أدوات التنقيب عن البيانات مفتوحة المصدر دراسة تحليلية تقييمية

د. أحمد فايز أحمد سيد

مدرس تكنولوجيا المعلومات قسم المكتبات والمعلومات

كلية الآداب والعلوم الإنسانية- جامعة قناة السويس- مصر

Afayez2003@yahoo.com

تاريخ التحكيم: ٩/ ١٠/ ١٤٣٦ تاريخ الإجازة: ٥/ ٢/ ١٤٣٧

المستخلص:

لقد ظهر مع التنقيب عن البيانات أدوات وبرمجيات تساعد في التنقيب عن الكم الهائل والمتزايد من البيانات للوصول إلى المعرفة في قواعد البيانات المختلفة، وتيسر هذه الأدوات العمل على معظم التخصصات العلمية ومنها علوم المكتبات والمعلومات. لذا تهدف هذه الدراسة إلى دراسة ماهية التنقيب عن البيانات ومهامه وتطبيقاته، وتحليل أدوات التنقيب عن البيانات مفتوحة المصدر وتقييمها، ومن ثم عقد مقارنة بين أدوات التنقيب عن البيانات مفتوحة المصدر. وتوصلت الدراسة للعديد من النتائج أهمها: هناك مزية يتصف بها بعض الأدوات والتي تتضح من خلال الاستخدام وهي توفير نموذج السحب والإفلات أثناء عملية

التركيب والبناء للتنقيب عن البيانات وهي تتوافر بأربع أدوات KNIME، Orange، Weka، RapidMiner.

الكلمات المفتاحية:

التنقيب عن البيانات، التنقيب عن الويب، التنقيب عن المعلومات، التنقيب الببليوجرافي.

مقدمة الدراسة

١/٠ تمهيد

لقد بدأ الاهتمام بالتنقيب عن البيانات عام ١٩٨٩م أثناء انعقاد ورشة عمل حول اكتشاف المعرفة في قواعد البيانات،(١) ومن ذلك الحين تم عقد هذه الورشة بصفة مستمرة سنويا حتى عام ١٩٩٤م، أما في ١٩٩٥م أصبح المؤتمر الدولي لاكتشاف المعرفة والتنقيب عن البيانات من أهم الأحداث السنوية، ومن ثم بدأ تخطيط الإطار العملي للتنقيب عن البيانات واكتشاف المعرفة في كتابين: اكتشاف المعرفة في قواعد البيانات، والتقدم في اكتشاف المعرفة والتنقيب عن البيانات ثم فاقت إمكانية تخزين كميات هائلة من البيانات قدرة العنصر البشري على التحليل والفهم بعد عام ٢٠٠٠م؛ ولم يكن هناك أداة مناسبة لاشتقاق المعلومات والمعرفة من البيانات، ويمكن إيجاد نماذج محددة وقواعد بواسطة أدوات التنقيب عن البيانات في ظل كم هائل من البيانات، والذي يوفر المعلومات الضرورية للأنشطة التجارية، والاكتشافات العلمية، والبحث الطبي وغيرها من المجالات.

لذا يعد التنقيب عن البيانات من أسرع المجالات نموا في تخصصات علم

(١) Piatesky-Shapiro,G. (Jan. 1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop, *AI Magazine*, 11: 5, pp. 68-70. Available at: https://www.google.com.eg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CCEQFjAAahUKEwiWh9nL-djGAhVMPhQKHQUjCDk&url=https%3A%2F%2Fwww.aaai.org%2Fojs%2Findex.php%2Faimagazine%2Farticle%2Fdownload%2F873%2F791&ei=AyCkVdb5Acz8UIXGoMgD&usg=AFQjCNFB_-Qrs8RdlFxNlNI9jhuh61eiXw&bvm=bv.97653015,d.ZGU

الحاسب الآلى، ولقد جاءت شهرته وانتشاره من الحاجة المتزايدة لأدوات تساعد في تحليل الكميات الهائلة من البيانات وفهمها، وتنتج هذه البيانات يوميا بواسطة المؤسسات المختلفة مثل البنوك، شركات التأمين، ومستودعات البيع وعلى شبكة الإنترنت، وصاحب هذا الانفجار في البيانات زيادة هائلة أيضا في استخدام الحاسبات الآلية، والماسحات الضوئية، والكاميرات الرقمية، والباركود وغيرها.

ولقد ظهر مع التنقيب عن البيانات أدوات وبرمجيات تساعد في التنقيب عن الكم الهائل والمتزايد من البيانات للوصول إلى المعرفة في قواعد البيانات المختلفة، وتيسر هذه الأدوات العمل على معظم التخصصات العلمية ومنها علوم المكتبات والمعلومات.

٢/٠ مشكلة الدراسة

مع التطور الهائل في البيانات والمعلومات المحدثة باستمرار سواء التى يمكن الوصول إليها بصفة مباشرة عن طريق البحث البسيط أو المتقدم بمحركات البحث المختلفة، إلا أنه توجد قواعد بيانات وقواعد محركات لا يمكن الوصول لبياناتها إلا بالتنقيب، وليست عملية التنقيب سهلة بما كان ليستطيع أي مستفيد أو باحث القيام بها يدويا أو إلكترونيا بدون أدوات مساعدة لذلك، ومن ثم سعت الشركات المختلفة والباحثين في مجال التنقيب عن البيانات لإنتاج أدوات تساعد في التنقيب عن البيانات بقواعد البيانات المختلفة، منها ما هو متاح مجانا أو بمقابل، لذا تحاول هذه الدراسة تحليل أهم أدوات التنقيب عن البيانات مفتوحة المصدر وتقييمها لتقرير مزاياها وعيوبها والوظائف التي يمكن القيام بها.

٣/٠ أهمية الدراسة ومبرراتها

تتمثل أهمية الدراسة ومبرراتها فيما يلي:

١) قلة الدراسات العربية المعنية بالتنقيب عن البيانات، وخاصة الأدوات مفتوحة المصدر.

٢) دراسة ماهية التنقيب عن البيانات ومهامه وتطوره.

٣) تحليل أدوات التنقيب عن البيانات مفتوحة المصدر وتقييمها.

٤/٠ أهداف الدراسة

تعمل الدراسة على تحقيق الأهداف التالية:

١) دراسة ماهية التنقيب عن البيانات ومهامه وتطبيقاته.

٢) تحليل أدوات التنقيب عن البيانات مفتوحة المصدر وتقييمها.

٣) عقد مقارنة بين أدوات التنقيب عن البيانات مفتوحة المصدر.

٥/٠ تساؤلات الدراسة

تهدف هذه الدراسة إلى الإجابة على التساؤلات التالية:

١) ما المفهوم العلمي للتنقيب عن البيانات؟

٢) إلى متى تعود بدايات التنقيب عن البيانات؟

٣) كم عدد المهام التى تقوم بها التنقيب عن البيانات؟

٤) إلى أي مدى يمكن تطبيق التنقيب عن البيانات؟

٥) ما الفروق العامة بين أدوات التنقيب عن البيانات مفتوحة المصدر؟

٦) بمَ تختلف أدوات التنقيب عن البيانات مفتوحة المصدر من الناحية الوظيفية؟

٧) أي من أدوات التنقيب عن البيانات مفتوحة المصدر سهلة الاستخدام

والتطوير؟

٦/٠ حدود الدراسة

١/٦/٠ الحدود الموضوعية: تركز الدراسة على أدوات التنقيب عن البيانات مفتوحة المصدر.

٢/٦/٠ الحدود النوعية: تهدف الدراسة إلى تحليل أدوات التنقيب عن البيانات مفتوحة المصدر والمتاحة على شبكة الإنترنت.

٣/٦/٠ الحدود اللغوية: تركز الدراسة على الإنتاج الفكري باللغة الإنجليزية والعربية.

٧/٠ مجتمع البحث والعينة

تم اختيار عينة قصدية من أدوات التنقيب عن البيانات مفتوحة المصدر، وهم: Knime، Rapidminer، Weka، Orange، Tangra، Rattle، والتى يتوافر بها الشروط التالية:

١) توافر موقع لها مع إمكانية التحميل من الموقع.

٢) تحديث المواقع باستمرار ومن ثم تحديث الأدوات ذاتها.

٣) أن تقوم بكل وظائف التنقيب عن البيانات وليس وظيفة محددة.

٤) أن تكون متاحة مجانا ولكل المستفيدين دون قيود.

٨/٠ منهج الدراسة، وأدواتها

١/٨/٠ مناهج الدراسة:

تقوم هذه الدراسة على ثلاثة مناهج:

١) المنهج التاريخي: لدراسة التنقيب عن البيانات وبداياته وتطوره وتطبيقاته.

٢) المنهج الوصفي التحليلي لدراسة أدوات التنقيب عن البيانات الست مفتوحة المصدر، ومن ثم تحليلها ومقارنتها.

٣) المنهج المقارن: للمقارنة بين أدوات التنقيب عن البيانات الست مفتوحة المصدر، والخروج بنتائج توضح أي من الأدوات يمكن استخدامها مع المستفيدين المبتدئين أو المتقدمين، وأي التطبيقات تصلح مع هذه الأدوات.

٠/ ٨/ ٢ أدوات جمع البيانات:

اعتمدت الدراسة على ثلاث أدوات:

١) أداة البحث الوثائقى ومصادر المعلومات الرقمية سواء كانت قواعد بيانات أو دوريات إلكترونية أو كتبا إلكترونية على شبكة الإنترنت، حيث تم استقراء أدبيات الإنتاج الفكري العالمي حول التنقيب عن البيانات ومفاهيمه وبداياته وتطوره ومهامه وتطبيقاته.

٢) الإبحار التفاعلى للويب لدراسة مواقع أدوات التنقيب عن البيانات مفتوحة المصدر وتحميلها واستخدامها للتعرف على خصائصها والمقارنة فيما بينهم.

٠/٩ الدراسات السابقة والمثيلة

هناك المئات من الدراسات حول التنقيب عن البيانات، لكنها تتركز في الأغلب على الأعمال التجارية والاحصائيات، وهناك القليل من الدراسات التى تتناول تقييم برمجيات وأدوات التنقيب عن البيانات وكيفية اختيارها، إلا أنها دراسات عامة تصلح على أي البرمجيات وليست التنقيب عن البيانات فقط، وتركز

هذه الدراسات على البرمجيات التجارية نظرا لكثرة الخدمات التى تقدمها وصعوبتها في البناء وتعاملها مع الكميات الهائلة من البيانات، ومن ثم تم حصر بعض الدراسات التى تتناول تقييم برمجيات وأدوات التنقيب عن البيانات وهو موضوع الدراسة الحالية:

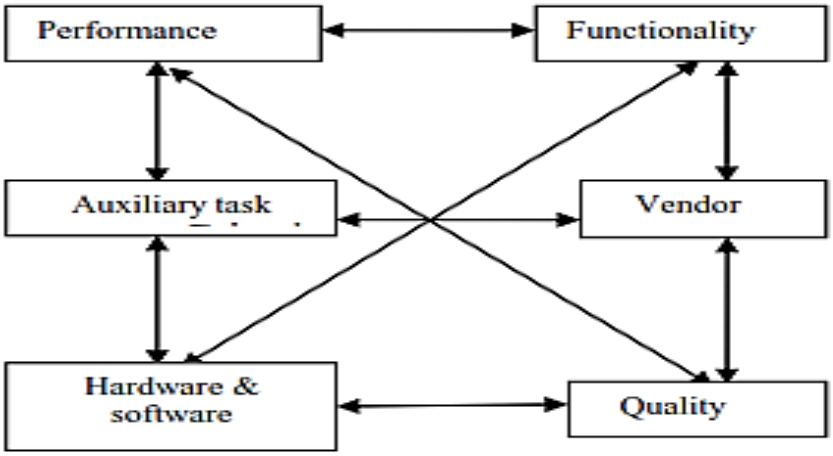
١) Lyras,D., Panagiotakopoulos,T., Kotinas,I., Panagiotakopoulos, C., Sgarbas,K. and Lymberopoulos,D. (Jun. 2014). Educational Software Evaluation: A Study From An Educational Data Mining Perspective. The International Journal of Multimedia & Its Applications (IJMA), 6 (3). Available at: http://airccse.org/journal/jma/6314ijma01.pdf

تهدف هذه الدراسة إلى مراجعة وتنقيح واختبار البرامج التعليمية من عدة اتجاهات. وتم استخدام وسائل تقنيات التنقيب عن البيانات التعليمية في الدراسة الحالية الخاصة (١٧٧) من أشهر معايير التقييم التى اقترحها العديد من الباحثين وتم اختبارها وتقييمها مع مراعاة درجة التأثير على كفاءة البرنامج التعليمى. وبواسطة تقنيات التنقيب وخاصة التنبؤ واختيار المزايا تم التحري عن العلاقة الخفية في البيانات المجمعة من التجارب التى تمت بقسم التعليم في جامعة باتراس والتي تتعلق بمهمة تقييم البرامج ومن ثم تم تقديم نتائج هذه الدراسة ومناقشتها بطريقة كمية وكيفية.

٢) Bhargava,N., Aziz,A. and Arya,R. (2013) Selection Criteria for Data Mining Software: A Study. IJCSI International Journal of Computer Science Issues, 10 (3): 308-312. Available at:

http://ijcsi.org/papers/IJCSI-10-3-2-308-312.pdf

تتزايد أعداد برامج التنقيب عن البيانات في الأسواق بشكل كبير، لذا فهناك حاجة لاختيار معيار لحزمة برمجيات التى يمكن إتاحتها للمستفيدين الأراد والمنظمات. ومع استمرار الزيادة في أعداد البرامج والمزايا الإضافية التى تتضمنها البرامج الأحدث، يصبح من الأصعب اختيار حزمة البرامج المناسبة، فقد يتم التوصل لقرار خاطئ مع فقد الكثير من الوقت والمال، لذا تم إجراء العديد من الدراسات حول العالم لتقييم البرامج؛ إلا أن الباحثين لم يصلوا للمستفيدين لتعميم معايير الاختيار والتقييم. إن الاختيار غير اللائق لحزمة البرامج قد يؤدي أن يكون مكلفا للغاية ويؤثر سلبا على العمل. وتتمثل المعايير التي تم اتباعها في الدراسة بالشكل التالي:



شكل رقم (١) معايير اختيار برامج التنقيب عن البيانات

(٣ Androni,M. and Crisan,D. (2010). Commercially Available Data Mining Tools Used in Economic Environment. Database System Journal, 1(2). Available at: http://www.dbjournal.ro/archive/2/6_Andronie_Crisan.pdf

تقدم هذه الدراسة بعض من أدوات التنقيب عن البيانات التجارية المتاحة، مع أهم ملامحها، جنبا إلى جنب مع بعض الاعتبارات التى تتعلق بتقييم أدوات التنقيب عن البيانات بواسطة الشركات التى ترغب في الحصول على مثل هذه النظم. ومن أهم العوامل التى يجب على الشركات مراعاتها كميات البيانات المتاحة، وكيف يمكن تخزينها، ومهام التنقيب عن البيانات التى يجب تنفيذها، هذا ويجب التنويه إلى أن تكلفة نظام التنقيب عن البيانات مهم للشركة، والتى لها تأثير محدود على توسعة سوق منتجات التنقيب عن البيانات بالنسبة للشركات الصغيرة.

٤) Qiu,M., Davis,S. and Ikem,F. (2004). Evaluation Of Clustering Techniques In Data Mining Tools. Issues in information systems, 7 (1): 254-260. Available at: http://iacis.org/iis/2004/QiuDavisIkem.pdf

تقسم المجموعات الكثافة المتماثلة إلى عدد من المجموعات الفرعية المتماثلة أو مجموعات تعكس قطاعات مجموعة البيانات مثل النماذج. ويبين هذا البحث كيف يمكن لاطار تقييم البرامج أن يتناسب وتقييم أدوات التنقيب عن البيانات التجارية لبيئة محددة من المستفيدين. وتطبق هذه الدراسة تقييم أداتين رئيستين من الأدوات التجارية للتنقيب عن البيانات وهما: SAS (EM) Enterprise Miner وIM) IBM DB2 Intelligent Miner) لاستخدامهما في البيئة الجامعية. ولقد استخدم (٤) أربعة معايير لتقييم تقنيات المجموعات في أدوات التنقيب عن البيانات:

١) الأداء والمقصود به القدرة على معالجة مصادر بيانات مختلفة بطريقة كفؤة، ويتمثل معيارها في تركيب البرنامج، والوصول إلى البيانات المتجانس.

٢) الوظيفة: القدرة على تضمين مجموعة من الإمكانات والتقنيات والطرق

للتنقيب عن البيانات، ومعاييرها: تنوع الخوارزميات، منهجية تم وصفها مسبقا.

