أروى عيسى الياسري

دكتوراة معلومات ومكتبات/العراق - بغداد
استاذة جامعية في كلية المنصور الجامعة / قسم
الحاسبات ونظم المعلومات
البريد الإلكتروني :
Arwa_alyasiri2005 @yahoo.com
arwaissa@hotmail.com

عضو الهيئة الإدارية للجمعية العراقية
لتكنولوجيا المعلومات
عضو في الهيئة العامة للجمعية العراقية
للمعلومات والمكتبات
عضو الهيئة الإدارية لمركز بغداد لدراسات
حقوق الإنسان

استخراج البيانات
Data Mining (DM)
اتجاه جديد في استرجاع المعلومات

مع وجود كميات هائلة من البيانات المخزنة في قواعد البيانات data bases ومستودعات البيانات الضخمة data warehouses ازدادت الحاجة إلى تطوير أدوات تمتاز بالقوة لتحليل البيانات واستخراج المعلومات والمعارف منها، ومن هنا ظهر ما يسمى باستخراج البيانات Data Mining كتقنية تهدف إلى استنتاج المعرفة من كميات هائلة من البيانات.

ومن هنا يمكن القول أن تقنية استخراج البيانات DM ما هي إلا اتجاه جديد في استرجاع البيانات Information Retrieval (IR) وخاصة تلك المنشورة على شبكة الإنترنت.

لقد ظهرت العديد من التعريفات لهذا المفهوم نذكر منها "التنقيب في مجموعة ضخمة من مجلدات البيانات فضلا عن اكتشاف العلاقات بينها أو الاجابة على الأسئلة المتخصصة التي تكون واسعة جدا عند استخدام أدوات الاستعلام التقليدية(6) أو أنها" عملية إستكشاف وتحليل كميات كبيرة من البيانات بإستخدام أساليب آليه أو شبه إليه اعتمادا على اكتشاف نماذج وقواعد ذات مغزى (1).

يعد استخراج البيانات Data Mining مرحلة من مراحل اكتشاف المعرفة في قواعد البيانات Knowledge Discovery in Database (KDD) التي تشير إلى استخراج المفاهيم الضمنية غير الاعتيادية والتي لم تكن معروفة سابقا، وعملية اكتشاف المعرفة في قواعد البيانات Knowledge Discovery in Database (KDD) تتضمن عدد من المراحل تبدأ من جمع البيانات الخام إلى مرحلة الحصول على المعرفة الجديدة، (4) وفيما يأتي عرض لهذه المراحل:

- تنقية البيانات : Data Cleaning وهي مرحلة عزل البيانات التي تحتوي على تشويش أو شوائب Noise من مجموعة البيانات.
- توحيد البيانات : Data Integration هذه المرحلة غالبا ما تكون مصادر معالجة البيانات متغيرة العناصر وربما تكون مجتمعة في مصدر شائع.
- اختيار البيانات : Data Selection في هذه المرحلة، يتم تحديد وإسترجاع البيانات الملائمة من مجموعة البيانات .
- نقل البيانات : Data Transformation وهي عملية نقل البيانات التي تم إختيارها الى شكل ملائم لاجراءات البحث والاسترجاع.
- استخراج البيانات : Data Mining في هذه المرحلة سيتم تطبيق أسلوب ذكي لاستخراج نماذج مفيدة قدر الإمكان.
- تقييم النموذج : Pattern Evaluation بعد استخراج النماذج المهمة والتي تمثل المعرفة يتم تقييمها استنادا الى مقاييس محددة.
- تمثيل المعرفة : Knowledge Representation وهي المرحلة الاخيرة من مراحل اكتشاف المعرفة في قواعد البيانات وهي المرحلة التي يراها المستفيد ، هذه المرحلة الأساسية تستخدم الاسلوب المرئي لمساعدة المستفيد في فهم و وتفسير نتائج استخراج البيانات.

ويمكن ان تنجز مرحلتين في ان واحد وعلى سبيل المثال يمكن انجاز كل من مرحلة تنقية البيانات ومرحلة توحيد البيانات مع بعضها ويمكن ان تشترك مرحلة اختيار البيانات مع مرحلة نقل البيانات.

يتضمن استخراج البيانات DM عدد من الأساليب الرئيسية التي يمكن من خلال استخدامها الوصول الى الهدف من استخدام هذا الإتجاه وهي:

1- قاعدة الارتباط Association Rule:

قواعد الإرتباط Associations Rule هي أحد الواجهات الواعدة من Data Mining كأداة من أدوات اكتشاف المعرفة KDD ولديها القدرة على تصفح كميات هائلة من البيانات، وهي تسمح بالتقاط كل القوانين الممكنة التي تشرح بعض الصفات الموجودة اعتمادا على وجود الصفات الأخرى[2].

وبمعنى آخر هي قواعد ارتباطية معينة بين مجموعة من البيانات في قاعدة البيانات وتتضمن إيجاد large Item set من خلال المعادلة التالية:

X ⟶ Y تتضمن إيجاد درجة الوثوقية لهذا الإرتباط.

