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

العدد 34، مارس 2014

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

Towards to Knowledge Integration by Ontology-based Web Mining:
An Analytical Study

إعداد

مؤمن النشرتي

مدرس مساعد بقسم المكتبات والوثائق والمعلومات

كلية الآداب – جامعة القاهرة

Mo'men Sayed Osman Elnasharty

Teaching assistant – Library, documents and Information Dept.

Faculty of Arts – Cairo University

Navigator001@gmail.com

Cybrarians Journal
ISSN 1687-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 34، مارس 2014

المحتويات

6- نتائج الدراسة.
7- توصيات الدراسة.

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.

This study aimed to clarify the concept of knowledge integration in the information science context, aimed to study the most famous technique initiatives that contributed to achieve knowledge integration at the level of the available content on the web, and monitoring ability to Ontology initiative within the framework of data mining in order to take advantage of them in Web Ontology - based Web mining system to achieve the knowledge integration.

This study based on survey method to monitor algorithm and linguistic challenges facing the web in the knowledge integration path.

As well as the study is based on the analytical method in analyzing the capabilities of the ontology to employ it to achieve the knowledge integration.

The most important result of the study was the ability of the web mining to achieve a high level of interoperability between web resources to provide knowledge integration.

Cybrarians Journal
ISSN 1687-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 34، مارس 2014

1- الأطار النظري للدراسة:

المقدمة:

يأتي التكامل المعرفي كأحد أهم التطلعات التي سعت اليها الكثير من المبادرات التكنولوجية لتحقيقه على صعيد محتوى شبكة الويب، فقد أمست الويب (الشبكة العنكبوتية العالمية - The world wide web) منصة عمل فريدة من نوعها، لم تشهد نظم المعلومات مثيلًا لها من قبل، ومستودعا أساسيا لزخم متراكم من مصادر المعلومات، والتي وجدت طريقها للنشر والإتاحة في بيئة اتسمت بفجوة كبيرة في تحقيق التكاملية بين مصادرها المختلفة وبين احتياجات المستفيدين منها، موسومة في ذلك بنمو مطرد لمحتواها، وارتفاع ديناميكي في معدلات التغيير والتحديث ، الأمر الذي جعل البعض يصف هذة الديناميكية بإن الحديث عن الويب بطرق موثقة أو شفهية لايتسم بالإستقرار، فما يكاد أن يلبث إلا أن يتقادم محتوى هذا الحديث خلال فترة زمنية قصيرة، كما جاء تفرد الويب في طبيعتها المعمارية والتكوينية، فلم تصمم البنية المعمارية للويب على أن تعمل وفقا لمنطق نظم إدارة قواعد البيانات Database Management Systems من حيث الهيكلة والتنظيم المطرد للبيانات والمعلومات، ولم تصمم الويب أيضا على أن تصبغ بهيئة الفهارس الببليوجرافية Online Public Access Catalogue من حيث ما تقدمه الثانية من عناصر تنظيمية واسترجاعية، وصيغ للبيانات وأشكال للاتصال Communication Forms كصيغة مارك MaRC تلك الصيغ التي كفلت القدرة على ضمان النسق والتوحيد في الإدخال للبيانات والمعالجة والتشغيل البيني بين الفهارس، فضلا جذرية تغيير أحدثتها في سلوك المستفيدين منها تجاه تعاملهم معها ومع مصادرها.

بل صممت الويب لتتيح من خلالها كل شئ عن أي شئ، مما استتبع في أن تكون بمثابة الحياة البرية للمعلومات - (فقد قدر حجم المصادر المعلوماتية المتاحة على الويب في اغسطس عام 2000 بنحو 7 ملايين صفحة بعدد مستخدمين لها قدر بـ 500 مليون مستخدم، ليصل حجم الشبكة في اغسطس 2010 إلى نحو 7.74مليار صفحة بعدد مستخدمين قدر بنحو 2مليار مستخدم[1])- فضلا عن التباين والتنوع الموضوعي واللغوي والنوعي والشكلي والجغرافي لما تشمله من مصادر، ولعل المرجعية الأساس وراء ما تعانية الويب من تحديات يعود إلى الإزدواجية Duplication في تكوينها، فهي بيئة استرجاعية تعمل في نفس الوقت كبيئة للنشر والاتاحة الحرة، مما جعل أمر ضبط مصادرها وتنظيمها أمرًا يكاد أن يكون مستحيلاً في اكتماله، هذا الأمر كان بمثابة دافعا نحو التفكير في تطوير العديد من التقنيات التي تكفل القدرة على استثمار هذا الحجم الهائل من المحتوى لتحقيق التكامل المعرفي ، وعلى هذا جاءت العديد من المبادرات التي تسعى إلى ضبط مصادر المعلومات المتاحة على شبكة الويب رامية بذلك نحو التكامل المعرفي، فجاءت مبادرة استخراج البيانات Information Extraction، ثم مبادرة التنقيب عن البيانات Web Mining، وتلي ذلك مبادرة الانطولوجيا Ontology والتي تعد من أبرز هذة المبادرات وأكثرهما توظيفاً.

[1]Kunder, M. d. (n.d.). WorldWideWebSize.com | The size of the World Wide Web (The Internet). RetrievedSeptember 21, 2011, fromhttp://www.worldwidewebsize.com.

يشير مصطلح التنقيب على شبكة الويب Web Mining الى القدرة على أكتشاف المعرفة من واقع البيانات المتاحة على شبكة الويب، مرتكزة في ذلك على ثلاثة قطاعات رئيسية:

- 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.

وتلعب الانطولوجيا Ontology دورا رئيسيا في كلا من القطاعات الثلاث إذ تعمل على تمثيل المحتوى وفقا للمفاهيم والدلالات المراد التعبير عنها معتمدة في ذلك على انماط المنطق والإستدلال الرياضي ، كما تعمل على تحديد العلاقات والروابط بين البيانات المختلفة على الويب، وتعمل على تصنيف المحتوى لفئات وفقا لموضوعية استخدامه.

وفي هذا تعمل هذة الدراسة على رصد توظيف مبادرة الانطولوجيات في أطار التنقيب عن البيانات سعياً بذلك إلى الاستفادة منهما فيما يعرف بنظم الانطولوجيا للتنقيب على الويب Ontology-based Web mining كل من المبادرتين لتحقيق التكامل المعرفي لمحتوى شبكة الويب.

مشكلة الدراسة:

يمكن بلورة مشكلة الدراسة الأساسية في:

1- عجز **تطبيقات وبرمجيات ومحركات شبكة الويب الحالية عن تحقيق التكامل المعرفي على صعيد محتواها،** وذلك في ظل ما تتسم به شبكة الويب الحالية من قضايا شائكة ومشكلات متفاقمة.
2- دراسة **امكانيات وقدرات الانطولوجيات على تحقيق التكامل المعرفي** على صعيد كل من محتوى الويب وروابطه وأنماط الاستخدام.

أهداف الدراسة:

1- رصد التحديات الداخلية والخارجية التي تواجه تحقيق التكامل المعرفي على صعيد محتوى الويب من حيث القدرة على اكتشاف المصادر ذات العلاقة والصلة الموضوعية فيما بينها، ومن حيث معالجتها ومن حيث القدرة على التشغيل البيني بين أدوات وانظمة البحث والاسترجاع المختلفة.
2- التأصيل النظري للانطولوجيات في أطار التنقيب عن البيانات.
3- التعرف على أبرز ملامح تقنيات وخوارزميات التنقيب عن البيانات والتي تمثل الأطار النظري لهذة الدراسة.
4- مدى قدرة الانطولوجيات على تحقيق التكامل المعرفي على صعيد محتوى الويب.

منهج الدراسة:

تعتمد 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.

تساؤلات الدراسة:

تحاول الدراسة الإجابة على التساؤلات التالية:

- ما هي ابرز التحديات الداخلية والخارجية التي تواجه برمجيات ونظم الويب في تحقيق التكامل المعرفي؟
- ماهي أبرز سمات وملامح التنقيب عن البيانات على الويب؟
- كيف يمكن للانطولوجيات أن تساهم في حل قضايا ومشكلات الويب في تحقيق التكامل المعرفي من خلال ما تم اقتراحه لها من معايير ومواصفات؟

حدود ومجال الدراسة:

تأتي الدراسة وفقا للحدود الآتية:

1. الحدود الموضوعية:

تتناول الدراسة رصد قدرة الانطولوجيات في تحقيق التكامل المعرفي.

2. الحدود النـوعيـــــة:

حيث تركز الدراسة على تناول الانطولوجيات في اطار التنقيب عن البيانات دون أن تتطرق الى الأطر الأخرى - كالويب الدلالي - التي قامت هي الأخرى بتوظيف الانطولوجيات في سياقها.

التكامل المعرفي على صعيد محتوى الويب:

مفهوم التكامل المعرفي:

يجدر البيان، بأن قبيل التطرق الى مفهوم التكامل المعرفي وواقع وجوده على صعيد محتوى الويب، ستقوم الدراسة بتوضيح مفهوم المعرفة ودور النظم في معالجتها وتيسير الإفادة منها.

كانت المهمة الأساس للحاسب الالي وبرمجيات المعالجة الالية للنصوص والأرقام ونظم البحث والاسترجاع، ومنصات التشغيل المختلفة على السواء، في القرن المنصرف، هو الحصول على المعلومات القيمة والقابلة للتطبيق من واقع تحويل أنماط البيانات إلى نسق من المعلومات، وعلى هذا أصطبغ هذا القرن بهوية "عصر المعلومات"، ولكن يأتي القرن الواحد والعشرون حاملًا معه تحولاً جوهرياً في تحويل تلك النظم والبرمجيات من مجرد معالج للبيانات ومقدم للمعلومات وعارض لها، الى أنظمة لديها القدرة على استخراج المعرفة واستنباطها من واقع ما أتيح لها من معلومات لينشئ في هذا السياق ما يعرف بـ "عصر المعرفة".

وفي هذا، عمدت تقنيات ونظم عصر المعرفة إلى توجه جديد، وهو الانتقال بالمعلومات من ذاتية البرمجة وفردية النظم وعزلة المستودعات، إلى نطاقات التكامل والتقاسم والتشارك، هادفة بذلك إلى زيادة حجم

المعرفة وتوسيع عمق الاستدلال والاستنباط للمعلومات، هذا الأمر الذي يسفر عن أنماط جديدة من الذكاء الاصطناعي.

يعرف قاموس ويبستر المعرفة بإنها "مجموع تحصيل الإنسان للحقائق والمعلومات والمبادئ والقواعد التي اثمرت خلال فترة حياته" [2].

ويعرفها كل من Sunasee and Sewery الخبرة البشرية المكتسبة من واقع تجاربه وتفاعله وقرءاته وتعلمه في البيئة المحيطة به والمخزنة في العقل البشري"[3].

تصنف المعرفة في الاسانيد الفكرية وادبيات الموضوع الى الأنماط الاتية:

- المعرفة القائمة على التصنيف.
- المعرفة الموجهة لاتخاذ القرارات.
- المعرفة الوصفية.
- المعرفة الإجرائية.
- المعرفة المنطقية.
- المعرفة الاستنباطية[4].

وفي هذا الصدد، تعددت الرؤي في مفهوم التكامل المعرفي Knowledge Integration فيما بينها، وذلك لتعدد السياقات والمجالات المعرفية التي هدفت لتحقيق التكامل في مجالها، ففي سياق العلوم الإدارية ينظر للتكامل المعرفي بإنه قدرة المنظمة على الإستغلال الأمثل للمعلومات والقيم المتوافرة بغية استكشاف أنماط جديدة للمعلومات.

أما في سياق الذكاء الاصطناعي، فينظر اليه في كونه أحد أهم المهام والمتطلبات الأساسية لبناء نظام معرفي يكفل القدرة على تفاعل المستفيد مع الآلة بصورة ذكية.

