مركز البحوث والدراسات متعدد التخصصات

Multi Disciplines Research and Studies Center

التحليل المتقدم وتنقيب البيانات

Advanced Analysis and Data Mining

د. م. مصطفى فؤاد عبيد

www.mdrscenter.com

مركز البحوث والدراسات متعدد التخصصات
Multi Disciplines Research and Studies Center

التحليل المتقدم وتنقيب البيانات

Advanced Analysis and

Data Mining

د. م. مصطفى فؤاد عبيد

النسخة الإلكترونية المجانية
غزة – فلسطين – ٢٠٢٦

www.mdrscenter.com

بسم الله الرحمن الرحيم

اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ

وَقُل رَّبِّ زِدْنِي عِلْمًا

وَمَا أُوتِيتُمْ مِنْ الْعِلْمِ إِلاَّ قَلِيلاً

ولَئِن شَكَرْتُمْ لَأَزِيدَنَّكُمْ

المقدمة

يعتمد الإنسان على البحث العلمي المنظّم بهدف حل المشكلات بكافة أنواعها والمضي قدمًا في مسيرة التطوير والتحديث والتنمية في شتى المجالات الحيوية في المجتمع، وذلك من خلال استكشاف المعرفة التي من شأنها المساهمة في وضع حلول للمشكلات على أسس علمية ودعم اتخاذ القرار المناسب في الوقت المناسب ومنع وقوع المشكلات المستقبلية، والابتكار والإبداع في شتى المجالات.

وبالرغم من التطور التكنولوجي الكبير الذي حدث في العقود السابقة والمزايا التي وفّرتها الأدوات التكنولوجية الحديثة للبحث العلمي، والتي تساهم بشكل كبير في تيسير الإجراءات البحثية وتوفر أساليب إلكترونية سريعة لحفظ وأرشفة وتحليل البيانات الخام في البحوث بكافة أشكالها، إلاّ أنه ساهم، ربما بشكل أكبر، في نشوء معضلة الحاجة إلى إجراء المزيد من البحوث في كل مجتمع من أجل الاستمرار في مسيرة التطوير والتنمية فيه، وذلك نظراً لتضخم وتشعّب المشكلات وتداخل وتشابك الكثير من التخصصات العلمية والإنسانية، علاوة على تضخّم البيانات الخام التي تنمو بشكل أسّي في هذا العصر ويتم حفظها وأرشفتها وتداولها يومياً.

ومع أن بعض أساليب البحث العلمي الكلاسيكية، بخاصة تلك التي تتميز بالأسلوب الاستقرائي، كانت تفتقر في كثير من الأحيان إلى الثقة في النتائج

والقدرة على التعميم بسبب محدودية البيانات التي يتم دراستها باعتبارها عينات إحصائية لا تُعبر عن المجموع الكلي بشكل دقيق، إلاّ أن هذا الأمر سيبدو معكوسًا في أساليب البحث العلمي التي يمكن أن يُطلق عليها أساليب البحث العلمي العصري في هذا العصر، والتي نتطرق لدراسة وتحليل كميات هائلة من البيانات تفوق ما كان يتم دراسته في العينات الإحصائية في البحوث الكلاسيكية، بحيث ستزداد الثقة في النتائج ويصبح التعميم وكأنه حاصل ضمنيًا في نتائج البحث نظرًا لضخامة البيانات التي يتم دراستها مقارنة مع ما كان يتم دراسته من عينات إحصائية صغيرة نسبيًا في البحوث الكلاسيكية.

فالبحث العلمي الكلاسيكي الذي يهدف لاستكشاف مستوى التحصيل الدراسي لدى طلاب المرحلة الابتدائية في أحد المجتمعات مثلًا قد يلجأ فيه الباحث لاختيار مجموعة من الطلاب في بعض المدارس لقياس تحصيلهم الدراسي باعتبارهم عينة إحصائية تمثل نسبة محددة من المجموع العام للطلاب في تلك المرحلة في المجتمع، ومن ثم يقوم بتعميم النتائج على كل الطلاب مع تحديد نسبة الثقة في تلك النتائج ومدى قابليتها للتعميم وذلك بحسب منهجية البحث التي تم اتباعها. أما في هذا العصر فإنه يكفي استخدام اختبار موحد لقياس مستوى التحصيل الدراسي لدى جميع الطلاب في تلك المرحلة وفي كل المدارس ومن ثم جمع وتحليل نتائج هذا الاختبار بسهولة ويُسر وبلمح البصر باستخدام الأدوات التكنولوجية الحديثة، وبطبيعة الحال سوف يأخذ هذا البحث طابعًا جديدًا ويُصبح وكأنه قد تحوّل إلى دراسة مسحية تتمتع

بالصدق التام للنتائج وقابليتها للتعميم، التي تمت بالفعل، بل والأكثر من ذلك أنه ربما تُصبح قابليتها للتعميم واردة في مجتمعات أخرى شبيهة من حيث العوامل التي تمت دراستها في اختبار قياس التحصيل الدراسي للطلاب في هذا المجتمع!

