خوارزم مقترح للتنقيب في الويب لتمييز أنماط المستخدمين والتنبؤ بمتطلباتهم *

إعداد

أ.د. عطا إبراهيم إمام الألفي
أستاذ الحاسب الآلي والنظم المعلوماتية المتفرغ
كلية التربية النوعية - جامعة المنصورة

أ.د. محي الدين إسماعيل موسى
أستاذ الحاسب الآلي والنظم المعلوماتية
كلية التربية النوعية - جامعة المنصورة

د. نبيل عبد المحسن أحمد موسى
مدرس الحاسب الآلي
كلية التربية النوعية - جامعة المنصورة

م.م. حنان الرفاعي عبد القادر
مدرس مساعد الحاسب الألى
كلية التربية النوعية - جامعة المنصورة

مجلة بحوث التربية النوعية ـ جامعة المنصورة
عدد (٣٣) ـ يناير ٢٠١٤

* بحث مستل من رسالة دكتوراه

خوارزم مقترح للتنقيب في الويب لتمييز أنماط المستخدمين والتنبؤ بمتطلباتهم

إعداد

أ.د. عطا إبراهيم إمام الألفي*
أ.د. محي الدين إسماعيل موسى**
د. نبيل عبد المحسن أحمد موسى***
م.م. حنان الرفاعي عبد القادر****

مقدمة :

تعتبر شبكة المعلومات العالمية World Wide Web أكبر مستودعات الوثائق المعروفة، ومنذ بدايتها تتزايد كمية البيانات بمعدل كبير، ومع هذا النمو الهائل تزداد صعوبة وصول المستخدم للمعلومات التي يحتاج إليها، كما تزداد صعوبة تصنيف وفهرسة تلك المعلومات الأمر الذي يجعل استعراض كل النتائج مضيعة لوقت المستخدم (Anthony. S, 2005, 20). وترجع ضخامة حجم المعلومات الموجودة على الشبكة إلى الحرية في تأليف ونشر المحتوى وعدم وجود سلطة تتحكم أو تسيطر على نشر المعلومات على الانترنت مما أدى إلى زيادة حجم المعلومات وتكرارها، وبسبب هذا فإنه على خلاف قواعد البيانات العلائقية فإن المعلومات على الويب ضعيفة البنية (Derar. H, 2009, 35).

وتلعب نظم تكامل المعلومات Information Integration Systems مثل محركات البحث دوراً هاماً في جعل هذا الكم الهائل من المعلومات قوي البنية وسهل الوصول إليه بسهولة وبشكل أكثر فائدة ورغم نجاح تلك الأنظمة في معالجة العديد من التحديات ألا أنها ما زالت تواجه العديد من القيود، خاصة عند استخراج واستكمال المحتوى من قواعد بيانات الويب التي تكمن وراء واجهات البحث، ولمعالجة تلك القضايا لابد من توجيه أنظار الباحثين إلي الاهتمام بتقنيات التنقيب في الويب Web Mining كمجال هام للبحث والدراسة (Zhongming. M, 2007, 23).

ويعد التنقيب في الويب Web Mining أحد تطبيقات تعلم الآلة Machine Learning على بيانات مبنية على الويب من أجل التعلم واستخراج المعرفة. التنقيب في الويب يمكن تصنيفه إلي ثلاث تصنيفات مميزة وهي التنقيب في استخدام الويب Web Usage Mining، التنقيب في بنية الويب Web Structure Mining، التنقيب في محتوى الويب Web Content Mining (Fatih. G, 2007, 18).

* أستاذ الحاسب الآلي والنظم المعلوماتية المتفرغ كلية التربية النوعية - جامعة المنصورة
** أستاذ الحاسب الآلي والنظم المعلوماتية كلية التربية النوعية - جامعة المنصورة
*** مدرس الحاسب الآلي كلية التربية النوعية - جامعة المنصورة
**** مدرس مساعد الحاسب الآلى كلية التربية النوعية - جامعة المنصورة

ويتعامل نظام التنقيب في الويب مع البيانات المدخلة من خلال ثلاث مراحل مختلفة حتى تصل إلي النتيجة النهائية، وتتمثل تلك المراحل في مرحلة ما قبل المعالجة Pre-processing، مرحلة التنقيب في البيانات Data Mining، مرحلة بعد المعالجة Post Processing. عندما تكون البيانات مدخلة في نظام التنقيب في البيانات فإن مرحلة ما قبل المعالجة تصبح ضرورية لنقل البيانات الخام إلي صيغة مقبولة وتلك المرحلة ربما تشمل تقليل الحقول غير المرتبطة وتنقية البيانات من المعلومات المشوهة Noisy Information وهذه الخطوة ضرورية جدا حيث أن صفحات الويب تحتوي على نسبة عالية من المعلومات المشوهة (John.L, Michel . V, 2007, 54-55).

مشكلة البحث:

يتميز هذا العصر بالتغيرات السريعة الناجمة عن التقدم العلمي والتكنولوجي وتقنية المعلومات، لذا تهتم المؤسسات التعليمية دائما بالحديث في مجال تكنولوجيا المعلومات والإنترنت حيث أنها تعتمد بصورة كبيرة حاليا على شبكة الإنترنت وخدماتها في التعامل مع المعلومات ونقلها من جهة إلى أخرى، خاصة بعد التطور الهائل في مجال التعليم الإلكتروني والإدارة الإلكترونية.

فعلى سبيل المثال وليس الحصر فجامعة المنصورة تعتمد في عملها على الإدارة الإلكترونية والتعليم الإلكتروني وهناك العديد من النظم الإلكترونية التي تستخدمها مثل (نظام ابن الهيثم لإدارة الدراسات العليا، نظام المستقبل لإدارة المكتبات، نظام الحسابات الخاصة، نظام الموازنة العامة لجامعة المنصورة، نظام حفظ المستندات لجامعة المنصورة، نظام الأمين لإدارة المخازن والعهد، نظام التصحيح الإلكتروني)، كذلك قامت بتحويل العديد من المقررات إلى مقررات تعليم إلكتروني.

