Cybrarians Journal
ISSN 1687-2215
دورية إلكترونية فصلية محكمة متخصصة في مجال المكتبات والمعلومات

العدد 35، سبتمبر 2014

نحو التكامل المعرفي من واقع توظيف الأنطولوجيات في أطار التنقيب عن البيانات : دراسة تحليلية . 2

إعداد
مؤمن النشرتي
مدرس مساعد، قسم المكتبات والوثائق والمعلومات
كلية الآداب، جامعة القاهرة، مصر
Navigator001@gmail.com

المستخلص

سعت هذة الدراسة إلى توضيح مفهوم التكامل المعرفي في سياق تقني، وأشهر المبادرات التقنية التي ساهمت في تحقيقه على صعيد المحتوى المتاح على الشبكة العنكبوتية، ثم تناولت الدراسة رصد القدرة على توظيف مبادرة الانطولوجيات في أطار التنقيب عن البيانات سعياً بذلك إلى الاستفادة منهما فيما يعرف بنظم الانطولوجيا للتنقيب على الويب Ontology-based Web mining لتحقيق التكامل المعرفي لمحتوى شبكة الويب.

اعتمدت الدراسة على المنهج المسحي لرصد التحديات الخوارزمية واللغوية التي تواجه شبكة الويب في تحقيق التكامل المعرفي، كذلك تعتمد الدراسة على المنهج التحليلي في رصد وتحليل واقع قدرات الانطولوجيا في أن توظف لتحقيق التكامل المعرفي للمحتوى المتاح على الويب.

وكان من اهم نتائج الدراسة هي قدرة التنقيب على الويب على تحقيق مستوى مرتفع من التشغيل المتبادل على صعيد محتوى الويب بأكمله لتوفير التكامل المعرفي.

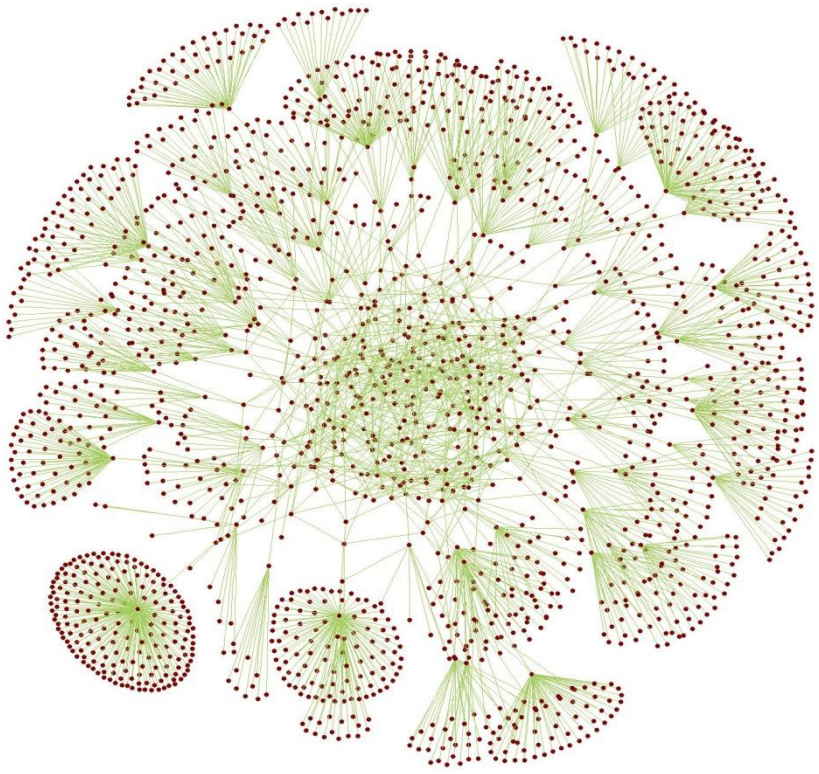
الاستشهاد المرجعي

نشرتي، مؤمن. نحو التكامل المعرفي من واقع توظيف الأنطولوجيات في أطار التنقيب عن البيانات : دراسة تحليلية .2 .– Cybrarians Journal .– ع 35، سبتمبر 2014 .– تاريخ الاطلاع <أكتب هنا تاريخ الاطلاع على المقال> .– متاح في: <أنسخ هنا رابط الصفحة الحالية>

مفهوم التنقيب على الويب:

وردت العديد من التعريفات والتي دارت جميعها في فلك كون التنقيب على الويب Web Mining واحدة من تقنيات الحوسبة الذكية التي تعمل في سياق إدارة بيانات الويب كوسيلة تهدف اكتشاف البيانات واستخراجها بغية تحويل الويب من وضعه الراهن الى شبكة معرفية. وفي هذا قام Jaideep Srivastava بتعريف التنقيب على الويب بأنه.."***أحد تطبيقات الذكية لتقنيات التنقيب عن البيانات والتي تهدف إلى استخراج المعرفة من محتوى وهيكل وسجلات الويب*** Web Log ***معتمدة في ذلك على منهجيات وخوارزميات التنقيب عن البيانات ونماذج بيانات الويب***[1]".

كان أحد أهم الاسهامات التي جعلت شبكة الويب مستودعا للبيانات ومنصة عمل فريدة في النشر والإتاحة هو بناء ما يعرف بالبنية البنائية Web Graph لشبكة الويب، هذة البنية التي تعمل على نمذجة البنية التكوينية للويب (مكونات الويب) في صورة رسم بياني يتكون من عنصرين اساسين الاول منهما هو العقد Nodes وهي المواقع او الصفحات والثاني هو الحواف Edges او Directed graph وتشير إلى الروابط بين هذة المواقع والصفحات كما هو موضح في الشكل الاتي:

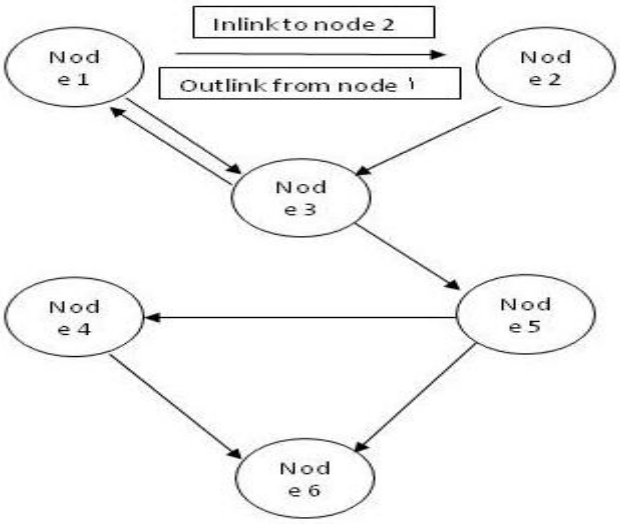


شكل رقم (7) يوضح البنية البيانية لشبكة الويب حيث تشير النقاط الحمراء إلى العقد (الصفحات) بينما تشير الخطوط الخضراء إلى الحواف (الروابط).

وقد أقدم Witten, I. H على دراسة المخطط البياني للويب بهدف وضع توصيف لاساسيات نموذج المخطط البياني للويب، والذي يعد ايضا بمثابة اساسيات للخوارزميات الخاصة للتنقيب على الويب، وقد جاء على النحو الاتي:

[1] J. Srivastava, P. Desikan, and V. Kumar, "Web Mining: Accomplishments and Future Directions," *Proc. US Nat'l Science Foundation Workshop on Next-Generation Data Mining* (NGDM), Nat'l Science Foundation, 2002.

- Web-graph: وهو ذلك المخطط البياني الموجه لتمثيل شبكة الويب.
- Node: اية وثيقة (سواء كانت صفحة او موقع) على العنكبوتية يشار اليها بانها عقدة على المخطط البياني.
- Edge (Directed graph) :تشير الي اية رابط فائق يربط بين موقعين أو اكثر.
- Indegree: تشير إلى عدد الروابط التي تشير إلى العقد وهي تعرف ايضا ب Inlink.
- Outdegree: تشير إلى عدد الروابط التي تنشا في العقدة إلى عقد اخرى وتعرف ايضا ب outlink[2].



شكل رقم (8) يوضح هذا الشكل أن كل عقدة B لديها (INDGREE) ولديها (OUTDGREE)[3].

أقسام التنقيب على الويب:

يعتمد التنقيب على الويب Web Mining على البنية المعمارية والبيانية لشبكة الويب الحالية، Web Graph هادفة بذلك إلى أكتشاف المعرفة من واقع كل من البيانات المتاحة على الويب – بنية الروابط الفائقة وما تحمله من دلالات بينية بين بيانات وكيانات الويب المختلفة – معدلات الاستخدام وأنماط وسلوك المستفيدين والمتعاملين مع الويب، وعلى الرغم من أعتماد التنقيب على الويب Web Mining على مفهوم التنقيب على البيانات Data Mining الإ أنها تتخلي عن تقنياتها وخوارزمياتها نظرا لطبيعة البيانات التي تتعامل معها فالتنقيب عن البيانات يعتمد على بيانات تتسم بالهيكلة في ظل وجودها داخل قواعد البيانات أما البيانات على الويب فتتسم بعدم الهيكلة في ظل اعتمادها على تقنيات التمثيل والعرض للمحتوى دون تقنيات التنظيم والهيكله، وعلى هذا أوضح Kosala and Blockeel أن التنقيب على الويب يأتي في 3 قطاعات بناءاً على الهدف من عملية التنقيب:

1- التنقيب اعتمادا على بنية الروابط المتاحة على الويب Web structure mining:

ويعمل هذا القطاع أو النوع من فئات التنقيب على الويب على أكتشاف المعرفة من واقع تحليل الروابط الفائقة على الويب Hyperlinks والتي تمثل أحد أهم أطراف بنية الويب المعمارية والبيانية، فمن واقع عملية التحليل

[2] Witten, I. H., Gori, M., & Numerico, T. (2007). Web dragons: inside the myths of search engine technology. Amsterdam: Morgan Kaufmann.
[3] Ibid.