٣) الاستخدام: تتناسب مع مستويات وأنواع مختلفة من المستفيدين دون فقد أي شيء من الوظيفة أو عدم الفائدة، ومعاييرها: أنواع المستفيدين، وعرض البيانات.

٤) دعم المهام المساعد: يتيح للمستفيد القيام بتنظيف البيانات وتنميتها وتحويلها وعرضها والكثير من المهام الأخرى التى تدعم التنقيب عن البيانات، ومعاييرها: فلترة البيانات، واشتقاق الخصائص.

٥) Collier,K., Carey,B., Sautter,D., Marjaniemi,C. (1999). A Methodology for Evaluating and Selecting Data Mining Software. Proceedings of the 32nd Hawaii International Conference on System Sciences. Available at: http://www.computer.org/csdl/proceedings/hicss/1999/0001/06/00016009.pdf

لقد تطور التنقيب عن البيانات ولازال يتطور ودخل في العديد من الممارسات التجارية، إلا أن برامج التنقيب عن البيانات الحالية وبرامج دعم القرار مرتفعة السعر واختيار الأداة الخطأ يؤدي إلى تكاليف باهظة بعدة طرق، لذا تحاول هذه الورقة البحثية إلى تقديم اتجاه ومعلومات صنع القرار حول الممارسة المهنية، وذلك من خلال تقديم إطار عمل لتقييم أدوات التنقيب عن البيانات ومنهجية تصف هذا الاطار، وتعرض الورقة البحثية في النهاية دراسة حالة لعرض كفاءة الطريقة، حيث تمثل هذه المنجية خبرة أولية باستخدام العديد من أدوات التنقيب عن البيانات الرائدة مقارنة بالبيانات التجارية في مركز البيانات الداخلية بجامعة شمال أريزونا Insight (CDI) at Northern Arizona University (NAU) Center for Data. والجدير بالذكر أن هذه الورقة البحثية ليست

مراجعة شاملة للأدوات التجارية وإنما تقدم طريقة ومرجع لاختيار أفضل أداة برمجية لمشكلة محددة، ولقد أوضحت الخبرة أنه ليست هناك أفضل أداة للتنقيب عن البيانات لكل الأغراض. ولقد تم تصميم هذه الأداة لتتلاءم مع الاختلاف في البيئات ونطاقات المشكلة، ومن المتوقع أن هذه المنهجية سيتم استخدامها لنشر مقارنات الأداة والنتائج المعيارية.

ويتبين من خلال الدراسات السابقة مدى اختلاف الدراسة الحالية عنها والتى يمكن توضيحها في: تركز الدراسة الحالية على توضيح بدايات التنقيب عن البيانات وتطوره وعلاقته بعلوم المكتبات والمعلومات، وتحليل أدوات التنقيب عن البيانات مفتوحة المصدر وتقييمها، ولقد تم اختيار أنسب المعايير من بين عدة دراسات حتى تتناسب والدراسة الحالية، كما تم الخروج بنتائج عدة يمكن من خلالها تحديد أي من الأدوات التى يمكن استخدامها للمبتدئين أو المتقدمين تكنولوجيا، وأي المجالات التى يمكن استخدامها.

المبحث الأول
التنقيب عن البيانات
مفهومه، بداياته وتطوره، مهامه، أنواعه، تطبيقاته

١/٠ تمهيد

أدى التطور في العلم والاقتصاد وتكنولوجيا المعلومات والاتصالات إلى زيادة كمية البيانات الرقمية، ومع هذه الكميات الهائلة من البيانات لم تعد وسائل التحليل التقليدية (الإحصائية مثلاً) قادرة على التعامل معها. لذا ظهرت العديد من الدراسات منذ أواخر الثمانيات في محاولة لحل تلك المشكلات مع البحث عن حلول تجمع بين عدة تخصصات سواء كانت الإحصاءات أو قواعد البيانات

والذكاء الصناعى وتمييز النماذج المختلفة والحوسبة التناظرية،[١] ومن ثم تم التوصل إلى التنقيب عن البيانات Data mining واكتشاف المعرفة اللذين أثبتا وجودها كأحد الحلول الناجحة لتحليل كميات ضخمة من البيانات وذلك بتحويلها من بيانات متراكمة وغير مفهومة إلى معلومات قيمة يمكن استغلالها والاستفادة منها بعد ذلك لتصبح معرفة.

لذا يعد التنقيب عن البيانات من أسرع المجالات نموا في تخصصات علم الحاسب الآلي، ولقد جاءت شهرته وانتشاره من الحاجة المتزايدة لأدوات تساعد في تحليل الكميات الهائلة من البيانات وفهمها، وتنتج هذه البيانات يوميا بواسطة المؤسسات المختلفة مثل البنوك، شركات التأمين، ومستودعات البيع وعلى شبكة الإنترنت، وصاحب هذا الانفجار في البيانات زيادة هائلة أيضا في استخدام الحاسبات الآلية، والماسحات الضوئية، والكاميرات الرقمية، والباركود وغيرها.

١/١ مفهوم التنقيب عن البيانات

يعد التنقيب عن البيانات عملية متطورة تقوم باشتقاق البيانات المطلوبة والفعالة والشاملة من كم هائل من البيانات طبقا لأهداف مسبقة تجارية،[٢] ويعد البعض التنقيب عن البيانات مصطلحا شائعا لاكتشاف المعرفة، في حين يضع البعض التنقيب عن البيانات كخطوات أساسية في عملية اكتشاف المعرفة. فلقد ظهر التنقيب عن البيانات في أواخر الثمانينات والذي كان مجالا حديثا مع قيمة بحثية

(1) Zhao,Y., Chen,Y. and Yao,Y. (2006). User-Centered Interactive data Mining. Proceedings of the Sixth IEEE International Conference on Cognitive Informatics (ICCI'06): 457-466. Available at: http://www2.cs.uregina.ca/~yanzhao/icci06.pdf

(2) Durkin,J. and Jingfeng,C. (2005) CAI Zixing. Decision Tree Technology And Its Current Research Direction[J].Control Engineering.12 (1).

مرتفعة في دراسة قواعد البيانات والذكاء الصناعى وتكنولوجيا قواعد البيانات وتعلم الآلة والإحصائيات وعرض البيانات وغيرها من المجالات النظرية والتكنولوجية.(١)

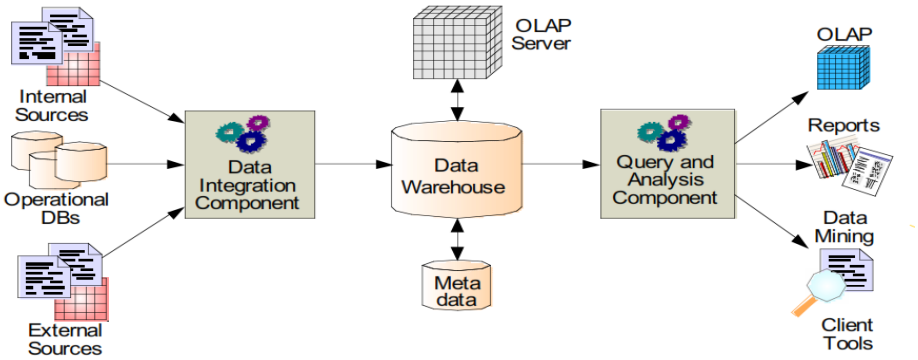
في الحقيقة هناك خطا في التسمية حيث إن المقصود بالتنقيب هو استخلاص المعرفة من كميات كبيرة من البيانات؛ حيث يمكن أن تكون تسميتها المناسبة هى تنقيب المعرفة من البيانات ولكن هذا المصطلح طويل بعض الشيء فيمكن تسميتها بمصطلح أقصر التنقيب عن المعرفة؛ لكن هذا المصطلح لا يعكس التوكيد على أن التنقيب من كميات كبيرة من البيانات.لهذا وجد أن مصطلح التنقيب عن البيانات (Data mining) المصطلح المناسب لهذا العلم.

فيمكن القول إن التنقيب عن البيانات هى اكتشاف المعرفة من البيانات أو هى التنقيب عن البيانات (أحيانا تسمى اكتشاف المعرفة) هى عملية تحليل البيانات من منظورات مختلفة واستخلاص علاقات بينها وتلخيصها إلى معلومات مفيدة، مثل معلومات يمكن أن تسهم في زيادة الربح، تخفيض التكاليف، أو كليهما معا. أو هو عملية الكشف والعثور على معلومات ذات فائدة من خلال استعمال مجموعة من الأدوات المعقدة. بعض من هذه الأدوات تشمل أدوات الإحصاء الاعتيادية والذكاء الاصطناعى والرسوم البيانية من صنع الكمبيوتر.

تبدو دورة حياة التنقيب عن البيانات كنوع من التكنولوجيا غير واضحة، حيث كان على الخبراء استغراق وقت طويل وجهد للبحث والتطوير للوصول إلى مرحلة النضج والقبول، فهو نوع من التكنولوجيا التى تجمع بين طرق معالجة البيانات التقليدية بخوارزميات مختلفة لتحليل أنواع البيانات الجديدة واشتقاق

(1) Han, J. and Kamber,M. Data Mining: Concepts and Techniques. Available at: https://cs.wmich.edu/~yang/teach/cs595/han/ch01.pdf

المعرفة من كميات هائلة من البيانات، وهناك نوعان من المعرفة في ظل هذا الكم الهائل من البيانات: الأول معالجة تحليلية على الخط المباشر On-Line Analytical Processing (OLAP) والثاني التنقيب عن البيانات Data Mining (DM)، وكلا النوعين أدوات تحليلية تعتمد على مستودعات البيانات، لكن ظهرت المعالجة التحليلية للبيانات على الخط المباشر قبل التنقيب عن البيانات، والتى تعتمد على عرض متعدد الأبعاد مؤكدا على كفاءة وسرعة تلبية رغبات المستفيدين؛ ويشير التنقيب عن البيانات للنموذج المفيد وغير المرئى للمستفيدين في أعماق البيانات ويتم بطريقة آلية دون مشاركة المستفيدين. ويوضح الشكل التالى الفرق بين المعالجة التحليلية على الخط المباشر والتنقيب عن البيانات بمستودعات البيانات.(١)



شكل رقم (٢) المعالجة التحليلية على الخط المباشر والتنقيب عن البيانات بمستودعات البيانات

يمكن للتنقيب عن البيانات اختبار أي نوع من البيانات وتدفق المعلومات، وتكمن صعوبته في نوع قاعدة البيانات:

(1) Eltabakh,M. (2010). OLAP & Data Mining. Available at: http://web.cs.wpi.edu/~cs561/s12/Lectures/IntegrationOLAP/OLAPandMining.pdf

١. التنقيب عن قواعد البيانات العلائقية: وهي مجموعة من الجداول، ويتكون كل جدول من مجموعة من العلاقات، مكونة عدد من الجداول الضخمة، وعادة ما تستخدم نموذجا يمثل العلاقة بين قاعدة البيانات والواقع. ويمكن الحصول من تنقيب قواعد البيانات العلائقية على الاتجاهات ونموذج البيانات، مثل: دخل المستفيد، عمره، مستوى تعليمه، ومعلومات أخرى يمكن الحصول عليها، وبواسطة قاعدة البيانات العلائقية التجارية يمكن تحديد التسويق المستهدف للمستفيدين وتجنب الاحتيال وتشكيل استراتيجية للشركة.(١)

٢. تنقيب مستودعات البيانات: المستودع عبارة عن مجموعة من البيانات الموجهة والمتكاملة والثابتة ومناسبة، تستخدم في دعم قرار الشركات، كما يمكن استخدامه كمصدر متكامل واحد للبيانات لمعالجة المعلومات، فهو يودع البيانات المجمعة التي تم معالجتها لإيجاد النماذج الخفية والعلاقة لتكوين نموذج تحليلي هيكلي لتصنيف البيانات ووضع التوقعات المحتملة.(٢)

٣. التنقيب عن قاعدة بيانات جديدة: تتضمن قاعدة البيانات الجديدة قاعدة بيانات مكانية، وقاعدة بيانات وقتية، وقاعدة بينات نصية، وقاعدة بيانات وسائط

(1) Netz, A., Chaudhuri,S. Bernhardt, J. and Fayyad, U. (2000) Integration of Data Mining and Relational Databases Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt. Available at: http://121.241.184.234/trddc_website/PastTecsweeks/2007/integration-of-data-mining.pdf

(2) Wang,J. (2008). Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications. Information Science reference. Available at: http://www-db.deis.unibo.it/~srizzi/PDF/isr08-1.pdf

متعددة، وتتضمن هذه البيانات بيانات مكانية، نصية، صورا، وصوتا، وبيانات ويب، وبنية البيانات عادة ما تكون أكثر تعقيدا وتغيرا ديناميكيا، ومن الصعب معالجتها، فعلى سبيل المثال: يمكن إيجاد تطور خصائص العنصر أو الشيء واتجاهاته، كما يتم تجميع البيانات المتدفقة ومقارنتها لإيجاد النماذج المهمة.(١)

ومن خلال ما سبق يمكن القول إن التنقيب عن البيانات يعتمد على أربعة عناصر أساسية وهي:

١. البيانات: هي عبارة عن الحقائق والأرقام والنصوص التي يمكن أن تعالج من قبل الحاسب.

٢. المعلومات: النماذج والعلاقات بين تلك البيانات التي تشكل معلومات مفيدة.

٣. المعرفة: المعلومات السابقة يمكن أن تحول إلى معرفة حول الأنماط التاريخية أو التوقعات المستقبلية، مثال معلومات عن حركة المبيعات والمشتريات للزبائن يمكن أن تزودنا بمعرفة عن سلوكهم الشرائي، فيساعدنا ذلك في معرفة أي من المواد تحتاج إلى ترويج أكثر.

٤. مستودعات البيانات: المستخدمة في التحليلات الزمنية واكتشاف المعرفة واتخاذ القرارات، فهي مصممة خصيصا لاستخلاص البيانات ومعالجتها وتمثيلها وتقديمها بصورة مناسبة لهذه الأغراض، وتخزن كمية ضخمة من البيانات قد تكون من مصادر مختلفة، مثلا عدة قواعد بيانات من عدة نماذج.

---

(1) Agrawal,R., Imielinski,T. and Swami,A. Database Mining: A Performance Perspective. Available at: http://www.rakesh.agrawal-family.com/papers/tkde93mining.pdf

٢/١ بدايات التنقيب عن البيانات وتطوره

لقد تطورت تكنولوجيا قاعدة البيانات وتكنولوجيا المعلومات تدريجيا في الستينيات من نظام معالجة الوثائق الأساسي إلى نظام قاعدة البيانات الأكثر تعقيدا وقوة مثل: قاعدة البيانات الهرمية والشبكية اللتين تمثلان هذه المساحة مع استقلال للبيانات وتلخيصها، ثم ظهرت قواعد البيانات العلائقية في السبعينيات التي تتيح للمستفيدين الوصول لبيانات مرنة قادرة على الوصول للغة والواجهة في السبعينيات؛ أما في منتصف الثمانينات ظهر نظام قاعدة البيانات القوية، وتم وضع العديد من نماذج البيانات المتقدمة، على سبيل المثال توسعة نموذج العلاقات ونموذج الشيء الموجه وتفسير النموذج وغيرهم. ولقد تطورت نماذج البيانات المتقدمة وتطبيق قاعدة البيانات الموجهة مع نهاية الثمانينات.