وتعتمد أنظمة الإدارة الالكترونية وبرامج التعليم الإلكتروني في المقام الأول على استخدام شبكة الانترنت وتبادل كم هائل من المعلومات بينها، عن طريق موقعها الذي يقدم خدماته لجميع مستخدمي الموقع سواء من داخل الحرم الجامعي أو من خارجه فإنه تعتريها ساعات ذروة نظرا لكثرة مستخدمي الموقع وكذلك ضخامة حجم البيانات المرتبطة بالموقع مما قد يتسبب في مجموعة من المشكلات مثل تعذر الدخول إلى الموقع أو الدخول ولكن لا يتم تنفيذ بعض متطلبات المستخدم كتحميل البرامج والأنظمة للمستخدمين. الأمر الذي يتطلب مساعدة المستخدمين لتخفيف العبء عن الموقع عن طريق استخدام تقنيات جديدة مثل تقنيات التنقيب في الويب للمساعدة في حل تلك المشاكل.ومن ثم فالمؤسسات التعليمية تفتقد لنظم التنقيب في الويب لتمييز أنماط المستخدمين والتنبؤ بمتطلباتهم لتحسين أداء المواقع الالكترونية، وبالتالي يمكن صياغة مشكلة البحث في التساؤل الرئيسي التالي:

كيف يمكن بناء نظام تنقيب في الويب لتمييز أنماط المستخدمين والتنبؤ بمتطلباتهم لتحسين أداء المواقع الالكترونية؟

ويتفرع من هذا السؤال الرئيسي التساؤلات الفرعية التالية:

١. ما الخوارزميات المستخدمة للتنقيب في الويب؟

٢. كيف يمكن تمييز أنماط مستخدمي المواقع الالكترونية إلي فئات متشابهة؟

٣. كيف يمكن التنبؤ بمتطلبات مستخدمي المواقع الالكترونية؟
٤. كيف يمكن تحسين أداء المواقع الالكترونية باستخدام تقنيات التنقيب في الويب؟

أهداف البحث:

١. التعرف على أهم الاتجاهات الحديثة في مجال التنقيب في الويب.
٢. تحديد أسس تصميم وبناء نظم التنقيب في الويب.
٣. التعرف على نوعية البيانات التي يمكن استنباط معرفة منها وتصنيفها طبقاً لاهتمامات مستخدمي موقع جامعة المنصورة.
٤. بناء خوارزم مقترح للتنقيب في الويب.
٥. توفير نظام تنقيب في الويب لتمييز أنماط المستخدمين والتنبؤ بمتطلباتهم لتحسين أداء المواقع الالكترونية.
٦. التعرف على مدى الاستفادة من استخدام النظام المقترح في تحسين أداء المواقع الالكترونية.

أهمية البحث:

١. توجيه أنظار الباحثين إلي الاهتمام بتقنيات التنقيب في الويب Web Mining كمجال هام للبحث والدراسة.
٢. تقديم تقنية جديدة مقترحه تعمل على تحسين عملية التنقيب في مواقع الويب لتقليل التحميل الزائد على المواقع الالكترونية لتوفير وقت وجهد المستخدم.
٣. توفير خوارزم تنقيب في الويب لتمييز أنماط المستخدمين والتنبؤ بمتطلباتهم لتحسين أداء المواقع الالكترونية يمكن استخدامه في تطوير نظم مشابهة.

منهج البحث:

يتبع البحث منهجين هما:

١. **المنهج الوصفي:** ويرتبط مفهوم المنهج الوصفي بتوضيح واقع الأحداث ولا يتوقف عند وصف الواقع على تقرير حقائقه الحاضرة كما هي، بل يتناولها بالتحليل والتفسير لغرض الاستنتاج لتصحيح الواقع أو تحديثه أو استكماله (محمد زياد حمدان، ١٩٩٩، ٦٦). وقد استخدم المنهج الوصفي لمعالجة الإطار النظري الخاص بالبحث من خلال وصف وتفسير وتحليل المفاهيم الخاصة بالتنقيب في الويب وكذلك تمييز الأنماط من اجل تحسين أداء المواقع.
٢. **المنهج التجريبي:** لتصميم وإنتاج نظام تنقيب في الويب لتمييز أنماط المستخدمين والتنبؤ بمتطلباتهم، وقياس فعاليته في تحسين أداء المواقع الالكترونية.

أدوات البحث:

١. خوارزم مقترح للعنقدة (Clustering).
٢. خوارزم مقترح للتصنيف(Classification) واستخلاص القواعد Rules Extraction.

مصطلحات البحث:

١ـ التنقيب في الويب Web Mining:

يعرفه (Anthony Scime,2005,20) على أنه الانتقال بشبكة المعلومات العالمية تجاه بنية أكثر فائدة حيث يستطيع المستخدمين إيجاد المعلومات التي يحتاجون إليها بسرعة وسهولة. وهو يتضمن استكشاف وتحليل البيانات، الوثائق والوسائط المتعددة من شبكة المعلومات العالمية.وتستخدم تلك التقنية محتويات الوثائق، بنية الروابط الفائقة واستخدام الإحصائيات لمساعدة المستخدمين في تلبية احتياجاتهم من المعلومات.

ويعرف البحث الحالي التنقيب في الويب بأنة "عملية اكتشاف وتحليل واستخراج المعلومات المفيدة من شبكة المعلومات العالمية من أجل تقديم خدمة أكثر سرعة وكفاءة".

٢ـ تمييز الأنماط Pattern Recognition

يعرفه (Richard . D, and et al., 2001, 35) بأنة عملية تصنيف البيانات إلى مجموعة من الفئات إما على أساس معرفة مسبقة أو على المعلومات الإحصائية المستخرجة من الأنماط، البيانات المصنفة عادة ما تكون مجموعات من القياسات أو الملاحظات.

٣ـ ملف دخول المستخدم Log file

يعرفه (Vel'asquez, J. D., Palade V.,2008) بأنه ملف يستخدم لتسجيل كافة بيانات المستخدم خلال إبحاره داخل الموقع وتستخدم تلك البيانات لاستخراج معرفة وأنماط تساعد على تلبية متطلبات المستخدمين.

٤ـ العنقدة (التجميع Clustering)

يعرفه (Hannah, H., Thangavel, K., 2009, 3) بأنها تقنية لجمع مجموعة العناصر التي تمتلك صفات متشابه في مجموعة واحدة. والعنقود هو مجموعة من العناصر المتشابهة فيما بينها، وغير مشابهة للعناصر المنتمية إلى عناقيد أخرى.

٥ـ التصنيف

يعرفه ((Kamber,M . and Han, J.(2006) عملية ربط عنصر بيانات بأحد الفئات المعرفة مسبقا وقد تشير إلى العملية التي يتم بها التعرف على الأفكار والأشياء متباينة ومفهومة. خوارزميات التصنيف الأكثر شيوعا هي شجرة القرار والشبكات العصبية. وهناك أيضا أساليب أخرى لاستخراج أنماط الاستخدام من سجلات ويب.

