21]

3-نمط الاستفسار :

إن 50% من اجمالي الاستفسارت المقدمه لأحد محركات البحث كانت تبدأ بعبارات استفهامية مثل "Where do I find . . .? كيف أجد" وأن 25% من المستفيدين يبدؤن بحثهم بعبارات طلبية مثل "احضر لي معلومات عن Get me information"، في حين أن معظم محركات البحث تعتمد في المقام الأول على الكلمات المفتاحية مما يسفر في الحالات السابقة عن الكثير من النتائج غير المرضي عنها من قبل المستفيد.

4-معدلات توزيع مفردات ومصطلحات البحث:

أوضحت إحدى الدراسات البحثية أن معدل توزيع تردد مصطلحات الاستفسار يتسم بالانحراف من إجمالي عمليات البحث فقد كشفت هذة الدراسة أن بعض المصطلحات تستخدم بشكل متكرر في مقابل الكثير من المصطلحات التي تستخدم مرة واحدة فقط ففي احدى الدراسات تم الكشف عن ان 63 مصطلح حظي بتردد ظهور بلغ أكثر من 100 مرة، في حين كون هذة المصطلحات تمثل اقل من 1% من اجمالي المصطلحات. مما يظهر طبيعة أن البحث على العنكبوتية يمكن وصفه بانه بحث بمصطلحات تتسم بنسبة منخفضة في تردد الظهور مقابل مصطلحات تتسم بتردد عالي في الظهور.

5-إعادة صياغة الاستفسار:

أن معظم الباحثين على محركات البحث يلجاون إلى صياغة حاجتهم البحثية في صورة استفسار واحد فقط دون اللجوء إلى اعادة صياغته مرة اخرى بمعدل بلغ اثنين من 3 باحثين وبصورة عامة بلغت نسبة المستفيدون الذين يعتمدون على تعديل استفسارتهم في البحث على العنكبوتية نحو 44% بينما بلغت نسبة من يقوم بتوجية اكثر من 3 استفسارت في عملية البحث الواحدة نحو 25%.

6-الاعتماد على الروابط البولينية والبحث المتقدم:

بلغت درجة الاعتماد على الروابط البولينية في عمليات البحث على درجة تكاد تصل إلى الندرة في عمليات البحث داخل محركات البحث فشخص واحد من اجمالي 18 شخص يلجا إلى الاعتماد على الروابط البولينية فضلا عن شخصان من اصل 3 اشخاص يستخدم هذة الروابط بصورة خاطئة.[22]

[21] Jansen, B. J., & Spink, A. (2003). An analysis of web information seeking and use: Documents retrieved versus documents viewed. In Proceedings of the 4th International Conference on Internet Computing, pp. 65-69. Las Vegas, Nevada. 23-26 June.

نتائج والتوصيات:

أولا: النتائج:

تتمثل نتائج الدراسة في حصر التحديات التي تواجه محركات البحث بعد التعرض لها بالحصر والتحليل فيما سبق ويمكن اجمال التحديات على هذا النحو:

اولا: التحديات الداخلية النابعة من خوارزميات محرك البحث والتي تعوقه في استرجاع المحتوى:

1. الزواحف:

- لايستطيع الزاحف تحميل وتجميع كافة الصفحات المتاحة على العنكبوتية، وفي ظل ذلك يجدر على الزاحف ان يحدد الصفحات التي يجب زيارتها وذلك وفقا لأهميتها.
- صعوبة تحديد الحد الادنى من التحميل والتجميع للمحتوى.
- صعوبة الكشف عن وجود مكررات على العنكبوتية على صعيد المحتوى.
- صعوبة تحديد المجموعات البذرية للمحتوى اي ماهي الصفحات التي يجب ان تزار اولا.

2. التكشيف:

- 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.
- ما تقوم به محركات البحث من أستبعاد لبعض الكلمات في المحتوى والتي قد تحمل دلالة ضئيلة ولكنها في ذات الوقت تلعب دورا محوريا.
- 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.
- مفاضلة محركات البحث بين صغر حجم الكشاف وبين القدرة على إجابة الاستفسارات المعقدة من محتوى الوثائق.
- بناء الكشافات الفرعية ضمن الكشاف المقلوب مما يسمح بوجود تداخل وتكرار في نتائج المحتوى.

