الفصل الخامس

تنقيب البيانات

مع كثرة البيانات الموجودة والمخزنة في ما يسمى بقواعد البيانات (Data base) ، أصبحت موضع تساؤل من عديد من الباحثين للاستفادة منها، ومع زيادة انتشار مستودعات التخزين الضخمة ما يدعى (Data Warehouses)، أصبح من الضروري إيجاد تقنيات وطرق ووسائل لاستخلاص المعلومات والمعرفة من مثل هذه البيانات المكدسة واستغلالها في حل المشاكل واتخاذ القرارات، باستخدام تطبيقات الحاسوب الحديثة والتي تعتبر تكنولوجيا حديثة ذكية قائمة على جعل الحاسوب ((يفكر كما يفكر الإنسان ويفعل كما يفعل الإنسان)) وهو ما يعرف بالذكاء الاصطناعي ، جاءت فكرة الكشف والتنقيب على هذه البيانات بطرق ذكية للمساعدة في حل المشاكل واتخاذ القرارات. تنقيب البيانات عملية تحليلية تمزج بين علم الذكاء الاصطناعي والإحصاء وتعلم الآلة وقواعد البيانات. وتعتبر خطوة من خطوات استكشاف المعرفة من قواعد البيانات **(محمد، 2003: 53)**.

اولاً: تاريخ تنقيب البيانات

تعتبر الخمسينيات بداية تقنيات التنقيب ، والتي قام بذلك مجموعة من علماء الرياضيات والمنطق والحاسوب. بمحاولات وضع أسس الذكاء الاصطناعي، حتى تبلورت الفكرة تماما" عام 1956 من قبل ((جون مكارثي ومار فن منسكاي)) وذلك في معهد ماسا تشوسش للتكنولوجيا وجامعة كارنيجي ميلون وستانفورد، وفي بداية الستينيات استمرت هذه التطبيقات مع مزج الذكاء الاصطناعي بتطبيقات اخرى سواء اكان في النظريات التطبيقية الرياضية أو الجبرية ، ثم استمر التسلسل التاريخي حتى وقتنا الحاضر. والجدول التالي يمثل التطور التاريخي لتقنيات وتطبيقات تنقيب البيانات **(سلمان، 2010: 5)**.

جدول (5)
التطور التاريخي لتقنيات وتطبيقات تنقيب البيانات

الفترة الزمنية	التقنيات والتطبيقات لتنقيب البيانات
s1950	بروز الذكاء الاصطناعي من قبل جون مكارثي وهربرت سيمون وآخرون
s1960	* ظهور قواعد البيانات **استخدام الذكاء الاصطناعي وتطبيقاته
s1970	*قواعد بيانات نموذجية **ظهور النظم الخبيرة ونجاحها في القواعد التجارية
s1980	*بدايات وجود مستودعات البيانات **انتعاش تطبيقات الذكاء الاصطناعي
s1990	*التنقيب عن البيانات *قواعد البيانات الهدفية *قواعد البيانات الموزعة *ظهور خدمات الانترنيت
s2000	*ظهور تطبيقات الشبكات العصبية *متاجر البيانات التنقيب في الشبكة Web Mining
المصدر: سلمان، عبد الستار شاكر (2010) التنقيب واعادة هندسة قواعد البيانات، المؤتمر السادس لجمعية المكتبات والمعلومات السعودية بعنوان البيئة المعلومات الامنة: المفاهيم والتشريعات والتطبيقات، ص5.	

ويمكن توضيح اهم الاسباب التي ادت الى نمو تنقيب البيانات من الآتي:

1. كمية البيانات الموجودة في مخزن البيانات وسوق البيانات تنمو بصورة سريعة (Exponential).ومن أجل ذلك، فإن المستخدم يحتاج إلى أدوات متطورة مثل التنقيب في البيانات من أجل استخلاص الفائدة والمعرفة من هذه البيانات.
2. الكثير من أدوات التنقيب عن البيانات بدأت تظهر مؤخرا"، وكل أداة أفضل من الأخرى.
3. المنافسة الشديدة الموجودة في السوق تدفع الشركات إلى الاستفادة القصوى من البيانات التي بيدها. عمليات التنقيب في البيانات تفعل ذلك تماما".