ولقد بدأ الاهتمام بالتنقيب عن البيانات عام ١٩٨٩م أثناء انعقاد ورشة عمل حول اكتشاف المعرفة في قواعد البيانات،[1] ومن ذلك الحين تم عقد هذه الورشة بصفة مستمرة سنويا حتى عام ١٩٩٤م، أما في ١٩٩٥م أصبح المؤتمر الدولي لاكتشاف المعرفة والتنقيب عن البيانات من أهم الأحداث السنوية، ومن ثم بدأ

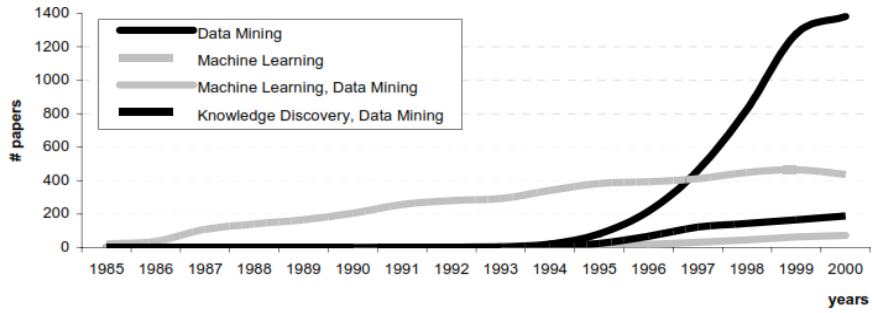
(1) Piatesky-Shapiro,G. (Jan. 1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop, AI Magazine, 11: 5, pp. 68-70. Available at: https://www.google.com.eg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CCEQFjAAahUKEwiWh9nL-djGAhVMPhQKHQUjCDk&url=https%3A%2F%2Fwww.aaai.org%2Fojs%2Findex.php%2Faimagazine%2Farticle%2Fdownload%2F873%2F791&ei=AyCkVdb5Acz8UIXGoMgD&usg=AFQjCNFB_-Qrs8RdlFxNlNI9jhuh61eiXw&bvm=bv.97653015,d.ZGU

تخطيط الإطار العملي للتنقيب عن البيانات واكتشاف المعرفة في كتابين: اكتشاف المعرفة في قواعد البيانات،(١) والتقدم في اكتشاف المعرفة والتنقيب عن البيانات(٢) ثم فاقت إمكانية تخزين كميات هائلة من البيانات قدرة العنصر البشري على التحليل والفهم بعد عام ٢٠٠٠م؛ ولم يكن هناك أداة مناسبة لاشتقاق المعلومات والمعرفة من البيانات، ويمكن إيجاد نماذج محددة وقواعد بواسطة أدوات التنقيب عن البيانات في ظل كم هائل من البيانات، والذي يوفر المعلومات الضرورية للأنشطة التجارية، والاكتشافات العلمية، والبحث الطبي وغيرها من المجالات. ولقد أصبح الذكاء التجاري القائم على التنقيب عن البيانات الوجه الجديد لصناعة تكنولوجيا المعلومات، ويتم تطبيق التنقيب عن البيانات حاليا بنجاح في تحليل بيانات تجارة البضائع، والتنبؤ بالمخاطر المالية، وجودة المنتج، والهندسة الجينية، واكتشاف نماذج الوصول لمواقع الإنترنت، والبحث عن المعلومات والتصنيف وغيرها من المجالات.

---

(1) Piatetsky-Shapiro, G., and Frawley,W., (Eds) (1991). Knowledge Discovery in Databases, AAAI/MIT Press, 1991. Available at: http://aaaipress.org/ojs/index.php/aimagazine/article/viewFile/1011/929

(2) Fayyad,U., Piatesky-Shapiro,G., Smyth, P., and Uthurusamy, R., (1996) Advances in Knowledge Discovery and Data Mining, AAAi/MIT Press, 1996. Available at: https://www.google.com.eg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0CCgQFjABahUKEwjlloCr-tjGAhVJOBQKHUkpBTM&url=https%3A%2F%2Fmitpress.mit.edu%2Fbooks%2Fadvances-knowledge-discovery-and-data-mining&ei=yiCkVeWJN8nwUMnSlJgD&usg=AFQjCNHeKXUmv7YO5vM2g_UcfrHrg6hsEw&bvm=bv.97653015,d.ZGU



شكل رقم (٣) تطور حقول التنقيب عن البيانات، واكتشاف المعرفة والتنقيب عن البيانات(١)

وفيما يتعلق باستفادة مجال المكتبات والمعلومات لهذه التقنية، فقد تم بالفعل استخدام هذه التقنية مع تخصصنا فيما يتعرف بالتنقيب عن البيانات في المكتبات، وقد استخدم مصطلح التنقيب عن البيانات في مجال المكتبات والمعلومات للمرة الأولى عام ١٩٩٨م، ونظراً لما حدث من التباس عند الاسترجاع من جانب الباحثين في مجال المكتبات؛ حيث كانت نتائج البحث تسترجع المؤلفات عن مكتبات البرمجيات المتعلقة بالتنقيب عن البيانات، وليس التنقيب عن البيانات للمكتبات، لذلك صك مصطلح آخر لفض هذا الالتباس عام ٢٠٠٣م مصطلح "التنقيب الببليوجرافي Bibliomining" الذي يعني بتطبيق الأدوات الإحصائية وأدوات التعرف على الأنماط في كم كبير من البيانات المرتبطة بنظم المكتبات من أجل المساعدة في اتخاذ القرارات، أو تبرير الخدمات المقدمة وتطويرها خاصة في المكتبات الرقمية، وبمعنى آخر فإن التنقيب الببليوجرافي عبارة عن إعادة رؤية للبيانات من منظور مختلف لتحقيق قيمة مضافة".(٢)

(1) Cios, K. and Kurgan, L. Trends in Data Mining and Knowledge Discovery. Available at: http://www.cioslab.vcu.edu/Publications/Papers/chapterTrendsDM2003.pdf

(٢) **وسام محمود أحمد درويش**. نحو رؤية جديدة لإدارة المكتبات باستخدام تقنية التنقيب عن =

٣/١ مهام التنقيب عن البيانات

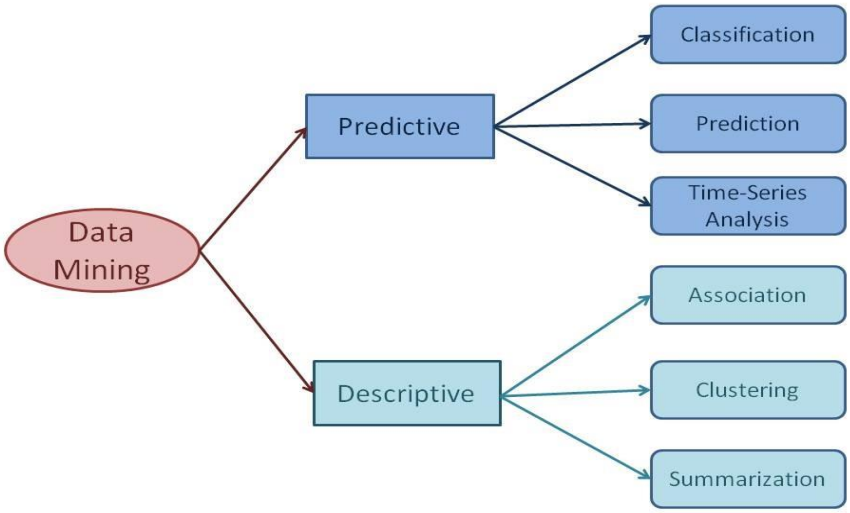
يقوم التنقيب عن البيانات بعمليتين أساسيتين متمثلتين فيما يلى:

١) التنبؤ: يهدف التنقيب عن البيانات إلى وضع توقعات مع الإحالة للسمة العامة أو سمات الكائن لبيانات التصنيف غير المعروفة، ويستخدم نموذج التعلم المتاح للتنبؤ، ويعد التصنيف والانحدار نوعين أساسيين من نموذج التنبؤ، فيستخدم الأول للتنبؤ بالقيمة المنفصلة أو الرمزية، أما الانحدار فيستخدم للتنبؤ بالقيم المستمرة، أي الإجابة عن سؤال حول شراء السلع عبر الانترنت إما أن يكون س أو ص وهذا ينطبق على الحالة الأولى وهى التصنيف، أما في حالة التنبؤ بأسعار الأسهم والاتجاهات فذلك من خلال مهام الانحدار. يمكن لنماذج التنبؤ تحديد فوائد السوق ومخاطره، كما يمكن التنبؤ بمعدلات استهلاك موارد الأرض.

٢) الوصف: يعد نموذج البيانات المحتمل المتاح الذي يلخص العلاقات الدور التوثيقى والتفسيري، يستخدم تحليل العلاقة عادة لوصف نموذج بخصائص علائقية قوية لاشتقاق النماذج المهمة لإيجاد العلاقة بين البيانات. يعبر اشتقاق خصائص الصيغ عن الخصائص العامة لمجموعة البيانات من مستودع البيانات، أو إيجاد ملامح أخرى للتمييز بين خصائص الأسلوب الواحد، مثل: اشتقاق الملامح وتمييزها عن الحالات الأخرى. وعلى الرغم من منطقها، إلا أنه يمكن لتنقيب دور التجميع إيجاد الكثير من التفاعلات المهمة، وهذا ما يطلق عليه تحليل سلة السوق الشهير، الذي كان سلاحا سريا في المتاجر الكبرى، حيث يمكن لتحليل سلة السوق المساعدة في إيجاد عمليات بيع المخازن والبضائع التابعة لها.[١]

= البيانات.- cybrarians journal.- ع ١٩ (يونيو ٢٠٠٩).
http://www.journal.cybrarians.org/index.php?option=com_content&view=article&id=437:-data-mining-&catid=164:2009-05-20-10-02-29&Itemid=60

(1) Padhy,N. Mishra,P. and Panigrahi, R. (June 2012) The Survey of Data =



شكل رقم (٤) مهام التنقيب عن البيانات(١)

ويمكن تقسيم تصنيف التنقيب عن البيانات إلى قطاعين: التنقيب عن البيانات المباشر وغير المباشر؛ حيث يتمثل الهدف من التنقيب عن البيانات المباشر في استخدام البيانات المتاحة لإنشاء نموذج مع وصف للمتغيرات؛ أما الهدف من التنقيب عن البيانات غير المباشر هو عدم توافر اختيار لمتغير محدد، لكن بناء علاقة بين كل المتغيرات. هذا ويندرج التصنيف والتقدير والتنبؤ ضمن التنقيب عن البيانات المباشر؛ أما دور التجميع والاتحاد والوصف والعرض فتندرج ضمن التنقيب عن البيانات غير المباشر. دور التجميع غير معروف مسبقا ما المعرفة

= Mining Applications And Feature Scope International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2 (3). Available at: http://arxiv.org/ftp/arxiv/papers/1211/1211.5723.pdf

(1) Data Mining Tasks. Available at: http://wideskills.com/data-mining/data-mining-tasks

التى يجب الحصول عليها، ما يمكن الحصول عليه بعد تحليل البيانات، مثل المستفيد يشتري منتج (أ) بمنتج (ب)؛ أما العنقود فهى تجميع التسجيلات المتشابهة ووضعها معا في مجموعة، والاختلاف بين العنقود والتصنيف أن العنقود لا يعتمد على تصنيفات محددة مسبقا، ولا مجموعة مدربة، أما الوصف والعرض فهمها تمثيل لنتائج التنقيب عن البيانات.(١)

٤/١ أنواع التنقيب عن البيانات

يوجد العديد من أنواع التنقيب أو طرق التنقيب والتي يمكن تلخيصها فيما يلى:

١/٤/١ تحليل الارتباط Correlation analysis: أي اكتشاف المعرفة ذات العلاقة والمفيدة من مجموعة كبيرة من البيانات، وتكمن الفكرة الأساسية في أن و > ب، حيث يعبر (و) عن مجموعة الصفات، ويمثل (ب) السمات منفردة، وتقوم القواعد بتفسيرها إذا كان (و) قيمته صحيحة، فإن (ب) كقيمة مفردة لديه إمكانية واتجاه للقيمة الصحيحة في قائمة قاعدة البيانات. ويمكن توضيحها بأنه بعد شراء سلعة، فما مدى احتمال الاستمرار في شراء سلعة (ب)؟(٢).

(1) Weiping,F. and Yuming,W. (Dec. 2013) The Development of Data Mining International Journal of Business and Social Science, 4 (16). Available at: http://ijbssnet.com/journals/Vol_4_No_16_December_2013/14.pdf

(2) Jensen, D. and Neville,J. Correlation and Sampling in Relational Data Mining. Available at: https://www.cs.purdue.edu/homes/neville/papers/jensen-neville-interf2001.pdf

١/ ٤/ ٢ شجرة القرارات Decision Tree: تتكون شجرة القرارات من سلسلة من العقد والأفرع، ثم تتفرع العقد إلى عقد فرعية بواسطة الأفرع، حيث تمثل العقد السمات التى يجب اعتبارها في عملية صنع القرار، ثم تأتي القيم المختلفة للسمات من الأفرع المختلفة؛ وباستخدام نموذج شجرة القرارات في صنع القرارات يمكن البحث من الجذر إلى الأوراق؛ فتحتوي عقد الأوراق على نتائج كل تصنيف.(١)

١/ ٤/ ٣ الخوارزميات الجينية Genetic Algorithm: إن الخوارزميات الجينية بحث احتمالات لإيجاد العملية المثلى، ونتجت عن مجموعات محددة أو عشوائية، وفقا لقواعد معينة من العملية لاستمرار الحساب التكراري، مثل الاختيار، والإنتاج، والتبادل والتغيير وغيرها، وهى عملية الاحتفاظ بالمتغيرات الجيدة، والقضاء على المتغيرات الرديئة، وتوجيه البحث للاقتراب من الحل الأمثل طبقا لمتطلبات كل شخص، ويتطلب تنفيذ الخوارزمية الجينية اثنين من عمليات تحويل البيانات، وهما: فك التشفير والترميز، حيث يتمثل الترميز في تحويل معاملات مسافة البحث إلى كروموزوم أو أفراد من المساحة الجينية؛ أما فك التشفير فيتمثل في تحويل الكروموزوم أو أفراد المساحة الجينية إلى معاملات لمساحة البحث، هذا ولقد تطورت الخوارزمية الجينية بناء على محاكاة علم الوراثة، للعمل مباشرة على

(1) Sharma, P., Bhartiya, R. (Dec. 2012) Implementation of Decision Tree Algorithm to Analysis the Performance International Journal of Advanced Research in Computer and Communication Engineering, 1(10). Available at: http://www.ijarcce.com/upload/december/24-Implementation%20of%20Decision.pdf

هيكل الكائنات، فهى لديها قوة كافية بدون قيود للقيام بعملية الاشتقاق والوظيفة.(١)

١/ ٤/ ٤ شبكات النظرية الافتراضية Bayesian Networks: تعتمد شبكات النظرية الافتراضية على نموذج رياضي للاستدلال الاحتمالي، ويتم الاستدلال الاحتمالي من خلال بعض المعلومات للحصول على احتمالات للمتغيرات الأخرى، وتعتمد شبكات النظرية الافتراضية على أساس الاستدلال الاحتمالي لحل مشكلة عدم اليقين وعدم الاكتمال، ولها أفضل مزية لحل الأخطاء الناجمة عن عدم اليقين الصعب والارتباط، والمستخدم على نطاق واسع في العديد من المجالات. ويمكن باستخدام بنية شبكة النظرية الافتراضية وجداول الاحتمالات الشرطية حساب الاحتمالات لقيم عقدة معينة بعد تقديم الأدلة.(٢)

١/ ٤/ ٥ مسار المجموعة الخام Rough Set Approach: تعد نظرية المجموعة الخام طرقا رياضية لمعالجة الغموض وعدم اليقين باستخدام طريقة مجموعة الخام التي تمكن من تحليل جدول القرارات، وتقييم أهمية سمات محددة، والتقليل من مجموعة الخصائص والطاقة النووية والتخلص من الخصائص الإضافية الزائدة من جدول القرارات وقواعد التصنيف التي تظهر من جدول التقليل لصناع القرار، وتعتمد الفكرة الرئيسة للمجموعة الخام على المعرفة

(1) Flockharta, I. and Radclieab, N. (1996) A Genetic Algorithm Based Approach to Data Mining Presented at "AAAI: Knowledge Discovery and Data Mining", Portland, Oregon. Available at: http://www.stochasticsolutions.com/pdf/kdd96.pdf

(2) Heckerman,D. (1997) Bayesian Networks for Data Mining Data Mining and Knowledge Discovery,1: 79–119. Available at: http://machinelearning101.pbworks.com/f/Tutorial-BayesianNetworks.pdf

الموجودة لمشكلة معينة، من خلال تصنيف إدارة البيانات الفعلية، وتقسيم نطاق المشكلة، وتقليل البيانات في إطار فرضية الاحتفاظ بالمعلومات المهمة، وتقليل نووية المعرفة، وتقييم استقلالية البيانات، واشتقاق قواعد تصنيف المفهوم.(١)

١/ ٤/ ٦ الشبكة العصبية Neural Network: هي نظام ديناميكي بهيكل طوبولوجي لتوجيه الرسم البياني، فهي تتعامل مع المعلومات من خلال الاستجابة لحالة الإدخال المستمرة أو المتقطعة، ويتكون نظام الشبكة العصبية من وحدات معالجة بسيطة وكبيرة، من خلال الربط ببعضها البعض عل نطاق واسع وتشكيل شبكة معقدة من النظم. هذا وعلى الرغم من بنية ووظيفة كل خلية بسيطة جدا، إلا أن سلوك نظام الشبكة يتكون من عدد كبير من الخلايا الملونة والمعقدة. هذا وتتناسب الخوارزمية مع تجميع البيانات والتى يمكن أن تقدم الكثير من المعلومات المعقدة والبيانات العادية والمنظمة، لإيجاد العلاقة الداخلية بين البيانات من خلال تشابه الزمان والمكان.(٢)

١/ ٤/ ٧ التحليل الاحصائى Statistical Analysis: هو طريقة دقيقة للتنقيب عن البيانات بالاعتماد على نظرية الاحتمالات الإحصائية، مثل: تحليل الانحدار وتحليل العوامل من خلال نماذج من الكائنات والعثور على استنتاجات، وعادة ينقسم إلى الخطوات التالية: وصف طبيعة البيانات التحليلية، والمجموعة البحثية من علاقات البيانات، وبناء النموذج، وملخص البيانات، وعلاقة المجموعة

(1) Pawlak, Z. Rough Sets And Data Mining. Available at: http://bcpw.bg.pw.edu.pl/Content/1884/RSDMEAK.pdf

(2) SINGH,Y. and Chauhan, A. Neural Networks In Data Mining Journal of Theoretical and Applied Information Technology, 5 (6). Available at: http://jatit.org/volumes/research-papers/Vol5No1/1Vol5No6.pdf

الأساسية، وشرح صلاحية النموذج، وأخيرا التنبؤ بالتنمية المستقبلية. ويستخدم SPSS وSAS على نطاق واسع كبرامج تطبيقية للإحصائيات.(١)

هذا ويتضح مما سبق أنه يوجد سبعة أنواع من التنقيب عن البيانات والتي تتمثل في: تحليل الارتباط، شجرة القرارات، الخوارزميات الجينية، شبكات النظرية الافتراضية، مسار المجموعة الخام، الشبكة العصبية، التحليل الاحصائي.