يمكن النظر إلى مفهوم التكامل المعرفي بإنه عملية تنطوي على دمج وإدراج معلومات جديدة في كيان معرفي قائم يشتركا في بنية مفاهيمية واحدة أو درجة من الصلة أو العلاقة الموضوعية، حيث يعمل التكامل المعرفي على تحدد كيفية أن تتكامل المعلومات الجديدة مع الكيان المعرفي القائم، وكيف يمكن أن يتم تعديل هذا الكيان المعرفي لاستيعاب المعلومات الجديدة، وكيف يمكن تعديل المعلومات الجديدة في ضوء المعرفة الحالية[5].

[2] Merriam-Webster (2001). Merriam-Webster, n.d. Web. 28 Sept. 2012. <http://www.merriam-webster.com/>.
[3] Sunasee and Sewery (2002). Introduction to Knowledge Modeling Available at www.makhfi.com/KCM_intro.htm
[4] Ibid.
[5] Murray, K. S. (1996) KI: A tool for Knowledge Integration. Proceedings of the Thirteenth National Conference on Artificial Intelligence.

كذلك يمكن فهم مفهوم التكامل المعرفي في ضوء مقارنته بالتكامل المعلوماتي information integration، حيث ينطوي التكامل المعلوماتي على دمج المعلومات ذات مخططات البناء المختلفة في حين يهدف التكامل المعرفي الى التركيز وبصورة أساسية على توليف وجهات النظر المختلفة القريبة والبعيدة لفهم موضوع محدد، أو بعبارة أخرى يركز التكامل المعرفي على العلاقات المفاهيمية والدلالية في حين يركز التكامل المعلوماتي على دراسة العلاقات الترابطية بين المعلومات بعضها البعض[6].

ولكن يمكن أن ينظر الى التكامل المعرفي في سياقه المجرد بأنه " **العملية التي تكفل القدرة على تجميع وتوليف وتلخيص نماذج من البيانات** Data Models **داخل نموذج مشترك أو نموذج واحد يوفر حالة معرفية فريدة[7]".**

واقع التكامل المعرفي على الويب:

يوضح التعريف السابق أن جوهر التكامل المعرفي يتجلى في اعتماده على نماذج البيانات Data Model والتي تعرف بإنها منجهية أو بنية Method or Structure تعمل على تنظيم البينات وفقا لنسق وهيكل محدد هذا النسق يساعد على تخزين وتيسير البحث والاسترجاع عن المعلومات بصورة تتعالى فيها الكفاءة والفاعلية.

هذة النماذج تتباين في بنيتها وخوارزمياتها ومنهجية عملها وأغراض انشائها، ومن أمثلتها نماذج العلائقية للبيانات المستخدمة في نظم إدارة قواعد البيانات DBMS ونماذج الاتصال المعيارية في فهارس المكتبات كنموذج MARC21 ومخططات اللغات التكويدية XML Schemas، الأمر الذي يسفر عن أن يكون كل نموذج بمثابة جزيرة منفصلة بما يشمله من معلومات، ويلزم كل من يتعامل معه بالبحث والاسترجاع أن يعمل وفقا لقواعد النموذج وخوارزمياته.

وعلى الرغم من تعدد هذا النماذج وأهدافها الإ ان سمة قاسمين مشتركين بينهم يتمثلا في:

- كونهم قوالب تشتمل على المضمون والمحتوى والمعلومات.
- أن كل منهم – بالرغم من اختلاف بنيته - لديه القدرة على أن يعمل تحت منصة الويب.

هذا الأمر الذي كان بمثابة دافع أساسي في التفكير لإيجاد طريقة أو منهجية ما تعمل على توفير التشغيل البيني بين نماذج البيانات المختلفة، وقد تمثلت هذة المنهجية في مبادرات متعاقبة، يأتي على رأسها مبادرة الشبكة العنكبوتية العالمية The World Wide Web.

في عام 1989 قام Tim Berners Lee بأنشاء العنكبوتية كمنصة عمل على شبكة الانترنت تخلص إلى نظام عالمي، حيث تعتمد على وضع معرفات فريدة للوثائق ونماذج البيانات وربطها بعضها البعض وتمثيلها

[6] Ibid.
[7] Linn, M. C. (2006) The Knowledge Integration Perspective on Learning and Instruction. R. Sawyer (Ed.). In The Cambridge Handbook of the Learning Sciences. Cambridge, MA. Cambridge University Press.

بصورة تسمح بالتشغيل البيني، وفي عام 1990 قام باختبار رؤيته والذي اثبتت جدارة ونجاح، وفي عام 1991 قام بانشاء اول خادم للعنكبوتية عرف بأسم world wide web ومنه اشتق اسم النظام، اما الدور الذي تميزت به الويب فتمثل في تكوين بنية معمارية اساس والتي تجلت فيها مفهوم التشغيل البيني والمتبادل وهي سمة النجاح التي تميز بها هذا النظام ولكي تتحقق سمة التشغيل البيني بين الأنظمة حدد منشئها ثلاثة مبادئ اساسية لها وهي:

- محددات الهوية Identification: ويرمز لها بـ URI والتي تهدف الى إكساب نماذج البيانات محددات ومعرفات فريدة على صعيد العالم بأسره بغية تحديد هويتها وتميزها عن بعضها البعض ويتم ذلك من خلال الاعتماد على أحد السمات المميزة لها وليكن موقعها على الشبكة (URL) او رقمها الدولي (DOI) او اسم مميز لها (ISBN) كمعرف لها.

- التفاعلية Interaction: ويقصد بها الأدوات التي تمكن اطراف النظام من التفاعل عبر عملية اتصالهم ببعضهم البعض، حيث يتم ذلك من خلال مجموعة من البوتكولات التي تحكم وتنظم قواعد الاتصال والتفاعل بين المرسل والمستقبل لمصدر المعلومات، ويعد اشهر نماذجها بروتكول TCP/IP وبروتكول HTTP.

- الصيغ أو التنسيقات Formats : ويقصد بها منهجيات تمثيل وتكويد المحتوى بصيغ تسمح بالتشغيل البيني وتمكن بروتكولات النقل من نقلها، وتمكن نظم التشغيل المختلفة من حفظها, وبرامج التطبيقات من عرضها ويعد اشهر الصيغ لهذا الامر هي صيغ النصوص الفائقة Hypertext.

وبهذا تتراءى شبكة الويب على أنها نظام عالمي لنشر نماذج البيانات المختلفة، وشبكة للربط المصادر، ومستودع مفتوح للوثائق الموزعه، وارشيف متكامل للمحتوى، وعليه استلزم الامر لهذة المنظومة أن توفر أدوات البحث والاسترجاع لما تمده من محتوى ومعلومات ولتكون البوابة الأساسية لتحقيق التكامل المعرفي وعليه ظهر مفهوم البنية الاسترجاعية العنكبوتية للمعلومات أو مايعرف بنظم استرجاع العنكبوتية للمعلومات Web Information Retrieval (Web IR).

وقد حُملت على هذة الأدوات الأمل في تحقيق التكاملية المعرفية للمحتوى المتاح على الويب، ولكن تضاءل ها الأمر فقد عملت كل من هذة الأدوات بمنأي عن بعضها البعض. فجاءت أدوات البحث والاسترجاع على المعلومات متخذة أشكال عدة وسمات مختلفة في بنائها وخورزميات عملها فجاء منها:

- الأدلة الموضوعية. Web directories.
- محركات البحث على الشبكة العنكبوتية Web search engines.
- ماوراء محركات البحث Meta search engines
- بوابات الشبكة العنكبوتية Web portals
- فهارس العنكبوتية غير المرئية Invisible web catalogue.

- فهارس المكتبات على العنكبوتية Web public access library catalogue .

وقد اوضح ضياء عبد الواحد[8] ان المرجعية في التعددية بين أدوات البحث يعود إلى تقديم كل أداة للشكل من المعلومات فضلا عن التفاوت في خبرات المستفيدين في التعامل مع كل أداة وطبيعة احتياجاتهم البحثية التي تفرض التوجه للأداة محددة دون غيرها لكونها مناسبة لطبيعة بحثه. ويضيف الباحث على هذا، أن المسعى في تعددية هذة الادوات يرجع إلى عجز كل أداة على حدى في تحقيق التكامل بين طبيعة ما تقدمه من بيانات ومعلومات ومحتوى وبين نمط آخر، فعلى سبيل المثال يعجز محرك البحث عن تقديم البيانات والمعلومات في صورة ببليوجرافية كالتي يقدمها الفهرس المتاح على الويب، كذلك يعجز فهارس العنكبوتية غير المرئية من تقديم المعلومات في صورة مبوبة ومتدرجة كالتي تقدمها الأدلة الموضوعية.

وعلى الرغم من أنه قد تم أرساء شبكة الويب كمنصة عمل لاستخراج محتوى نماذج البيانات المختلفة (DBMS – MARC21… وغيرها) وتجميع مضامينه وإعادة هيكلته بغية التشغيل البيني وتحقيقا للتكامل المعرفي من واقع ربط المضامين ذات الصلة بعضها البعض من خلال أدوات البحث المختلفة، إلا أن هذة أدوات البحث على الويب قد واجهت تحدي رئيسي يقف في طريق تحقيق التكامل المعرفي على صعيد الويب وهو أن شبكة الويب بجانب ما تحمله من نماذج للبيانات تيسر البحث والاسترجاع فيها ومن ثم تحقيق التكامل المعرفي، الإ أنها تحمل في بنيتها شكل آخر من أشكال قوالب ونماذج البيانات، وهو شكل لم يخضع بأي صورة للهيكلة أوالتنظيم، هذا الشكل يعرف تعرف بإسم البيانات غير المهيكله Unstructured Data والتي تمثل نحو 46% من أجمالي حجم محتوى الويب والمتمثلة في المواقع المحررة بلغة النص الفائق التشعبي Hypertext Markup Language وخطابات البريد الالكتروني فضلا عن المحتوى ذات الطبيعة الانية Real-time content وهي معلومات تتسم بإنها ديناميكة في وجودها وفي تعطل محتواها، كجداول الطيران واسعار الاسهم وغيرها.

كان أحد أهم الادوات التي اعتمد عليها مخترع الويب Tim Berners Lee في تحقيق رؤيته هي لغات النص الفائق HTML والتي سرعان ما أتضح أنها أعظم التحديات التي تقف في طريق تحقيق التكامل المعرفي، نظرا لعدم صلاحية هذة اللغة في تحقيق هيكلة للبيانات في ظل انتمائها إلى عائلة لغات التمثيل غير المهيكل للبيانات والتي يصعب من خلالها استثمار المحتوى.

وفي هذا يجدر الإشارة إلى طبيعة بنية البيانات في الويب تتخذ نمطين أساسين هما:

- البيانات المهيكلة: وهي تلك البيانات التي تسكن نماذج البيانات وهي بذلك خضعت لتنظيم ما وفقا لمعرفات أو محددات منطقية بحيث تسمح بتحديد أجزاء معينة من المعلومات مما يسهل عملية التعرف عليها والبحث عنها واسترجاعها كما هو الحال في قواعد البيانات العلائقية، وتعتمد هذة النوعية على سمات وخصائص تعكس من

[8] ضياء عبد الواحد (2005). محركات البحث المتخصصة دراسة تجريبية . القاهرة: جامعة حلوان قسم المكتبات والوثائق والمعلومات.رسالة دكتوراه. ص 26

خلالها المحتوى والمعنى والاستخدام وتشكل هذة النسبة 15% من محتوى الويب وزعد لغة structure query language اكثر اللغات المستخدمة لهذا الغرض.