دراسات سابقة:

١- **دراسة (David. S, Antonio. M, 2005)** بعنوان " تقنيات التنقيب في الويب من أجل الاستكشاف الأوتوماتيكي للمعرفة الطبية" استهدفت هذه الدراسة توضيح أهمية شبكة المعلومات العالمية كأداة حيوية للباحثين ومهندسي المعلومات والشركات الطبية والممارسين من أجل استخراج المعرفة.فتقترح هذه الدراسة طريقة أتوماتيكيه ومستقلة من اجل استخراج التصنيفات من

المصطلحات على الويب وتقدم وثائق الويب المسترجعة بترتيب ذو مغزى. كما قدمت هذه الدراسة طريقة جديدة لاكتشاف المرادفات والlexicalization والنتائج التي تم التوصل إليها كانت مفيدة جدا لتسهيل عملية الوصول إلي مصادر الويب في أي مجال طبي أو عروض Ontological.

٢- **دراسة (Srivastava.J , et al, 2005)** بعنوان " التنقيب في الويب – المفاهيم والتطبيقات واتجاهات البحث.استهدفت الدراسة توضيح أهمية استخراج معلومات ذات قيمة من الويب وأن تقنيات التنقيب في البيانات لاستخراج المعرفة من المحتوى ، والبنية والاستخدام هو تجميع لمجموعة من التكنولوجيات لتحقيق تلك الإمكانية. وأن الاهتمام بمجال التنقيب في الويب قد نمى بسرعة في مجالات البحث والمجالات الاجتماعية الأخرى. فتقدم هذه الدراسة لمحة مختصرة عن الانجازات في هذا المجال سواء من حيث التقنيات والتطبيقات ويحدد التوجهات الرئيسية للبحث في بحوث المستقبل. وتوصلت هذه الدراسة إلي أهمية تحليل البيانات واستخراج جميع أنواع المعرفة المفيدة منه. وتوصلت الدراسة أيضاً إلى وصف مساهمات العملية الأساسية للحاسب التي ساهمت في نمو هذا المجال وتطوره.

٣- **دراسة (Lan.Y, Bing. L, 2006)** بعنوان " تنقية صفحة الويب للتنقيب في الويب خلال وزن الحقول".

استهدفت هذه الدراسة إيضاح أنة على خلاف البيانات التقليدية أو النصية فإن صفحات الويب تحتوى على كمية كبيرة من المعلومات والتي لا تعتبر جزء من المحتويات الرئيسية لتلك الصفحات مثل أشرطة الإبحار، الإعلانات ، ومعلومات حقوق النشر، هذه المعلومات غير زى علاقة وتسمى ضوضاء صفحة الويب Web page noise مثل العنقده Clustering والتصنيف Classification. تقترح هذه الدراسة تقية جديدة لوزن الحقول Feature Weight للتعامل مع ضوضاء صفحة الويب من أجل تحسين التنقيب في الويب.هذه الطريقة تنشئ ما يسمى شجرة التركيب المضغوطة Compressed Structure Tree وتقوم باستخدام المعلومات المبنية على القياس لتقييم أهمية كل عقدة node في شجرة التركيب المضغوطة. وبناء على الشجرة وأهمية قيم عقدها فإن تلك التقنية تخصص وزن لكل حقل في كل محتوى. الأوزان الناتجة تستخدم في التنقيب في الويب ويتم تقييم التقنية المقترحة عن طريق مهمتين للتنقيب في الويب هما:عنقدة صفحة الويب Web Page Clustering - تصنيف صفحة الويب Web Page Classification. وتوصلت النتائج إلى أن طريقة القياس المستخدمة في الدراسة استطاعت أن تحسن نتائج التنقيب بشكل ملحوظ.

٤- **دراسة (David. J, et al 2007)** بعنوان (التنقيب المرئي للويب)تتضمن تحليل بيانات استعمال موقع الويب تحديين هامين:أولاً حجم البيانات الناشئ عن نمو الويب، وثانياً، التعقيد الهيكلي لمواقع الويب استهدفت الدراسة تطبّيق تقنيات التنقيب في البيانات والمعلومات المرئية إلى مجال الويب لكي يستفيد من قوة كل من فهم الرؤية البشرية واستعمال الحاسبات نعين هذا التنقيب في الويب البصري ردا على التحديين حيث نطبق تقنيات التنقيب في البيانات إلى مجموعات

بيانات الويب الضخمة وتستهدف الدراسة تلخيص التقنيات المستخدمة، وتستخدم أساليب تصور المعلومات على النتائج. الهدف هو ربط نتائج التنقيب في تسجيلات استخدام الويب وبنية الويب المستخرجة عن طريق تركيب النتائج بشكل بصري. وتقترح الدراسة مجموعة من تخطيطات المعلومات المرئية الجديدة وتحليل أدواتهم والتحكم في بنية نموذج التطبيق.

٥- **دراسة (Cheng. Y, et al, 2009)** بعنوان " تقنيات التنقيب في الويب لتسويق الكتب على الانترنت"

استهدفت الدراسة توضيح للعملاء المحتملين للكتب على الانترنت عن طريق تنقيب محتوى الويب. فهذه الدراسة أولا تقوم بعمل قائمة من العلماء الذين يرتبط مجال بحثهم في تكنولوجيا المعلومات، ثم بعد ذلك يتم استخدام محرك البحث لحساب عدد صفحات الويب المرتبطة بخبرة العلماء. هذه البيانات يتم معالجتها قبلياً بواسطة ثلاث خطوات رئيسية قبل استخدامها وهي:

١. ترشيح أو فرز البيانات الغير طبيعية.
٢. تطبيع البيانات Normalizing Data
٣. توليد البيانات الثنائية.