يتم أكتشاف وتحديد أي المواقع تحظى بأهتمام وأهمية في ظل الروابط التي تشير اليها، كذلك يمكن من خلال تحليل الروابط الفائقة على الويب الكشف عن مجموعات الاهتمام، ويجدر القول في هذا السياق أن تقنيات التنقيب عن البيانات Data Mining لا تعتمد على هذا النمط او النوع في عمليات أكتشاف المعرفة وذلك في ظل عدم أشتمال جداول قواعد البيانات (التي تقوم بتحليلها) على الروابط الفائقة.

2- التنقيب عن محتوى الويب Web content mining.

يعمل هذا النمط من انواع التنقيب على الويب على أكتشاف المعرفة من واقع تحليل محتوى صفحات ومواقع الويب، ينطوي التنقيب والتحليل في هذا النمط على عنقدة وتبويب المواقع والمحتوى وفقا لموضوعاتها، كذلك تنطوي هذة الفئة على أكتشاف المعرفة من واقع تحليل التعليقات والتغذية المرتدة الناتجة من قبل القارئين والمستفيدين للخروج بمؤشرات أهتمام يمكن أن تستثمر في جوانب معرفية مختلفة، جدير بالذكر أن هذا الأمر لاينطبق على التنقيب عن البيانات أذ لاتتوافر في جداول قواعد البيانات قدرة أضافة التعليقات أو التغذية الراجعة للمحتوى.

3- التنقيب من واقع انماط الافادة والاستخدام على الويب Web usage mining:

يشير هذا النوع من انواع التنقيب على الويب الى أكتشاف أنماط وصول المستفيد للمحتوى من واقع تحليل ما يعرف بسجلات الويب Web Logs، هذة السجلات التي تقوم بتسجيل كافة التفاعلات من نقر وتسجيل للبيانات يقوم بها المستفيد أثناء تصفحه للمحتوى، وفي هذا تعتمد هذة الفئة على الكثير من خوارزميات التنقيب عن البيانات والتي سبق الإشارة اليها.

إن جوهر الإختلاف بين كل من التنقيب على الويب Web Mining وبين التنقيب على البيانات Data Mining يكمن في مرحلة جمع البيانات Data Collect، ففي نظم التنقيب عن البيانات التقليدية تحظى البيانات بالحفظ والتنظيم والهيكلة في نظم إدارة قواعد البيانات Database Management System أو في مستودعات البيانات Data Warehouse مما يكفل القدرة على معالجتها بصورة أيسر وأسرع وتحقيق التكاملية من خلالها على نطاق كبير. ولكن يختلف وضع التجميع في التنقيب على الويب فهي تعد مهمة في غاية الصعوبة والمشقة والتداخل والتكرار نظرا لما تكفله الويب من ديموقراطية للنشر ونظرا أيضا لعدم هيكلة البيانات الويب ضمن قوالب أو أشكال محددة، هذا الأمر يعمل على توفير آليات وخوارزميات للمعالجة تختلف بشكل كبير عن المستخدمة في التنقيب داخل قواعد البيانات[4].

آليات التنقيب على الويب:

1- زواحف الويب Web Crawlers:

وتعرف أيضا بالعناكب Spiders والآليات Robots وهي عبارة عن برنامج يعمل على تجميع والتقاط المحتوى من على شبكة الويب، حيث يقوم بتصفح شبكة الويب من خلال تتبع الروابط الفائقة بين المواقع ومصادر

[4] Kosala, R., and Blockeel, H.(2000), "Web Mining Research: A Survey," SIGKDD Explorations, 2(1), June 2000. Available at http://www.umiacs.umd.edu/~joseph/classes/enee752/Fall09/survey-2000.pdf

المعلومات للوصول إلى الصفحات التي تشتمل على المحتوى. ثم استخراج المحددات الفريدة للمصادر URLs وإعطائها إلى وحدة التحكم للزاحف.

جدير بالذكر أن هذه البرامج تتعدد أوجه وأغراض استخدامها وقد قام بتحديدها كل من Christopher Olston & Marc Najork على النحو الاتي:

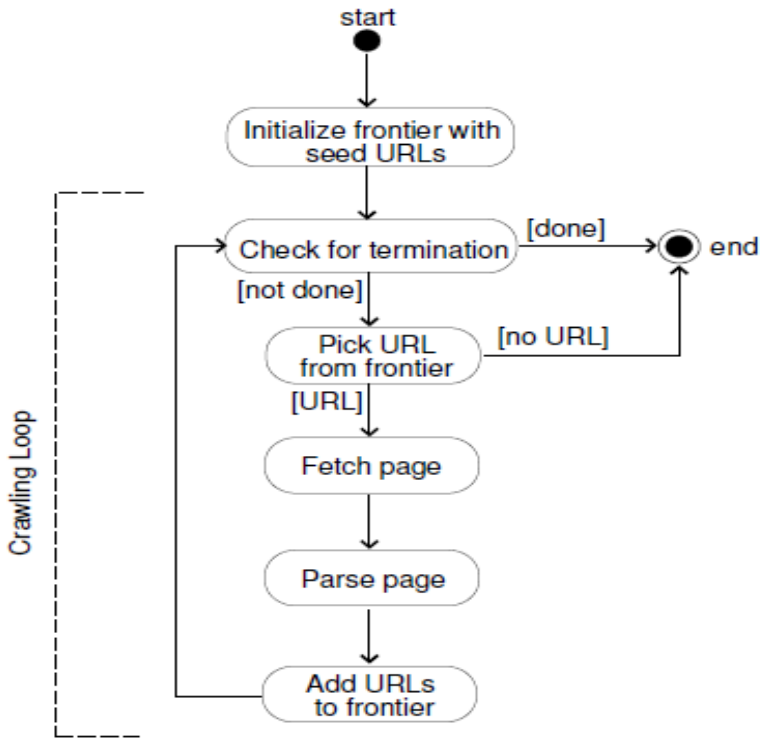
- أحد المكونات الاساسية لمحركات البحث لتجميع صفحات العنكبوتية.
- ارشفة محتوى الويب.
- التنقيب عن البيانات على الويب.
- لرصد نمو الويب للخروج بدلالات احصائية[5].

تبدأ عملية الزحف بأن يقوم الزاحف بتحديد أولوياته من عملية تجميع المحتوى وذلك وفقا للغة أو المكان أو الموضوعات، وذلك في ظل الضخامة التي تتمتع بها شبكة الويب من حيث حجم محتواها المتاح، الأمر الذي لا يجعل الزاحف قادرًا الإ على تجميع جزء ضئيل جدًا من محتوى الويب خلال فترة زمنية محددة[6].

ولا يقتصر الأمر على ضخامة حجم الويب فحسب، بل يزداد الأمر سوءاً في ظل ارتفاع معدلات التغيير والتحديث للمواقع خلال فترات زمنية متلاحقة، وعليه قد ينقضي الأمر بأن الصفحة التي قد قام الزاحف بتجميعها تخضع لإحتمال أنها قد حذفت أو عدلت أو حدثت قبل أن يتم تكشيفها (كصفحات الأرصاد الجوية والعملات مثلا)[7].

يعد من الضروري أن يتمتع الزاحف ببنية معمارية ذات أداء قوي ومرتفع، ومع ذلك فإن بناء زاحف قوي يرتبط بالعديد من التحديات المتعلقة بكفاءة الشبكة المكشفة وفاعليتها، ورغم ذلك فإن خوارزميات الزاحف وتفاصيل أداءه تحاط بسرية خاصة من قبل البرمجيات والنظم المستخدمة لها، حيث لا تعلن الشركات عن خوارزميات زحفها، حتى أنه عنما يتم نشر تصاميم الزاحف فكثير من التفاصيل الهامة في بنيته لا يتم نشرها أو ذكرها، ولعل المرجعية في ذلك ترجع إلى أن سمة تخوف لدى مطوري البرمجيات والنظم من الإعلان عن خوارزميات زواحفها فتصبح فريسة سهلة لبرمجيات خداع محركات البحث Search engine Spammer. اما البنية النموذجية لزواحف محركات البحث فتتمثل في الشكل الاتي:

[5]Christopher Olston and Marc Najork.Web Crawling.Foundations and Trends in Information Retrieval.Vol. 4, No. 3 (2010).
[6] 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
[7] Ibid.



شكل رقم 9 يوضح خريطة تدفق توضح بنية الزاحف منهجيته في الزحف[8]

تبدأ عملية الزحف من قبل مطوري برمجيات ونظم التنقيب على الويب من خلال تزويد الزاحف بمجموعة من عناوين المواقع والتي لم تتم زيارتها list of unvisited URLs والتي تعرف باسم حدود أو جبهة الزاحف the frontier وهذه القائمة تهيئ كمحددات بذرية seed points حيث يتم توفيرها يدويا أو من خلال برنامج أخر كأدلة البحث yahoo، حيث أن كل عملية زحف تنطوي على اختيار المحدد التالي من جبهة الزاحف، ثم يتم بعد ذلك جلب fetching الصفحات الموافقة لـ URLs من خلال بروتكول النص الفائق HTTP، ثم يتم بعد ذلك وضع هذه 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 هذه العملية بالنسبة للصفحات الجديدة، ويمكن انهاء عملية الزحف عند عدد معين من الصفحات، أم اذا كان الزاحف لديه القدرة على الزحف فيمكن أن يكمل عملية الزحف، ولكن تتسم جبهته حين إذ بالفراغ، وذلك يؤدي إلى الوصول إلى طريق مسدود للزحف dead-end، وهو ما يشير إلى أن هناك عمقا لعملية التجميع Crawling Depth وقد وجد ان عمق التجميع المثالي يتراوح ما بين 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 هذه الروابط في قائمة الجلب للاسترجاع.

[8] 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

- الاستمرار في عملية الزحف حتى التوقف عند عدد معين من الصفحات محدد سلفا أو فراغ قائمة الجلب[9].

2- التأخيذ والتحليل Tokenization and Analysis:

تتمثل المرحلة الثانية في التنقيب على الويب هي استخلاص المعلومات Web information extraction ويتخذ استخلاص المعلومات على الويب نمطين النمط الأول هو استخلاص المعلومات من نص اللغات الطبيعية للمحتوى، والنمط الثاني هو استخلاص المعلومات البيانات المهيكلة من صفحات الويب، ويحظى النمط الأول بالاهتمام من جانب دراسات معالجة اللغة الطبيعية Natural language processing، بينما يمثل النمط الثانئ جوهر المعالجة في هذة الدراسة، وعادة ما يسمى البرنامج الذي يضطلع باستخراج هذة البيانات بأسم المجمع Wrapper[10].