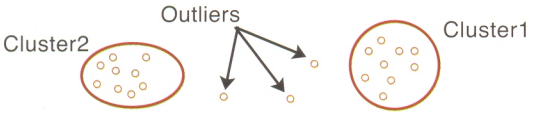
$$\text{Support} = \frac{\text{The Number of Transactionhat Conatains X and Y}}{\text{The Total Number of Transaction}}$$

2- العنقدة clustering:

وهي عملية تقسيم البيانات إلى مجموعة من الأصناف اعتمادا على اشتراكها بالخواص المتشابهة وأن العنقدة هي تقسيم غير موجه للبيانات [3]. وهي عكس التصنيف الذي سيرد لاحقا، كما أنها تساعد المستفيد على فهم التركيب الطبيعي للمجموعات من البيانات.

Unsupervised Classification = Clustering

هنا لانعرف الأصناف ولاعددها.



شكل (1) يبين عملية العنقدة

3- التصنيف Classification:

يستخدم التصنيف بشكل واسع في حل الكثير من المشكلات خاصة تلك التي تتعلق بالأعمال Business من خلال تحليل مجموعة من البيانات ووضعها على شكل أصناف أو أقسام يمكن استخدامها فيما بعد لتصنيف البيانات المستقبلية،(5) وهنا يكمن الفرق بين التصنيف والعنقدة. وهناك عدد من الطرق التي يمكن استخدامها في تصنيف البيانات باستخدام الخوارزميات مثل الخوارزميات الإحصائية Statistical Alg. وخوارزميات الشبكات العصبية Neural Network Alg. وخوارزميات الوراثة Genetic Alg. وطريقة الجار الأقرب Nearest neighbor method.

4- التحليل التسلسلي Sequential analysis:

في هذه الطريقة يتم البحث لاكتشاف نماذج تحدث بالتسلسل إذ تكون المدخلات عبارة عن بيانات تشكل مجموعة متسلسلة وكل سلسلة من البيانات هي قائمة منظمة من العمليات أو المصطلحات وعندما تكون العملية عبارة عن مجموعات من المصطلحات لابد أن يحسب معها الوقت المصاحب لكل عملية .(5) ولكن مشكلة هذا النموذج تكمن في إيجاد كل النماذج المتسلسلة مع أقل دعم يخصصه المستفيد عندما يكون الدعم لهذا النموذج هو نسبة تسلسل البيانات التي يتضمنها النموذج.

نماذج تطبيقية في إستخدام استخراج البيانات (DM) في علم المعلومات.

نظرا للمزايا التي يمتلكها هذا الاتجاه تم تنفيذ بعض من أساليبه على سبيل التجارب التطبيقية في موضوعات علم المعلومات ومنها تجربة بناء مكنز آلي باستخدام اسلوبي قاعدة الارتباط Association Rule والعنقدة clustering (7) إذ وبواسطة الأسلوب الأول تم تحليل مستخلصات بحوث علمية وتحديد مجموعة المصطلح الكبير large Item set وبواسطة الأسلوب الثاني تم تجميع المصطلحات في عناقيد ومنها تم الوصول إلى المصطلحات العريضة والمصطلحات الضيقة والمصطلحات المترابطة وبالنتيجة تم الحصول على مكنز آلي بإستخدام إساليب إستخراج البيانات DM أما التجربة الثانية فكانت تدور حول إستخدام أسلوب وخوارزميات التصنيف المستخدمة في إستخراج البيانات DM لغرض إيجاد طريقة جديدة في التصنيف الآلي لمصادر المعلومات في المكتبات(8). في هذه التجربة تم تحليل مجموعة من البحوث في موضوع تكنولوجيا المعلومات وباستخدام خوارزمية Classification - rule learning تم الحصول على أصناف محددة سيتم إتباعها في المستقبل عندما تصل المكتبة مصادر معلومات في هذا الموضوع.

وما زال موضوع تطبيق إستخراج البيانات DM في مجال علم المعلومات وتحديدا إسترجاع المعلومات أرضا بكرا بحاجة إلى المزيد من البحث وإجراء التجارب لغرض الحصول على العديد من الموضوعات والأفكار الجديدة التي من شأنها الإرتقاء بمستوى خدمات المعلومات.

المصادر

1- Adriaan & P. and D. Zanting. Data Mining . Addison-Wesley Harlow, England, 1996.

2- Al-Hamami ,Alaa H., abass F Kader ,Hussein K.Al-khefaji,"Desgin and Implementation of Genenrate of large Dense, or sparce Database to test Association rules Miners" (selected reachers papers), Scientific journal of Fedration of Arab Scintific Research Council, 2002 .

3- Botta, Marco "Clustering Techniques ",Dipartimento di Informatica Universitàdi Torino,www.di.unito.it/~botta/didattica/clustering.html,2003.

4- Fayyad, U., G. Piatetsky-Shapiro,P. Smyth, & R. Uthurusamy, Advance in Knowledge Discovery & Data Mining. Cambridge, MA (The AAAI Press/The MIT Press), 1996.

5- Joshi, Karuna Pande . Analysis of Data Mining Algorithms .available at : http://www.

6- Michael, J., A. Berry and Gordan S. Linoff, Mastering Data Mining. John Wiley & Sons, Inc, 2000.

7- الياسري، أروى عيسى ،هديل شوكت العبيدى. تجربة تصميم مكنز آلي بإستخدام أساليب إستخراج البيانات Data Mining _ بحث غير منشور ألقي في مؤتمر بلدية دبي الدولي الثالث للتوثيق والأرشفة الإلكترونية أيلول 2005.

8- الياسري، أروى عيسى، هديل شوكت العبيدي. التصنيف الآلي لمصادر المكتبة بإستخدام تقنيات التصنيف Classification Techniques في إستخراج البيانات Data Mining بحث غير منشور.