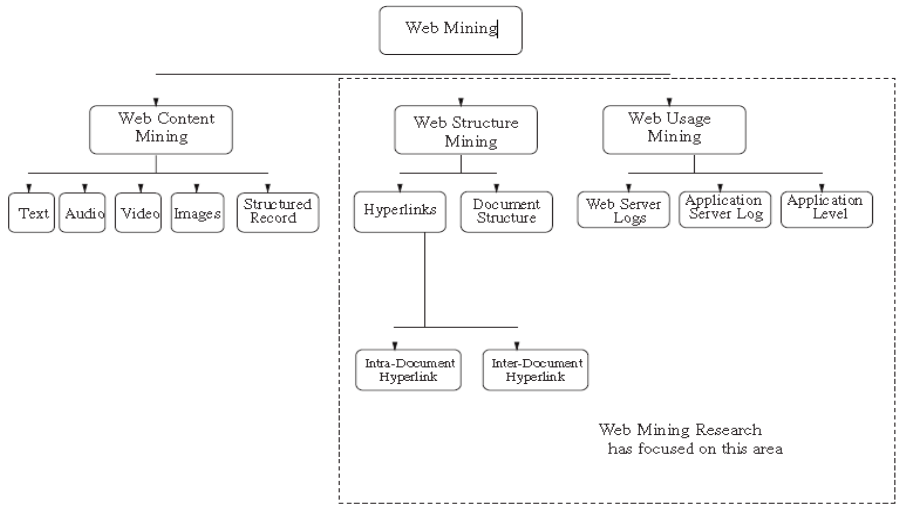
تم استخدام تحليل الارتباط والتحليل العنقودي الهرمي وذلك لتوليد عنقدة العلماء وخبرتهم من اجل اختبار دقة استخدام التنقيب في الويب للتنبؤ بالعلماء المهتمين بقوائم الكتب.وتوصلت النتائج إلى أن معدل الدقة لقوائم الكتب الموصي بها هام من الناحية الإحصائية.

٦- **دراسة (Tuncay. S, Necmi. D, Zafer. C, 2010)** بعنوان " التنقيب في الويب وبيئات التعلم الافتراضية" ، استهدفت هذه الدراسة توضيح أن البيئات التعليمية تنقسم إلى شكلين أساسين هما(بيئات التعلم التقليدية، بيئات التعلم الافتراضية. ويوجد مشاكل في تلك البيئات تشمل مشكلة التقييم وتحليل السلوك. ففي التعليم التقليدي تحليل السلوك يتم بسهولة من خلال أساليب الملاحظة بينما في التعليم الافتراضي فإن تحليل اتجاهات الطلاب وسلوكهم يعتبر مشكلة هامة. تطبيقات التنقيب في الويب تستخدم في إعطاء معلومات ذات معنى من خلال سلوكيات غير ذات مغزى. التجول داخل بيئات الويب هي الوسيلة للتغلب على تلك المشاكل. وأثبتت هذه الدراسة أن تطبيقات التنقيب في الويب في بيئات التعلم الافتراضية شكلت توافق مع طريقة البحث الوصفية ونتيجة لذلك تم إدخال التنقيب في الويب في بيئات التعلم.

الإطار النظري:

تصنيف التنقيب في الويب

ويمكن تقسيم التنقيب في الويب إلى ثلاث فئات متميزة، وفقا لأنواع البيانات التي يمكن التنقيب فيها. ويمكن إيضاح لمحة موجزة عن الفئات الثلاث في الشكل التالي ( ) Rathod, D. (2012)).



شكل (١) تصنيف التنقيب في الويب

التنقيب في محتوى الويب (Web Content Mining (WCM)

هو عملية استخراج معلومات مفيدة من محتويات مستندات ويب، ويمكن أن تتألف من النصوص والصور والصوت والفيديو، أو سجلات منظمة مثل القوائم والجداول. ويمكن معاملة التنقيب في محتوى الويب كامتداد لمحرك البحث الأساسي (Search Engine). معظم محركات البحث تعتمد على الكلمة المفتاحية. في حين أن التنقيب في محتوى الويب يمتد إلى تكنولوجيا استرجاع المعلومات التقليدية (Information Retrieval (IR)) من خلال بناء المفاهيم الهرمية، واستخراج ملفات تعريف المستخدمين وتحليل الروابط بين صفحات الويب (Taherizadeh, S., Moghadam N. 2009).

التنقيب في بنية الويب (Web Structure Mining (WSM))

هو العملية التي من خلالها نكتشف نموذج لبنية روابط صفحات الويب. وتوليد المعلومات مثل التشابه والعلاقات فيما بينها من خلال الاستفادة من الارتباط التشعبي. فيتم فهرسة الروابط ، وتوليد المعلومات مثل التشابه Similarity والعلاقات فيما بينها. يهدف التنقيب في الويب إلى ملخص هيكلي لموقع وصفحات الويب (Jain,R. Purohit, G. N.,2011, 0975 – 8887).

التنقيب في استخدام الويب (Web Usage Mining (WUM

التنقيب في استخدام شبكة الويب هو عملية تطبيق تقنيات التنقيب في البيانات لاكتشاف أنماط السلوك على أساس البيانات المبنية على شبكة الإنترنت، لمختلف التطبيقات(Kohavi. R) and Parekh. R, 2003.

مراحل التنقيب في استخدام الويب

يوفر التنقيب في استخدام الويب مدخلاً لجمع وتجهيز البيانات على شبكة الإنترنت، ويبني نموذج يمثل سلوك المستخدمين ومتطلباتهم. ويتكون التنقيب في استخدام الويب من المراحل الأساسية التالية(Gutschmidt, A., et al, 2008) :

١- جمع البيانات Data collection.
٢- معالجة البيانات Data preprocessing.
٣- اكتشاف الأنماط Pattern discovery.
٤- تحليل الأنماط Pattern analysis.

١- جمع البيانات Data collection.

جمع البيانات مهمة ضرورية في أي تطبيق لاستخراج البيانات و خلق مجموعة بيانات ذات هدف مناسب. قد تشمل عملية ما قبل معالجة البيانات الأصلية، ودمج البيانات من مصادر متعددة، وتحويل البيانات إلي شكل مناسب للاستخدام في عمليات تنقيب بيانات محددة. يتم جمع البيانات من خوادم الويب Web Servers، من العملاء المتصلين بالخادم، أو من مصادر وسيطة مثل خوادم بروكسيproxy servers (Dunham,H., ,2003)

٢- معالجة البيانات Data preprocessing

تهدف معالجة سجل دخول الويب (Log File) إلي إعادة تهيئة تسجيلات الدخول الأصلية لتحديد جميع جلسات (Sessions) الوصول إلى شبكة الإنترنت. خادم الويب عادة يسجل أنشطة دخول كافة المستخدمين للموقع وسجلات خادم الويب Web server logs. هناك العديد من أنواع سجلات الويب ويوجد العديد من المهام التي يتعين القيام بها على سجلات خادم الويب قبل تنفيذ خوارزميات التنقيب على الويب. وتشمل تلك المهام، تنقية البيانات data cleaning ، تعريف المستخدمين، تمييز المستخدمين user differentiations، تحديد الجلسات session identification(Ling Zheng, et al, 2010, . VI-19-VI-21.). سجلات الدخول الأصلية يتم تنقيتها، إعادة تهيئتها، ومن ثم تجميعها في مجموعات ذات معنى قبل أن يتم استخدامها من قبل التنقيب في استخدام الويب (SubMasthan, T., et al.. 2012, 307-312).