تأتي وثائق شبكة الويب تأتي في العديد من التنسيقات المختلفة من صيغ html , pdf وغيرها من تنسيقات الصفحات، وتتمثل المرحلة الاولى في عملية الاستخراج هو إيجاد رؤية منطقية أو نماذج models قياسية للوثائق، ويعد "نموذج حقيبة الكلمات Bag of Words" أكثر الرؤى المنطقية استخداما في استخراج المعرفة، 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 برمجيات التنقيب عن محتوى الويب.
ووفقا لهذاالنموذج ايضا 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 النصوص وفقا لشكل موحد، حيث تكفل هذه 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 وتشتمل عملية تطبيع النصوص وفقا للشكل على مجموعة من العمليات الفرعية:

- التأخيذ Tokenization :

وتشمل هذه العملية على تفتيت النصوص الكاملة لكلمات وتحديدها، وتعد هذه 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 لا يجدي الامر نفعا في لغات أخرى خاصة في لغات الشرق اسيوية حيث تتشابك بنيتها دون فواصل أو محددات مثل اللغة الصينية[11].

- قائمة الاستبعاد stop words:

وتتمثل في استبعاد الكلمات التي تحمل دلالات معلوماتية ولغوية ضئيلة في الوثيقة، وفي نظم استرجاع المعلومات عادة ما يتم التخلص من هذه الكلمات لأسباب تتعلق بالكفاءة.[12].

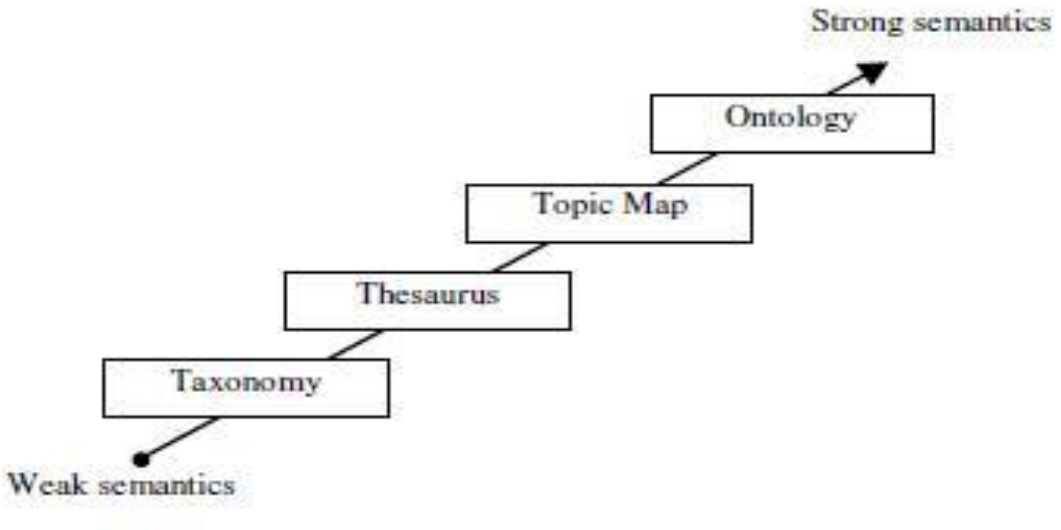
- جذور الكلمات stemming:

تعمل هذه المنهجية على استخراج الجذور الصرفية لكل كلمات الوثيقة،[13].

3- نماذج هيكلة البيانات Web Data Model:

[9]Ibid.
[10] Liu, B. (2007). *Web data mining exploring hyperlinks, contents, and usage data*. Berlin: Springer.
[11]Baeza-Yates, R. and Castillo, C.(2005) "Web Search." Proceedings of the third Workshop on Web Graphs (WAW), Vol. 3243 of Lec-ture Notes in Computer Science, pp. 156-167, Rome, Italy, Spring,..
[12] Ibid.
[13] Ibid.

سعى الباحثون في مضمار التنقيب عن المحتوى المتاح على الويب على توفير نماذج Models وتقنيات تعمل على تحقيق التكامل المعرفي والدلالي بصورة آلية، بحيث تهدف هذة النماذج بصورة أو بأخرى إلى تمثيل المعرفة البشرية في صورة لغة مهيكلة ذات بناء منطقي وذلك من خلال نمذجة المفاهيم والدلالات لتمكن ادوات وبرمجيات الويب المختلفة من التعامل معها وعلى هذا انصبت كثير من جهود العديد من المبادرات حول امكانية تمثيل المعرفة والمفاهيم في صورة نماذج دلالالية Semantic Modeling والتي يمكن تصوريرها كما هو في الشكل () [14].



شكل رقم (10) يوضح تقنيات ومنهجيات التنقيب عن المحتوى للوصول إلى التكامل المعرفي[15]

ويمكن تقسيم النماذج التي يعتمد عليها التنقيب عن المحتوى إلى فئتين:

- **الفئة الأولى: التقنيات التقليدية:**

وهي تلك التقنيات التي ظهرت قبل ظهور مفهوم التنقيب عن البيانات، حيث ظهرت في سياق تخصص المكتبات وعلوم المعلومات والتي كانت تهدف إلى ضبط مصادر المعلومات في سياق نظم الاسترجاع التقليدية كالفهارس وقواعد البيانات الببليوجرافية ونظم استرجاع الحقائق كبنوك المعلومات ويأتي على رأس هذة الأدوات:

1- قوائم رؤوس الموضوعات Subject Headings:

أحد اهم المنهجيات والأدوات التي هدفت إلى حصر المفاهيم والمعرفة في صورة مفردات لغوية، تنظم هذة المفردات اللغوية في نسق هجائي عادة، ما توفره هذة الأداة القدرة إلى توفير المترادفات اللغوية للمصطلح المستخدم للتعبير عن المفهوم وبالتالي يمكن استثمار هذة البينة والأداة في احالة استفسار المستفيد من المصطلح غير المستخدم في كشاف محرك البحث إلى المصطلح المستخدم والمكشف بها الوثيقة في بينة محركات البحث أي توجيه عملية البحث وفقا لسياق محدد[16].

[14] Lim, E. H., Liu, J. N., & Lee, R. S. (2011). Knowledge seeker ontology modelling for information search and management : a compendium. New York: Springer.
[15] Ibid.
[16] Ibid.

2- نظم التصنيف الموضوعية Taxonomy:

وهي نظم التصنيف التي تعمل على تمثيل المفاهيم والمعرفة في صورة هرمية تتداعى فيها القطاعات الموضوعية من الأعم إلى العام إلى الخاص إلى الاكثر خصوصية، وقد تستخدم الالفاظ في التعبير عن المفاهيم أو قد تستخدم رموز اخرى كالارقام كما هو الحال في تصنيف ديوي العشري، وتركز هذة الأداة على توضيح نمط علاقات البنوة والبوة والأخوة بين المفاهيم وبالتالي فمنطق ما ينطبق على القطاع الرئيسي ينسحب بطبيعة الحال على القطاع الفرعي، وأن القطاعات المناظرة لبعضها تشترك في بعض القواسم المشتركة ومن ثم القدرة على استثمار هذة البنية في تحديد المفاهيم المتصلة بموضوع الاستفسار المقدم لمحركات البحث، وقد طورت شركة SUN Microsystems أحد أهم النماذج التصنيفية التي تعمل على التكشيف المفاهيمي داخل محركات البحث، والتي عرفت فيما بعد بخوارزمية MMS في إعداد الفئات التصنيفية[17].

3- المكانز الموضوعية Thesaurus:

أحد اهم فئات المراجع التي تعمل على بناء شبكة من العلاقات بين المفردات اللغوية والمفاهيمية في مجال موضوعي محدد، وتتخذ هذة العلاقات صور عديدة منها الهرمية Hierarchical والتكافؤية Equivalence والترابطية Association، تنظم المفردات والواصفات في بنية المكنز وفقا للتشابه الدلالي في المعاني وقد تتوسع بنية المكنز لتشتمل على المرادفات والأضداد، ما يمثل قيمة يمكن لمحركات البحث أن تتثمرها يتمثل في استعلال بنية الروابط بين المفاهيم لتحقيق تكاملية في الاسترجاع، كذلك الاعتماد على المترادفات والاضداد لتحديد المفهوم المقصود المعبر عنه في استفسار المستفيد[18].

- الفئة الثانية: التتقنيات الحديثة:

وهي تلك التقنيات التي ظهرت في مضمار الاعتماد على نظريات الجبر الخطي، التفاضل والتكامل، ومجال حوسبة اللغة من جانب، ومن جانب اخر اشتملت هذة الفئة على التقنيات التي ظهرت في مضمار الويب الدلالي، والتي كان يهدف مطوريها منها أن تستخدم داخل بنية نظم التنقيب عن البيانات لتحسين كفاءة البحث والاسترجاع وتشتمل هذة التقنيات على:

1- التكشيف الدلالي الكامن: Latent Semantic Index:

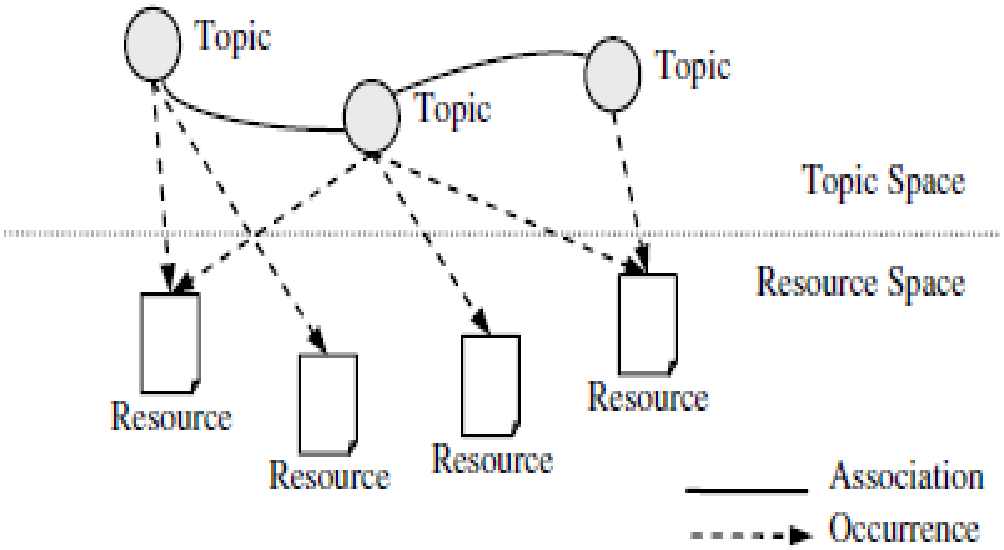
أحد منهجيات التكشيف والاسترجاع التي ظهرت قبل ظهور الويب، والتي تعمل على توظيف التقنيات الرياضية بصورة عامة وتقنية الجبر الخطي لتحقيق معدلات استرجاع مرتفعة، ويعد أحد اشهر المنهجيات التي قامت هذة التقنية بتوظيفها منهجية تحليل القيم المفردة Singular Value Decomposition (SVD) والتي تعمل على تحديد أنماط العلاقات بين المصطلحات الواردة في النص والمفاهيم الضمنية أو المعبر عنها من خلال هذة المصطلحات. أن المنطق الذي تلجأ اليه هذة التقنية هو تكشيف السياق الخاص بالنصوص المكشفة حيث تقوم هذة التقنية بإستخراج المحتوى المفاهيمي الوارد في النص ثم انشاء روابط بين المصطلحات المعبرة عن المفاهيم وبين ما تشير اليه من مفاهيم ، ومن جانب اخر تلجأ إلى عمل كشاف بالقيم Values المختلفة للمفاهيم وربطها بالمصطلح المستخدم مما يؤدي إلى ان استرجاع أي منهما يؤدي إلى استرجاع الاخر. ما كفلته هذة المنهجية من مميزات تمثل في قدرتها على تمثيل الابعاد المختلفة للمفهوم مما يوفر القدرة على تحقيق التكامل على صعيد المفهوم بين المصطلحات الواردة في الوثيقة الواحدة وبين نظائرها ومرادفاتها الواردة في الوثائق الاخرى. تعمل هذة التقنية ايضا على تحديد درجات مسبقة للالفاظ المستخدمة في التعبير عن المفهوم الواحد بين

[17] Ibid.
[18] Ibid.

ماهو شائع منها في الاستخدام وبين ما هو نادر الاستخدام وذلك وفقا لمعادلات رياضية تعتمد على السياقات التي وردت فيها هذه الالفاظ. كفلت هذه المنهجية ايضا البعد عما يعرف باستقلالية المصطلح Term Depended والتي امتازت به التقنيات والمنهجيات السابقة عليها كنموذج البوليني ونموذج Vector Space وغيرها من النماذج التي عملت على تكشيف الالفاظ في استقلالية تامة عن السياق التي وردت فيه، الأمر الذي يغفل فرضية أن المصطلحات ذات القرابة وذات الصلة تتصل ببعضها البعض في نسق مفاهيمي في ظل اشتراكها في ذات السياق وهو الأمر الذي قامت منهجية التكشيف الدلالي باستثماره. على الرغم من اتسام تقنية التكشيف الدلالي الكامن بالكثير من المميزات إلى أن يكتنفها جوانب قصور تتمثل في قابلية التوسع لتشمل أو تتعامل مع مجموعات ضخمة من الوثائق، فما تقوم به هذه التقنية يتمثل في ربط المصطلحات بعضها البعض من جانب وربط المصطلحات المختلفة بالمفهوم الواحد الأمر الذي يستتبع إنشاء كشافات وقواعد بيانات ضخمة تكاد تكون بليونية، وبالتالي يصعب تطبيق هذة التقنية في ظل ضخامة وحجم الويب واتساع حجمه وتنوع لغات وأنواع وثائقه الأمر الذي يجعل هذة التقنية تقف عاجزة عن تحقيق غايتها أمام هذا الكم من مصادر المعلومات[19].

2- خرائط الموضوعات Topic Maps:

هي مواصفة معيارية دولية صدرت عن المنظمة الدولية للمواصفات القياسية ISO، هدفت هذة المواصفة الي تمثيل المعلومات من خلال نموذج مهيكل structured information model وتمثيل العلاقة بين المفاهيم المجردة ومصادر المعلومات، تاتي بنية هذا النموذج في مستويين كما هو موضح في الشكل رقم (3/2) المستوى الأول يعرف باسم المستوى الموضوعي Topic Space والذي يشير إلى الموضوعات التي تمثل المفاهيم المجردة، المستوى الثاني يتمثل مستوى مصادر المعلومات Resource Space والذي يتكون من مصادر المعلومات الرقمية، ويتصل كلا من المستويين بعضهما البعض من خلال مجموعة من الروابط الموصفة والمحددة، فالموضوعات ترتبط مع بعضها البعض من خلال روابط تعرف باسم connections association وتتصل المصادر بالمفاهيم التي تعبر عنها بنمط من الروابط يعرف باسم occurrence connection[20].



[19] Ibid.
[20] Ibid.

شكل رقم (11) يوضح بنية خرائط الموضوعات[21]

3- الشبكات الدلالية Semantic Network:

وهي أحدى المنهجيات التي تعتمد على معالجة اللغة الطبيعية NPL من أجل انشاء هياكل ومخططات للمفاهيم، حيث تعتمد على تكشيف الكيانات في صورة شبكة ترتبط مع بعضها البعض من خلال العلاقات بين المفاهيم المعبرة عنها، بل يمتد الأمر في أنها تعمل على أعادة هيكلة الأستفسار المقدم لها وفقا لتقنيات معالجة اللغة الطبيعية ويعد أشهر المحركات البحث التي تعتمد على هذة التقنية هو محرك LexiGuide[2223].

4- وثائق انطولوجيات الويب :

تعمل لغة انطولوجيا الويب Web Ontology Language على اثراء محتوى الويب من خلال جعله اكثر تعبيرا ودلالة، وذلك من خلال العمل على تسكين المصادر في صورة كيانات وفقا لخصائصها Property داخل فئات أوقطاعات موضوعية Classes ثم تحديد الروابط بين هذة الكيانات وتوصيفها في صورة تسمح للبرمجيات والآليات من أدراكها والتعامل معها من خلال قدرة هذة اللغة على التعامل بصورة منطقية تحكمها قواعد المنطق الجبري والاستدلالي. وسيفرد له الباحث الجزء الأخير من الدراسة.

خوارزميات التنقيب على الويب:

وفقا لرؤية Ricardo Baeza Yates فجوهر التحديات التي تواجه التكامل المعرفي على الويب يكمن في تحليل بنية الويب من الروابط الفائقة والمحتوى حيث أوضح " أن المشكلة الرئيسية في التنقيب على الويب بشكل خاص، تكمن في قضية التنبؤ بتحديد أي من محتوى الوثائق يتسم بالصلة، وأي منها لا يتسم بالصلة.[24]".

ومثل هذا القرار لا يخضع إلى الحدس أو التخمين بل يستند وبشكل أساسي على خوارزميات تعمل على تحليل هيكلية الويب من الروابط والمحتوى.

ويوضح B.Liu "أن يجب الأخذ في الأعتبار أن تقييم وفرز وترتيب المحتوى أعتمادا على الروابط القائقة لايقتصر استخدامه على محركات البحث فحسب، بل تشكل هذة الروابط جوهر مؤشرات التنقيب على الويب نظرا لما يمكن أن تحمله من دلالات تساعد على تحقيق التكامل المعرفي[25]".

حيث تساعد الروابط الفائقة في إيجاد ما يعرف بمجموعات الويب، تعرف مجموعات الويب Web communities بإنها مجموعة من الصفحات المرتبطة مع بعضها البعض لتمثيل مجال اهتمامي لمجموعة من الأفراد، وبناء علي تحليل هذة الروابط يمكن الخروج بمؤشرات حول كيانات وظواهر اجتماعية وغيرها[26].

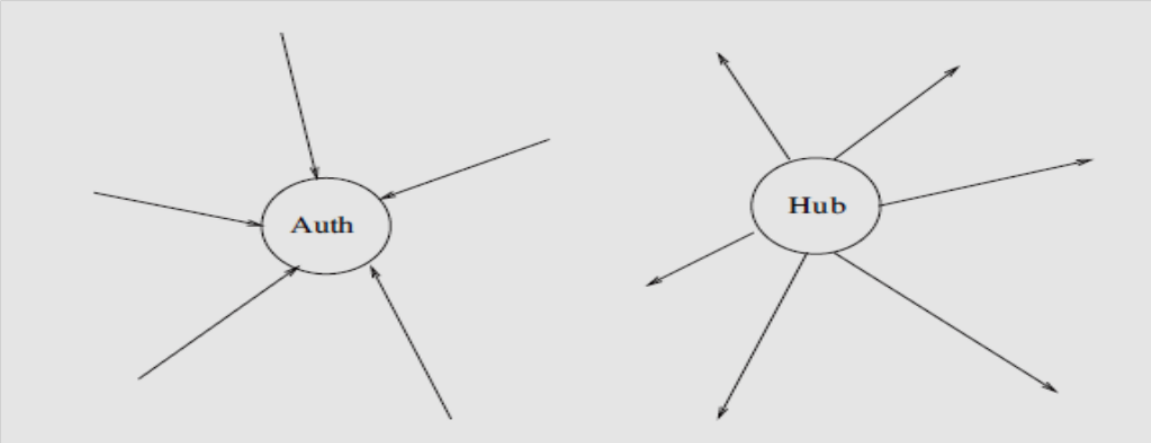
1- خوارزمية الروابط الفائقة الناتجة عن البحث الموضوعي (hyperlink-induced topic search) (HITS):

طورت هذه الخوارزمية عام 1997 على يد Jon Kleinberg في نفس الوقت الذي طورت فيه خوارزمية الترتيب الطبقي Page Rank، وتعتمد هذه الخوارزمية على اكتشاف وترتيب محتوى الوثيقة ذات الصلة بموضوع محدد – وتعد هذه الخوارزمية الآن جزءا أساسي في خوارزميات محرك البحث

[21] Lim, Edward H. Y., J. N. K. Liu, and R. S. T. Lee. Knowledge seeker ontology modelling for information search and management : a compendium. New York: Springer, 2011. Print.
[22] http://www.lexiquest.com/
[23] Ibid.
[24] Yates, R., & Neto, B. (1999).*Modern information retrieval* . New York: ACM Press.
[25] B. Liu (2011), *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Data-Centric Systems and Applications, DOI 10.1007/978-3-642-19460-3_7, © Springer-Verlag Berlin Heidelberg 2011
[26] Ibid.