٥/١ تطبيقات التنقيب عن البيانات

يلعب التنقيب عن البيانات دورا أساسيا في البنوك، والتأمين والنقل والتجارة، ويمكن للتنقيب عن البيانات حل كثير من المشاكل الحسابية، وزيادة الأرباح وصنع قرارات حكيمة. ولم تكن العمليات التجارية التطبيق الأول لتقنيات التنقيب عن البيانات إنما كانت من المجالات المهمة، لأن العمليات التجارية بها الكثير من بيانات المبيعات، مثل تسجيلات الشراء الخاصة بالمستفيدين، ومعلومات المستهلكين، ومعلومات الخدمة وغيرها. يمكن للشركات استخدام البيانات لتصنيف المستفيدين من مجموعات المستفيدين الأساسية وإيجاد الخصائص المشتركة للمستفيدين ورغباتهم المستقبلية، وتقديم منتجات كافية وخدمات تلبي رغباتهم. وعند استخدام تطبيقات برامج التنقيب عن البيانات، يجب اختيار الخوارزمية المناسبة، والمعرفة فيما وراء البيانات التي يمكن إيجادها. وبما أن التنقيب عن البيانات يقوم بعدة مهام مثل: تجميع البيانات وتخزينها وتنظيمها لذا فهو يستخدم في عدة مجالات مثل الطب والمالية والذكاء الصناعي والقانون والدفاع والتعليم وعمليات التحكم وغيرها، وتستخدم معظم التطبيقات التنقيب

(1) Friedman, J. Data Mining and Statistics: what's the Connection? Available at: http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf

عن البيانات للدعاية والتسويق والمبيعات، كما يمكن استخدامه في التشخيص، ومن الأمثلة على هذه التطبيقات:

١) مكاتب الائتمان على القروض: تعتمد على ملاحظات الأفراد المتشابهين في نماذج الشراء، والدخل، والقروض، كما يمكن تطوير وإنشاء تقارير موجزة عن الزبائن المهمين وعن بطاقات الائتمان.

٢) السوبر ماركت: ينظم بضائعه طبقا لنماذج البيع والمعلومات حول الجمعيات بين المنتجات، والتسويق لفئة معينة لإيجاد الزبائن من أجل منح التخفيضات لهم لسبب معين، وايجاد السلع التي تباع مع بعضها.

٣) شركات الأدوية: تقوم بتحليل الوصفات الطبية (الروشتات) لإرسال المواد الترويجية للزبائن المستهدفين، وايجاد منوال معين لاستعمال الخدمات والسلع.

٤) وكالة الاستخبارات: تستعرض نماذج الإنفاق وبيانات السفريات للكشف عن السلوكيات غير الطبيعية من موظفيها.

٥) طبيب التحاليل: يقوم بتحليل صور الأشعة السينية لاكتشاف الأنماط غير الطبيعية.

٦) نظام حجز الطيران: يستخدم معلومات حول نماذج السفر والاتجاهات لتحقيق أقصى قدر من استخدام المقاعد.

٧) تطبيقات تكنولوجيا المعلومات: يساعد التنقيب عن البيانات في التأكد من جودة البيانات، فعلى سبيل المثال في يساعد التنقيب عن البيانات في التطبيقات اللوجستية في اختيار الأفراد المناسبين للعمل في مشروعات محددة.

٨) البنوك: تستخدم التنقيب عن البيانات للحصول على بيانات تساعد على جذب الزبائن.

وعلى الرغم من استخدام هذه التطبيقات لبعض الوقت بشكل تام، إلا أنها تعتمد على التحليل الاحصائي يدويا، ولقد بدأ الموظفون مؤخرا باستخدام تكنولوجيا التنقيب عن البيانات لتحليل البيانات وإنشاء علاقات متبادلة ووضع تنبؤات.(١)

٩) تطبيقات التنقيب عن البيانات بالمكتبات

أ) **إدارة مقتنيات المكتبة**: فمن خلال استخدام تقنية التنقيب عن البيانات يمكن للمكتبة إدارة مقتنياتها بشكل جديد في أكثر من جانب يذكر منها:

معرفة الثغرات في مقتنيات المكتبة، حيث من خلال استخدام خوارزميات التنقيب عن البيانات يمكن الحصول على أنماط معرفية جديدة ودقيقة لم تكن معروفة من قبل؛ تبين لنا أوجه القصور في المقتنيات وأوجه الزيادة، مما يصبح أمام متخذي القرار الفرصة في تقييم مقتنياتهم في أكثر من جهة، وهذا يساعد أيضا في فتح الباب أمام المكتبة للمشاركة في المصادر مع المكتبات الأخرى لسد هذه الفجوات لديها ولمساعدة المكتبات الأخرى في مشاركتها في الجوانب الأخرى التى تم تغطيها بشكل جيد.

إعادة تقسيم مجموعات المكتبة وتكاملها، يساعد ذلك متخذي القرار في

(1) Thuraisingham,B. (2000). A Primer for Understanding and Applying Data Mining. IT Pro IEEE Xplore. Available at: https://www.utdallas.edu/~bxt043000/Publications/Journal-Papers/DS-DM/J71_A_Primer_for_Understanding_and_Applying_Data_Mining.pdf

الحد من تكرار المقتنيات ويحدث ذلك على مستوى الموضوع الواحد و/أو الموضوعات الشبيه وذات الصلة؛ فمن خلال التحليل الدقيق والذكى لمقتنيات المكتبة وباستخدام الآليات المختلفة للتنقيب عن البيانات يتم استنباط أنماط معرفية (تمكن من وجود علاقات تربط موضوعات ببعضها البعض لم يكن واضح من ذي قبل أمام المسئولين ومتخذي القرار بوجود مثل هذه العلاقات بينهم)، ومن هنا تحدث التكاملية بين الموضوعات والمقتنيات؛ فبدلاً من شراء مقتنيات جديدة لموضوع ما يتم استبدال ذلك بوضع رؤى وتقسيمات جديدة لمقتنيات موجودة بالفعل من الممكن أن يتم الاستفادة منها وتقديمها للمستفيدين في هذا الموضوع.

ب) **قواعد بيانات خاصة بالمستفيدين**: من المعروف أنه يتوافر لدى المكتبة العديد من المعلومات التى تتعلق بالمستفيدين، وتعد دراسات سلوك المستفيدين تجاه مجموعات المكتبة سواء في البيئة التقليدية أو بيئة الشبكات الرقمية، من المعلومات القيمة لتطويرات مبتكرة في كيفية عرض وإتاحة المعلومات بالشكل الذي قد يرغبه المستفيدين، ومن هذه المعلومات الدراسات الاستقصائية، وبيانات الإعارة، ومرات الولوج الى غير ذلك من المصادر، ولا سيما إذا تم وضع كل هذه البيانات في قاعدة بيانات واحدة واستخدم فيها آليات التنقيب عن البيانات وتم ربطها مع مقتنيات وأنشطة وخدمات المكتبة، يصبح أمام المسئولين مادة خصبة يمكن من خلالها استخراج معلومات تفيد في كافة الأوجه بالمكتبة.

ج) **تنمية الموارد البشرية**: يوجد بالمكتبة موارد بشرية متخصصة وغير متخصصة، ويمكن تقسيم العاملين غير المتخصصين لأكثر من تخصص، وبالتالى يصبح هناك العديد من التخصصات والمؤهلات والدرجات العلمية داخل المكتبة الواحدة، ومع زيادة أعداد العاملين يزداد الأمر صعوبة أمام المسئولين ومتخذي القرار في الإلمام بجميع جوانبهم المختلفة. ولكن في حالة توافر قاعدة

بيانات واحدة تشتمل على كافة بيانات العاملين بالمكتبة من حيث (أنواعهم – حالاتهم الاجتماعية – مسكنهم – تخصصاتهم – مؤهلاتهم – هواياتهم – ظروفهم الاقتصادية – خبراتهم...إلى غير ذلك من بيانات يمكن تجميعها عن العاملين) مستخدمة في ذلك آليات التنقيب عن البيانات، يمكننا استخراج علاقات ورابط قوية بين هذه البيانات واستنباط أنماط معرفية ومعلوماتية تربط كل ذلك ببعضه البعض، وهذه المعلومات الجديدة لم يكن من الواضح اكتشافها أو الحصول عليها إلا من خلال قاعدة التنقيب عن البيانات، مما تسنح الفرصة أمام المسئولين ومتخذي القرار من اكتشاف الموارد البشرية المتاحة لديهم بصورة جديدة ومبتكرة تساعدهم في إعادة توزيعهم داخل الأقسام والأنشطة المختلفة في المكتبة ـ ليس فقط وفق تخصصاتهم ولكن وفق الأنماط المعرفية الذكية المكتشفة ـ مما يساعد كل من المكتبة وعامليها على حد سواء.

د) **خدمات المعلومات المتاحة بالمكتبة**: ما سبق ذكره من حيث إعادة تقسيم المقتنيات وتكاملها وتنمية الموارد البشرية المتاحة بالمكتبة وحُسن توزيعها ودراسات سلوك المستفيدين، يصبح أمام المسئولين الرؤية الواضحة لكيفية تقييم الخدمات المقدمة ومدى جدواها وعلاوة على ذلك التخطيط الصحيح لتقديم خدمات جديدة.

ه) **أوجه صرف الميزانية**: فمن الطبيعى بعد إعادة الهيكلة الجديدة لإدارة المكتبة بمساعدة آليات هذه التقنية الحديثة، يصبح هناك بعض الترشيد في أوجه صرف الميزانية وتوفيرها لجوانب جديدة كان من الممكن عدم التفكير فيها مع قصور الميزانية لتغطيتها.

ويتضح مما سبق أنه يمكن تطبيق التنقيب عن البيانات في العديد من المجالات ومن هذه المجالات: مكاتب الائتمان على القروض، السوبر ماركت، شركات الأدوية، وكالة الاستخبارات، طبيب التحاليل، نظام حجز الطيران،

تطبيقات تكنولوجيا المعلومات، البنوك، المكتبات ومراكز المعلومات.

هذا ويمكن تلخيص التطبيقات طبقا لما يمكن للتنقيب عن البيانات القيام به، وليس طبقا للتخصصات العلمية التى تم التنويه عنها:

١. كتابة تقرير مختصر عن فئة معينة Profiling Populations: تطوير وإنشاء تقارير موجزة عن الزبائن المهمين وعن بطاقات الائتمان.

٢. تحليل النزعة التجارية Analysis of Business Trend: ايجاد الأسواق ذات قدرات النمو القوية أو الضعيفة.

٣. التسويق لفئة معينة: Target Marketing إيجاد الزبائن من أجل منح التخفيضات لهم لسبب معين.

٤. تحليل الاستعمال Usage Analysis: ايجاد منوال معين لاستعمال الخدمات والسلع

٥. فعالية الحملة Campaign Effectiveness: مقارنة استراتيجيات الحملات مع بعضها البعض من أجل إيجاد أكثرها فعالية وتأثيرا.

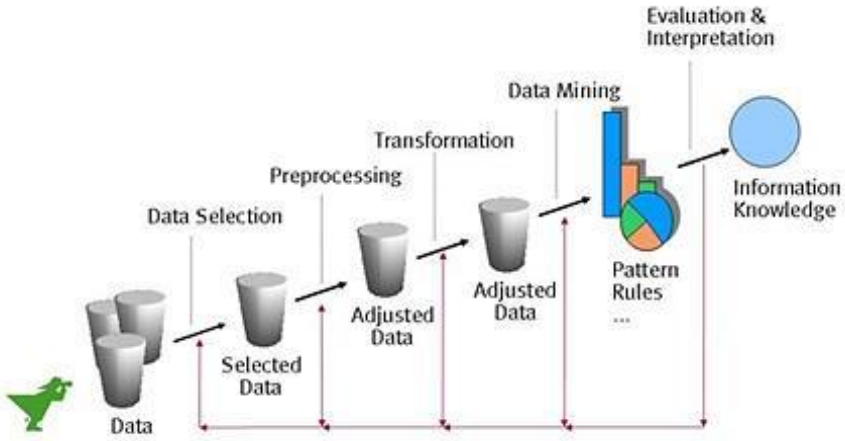
٦. جاذبية السلعة: إيجاد السلع التى تباع مع بعضها.

المبحث الثاني
أدوات التنقيب عن البيانات وتقييمها

تمهيد:

على الرغم من حداثة مجال التنقيب عن البيانات، إلا أنه كنتيجة للأبحاث في السنوات الأخيرة، ظهرت العديد من الخوارزميات والطرق والتقنيات التى تتيح للمستفيدين القيام بالعديد من الوظائف باستخدام هذه التقنيات. هذه الوظائف مهمة للشركات التي لديها حرية اختيار نظام التنقيب عن البيانات لتطبيقه. إن

لتقنيات التنقيب عن البيانات وخوارزمياتها وطرقها أهداف لتصميمها وأدائها مهمة محددة، ومن أكثر المهام الشائعة التى يقوم بها نظام التنقيب عن البيانات العناقيد والتجميع والتصنيف والتنبؤ وتحليل سلاسل البيانات والتحليل الخارجى. هذا ويمكن تخصيص نظام التنقيب عن البيانات لمهمة واحدة أو العديد من المهام العامة، وتقوم معظم أدوات التنقيب عن البيانات بتطبيق العديد من الخوارزميات والقيام بالعديد من المهام الكافية لتلبية احتياجات المستفيدين.

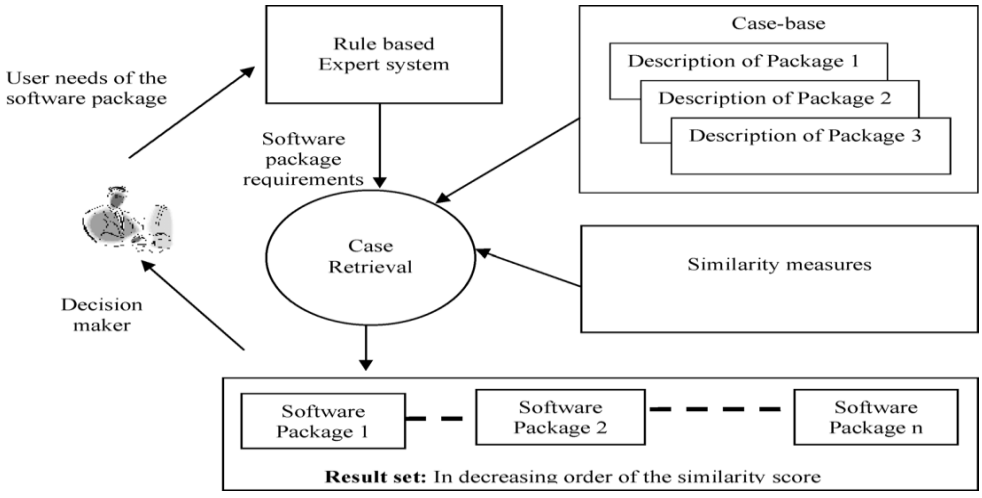


شكل رقم (٥) نموذج التنقيب عن البيانات[١]

ولقد قام العديد من الباحثين في الآونة الأخيرة بنشر الخطوط العريضة لمعايير الاختيار والتقييم لحزم البرمجيات، ومعظم الأوراق البحثية عبارة عن مراجعة لمعايير اختيار عامة للعديد من البرمجيات مثل المحاسبة وأداة المحاكاة وغيرها، أما الأوراق البحثية التى تناقش معايير اختيار البرمجيات لأي مجال بشكل خاص مثل التنقيب عن البيانات فهي قليلة، كما أن هناك القليل من العمل الذي تم

(1) Knitting and Crochet Patterns. (2015). Pattern Discovery In Data Mining Available at: http://ktuliuepatt.com/pattern-discovery-in-data-mining/

في إطار صنع القرار، وعلى أية حال يمكن القول بأن المعايير المتعلقة بمتطلبات الأداء والجودة والعتاد والبرمجيات يمكن أن يتم تقييمها لأي برامج. ولقد أصبحت مهمة اختيار حزمة البرمجيات أصعب بسبب الصعوبة في الوصول لإمكانية تطبيق حزمة البرمجيات طبقا للاحتياجات التجارية للمنظمة طبقا لإتاحة العدد الهائل من عدد الحزم في السوق؛ وبسبب عدم توافق بين العتاد وحزمة البرمجيات في السوق؛ وكذا نقص المعرفة التقنية والخبرة لدى صانعى القرار؛ هذا بالإضافة إلى التطورات المستمرة في تكنولوجيا المعلومات.ويوضح الشكل التالى مسار تقييم واختيار حزم البرمجيات الخاصة بالتنقيب عن البيانات(١).