- البيانات غير المهيكلة: تشير إلى البيانات التي لم تخضع لآية صيغة من الهيكلة أو التنظيم أو التحديد لها بشكل مسبق عند انشائها كما لايظهر فيها المعرفات أو المؤشرات مما يجعل طبيعتها بمثابة كتلة من البيانات المجمعة لها بداية ونهاية دون أية إشارة إلى أقسامها أو القطاعات التي تشتمل عليها وتمثل هذة الفئة نسبة 85% من أجمالي محتوى الويب ، ويشكل البريد الالكتروني والمدونات والمجموعات الاخبارية والمنتديات والمحادثات والدراسات المسحية والابحاث والعروض التقديمية والمواقع نسبة 65% من هذة البنية، أما عن النسبة الباقية 35% فهي اسيرة التنظيم في الجداول الالكترونية وبرامج تحليل البيانات النصية وغيرها، وتعد لغة html أحد أكثر اللغات التكويدية انتشارا - فقد بلغت عدد المواقع المكودة بلغة html نحو 85% من اجمالي مواقع الويب - والتي تستخدم لغرض التمثيل دون الهيكلة أو التنظيم. فمصممي ومنشئي المحتوى على العنكبوتية يلجاون إلى تنظيم محتواهم بشكل انطباعي ذاتي أو بطرقهم الخاصة دون اللجوء إلى هيكلة منطقية لها والذي يعد أمرا غير مستحب في معالجة المحتوى، ورغم ما توفره لغة HTML من منهجيات أولية لهيكلة البيانات كالتيجان TAGS ونقاط الارساء ANCHORS إلى أنها تتسم بالمحدودية وعدم الجدوى نسبيا في هيكلة المحتوى.

أما عن أوجه القصور التي تعتلي لغة Hypertext Markup Language فيتحقيق التكامل المعرفي فتتمثل في:

- لاتحمل هذة اللغة أية نماذج مفاهيمية من شانها أن تكفل التنظيم أو الدلالة أوالمعنى بل يقتصر دورها على التمثيل البصري للمعلومات.
- أن هذة اللغة بمثابة كتلة من البيانات غير المهيكلة، حيث لا توفر هذه اللغة توصيف لبنية المحتوى, فهي تمثيل أصم للبيانات.
- تعد هذه اللغة بمثابة عالم فوضوي من الروابط الصماء والتي لا تعني للبرامج والتطبيقات البحثية إلا جملة "هذا الموقع مرتبط مع هذا الموقع" فقط من دون أن تحمل هذه الجملة أي دلالة كأن يشير الرابط إلى أن الموقع مرتبط بآخر بعلاقة "هو نوع من" أو "جزء من" وغيرها من العلاقات الوصفية (رغم ذلك تعتمد أقوى محركات البحث على هذا العنصر في بناء خوارزمياتها وترتيب نتائجها للمستفيد).
- قصور التكامل بين البيانات في ظل التطوير البرمجي المستقل للبرامج المستخدمة لتحرير النصوص بلغة html.
- ان غالبية صفحات html تبنى بشكل غير معياري بمعنى انها منشئوها لايلتزمون بمواصفات الانشاء والتحرير معتمدون في ذلك على قدرة المتصفحات في دعم المواصفات الناقصة ولكن الامر يختلف بشكل كبير عند المعالجة في أدوات البحث لتحقيق التكامل المعرفي.

Cybrarians Journal
ISSN 1687-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 34، مارس 2014

وعليه، كان على أدوات البحث المختلفة على الويب أن تفاضل فيما بين الفئتين السابقتين من نماذج البيانات (البيانات المهيكلة – البيانات غير المهيكلة)، وذلك من حيث أيهما يمثل حقل عمل هذة الأدوات؟ وأيهما ستركز عليه في التجميع والتنظيم والإفادة؟ ومن ثم التكامل المعرفي؟ وقد كان الترجيح لصالح البيانات غير المهيكلة ومن ثم فإن أدوات البحث الحالية مهما كانت تعمل على تطوير قدراتها في البحث الذكي عن المعلومات وتحليلها والتكامل المعرفي فإنها محدودة في النهاية بقيود البحث الشكلي وفقا لطبيعة تكويد محتوى الوثائق بلغة النص الفائق التشعبي، ومن هنا كانت محدودية هذة اللغة منبع للتحديات التي تواجه التكامل المعرفي في ظل أفتقار برمجيات وتطبيقات الويب النماذج المفاهيمية التي من شانها ضمان الاتساق بين المحتوى.

التحديات التي تواجه شبكة الويب في تحقيق التكامل المعرفي:

1- منهجيات خداع محتوى Content Spamdexing:

يمكن القول بأن بعض التحديات والمشكلات التي تواجه الويب في التكامل المعرفي ارتبط بظهور ووجود الويب نفسها، والبعض الآخر كان بمثابة تغيرات مستحدثة لمشكلات حظيت من قبل بالدراسة والاهتمام العلمي في أدبيات معالجة البيانات ومجال استرجاع المعلومات والشبكات والدراسات الببلومترية أي قبل ظهور الويب.

ولقد ارتبطت عمليات البحث على المعلومات على الويب من خلال أدوات البحث المختلفة من فهارس وأدلة وكشافات بتقنيات التنقيب عن المحتوى ومدي أسهامها في دقة النتائج المسترجعة من قبل أدوات البحث،وقد اتسم نمط تعامل وسلوك المستفيدين في إيجاد والبحث عن المعلومات إلى اللجوء والاعتماد بصورة قصوى على النتائج الأولى داخل الصفحات الاولى من نتائج نظام استرجاع المعلومات وقصر التصفح عليها دون النظر إلى باقي النتائج في الصفحات الاخرى، فقد أوضحت Silverstein أن 85% من عمليات البحث يقتصر فيها التصفح على النتائج الاولى فقط، ولعل المرجعية في ذلك إلى قناعة ذاتية من جانب المستفيدين في أن النتائج ذات الصلة باستفساره لابد وأن تظهر أولا، وأنه كلما توغل في النتائج الاخرى ابتعدت به عن مجال استفساره[9].

وعليه ادرك مديرون المواقع والقائمين على إدارة المحتوى تبعية منطقية مفادها في أن تضمين المواقع داخل النتائج العشر الاولى لأدوات البحث يؤدي إلى ما يعرف بارتفاع معدل المرور إلى محتوى الموقع traffic to web site[10] ، 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.

[9] C. Silverstein, M. R. Henzinger, J. Marais, and M. Moricz.(1999)."Analysis of a very large AltaVista query log." ACM SIGIR Forum, 33:P6-12.

[*] 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.

وبالتالي يعمل أغلب موفري المحتوى على الويب – خاصة المواقع التجارية – على جعل محتوى صفحات مواقعهم تحظى بترتيب طبقي عالي في نتائج أدوات داخل محركات البحث العامة والأكثر استخداما . ويتم تحقيق ذلك عن طريق منهجية واضحة ومباشرة وهي تحسين جودة صفحات الموقع سواء كان في المحتوى أو الشكل من خلال ما يعرف بمنهجيات تحسين الأداء(SEO- Search Engines Optimization) ولكن تكتنف هذه الطريقة الكثير من المال والوقت والجهد والموارد الأخرى، وعوضًا عن كل ذلك لجأ بعض موفري المحتوى إلى طرق أخرى مختصرة تتمثل في التلاعب في نتائج محركات البحث من خلال استخدام أساليب غير اخلاقية عند بناء محتوى المواقع أو تصميم الصفحات، فيما تعرف هذه المحاولة بتضليل خوارزمية الترتيب الطبقي لنتائج محركات البحث بمنهجيات خداع محركات البحث search engine spam.

تعرف منهجيات خداع محركات البحث ب "المنهجية أو اسلوب الذي يعتمد على استخدام بعض الاليات المصممة عمدا لرفع ترتيب المواقع أو الصفحات في نتائج محركات البحث[11]".

أما عن أوجه الضرر التي قد تلحقه منهجيات خداع المحتوى بتقنيات التنقيب عن الويب فتتمثل في:

- تصعيب مهمة المستفيدين في ارضاء حاجتهم المعلوماتية information needs.
- إرساء خلفية سيئة تجاه محرك البحث ونتائجه وعدم مصداقيته.
- تلويث كشاف محركات البحث بمواقع زائفة.
- حرق النطاق الترددي للزواحف أو ما يعرف بBandwidth.
- تشوية نتائج محركات البحث.
- قدرة الخادعات على توليد العديد من شبيهاتها.
- قد يمتد الضرر إلى مزودي محتوى الويب ومديري 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 ادوات البحث الملوثة مما يستتبع أن يجعل محتوى الويب موجه إلى أدوات البحث بدلا من أن يوجه إلى مستفيدي شبكة الويب الامر الذي يشكل كارثة في نهاية المطاف على الشبكة العنكبوتية[12].

فئات برمجيات خداع محركات البحث:

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

- § خداع محتوى المواقع.
- § خداع روابط الصفحات.

[11] Z. Gy ongyi and H. Garcia-Molina.(2005) Web spam taxonomy.In First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb). Retrieved July 21, 2011, from http://www.airweb.cse.lehigh.edu/2005/gyongyi.pdf

[12] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly.(2006). Detecting spam web pages through content analysis. In Proceedings of the World Wide Web conference, pages 83-92.avaliable at http://www.research.microsoft.com/apps/pubs/default.aspx?id=65140

§ خداع معتمد على اخفاء أو تضمين الصفحات.

شكل رقم (1) يوضح أنواع منهجيات خداع المحتوى.

1- خداع المحتوى:

تعتمد هذه الفئة على تغيير محتوى صفحات الويب للحصول على رتبة أعلى في نتائج محركات البحث، إن اغلب خوارزميات محركات البحث المستخدمة في ترتيب النتائج تعتمد وبشكل أساسي على استخدام أنماط مختلفة من المعادلات والحسابات المختلفة ولكن تعد خوارزمية tf-idf القاسم المشترك بين اغلب هذه 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 هذه المعادلة لتحقيق هدفين اساسين:

جعل صفحاتهم تتصل بعدد كبير من الاستفسارات وذلك من خلال جعل صفحاتهم لا تحصل على الرقم صفر كناتج لهذه المعادلة ويتم تحقيق ذلك من خلال تضمين مجموعة كبيرة وضخمة من المصطلحات بنية الوثيقة.

جعل الوثيقة أكثر 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.

2- الويب غير المرئي The Invisible Web:

يمكن النظر إلى الويب وفقا لمنظور التنقيب على الويب **Web Mining** على أنها تنقسم إلى قسمين:

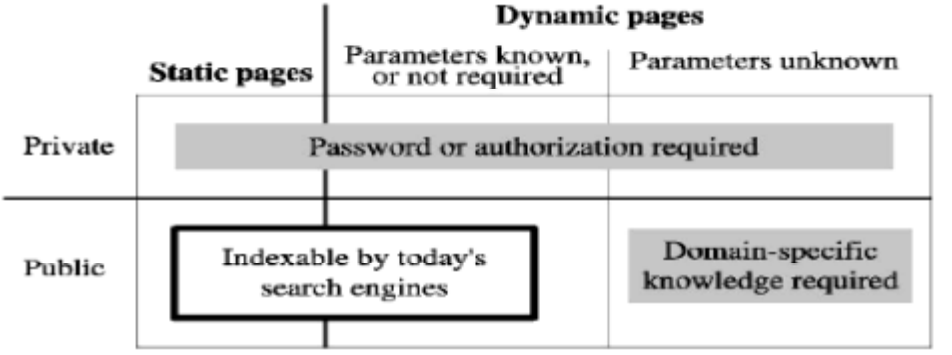
- القسم الأول: الويب السطحية **Surface web**:

أو الويب القابلة للتكشيف **Indexable web Publically** أو الويب المفتوحة **Open Web**: وهي ذلك المحتوى والصفحات والمواقع التي تخضع للتجميع والتكشيف والاسترجاع والتحليل والمعالجة من قبل تقنيات التنقيب على الويب وذلك من خلال تتبع الروابط.