٢,١ مكونات ملف الدخول Log File:

يتكون ملف دخول المستخدم من مجموعه من العناصر مثل رقم العنوان Ip adress : وهو رقم الجهاز الذي يتم الدخول على الموقع من خلاله، اسم العميل Client name وهو اسم المستخدم في حاله وجود حماية بكلمة سر، الوقت والتاريخ Time stamp وهو وقت وتاريخ الزيارة كما يتم رؤيتها من خادم الويب، access request ويشمل اسم الرابط الذي تم طلبه والبرتوكول المستخدم، رابط المرجع Refferr Url ويشير إلى الرابط المصدر الذي تم الدخول من خلاله ، وكيل المستخدم User Agent ويشمل الوكيل أو المستعرض المستخدم في التصفح وليكن الانترنت

اكسبلور Internet explorer أو موزيلا Mozilla. (Vel'asquez, J. D., Palade V. 2008)

٣- اكتشاف الأنماط Pattern discovery

في تلك المرحلة يتم اكتشاف الأنماط التي تكون مطلوبة من أجل تخصيص الويب والتي تستجيب لسلوك واهتمامات المستخدمين، يتم اكتشاف المعرفة عن طريق تطبيق العديد من التقنيات مثل العنقده (التجميع) Clustering ، التصنيف Classification، قواعد الارتباط Association Rules واكتشاف النمط المتسلسل sequential(:(2011).M ,Kiruthika) pattern discovery.

٣,١ العنقدة أو التجميع Clustering :

يمكن أن تتم العنقدة في مجال التنقيب في استخدام الويب على المستخدمين user clusters ،عنقدة الصفحة page clusters وعنقدة الجلسات sessions clusters من ملف الدخول.عنقدة المستخدمين تهدف إلى إنشاء مجموعات من المستخدمين يمتلكون اهتمامات مشتركة بناء على سلوكهم على الموقع. وعلى الجانب الآخر عنقدة الصفحات تكتشف مجموعات من الصفحات تمتلك محتويات متشابهة من أجل تحسين محركات البحث search engines في الويب. بالإضافة إلى ذلك يمكن أن تطبق العنقده على الجلسات sessions حيث كل منها يمكن أن يعرض موضوع واحد هام خلال الموقع (Suresh, K.,et al.,(2011).

٣,٢ التصنيف Classification

يكون الاهتمام في مجال الويب بتطوير الملف الشخصي Profile للمستخدمين المنتمين لفئة معينه.وهذا يتطلب تحديد واستخراج السمات features التي تصف خصائص الفئة المعطاة بأفضل ما يكون.يمكن القيام بالتصنيف عن طريق باستخدام العديد من التقنيات مثل شجرة القرارات decision tree ، K-nearest neighbor classifier ، Support Vector Machine الخ (Usama Fayyad, et al,(2009).

٤- تحليل الأنماط Pattern analysis

تحليل الأنماط هو أخر مرحلة من مراحل التنقيب في الويب والهدف وراء تلك المرحلة هو فرز القواعدrules أو الأنماط patterns الغير هامه من مجموعة الأنماط التي تم الحصول عليها في مرحلة اكتشاف الأنماط.

الإطار التطبيقي:

أولا تصميم وبناء الخوارزم المقترح:

١- ماهية الخوارزم المقترح:

يعتمد النظام المقترح على خوارزمين هما:

- خوارزم مقترح للعنقدة clustering algorithm: وذلك لتجزئة مصفوفة جلسات العمل الـ sessions التي تم إنشاؤها في مرحلة المعالجة وتحويلها إلى مكونات مترابطة (عناقيد clusters).ولتحديد المكونات المترابطة وتخصيصها لعنقود (cluster).
- خوارزم مقترح للتصنيف Classification: وذلك لتصنيف الصفحات إلى أحد المجموعات Clusters التي تم تحديدها مسبقا.

٢- مراحل الخوارزم المقترح:

يمر الخوارزم المقترح بمجموعة من الخطوات، يتم توضيحها كما يلي:

١. مرحلة جمع البيانات Data Collection : في هذه المرحلة، بيانات المستخدم تم تسجيلها في ملف الدخول log file يستخدم في متابعة سلوك المستخدم خلال الإبحار في الموقع وبيانات جلسات العمل Sessions ويتكون من الجدول التالي:

أ- جدول مصفوفة الصفحات Page_array table : يتتبع سلوك المستخدم خلال الإبحار في تسجيلات دخول الموقع (رقم المستخدم user_id - رقم الجلسة session_id - الروابط المطلوبة Urls- والوقت المستغرق timestamp).

جدول (١) جدول دخول الصفحات

جدول : pagearray

session_id	user_id	pagearray	date
hyjy5m5rbx	eng_s	p1,p2,p3,p4,p7,p8,p10	8/21/2012 7:22:07 PM
hjuj1pekib45	dr_hamdy	p1,p3,p4,p5,p6,p7,p10	7/25/2012 11:59:03 AM
bh45lktel13z	i_risho	p2,p3,p4,p6,p7,p9	7/16/2012 12:09:50 AM
se453s2z34nq	dr_hamdy	p1,p3,p4,p5,p6,p8,p10	7/15/2012 11:09:01 PM
kyvxgagf51z2	dr_hamdy	p1,p5,p7,p8,p10	7/15/2012 7:33:03 PM
uuebtp24faje	i_risho	p2,p3,p6,p9	7/15/2012 6:11:12 AM
sl45lc4xwemm	i_risho	p2,p3,p5,p6	7/14/2012 6:01:19 PM
rdgra55tlt45vey	i_risho	p2,p3,p5,p8	7/13/2012 2:21:18 PM
eor45hof4s445	dr_hamdy	p3,p5,p7,p8,p10	7/12/2012 8:01:12 AM
umhea5ua555	i_risho	p3,p5,p7,p10	6/20/2012 9:05:40 PM

السجل: 11 من 11

٢. معالجة البيانات Data Preprocessing: تشمل هذه المرحلة تنقية البيانات Data cleaning، تحديد الجلسة session identification بناء مصفوفة علاقة التشابه بين الجلسات Sessions Similarity.