شكل رقم (٦) مسار تقييم واختيار حزم البرمجيات

(1) Jadhav, A., and Sonar, R. (2011). Framework for evaluating and selection of the software packages: A hybrid knowledge based system approach. the journal of system and software, 84 (8): pp1394-1407. Available at: http://romisatriawahono.net/lecture/dm/paper/classification/Jadhav%20-%20Framework%20for%20evaluation%20and%20selection%20of%20the%20software%20packages%20-%202011.pdf

هناك العديد من الدراسات التي تقترح معايير لتقييم أدوات التنقيب عن البيانات سواء التجارية أو مفتوحة المصدر أو المجانية، منها دراسة كين كولير عام ١٩٩٩م Ken Collier وآخرون والتي تم فيها وضع منهجية أولية لتقييم أدوات التنقيب عن البيانات، وتوضح الدراسة أن من خلال الخبرة البحثية تبين أن هناك أربعة تصانيف أساسية لمعايير تقييم أدوات التنقيب عن البيانات: الأداء، والوظيفة، والقدرة على الاستخدام، ودعم الأنشطة الثانوية، وهذه الخبرة تدعمها العديد من الدراسات السابقة. وتشكل هذه التصانيف اطار عمل تقييم في مركز البيانات الداخلية بجامعة شمال أريزونا Insight (CDI) at Northern Arizona Center for Data University (NAU)

أما بارجافا **Bhargava,N**. وآخرون عام ٢٠١٣م قسموها إلى (٦) ستة معايير أساسية، وهي: أداء البرنامج، الوظيفة، ومهامه الثانوية، العتاد، ومتطلبات البرمجيات لتشغيل حزمة البرمجيات بكفاءة، ومسئوليات البائع، والجودة أو القدرة على معالجة النقص في البيانات. ويقرر أن كل المعايير مترابطة؛ فلا يمكن إعطاء وزن أعلى لأي معيار وتجاهل المعايير الأخرى. فعلى سبيل المثال إذا تم تطبيق برامج التنقيب عن البيانات في صناعة الهندسة، فهذا سيوفر الوقت والتكلفة في مجالات التصميم والصناعة والصيانة وغيرها، وقد يحدث ذلك باختيار البرنامج الصحيح للتنقيب عن البيانات والذي قام بتحليل كم هائل من البيانات في الصناعة والتنبؤ بها.(١)

كما قسمها مايك فيرجوسن Mike Ferguson إلى (٩) تسعة معايير أساسية والتي يندرج أسفلها بعض المعايير الفرعية التي تفسر وتوضح المعايير

(1) Bhargava,N., Aziz,A. and Arya,R. (2013) Op. Cit.

الأساسية والتي تتمثل في: هيكل المنتج، وتكامل مستودع البيانات، والأداء، والوظيفة، والتمثيل، ومصادر البيانات، وإعداد البيانات، والبيئة، والإدارة.(١)

هذا ولقد اتفق الباحثون على أن هناك أربع معايير أساسية والتي يمكن استخدامها في تقييم وتصميم اطار عمل اختيار أداة التنقيب عن البيانات، وتتمثل هذه المعايير في الأداء والوظيفة والاستخدام والدعم.

يعد المعيار الأول والأهم، الأداء فهو يركز على نقاط الجودة في قدرة الأداة وسهولة معالجة البيانات تحت ظروف مختلفة وليس متغيرات الأداء الناتجة عن تركيب العتاد أو صفات الخوارزمية، ويجب القيام بعملية التقييم مع الحفاظ على ملامح تركيب العتاد؛ لأن مرحلة تجربة التجريب لها تأثير رئيس على أداء الأداة، وبشكل إجمالي يتعامل هذا المجال من الاهتمام المتوقع مع وجهات النظر الحاسوبية. ويتمثل المعيار الثاني في الوظائف والذي يعالج إمكانيات الأداة، فهو يتضمن مجموعة من الإمكانيات، والتقنيات والمنهجيات للتنقيب عن البيانات، وتساعد وظائف البرامج في تقييم مدى نجاح الأداة في التكيف مع المشاكل المختلفة للتنقيب عن البيانات مثل النطاقات. أما المعيار الرئيس الثالث يناقش مرحلة بناء هذا الاطار وهو الاستخدام، والمقصود بالاستخدام تحديد عدد المستفيدين الذين يمكنكم استخدام الأداة دون فقد في الوظيفة أو خسارة في الاستخدام، فسهولة الاستخدام وسوء الاستخدام يسيران جنبا إلى جنب، فإذا وفرت الأداة فهما سريعا

(1) Ferguson, M. Evaluating And Selecting Data Mining Tools. InfoDB, 11 (2): pp: 1-10.- Available at: http://www.evaltech.com/admin/upload/Evaluating_Data_Mining_Tools.pdf

وسهلا، فيجب أن تركز أيضا على جودة البيانات في نماذجها. هذا ويتمثل المعيار الرابع في الدعم والمقصود به أداء العديد من الوظائف الثانوية المطلوبة في عملية التنقيب عن البيانات، وتتضمن هذه المهام اختيار البيانات، وتنظيفها، وإثراءها، وترشيحها، وتعديلها أو حذفها، ويمكن اعتبار عملية تطويع البيانات لتقييم الأداة لدعم الاهتمامات لأنه من غير اللائق توقع مجموعة بيانات نظيفة لديها درجة عالية من التطويع.[١]

والجدير بالذكر أن هناك العديد من المعايير الفرعية التي تندرج أسفل المعايير الرئيسة[٢]التي يمكن توضيحها وتطبيقها على أدوات التنقيب عن البيانات فيما يلي:

أولا: معايير الأداء: تعالج هذه المعايير الأداء الحاسوبي والمهام التي تتناسب مع الأداة، مثل نظام التشغيل والتركيب وحجم البيانات وكفاءتها وغيرها.

(1) Piatetsky -Shapiro,G., Brachman,R., Khabaza,T., Kloesgen,W. and Simoudis,E. (1996) An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications. Proceedings KDD -96, AAAI Press, Portland, Oregon August 2 -4, 1996. Available at: http://aaaipress.org/Papers/KDD/1996/KDD96-015.pdf

(2) Verma, V. and Dhawan, S. (May 2014) Methodology for Selection of a Data Mining Tool. International Journal of Software & Hardware Research in Engineering, 2 (5): pp. 189- 192. Available at: http://ijournals.in/ijshre/wp-content/uploads/2014/05/IJSHRE-2552.pdf

جدول رقم (١)

معايير الأداء لأدوات التنقيب عن البيانات

المعيار	الوصف
نظام التشغيل	هل البرنامج يعمل على نظم تشغيل متنوعة؟ وهل يعمل على نظم تشغيل المستفيدين التجارية النموذجية؟
تركيب البرنامج	هل يستخدم البرنامج تركيب الخادم – العميل؟ أو يستخدم تركيبا وهيكلا مستقلا؟ وهل يحق للمستفيدين اختيار الهيكل أو التركيب المرغوب؟
الوصول إلى البيانات	ما واجهة البرنامج المطلوبة؟
حجم البيانات	هل البرنامج يتناسب مع مجموعات البيانات الكبيرة؟
الكفاءة	هل يظهر البرنامج نتائج خلال وقت مناسب؟
التوافقية	هل تتوافق واجهة الأداة مع الأدوات الأخرى أو مجموعة الأدوات؟
المتانة	ما درجة تماسك الأداة؟ كم عدد المرات التي يتم فيها عطل الأداة أو انهيارها؟

ثانيا: معايير الوظيفة: تعالج هذه المعايير عوامل مختلفة مثل الإمكانيات والتقنيات والمنهجيات المختلفة، وتختبر هذه العوامل الأداة مقابل مشكلة التنقيب عن البيانات، لذا يمكن معرفة مدى تأقلم الأداة مع الظروف المختلفة، كما يتيح اختبار وظائف الأداة الأساسية مع النظر لمسار تركيب خوارزمية مشكلة التنقيب عن البيانات.

جدول رقم (٢)

معايير الوظيفة لأدوات التنقيب عن البيانات

المعيار	الوصف
التنوع	هل يوفر البرنامج مجموعة مختلفة من تقنيات التنقيب والخوارزميات لدعم القرارات؟
المنهجية	هل يساعد البرنامج المستفيد بتقديم منهجية التنقيب خطوة بخطوة؟
الصلاحية	هل تدعم الأداة نموذج الصلاحية بالإضافة إلى إنتاج نموذج؟
نوع البيان	هل تطبيق الخوارزميات المدعمة تعالج مجموعة من أنواع البيانات؟
القدرة على التعديل	هل يمكن للمستفيد تعديل وضبط الخوارزميات؟
عينة البيانات	هل يمكن للأداة أخذ عينة عشوائية من البيانات لنموذج التنبؤ؟
التقرير	هل تظهر نتائج تقارير تحليل التنقيب بعدة طرق مختلفة؟
تصدير النموذج	هل من الممكن تصدير النموذج لصيغ أخرى من الأدوات مثل اكسيل أو اس كيو ال SQL

ثالثا: معايير الدعم: تستخدم لتتبع معايير الدعم ومكان مصادر الدعم التي يتم قياسها بهذا المعيار، ويستخدم هذا المعيار أيضا لبناء خصائص تساهم في دعم النظام، ومن هذه الخصائص: تنظيف البيانات واستبدالها وفلترتها وحذفها وغيرها.

جدول رقم (٣)

معايير الدعم لأدوات التنقيب عن البيانات

المعيار	الوصف
تنظيف البيانات	هل تتيح الأداة إمكانية تعديل القيم الخطأ في مجموعة البيانات أو أداء عمليات أخرى مصممة لتنظيف البيانات؟
استبدال البيانات	هل تتيح الأداة التبديل الشامل لأحد قيم البيانات أو مجموعة من القيم؟
فلترة البيانات	هل تتيح الأداة اختيار المجموعات الفرعية من البيانات بناء على معايير اختيار المستفيد؟
العشوائية	هل تتيح الأداة عشوائية البيانات طبقا لنموذج البناء؟
حذف التسجيلات	هل تتيح الأداة حذف كل التسجيلات أو حذف بعض منها؟
معالجة الفراغات	هل تعالج الأداة الفراغات لتجنب فساد البيانات؟
معالجة واصفات البيانات	هل تقدم الأداة للمستفيد توصيفات للبيانات وأنواعها؟
التغذية المرتدة الناتجة	هل تتيح الأداة باستخراج النتائج من التحليل الداخلي؟

رابعا: معايير الاستخدام: وتستخدم لتتبع معايير الاستخدام وسهولة الاستخدام، ويتم استخدام هذه المعايير لبناء صفات تساهم في استخدام النظام، وتتضمن هذه الصفات منحنى تعلم واجهة المستفيد، أنواع المستفيدين، عرض البيانات وغيرها.

جدول رقم (٤)

معايير الاستخدام لأدوات التنقيب عن البيانات

المعيار	الوصف
واجهة المستفيد	هل تقدم الواجهة النتائج بطريقة مفهومة وواضحة؟
منحنى التعلم	هل الأداة سهل تعلمها؟
أنواع المستفيدين	هل الأداة مصممة للمبتدئين والمتوسطين والمتقدمين من المستفيدين أو دمج بين أنواع المستفيدين؟
رؤية البيانات وعرضها	هل الأداة تمثل البيانات؟
تقارير الخطأ	هل تم الإعلان عن الخطأ بطريقة مفهومة وواضحة؟
تاريخ العمل	هل تحتفظ الأداة بتاريخ الأفعال التي تم اتخاذها في عملية التنقيب؟
تنوع النطاق	هل يمكن للأداة أن تستخدم في مجموعة من التطبيقات والصناعات لحل المشكلات المختلفة

والجدير بالذكر أن هناك العديد من الباحثين والمنظمات الذين قاموا بمراجعة أدوات التنقيب عن البيانات وبعمليات مسحية حول منقبي البيانات، وأنتجت هذه الدارسات مجموعة من حزم البرمجيات التي لها مزاياها وعيوبها، وهذه الدارسات تقع ما بين عام ١٩٩٩م – ٢٠١١م[1] ومن خلال هذه الدراسات تم

(1) Mikut, R., Reischl, M. (September–October 2011). Data Mining Tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (5): 431–445. Available at: http://onlinelibrary.wiley.com/doi/10.1002/widm.24/pdf

- Rexer, K., Allen, H. and Gearan, P. (2011). Understanding Data Miners. =

تقسيم أدوات التنقيب عن البيانات مفتوحة المصدر إلى ما يلي:

١. محررون R IDE/Editors: ومن أمثلتها مشروع آر R[1]، آر ستوديو

=

Analytics Magazine, INFORMS: Institute for Operations Research and the Management Sciences. Available at: http://www.analytics-magazine.org/may-june-2011/320-understanding-data-miners

- Kobielus, J. (Jul. 2008) The Forrester Wave: Predictive Analytics and Data Mining Solutions, Q1 2010, Forrester Research. Available at: https://www.forrester.com/rb/Research/wave™_predictive_analytics_and_data_mining_solutions,/q/id/56077/t/2
- Herschel, G. (2008) Magic Quadrant for Customer Data-Mining Applications, Gartner Inc.. Available at: http://www.gartner.com/technology/research/media_products/overview.jsp
- Nisbet, R. (2006). Data Mining Tools: Which One is Best for CRM? Part 1. Information Management Special Reports. Available at: http://www.information-management.com/specialreports/20060124/1046025-1.html
- Haughton, D., Deichmann,J., Eshghi,A., Sayek,S., Teebagy,N., and Topi, H. (2003). A Review of Software Packages for Data Mining, The American Statistician, 57 (4) pp. 290–309. Available at: http://www.jstor.org/stable/30037299
- Goebel, M., and Gruenwald,L. (1999). A Survey of Data Mining and Knowledge Discovery Software Tools, SIGKDD Explorations, 1 (1): pp.20–33. Available at: https://wwwmatthes.in.tum.de/file/1klx69ggd5riv/Enterprise%202.0%20Tool%20Survey/Paper/A%20survey%20of%20data%20mining%20and%20knowledge%20discovery%20software%20tools.pdf

(1) http://www.r-project.org/

RStudio[١]، وتين آر Tinn-R[٢].

٢. برامج التنقيب عن البيانات Data Mining Software: وهي البرامج التي تقوم عليها الدراسة الحالية، وهي: Weka[٣]، RapidMiner[٤]، KNIME[٥]، Orange[٦]، Rattle[٧]، TANAGRA[٨].

٣. العناقيد Clustering: وهي البرامج التي تقوم بأحد مهام التنقيب عن البيانات فقط وهي العنقدة، مثل: CLUTO[٩]، fastcluster[١٠].

٤. أدوار التجميع Association Rules: والمقصود بها البرامج التي تقوم بأدوار التجميع فقط وما يتعلق بها، مثل: arules[١١]، ARMiner[١٢]، A C++ Frequent Itemset Mining Template Library[١٣]،

(1) https://www.rstudio.com/products/RStudio/
(2) http://sourceforge.net/projects/tinn-r/
(3) http://www.cs.waikato.ac.nz/ml/weka/
(4) https://rapidminer.com/
(5) http://www.knime.org/
(6) http://orange.biolab.si/
(7) http://rattle.togaware.com/
(8) http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
(9) http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview
(10) http://math.stanford.edu/~muellner/fastcluster.html
(11) http://cran.r-project.org/web/packages/arules/index.html
(12) http://www.cs.umb.edu/~laur/ARMiner/
(13) http://www.cs.bme.hu/~bodon/en/fim_env/index.html

Frequent Itemset Mining Implementations Repository[١].

٥. تحليل التسلسل Sequence Analysis: مثل: TraMineR[٢].

٦. تحليل الشبكات الاجتماعية Social Network Analysis: مثل: Gephi[٣]، Pajek[٤]، CFinder[٥].