- القسم الثاني: الويب غير المرئية **Invisible web**:

وتعرف ايضا بالويب العميقة the deep web، ، أو الويب المظلمة Dark net أو ما تحت الويب Under net: وهي شبكة الويب التي تشتمل على مصادر المعلومات لإدوات البحث أو تقنيات التنقيب على الويب من الوصول اليها ومن ثم أخضاعها للتجميع أو التكشيف أو التسكين في مستودعاتها[13].

فعجز تقنيات التنقيب على الويب وأدوات البحث لا يتمثل في عدم قدرتها على التعامل مع الويب الخفية وحسب، بل يمتد ليتضح أن تغطيتها لما هو متاح على الويب المرئية لا يتجاوز من 20% إلى 50% .[14]



شكل رقم (2) رسم توضيحي يوضح فئات الوثائق من خلال انواع الويب[15].

ويمكن تحديد خصائص الويب غير المرئية على النحو الاتي:

- يقدر حجم الويب غير المرئية من 400 - 500 مرة من الويب المرئية.
- تشتمل الويب على ما يقرب من 70 الف تيرابايت من المعلومات مقارنة ب9 الاف تيرابايت متاحة على العنكبوتية القابلة للتكشيف.
- تشتمل الويب غير المرئية على ما يقرب من 550 بليون وثيقة منفردة مقارنة ببليون وثيقة متاحة على نظيرتها.

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

- يشتمل الويب غير المرئية على أكثر المصادر تتمتع بمصداقية وموثوقية عما هو متاح على العنكبوتية السطحية.
- تشتمل غير المرئية على أغلب المصادر التي يمكن أن تحقق نسبة مرتفعة من تلبية الحاجات المعلوماتية.
- 95% من معلومات الويب غير المرئية متاح مجانا[16].

3- سلوك المستفيدين البحثي:

ذهب Broder إلى أن نشأة شبكة الويب حمل معها العديد من الأنماط الجديدة من الحاجات المعلوماتية للمستفيدين سواء كان على صعيد البحث أوالاسترجاع ، وقد امتدت هذه الانماط لتشكل فئات جديدة من الاستفسارات المعلوماتية لدى المستفيدين، ولذا كان لزاما على تقنيات التنقيب على الويب ومبادرات التكامل المعرفي لمحتوى الويب التوجه إلى حصر هذة الحاجات المعلوماتية تمهيدا لتوظيف التقنيات التي تعمل على أشباعها، ففي سياق الويب لا تقتصر الحاجة من وراء الاستفسار على استرجاع الوثائق فحسب - كما هو الحال في نظم استرجاع المعلومات وقواعد البيانات- بل أمتدت هذة الحاجة لتشمل 3 اقسام تتمثل في:

- الحاجة الملاحية Navigational need.
- الحاجة المعلوماتية Informational need.
- الحاجة الاجرائية Transactional need .
- الحاجة الاستنتاجية Deductive need[17].

1- الحاجة الملاحية:

يعد نمطا حديثا من الاستفسارات، فالغرض منه هو الوصول إلى موقع أو مصدر معلومات معين لدى المستفيد، فالحاجة من وراء هذا الاستفسار تتمثل في أن المستفيد ربما قد زار أو تعرض لهذا الموقع من قبل، أو على فرضية أن مثل هذا الموقع يمكن ان يكون موجودا على الويب. هذا النمط من الحاجات أو البحث كان يشار اليه سالفا بالبحث عن "المفردة المعروفة" بمعنى أن يكون المستفيد على دراية وعلم بأن هذه المفردة - سواء كانت هذه المفردة وثيقة أو ملف صوتي أو موقع ..الخ - موجودة على الويب وجدير بالإشارة إلى أن هذا النمط من الحاجات قد ظهر في نظم الاسترجاع الكلاسيكية ولكن كان الغرض منه يتمثل في تقييم أداة الاسترجاع.

2- الحاجة المعلوماتية:

[16] MICHAEL K. BERGMAN .The Deep Web: Surfacing Hidden Value.avaliable: http://brightplanet.com/images/uploads/12550176481-deepwebwhitepaper.pdf
[17] A. Broder, "A taxonomy of web search," presented at SIGIR Forum, 2002. Available at: http://www.sigir.org/forum/F2002/broder.pdf

يعد هذا النمط تقليديا إلى حد ما فالغرض منه لا يختلف كثيرا عن الغرض من الحاجات المعلوماتية من نظم الاسترجاع التقليدية وهو ايجاد المعلومات التي يفترض أن 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 القاهرة في بعض محركات البحث تكفل خوارزميات هذه المحركات القدرة على توفير مصادر المعلومات التي تتناول القاهرة ثم تستطرد لتسترجع اسماء فنادق القاهرة والمعالم الاثرية بها... الخ .

3- الحاجة الاجرائية:

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

4- الحاجة الاستنتاجية:

نمط أخر من الحاجات المعلوماتية يعتمد عليه متخذي القرار، وهو القدرة على الاستدلال والاستنتاج من خلال ما هو متاح من معطيات وبيانات ومعلومات للحصول على نتائج جديدة ومستحدثة.

Arabic Speaking Internet Users					
COUNTRIES	Population (2009 Est.)	Internet Users, Latest Data	Penetration (% Population)	User Growth 2000-2009	% User in Tabl
Algeria	34,178,188	4,100,000	12.0 %	8,100.0 %	6.
Bahrain	728,709	402,900	55.3 %	907.3 %	0.
Comoros	752,438	23,000	3.1 %	1,433.3 %	0.
Djibouti	724,622	19,200	2.6 %	1,271.4 %	0.
Egypt	78,866,635	16,636,000	21.1 %	3,596.9 %	27.
Iraq	28,945,569	300,000	1.0 %	2,300.0 %	0.
Jordan	6,269,285	1,595,200	25.4 %	1,153.4 %	2.
Kuwait	2,692,526	1,000,000	37.1 %	566.7 %	1.
Lebanon	4,017,095	945,000	23.5 %	215.0 %	1.
Libya	6,324,357	323,000	5.1 %	3,130.0 %	0.
Mauritania	73,129,486	60,000	1.9 %	1,100.0 %	0.
Morocco	31,285,174	10,442,500	33.4 %	10,342.5 %	17.
Oman	3,418,085	557,000	16.3 %	518.9 %	0.
Qatar	833,285	436,000	52.3 %	1,353.3 %	0.
Saudi Arabia	28,686,633	7,761,800	27.1 %	3,780.3 %	12.
Somalia	9,832,017	102,000	1.0 %	50,900.0 %	0.
Sudan	41,087,825	4,200,000	10.2 %	13,900.0 %	7.
Syria	21,762,978	3,565,000	16.4 %	11,783.3 %	5.
Tunisia	10,486,339	3,500,000	33.4 %	3,400.0 %	5.
United Arab Emirates	4,798,491	3,558,000	74.1 %	384.1 %	5.
Palestine	2,461,267	355,500	14.4 %	915.7 %	0.
Yemen	22,858,238	370,000	1.6 %	2,366.7 %	0.
TOTAL	344,139,242	60,252,100	17.5 %	2,297.7 %	100.

جدول رقم (**1**) يوضح الهوية المعلوماتية العربية على الويب من حيث عدد المستخدمين ومعدلات النمو في الاستخدام[18]

يمكن أجمال التحديات التي تتعلق بتفاعل المستفيدين مع محتوى الويب والتي تعوق تحقيق التكامل المعرفي في العناصر الآتية:

1- حجم الاستفسارات:

فحجم استفسارات المقدمة إلى الويب 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 الموجهة لمحرك البحث جوجل أكثر من 200 مليون استفسار يوميا عام 2003، لتصل هذه النسبة في عام 2010 إلى

[18] Internet Statstics (2012). Available at : www.internetworldstats.com

3,5 مليار استفسار يوميا بمعدل يصل إلى 40,000 استفسار في الثانية، وبالتالي تعد هذه المؤشرات بمثابة تحد كبير يواجه محركات البحث[19].

2- صياغة الاستفسار:

عادة ما يميل المستفيدون إلى تقديم استفسارات أعم بكثير من الحاجة الفعلية إلى المحتوى، والمرجعية وراء ذلك تكمن في ظن المستفيد بأن أداة البحث يمكن أن تعمل على مستوى أعرض وليس على مستوى التخصيص، هذا من ناحية ومن ناحية أخرى كشفت الدراسات المقارنة – القائمة على تحليل سجلات نظم أدوات البحث – أن سلوك المستفيدين في البحث على المعلومات ينطوي على أن معدلات الاستفسارات تتسم بالقصر. فمعظم المستفيدين يدخل ما بين مصطلحين إلى ثلاثة مصطلحات في الاستفسار الواحد، ويدخل ما بين استفسارين إلى ثلاثة استفسارات في عملية البحث الواحدة.[20]

3- 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 فقد تم الكشف عن أن 63 مصطلح حظي بتردد ظهور بلغ أكثر من 100 مرة، في حين كون هذة المصطلحات تمثل أقل من 1% من اجمالي المصطلحات القابلة للبحث. مما يظهر طبيعة أن البحث على الويب يمكن وصفه بانه بحث بمصطلحات تتسم بنسبة منخفضة في تردد الظهور، مقابل مصطلحات تتسم بتردد عالي في الظهور.

4- إعادة صياغة الاستفسار:

فمعظم الباحثين على أدوات 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%.

5- نمط الاستفسار :

إن 50% من إجمالي 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

[19]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.
[20]Ibid

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.

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 أدوات البحث، فشخص واحد من أجمالي 18 شخص يلجأ إلى الاعتماد على الروابط البولينية فضلاً عن شخصان من أصل 3 اشخاص يستخدم هذة الروابط بصورة خاطئة.[21]

[21] 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.

Cybrarians Journal
ISSN 1687-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 34، مارس 2014

2- التنقيب عن البيانات Data Mining:

النِشاة والمفهوم:

يأتي مفهوم التنقيب عن البيانات على الويب Web Mining كأحد تطبيقات مفهوم التنقيب في قواعد البيانات Data Mining، وكلاهما ينتميان إلى مجال أكتشاف المعرفة من داخل قواعد البيانات Knowledge Discovery from Database وبالتالي يعد من الضروري قبيل التطرق إلى التنقيب عن البيانات على الويب Web Mining، التعرف على مفهوم التنقيب في قواعد البيانات Data Mining والذي يعد جوهر استخراج المعرفة من على الويب.

طور مفهوم التنقيب عن البيانات Data Mining في سياق مجال أكتشاف المعرفة في قواعد البيانات "Knowledge Discovery in Databases" وهو مجال معرفي عمل على توظيف الإحصاء والذكاء الاصطناعي ونظم قواعد البيانات والتعلم الآلي في معالجة محتوى قواعد البيانات، حيث هدف إلى الخروج من مجموعة البيانات المختزنة في قواعد البيانات بمعلومات ذات دلالات ومؤشرات، وتحويلها إلى بنية مفهومة تمهيدا لاستثمارها مرة أخرى والتنبوء من خلالها بعوامل وفرضيات جديدة في ظل المعطيات المتوافرة، إذ يعمل هذا المجال على تحديد وتعيين البيانات في أدنى مستوى لها وتحويلها الى معلومات أكثر إحكاما وتجريدا وفائدة حاملة معها القدرة على توليد نموذج تنبؤي لتقدير الأوضاع في المستقبل لمجال ما أو منظمة ما[22].