ثانياً: مفهوم تنقيب البيانات واهميته

لقد اختلف الباحثين والكتاب في تعريفهم لمفهوم تنقيب البيانات وذلك حسب توجهات كل منهم ووجهات النظر الخاصة بهم ويمكن تحديد بعض من هذه التعاريف

وكما ياتي: " عملية بحث محوسب ويدوي عن معرفة من البيانات دون فرضيات مسبقة عما يمكن أن تكون هذه المعرفة " **(نورين، 2011: 1)**. كما يعرف التنقيب في البيانات على أنه : "عملية تحليل كمية بيانات عادة ما تكون كمية كبيرة لإيجاد علاقة منطقية تلخص البيانات بطريقة جديدة نكون مفهومة ومفيدة لصاحب البيانات" **(محمد، 2003: 54)**. ومن جهة اخرى عرف بانه: " عملية استخراج المعرفة من كمية كبيرة من البيانات (مستودعات البيانات) وذلك باستخدام أساليب جديدة سواء كانت الخوارزميات أم أساليب الذكاء الاصطناعي أم شجرة القرارات وبمراحل عدة" **(سلمان، 2010: 4)**. ويمكن تحديد ابرز انواع التنقيب بالاتي:

1. **التنقيب الوصفي:** يعتمد على إعادة تنظيم البيانات، والتنقيب فيها لاستخراج النماذج الموجودة منها، كتشابه الزبائن الذي يسمح لك بإنشاء وصف بسيط عن مجموعة زبائن متشابهين، ولا يستوجب وجود هدف لمثل هذه البيانات.
2. **التنقيب التنبؤي:** يحاول إيجاد أفضل التنبؤات اعتمادا" على المعطبات، كمعرفة المنتج الأفضل لزبون معين، باختصار يعتمد هذا التنقيب على استخدام المعلومات القديمة لتوقع ما سيحدث في المستقبل وتكون لدى مثل هذه البيانات هدف.

اما اهمية التنقيب في قواعد البيانات فيمكن ابرازها من خلال الاتي **(نورين، 2011: 1)**، (محمد، 2003: 55-56)**:**

1. إن التنقيب في قواعد البيانات يهدف إلى استخلاص المعلومات المخبأة فيها، وهي تكنولوجيا حديثة فرضت نفسها بقوة في عصر المعلوماتية وفي ظل التطور التكنولوجي الكبير وانتشار استخدام قواعد البيانات واستخدامها يوفر للمؤسسات وأجهزة الأمن في جميع المجالات القدرة على استكشاف، والتركيز على أهم المعلومات في قواعد البيانات.
2. تركز تقنيات التنقيب على بناء التنبؤات المستقبلية واستكشاف السلوك والاتجاهات، مما يسمح بتقدير القرارات الصحيحة واتخاذها في الوقت المناسب. تجيب تقنيات التنقيب على العديد من الأسئلة، وفي وقت قياسي، بخاصة تلك النوعية من الأسئلة التي كان من الصعب الإجابة عليها.
3. استخدامها يوفر للمؤسسات وأجهزة الأمن في جميع المجالات القدرة على استكشاف المعلومات، والتركيز على أهم المعلومات في قواعد البيانات.

ثالثاً: أهداف تنقيب البيانات

يمكن تحديد ابرز أهداف تنقيب البيانات بالاتي **(نورين، 2011: 5)**:

1. تعليل بعض الظواهر المرئية.
2. التثبت من نظرية ما. مثال: التثبت من النظرية التي تقول بأن الأسر الكبيرة تهتم بالضمان الصحي أكثر من الأسر الصغيرة عددا".
3. تحليل البيانات للحصول على علاقات جديدة وغير متوقعة.مثال: كيف سيكون الأنفاق العام إن كان ملازما" لعمليات خداع واسعة من قبل البطاقات الانتمائية.