٧. معالجة التنقيب Process Mining: مثل: ProM[٦].

٨. تحليل البيانات الفضائية: Spatial Data Analysis، مثل: GeoDa[٧]، CLAVIN[٨].

وستحاول هذه الدراسة تحليل وتقييم (٦) ستة أدوات للتنقيب عن البيانات مفتوحة المصدر والتي يمكن الوصول إليها على شبكة الإنترنت، وتم دراسة هذه النظم من أربعة اتجاهات: المعلومات العامة، مصادر البيانات، وظائف التنقيب عن البيانات، إمكانية الاستخدام، وأسفرت نتائج عمليات المسح عن الجداول التالية:

أولا: المعلومات العامة: ويقصد بها بعض الصفات العامة للنظام، منها: الطبعة أو الإصدارة المتاحة، الشركة الراعية ومكانها، الرخصة، واللغة المتاح بها الأداة، إمكانية التحميل، التكلفة، وتحديث البرنامج.

(1) http://fimi.ua.ac.be/src/
(2) http://traminer.unige.ch/
(3) http://gephi.github.io/
(4) http://mrvar.fdv.uni-lj.si/pajek/
(5) http://cfinder.org/
(6) http://www.promtools.org/doku.php
(7) http://geodacenter.asu.edu/projects/opengeoda
(8) https://clavin.bericotechnologies.com/

جدول رقم (٥)

مقارنة بين المعلومات العامة حول أدوات التنقيب عن البيانات مفتوحة المصدر

م	المعيار	RapidMiner	Weka	Orange	KNIME	Rattle	TANAGRA
١.	الطبعة/ الإصدارة	Community edition، version 5.0	٣,٦,٥/ ٣,٧,٤	Version 2.0	desktop edition، version 2.1.2	version 2.5.21	Tanagra 1.4.24
٢.	الشركة/ المنظمة/ الدولة	Rapid-I (Germany)	University of Waikato , New Zealand.	University of Ljubljana (Slovenia)	KNIME.com GmbH (Switzerland)	Togaware (Australia)	Lumière University Lyon - FRANCE
٣.	الرخصة	OSS،[1] GNU AFFERO GPL, v.3[2]	OSS, GNU GPL v.2[3]	OSS, GNU GPL v.3	OSS, GNU GPL v.3	OSS, GNU GPL v.2	OSS, GNU GPL v.2
٤.	اللغة	Java	Java 5.0	C++ Python	Java	R	+C+
٥.	التكلفة	مجاني	مجاني	مجاني	مجاني	مجاني	مجاني
٦.	التحديث	باستمرار	باستمرار	باستمرار	باستمرار	باستمرار	باستمرار
٧.	نظام التشغيل **OS platform**	Windows، Mac OS X، Linux	Windows، Mac OS X، Linux	Windows، Mac OS X، Linux	Windows، Mac OS X، Linux	Windows، Mac OS X، Linux	Windows

(1) New Media Rights. (Dec. 2008) Open Source Licensing Guide. Available at: http://www.newmediarights.org/open_source/new_media_rights_open_source_licensing_guide

(2) Free Software Foundation. (Nov. 2007). Gnu Affero General Public License, V.3.. Available at: http://www.gnu.org/licenses/agpl-3.0.html

(3) Free Software Foundation. Gnu Affero General Public License, V.2.. Available at: http://www.gnu.org/licenses/gpl-2.0.html

ويتضح من الجدول السابق ما يلي:

١. تشابه موقع الدول المنتجة لهذه البرامج وهي قارة أوربا فيما عدا Rattle بأستراليا وWeka بنيوزلندا.

٢. تستخدم الأدوات لغة واحدة للعرض فيما عدا Orange تستخدم لغتين، كما تتشابه ثلاث أدوات في استخدام لغة الجافا وهم: RapidMiner، Weka، KNIME.

٣. بما أن أدوات الدراسة مفتوحة المصدر، لذا فهي تستخدم رخصة البرامج مفتوحة المصدر والرخصة المجانية، كما أن تحميلهم بالمجان.

٤. تدعم معظم الأدوات العديد من نظم التشغيل مثل لينكس وماك والويندوز فيما عدا تانجرا Tangra.

ثانيا: مصدر البيانات (برامج قواعد البيانات): عادة تتوافر تطبيقات البيانات من مصادر مختلفة وبصيغ متنوعة، وتعد إمكانية الوصول لصيغ هذه البيانات من الأهمية بما كان في اختيار نظام المصدر المفتوح، لذا يوضح الجدول التالي مصادر بيانات التي تميز النظم التي يمكن الوصول إليها وإلى مصادر البيانات الخاصة بها، ويرتبط بذلك حجم البيانات لسهولة التعامل مع النظام أو الأداة.

جدول رقم (٦)

مقارنة بين مصادر بيانات أدوات التنقيب عن البيانات مفتوحة المصدر

م	المعيار	RapidMiner	Weka	Orange	Rattle	KNIME	TANAGRA
١.	Oracle	✓	×	×	×	×	×
٢.	Sybase	✓	×	×	×	×	×
٣.	SQLServ	✓	×	×	×	×	×
٤.	MySQL	✓	×	✓	✓	✓	×

م	المعيار	RapidMiner	Weka	Orange	Rattle	KNIME	TANAGRA
٠٥	Access	×	×	×	✓	✓	×
٠٦	ODBC	×	×	×	✓	✓	×
٠٧	JDBC	✓	✓	×	×	✓	×
٠٨	ARFF	✓	✓	×	×	✓	✓
٠٩	CSV	✓	✓	×	✓	✓	×
١٠	Excel	✓	×	×	✓	×	✓
١١	حجم البيانات	متوسط	متوسط	متوسط	كبير	متوسط	متوسط

ويتضح من الجدول السابق ما يلي:

١) شركة أوراكل هي واحدة من أضخم وأهم شركات تقنية المعلومات بشكل عام وقواعد البيانات بشكل خاص. تأسست شركة أوراكل في العام ١٩٧٧ على يد "لاري اليسون" ولدى الشركة مراكز خدمة للعملاء في أكثر من ١٤٥ دولة[١]. وتستخدمها الأداة RapidMiner فقط.

٢) أصبحت سايبيس ثاني نظام لقواعد البيانات بعد أوراكل، وبعد إجراء صفقة مع مايكروسوفت لكى تستطيع مايكروسوفت التسويق على نظام التشغيل OS2 مزود الخدمة في ذلك الوقت، أطلقت على سايبيس خادم قاعدة البيانات "Sybase SQL Server"، حتى الإصدار ٤,٩.[٢] وتستخدمها الأداة RapidMiner فقط.

٣) ميكروسوفت إس كيو إل سيرفر: Microsoft SQL Server هو

(1) ORACLE. Available at: http://www.oracle.com/index.html

(2) Sybase Products. (2014). Available at: http://www.sybaseproducts.com/

برنامج لقواعد البيانات العلائقية من إنتاج مايكروسوفت، لغة الاستعلام الرئيسية فيه هي إس كيو إل وT-SQL[١].وتستخدمها الأداة RapidMiner فقط.

٤) ماي إس كيو إل MySQL هو نظام إدارة قواعد البيانات علائقي يعتمد التعامل معه على لغة إس كيو إل. وسمي بهذا الاسم تبعا لابنة مبرمجه الأصلي Michael Widenius، والتي اسمها My. ماي إس كيو إل هو من المنتجات مفتوحة المصدر ينشر كوده المصدري تحت رخصة جنو العامة بالإضافة إلى بعض الاتفاقيات الاحتكارية، فكانت الشركة الربحية السويدية MySQL AB تمتلكه، وانتقلت الملكية إلى شركة صن ميكروسيستمز وهي فرع من شركة أوراكل.[٢] وتستخدمها أربع أدوات فيما عدا تانجرا وويكا.

٥) مايكروسوفت أكسس (Microsoft Access) هو برنامج لإدارة قواعد البيانات من تطوير شركة مايكروسوفت يأتي البرامج مرافقا لحزم مايكروسوفت أوفيسMicrosoft Office كجزء منها وله واجهة رسومية. كانت هناك عدة إصدارات للبرنامج، فأولها كان مع أوفيس ٩٧ ثم أوفيس ٢٠٠٠ واوفيس ٢٠٠٣ وآخر إصدار هو مع أوفيس ٢٠١٣. يتميز البرنامج بقدرته على استدعاء البيانات من نظم مختلفة لقواعد البيانات، كقواعد بيانات أوراكل وSQL وأي قاعدة بيانات مفتوحة الاتصال (ODBC)[٣]. وتستخدمه أداتان فقط من أدوات الدراسة وهما: Rattle وKnime.

---

(1) Microsoft. SQL Server 2014. Available at: http://www.microsoft.com/en-us/server-cloud/products/sql-server /

(2) Oracle Corporation. MySQL The world's most popular open source database. Available at: https://www.mysql.com/

(3) Microsoft. Access.. Available at: https://products.office.com/en-us/access

٦) رابط قاعدة بيانات مفتوح لميكروسوفت Microsoft Open Database Connectivity (ODBC) هو واجهة للغة برمجة السي التي تجعل من الممكن للتطبيق الوصول لمجموعة من نظم إدارة قواعد البيانات، فهي واجهة عالية الأداء منخفضة المستوى مصممة لمخازن البيانات العلائقية.(١) وتستخدمه أداتين فقط من أدوات الدراسة وهما: Rattle وKnime.

٧) ربط قواعد بيانات الجافا The Java Database Connectivity (JDBC): هي معيار صناعة لربط قواعد البيانات المستقلة بين لغة برمجة الجافا ومجموعة كبيرة من قواعد البيانات. وتتيح هذه التكنولوجيا استخدام لغة برمجة الجافا لإظهار إمكانيات تطبيقات الكتابة مرة واحدة والتشغيل في أي مكان والتي تتطلب الوصول لبيانات الشركة.(٢) وتستخدمه ثلاث أدوات من أدوات الدراسة وهم: RapidMiner، Weka، Knime.

٨) صيغة ملف علاقة الخاصية ARFF (Attribute-Relation File Format): هي عبارة عن ملف نصي بصيغة آسكي يصف قائمة بمجموعة من الخصائص المترابطة، ولقد طوره مشروع تعليم الآلة بقسم علم الحاسوب بجامعة ويكاتو The University of Waikato للاستخدام مع برنامج تعلم الآلة ويكا Weka(٣) تستخدمه أربع أدوات فيما عدا Orange وRattle.

٩) القيم المنفصلة كوما The CSV ("Comma Separated

(1) Microsoft Open Database Connectivity (ODBC). Available at: https://msdn.microsoft.com/en-us/library/ms710252(v=vs.85).aspx
(2) Java SE Technologies - Database. Available at: http://www.oracle.com/technetwork/java/javase/jdbc/index.html
(3) Attribute-Relation File Format (ARFF). (2008). Available at: http://www.cs.waikato.ac.nz/ml/weka/arff.html

"Values): هي صيغة للملفات التي تستخدم في الأغلب لتبادل البيانات بين التطبيقات المتشابهة والمختلفة، وتستخدمها العديد من التطبيقات لتصدير واستيراد البيانات[1]. تستخدمها أربع أدوات فيما عدا Orange وTangra.

١٠) مايكروسوفت أوفيس أكسل: Microsoft Office Excel هو أحد البرامج الموفرة ضمن حزمة أوفيس وهو مخصص للعمليات الحسابية حيث أنه عبارة عن أوراق افتراضية يمكن أضافة معادلات حسابية عليها ومن ثم إضافة الارقام حيث يقوم البرنامج بالعمليات الحسابية بشكل آلي وفي نفس الوقت يمكن أن تستخدم لتخزين البيانات الإليكترونية حيث يمكن الاحتفاظ بها أو طبعها على طلائع ورقية. يستخدمه ثلاث أدوات فقط وهم: **RapidMiner**، وRattle، وTangra، يستخدم برنامج الاكسيل ثلاث أدوات فقط وهم: **RapidMiner**، وRattle، وTangra.

١١) تنفرد الأداة Rattle بتضمينها حجم بيانات كبير مقارنة بالأدوات الأخرى التي تتضمن بيانات متوسطة.

ثالثا: الوظائف: يقصد بمعايير الوظائف أي القيام بوظائف التنقيب عن البيانات لحل المشكلات المختلفة، لذا تعد الوظائف من أهم ملامح نظم التنقيب عن البيانات مفتوحة المصدر، هذا ويمكن تقسيم الوظائف لست مجموعات، وهي: المعالجة الأولية للبيانات، والتصنيف، والتنبؤ، والعنقدة، وقواعد التجميع، والتقييم، والعرض. ولقد تم وضع رقم كهدف لكل منهم، حيث يعد رقم (٣) هو أعلى معدل و(٠) أي لا يدعم هذه الوظيفة، كما تم إضافة تقنيات حديثة أخرى وهي دعم الآلات الموجهة، والزيادة في البيانات، والكم العشوائي، ويتضح كما في

(1) Edoceo, Inc. (2014). Comma Separated Values (CSV) Standard File. Available at: http://edoceo.com/utilitas/csv-file-format

الجدول أن نظم التنقيب عن البيانات الحالية تستخدم بصفة عامة وظائف التنقيب عن البيانات الشائعة.

جدول رقم (٧)

معايير الوظيفة لأدوات التنقيب عن البيانات

م	المعيار	RapidMiner	Weka	Orange	Rattle	KNIME	TANAGRA
١	المعالجة الأولية Preprocess	٣	٣	٣	٢	٣	٣
٢	شبكة Bayes Network	✓	✓	✓	✓	✓	✓
٣	شجرة القرارات Decision Tree	✓	✓	✓	✓	✓	✓
٤	الشبكة العصبية Neural network	✓	✓	×	✓	✓	✓
٥	دعم الآلات الموجهة SVM support vector machines	✓	✓	✓	✓	✓	✓
٦	ملامح الاختيار Feature Selection	✓	✓	✓	✓	✓	✓
٧	العنقدة Clustering	✓	✓	×	✓	✓	✓
٨	قواعد الربط (التجميع) Association Rules	✓	✓	✓	✓	✓	✓
٩	معلومات النموذج Model Information	✓	✓	✓	✓	✓	✓
١٠	التقييم Evaluation	✓	✓	✓	✓	✓	✓

م	المعيار	RapidMiner	Weka	Orange	Rattle	KNIME	TANAGRA
١١	الزيادة في البيانات boosting	✓	✓	✓	✓	✓	✓
١٢	الكم العشوائي random forests	✓	✓	✓	✓	✓	×
١٣	عرض البيانات Data Vis	٣	٣	٣	٣	٣	١
١٤	عرض النموذج Model Vis	٣	٣	٣	٣	٣	١

يتضح من الجدول السابق قيام أدوات الدراسة الست بمعظم وظائف التنقيب عن البيانات فيما عدا الأداة Orange لا تقوم بوظيفتين وهما الشبكة العصبية والعنقدة، أما أداة Tangra فلا تقوم بالكم العشوائي، كما تقوم معظم الأدوات بالمعالجة الأولية وعرض البيانات وعرض النموذج بكفاءة عالية، فيما عدا أداة Tangra تقوم بعرض البيانات والنموذج بشكل ضعيف، أما الأداة Rattle تقوم بالمعالجة الأولية بدرجة متوسطة.

رابعا: الاستخدام: يصف الاستخدام سهولة استخدام نظام التنقيب عن البيانات مفتوح المصدر والذي يمكن استخدامه في حل المشكلات التجارية العالمية وفي بيئات النظام، لذا تم انتقاء تسعة معايير تتعلق بالاستخدام والتي تتمثل في الجدول التالي؛ حيث يشير التفاعل البشري إلى كم التفاعل المطلوب مع عملية الاكتشاف، ومقسمه إلى يدوي وموجه، حيث يدل يدوي على إمكانية القيام بعملية التنقيب عن البيانات دون الحاجة إلى توجيه، أما الموجه أي يوفر النظام مساعدة واضحة أثناء عملية التنقيب؛ وتشير التوافقية إلى مدى دعم النظام لنفسه ذاتيا، أم يعتمد على لغة تكويد النمذجة التنبؤية PMML (Predictive Modeling Markup Language، كما أن هناك العديد من أوجه التشابه والاختلاف من

ناحية استخدام النظم مفتوحة المصدر الخاصة بالتنقيب عن البيانات، وتم وضع رقم (٣) لأعلى معدل، (٢) متوسط، (١) ضعيف.