3. الترتيب للمحتوى المسترجع:

[22] Jansen, B. J., & Spink, A. (2005). How are we searching the World Wide Web? A comparison of nine large search engine transaction logs. Information Processing and Management, 42(1), 248-263.

- الاعتماد على المقياس الثنائي في الاسترجاع والذي يعتمد على أن ترجيح المحتوى المسترجع بين كفتين فقد أما متصل باستفسار المستفيد وأما غير متصل.
- الاعتماد على التكشيف المعجمي وما يحمل من مشكلات تتعلق بالترادف اللغوي أو التجانس اللغوي.
- الإعتماد على نظرية الاحتمال في الاسترجاع.

ثانيا: التحديات الخارجية التي تعوق استرجاع المحتوى في محركات البحث:

1. خادعات المحتوى لمحركات البحث:

وما تحمله من قضايا تؤرق محركات البحث والتي تتمثل في:

- تصعيب مهمة المستفيدين في ارضاء حاجتهم المعلوماتية information needs.
- ارساء خلفية سيئة تجاه محرك البحث ونتائجه وعدم مصداقيته.
- تلويث كشاف محركات البحث بمواقع زائفة.
- حرق النطاق الترددي للزواحف او ما يعرف بBandwidth.
- تشوية نتائج محركات البحث.
- قدرة الخادعات على توليد العديد من شبيهاتها.

وقد يمتد الضرر إلى اصحاب المواقع، فقد يدفع مدراء المواقع إلى التخلي عن جهودهم في توفير صفحات ومحتوى ذو جودة مرتفعة، اخذين على عاتقهم مهمة انشاء محتوى يستطيع ان يتعامل مع محركات البحث مما يستتبع ان يجعل محتوى الشبكة العنكبوتية موجه إلى محركات البحث بدلا من ان يوجه إلى مستفيدي الشبكة العنكبوتية الامر الذي يشكل كارثة في نهاية المطاف على الشبكة العنكبوتية.

2. العنكبوتية الخفية:

يتمثل التحدي الرئيسي لمحركات البحث في التعامل مع محتوى العنكبوتية الخفية في عجزها الكامل في الوصول اليه ومن ثم عدم تكشيفه أو نظمة أو معالجته وذلك على الرغم من كون العنكبوتية الخفية تشتمل على 5 اضعاف العنكبوتية المكشفة.

3. تفاعل المستفيد مع محركات البحث في استرجاع المحتوى:

يتخذ هذا التفاعل صورة الاستفسارات الموجهه من المستفيد الى محرك البحث والتي تحمل معها الحاجات المعلوماتية أو المحتوى الفعلي المطلوب، أما عن سمات استفسارات المستفيدين فهي تتفاوت بين عناصر ومؤشرات تصعب على محرك

البحث الاسترجاع بالصورة التي يرضى عنها المستفيد، والمرجعية في ذلك تعود الى عدم المام المستفيد بحاجته الفعلية أو عدم قدرته على التعبير بصورة صحيحة عن المحتوى الذي يطلبه، أو عدم فهمه لإمكانيات محرك البحث الذي يتعامل معه.

ثانيا: التوصيات:

يوصي الباحث في نهاية الدراسة بما يلي:

1. العمل على التخلي عن البنية التكويدية الحالية للشبكة العنكبوتية المتمثلة في لغة (HTML) والتوجه نحو الاعتماد على التكويد بلغة XML.
2. الاعتماد على التشفير بالنظام UTF-8 في صياغة محددات المصادرURI.
3. التوجه نحو نمذجة المحتوى المتاح على الشبكة العنكبوتية في مجموعة من نماذج للبيانات DATA MODEL.
4. اثراء المحتوى العربي بالانطولوجيات العامة والمتخصصة.
5. توفير نطاقات اسماء عامة ومتخصصة ترعاها رابطة اتحاد الويب من حيث التحديث وجعلها مظلة للمحتوى المتاح على الويب.
6. اثراء المحتوى العربي على الانترنت بالبحوث الاصيلة العلمية في مختلف المجالات بهدف رفع الرتبة اللغوية العالمية للغة العربية على الويب .