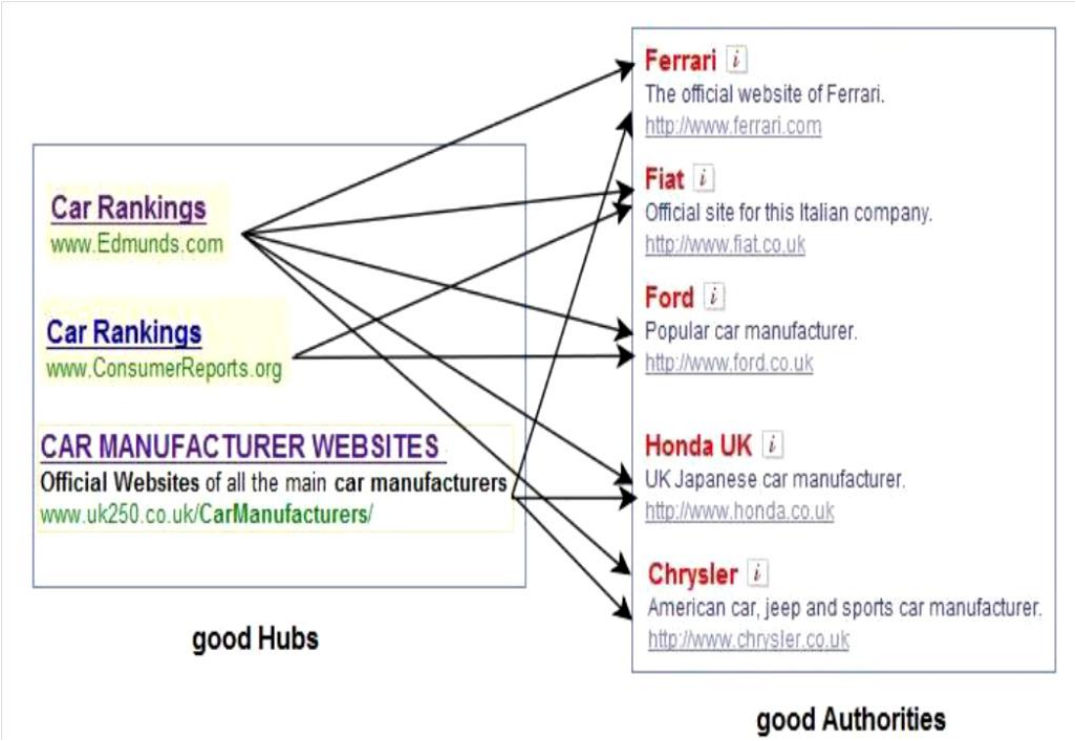
Ask(www.ask.com)–هذه الخوارزمية تعتمد على أن يوجه المستفيد أولا الاستفسار لأداة البحث ثم تسترجع النتائج من الكشاف أو قاعدة البيانات لبدء مرحلة الترتيب للنتائج وفقا لعنصرين أساسين هما:

- المواقع الارتكازية Hubs Nodes: وهي المواقع التي تشتمل على محتوى يصدر منه الرابط في اشارة منها لمواقع الاستنادية.
- المواقع الاستنادية أو ذات الموثوقيةAuthorities Nodes: ويقصد بها المواقع التي يرد اليها الرابط من قبل المواقع المحورية.



شكل رقم 12 يوضح بنية المواقع الارتكازية والمواقع الاستنادية.

فلو افترضنا أن الصفحة (i) تشتمل على محتوى ذو موثوقية authority للاستفسار المقدم لمحرك البحث عن " اشهر صناع المركبات (السيارات)" ، حيث تعد الصفحات الرسمية لمنتجي السيارات بمثابة صفحات استناديه ذات موثوقية لهذه العملية البحثية كموقع تويوتا ومرسيدس وفيات وغيرها، كما تعد مواقع وصفحات وكلاء المبيعات لهذه السيارات بمثابة صفحات استناديه ايضا للموضوع المشار اليه ويتم 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 أخرى كما هو موضح في الشكل الاتي:



شكل رقم 13 يوضح الاسترجاع وفقا للمواقع الارتكازية

وعلى هذا فإن كل محتوى يحظى بدرجتين واحدة للمواقع التي تشير اليها - وهي في هذه الحالة نقطة ارتكازية - واخرى للمواقع التي تشير إلى هذه الصفحة، ويحدد على أساسهما 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[27].

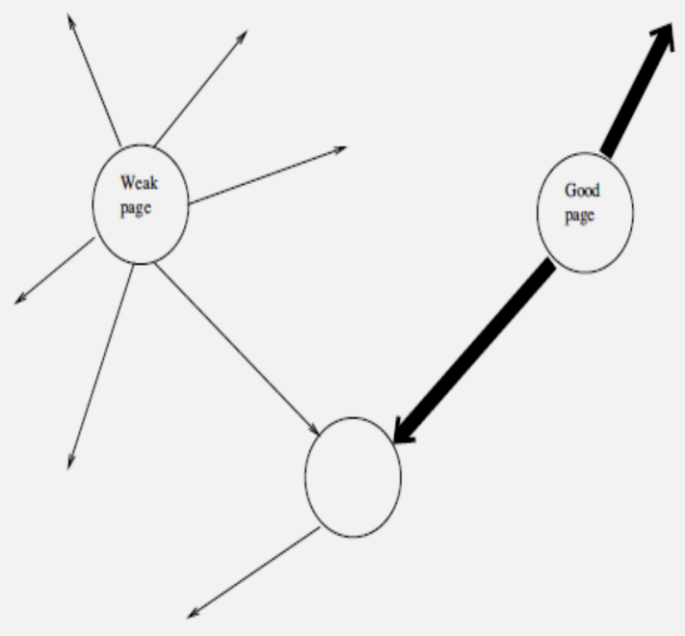
2- خوارزمية الترتيب الطبقي للصفحات PageRank:

طور هذا النموذج على يد كلا من Sergey Brin و Lawrence Page عام 1997، وقد عرف هذا النموذج بأنه المنهجية التي تعني بحساب رتبة محتوى كل صفحة على الويب اعتمادا على نمذجة الويب في مخطط بياني (web Graph) قائم على الروابط والمواقع. ولقياس جدوى هذه الخوارزمية قاما كلا من Brin و Page بتصميم محرك البحث الشهير Google.

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

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

- تمثل الروابط مؤشرات جيدة لتحديد أهمية محتوى الوثيقة التي تشير اليها.
- الروابط الصادرة من الوثائق التي تحظى بأهمية في موضوعها تعد مؤشرا جيدا لجودة الوثيقة التي تشير اليها، عن الوثيقة التي يشار اليها من قبل وثائق أقل في الأهمية والجودة كما هو موضح في الشكل الاتي:



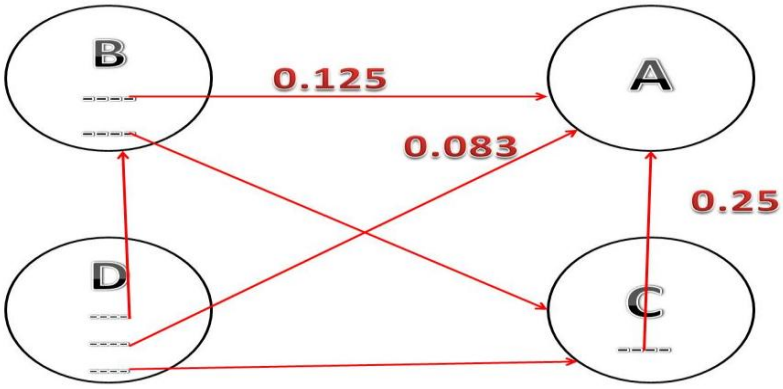
شكل رقم 14 يوضح منهجية PageRank ويوضح دلالة ان الرابط الفائق يكتسب قوة من قوة الصفحة ومحتواها.

الخوارزمية:

بداية يجدر الإشارة إلى ان خوارزمية الترتيب الطبقي (the PageRank) صدرت في أكثر من صيغة وأكثر من معادلة متتالية، وسيتعرض الدراسة في هذا المقام إلى الصيغة البسيطة من هذه المعادلات.

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

وبافتراض وجود بنية بيانية لشبكة عنكبوتية تتكون من 4 نقاط (NODES) (أربع صفحات) A,B,C and D وأن أهمية هذه 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 هذه الصفحات مجموعة من الروابط والتي سيتم الاعتماد عليها لحساب رتبة الصفحة A، هذه الروابط تتمثل في الشكل الاتي:



شكل رقم 15 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 هذه 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:

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

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

1- الانطولوجيات والتنقيب على الويب Ontology-based Web mining**:**

أظهرت تقنيات التنقيب على الويب – السابق الإشارة إليها - نتائج واعدة في سياق تحقيق التكامل المعرفي على صعيد محتوى الويب، ولكن في المقابل أتسمت هذة التقنيات بالمحدودية في المعالجة نظرا لعجزها الكامل عن توفير الدلالات والمفاهيم المرجوة في معالجة كل من نص المحتوى، والروابط بين كيانات الويب، وفي هذا الصدد تأتي أحد أهم التقنيات التي طورت في سياق نظم الذكاء الاصطناعي والمتمثلة في الانطولوجيا Ontology حاملة معها القدرة توفير أطر مفاهيمية ودلالية وعلائقية تعمل من خلالها على توصيف المعلومات وتحديد اوجة العلاقات بين كيانات المعلومات في سياق أو مجال موضوعي محدد، ومن ثم عمل التنقيب على الويب على توظيفها لسد ما تتسم به تقنياته من عجز على صعيد الدلالات والمفاهيم، حيث أرتكو توظيف الانطولوجيا في سياق التنقيب على الويب على توفير القدرة على تصنيف الموضوعي والدلالي لمحتوى الويب وربط النتائج بالكيانات مما يوفر فرص عظيمة لفهم والتبادل للمعلومات على السواء بين البشر وبرمجيات الويب، وعلى هذا تعمل هذة الورقة البحثية على تقديم وصفا كاملا عن مفهوم الانطولوجيا وأوجه استثمارها في التنقيب على الويب.