تاريخياً، حظيت فكرة استثمار محتوى قواعد البيانات بالاهتمام في مجال صناعة المعلومات فضلا عن أجتذابه لقدرا كبيرا من الإهتمام من جانب مجالات معرفية مختلفة، نظرا لما تمتاز به هذة البيانات من طبيعة هيكلية وتنظيمة تسمح بتشغيلها في سياقات ونسق مختلفة، ونتيجة لتزايد عدد قواعد البيانات وضخامة حجم بيانات المختزنة بها، ووجود حاجة ملحة لتحويل هذة البيانات إلى معلومات يمكن أن تستثمر للخروج بها الى معرفة جديدة مجردة، ظهر مفهوم التنقيب عن البيانات Data Mining.

حظي مفهوم استخراج المعرفة من قواعد البيانات بالعديد من المسميات والتي تنوعت بين (استخراج المعرفة – أكتشاف المعلومات – جمع المعلومات – معالجة انماط البيانات – التقيب عن البيانات)، الى أن تم الإستقرار على مسمى واحد لها وهو Data Mining ليدل على مفهوم واحد وهو اكتشاف المعرفة في قواعد البيانات KDD في ولكن يجب الإشارة إلى أن الأستخدم الأولي لمفردة التنقيب عن البيانات Data Mining قد ورد في سياق فعاليات ورشة العمل الأولى لاكتشاف المعرفة في قواعد البيانات على هامش اجتماعات اللجنة الدولية المشتركة للذكاء الاصطناعي International Joint Conferences on Artificial Intelligence والمنعقدة في ولاية ميتشجان في الولايات المتحدة.

[22] U. Fayyad, G. P.-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, Vol. 17 No. 3, pp. 37-54, Fall 1996.

يوضح كل من U. Fayyad, G. P.-Shapiro, and P. Smyth أن العلاقة بين مجال أكتشاف البيانات KDD وبين Data Mining هي علاقة أشتمال من ألاولي للثانية، حيث أن كينونة KDD تتضح في أنها عملية تهدف الى أكتشاف المعرفة من واقع البيانات، في حين يشير التنقيب عن البيانات إلى كونه خطوة أو مرحلة معينة في أطار العملية السابقة.

وفي ضوء ذلك، يعرف التنقيب عن البيانات Data Mining بأنه:

" آلية تعمل على تحليل مجموعات من البيانات المجردة من خلال توظيف خوارزميات معينة تعمل على استخراج انماط محددة لإيجاد علاقات غير متوقعه أو لتلخيص البيانات بطرق جديدة وبصورة مفهومة وقابلة للاستثمار في اتخاذ القرار[23]".

لاتعد مرحلة التنقيب عن البيانات هي المرحلة الوحيدة أو الفردية في عملية اكتشاف المعرفة في قواعد البيانات KDD إذ ينطوي مجال اكتشاف المعرفة على عدد من الخطوات والمراحل تحدد من خلالها كيفية اكتشاف المعرفة من البيانات المتاحة، وماهي منهجية حفظ وتخزين البيانات، وما هي أنسب الطرق لأتاحة هذة البيانات، وكيف يمكن أت تقوم خوارزمياتها بتعزيز المعرفة، بحيث تكفل هذة المراحل أكتشاف وتوفير وأشتقاق المعرفة بصورة سليمة من واقع البيانات المتوفرة.

مراحل التنقيب عن البيانات:

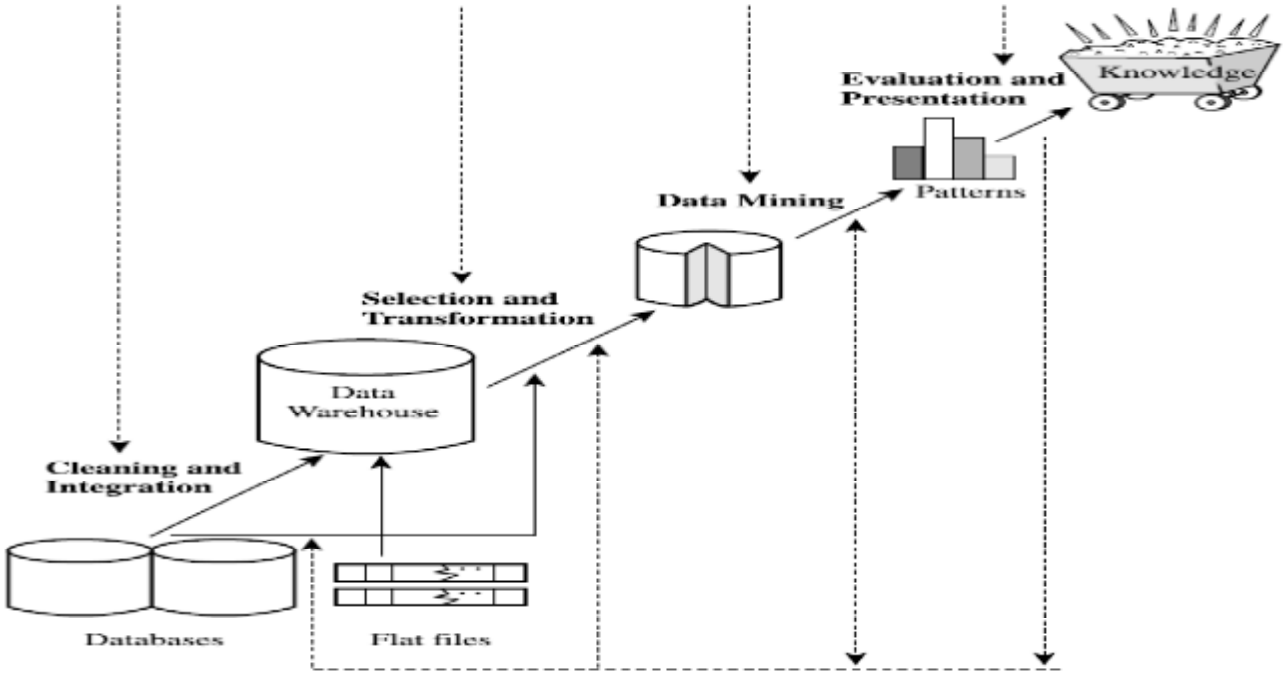
وردت العديد من الرؤي في تحديد مراحل أكتشاف المعرفة في قواعد البيانات (موقع التنقيب عن البيانات في منظومة أكتشاف المعرفة) ولكن تعد الرؤية المعيارية في هذا الصدد هي الرؤية التي قام بوضعها كل من Jiawei Han, Micheline Kamber, Jian Pei فعملية أكتشاف المعرفة لديهم هي عملية تفاعلية وتكرارية، تنطوي على خطوات محددة تتتمثل في:

1- Data cleaning (تنقية البيانات): تختص هذه العملية بحذف البيانات غير المهمة والمكررة.
2- Data Integration (تكامل البيانات): وتعمل هذة المرحلة على جمع البيانات من المصادر المختلفة.
3- Data Selection (إختيار البيانات): وفي هذه المرحلة يتم إختيار البيانات التى سيتم عليها التحليل.
4- Data Transformation (تحويل البيانات): أى عملية توحيد/إدماج البيانات في أشكال محددة مناسبة حتى تكون مهيئة لعملية التنقيب.
5- Data mining (التنقيب عن البيانات): وتعد الخطوة الجوهرية حيث يتم استخدام التقنيات الذكية لاستنباط أنماط مفيدة من المعلومات.
6- Pattern Evaluation (تقييم الأنماط): أي تحديد الأنماط التى تمثل المعرفة وفق المقاييس المعطاة.

[23] David Hand, Heikki Mannila, and Padhraic Smith, *Principles of Data Mining*, MIT Press, Cambridge, MA, 2001

Cybrarians Journal
ISSN 1687-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 34، مارس 2014

7- Knowledge Representation (تمثيل المعرفة): وهي المرحلة الأخيرة حيث يتم اكتشاف المعرفة الجديدة، وفي هذه المرحلة يتم استخدام التقنيات المصورة Visualization لمساعدة المستفيدين على فهم وتفسير النتائج المستخرجة[24].



شكل رقم (3) يمثل خطوات اكتشاف المعرفة من واقع تحليل البيانات، ويوضح موقع دور التنقيب عن البيانات كمرحلة جوهرية داخل هذة العملية[25].

معمارية التنقيب عن البيانات:

وفي ضوء هذة المراحل والخطوات اقترح كل من Jiawei Han, Micheline Kamber, Jian Pei تصورا لمعمارية مقترحة لما يجب أن تكون عليه نظم التنقيب عن البيانات وجاءت وحداتها على النحو التالي:

1- مستودعات البيانات Data Repository: وتتمثل في قواعد البيانات Data Base و شبكة الويب العالمية World Wide Web ومخازن البيانات Data Warehouse وتلعب كلا من مرحلة تنقية وتكامل البينات الدرو الرئيسي في هذة الوحدة.
2- خوادم مستودعات البيانات: وهي وحدة مسؤولة عن جلب وإحضار البيانات ذات الصلة بناء على طلب المستفيد في أطار التنقيب عن البيانات.

[24] Han, J., & Kamber, M. (2001). *Data mining: concepts and techniques*. San Francisco: Morgan Kaufmann Publishers.
[25] Ibid.

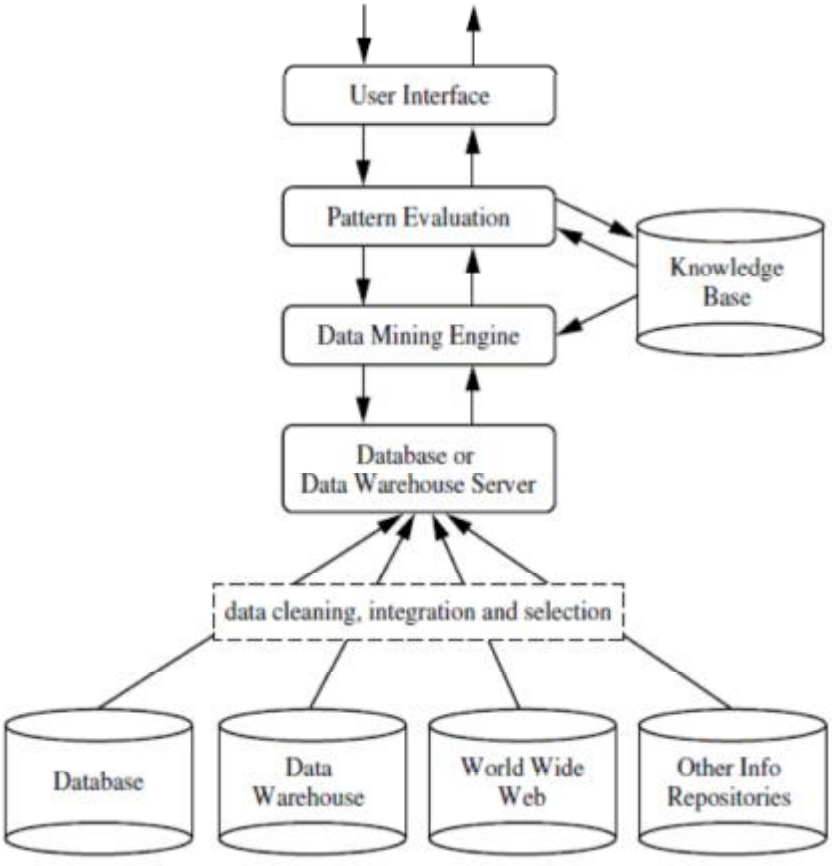
3- قواعد المعرفة Knowledge Base: ويقصد بها المجال المعرفي المستخدم في توجيه عملية البحث وتقييم درجة الصلة الناتجة عن استرجاع النماذج والأنماط، وتعمل هذه الوحدة على إنتاج الأنماط التي سيتم في ضوئها تحليل البيانات واستنباط المعرفة كما تعمل هذه الوحدة على تنظيم المعرفة في مستويات مجردة الصورة الهرمية التي تتدرج فيها من العام الى الخاص، ويعد أشهر نماذج المعرفة التي تستخدم في هذة الوحدة هي الميتاداتا.