أ- البيانات Data Cleaning، تحديد الجلسة session identification: وتتم عملية تنقية ملف الدخول log file حيث يتم تحديد مصفوفة صفحات الجلسات sessions فخلال إبحار المستخدم يقوم النظام بتسجيل الطلبات requests ويبنى مصفوفة من الصفحات

والجلسة session (sessions_pages). هذه المصفوفة تتكون من كل الصفحات التي تم الوصول إليها بواسطة المستخدم خلال الجلسة session ويتم تخزين تلك المصفوفة في جدول يسمى array_page فيتم تحويل البيانات في ملف الدخول إلى شكل ثنائي(١،٠) حيث أن الموقع يتكون من ١٠ صفحات (p1–p10) والتعامل مع sessions من (s1– s10) حيث يتم إعطاء الصفحة التي تم الدخول عليها القيمة (١) والصفحة التي لم يتم الدخول عليها القيمة (٠) كما هو موضح بالجدول (2) والعمود الأخير(Count) يدل على عدد الصفحات التي تم الدخول عليها.

جدول (٢) مصفوفة الصفحات وجلسات العمل

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Count
S1	1	1	1	1	0	0	1	1	0	1	7
S2	1	0	1	1	1	1	1	0	0	1	7
S3	0	1	1	1	0	1	1	0	1	0	6
S4	1	0	1	1	1	1	0	1	0	1	7
S5	1	0	0	0	1	0	1	1	0	1	5
S6	0	1	1	0	0	1	0	0	1	0	4
S7	0	1	1	0	1	1	0	0	0	0	4
S8	0	1	1	0	1	0	0	1	0	0	4
S9	0	0	1	0	1	0	1	1	0	1	5
S10	0	0	1	0	1	0	1	0	0	1	4

ب- بناء مصفوفة علاقة التشابه بين الجلسات Sessions Similarity: بعد تحديد مصفوفة الـ session_page، تستخدم لبناء مصفوفة M = n*n حيث n عدد الـ sessions في الموقع. وكل similarity(s1,s2) تقيس التشابه بين الـsessions وبعضها عن طريق المعادلة التالية:

$$Sim \int_{0}^{1} \quad Pm(Si) = Pm(Sj) = 1$$
$$Other\ wise$$

Where:

m=1,….,n1

i=1,….,n2

j=j+i and j>0

$$\text{Simllarty}(Si, Sj) = \frac{\sum_{m=1}^{n1} Sim(Pm)}{Max(\sum_{m=1}^{n1} Pm(Si), \sum_{m=1}^{n1} Pm(Sj))}$$

Where:

Si,Sj ➔ Sessions

Pm ➔ Pages

ليكون الناتج كما هو موضح بالجدول التالي:

جدول (٣) مصفوفة علاقة التشابه بين الجلسات

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	1	0.71	0.57	0.71	0.571	0.29	0.286	0.43	0.57	0.429
S2	0.71	1	0.57	0.86	0.571	0.29	0.429	0.29	0.57	0.571
S3	0.57	0.57	1	0.43	0.167	0.67	0.5	0.33	0.33	0.333
S4	0.71	0.86	0.43	1	0.571	0.29	0.429	0.43	0.57	0.429
S5	0.57	0.57	0.17	0.57	1	0	0.2	0.4	0.8	0.6
S6	0.29	0.29	0.67	0.29	0	1	0.75	0.5	0.2	0.25
S7	0.29	0.43	0.5	0.43	0.2	0.75	1	0.75	0.4	0.5
S8	0.43	0.29	0.33	0.43	0.4	0.5	0.75	1	0.6	0.5
S9	0.57	0.57	0.33	0.57	0.8	0.2	0.4	0.6	1	0.8
S10	0.43	0.57	0.33	0.43	0.6	0.25	0.5	0.5	0.8	1

٣. اكتشاف الأنماط Pattern Discovery : لاكتشاف الأنماط يتم استخدام خوارزم العنقدة clustering algorithm وذلك لتجزئة مصفوفة الجلسات sessions التي تم إنشاؤها في مرحلة المعالجة وتحويلها إلى مكونات مترابطة (عناقيد clusters).ولتحديد المكونات المترابطة وتخصيصها لعنقود(cluster) باستخدام threshold ولذلك لتنقية الروابط الضعيفة بين الـ sessions حيث تم تنقية وفرز الروابط بين الـ sessions عند threshold =0.6 كما هو بالجدول التالي:

جدول (٤) الجلسات ذات الارتباط المرتفع

S10	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
0	1	0.71	0	0.71	0	0	0	0	0	0
0	0.71	1	0	0.86	0	0	0	0	0	0
0	0	0	1	0	0	0.67	0	0	0	0
0	0.71	0.86	0	1	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0.8	0
0	0	0	0.67	0	0	1	0	0	0	0
0	0	0	0	0	0	0.75	1	0.75	0	0
0	0	0	0	0	0	0	0.75	1	0	0
0.8	0	0	0	0	0.8	0	0	0	1	0.8
1	0	0	0	0	0	0	0	0	0.8	1

ومن خلال الجدول السابق يمكن استخلاص ثلاث مجموعات clusters حيث تشمل المجموعة الأولى (s1,s2,s4) ، المجموعة الثانية (s3,s6,s7,s8) ، المجموعة الثالثة (s5, s9,s10). كما هو موضح بالجدول التالي:

جدول (٥) المجموعات التي تم اكتشافها

Cluster1	S1	S2	S4	
Cluster2	S3	S6	S7	S8
Cluster3	S5	S9	S10	

٤. استخراج القواعد Rules Extraction وتستخدم عن طريق استخدام خوارزم لتصنيف الصفحات بناء على المجموعات Clusters التي تم الحصول عليها. فمن خلال إضافة كل مجموعة Cluster كهدف للجلسة session الذي ينتمي إليها كما هو مبين بالجدول التالي فعلى سبيل المثال ينتمي s1 إلى Cluster1 و s3 إلى Cluster2 مع تكرار الخطوات السابقة.