مفهوم الانطولوجيات The Ontology**:**

تأتي أول الخطوات الأساسية لتجهيز المحتوى للتحقيق التكامل المعرفي هي خطوة إعداد ما يعرف بالأنطولوجيات Ontology، فيجب أن يتم إعداد محتوى الويب سلفاً حتى تتمكن البرمجيات والتطبيقات كمحركات البحث وقواعد البيانات من التعامل معه، ويتم ذلك من خلال الأنطولوجيات التي تعمل على أن توفر للآلة الفهم الخاص بدلالات ومعاني البيانات، ومن ثم القدرة على تشغيلها واستثمارها وتحقيق التكامل فيما بين البيانات، وهو ما يعرف بالتوافق الدلالي semantic compatibility والذي يعني أن ما يفهمه التطبيق (أ) هو نفس ما يفهمه التطبيق (ب) والتطبيق (ج)، هذا بجانب مايعرف بالتوافق التركيبي syntax compatibility والذي يعمل على التعرف على بنى البيانات بشكل صحيح.

يرجع تاريخ لفظة الأنطولوجيا ontology إلى مجال الفلسفة حيث يمثل مسمى لفرع أساسي من فروع علم الميتافيزيقا والذي يركز على تحليل أنماط الوجود والموجودات Entity، والبحث في كينونة الكيانات المختلفة والاشياء وانواعها وبنيتها وخصائصها ووقائعها وعملياتها.

- أولاً: التعريف اللغوي:

جاء التأصيل اللغوي لكلمة الأنطولوجيا ontology على يد فلاسفة العصور القديمة وبالتحديد فلاسفة المدرسة الآرسطية الفلسفية التي اعتمدت على تحكيم المنطق في الوصف والتحليل وعملت على تصنيف الحياة إلى فئتين :-

1. الوجوديات:اتسمت ملامحها بدراسة الوجود والتكوين الفعلي المادي فجاءت الأنطولوجيا معبرة عن هذا القصد.

2. الغيبيات: التي ارتسمت ملامحها بالخفية والتكوين الهلامي فجاءت الميتافيزيقا للدلالة على هذا المعنى.

وفي العصر الحديث صك مصطلح الأنطولوجيا في الدراسات اللاتينية على يد كلاً من الفيلسوف Rudolf Göckel و Jacob Lorhard في المعجم الفلسفي عام 1613م[28].

ظهرت كلمة ontology في اللغة الإنجليزية عام 1721 في قاموس Bailey's والذي عرفها بأنها إحصاء للموجودات في العالم الحقيقي وفي هذا أرجع قاموس ويبيستر WEBESTER لظهورها لنفس العام وقد عرفها قاموس اكسفورد بانها: فرع من فروع الفلسفة معني بدراسة طبيعة الوجود.

ثانيا: التعريف الاصطلاحي:

على الرغم من أن جاء مصطلح الأنطولوجي فلسفي المولد والنشأة والدلالة اللغوية، إلا أنه في القرن العشرين استعانت به العديد من المجالات المعرفية في تطوير أسسها النظرية واختباراتها، فضلاً عما تمثله الأنطولوجيا الآن من كونها أطر معرفية لكل من (علوم الجبر وعلوم المعلومات والويب الدلالي والطوبولوجيا وغيرها) إلا أن مجال هندسة الحاسبات وعلوم المعلومات كان أوفر حظاً من غيرها، فقد أدخل هذا المصطلح من قبل أوائل الباحثين في مجال الذكاء الاصطناعي في فترة الثمانينات من القرن العشرين في إشارة إلى نظرية نمذجة تمثيل المعرفة Knowledge Representation.

[28]Smith, B. and Welty, C. (2001) Ontology-towards a new synthesis. Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS2001). ACM Press,.

وفي أوائل التسعينات أثمرت العديد من الجهود في مجال الحاسبات والذكاء الاصطناعي إلى إرساخ مفهوم الأنطولوجيا في العديد من المجالات منها هندسة النظم – المعلومات الحيوية bio-information– ومجال المكتبات والمعلومات كمصطلح موضوعي في هذه المجالات

ويرجع الفضل في ذلك إلى دراسة Gruber التي كانت أولى الدراسات التي أرست هذا المصطلح والتي نشرت عام 1992 وقد عرفته على أنه نمذجة تسعى إلى التمثيل المفاهيمي للمجالات والقطاعات المعرفية، وما تشمله من معرفة فرعية وما بداخلها من كيانات وعلاقات تربط بين هذه الكيانات[29].

كما عرفها قاموس Harrods للمكتبات على انها " المجال الذي يعمل توفير مجموعة مشتركة من المصطلحات والمفاهيم وتحديد العلاقات بينهما بدقة شديدة من أجل توصيف مجال موضوعي أو معرفي محدد بطريقة تمكن نظم الحاسبات من الاتصال والتواصل مع بعضها البعض بحيث تكون بمنأى عن نظم التشغيل الفردية ومعماريات المعلومات ونطاقات البرامج[30]".

تعد الأنطولوجيات بمثابة حلقة الوصل بين فهم الاليات وفهم البشر فبناتج توصيف العلاقات والمفاهيم والمصطلحات تستطيع الآلة أن تفسر وتعالج البيانات في أطار من المنطقية والاستدلال وبالنسبة للبشر فإن أدراك المخططات التفصيلية والتقسيمات الموضوعية يعد أسهل من إدراك النصوص المسردة وعلى هذا فالأنطولوجيا تعمل على توفير هيكل مفاهيمي للمجالات المعرفية لكل من الآلة والأنسان ولهذا تأتي طبقة الأنطولوجيا في معمارية الويب الدلالي أعلى من كلا من RDF& RDF schema.

تعمل الأنطولوجيا على التعريف بجانبين مهمين في أي مجال معرفي:

- الجانب الأول: هو تصنيف المعرفة البشرية وكيف تتداعى المفاهيم المعبر عنها بالمصطلحات والمفردات تداعيا منطقيا داخل مجال موضوعي محدد.

- الجانب الثاني: هو القدرة على الاستدلال من خلال منطقية التوصيف لكل من المفاهيم والعلاقات بصورة تتنامى من خلالها القدرة على الاستدلال واستنباط معلومات جديدة من واقع ما هو متاح من معلومات موصوفة وهيكلة لها.

البنية البنائية Syntax للأنطولوجيا:

يعد القاسم المشترك بين مجال الفلسفة ومجال الحاسب الالي في استخدام مصطلح الأنطولوجيا هو تمثيلهما للمعرفة في صورة:

- كيانات Entity: والتي تعرف في مجال التكامل المعرفي ايضا بالمفردات Individuals.
- الافكار Ideas: والتي تعرف في مجال التكامل المعرفي بالفئات Classes.
- الخصائص Properties: والتي تعرف ايضا بالسمات Attribute.

1- المفردات Individuals: يعد المكون الاساسي في بنية الأنطولوجيا وتمثل المستوى الأول داخل الأنطولوجيا وتشير كلمة المفردات Individuals إلى مختلف الكيانات المادية والمجردة.

[29]Gruber, T. R.,(1992) Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies*, 43(5-6):907-928,.

[30] Prytherch, R. J. (2005). *Harrod's librarians' glossary and reference book: a directory of over 10,200 terms, organizations, projects and acronyms in the areas of information management, library science, publishing and archive management* (10th ed.). Aldershot, Hants, England: Ashgate.

2- الفئات Classes:يعد المكون الثاني في بنية الأنطولوجيا وتشير عبارة الفئات Classes إلى التصنيف الاساسي في مجال ما، حيث تشتمل الفئة على مجموعة من المفردات individuals أو الكيانات object والتي تجمع صفات مشتركة جعلتهم ينتمون إلى هذه الفئة التي تم تسكينهم فيها.ولا يقتصر اشتمال الفئات على المفردات فحسب بل تمتد لتشمل ايضا فئات فرعية subclasses لتكون بذلك فئة رئيسية تشتمل على فئات فرعية في تسلسل هرمي وبالتالي فإن اعضاء الفئة الفرعية ينتمون بحكم المنطق إلى الفئة الرئيسية.

3- السمات Attributes: توصف كلا من الفئات Classes والمفردات Individuals في الأنطولوجيا وفق الخصائص المميزة لها عن غيرها والمحددة لذاتها، فبالنسبة للمفردات تقوم السمات بتحديد القيم الخاصة بكل مفردة، وبالنسبة للفئة تقوم بتحديد السمات الاساسية لهذه الفئة والتي يدرج تحتها المفردات.لا يقتصر أمر السمات على توصيف المفردات والفئات بل يمتد ليقوم بتوصيف العلاقات التي تربط هذه الكيانات كلها مع بعضها البعض، حيث تقوم بتحديد طبيعة ونوع العلاقة التي تربط مفردة بمفردة وفئة بفئة ومفردة بفئة.

4- العلاقات Relationship: تعد العلاقات أحد أهم السمات التي تميز الأنطولوجيا حيث تتسم العلاقات في بنية الأنطولوجيا بأنها تتمتع بالتوصيف والمسميات والدلالات الأمر الذي يكفل للحاسب تحقيق التكامل المعرفي بين الكيانات المختلفة.

تنقسم العلاقات في الأنطولوجيا إلى نوعين:

- النوع الأول: علاقة الاستيعاب Subsumption (التضمين الفئوي) ويعد هذا النوع من اهم انواع العلاقات في بنية الأنطولوجي ويعرف هذا النوع بأسم "Is a" "واحد من" والذي يعمل على التعريف ان هذا الكيان عضو في الفئة التي ينتمي اليها، فعلى سبيل المثال: لو لدينا فئة class يعرف بأسم البشر human وهو يستوعب فئتين الذكر maleوالأنثىfemale وذلك بطبيعة حكم المنطق.

- اما النوع الثاني فهي العلاقة المخصصة Custom: والتي يقوم منشئ الأنطولوجيابإعدادها من خلال مفهومه ومنطقه ورؤيته التصورية للكيانات وما يمكن أن ينتمي اليها وتتنوع فيها مسميات العلاقة.