رابعاً: مهام تنقيب البيانات

هناك عدد من المهام التي يتم انجازها من خلال تنقيب البيانات يمكن توضيحها بالآتي:

1. **التصنيف:** يستخدم التصنيف بشكل واسع في حل الكثير من المشكلات من خلال تحليل مجموعة من البيانات ووضعها على شكل أصناف أو أقسام يمكن أستخدمها فيما بعد لتصنيف البيانات المستقبلية، وهناك عدد من الطرق التي يمكن استخدمها في تصنيف الشبكات العصبية، والعديد من خوارزميات تعلم الآلة.
2. **العنقدة:** وهي عملية تقسيم البيانات إلى مجموعة من الأصناف اعتمادا" على اشتراكها بالخواص المتشابهة وأن العنقدة هي تقسيم غير موجه للبيانات وهي عكس التصنيف، كما أنها تساعد المستفيد على فهم التركيب الطبيعي للمجموعات من البيانات.
3. **قاعدة الارتباط:** هي أحدى التقنيات الواعدة منData Mining كأداة من أدوات اكتشاف المعرفة KDD ولديها القدرة على تصفح كميات هائلة من البيانات، وهي تسمح بالتقاط كل القوانين الممكنة التي تشرح بعض الصفات الموجودة اعتمادا" على وجود الصفات الأخرى، وبمعنى آخر هي قواعد ارتباطيه معينة بين مجموعة من البيانات في قاعدة البيانات.
4. **التحليل التسلسل:** في هذه الطريقة يتم البحث لاكتشاف نماذج تحدث بالتسلسل إذ تكون المدخلات عبارة عن بيانات تشكل مجموعة متسلسلة وكل سلسلة من البيانات هي قائمة منظمة من العمليات أو المصطلحات وعندما تكون العملية عبارة عن مجموعات من المصطلحات لابد أن يحسب معها الوقت المصاحب لكل عملية. ولكن مشكلة هذا النموذج تكمن في إيجاد كل النماذج المتسلسلة مع أقل دعم يخصصه المستفيد عندما يكون الدعم لهذا النموذج هو نسبة تسلسل البيانات التي يتضمنها النموذج.

خامساً: أدوات تنقيب البيانات

أدوات تنقيب البيانات هي عبارة عن ادوات السوفت وير والتي تقوم باستخدامها للاستفسار عن المعلومات في قواعد البيانات، ويمكن تحديد ابرز انواع ادوات تنقيب البيانات من خلال الآتي **(نورين، 2011: 2-3)**:

1. **خوارزمية الجار الأقرب:** تعتبر من تقنيات التنقيب في البيانات التي تهدف للتنبؤ عن طريق مقارنة السجلات الشبيهة بالسجل المراد التنبؤ له وتقدير القيمة المجهولة لهذا السجل بناء على معلومات لتلك السجلات.

كثيرا" ما تستخدم خوارزمية الجار الأقرب في مجال الأعمال ومن الأمثلة الشائعة الاستخدام تلك التي تساعد المستخدمين في الشراء عن طريق اختيار السلع الأقرب لاحتياجاتهم مقارنة بسلع قد تم شراؤها بالفعل، في المجال الأمني، يمكن استخدام هذه الخوارزمية للكشف عن مرتكبي جريمة ما وذلك بأن يتم استخدام المعلومات الخاصة بالجرائم الشبيهة التي تم ارتكابها سابقا" بهدف تحديد هوية مرتكب الجريمة الحالية عن طريق تحديد عدد من السجلات التجريبية تم استخدامها بهدف التنبؤ بالقيمة المطلوبة.

