مدخل إلي علم تنقيب البيانات: 1-5

إن علم تنقيب البيانات علم مستقل بحد ذاته، كما يري بعض العلماء أنه مرتبط بعلم جلب المعلومات Information Retrieval ،لأنه يعتبر أقدم بكثير من علم الData mining ،ويعتبر علم تنقيب البيانات هو العلم المسئول عن أساليب وطرق إنتاج المعلومات الهامة وقواعد المعرفة من خلال كم كبير من البيانات المراد التنقيب فيها وربطها بأساليب علمية لنخرج بمعلومة أو معرفة مفيدة وجديدة.

والسبب الرئيسي في ظهور علم تنقيب البيانات هو وفرة البيانات داخل المؤسسات والشركات من حولنا،كما أن أدوات تنقيب البيانات لها مقدرات تحليلة عالية لإجراء التحليل العميق مثل التجميع والتصنيف ،وقامت الData mining بدمج وتكامل التقنيات المتقدمة ومن عدة مجالات مثل مخزن البيانات ،وتمييزالبيانات ،والشبكات العصبية ،وتحليل البيانات المكانية.

وقدم هذا العلم فائدة كبيرة في استغلال مصادر نظام الحاسب الالي مثل الذاكرة وسعات الأقراص التشغيلية.

وتوفر أنظمة التنقيب بيانات واضحة لتمكن المستخدمين من تعريف إحتياجاتهم وإكتشاف وإكتساب المعرفة.

2-5 تعريف تنقيب البيانات:-

هي عملية بحث يدوي أو محوسب عن معرفة داخل البيانات دون فرضيات مسبقة عن هذه المعرفة، وهي عبارة عن تقنية تستخدم لإستخراج المعرفة والمعلومات العالية المستوي أو المهمة من قاعدة بيانات عملت مثل:-

Data ware house

Working with data mining

ونستطيع من خلال تنقيب البيانات الوصول إلي المعلومات الهامة التي نحتاجها في صنع القرارات الذكية، وذلك لحل مجموعة من المشاكل وإيجاد الحلول لها.

وهي تخدم الجميع وكل المؤسسات مليئة بالبيانات الكبير جداً لكن لايوجد معلومات وهذا أدى إلى ظهور الData mining لتقدم معلومات مفيدة للإستفاد منها في إتخاذ القرارات المناسبة لكي تساعد في عمليات التطوير والنماء وحل المسائل الصعبة.

وتعمل الData mining على إستخراج المعرفة المخفية داخل مستودع البيانات ،وتساعد المنظمات والشركات بالتركيز على المعلومات الهامة والتي تعتمد علي سلوك البشر.

1-2-5أنواع تنقيب البيانات:

1. التنقيب الإشرافي : وينتج عنه نموذج عن النظام الذي تصفه البيانات المستخدمة في عملية التنقيب.
2. التنقيب الوصفي : أما هذا النوع فينتج عنه معلومات جديدة بناء علي المعلومات المستخدمة في عملية التنقيب.

5-2-2 التقنيات الأساسية لتنقيب البيانات:

1. التصنيف :-

وهي مجموعة من التقنيات التي تستخدم في تصنيف البيانات وتوزيع الخصائص المتعلة بكل صنف.

1. التجميع او(العنقدة):-

وهي تقوم بتقسيم قاعد البيانات إلي مجموعة من القطع أوالعناقيد ذات السجلات المتشابه والتي تشترك في عدد من الخصائص لتحسين الدقة وإكتشاف التجانس في مجتمع قواعد البيانات.

1. نماذج الاعتماد:-

والغرض منها إيجاد التعلق والإرتباط الموجود بين عدد من السجلات المنفردة أو مجموعة من السجلات في قاعدة البيانات ، وتستخدم لدراسة سلوك الزبائن خلال فترة زمنية طويلة.

1. الكشف عن التغيرات والانحرافات:-

وتعتبر مصدر للإكتشاف الصحيح ،ويمكن ايجادها بإستخدام التقنيات الإحصائية مثل الإنحدار الخطي ويمكن عمل تمثيل رسومي وعرض مخططات والتي تجعل عملية اكتشاف الانحراف أسهل.

1. نماذج التنبؤ:-

وهي تشبه خبرة الإنسان في التعلمُ وذلك من خلال الملاحظة لصياغة نماذج لخصائص هامة لبعض الظواهر النادرة،وهي تستخدم مجموعه من قواعد البيانات التاريخية الموجودة لكي نصل إلي البيانات الهامة أو المعرفة.

ملاحظة:-

تعتبر الData mining أشمل وأوسع من طرق التحليل الإحصائي في عملية بناء النماذج ومعالجة البيانات للوصول إلي المعرفة، لأنها تعتمد علي البيانات التاريخية دون فرضيات مسبقة.