المصادر:

المصادر الأجنبية:

1. Andrew Hammond. Arabic search engine may boost content. http://www.abc.net.au

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

3. 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.

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

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

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

7. Christopher Olston and Marc Najork. Web Crawling. Foundations and Trends in Information Retrieval. Vol. 4, No. 3 (2010).

8. Craig Silverstein ,Hannes Marais ,Monika Henzinger,Michael Moricz. Analysis of a very large web search engine query log. ACM SIGIR Forum, 1999 - portal.acm.org.

9. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (September,1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6): 391-407.

10. Detecting Spam Web Pages. Marc Najork. Microsoft Research, Silicon Valley.

11. Dominich, S. (2008). The modern algebra of information retrieval . Berlin: Springer.

12. Eric Convey.Porn sneaks way back on Web.The Boston Herald, May 22, 1996.

13. G. Madhu, A. Govardhan, T. V. Rajinikanth: Intelligent Semantic Web Search Engines: A Brief Survey CoRR abs/1102.0831: (2011).

14. Gyongyi, Z., & Garcia-Molina, H. (n.d.). Web Spam Taxonomy. Web Spam Taxonomy. Retrieved July 21, 2011, from airweb.cse.lehigh.edu/2005/gyongyi.pdf.

15. Gyongyi, Z., & Garcia-Molina, H. (n.d.). Web Spam Taxonomy. *Web Spam Taxonomy*. Retrieved July 21, 2011, from airweb.cse.lehigh.edu/2005/gyongyi.pdf.

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

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

18. Internet world stat.http://www. Internetworldsta.com

19. Jansen, B. J., & Spink, A. (2005). How are we searching the World Wide Web? A comparison of nine large search engine transaction logs. Information Processing and Management, 42(1), 248-263.

20. K. Sparck Jones. A probabilistic model of information retrieval: development and comparative experiments. Information Processing and Management 36 (2000) 779±808.

21. Langville, A. N., & Meyer, C. D. (2006). Google's PageRank and beyond: the science of search engine rankings. Princeton, N.J.: Princeton University Press.

22. 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.

23. 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: Harper-Business.

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

25. Levene, M. (2010). An introduction to search engines and web navigation (2nd ed.). Hoboken: Wiley.

26. Maron, M. E., & Kuhns, J. (1960). On relevance, probabilistic indexing and information retrieval. Journal of the Association for Computing Machinery, 7(3), 216–244.

27. 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.

28. MICHAEL K. BERGMAN. The Deep Web: Surfacing Hidden Value.avaliable: http://brightplanet.com/images/uploads/12550176481-deepwebwhitepaper.pdf.

29. Ozgener, Isil. (2005). Publishing content on the web. : Stanford university.

30. 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.

31. 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.

32. S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Sciences*, 27(3):129-146, 1976.

33. Sherman, C., & Price, G. (2001). *The Invisible Web: uncovering information sources search engines can't see*. Medford, N.J.: CyberAge Books.

34. Stuckenschmidt, Heiner, and Frank Harmelen. *Information sharing on the semantic Web* . Berlin: Springer, 2005.

35. The Search Engine Industry. *Tommaso Buganza and Emanuele.springer .2010.*

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

37. 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

38. Wu, B. FINDING AND FIGHTING SEARCH ENGINE SPAM. *CiteSeerX*. Retrieved July 21, 2011, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.6961.

39. Yates, R., & Neto, B. (1999). Modern information retrieval . New York: ACM Press ;.

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

41. Zdravko, Markov & Daniel T. Larose. Data-mining the Web : uncovering patterns in Web content, structure, and usage. John Wiley & Sons, Inc.2007.

المصادر العربية:

1-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
2-نبيل علي. العرب وعصر المعلومات.عالم المعرفة.الكويت:المجلس الوطني للثقافة والفنون والاداب.1994.ص333.