2. **التحليل بالتجزئة العنقودية:** هي عملية تجميع السجلات المتشابهة في مجموعات، ويتم ذلك بهدف الاستكشاف العالي المستوى لما يجري داخل قاعدة البيانات. تهدف التجزئة العنقودية الى وضع العناصر المتجانسة في مجموعات منفصلة، القاعدة العامة لضم أي عنصر في مجموعة هي أن يكون العنصر مائلا" للتشابه بعنصر منها أكثر من أن يكون شبيها" لعنصر من مجموعة أخرى.
3. **شجر القرار:** شجرة القرار هي نموذج استكشافي يظهر على شكل شجرة، كما يعبر اسمها، وبشكل دقيق يمثل كل فرع من فروعها سؤالا" تصنيفيا" وتمثل أوراقها أجزاءا" من قاعدة البيانات تنتمي للتصنيفات التي تم بنائها. ومن المهم جدا" عند بناء خوارزمية شجرة القرار إن يؤخذ بعين الاعتبار إن تكون قابلة للتطبيق بقدر الامكان وبشكل مثالي على كل البيانات المتوفرة، القاعدة الأساسية في بناء شجرة القرار هي إيجاد أفضل سؤال عند كل فرع من فروع الشجرة بحيث يقسم هذا السؤال الى قسمين، القسم الأول منها ينطبق عليهم السؤال والقسم الثاني لا ينطبق، وهكذا يتم من خلال سلسلة من الأسئلة بناء شجرة القرار بفروعها المتسلسلة. تهدف تقنية شجرة القرار إلى تقسيم قاعدة البيانات بهدف معين سبق وأن تم تحديده. وجود عنصر معين في إحدى المجموعات.
4. **الشبكات العصبونية:** تعتبر الشبكات العصبونية هي وأشجار القرار من أهم تقنيات التنقيب في البيانات، نظرا" للنتائج الدقيقة التي يتم التوصل إليها باستخدام هذه الخوارزميات ولإمكانية تطبيقهما في حل العديد من المشاكل وبكافة الأنواع. خوارزمية الشبكة العصبونية تشبه في تركيبتها تركيبة مخ الإنسان، فهي تعمل بنفس الطريقة كما يعمل المخ في نقل ومعالجة المعلومات والتوصل إلى الاستنتاجات واكتشاف الأنماط والتنبؤات ونستطيع من خلالها تطبيق بعض ما يطبقه المخ الطبيعي.

5. **استقراء القاعدة:** أن تقنية استقراء القاعدة من التقنيات الأساسية في التنقيب في البيانات وأكثرها شيوعا" في مجال استكشاف المعرفة، وهي أقرب ماتكون إلى مايسمى بعملية التنقيب بحد ذاتها، والذهب في هذه الحالة هو (القاعدة) تبين هذه القاعدة مايجري داخل قاعدة البيانات وتظهر لنا مالم نكن نعرفه من قبل، وربما أيضا" ما لن نستطيع أن نعرفه إلا من خلالها.
6. **ادوات الاستفسار واعداد التقارير:** وهي الادوات التي تستخدم من اجل توليد الاستفسارات والتقارير البسيطة من خلال استخدام المعلومات الموجودة في مستودع البيانات.
7. **وكلاء الذكاء:** يقوم وكلاء الذكاء باستخدام ادوات الذكاء الاصطناعي المتنوعة من اجل اكتشاف المعلومات وبناء ذكاء الاعمال.
8. **ادوات التحليل المتعددة الابعاد:** وهي عبارة عن تقنيات التي تسمح لك باستعراض المعلومات المتعددة الابعاد من منظور مختلف.
9. **الادوات الاحصائية:** تساعد في تطبيق النماذج الرياضية المتنوعة على المعلومات المخزونة في مستودع البيانات وذلك لغرض اكتشاف المعلومات الجديدة.

اختيار تقنية التنقيب المناسبة: لا توجد نظرية محددة يتم بناء عليها اختيار تقنية من تقنيات التنقيب، ويتم الاختيار عادة بناءا" على الخبرة في هذا المجال والتجربة الفعلية للتقنيات ومدى فاعليتها، ومن جهة أخرى قد تكون المفاضلة أيضا" بين التقنيات التقليدية والتقنيات الحديثة بقدر مايكون هناك توفرا" للأدوات المناسبة، ومع ازدياد الخبرة نستطيع أن نقيم الخيارات ونحدد منها المناسب ونطبقه.