3-2-5 أهداف تنقيب البيانات:

1. التعليل عن بعض الظواهر المرئية مثال التعليل عن سبب زيادة نسبة المدخنين في الوطن العربي.
2. إثبات نظرية ما مثال إثبات النظرية التي تقول بأن الأسر الكبيرة تهتم بالضمان الصحي أكثر من الأسر الصغيرة العدد.
3. تحليل البيانات للحصول علي علاقات جديدة وغير متوقعة مثال كيف يكون الإنفاق العام إن كان ملازما لعمليات خداع واسعة من قبل البطاقات الائتمانية.

4-2-5 الوسائل المستخدمة في تنقيب البيانات:-

هناك عدة وسائل مستخدمة من أجل تنقيب البيانات ، ويعتمد اختيار الوسيلة المناسبة علي طبيعة البيانات تحت الدراسة وعلي حجمها، ومن هذه الوسائل المستخدمة في تنقيب البيانات هي:-

1. التفكير واستخلاص النتائج والقوانين من أمثلة حية Case-Based reasoning.
2. الكشف عن قانون Rule Discovery: البحث عن علاقة معين في جزئية كبير من البيانات.
3. معالجة الإشارات Signal Processing :ايجاد الظواهر المشابهه بعضها مع بعض.
4. الشبكات العصبية Neural Network: وتستخدم في تطوير نماذج قابلة لتنبؤ النتائج، وهذه النماذج تم استباطها من عقل الانسان.
5. منحنيات غير ثابتة Fractals : تقوم بتغير البيانات الكبيرة من دون ضياع للمعلومات .

5-2-5 تطبيقات تنقيب البيانات:-

وسائل تنقيب البيانات تستعمل وبنجاح في الكثير من التطبيقات الحقيقية حول العالم منها:-

1. كتابة تقرير مختصر عن فئة معينة ويستخدم في تطوير وإنشاء تقارير مؤجزة عن الزبائن المهمين وعن بطاقات الائتمان.
2. التسويق لفئة معينة لإيجاد الزبائن وذلك لمنح التخفيضات لهم لسبب معين.
3. تحليل الإستعمال وذلك لإيجاد علاقة معينة لإستخدام السلع والخدمات.
4. فعالية الحملة وذلك بمقارنة الحملات بعضها مع بعض وإيجاد الأكثر فعالية.
5. جاذبية السلعة وذلك لإيجاد السلع التي تباع بعضها مع بعض.
6. الحوادث والأمراض التي يتعرض لها الأشخاص، وحالتهم الصحية، أو بوجهٍ عام تحليل السجل الصحي الخاص بكل شخص

3-5الأسباب التي أدت إلي نمو تطبيقات تنقيب البيانات:-

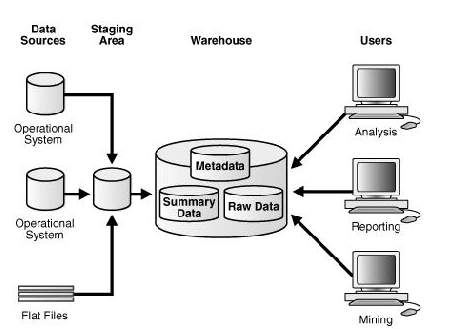
1. كمية البيانات الكبيرة الموجودة في مخزن البيانات وسوق البيانات ينمو بصور كبيرة ومن أجل ذلك فإن المستخدم يحتاج إلي أدوات متطورة مثل تنقيب البيانات من أجل إستخلاص الفائدة والمعرفة من هذه البيانات.
2. الكتير من أدوات تنقيب البيانات بدأت تظهر مؤخراً ، وكل أداة أفضل من الاخري.
3. المنافسة الشديدة الموجودة في السوق تدفع الشركات إلي الإستفادة القصوى من البيانات التي بيدها، وعمليات تنقيب البيانات تفعل ذلك تماماً.

5-4 مفهوم مستودعات البيانات (Datawerehouse):

عبارة عن "مجموعة من بيانات دائمة تاريخية متكاملة تساعد في اتخاذ القرارات الإدارية ، للمساعد في الوصول إلى البيانات لأغراض التحليلات الزمنية واكتشاف المعرفة واتخاذ القرارات، فهي مصممة خصيصاً لإستخراج واستخلاص البيانات ومعالجتها وتقديمها وتمثيلها في صورة مناسبة لهذا الغرض، وتتضمن كميات ضخمة من البيانات تكون من مصادر مختلفة، أو من عدة قواعد مختلفة من أنظمة وأماكن مختلفة.