4- محرك التنقيب عن البيانات (Data mining engine): والذي يعد قلب وجوهر نظم التنقيب عن البيانات، ويتألف من وحدات وظيفية فرعية تعمل بالقيام بالمهم الآتية: التوصيف – الربط – تحليل الروابط – التصنيف – التنبوء- التحليل العنقودي للبيانات – تحليل النماذج – تحليل الاستدلال.

5- وحدة تقييم الانماط: تهدف هذة الوحدة على توظيف مقاييس تهدف الى تحديد أي الموضوعات تحظى بالاهتمام داخل مجموعات البيانات المختلفة.

6- وحدة واجهة المستفيد: وهي الوحدة المسئولة عن تحقيق التفاعل بين المستفيدين ونظم التنقيب عن البيانات، وذلك من خلال توفير القدرة للمستفيد أن يطرح الاستفسارات والتي سيتم ضوءها التنقيب عن البيانات، كما تعمل هذة الوحدة على توفير إمكانية السماح للمستفيدين بتصفح مخططات قواعد البيانات المحللة، ومستودعات البيانات وبني البيانات[26].

[26] Ibid.



شكل رقم (4) يوضح البنية المعمارية الأساس لنظام التنقيب عن البيانات[27].

مهام التنقيب عن البيانات:

حدد كل من Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic المهام الرئيسية التي يمكن أن تضطلع بها نظم وتطبيقات التنقيب عن البيانات في ستة قطاعات رئيسية:

1- الكشف الخاص Anomaly detection: وتشير هذة المهمة الى الكشف عن أنماط في مجموعة من البيانات والتي تتطلب مزيدا من التحقيق والتحليل نظرا لعدم أتساقها مع بعضها البعض.
2- الربط وفقا لقواعد محددة Association rule learning: أو تعرف أحيانا بالنمذجة ذات التبعية Dependency modeling وتشير هذة المهمة إلى البحث عن القات بين المتغيرات المختلفة فعلي سبيل المثال تعمل هذة المهمة على جمع البيانات عن عادات وأنماط سلوك المستفيدين في البحث عن البيانات

[27] Ibid.

وتحليلها وفي ظل هذة المهمة يقوم النظام بدفع بعض المواقع ومصادر المعلومات للمستفيدين من واقع تحليل سجلات تاريخ البحث للمستفيدن وربطها معا.

3- العنقدة Clustring: تعمل هذة المهمة على ان تكتشف داخل مجموعات البيانات عن الصلات أو درجات التشابه بين البيانات تمهيدا لجمعها معا في نسق محدد يعرف بالعنقدة.

4- التصنيف Classification: تعمل هذة المهمة على تعميم بنية محددة لتطبق على البيانات الجديدة لتسكن وفقا لهويتها أو موضوعها فعلي سبيل المثال يستخدم البريد الالكتروني نموذج محدد وهيكل معين لتحديد وتسكين البريد الوارد في قطاعين المشروع والمزعج.

5- الإنحدار Regression : حيث تشير هذة المهمة في فهم كيفية تأثير المتغيرات الثابتة على قيم المتغيرات المتغيرة أو بمعنى أخر يساعد على فهم تبعية التفيير التي تحدث في المتغير التابع من جراء التغيير في المتغير المستقل.

6- التلخيص Summarization: أذ تعمل هذة المهمة على تحديد وتمثيل البيانات في صورة أكثر ايجازا تمهيدا لتوليد التقارير لمساعدة متخذي القرار. [28]

آليات التنقيب عن البيانات:

أوضح كل من Fayyad, Usama; Piatetsky-Shapiro أن مختلف خوارزميات التنقيب عن البيانات تدور في فلك 3 عوامل أساسية:

1- تمثيل النماذج Model representation: ويقصد بها اللغة التي تستخدم لوصف أنماط التنقيب التي تعمل على اكتشاف المعرفة وتحليلها، او بمعنى أخر هي لغة صياغة انماط Patterns.

2- معايير تقييم النماذج Model-evaluation criteria: وهي منهجية كمية أو مجموعة من الدوال التي تهدف الى تقييم مدى نجاح الأنماط المحددة في نظام التنقيب عن البيانات، أو بمعنى أخر تعمل هذة الدالات على أختبار ما إذا كانت الانماط Parameters تلبي أهداف قواعد بيانات اكتشاف المعرفة.

3- منهجيات البحث Search Method: وتعمل هذة المنهجية على تكوين عنصرين مهمين في نظم التنقيب عن البيانات وهما نماذج البحث ومحددات البحث والذان يتم توظيفهما بغية تحقيق التكامل المعرفي على صعيد البيانات المحللة.

خوارزميات التنقيب عن البيانات.

في عام 2006 قام معهد مهندسي الكهرباء والالكترونيات Institute of Electrical and Electronics Engineers IEEE بتكليف فريق بحثي لإعداد دراسة تهدف الى تحديد أشهر الخوارزميات الأساسية

[28] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008. Available at: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

المستخدمة في مجال التنقيب عن البيانات ذات التأثير وألاستخدام داخل الأوساط البحثية، لتوثق هذة الدراسة في فعاليات المؤتمر الدولي للتنقيب عن البيانات International Conference on Data Mining (ICDM) وقد جاءت هذة الخوارزميات على النحو الآتي:

1- خوارزمية الموجهات في الفراغ Vector space model:

طورت هذة الخوارزمية على يد Gerard Salton عام 1975 في أطار تطويره لنظام أحصائي لتحليل البيانات عرف بإسم SMART (System for the Mechanical Analysis and Retrieval of Text ، قدم هذا النموذج أطار عمل جديدًا بحيث ينظر إلى محتوى الوثيقة على أنه حقيبة كلمات Bag of words بمعنى أن محتوى الوثيقة يشتمل على مصطلحات غير مرتبة وذات ترددية غير منتظمة داخل محتوى الوثيقة.

كان الاستخدان الأولي لهذة الخوارزنية في سياق نظم استرجاع المعلومات حيث تمثلت رؤيته في حساب 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:

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

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 أن 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 tfi, j إلى حساب تردد المصطلح.
§ تشير ni,j إلى عدد مرات ظهور المصطلح ti في محتوى الوثيقة dj.
§ وتشير ∑k nk,j إلى مجموع عدد المصطلحات في اجمالي الوثيقة.

مثال: إذا افترضنا أن 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 $\mathrm{idf_i}$ إلى حساب تردد الوثيقة العكسي.
- بينما تشير log إلى حساب لوغاريتم 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 افترضنا أن عدد الوثائق المختزنة في النظام تبلغ **1000000** وثيقة ويظهر مصطلح المكتبات في 1000 وثيقة من إجمالي عدد الوثائق وبالتالي يحسب **3=log(1000000/1000)**.
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 **0.12=0.04X3**إي أن رتبة الوثيقة يساوي 0.12، ولعل من الملاحظ أن إجمالي 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:

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

2- خوارزمية C4.5 and beyond:

تأتي هذة الخوارزمية كأحد أشهر وأهم الخوارزميات المستخدمة في بناء نظم التصنيفات هذة النظم تعتمد مجموعة من الحالات Cases كمدخل لها، حيث أن كل حالة تنتمي الى عدد قليل الفئات وتوصف من خلال قيمها لتعمل هذة الخوارزمية على بناء هيكل أو شجرة قرارات تساهم على التنبوء بتحديد وتسكين أي من الحالات قد ينتمي لفئة من الفئات، طورت هذة الخوارزمية على يد روس كوينلان Ross Quinlan.

تقوم منهجية هذة الخوارزمية على أفتراض أن البيانات تعيين في مجموعة حيث $S = s_1, s_2, \ldots$ والتي تصنف بالفعل في عينات Sample بحيث تشتمل كل عينة على مجموعة من المحددات Vector لتأخذ الشكل الآتي $s_i = x_1, x_2, \ldots$ حيث أن كل محدد $x_1, x_2, \ldots$ يمثل خواص عينة من العينات، تتمثل المرحلة التالية في تزويد البيانات بمحددات تعمل على تسكينها داخل الفئات التي تنتمي اليها.

3- خوارزمية The k-means algorithm:

تنتمي هذة الخوارزمية الى منهجيات التحليل العنقودي لمجموعات البيانات أذ تعمل على تقسيم مجموعات البيانات المحددة الى مجموعات عنقودية محددة تعرف ب K وقد مر تطوير هذة الخوارزمية بالعديد من المراحل لما يقرب من عشرة سنوات.

تعتمد هذة الخوارزمية توافر مجموعة من البيانات ($x_1, x_2, \ldots, x_n$) حيث يتم أعطاء لكل بيان محدد ذات أبعاد d- dimensional real vector ليتم تسكين هذة المجموعات في عنقاقيد K وفقا لانتمائها الموضوعي.

4- خوارزمية Naive Bayes:

وهي أحد الخوارزميات التصنيف الاحتمالي المعتمدة على نظرية تعرف بنظرية Bayes' theorem تعد هذة النظرية احد نظريات مجال الاحتمالات في علم الرياضيات والتي تعني بقياس العلاقة بين احتمالين شرطيين والذي عادة ما يعلوهما التناقض فيما بينهم، أن المنهجية الأساسية التي تستند عليها هذة الخوارزمية يتمل في أن يتم توصيف الكيانات المختلفة حيث أن كل كيان ينتمي إلى فئة معروفة، وكل كيان لديه مجموعة من المحددات المعروفة بحيث تهدف هذة الخوارزمية الى بناء قاعدة تسمح بتحديد وتعيين الكيانات المستقبلية داخل المجموعات في ظل ما توافر من محددات ومتغيرات.

3- التنقيب على الويب Web Mining:

الدوافع وراء التوجه للتنقيب عن المحتوى على الويب:

اتضح من واقع دراسة البنية المعمارية لنظم التنقيب عن البيانات ضرورة توافر منصات الحفظ ومستودعات التخزين للبيانات كركيزة أساس لتلك النظم، فهي بمثابة الحاويات التي تستورد منها نظم التنقيب عن البيانات مدخلاتها وتقوم علي تقديم أوجه المعالجة المختلفة لاكتشاف واستخراج المعرفة منها ومن ثم تحقيق التكامل المعرفي، ولما كان الويب يمثل أكبر مستودعات للبيانات في العالم[29] سواء المهيكل منها أو غير المهيكل، كان

[29] Liu, B. (2007). *Web data mining exploring hyperlinks, contents, and usage data*. Berlin: Springer.

لزاما على الباحثين والعاملين في مجال التنقيب عن البيانات التطرق بالبحث والدراسة لاستثمار هذا الكم بغية تحقيق التكامل المعرفي على صعيد محتواه وأنماط استخدامه وبنيته.

يرى تيم بيرنرز Tim Berners lee - مبتكر شبكة الويب - أن المشكلات والتحديات التي تواجه الويب في تحقيق التكامل المعرفي يكمن في طبيعة أدوات البحث (من أدلة موضوعية ومحركات بحث ...وغيرها) ، فلقد صممت هذه الأدوات للإجابة على تساؤل واحد " ماهي الوثائق التي تشتمل على الكلمات أو الجمل الآتية" دون النظر إلى اعتبارات أخرى كالعلاقات بين المحتوى ذات الصلة وكمصداقية وموثوقية المحتوى أو تكاملية المعرفة، وقد اوضح ذلك في قوله...

"... ***اذا استطاع محرك البحث على الويب أن يجمع في تقنياته بين محرك الاستدلال*** reasoning engine ***ومحرك البحث*** Search engine ***فإننا يمكن أن نحظى بالتكامل المعرفي على صعيد شبكة الويب***...." [30].

"اننا لانزال في عصر ما قبل الويب ".