سادساً: تخطيط عمليات التنقيب في قواعد البيانات:

إن تخطيط عمليات التنقيب في قواعد البيانات من الأمور المهمة للحصول على أفضل النتائج، فالتخطيط الجيد يؤدي للنتائج الجيدة. ويمكن تلخيص الخطوات الأولية للتنقيب في قواعد البيانات في ما يلي **(نورين، 2011: 4)**:

1. تحديد المشكلة المراد بحثها وإيجاد الحلول لها.
2. بناء قاعدة بيانات التنقيب.
3. استكشاف البيانات.
4. تحضير البيانات للتنقيب.
5. بناء نموذج التنقيب المناسب.
6. تطبيق النموذج.
7. استخراج النتائج.

سابعاً: خطوات تنقيب البيانات

يمكن تحديد ابرز خطوات تنقيب البيانات بالاتي **(محمد، 2003: 56-58)**:

1. **اختيار البيانات:** في هذه المرحلة يتم تحديد واسترجاع البيانات من مجموعة البيانات.
2. **تهيئة البيانات:** عملية التنقيب هي عبارة عن عملية معالجة ولكن تهيئة البيانات هي قلب المعالجة فالبيانية هي المادة الخام المراد تحليلها وفي هذه المرحلة يتم تجهيز وعزل البيانات التي تحتوي على بيانات مبهمة او مفقودة من مجموعة البيانات وهناك بعض الطرق للقيام بها:
   - **تنظيف البيانات:** فالبيانات ليست دائما" خالية من العيوب أو من فقدان بعضها وقد تحوي بيانات قديمة غير صحيحة في الوقت الحالي أو ربما كانت أخطاء مطبعية.
   - **البيانات المفقودة:** تنقيب البيانات يتطلب أن تكون هذه البيانات كاملة وإذا لم تتحقق هذه فمن الصعب الوصول إلى نتائج مفيدة.
   - **اشتقاق البيانات:** أحيانا" في عمليات التهيئة قد يحدث اشتقاق بعض الأعمدة مما يساعد في الحصول على بيانات مفيدة.
   - **دمج البيانات:** أيضا" بعض الأوقات نحتاج إلى دمج بعض الأعمدة للحصول على نتائج أفضل وذلك يرجع إلى المشرف على هذه البيانات.
3. **دراسة البيانات:** تتبلور هذه المرحلة في دراسة البيانات من حيث طريقتها أو مدى استجابتها للتعلم فإما أن تكون إشرافية أو غير إشرافية.
4. **قراءة البيانات وبناء النموذج:** أي نموذج يلخص كمية كبيرة من البيانات وذلك بتجميع المؤشرات مثل:
   - **الترددات:** إظهار كم في الغالب نسبة التأكد من القيمة المحدثة.
   - **الوزن أو التأثير:** يبين كيف أن بعض المدخلات المشار إليها تعطي المخرج المطلوب.
   - **الارتباطات:** بعض المدخلات لها أكثر من وزن معا".
   - **التمايز أو التفاضل:** تبين مدى أهمية معايير المدخلات لمخرج واحد ،كذلك تحديد مدى معين بين الأوزان.
5. **فهم النموذج:** نستطيع أن نقوم بفهم بعض النماذج عن طريق تطبيق بعض التقنيات مثل أشجار القرار أو الشبكات العصبية.

6. **تقييم النموذج**: بعد استخراج النماذج المهمة والتي تمثل المعرفة يتم تقييمها استنادا الى مقاييس محددة.
7. **تمثيل المعرفة:** وهي المرحلة الاخيرة من مراحل تنقيب البيانات وهي المرحلة التي يراها المستفيد وهذه الاساسية تستخدم الاسلوب المرئي لمساعدة المستفيد في فهم وتفسير نتائج استخراج البيانات.

اسئلة الفصل

س1: ناقش المراحل التاريخية لتطور تنقيب البيانات؟
س2: وضع مفهوم تنقيب البيانات، ومصادر اهيمتها في المنظمات المعاصرة؟
س3: هناك عدد من الاهداف التي تسعى المنظمات الى تحقيقها عند استخدام ادوات تنقيب البيانات، عددها؟
س4: وضع اهم مهام تنقيب البيانات، مع تعداد اهم الادوات المستخدمة في ذلك ؟
س5: ناقش اهم خطوات التخطيط في تنقيب البيانات؟

ملاحظة: عزز اجابتك بالاشكال البيانية اينما تجد ضرورة لذلك.