كما يعرفها البعض الآخر :-

"على أنها نوع من [قواعد البيانات](http://ar.wikipedia.org/wiki/ÙÙØ§Ø¹Ø¯_Ø§ÙØ¨ÙØ§ÙØ§Øª) تحوي كماً هائلاً من البيانات الموجهة للمساعدة على اتخاد القرارات داخل المؤسسة، وتتميز تلك النوعية من قواعد البيانات بتطابق بنيتها الداخلية مع ما يحتاجة المستخدم من مؤشرات ومحاور التحليل.



**شكل رقم (5-1):مستودع البيانات**

وقد برزت أهمية مستودعات البيانات مع وجود الشركات الكبرى؛ حيث أن كل قسم من تلك الشركات يدير قواعد بيانات خاصة به (تسويق، مالي، إدارة... الخ) والتي تتضمن كثيراً من البيانات المشتركة لذا تتطلب الحاجة إلى إيجاد بيانات متعلقة بموضوع محدد مثل العملاء على سبيل المثال من خلال البحث في قاعدة بيانات واحدة بدلاً من البحث في القواعد المختلفة، لذلك كان من الضروري تجميع البيانات في قاعدة واحدة تسمى مخزن البيانات مع الاحتفاظ بالقواعد الخاصة على أن تتم عملية تغذية دورية للمخازن إذا حدث أى تعديل أو تغير وتبديل في القواعد مع تصنيفها حسب الموضوع وذلك باستخدام برامج خاصة تسمى Middleware تستعمل بيانات دمج خاصة في كل موضوع تسمى Schema

integration يحدد فيها شكل وكيفية نقل البيانات.

ويمكن القول بأن مستودعات البيانات هي عبارة عن قاعدة بيانات كبيرة وضخمة يمكن من خلالها الوصول إلى جميع المعلومات المتعلقة بمؤسسة أو شركة ما، ويودع في هذا المستودع البيانات الواردة من قواعد البيانات المختلفة كما يحتوي على العديد من قواعد البيانات وعلى معلومات من مصادر مختلفة، وتلك البيانات تكون في أشكال مختلفة ويمكن الوصول إليها من خلال الكمبيوتر الخادم، ويكون الوصول للمستودع غير مرئي بالنسبة للمستخدم الذي يستعمل أوامر بسيطة في التحليل واسترجاع وجمع البيانات، ويحتوى على بيانات تشير إلى كيفية تنظيمه، وأماكن تلك المعلومات، وهي تستخدم لمساعدة الإدارة في كيفية اتخاذ القرارات داخل الشركات أو المؤسسات، كما تسمح للمؤسسة في تنظيم وتحديث وتنسيق تلك البيانات، وإظهار العلاقات بين المعلومات المجمعة من الإدارت المختلفة بالمؤسسة.

5-5 فائدة تنقيب البيانات في المجال الصحي:

تحتاج معظم المراكز الطبية الهامة إلى إجراء عمليات البحث في سجلات المرضى الطبية وتحليلها. ولعل من أهم أعمال التحليل المطلوبة، تحديد الحوادث والأمراض التي يتعرض لها الأشخاص، وحالتهم الصحية، أو بوجهٍ عام تحليل السجل الصحي الخاص بكل شخص. ولقد ازدادت التطبيقات التي يمكنها الاستفادة من هذا النوع من التحاليل. يمكن أن نذكر في هذا السياق أمثلة أهمها: تشخيص الأمراض، والتخطيط للعمليات الجراحية، واستقطاب الموارد البشرية اللازمة لإجراء الاختبارات الصحية، والتحقق من تطبيق الدليل الصحي، وتقييم مقاييس الجودة، وتحسين أداء الأبحاث الطبية.

يعتمد التشخيص الطبي على العديد من المعايير، كوجود عوارض معينة، أو وجود نتائج تحليل مختبر طبي، أو وجود عادة معينة (كالتدخين أو شرب الكحول مثلاً)، وغيرها. ويُحصل على هذه المعلومات عادة من السجل الطبي الخاص بالمريض، الذي يخزَّن في بعض الحالات ضمن قواعد بيانات بنيوية، لكنها تكون موجودة في معظم الحالات على شكل بيانات غير بنيوية، كالتقارير المخبرية، أو ملاحظات الطبيب، أو الصور الشعاعية، أو الكثير من المستندات الطبية الموجودة في مراكز العناية الصحية.

ونحن من هذا الجانب قمنا بالإستفادة من مصطلح data wearhouse الخاص بتنقيب البيانات في انشاء مخزن للبيانات يحتوي على السجلات الطبية الخاصة بالمرضي الذين يعانون من مرض الفشل الكلوي ،نسبة لانتشار هذا المرض بشكل كبير ،والتوصل الي الاسباب التي تؤدي الي الاصابة بهذا المرض والتوصل الي حلول للقضاء عليها والحد من انتشار هذا المرض.