لغات انطولوجيا الويب:

تعرف لغة الأنطولوجيا بأنها مجموعة من الواصفات الدلالية التي يمكن استخدامها للتعريف ولتحديد أية بنية (كيانات، فئات، مفاهيم، علاقات، ضوابط) في أي مجال معرفي. حيث تعمل على توفير مجموعة من المسلمات البديهية التي تمكن الآلة والتطبيقات من التفكير والاستدلال.

جدير بالذكر أن تطوير لغات انطولوجيا الويب لايعني استبدالها بلغات HTML و XML فالانطولوجيا تهدف الى تحقيق التكامل المعرفي على صعيد شبكة الويب بإسرها.

بالرغم من كون لغة html تمثل اللغة الاساس لتكويد المحتوى على الويب الحالية الا إنها لاتدعم أية خصائص أو مواصفات لازمة لبناء لغات انطولوجيا الويب وهذا الامر يمتد ليشمل لغة xml ايضا، فبرغم مما تلعبة هذة اللغة من دورا هاما في التكامل المعرفي الإ انه لايمكن النظر اليها بإنها لغة انطولوجيا للويب نظرا لافتقارها إلى القدرة على التعبير عن دلالات البيانات التي تقوم بتكويدها، كما أنه لايمكن الاعتماد عليها في الاستدلال.

1- لغة الاستدلال الآنطولوجي OIL ontology inference language:

تم تطوير هذة اللغة من قبل مشروع ontoknowledge أحد مشروعات الاتحاد الاوروبي وكانت تهدف هذة اللغة إلى توفير الفرصة لوجود التشغيل البيني الدلالي بين المصادر المتاحة على الويب.

2- لغة DAML : darpa agent markup language

قدمت هذة اللغة ضمن مشروع Defense Advanced Research Projects Agency عام 200 وكان مبتكر الويب الدلالي Tim Lee رئيس فريق تطوير هذة اللغة وهدفت هذة اللغة إلى تطوير لغة تكويدية على الويب تعمل على إكساب الآلة القدرة على تفسير معاني ودلالات مصادر المعلوات المتاحة على الويب.

3- لغة انطولوجيا الويب OWL ontology web language:

تعد هذة اللغة تطورا كبيرا وعلامة فارقة في مسيرة تنفيذ ونقل الويب الدلالي إلى الوجود. وهي اللغة التي حظيت على recomandation من w3c كلغة دلالية وصفية وترميزية للويب الدلالي وقد صممت خصيصا لإنشاء ونشر الآنطولوجيات على الويب الحالي تمهيدا للانتقال إلى الويب الدلالي. وقد إدراجها رسميا في معمارية الويب الدلالي عام 2004.

يعد الهدف من إنشاء هذة اللغة هي:

1- توفير وسيلة موحدة لترميز المحتوى بصورة دلالية على صعيد الويب.
2- توفير إمكانية تفسير المحتوى للالات والتطبيقات التي لاتعمل في بيئة الويب الدلالي.
3- العمل على أن تكون واجه تعامل الآلة مع المحتوى.

أما عن مميزاتها فقد حصرها Serge linckels على النحو الاتي:

1- تعمل على تحديد المفاهيم الاساسية والخاصة بالمجالات الموضوعية في صورة هرمية ذات كيانات وعلاقات وخصائص مما يكفل تمثيلا فعالا للمعرفة.
2- القدرة على تمثيل العلاقات المعقدة بين الكيانات والفئات والخصائص.
3- القدرة على تحقيق قدر مرتفع من منطق التوصيف للكيانات والعلاقات.
4- العمل على توفير النهج الاستدلالي والاستنباطي لتوليد معرفة جديدة والربط المعرفي بين المجالات المعرفية البشرية.
5- وجودها في 3 مستويات (OWL lite-OWL DL- OWL FULL) يكفل المرونةفي تمثيل المعرفة.
6- يمكن للمستفيد أو الشخص العادي أن يقوم بفهمها وإعداد انطولوجيات وفقا لمبادئها[31].

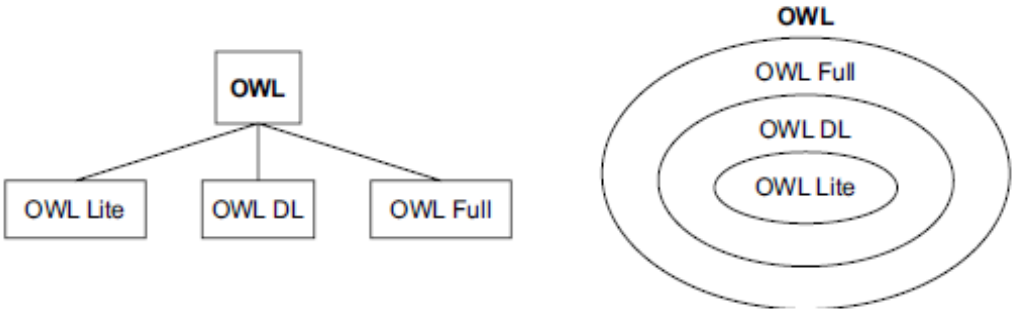
مستويات لغة انطولوجيا الويب:

يرجع وجود تنوع ومستويات للغة OWL بهدف توفير القدرة لدى المستفيد على تمثيل المعرفة بالرؤية التي يراها مناسبة ووفقا للقيود والضوابط التي تفرضها كل مستوى على حدى.

1. المستوى الأول: OWL Lite:

تعد من أبسط مستويات لغة owl، حيث تعمل على تدعيم هؤلاء المستخدمينالذين يحتاجونفي المقام الأول إلى عمل تسلسل الهرمي للمجال الموضوعي في صورة فئاتكما توفر نمطا من القيوديتسم بالبساطة.

[31]Linckels, S., & Meinel, C. (2011). E-librarian service user-friendly semantic search in digital libraries.Heidelberg: Springer.



شكل رقم 16 يوضح مستويات لغة owl[32]

2. المستوى الثاني: OWL DL

يعد هذا المستوى من أفضل المستويات في لغة OWL (وقد اعتمد عليه الباحث في بناء انطولوجيا الخاصة الدراسة) حيث تعتمد على توصيف المنطق description logics الذي يمكن أن يدركه البشر كطبيعة الكيانات وطبيعة العلاقات بينهما للآلة في صورة أقرب لإكساب الآلة الفهم البشري ولذلك تفرض هذة اللغة الكثير من القيود في عملية تحرير الآنطولوجيا على المنشئ.

3. المستوى الثالث: OWL FULL

يعد أكثر المستويات ضبطا وفرطا للمنطق بصورة اقرب للتعقيد فهي تعمل على رفع مستوى القيود الخاصة بالمنطق لاقصى درجة[33].

وظائف الانطولوجيات في أطار التنقيب على الويب:

سبق وأن أشارت الدراسة إلى أن التقيب على الويب يتخذ 3 أشكال رئيسية وهي:

- التنقيب عن محتوى الويب Web content mining.
- التنقيب اعتمادا على بنية الروابط المتاحة على الويب Web structure mining.
- التنقيب من واقع انماط الافادة والاستخدام على الويب Web usage mining.

وقد أوضح كل من Ee-Peng Lim and Aixin Sun أن التنقيب عن محتوى الويب في سياق التنقيب على الويب يضطلع ب 3 مهام أساسية والتي تتمثل في:

- تصنيف صفحات الويب Web page classification: حيث ينطوي هذا المحور على تصنيف صفحات الويب تحت مجموعة من الفئات المحددة سابقا والمسبق تعريفها والتي قد تتخذ صورة تدرج هرمي لمجال معرفي محدد.
- عنقدة صفحات الويب Web clustering: حيث تنطوي هذة الركيزة على تجميع Grouping صفحات الويب مع بعضها وفقا لأوجه التشابه فيما بينها، على أن تشتمل كل مجموعة مكونة على صفحات ويب متماثلة معاً.

[32]Geroimenko, V. (2004).*Dictionary of XML technologies and the semantic Web* .London: Springer.

[33]McGuinness, D. L., & Harmelen, F. v.W3C.OWL Web Ontology Language Overview.avaliable at: www.w3.org/TR/owl-features/date:8/3/2012.

- استخراج محتوى الويب Web extraction: وينطوي هذا العنصر على استخراج مؤشرات الصفحات من محددات النص الفائق HTML elements، والعبارات المؤشرية أو بالأخرى تلك البيانات التي يتم من خلالها توصيف الكيانات كأسم الشخص، أو المكان، أو رقم التسجيلة[34].

وفي ضوء ما تعرضته له الدراسة من شرح لخصائص الانطولوجيات ومكوناتها، وما تم تحديده من مهام للتنقيب عن محتوى الويب، نجد أن الانطولوجيات يمكن أن تستثمر لتحقيق وظائف ومهام التنقيب على الويب بصورة كاملة ومن ثم قدرتها على الوصول الى التكامل المعرفي على صعيد محتوى الويب حيث يمكن استثمار الانطولوجيات لتحقيق مهام ووظائف التنقيب على الويب Web Mining وفقا لثلاثة منهجيات أساسية هي:

- النهج الأول: عنقدة الويب اعتمادا على الانطولوجيات Ontology-based Web clustering: ففي ظل ما تقدمه الانطولوجيات من تعريف بالمجالات المعرفية أو الموضوعية والتعريف بكيانات هذا المجال، تتوافر فرص الحصول على دلالات أضافية حول البيانات والمعلومات، مما يسهل على تقنيات التنقيب على الويب من أكتشاف المعرفة أعتمادا على انتمائها لمفهوم معرف مسبقا.
ويتضح هذا النهج من خلال أعداد عنقودية وبيبة Web Clustering ولتكن عنقودية (A) يحدد في هذة العنقودية الخصائص والسمات التي يجب أن تتوافر في الصفحات والمفردات والكيانات التي تنتمي لهذة العنقودية، هذا الأمر الذي يسفر عن سهولة في تجميع المحتوى والصفحات التي تشتمل خصائص العنقودية (A) وتسكينه ايضا وفقا لاشتماله على هذة الخصائص داخل العنقودية.
وتعتمد العنقودية الويبية القائمة على الانطولوجيات Ontology-based Web clustering على توظيف محددات لغة النص الفائق HTML elements في النص لأن تكون السمات والخصائص التي يتم في ضوءها تجميع عدد أكبر من صفحات الويب وبصورة دقيقة.
كذلك يمكن للانطولوجيا أن تعمل على منهجية للتنقيب على الويب تعتمد على تحليل الروابط الفائقة في بنية الصفحة Ontology-based Web site structure mining أذ تعتمد هذة المنهجية على توظيف الروابط كأنماط اشتقاق وتجميع وتسكين للصفحات.