"أن الشبكة الويب الآن بصورتها الحالية مفهومة لنا نحن البشر ولكنها بالنسبة إلى الحاسب الالي عبارة عن صفحات ممثلة بالصفر والواحد ولاتعني لها شيئا، إن ما نريده من مبادرات التكامل المعرفي أن يجعل التطبيقات والبرمجيات قادرة على فهم ماذا تعني محتويات الصفحة التي يعرضها وما هو موجود في الويب من معلومات يمكن من خلالها تحقيق التكامل المعرفي. "

"لايمكن أن ينكر أحد أن الويب قد اضفت سمة من التغيير الشامل في طرق اكتساب المعرفة والتواصل والنجاح على مختلف الأصعدة وأنه كان المحرك الأساس للاتجاه نحو اقتصاديات المعرفة ومجتمعات المعرفة بما قد ضمنه من قدرات وامكانيات كفلت نقل المحتوى بيسر واتحته بسهولة، ولكن لم تصل الويب من خلال هذا المحتوى أو به إلى مرحلة النضج فهي مازالت قائمة على تطبيقات ومستودعات للمحتوى منعزلة ومنفصلة تتمثل في كيانات مستقلة تعمل دون تكاملية في المحتوى، فضلا عما يعلوها من عجز في التعالم مع هذا المحتوى وفهمه وتحليله بناءا على دلالاته ومضمونه ".

ويرى ريكاردو بيزا Ricardo Baeza-Yates أن المشكلة الرئيسية في الوصول للتكامل المعرفي على الويب يكمن في قضية التنبؤ بتحديد أي من الوثائق قد يتسم بالصلة بعضها البعض وأي منها لا يتسم بالصلة[31].

كذلك أوضح كلا من كارلو تاسو & بيتر بيرسيل كوفPETER BRUSILOVSKY و CARLO TASSO أن كافة التحديات التي تواجه الويب في الوصول الى التكامل المعرفي تدور في فلك عاملين أساسين هما:1- المعالجة اللغوية: حيث أن غالبية برمجيات الويب تفتقر في معالجتها لمحتوى الويب على وجود أو

[30]Alesso, H. P., & Smith, C. F. (2006). Thinking on the Web: Berners-Lee, Go del, and Turing. Hoboken, N.J.: Wiley-Interscience.(p.67).
[31]Yates, R., & Neto, B. (2011). Modern information retrieval: the concepts and technology behind search (Second ed.). New York: Addison Wesley.(p.11).

غياب الكلمات المفتاحية في النص دون أية محاولة لتحليل المحتوى أو تحديد المفاهيم المشار اليها في النص وهو السبب الرئيسي وراء انخفاض الدقة في عمليات البحث والاسترجاع فضلا عن الظواهر اللغوية الاخرى كالترادف والتجانس.و2- محدودية الآليات والخوارزميات: وتتجلى هذه المحدودية في عجز التطبيقات والبرمجيات من فهم المحتوى والمعلومات المقدمة نظرًا لعدم وجود نسق عالمي للتشغيل المتبادل فضلا عن اتساع الفجوة بين الجانبين الأساس لشبكة الويب المتمثلان في المحتوى (وهو المضمون الذي تشمله صفحات المعلومات على شبكة الويب) والبرمجيات (المسئولة عن معالجة هذا المحتوى واسترجاعه)، وبالأحرى افتقار البرامج لمعالجة المحتوى. فهما يفتقران إلى التكامل لينسجا معا نسيج الويب [32].

بينما أوضح ماركوف Markov أن تحديات التي تواجه الويب في الوصول الى التكامل المعرفي منبعها يعود إلى المحدودية الدلالية لشبكة الويب ذاتها، فصفحات الويب لا تحمل دلالة لمحتواها ولكن تحمل تنسيقا جيد وتمثيل عظيما للبيانات، أما الروابط فتكاد تنعدم دلالاتها على الويب والدلالة الوحيدة التي تحملها في أطارها هي أن الموقع هذا يرتبط بالموقع ذاك دون أية تحديد لدلالة الربط أو نوع الارتباط[33].

ويرى كلا من فان هيرميلين و ستنكشميدت Van Harmelen & Stuckenschmidt أن التحديات التي تواجهها تكمن في افتقار الويب إلى النماذج المفاهيمية لمصادر المعلومات وعدم وضوح حدود وملامح الويب في ظل ديناميكيتها المفرطة[34].

محدودية الويب في تحقيق التكامل المعرفي:

بالرغم من أتضاح بنية الويب المعمارية والتكوينية، الإ أن تحقيق التكامل المعرفي على صعيد محتواه يتسم الصعوبة للاسباب الاتية[35]:

1- ضخامة حجم المحتوى المتاح على الويب:

فقد قدر حجم المصادر المعلوماتية المتاحة والقابلة للتكشيف على الويب في اغسطس عام 2000 بنحو 7 ملايين صفحة بعدد مستخدمين لها قدر بـ 500 مليون مستخدم، ليصل حجم الشبكة في اغسطس 2010 إلى نحو 7.74مليار صفحة وبعدد روابط قدر بنحو 4 مليار رابط وبعدد مستخدمين قدر بنحو 2مليار مستخدم[36] - فقد قامت الويب بالسماح

[32] 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.
[33]Markov, Z., & Larose, D. T. (2007). Data mining the Web: uncovering patterns in Web content, structure, and usage. New York: Wiley-Interscience.
[34]Stuckenschmidt, H., & Harmelen, F. v. (2005). Information sharing on the semantic web: New York: Springer.
[35] Berners-Lee, T. (n.d.). The Semantic Web: Scientific American. *Science News, Articles and Information | Scientific American*. Retrieved August 2, 2011, from http://www.scientificamerican.com/article.cfm?id=the-semantic-web
[36]Kunder, M. d. (n.d.). WorldWideWebSize.com | The size of the World Wide Web (The Internet). RetrievedSeptember 21, 2011, fromhttp://www.worldwidewebsize.com.

لمختلف الأدوار والأفراد بإضافة المحتوى والتحليل والنقد، مما أثمر عن حجم هائل من المحتوى ذات التنوع اللغوي والموضوعي والشكلي، الأمر الذي أسفر عن تحديات غير مسبوقة على صعيد ضبطه وتنظيمه.

2- افتقار بنية شبكة الويب الداخلية إلى البنية السليمة لتمثيل المعلومات والمحتوى:

تلك البنية التي تقف وراء عجزها الكامل عن تحقيق التكامل المعرفي على صعيد محتواها سواء النصي أو ذو الوسائط المتعددة، ولعل المرجعية الأساس وراء هذا العجز والافتقار يعود إلى اعتماد شبكة الويب على لغات تمثيل ومواصفات يقتصر دورها فقط على التمثيل الأصم للمحتوى plain text، مع افتقار كامل إلى المعاني meaningful tags، والمؤشرات mark-up indicating، التي تيسر من معالجة المحتوى واستثماره وتحقيق التكامل على صعيده[37].

3- تعقد احتياجات المستفيدين والباحثين في التعامل مع محتوى الويب:

سواء كان ذلك على صعيد التأليف والتحرير أو على صعيد البحث والاسترجاع، فكثير من المستفيدين يقومون بطرح استفساراتهم بصورة قلما يعبروا من خلالهاعن حاجاتهم البحثية الفعلية, نظرا لافتقارهم إلى الفهم الصحيح للموضوع قيد البحث، أو عدم الالمام الشامل بمفردات حاجاتهم المعلوماتية، أو تقديم استفسارات أعم بكثير من الحاجة الفعلية إلى المعلومات، فضلا عن حجم الاستفسارات المقدمة إلى الويب والذي يتسم بالنمو المتزايد والمستمر فقد أوضحت إحدى الدراسات عن أن حجم الاستفسارات على الويب قد بلغت نحو 10 مليار استفسار عام 2008 موجه إلى 5 محركات بحث (Google, Yahoo!, Bing, AOL). وأن معدل توزيع تردد مصطلحات الاستفسار يتسم بالانحراف من إجمالي عمليات البحث فقد كشفت هذه الدراسة أن بعض المصطلحات تستخدم بشكل متكرر في مقابل الكثير من المصطلحات التي تستخدم مرة واحدة فقط،فقد تم الكشف عن أن 63 مصطلح حظي بتردد ظهور بلغ أكثر من 100 مرة، في حين كون هذه المصطلحات تمثل اقل من 1% من اجمالي المصطلحات. مما يظهر طبيعة أن البحث على الويب يمكن وصفه بانه بحث بمصطلحات تتسم بنسبة منخفضة في تردد الظهور مقابل مصطلحات تتسم بتردد عالي في الظهور[38].

4- الغموض المعلوماتي لمحتوى الويب Ambiguity of information:

وهي تلك الظاهرة التي أرستها بنية روابط شبكة الويب الحالية والناتجة عن ضعف توصيف وتحديد طبيعة الروابط والعلاقات بين المصادر المتاحة على الويب، فما تشير اليه مواقع الويب في شبكة الويب الحالية هو ان الموقع (أ) يرتبط بالموقع (ب) دون وجود لدلالة أو توصيف لطبيعة هذة العلاقة فهل هي علاقة ابوة وبنوة أم علاقة اشتمال أم اكتمال أو غيرها من انماط العلاقات بين الكيانات، وبمعنى أخر يمكن القول بأن الروابط في الويب هي روابط صماء.

5- صبغ عملية انتاج المحتوى على الويب بصبغة الهدف Target:

[37]Lim, E.P., Sun, A.(2005) : Web Mining - The Ontology Approach. In: Proceedings of The International Advanced Digital Library Conference (IADLC 2005), Nagoya, Japan (August 2005) available at:*iadlc.nul.nagoya*-http://iadlc.nul.nagoya-u.ac.jp/archives/IADLCpresen/Lim.pdfdate: 7/3/2012.
[38]Ibid .

فإنتاج المحتوى على الويب يتسم بطابع الهدف Target بمعنى أن المحتوى ينقسم إلى فئتين وفقا للهدف المرجو منه:

3. الفئة الأولى: يتمثل في المعلومات التي يتم انتاجها اساسا للاستخدام من قبل البشر والمستفيدين، المتمثلة في الرسالة الفكرية لدى المؤلفين والتي يقوم بنقلها من خلال الويب.

4. الفئة الثانية: فيتمثل في المحتوى الذي ينتج بهدف أن يستخدم من قبل البرمجيات واجهزة الحاسب ولا يمكن للبشر أن يستفيدوا منه، ويتمثل في بروتكولات الاتصال ولغات البرمجة البيبنة ونظم ادارة قواعد البيانات.

هذا الأمر الذي أسفر عن عجز التطبيقات والبرمجيات من فهم المحتوى والمعلومات المقدمة نظرًا لعدم وجود نسق عالمي للتشغيل المتبادل[39]

6- تحمل اللغات الطبيعية المعبر بها عن محتوى الويب الكثير من القضايا الشائكة:

التي تتعلق بالترادف والأضداد والتروية والجناس اللفظي، تلك القضايا التي تنسحب بطبيعة الحال على مجال المعالجة الآلية للمحتوى من جانب، وعلى تفاعل المستفيدين مع انظمة البحث والاسترجاع من جانب أخر في ظل الاعتماد على التعبير اللفظي والنصي عن الحاجة المعلوماتية.

7- محدودية المعالجة الذكية لمحتوى الويب:

فالمعالجة الآلية لمحتوى الويب تعتمد بالأساس على المعالجة المفردة للكلمات الأمر الذي يعرف بالمعالجة المعجمية واللغوية للمحتوى Lexicon Handling دون ان تعمل على المعالجة وفقا للمفاهيم والدلالات والسياقات الواردة بها كلمات النص الأمر الذي يصعب القدرة على تحقيق التكامل المعرفي المعتمد في الأساس على معالجة المعاني وفقا لفهم المضمون والسياق.