جدول رقم (٨)

مقارنة بين ملامح النظام لأدوات التنقيب عن البيانات مفتوحة المصدر

م	المعيار	RapidMiner	Weka	Orange	Rattle	KNIME	TANAGRA
١	التفاعل البشري Human Interaction	يدوي	يدوي	يدوي	يدوي	موجه	يدوي
٢	التوافقية Interoperability	ذاتي	ذاتي	ذاتي	PMML	PMML	ذاتي
٣	القدرة على الامتداد Extensibility	ممتاز	ممتاز	ممتاز	بسيط	ممتاز	بسيط
٤	التوثيق Documentation	٢	٣	١	٢	٣	٣
٥	سهل التعلم Easy-to-learn	١	٢	٣	٢	٢	٣
٦	سهولة الاستخدام Usability	٣	٢	٣	٣	٢	٣
٧	الدعم Support	٣	٣	١	١	١	٣
٨	الموثوقية Reliability	٣	٣	٣	٣	٣	٣
٩	التحميل Installation	٣	٣	٣	٢	٣	٣

يتضح من الجدول السابق ما يلي:

١. هناك تفاعل جيد للعنصر البشري؛ حيث تدعم كل الأدوات العنصر البشري أثناء عمليات الاكتشاف، وينفرد نايم KNIME في حاجته إلى دليل إرشادي لتوجيه العنصر البشري لما يجب القيام به؛ وهناك مزية يتصف بها بعض الأدوات والتي تتضح من خلال الاستخدام وهي توفير نموذج السحب والإفلات أثناء عملية التركيب والبناء للتنقيب عن البيانات وهي تتوافر بأربع أدوات

KNIME، Orange، Weka، RapidMiner.

٢. تدعم الأدوات ذاتيا فيما عدا KNIME، Rattle فهما يعتمدان على لغة تكويد النمذجة التنبؤية.

٣. يمكن لمعظم الأدوات الامتداد والتكيف مع الظروف المختلفة للبيانات بدرجة ممتازة فيما عدا Rattle، TANAGRA فيمكنها التكيف بدرجة بسيطة، كما يمكن لمعظم الأدوات القيام بعمليات جديدة في شكل تسجيل Plug-ins.

٤. يحظى ثلاث أدوات بقدرة ممتازة على التوثيق وهم: Weka، وNIME، وTANAGRA. كما تتميز الأداتان Orange و TANAGRA بقدرتهما العالية في سهولة التعلم، في حين تتشابه معظم الأدوات في سهولة استخدامها بدرجة كبيرة فيما عدا الأداتين Weka، وNIME فدرجة الاستخدام متوسطة.

٥. تتميز ثلاث أدوات بدعمهم العالي وهم RapidMiner، Weka، و TANAGRA، في حين تتميز الأدوات الست بقدرتهم العالية على الموثوقية والتحميل.

ومن خلال الجداول السابقة يمكن ذكر **مزايا أدوات التنقيب عن البيانات مفتوحة المصدر وعيوبها** فيما يلي:

١) تعد الأداة Weka من أشهر أدوات برمجيات التنقيب عن البيانات مفتوحة المصدر، والتي تتكون من العديد من مكونات التنقيب عن البيانات والتي تدخل في العديد من الأدوات الأخرى مثل RapidMiner، Rattle، KNIME. وتتكون من أربع تطبيقات رئيسية، وهي المستكشف لاستكشاف البيانات،

والمجرب للقيام بالتجارب والاختبارات الإحصائية، وتدفق المعرفة، وواجهة الأوامر لتنفيذ أوامر الويكا، لذا تتناسب هذه الأداة مع المبتدئين للبداية بمستكشف الويكا، كما توفر واجهة بسيطة سهلة التعلم للوصول لمكونات التنقيب عن البيانات.

٢) تم بناء الأداة RapiderMiner والتي كانت معروفة سابقا YALE على الويكا، وهي تتضمن وظائف قوية إضافية لتحليل البيانات مثل المعالجة الأولية للبيانات وعرضها وخوارزميات تعلم الآلة الإضافية. كما تتميز هذه الأداة بواجهة المستفيد؛ حيث إنها أكثر إبداعا من ويكا، ويمكن للمستفيد الذي لديه خبرة قليلة لعلم الحاسوب والبرمجة تعلم استخدام الأداة بصفة أساسية.

٣) تعد الأداة KNIME من أفضل الأدوات الداخلية التي تدعم الملامح على الخط المباشر والتي تساعد المستفيدين الجدد المنوطين بالقيام بعملية بناء التنقيب عن البيانات، وهي تدعم سكريبت R، Python. كما تتميز بتكاملها مع أداة التطور الكيميائية Chemistry Development Kit مع عقد إضافية لمعالجة الهيكل الكيمائي ومكوناته وغيرها.

٤) توفر الأداة Rattle واجهة مستخدم رسومية مخصصة للتنقيب عن البيانات، وعلى الرغم من فهم الأداة لا يتطلب البدء باستخدامها في الوظائف الأساسية للتنقيب عن البيانات، إلا أنها تتناسب مع المستفيدين المعتادين على R، كما تتكامل الأداة مع أداتين متخصصتين في تحليل البيانات الرسومية التفاعلية، وهما: GGobi، Latticist.

٥) تتميز الأداة Orange ببساطتها والواجهة الرسومية الإبداعية ولا تتطلب إلا معرفة محدودة بالتنقيب عن البيانات، ومقارنة بأدوات التنقيب عن

البيانات الأخرى، تتمثل قوتها في وظيفة العرض التفاعلي التي تتيح للمستفيدين وضع علامات العرض ثم اختيار نقاط البيانات أو العقد مباشرة من الرسوم البيانية.

٦) يمكن لكل باحث الوصول لكود المصدر الخاص بأداة تانجرا TANAGRA، فهي توفر للباحثين والطلاب برنامج تنقيب عن البيانات سهل الاستخدام مع واجهة رسومية وشرح طريقة الاستخدام، كما تتيح إمكانية تحليل البيانات. هذا بالإضافة إلى أن تانجرا تقترح للباحثين هيكل يتيح إضافة طرق التنقيب عن البيانات الخاصة بهم بسهولة، وحتى يمكنهم مقارنة أدائهم.

كما أن هناك بعض العيوب التي اتضحت من خلال الدراسة في هذه الأدوات، وهي:

١. نقص دعم مصادر بيانات مختلفة؛ حيث تدعم الأدوات knime، Rattle، RapidMiner معظم الصيغ وقواعد البيانات العلائقية.

٢. صعوبة التعامل مع الأحجام الكبيرة للبيانات: تستلم التطبيقات العالمية غالبا مجموعات ضخمة من البيانات، بملايين الصفوف، إلا أن معظم نظم التنقيب عن البيانات مفتوحة المصدر من الصعب عرض ومقارنة مجموعة من الخوارزميات على مجموعة صغيرة من مجموعات البيانات، لكنها لا تركز على التعامل مع مجموعات البيانات الكبيرة جدا.

٣. ضعف التوثيق والخدمات: تبين من خلال الدراسة أن معظم الأدوات لديها توثيق ضعيف وكذا الخدمات إذا ما قورنت بالأدوات التجارية، حيث تعتمد الخدمات على المستفيدين ودعمهم.

الخاتمة

يعد التنقيب عن البيانات نوعا حديثا من تكنولوجيا معالجة المعلومات الذكية، ومع الطفرة الهائلة في تكنولوجيا المعلومات، سنشاهد توسعا أفقيا ورأسيا في استخدام تطبيقات التنقيب عن البيانات، وخاصة في التطبيقات العسكرية والأمنية والاستخبارية والتجارية، وسيكون التنقيب عن البيانات على الأجهزة المحمولة الاتجاه المستقبلي للبيانات. لقد طغت قوة الإنترنت على مجموعات البيانات العلمية، والتسلسل الاجتماعي لمجموعة البيانات، والطوبولوجيا والهندسة والخصائص الأخرى بالأخص ربط التنقيب بيانيا، وتحليل الشبكات الاجتماعية. يواجه التنقيب عن البيانات قواعد البيانات الضخمة، لذا يجب أن تكون خوارزمية التنقيب عن البيانات كفؤة وقابلة للتطوير. معظم قواعد البيانات الحالية هي العلائقية؛ لذا يتطلب ظهور نماذج أخرى من قواعد البيانات القدرة على معالجة أنواع البيانات، ويمكن لمتخصصي التنقيب عن البيانات الإسراع من عملية التنقيب عن البيانات، أي يجب توفير واجهة تفاعلية للمستفيدين ملائمة للتعبير عن المتطلبات والاستراتيجيات؛ ومن ناحية أخرى تقوم الواجهة التفاعلية بتحويل النتائج المتنوعة للمستفيد، أي نظام تنقيب عن بيانات يتطلب تفاعلية أقوى. في نفس الوقت نجد أن البحث عن التنقيب عن البيانات قد يؤدي إلى توسع في البيانات غير القانونية وهذا يشكل مشكلة يجب حلها.

وقد توصلت الدراسة إلى العديد من النتائج والتوصيات التي يمكن توضيحها فيما يلي:

أولا: النتائج:

١) يمكن القول إن التنقيب عن البيانات هي اكتشاف المعرفة من البيانات

أو هى التنقيب عن البيانات (أحيانا تسمى اكتشاف المعرفة) هى عملية تحليل البيانات من منظورات مختلفة واستخلاص علاقات بينها وتلخيصها إلى معلومات مفيدة، مثل معلومات يمكن أن تسهم في زيادة الربح، تخفيض التكاليف، أو كليهما معا. أو هو عملية الكشف والعثور عن معلومات ذات فائدة من خلال استعمال مجموعة من الأدوات المعقدة. بعض من هذه الأدوات تشمل أدوات الإحصاء الاعتيادية والذكاء الاصطناعى والرسوم البيانية من صنع الكمبيوتر.

٢) لقد استخدم مصطلح التنقيب عن البيانات في مجال المكتبات والمعلومات للمرة الأولى عام ١٩٩٨م، ونظراً لما حدث من التباس عند الاسترجاع من جانب الباحثين في مجال المكتبات؛ لذلك صك مصطلح آخر لفض هذا الالتباس عام ٢٠٠٣م مصطلح "التنقيب البيليوجرافي Bibliomining" الذي يعنى بتطبيق الأدوات الإحصائية وأدوات التعرف على الأنماط في كم كبير من البيانات المرتبطة بنظم المكتبات من أجل المساعدة في اتخاذ القرارات، أو تبرير الخدمات المقدمة وتطويرها خاصة في المكتبات الرقمية

٣) يقوم التنقيب عن البيانات بعمليتين أساسيتين: التنبؤ: يهدف التنقيب عن البيانات إلى وضع توقعات مع الإحالة للسمة العامة أو سمات الكائن لبيانات التصنيف غير المعروفة. الوصف: يعد نموذج البيانات المحتمل المتاح الذي يلخص العلاقات الدور التوثيقي والتفسيري، يستخدم تحليل العلاقة عادة لوصف نموذج بخصائص علائقية قوية لاشتقاق النماذج المهمة لإيجاد العلاقة بين البيانات.

٤) يوجد سبعة أنواع من التنقيب عن البيانات والتي تتمثل في: تحليل الارتباط، شجرة القرارات، الخوارزميات الجينية، شبكات النظرية الافتراضية،

مسار المجموعة الخام، الشبكة العصبية، التحليل الاحصائى.

٥) يمكن تطبيق التنقيب عن البيانات في العديد من المجالات ومن هذه المجالات: مكاتب الائتمان على القروض، السوبر ماركت، شركات الأدوية، وكالة الاستخبارات، طبيب التحاليل، نظام حجز الطيران، تطبيقات تكنولوجيا المعلومات، البنوك، المكتبات ومراكز المعلومات.

٦) هناك العديد من الدراسات التى تقترح معايير لتقييم أدوات التنقيب عن البيانات سواء التجارية أو مفتوحة المصدر أو المجانية، منها دراسة كين كولير عام ١٩٩٩م Ken Collier وآخرون والتى تم فيها وضع منهجية أولية لتقييم أدوات التنقيب عن البيانات، وتوضح الدراسة أن من خلال الخبرة البحثية تبين أن هناك أربعة تصانيف أساسية لمعايير تقييم أدوات التنقيب عن البيانات: الأداء، والوظيفة، والقدرة على الاستخدام، ودعم الأنشطة الثانوية، وهذه الخبرة تدعمها العديد من الدراسات السابقة

٧) هناك العديد من الباحثين والمنظمات الذين قاموا بمراجعة أدوات التنقيب عن البيانات وبعمليات مسحية حول منقبي البيانات، وأنتجت هذه الدارسات مجموعة من حزم البرمجيات التي لها مزاياها وعيوبها، وهذه الدارسات تقع ما بين عام ١٩٩٩م – ٢٠١١م ومن خلال هذه الدراسات تم تقسيم أدوات التنقيب عن البيانات مفتوحة المصدر إلى ثمانية أقسام: محررون، **برامج التنقيب عن البيانات**، العناقيد، أدوار التجميع، تحليل التسلسل، تحليل الشبكات الاجتماعية، معالجة التنقيب، تحليل البيانات الفضائية

٨) تشابه موقع الدول المنتجة لهذه البرامج وهى قارة أوربا فيما عدا Rattle بأستراليا و Weka بنيوزلندا.

٩) تستخدم الأدوات لغة واحدة للعرض فيما عدا Orange تستخدم لغتين، كما تتشابه ثلاث أدوات في استخدام لغة الجافا وهم: **RapidMiner**، **Weka**، **KNIME**.

١٠) بما أن أدوات الدراسة مفتوحة المصدر، لذا فهي تستخدم رخصة البرامج مفتوحة المصدر والرخصة المجانية، كما أن تحميلهم بالمجان.

١١) تدعم معظم الأدوات العديد من نظم التشغيل مثل لينكس وماك والويندوز فيما عدا تانجرا Tangra.

١٢) تنفرد الأداة **RapidMiner باستخدام ثلاث قواعد** أوراكل وسايبيس وميكروسوفت إس كيو إل سيرفر، تستخدم أربع أدوات ماي إس كيو إل فيما عدا تانجرا وويكا، في حين تستخدم أداتين فقط من أدوات الدراسة وهما: Rattle و Knime قاعدة بيانات الاكسيس على الرغم من سهولة استخدامها، ورابط قاعدة بيانات مفتوح لميكروسوفت. كما تستخدم ثلاث أدوات من أدوات الدراسة وهم: RapidMiner، Weka، Knime ربط قواعد بيانات الجافا، في حين تستخدم أربع أدوات فيما عدا Orange وRattle صيغة ملف علاقة الخاصية، أما صيغة القيم المنفصلة كوما يستخدمها أربع أدوات فيما عدا Orange وTangra؛ في حين يستخدم برنامج الاكسيل ثلاث أدوات فقط وهم: **RapidMiner**، وRattle، وTangra.

١٣) تنفرد الأداة Rattle بتضمينها حجم بيانات كبير مقارنة بالأدوات الأخرى التي تتضمن بيانات متوسطة.

١٤) قيام أدوات الدراسة الست بمعظم وظائف التنقيب عن البيانات فيما عدا الأداة Orange لا تقوم بوظيفتين وهما الشبكة العصبية والعنقدة، أما أداة

Tangra فلا تقوم بالكم العشوائي، كما تقوم معظم الأدوات بالمعالجة الأولية وعرض البيانات وعرض النموذج بكفاءة عالية، فيما عدا أداة Tangra تقوم بعرض البيانات والنموذج بشكل ضعيف، أما الأداة Rattle تقوم بالمعالجة الأولية بدرجة متوسطة.

١٥) هناك تفاعل جيد للعنصر البشري؛ حيث تدعم كل الأدوات العنصر البشري أثناء عمليات الاكتشاف، وينفرد نايم KNIME في حاجته إلى دليل إرشادي لتوجيه العنصر البشري لما يجب القيام به؛ وهناك مزية يتصف بها بعض الأدوات والتي تتضح من خلال الاستخدام وهي توفير نموذج السحب والإفلات أثناء عملية التركيب والبناء للتنقيب عن البيانات وهي تتوافر بأربع أدوات KNIME، Orange، Weka، RapidMiner.

١٦) تدعم الأدوات ذاتيا فيما عدا KNIME، Rattle فهما يعتمدان على لغة تكويد النمذجة التنبؤية.

١٧) يمكن لمعظم الأدوات الامتداد والتكيف مع الظروف المختلفة للبيانات بدرجة ممتازة فيما عدا Rattle، TANAGRA فيمكنها التكيف بدرجة بسيطة.

١٨) تحظى ثلاث أدوات بقدرة ممتازة على التوثيق وهم: Weka، وNIME، وTANAGRA. كما تتميز الأداتين Orange و TANAGRA بقدرتهما العالية في سهولة التعلم، في حين تتشابه معظم الأدوات في سهولة استخدامها بدرجة كبيرة فيما عدا الأداتين Weka، وNIME فدرجة الاستخدام متوسطة.

١٩) تتميز ثلاث أدوات بدعمهم العالي وهم RapidMiner، Weka،

و TANAGRA، في حين تتميز الأدوات الست بقدرتهم العالية على الموثوقية والتحميل.