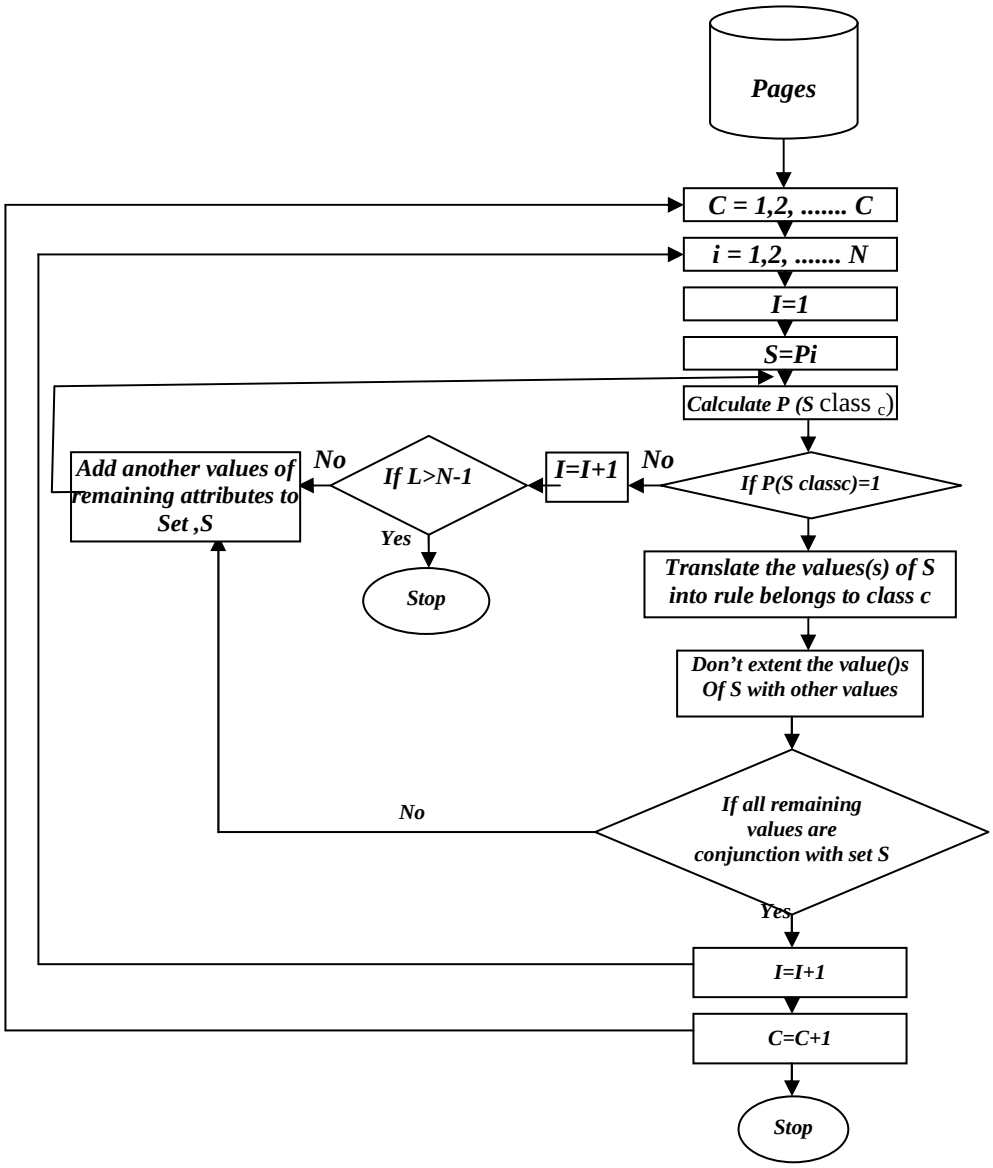
جدول (٦) الجلسات بعد إضافة المجموعات كهدف لاستخلاص القواعد

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Cluster1	Cluster2	Cluster3
S1	1	1	1	1	0	0	1	1	0	1	1	0	0
S2	1	0	1	1	1	1	1	0	0	1	1	0	0
S3	0	1	1	1	0	1	1	0	1	0	0	1	0
S4	1	0	1	1	1	1	0	1	0	1	1	0	0
S5	1	0	0	0	1	0	1	1	0	1	0	0	1
S6	0	1	1	0	0	1	0	0	1	0	0	1	0
S7	0	1	1	0	1	1	0	0	0	0	0	1	0
S8	0	1	0	0	1	0	0	1	0	0	0	1	0
S9	0	0	1	0	1	0	1	1	0	1	0	0	1
S10	0	0	1	0	1	0	0	0	0	1	0	0	1

ومن خلال تطبيق خوارزم التصنيف شكل (٢) يتم استخراج بعض القواعد التي تساعد المستخدم الجديد في تلبية متطلباته بتخصيص صفحات مرتبطة بالصفحات إلى يقوم بتصفحها وذلك لتوفير الوقت وتقليل عمليات البحث حيث يتم إضافة المجموعات clusters التي تم اكتشافها في المرحلة السابقة وهى (٣) مجموعات كهدف إلى الجدول السابق بحيث تتم قياس احتمالية وجود الصفحات كل على حدة ومجتمعة في مجموعة معينة من المجموعات (ارتباط الصفحة بالمجموعة)، فمثلا نرى أن الصفحة P1 تم الدخول عليها (٤) مرات منهم ثلاث مرات تنتمي إلى المجموعة الأولى cluster1 ومره واحده تنتمي إلى المجموعة الثالثة cluster3 وبالتالي يكون نسبة ظهورها في المجموعة الأولى ٧٥٪ ونسبة ظهورها في المجموعة الثالثة ٢٥٪ وبالتالي نحتاج إلى إضافة صفحة أخرى إلى الصفحة P1 لمعرفة احتمال ظهور الصفحتين في احد المجموعات فنقوم بتجربة P1,P2 لنجدها تظهر معا مره واحده في المجموعة الأولى cluster1 وبالتالي يكون احتمالية ظهورهما معا بنسبة ١٠٠٪ وبالتالي يمكن استخلاص القاعدة التالية:

If p1=1 and p2=1 then cluster1

ولقياس ارتباط P1, P3 معا نجد أنهم ظهرا (٣) مرات في المجموعة الأولى



شكل (٢) خوارزم التصنيف

نتائج البحث

يهدف تخصيص الويب إلى إمداد المستخدم بالبيانات التي يحتاجها بدون أن يطلبها بشكل مباشر. في هذا السياق اتضح أن تقنيات التنقيب في بيانات الويب من التقنيات المفيدة في هذا المجال لاستكشاف المعلومات المخبأة في البيانات المرتبطة بالويب وتحديدا تقنية التنقيب استخدام الويب والتي تستكشف المعلومات من بيانات دخول المستخدم للموقع باستخدام تقنيات التنقيب في البيانات .وتستخدم المعرفة المستخلصة من البيانات التاريخية للمستخدمين في تطوير نظام التخصيص أو النظام المقترح. ويقوم النظام المقترح باستخلاص البيانات من ملف دخول المستخدم log file ليتم استخدامها في اكتشاف أنماط المستخدمين عن طريق تطبيق تقنيات العنقدة أو التجميع (Clustering) حيث تم تقسيم جلسات العمل Sessionsعلى ثلاث مجموعات مترابطة وفقا لعلاقة التشابه فيما بينها حيث شملت المجموعة الأولى(S1,S2,S4) والمجموعة الثانية (S3,S6, S7,S8) والمجموعة الثالثة (S10, S9 ,S5) .وكذلك استخراج القواعد (rules) عن طريق استخدام تقنيات التصنيف classification وذلك من اجل إعطاء المستخدم مجموعه من التوصيات recommendations حيث تم حساب الارتباط بين الصفحات التي يقوم المستخدم الجديد بالدخول عليها والمجموعات الموجودة بالفعل من خلال عمليه العنقدة التي تساعده في الحصول على الصفحات الأكثر أهميه والأكثر تشابها توفيرا لوقت وجهد المستخدم وتخفيف الحمل الزائد على الموقع.

المراجع:

أولا المراجع العربية

١. محمد زياد حمدان (١٩٩٩) البحث العلمي كنظام، سوريا، دار التربية الحديثة، ص ٦.

ثانيا المراجع الأجنبية:

1. Anthony. S (2005): Web Mining Application and Techniques, United State of American, Idea Group Inc, ISBN 1-59140-9 (e-book), P.20,Available Online at http://books.google.com.eg/books, , Accessed On 11 May 2010
2. Derar.H (2009): Effectiveness of Template Detection on Noise Reduction and Websites Summarization, M.S, Department of Computer Science, University of Calgary, Alberta.
3. Zhongming. M (2007): Web Mining for Knowledge Discovery, Ph.d, Department of Bussiness Administration, University of Utah.
4. Fatih. G (2007): Effective Use of Term Relationships in Web Content Mining , Ph.d, Arizona State University, P.18.
5. John.L, Michel . V (2007): NonLinear Dimensionality Reduction, NewYork, USA, PP.45-55.

6. Richard . D, Peter. H, David. S (2001) :Pattern classification ,2nd ed, Wiley, New York, PP.35.
7. Vel'asquez, J. D., Palade V. (2008): "Adaptive Websites: A Knowledge Extraction From Web Data Approach," IOS Press, Amsterdam, NL.
8. Hannah, H., Thangavel, K.(2009): Rough set based User profiling for Web Personalization, International Journal of Recent Trends in Engineering, Vol. (2), No. (1).
9. Kamber,M . and Han, J.(2006) : Data mining Concepts and Techniques. All rights reserved by Elsevier Inc. 13:978-1-55860-901-3.
10. Yuefeng. L, Ning. Z (2004): Web Mining Model and Its Applications for Information Gathering, Journal of Knowledge-Based Systems, Vol (17), PP. 207–217.
11. David. S, Antonio. M (2009): Web Mining Techniques for Automatic Discovery of Medical Knowledge, Department of Computer Science and Mathematics Universitat Rovira, Available Online at http://www.Sciencedirect.com, Accessed On 27 march
12. Srivastava, J, et al (2005): Web Mining – Concepts, Applications and Research Directions, Journal of Computer Science, Vol (180), PP. 275–307.
13. Lan. Y, Bing. L (2006) : Web Page Cleaning for Web Mining through Feature Weighting, Available Online at http://www.Sciencedirect.com, Accessed On 27 march 2010.
14. David .J, et al (2007) : Visual Web Mining , Machine Learning Journal, Vol (42), PP.31–60.
15. Cheng. Y, et al (2009) : Applications of Web Mining for Marketing of Online Bookstores, Journal of Expert Systems with Applications, Vol(39), PP. 11249–11256.
16. Tuncay. S, Necmi. D, Zafer. C (2010): Virtual Education Environments and Web Mining, Journal of Procedia Social and Behavioral Sciences, Vol (2),PP. 5120–5124.
17. Rathod, D. (2012): A Review On Web Mining, International Journal of Engineering Research and Technology (IJERT), ISSN: 2278-0181, Vol. (1) Issue (2), April.
18. Taherizadeh, S., Moghadam N. (2009): Integrating Web Content Mining into Web Usage Mining for Finding Patterns and Predicting Users' Behaviors,

International Journal of Information Science and Management, Vol. (7), No. (1).

19. Jain,R. Purohit, G. N.(2011): Page Ranking Algorithms for Web Mining, International Journal of Computer Applications, Vol(13)– No.(5), pp.0975 – 8887.
20. Kohavi. R and Parekh. R (2003): Ten supplementary analyses to improve e-commerce web sites. In Proceeding of the Fifth WEBKDD workshop.
21. Gutschmidt, A., Cap, C. H., Nerdinger, F.W. , (2008): Paving the Path to Automatic User Task Identification. Workshop on Common Sense Knowledge and Goal-Oriented Interfaces, International Conference on Intelligent User Interfaces.
22. Dunham,H. (2003): "data mining introductory and advanced topic", pearson education INC.- Eirinaki,M. and Vazirgiannis,M.(2003)" Web mining for web personalization", http://www.soe.ucsc.edu/eirinaki/papers/Ev03_TOIT.pdf, may 2008.
23. Ling Zheng, Hui Gui and Feng Li,(2010) " Optimized Data Preprocessing Technology For Web Log Mining", IEEE International Conference On Computer Design and Applications(ICCDA), pp. VI-19-VI-21.
24. SubMasthan, T., Ravindra, Y., Satish, U., Sandeep, S., Srikanth, K. (2012): An Effective Framework For Identifying Personalized Web Recommender System By Applying Web Usage Mining, International Journal of Engineering Research and Applications (IJERA), Vol. (2), Issue (3), pp. 307-312.
25. Vel'asquez, J. D., Palade V. (2008): "Adaptive Websites: A Knowledge Extraction From Web Data Approach," IOS Press, Amsterdam, NL.
26. Kiruthika, M.(2011): Pattern Discovery Using Association Rules, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. (2), No. (12).
27. Suresh, K., Madana, R., Rama, A. (2011): Improved FCM algorithm for Clustering on Web Usage Mining, IJCSI International Journal of Computer Science, ISSN (Online): 1694-081, Vol. (8), No(1), www.IJCSI.org.
28. Usama Fayyad, et al(2009): From Data Mining to Knowledge Discovery in Databases, Morgan Kaufmann, USA.