8- صعوبة تحديد المصداقية والموثوقية لمحتوى الويب:

فكثير من محتوى الويب يعلوه سمة خلط بين الحقيقي وما هو زيف، ولا يمكن تحديد درجة أو معيار للموثوقية للمحتوى. على الرغم من كون الفكرة الأساسية للتكامل المعرفي هو القدرة تبادل المعلومات والتكامل البيني مطمئنة في ذلك الى مصدقيتها.

9- التعامل الذكي.

تقف الويب بمعماريتها وتقنياتها ثابتة أما ابواب التعامل الذكي مع المعلومات والمحتوى فليس لديها ملكة الذكاء في تحليله بناءا على مضمونه وكذلك تتلاشى قدرات الربط بين المعلومات واستنباط نتائج جديدة من واقع مما هو متاح.

[39]Sanjib kumar (march 2009), "TOWARDS SEMANTIC WEB BASED SEARCH ENGINES" National Conference on "Advances in Computer Networks & Information Technology (NCACNIT-09) March 24-25, available at :http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5974163 at date: 7/2/2012.

Cybrarians Journal
ISSN 1687-2215
دورية إلكترونية فصلية محكمة ومتخصصة في مجال المكتبات والمعلومات
العدد 34، مارس 2014

10- المحتوى المغلق.

على الرغم من أن الويب قد اتسمت بانها قاتل للتطبيقات Application killer بما قامت بتوفيره من برامج الويكي والمدونات وغيرها من التطبيقات التي كفلت التحرير الحي للمحتوى على الويب، الا ان هذا المحتوى عادة ما ينشر هذا المحتوى على الويب في صورة اقرب للصومعات مما يعيق عملية اعداة استخدامها أو تدوريها من قبل برمجيات اخرى على الويب.

11- افتقار الدلالة في مخططات ونماذج الوصف:

فالويب تشتمل على نماذج توصيفية (كالميتاداتا) تعتمد على وصف المصادر من خلال مخططات (كالدبلن كور) وتفتقر هذة المخططات والنماذج إلى الدلالة والمعاني، فيقتصر دورها فقد على توصيف المحتوى دون بعبارات حرة، بصورة قد لاتشير بالشكل الكامل إلى محتوى الموقع أو الصفحة فضلا عن عجزها في امكانية تحقيق التشغيل المتبادل.

12- ديناميكية البيانات كتحد لمحركات البحث Data Dynamic:

تعد احد السمات التي يتميز بها محتوى العنكبوتية انه محتوى ديناميكي الوجود، بسبب ديناميكية العنكبوتية والانترنت حيث ترتفع معدلات التعديل في محتوى العنكبوتية، ان المرجعية الاساس وراء ديناميكية محتوى العنكبوتية يعود إلى 3 اسباب:

- المحتوى ذات الطبيعة الحساسة للوقت: كالمواقع التي تشتمل على اسعر الاسهم وعناوين الاخبار.
- المحتوى المولد وفقا بصورة مخصصة لمستفيد ما: وتتمثل فيما يعرف بالمو اقع ذات سمات الخاصة لمستفيد ما personalization من (حيث الشكل والسمات والبنية والمحتوى) لتناسب مستخدم ما.
- المحتوى المولد بناءا على المدخلات: ويتمثل في المواقع التي تعتمد على استقبال مدخلات من قبل المستفيد كشاشات قواعد البيانات.

الجانب الآخر هو معدلات التعديل سواء كان بالحذف او الاضافة وقد قدر بنحو 80% يوميا، وفي دراسة قام بها Ntouls &Olson عن تحديث صفحات العنكبوتية وجدا أن 320 مليون صفحة جديدة تضاف اسبوعيا كما ان 20% من صفحات العنكبوتية اليوم سوف تختفي خلال عام واحد . كما ان 50% من محتوى تلك الصفحات سوف يتغير خلال نفس الفترة. ولا يقتصر الامر على المحتوى فحسب بل تمتد هذة الديناميكية لتشمل هيكلة البيانات وقوالبها والتي اتسمت بعدم الاستقرار والديناميكية والمرجعية في ذلك تعود إلى وفرة برامج ادارة المحتوى الرقمي على العنكبوتية، يمكن اجمال ديناميكية العنكبويتة من خلال المؤشرات الاتية:

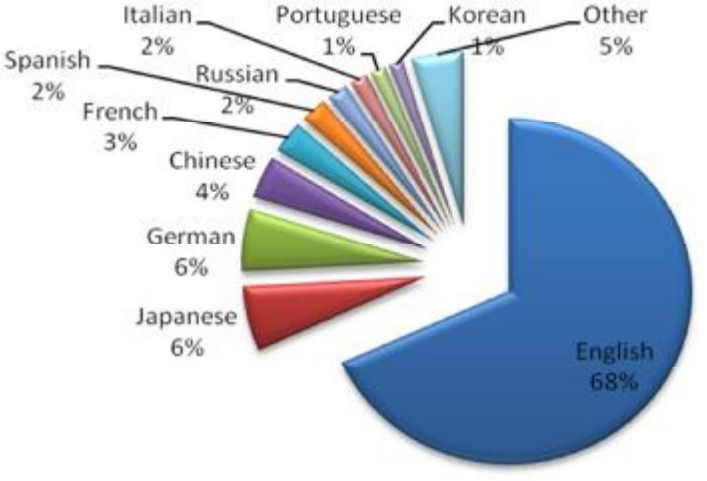
- لايزيد عمر 50% من اجمالي صفحات العنكبويتة عن 100 يوم بينما نسبة 25% من صفحات العنكبويتة تزيد عمرها عن سنة واحدة.

- 40% من الصفحات ذات النطاق com. تتغير كل يوم، بينما المتوسط العمري للصفحات ذات النطاق gov. و النطاق edu. لايزيد عن اربعة اشهر.
- متوسط عمر محتوى العنكبوتية عامان.
- متوسط عمر المحددات الفريدة للمواقع لايتجاوز 4 سنوات.
- متوسط عمر bookmarks لايتجاوز الشهران[40].

13- التباين اللغوي:

ففضلا عن التباين في التكويد وصيغ التكويد داخل الويب تتفاقم ظاهرة أخرى تحول بين التكامل المعرفي لمحتوى الويب وهي تعدد لغات المحتوى بين العربية والعبرية والفرنسية والصينية واليابانية وغيرها من اللغات الحية المنطوقة. الأمر الذي استتبع أن تقوم برمجيات معالجة المحتوى بالتركيز على لغة محتوى بعينه دون الاخر لتظهر بعد ذلك طوائف البرمجيات وفقا للغات ووفقا للنطاق الجغرافي للتركيز على المحتوى في لغات محددة وفي مناطق جغرافية محددة،[41].

اما عن واقع اللغة العربية على الويب فيتمثل في حجم المحتوى العربي المتاح والذي يشغل نسبة 0.2% من حجم المحتوى المتاح على الانترنت بواقع 100 مليون صفحة[42].

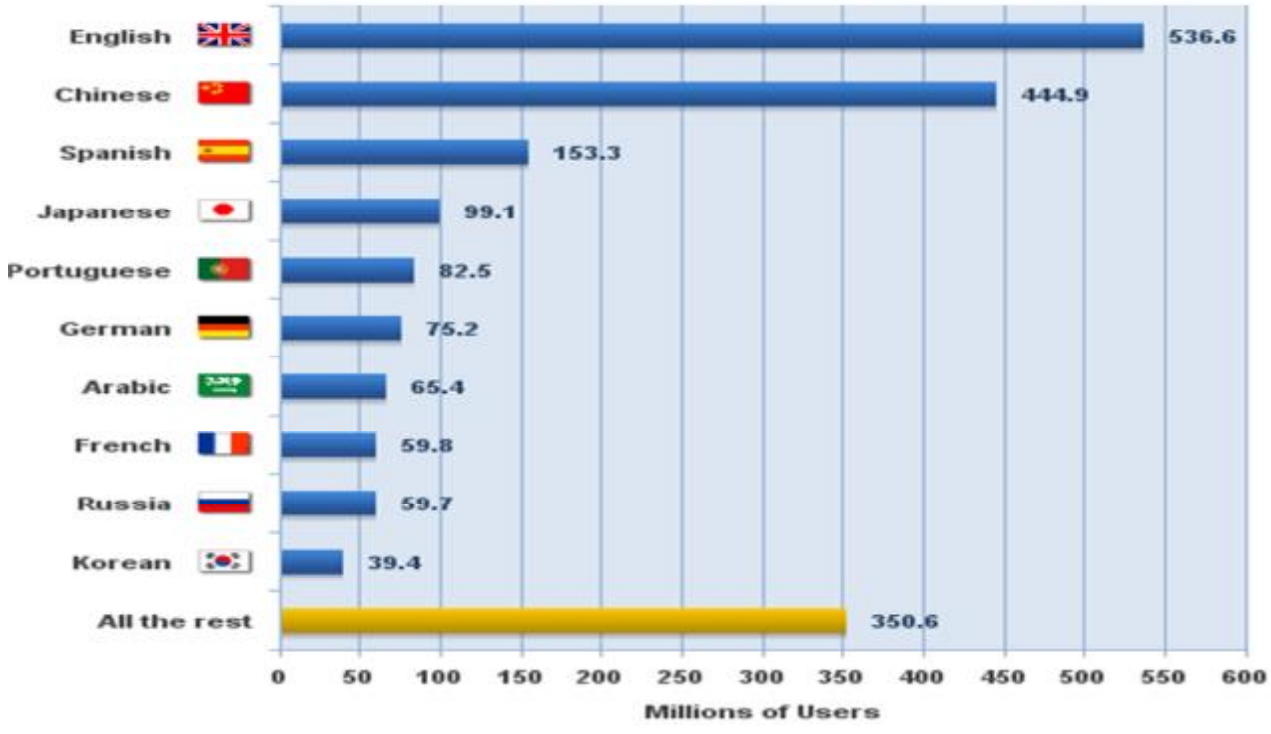


شكل رقم (5) يوضح واقع حجم لغات المحتوى على الويب.

[40] Terrence A. Brooks. Web search: how theWeb has changed information retrieval. Information Research, 8(3):(paper no. 154), April 2003.
[41] Andrew Hammond (2004). Arabic search engine may boost content available at . http://www.abc.net.au.
[42] http://www.translate-to-success.com/online-language-web-site-content.html

اللغة	جمالي نسبة المحتوى
glish	.40%
panese	90%
erman	80%
inese	90%
ench	00%
anish	40%
ssian	90%
lian	60%
rtuguese	40%
rean	30%
abic	60%

جدول رقم (1) يوضح حجم المحتوى المصاغ باللغة العربية على صعيد العالم[43].



شكل رقم (6) يوضح ترتيب اللغة العربية من حيث عدد مستخدميها على الانترنت[44]

[43] 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
[44] Ibid

14- الروابط الصماء:

على الرغم من ان بنية الويب الحالية تمتاز بالروابط بين نصوصها ومحتواها ألا أنها تعد من أكثر المشكلات في الويب، فالروابط داخل بنية الويب لاتحدد طبيعة أو نمط الارتباط بين المصادر والكيانات فهي تشير فقط أن الموقع (أ) يرتبط بالموق(ب) دون تحديد لنوع الربط ودرجته أو الموثوقية فيه، فوفقا لتعريف Tim Lee للنص الفائق فان أي شئ يمكن أن يرتبط بأي شئ وبالتالي ووفقا لهذا المفهوم يمكن أن ترتبط معلومات اكاديمية بمعلومات تجارية[45].

وعلى هذا يتضح أن شبكة الويب الحالية تواجه العديد من القضايا الشائكة والمشكلات المتفاقمة التي كانت دافعا للتوجه الى التنقيب الى البيانات Web Data Mining بغية التكامل المعرفي.

[45] Berners-Lee, T. (n.d.). The Semantic Web: Scientific American. *Science News, Articles and Information | Scientific American*. Retrieved August 2, 2011, from http://www.scientificamerican.com/article.cfm?id=the-semantic-web