- النهج الثاني: تصنيف محتوى الويب أعتمادا على الانطولوجيات Ontology-based Web classification: أذ تعمل الانطولوجيا على توفير بنية دلالية يمكن ان تستثمرها تقنيات التنقيب على الويب كخلفية ومرجعية لها في التجميع والتسكين والتحليل، وذلك في ظل ما تقوم به الانطولوجيات من تقسيم للمعرفة البشريه او لمجال موضوعي الى فئات، يمكن أن تستثمر هذة الميزة في تصنيف صفحات الويب وتسكينها داخل فئات موضوعية ككيانات تتواجد بينها قواسم مشتركه سمحت لها بأن تسكن بجوار بعضها البعض، كما تعمل الانطولوجيات على الحفاظ على الروابط والعلاقات الداخلية والخارجية للصفحة دون تقطيع لاواصلها تحت وطأة التسكين الفئوي، هذا الأمر الذي يسفر عن التكامل المعرفي في البحث والاسترجاع في ظل فتح اوجه معالجات الموضوع المستفسر عنه والذي قد يستفسر عنه من خلال زاوية واحدة فقط من قبل المستفيد.

- النهج الثالث: ويتمثل في استخراج المعرفة اعتمادا على الانطولوجيات ontology-based Web extraction حيث تعمل هذة المنهجية على توظيف الانطولوجيات في تحديد المؤشرات والمحددات التي سيتم في ضوءها علمية استخراج المعرفة من محتوى صفحات الويب، ولايقتصر الأمر على ذلك فحسب بل العمل على إيجاد المحددات ذات الصلة والتي يمكن أن تكون قواسم مشتركة يتم في ضوءها تحقيق التكامل المعرفي[35].

[34] Ee-Peng Lim and Aixin Sun (2005), "Web Mining - The Ontology Approach ", available at http://reference.kfupm.edu.sa/content/w/e/web_mining_____the_ontology_approach_61587.pdf
[35] Ibid.

دور الانطولوجيا في تحقيق التكامل المعرفي على الويب:

يكفل توظيف الانطولوجيات في أطار التنقيب على الويب الكثير من المزايا والفوائد في سياق تحقيق التكامل المعرفي هذة المزايا تتمثل في:

1- التكامل المعرفي على صعيد محتوى الويب: إذ تعمل الانطولوجيات على تحويل شبكة الويب بمحتواها وروابطها وهيكلها الى مستودع منظم للبيانات يكفل القدرة على البحث واسترجاع المحتوى وفقا للمفاهيم والدلالات وليس فقط بل تعمل الانطولوجيات على اكساب البرمجيات والتطبيقات القدرة على استنباط حقائق جديدة من واقع المصادر المتاحة من خلال استخدام المنطق وقواعد الاستدلال.

2- تحسين البحث على الويب Improved search to Web data: حيث تمكن الانطولوجيات من تحسين عمليات التكشيف في أدوات البحث وذلك في ظل ما توفره من محددات للمحتوى ومحددات للروابط تقدم دلالات ومفاهيم تعمل كقيمة مضافة لبيانات الويب، وتنقل عملية التكشيف داخل أدوات البحث من التكشيف اللفظي الى التكشيف المعتمد على المفاهيم والعلاقات بين الكيانات المختلفة.

3- القدرة على الإبحار في محتوى بصورة أكثر تكاملية Better browsing capabilities: فعلي غرار البحث تكفل الانطولوجيات القدرة على الإبحار في المحتوى أعتمادا على المفاهيم والعلاقات المنطقية التي تربط بين الكيانات المختلفة عوضا عن الاعتماد على الروابط الفائقة وحسب في عملية الابحار والتصفح لمحتوى الويب.

4- إضفاء الطابع الشخصي في إتاحة البيانات Personalization of Web data access: أن الهدف الأساسي من مفهوم الشخصنة أو أضفاء الطابع الشخصي هو إيجاد مجموعات محددة من البيانات تضاهي الملف الشخصي لاهتمامات مستفيد ما، ويتم ذلك الأمر من خلال تقيم توصيات بصفحات أو مواقع او محتوى يضاهي هذة الاهتمامات، أو عن طريق تصفية واستبعاد صفحات الويب او المحتوى والذي لايضاهي احتياجات المستفيد، ويتم ذلك الأمر من خلال أن تقوم الانطولوجيا بدراسة السجل التاريخي لتصفح المستفيد وهيكلته وتصنيفه وتسكينه في قطاع معرفي ليتم تزويد هذا المستفيد بالمواقع التي تسكن لدى الانطولوجيا في فئات موضوعية تضاهي احتياجات المستفيد[36].

[36] Ibid.

2- نتائج الدراسة:

في ضوء الأهداف التي اتضتها الدراسة لها جاءت نتائج الدراسة على النحو الآتي:

1- وجود العديد من التحديات التي تواجه شبكة الويب في تحقيق التكامل المعرفي على صعيد المحتوى والتي تتعلق بـ:

a. عدم قدرة زواحف الويب من تحميل وتجميع كافة الصفحات المتاحة على الويب، نظرا لضخامة الويب وعدم وجود سياسة واضحة للتجميع لدي الزاحف مما يسفر أن يقوم الزاحف بتحدد الصفحات التي يجب زيارتها وذلك وفقا لأهميتها.

b. صعوبة الكشف عن وجود مكررات على صعيد محتوى الويب.

c. لا تتناسب عمليات التأخيذ Tokenization – وهي أحد المراحل الأساسية - في التنقيب على الويب والتي تعتمد على تفتيت المحتوى الكامل إلى كلمات مستقلة - مع طبيعة بعض المحتويات الخاصة بالوثائق ففرضا إذا تم تفتيت معادلة حسابية فوفقا لهذا المبدء لايمكن لأدوات البحث ان تسترجع المعادلات الحسابية أو الرياضية.

d. تساهم برمجيات خداع المحتوى Spamdexingعلى تصعيب مهمة المستفيدين في ارضاء حاجتهم المعلوماتية information needs.

e. يتمثل التحدي الرئيسي للتنقيب على الويب في التعامل مع محتوى الويب الخفي نظرا لعجزه الكامل في الوصول اليه ومن ثم عدم تكشيفه أو نظمة أو معالجته وذلك على الرغم من كون الويب الخفي يشتمل على 5 اضعاف الويب القابل للتكشيف.

2- يعد التنقيب على الويب Web Mining أحد أهم المبادرات التي طرحت على الصعيد العالمي لتحقيق التكامل المعرفي على صعيد المحتوى المتاح.

3- تعمل الانطولوجيات من واقع بنيتها على جعل شبكة الويب بمثابة قاعدة بيانات معرفية تمكن المستفيدين من البحث والاسترجاع وتحديد المعلومات بسهولة ويسر وفاعلية وكفاءة.

4- تساعد الانطولوجيات على توفير بنية فئوية ذات تبويب موضوعي للمحتوى المتاح مما يكفل القدرة على استرجاع ذو كفاءة عالية.

5- توفر الانطولوجيات القدرة للتطبيقات والبرامج على إدارة ودمج وتكامل المعرفة بين الأقطاب والمصادر والمحتوى المتاح على الويب.

6- كفلت الانطولوجيات فرص اكساب البرمجيات والتطبيقات القدرة على استنباط حقائق جديدة من واقع المصادر المتاحة من خلال استخدام المنطق وقواعد الاستدلال.

7- يعمل التنقيب على الويب على توفير القدرة على تحقيق مستوى مرتفع من التشغيل المتبادل على صعيد محتوى الويب بأكمله.

3- التوصيات والمقترحات:

على ضوء النتائج السابقة يمكن الخروج بمجموعة من التوصيات، حيث يوصي الباحث في نهاية الدراسة بـ:

1. ضرورة ان تتخلي رابطة الويب العالمية W3C عن البنية التكويدية الحالية للشبكة العنكبوتية المتمثلة في لغة (HTML) والتوجه نحو الاعتماد على مواصفات أكثر دلالة كلغة XML وكنماذج وأطر لوصف المصادر، ولغات انطولوجيا الويب.

2. ان يعمل مصممي محتوى الويب على نمذجة المحتوى المتاح على الويب في مجموعة من نماذج للبيانات DATA MODEL اعتمادا على الانطولوجيا.

3. أن تتخلي ادوات البحث والاسترجاع ومختلف تطبيقات الويب عن منهجيات التكشيف التي تعتمد على تطويع النصوص وتفتيتها بهدف الفصل في الاسترجاع.

4. ان تعمل أدوات البحث على توظيف نماذج وتقنيات التنقيب على الويب لاستخراج المعرفة من مختلف مصادر المعلومات.

5. أن يعمل مطوري الويب على إيجاد نماذج مفاهيمية وأنطولوجية بهدف التنقيب عن المعرفة في الملفات صوتية والفيديو والصور ثابته وذلك في سياق التكامل المعرفي على الويب.

6. أن تتوجه نظم التنقيب على الويب إلى تطوير محركات للاستدلال والمنطق التي تكفل القدرة على التحليل والتكشيف المفاهيمي للوثائق.

7. أن تعمل المنظمات الإقليمية ذات التوجه المعلوماتي على اثراء المحتوى العربي بالانطولوجيات العامة والمتخصصة.

8. أن توفر رابطة الويب العالمية نطاقات اسماء عامة ومتخصصة بحيث تراعى فيها التحديث وجعلها مظلة للمحتوى المتاح على الويب.

9. أن تعمل شركات تقنيات المعلومات على تطوير برمجيات خاصة بتحرير الانطولوجيات تتسم بدعمها للغة العربية.

10. أن تثري المحافل العربية من مؤتمرات وندوات متخصصة بالمحتوى العربي على الانترنت بالبحوث الأصيلة العلمية في مختلف المجالات بهدف رفع الرتبة اللغوية العالمية للغة العربية على الويب .