٢٠) تعد الأداة Weka من أشهر أدوات برمجيات التنقيب عن البيانات مفتوحة المصدر، والتي تتكون من العديد من مكونات التنقيب عن البيانات التي تدخل في العديد من الأدوات الأخرى مثل RapidMiner، Rattle، KNIME. وتتكون من أربع تطبيقات رئيسية، وهي المستكشف لاستكشاف البيانات، والمجرب للقيام بالتجارب والاختبارات الإحصائية، وتدفق المعرفة، وواجهة الأوامر لتنفيذ أوامر الويكا، لذا تتناسب هذه الأداة مع المبتدئين للبداية بمستكشف الويكا، كما توفر واجهة بسيطة سهلة التعلم للوصول لمكونات التنقيب عن البيانات.

٢١) تم بناء الأداة RapiderMiner التي كانت معروفة سابقا YALE على الويكا، وهي تتضمن وظائف قوية إضافية لتحليل البيانات مثل المعالجة الأولية للبيانات وعرضها وخوارزميات تعلم الآلة الإضافية. كما تتميز هذه الأداة بواجهة المستفيد؛ حيث أنها اكثر إبداعا من ويكا، ويمكن للمستفيد الذي لديه خبرة قليلة لعلم الحاسوب والبرمجة تعلم استخدام الأداة بصفة أساسية.

٢٢) تعد الأداة KNIME من أفضل الأدوات الداخلية التي تدعم الملامح على الخط المباشر وتساعد المستفيدين الجدد المنوطين بالقيام بعملية بناء التنقيب عن البيانات، وهي تدعم سكريبت R، Python. كما تتميز بتكاملها مع أداة التطور الكيميائية Chemistry Development Kit مع عقد إضافية لمعالجة الهيكل الكيمائي ومكوناته وغيرها.

٢٣) توفر الأداة Rattle واجهة مستخدم رسومية مخصصة للتنقيب عن

البيانات، وعلى الرغم من فهم الأداة لا يتطلب البدء باستخدامها في الوظائف الأساسية للتنقيب عن البيانات، إلا أنها تتناسب مع المستفيدين المعتادين على R، كما تتكامل الأداة مع أداتين متخصصتين في تحليل البيانات الرسومية التفاعلية، وهما: GGobi، Latticist.

٢٤) تتميز الأداة Orange ببساطتها والواجهة الرسومية الإبداعية ولا تتطلب إلا معرفة محدودة بالتنقيب عن البيانات، ومقارنة بأدوات التنقيب عن البيانات الأخرى، تتمثل قوتها في وظيفة العرض التفاعلي التي تتيح للمستفيدين وضع علامات العرض ثم اختيار نقاط البيانات أو العقد مباشرة من الرسوم البيانية.

٢٥) يمكن لكل باحث الوصول لكود المصدر الخاص بأداة تانجرا TANAGRA، فهي توفر للباحثين والطلاب برنامج تنقيب عن البيانات سهل الاستخدام مع واجهة رسومية وشرح طريقة الاستخدام، كما تتيح إمكانية تحليل البيانات. هذا بالإضافة إلى أن تانجرا تقترح للباحثين هيكل يتيح إضافة طرق التنقيب عن البيانات الخاصة بهم بسهولة، وحتى يمكنهم مقارنة أدائهم.

٢٦) نقص دعم مصادر بيانات مختلفة؛ حيث تدعم الأدوات knime، Rattle، RapidMiner معظم الصيغ وقواعد البيانات العلائقية.

٢٧) صعوبة التعامل مع الأحجام الكبيرة للبيانات: تستلم التطبيقات العالمية غالبا مجموعات ضخمة من البيانات، بملايين الصفوف، إلا أن معظم نظم التنقيب عن البيانات مفتوحة المصدر من الصعب عرض ومقارنة مجموعة من الخوارزميات على مجموعة صغيرة من مجموعات البيانات، لكنها لا تركز على التعامل مع مجموعات البيانات الكبيرة جدا.

٢٨) ضعف التوثيق والخدمات: تبين من خلال الدراسة أن معظم الأدوات

لديها توثيق ضعيف وكذا الخدمات إذا ما قورنت بالأدوات التجارية، حيث تعتمد الخدمات على المستفيدين ودعمهم.

ثانيا: التوصيات:

١) عقد العديد من الندوات والمؤتمرات فيما يتعلق بالتنقيب عن البيانات في مجال علوم المكتبات والمعلومات، مع توضيح أوجه الاستفادة طبقا للتطورات التكنولوجية الحادثة في المجال.

٢) القيام بالعديد من الدراسات حول التنقيب عن البيانات في المكتبات فيما يتعلق بالمشكلات والتحديات التي تواجه تطبيقها، والعوامل المساعدة في تطبيق التنقيب عن البيانات في المكتبات، وما هي البرامج والأدوات التي تتناسب معها.

٣) ضرورة استخدام التنقيب عن البيانات في العديد من مجالات علوم المكتبات والمعلومات لسهولة تنظيم المعلومات والوصول إلى المعرفة بسهولة في ظل التضخم الهائل في البيانات.

قائمة المراجع

1. Agrawal,R., Imielinski,T. and Swami,A. Database Mining: A Performance Perspective. Available at: http://www.rakesh.agrawal-family.com/papers/tkde93mining.pdf

2. Attribute-Relation File Format (ARFF). (2008). Available at: http://www.cs.waikato.ac.nz/ml/weka/arff.html

3. Cios, K. and Kurgan, L. Trends in Data Mining and Knowledge Discovery. Available at: http://www.cioslab.vcu.edu/Publications/Papers/chapterTrendsDM2003.pdf

4. Data Mining Tasks. Available at: http://wideskills.com/data-mining/data-mining-tasks

5. Durkin,J. and Jingfeng,C. (2005) CAI Zixing. Decision Tree Technology And Its Current Research Direction[J].Control Engineering.12 (1).

6. Edoceo, Inc. (2014). Comma Separated Values (CSV) Standard File. Available at: http://edoceo.com/utilitas/csv-file-format

7. Eltabakh,M. (2010). OLAP & Data Mining. Available at: http://web.cs.wpi.edu/~cs561/s12/Lectures/IntegrationOLAP/OLAPandMining.pdf

8. Fayyad,U., Piatesky-Shapiro,G., Smyth, P., and Uthurusamy, R., (1996) Advances in Knowledge Discovery

and Data Mining, AAAi/MIT Press, 1996. Available at: https://www.google.com.eg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0CCgQFjABahUKEwjlloCr-tjGAhVJOBQKHUkpBTM&url=https%3A%2F%2Fmitpress.mit.edu%2Fbooks%2Fadvances-knowledge-discovery-and-data-mining&ei=yiCkVeWJN8nwUMnSlJgD&usg=AFQjCNHeKXUmv7YO5vM2g_UcfrHrg6hsEw&bvm=bv.97653015,d.ZGU

9. Ferguson, M. Evaluating And Selecting Data Mining Tools. InfoDB, 11 (2): pp: 1-10.- Available at: http://www.evaltech.com/admin/upload/Evaluating_Data_Mining_Tools.pdf

10. Flockharta, I. and Radclieab, N. (1996) A Genetic Algorithm Based Approach to Data Mining Presented at "AAAI: Knowledge Discovery and Data Mining", Portland, Oregon. Available at: http://www.stochasticsolutions.com/pdf/kdd96.pdf

11. Free Software Foundation. (Nov. 2007). Gnu Affero General Public License, V.3.. Available at: http://www.gnu.org/licenses/agpl-3.0.html

12. Free Software Foundation. Gnu Affero General Public License, V.2.. Available at: http://www.gnu.org/licenses/gpl-2.0.html

13. Friedman, J. Data Mining and Statistics: what's the Connection? Available at:

http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf

14. Goebel, M., and Gruenwald,L. (1999). A Survey of Data Mining and Knowledge Discovery Software Tools, SIGKDD Explorations, 1 (1): pp.20–33. Available at: https://wwwmatthes.in.tum.de/file/1klx69ggd5riv/Enterprise%202.0%20Tool%20Survey/Paper/A%20survey%20of%20data%20mining%20and%20knowledge%20discovery%20software%20tools.pdf

15. Han, J. and Kamber,M. Data Mining: Concepts and Techniques. Available at: https://cs.wmich.edu/~yang/teach/cs595/han/ch01.pdf

16. Haughton, D., Deichmann,J., Eshghi,A., Sayek,S., Teebagy,N., and Topi, H. (2003). A Review of Software Packages for Data Mining, The American Statistician, 57 (4) pp. 290–309. Available at: http://www.jstor.org/stable/30037299

17. Heckerman,D. (1997) Bayesian Networks for Data Mining Data Mining and Knowledge Discovery,1: 79–119. Available at: http://machinelearning101.pbworks.com/f/Tutorial-BayesianNetworks.pdf

18. Herschel, G. (2008) Magic Quadrant for Customer Data-Mining Applications, Gartner Inc.. Available at: http://www.gartner.com/technology/research/media_products/overview.jsp

19. Jadhav, A., and Sonar, R. (2011). Framework for evaluating and selection of the software packages: A hybrid

knowledge based system approach. the journal of system and software, 84 (8): pp1394-1407. Available at: http://romisatriawahono.net/lecture/dm/paper/classification/Jadhav%20-%20Framework%20for%20evaluation%20and%20selection%20of%20the%20software%20packages%20-%202011.pdf

20. Java SE Technologies - Database. Available at: http://www.oracle.com/technetwork/java/javase/jdbc/index.html

21. Jensen, D. and Neville,J. Correlation and Sampling in Relational Data Mining. Available at: https://www.cs.purdue.edu/homes/neville/papers/jensen-neville-interf2001.pdf

22. Knitting and Crochet Patterns. (2015). Pattern Discovery In Data Mining Available at: http://ktuliuepatt.com/pattern-discovery-in-data-mining/

23. Kobielus, J. (Jul. 2008) The Forrester Wave: Predictive Analytics and Data Mining Solutions, Q1 2010, Forrester Research. Available at: https://www.forrester.com/rb/Research/wave™_predictive_analytics_and_data_mining_solutions,/q/id/56077/t/2

24. Microsoft Open Database Connectivity (ODBC). Available at: https://msdn.microsoft.com/en-us/library/ms710252(v=vs.85).aspx

25. Microsoft. Access. Available at: https://products.office.com/en-us/access

26. Microsoft. SQL Server 2014. Available at: http://www.microsoft.com/en-us/server-cloud/products/sql-server/

27. Mikut, R., Reischl, M. (September–October 2011). Data Mining Tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (5): 431–445. Available at: http://onlinelibrary.wiley.com/doi/10.1002/widm.24/pdf

28. Netz, A., Chaudhuri,S. Bernhardt, J. and Fayyad, U. (2000) Integration of Data Mining and Relational Databases Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt. Available at: http://121.241.184.234/trddc_website/PastTecsweeks/2007/integration-of-data-mining.pdf

29. New Media Rights. (Dec. 2008) Open Source Licensing Guide. Available at: http://www.newmediarights.org/open_source/new_media_rights_open_source_licensing_guide

30. Nisbet, R. (2006). Data Mining Tools: Which One is Best for CRM? Part 1. Information Management Special Reports. Available at: http://www.information-management.com/specialreports/20060124/1046025-1.html

31. Oracle Corporation. MySQL The world's most popular open source database. Available at: https://www.mysql.com/

32. ORACLE. Available at:

http://www.oracle.com/index.html

33. Padhy,N. Mishra,P. and Panigrahi, R. (June 2012) The Survey of Data Mining Applications And Feature Scope International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2 (3). Available at:

http://arxiv.org/ftp/arxiv/papers/1211/1211.5723.pdf

34. Pawlak, Z. Rough Sets And Data Mining. Available at: http://bcpw.bg.pw.edu.pl/Content/1884/RSDMEAK.pdf

35. Piatesky-Shapiro,G. (Jan. 1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop, AI Magazine, 11: 5, pp. 68-70. Available at: https://www.google.com.eg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CCEQFjAAahUKEwiWh9nL-djGAhVMPhQKHQUjCDk&url=https%3A%2F%2Fwww.aaai.org%2Fojs%2Findex.php%2Faimagazine%2Farticle%2Fdownload%2F873%2F791&ei=AyCkVdb5Acz8UIXGoMgD&usg=AFQjCNFB_-Qrs8RdlFxNlNI9jhuh61eiXw&bvm=bv.97653015,d.ZGU

36. Piatesky-Shapiro,G. (Jan. 1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop, AI Magazine, 11: 5, pp. 68-70. Available at: https://www.google.com.eg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CCEQFjAAahUKEwiWh9nL-djGAhVMPhQKHQUjCDk&url=https%3A%2F%2Fwww.a

aai.org%2Fojs%2Findex.php%2Faimagazine%2Farticle%2Fdownload%2F873%2F791&ei=AyCkVdb5Acz8UIXGoMgD&usg=AFQjCNFB_-Qrs8RdlFxNlNI9jhuh61eiXw&bvm=bv.97653015,d.ZGU

37. Piatetsky -Shapiro,G., Brachman,R., Khabaza,T., Kloesgen,W. and Simoudis,E. (1996) An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications. Proceedings KDD -96, AAAI Press, Portland, Oregon August 2 -4, 1996. Available at: http://aaaipress.org/Papers/KDD/1996/KDD96-015.pdf

38. Piatetsky-Shapiro, G., and Frawley,W., (Eds) (1991). Knowledge Discovery in Databases, AAAI/MIT Press, 1991. Available at: http://aaaipress.org/ojs/index.php/aimagazine/article/viewFile/1011/929

39. Rexer, K., Allen, H. and Gearan, P. (2011). Understanding Data Miners. Analytics Magazine, INFORMS: Institute for Operations Research and the Management Sciences. Available at: http://www.analytics-magazine.org/may-june-2011/320-understanding-data-miners

40. Sharma, P., Bhartiya, R. (Dec. 2012) Implementation of Decision Tree Algorithm to Analysis the Performance International Journal of Advanced Research in Computer and Communication Engineering, 1(10). Available at: http://www.ijarcce.com/upload/december/24-Implementation%20of%20Decision.pdf

41. SINGH,Y. and Chauhan, A. Neural Networks In Data Mining Journal of Theoretical and Applied Information Technology, 5 (6). Available at: http://jatit.org/volumes/research-papers/Vol5No1/1Vol5No6.pdf

42. Sybase Products. (2014). Available at: http://www.sybaseproducts.com/

43. Thuraisingham,B.(2000). A Primer for Understanding and Applying Data Mining. IT Pro IEEE Xplore. Available at: https://www.utdallas.edu/~bxt043000/Publications/Journal-Papers/DS-DM/J71_A_Primer_for_Understanding_and_Applying_Data_Mining.pdf

44. Verma, V. and Dhawan, S. (May 2014) Methodology for Selection of a Data Mining Tool. International Journal of Software & Hardware Research in Engineering, 2 (5): pp. 189- 192. Available at: http://ijournals.in/ijshre/wp-content/uploads/2014/05/IJSHRE-2552.pdf

45. Wang,J. (2008). Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications. Information Science reference. Available at: http://www-db.deis.unibo.it/~srizzi/PDF/isr08-1.pdf

46. Weiping,F. and Yuming,W. (Dec. 2013) The Development of Data Mining International Journal of Business and Social Science, 4 (16). Available at: http://ijbssnet.com/journals/Vol_4_No_16_December_2013

/14.pdf

47. Zhao,Y., Chen,Y. and Yao,Y. (2006). User-Centered Interactive data Mining. Proceedings of the Sixth IEEE International Conference on Cognitive Informatics (ICCI'06): 457-466. Available at: http://www2.cs.uregina.ca/~yanzhao/icci06.pdf

٤٨. **وسام محمود أحمد درويش**. نحو رؤية جديدة لإدارة المكتبات باستخدام تقنية التنقيب عن البيانات.- cybrarians journal.- ع ١٩ (يونيو ٢٠٠٩)

http://www.journal.cybrarians.org/index.php?option=com_content&view=article&id=437:-data-mining-&catid=164:2009-05-20-10-02-29&Itemid=60

Data mining tools open source
Analytical evaluation study

Dr.. Ahmad Faiz Ahmed Sayed

Information Libraries and Information Technology teacher

Faculty of Arts and Sciences humanity, Suez Canal University, Egypt

Abstract

There are tools and software accompany with data mining to help in the exploration for and the huge amount of data increasing access to knowledge in different databases, and facilitate these tools work on most of the scientific disciplines, including the library and information science. Therefore, this study aims to study the nature of data mining, functions, applications and analysis and evaluation tools of open source data, and then make a comparison between prospecting for open source data tools. The study found many results and the most important results are: Four tools are characterized with drag and drop form: KNIME, Orange, Weka, RapidMiner.

Key words:

Data Mining, Web Mining, Information Mining, Bibliomininig, KNIME, Orange, Weka, RapidMiner, Tangra, Rattle.