هذا الكتاب يمكن اعتباره مدخلاً مبسطًا وخطوة من الخطوات اللازمة لتأسيس منهجية جديدة في البحث العلمي تستند إلى استخدام خوارزميات وتقنيات تنقيب البيانات، هذا العلم الذي نشأ حديثًا وبات يتطور مع مفرزات هذا العصر وأدواته التكنولوجية الحديثة التي ساهمت في حفظ وأرشفة كميات هائلة من البيانات الخام، ليس باعتبارها عينات إحصائية قد تؤدي لنتائج منقوصة أو يفتقر معها الباحث للقدرة على التعميم، بل باعتبارها بيانات شاملة ومتكاملة ومفصّلة في كل المجالات الحيوية في المجتمع، وبطريقة تؤسس لحل المشكلات والتطوير والتحديث والتنمية الإستراتيجية المستدامة بعيدة المدى، وفقًا لما توفره من ثقة في النتائج وقابلية للتعميم لم تكن متوفرة سابقًا في معظم البحوث الكلاسيكية.

وباعتبار أن هذا الكتاب هو الأول من نوعه في المجتمع العربي باللغة العربية فإنه يمكن أن يُشكل محاولة لسد الفجوة التي تكوّنت على مر عقود من الزمن بين المجتمع العربي والدول المتقدمة في مجال البحث العلمي بشكل خاص، كما يمكن اعتباره وسيلة تهدف لربط أساليب وإجراءات التحليل المتقدم وتنقيب البيانات والبحث العلمي العصري باللغة العربية نظرًا لما تتمتع

به هذه اللغة من مزايا خفية تساهم في تعزيز القدرات المنطقية والتحليلية لدى المتحدثين بها، تلك القدرات التي تُشكل الأساس القوي، اللازم والضروري، لرفع كفاءة البحث العلمي بشكل عام في المجتمع العربي.

من جهة أخرى، فإنه يمكن أن يُشكل هذا الكتاب حلقة وصل بين مجتمع البحث العلمي الأكاديمي وبين البحث العلمي المهني وتطبيقاته العملية في المجالات الحيوية المختلفة في المجتمع من خلال ما يقدمه من مفاهيم ونظريات وأمثلة توضيحية مبسطة لأحدث أساليب التحليل المتقدم وخوارزميات تنقيب البيانات وتطبيقاتها في العديد من المجالات باعتبارها أحد أهم الأدوات المستخدمة في استكشاف المعرفة المخبأة في قواعد البيانات وكُل البيانات الكبيرة، بما يمكن أن يساهم في سد الفجوة الحادثة بين مجتمع البحث العلمي الأكاديمي والتطبيق العملي له على أرض الواقع واللحاق بركب التطور والتقدم العلمي والتكنولوجي العالمي من منظور بحثي عصري يفتقد له المجتمع العربي.

وقد تم استعراض المادة العلمية المكونة لهذا الكتاب بما يتناسب مع أحدث البرامج الأكاديمية المتخصصة في تحليل وتنقيب البيانات والمعتمدة في المعاهد والجامعات في معظم دول العالم، بالإضافة لأحدث ما توصل إليه هذا العلم من الناحية العملية التطبيقية، بحيث تم الجمع بين المنظور البحثي الأكاديمي والمنظور التطبيقي العملي بشكل تكاملي، وهو بذلك يمكن أن يكون وسيلة فعّالة للعلماء والباحثين المتخصصين، الأكاديميين والمهنيين، من جهة

والمحللين والخبراء والإداريين في مستويات الإدارة الوسطى التنفيذية والإدارة العليا الإستراتيجية من جهة أخرى، بالإضافة لكافة المهتمين بعلم تنقيب البيانات، بحيث يمكن الاسترشاد به كمدخل إلى هذا العلم.

كما أنه يمكن أن يكون مقدمة وافية ومفيدة للدارسين والمتخصصين في مجالات البحث العلمي والإحصاء وعلم المعلومات وتحليل وهندسة البيانات وعلوم الإدارة وإدارة الأعمال والتسويق والتأمين وتحليل وإدارة المخاطر، وعلم الحاسوب والذكاء الاصطناعي، بحيث يمكن أن يساعدهم في توضيح المفاهيم الأساسية لعلم تنقيب البيانات والتطبيقات المختلفة له في المجتمع والاستخدامات المتعددة له في جميع المجالات، بالإضافة للاطلاع على مسارات بحوث التنقيب المختلفة بطريقة تساعدهم في اتخاذ القرار المناسب عند اختيار التخصص الدقيق الذي يتوافق مع اهتماماتهم وتفضيلاتهم.

وقد تطلب إعداد هذا الكتاب بهذا الشكل، ليلبي احتياجات الفئات المتعددة من القرّاء، التطرق لبعض المفاهيم والأساسيات المتعلقة بجوانب مختلفة تعزز من فهم خوارزميات التنقيب، كمبادئ الإحصاء وتحليل البيانات وسمات وخصائص البيانات والمتغيرات وقواعد البيانات، وبطريقة تساهم في تبسيط المحتوى وتجعل منه مادة سهلة ومتكاملة لغير المتخصصين. كما شكّل هذا الكتاب تحديًا كبيرًا، باعتباره الأول من نوعه في المجتمع العربي الذي يتناول موضوع تنقيب البيانات بشكل شامل ومعمّق وباللغة العربية، بالإضافة لكونه يتعلق بمادة علمية سريعة التطور وتزداد توسعًا وتشعبًا يوميًا،

لتغطي المزيد من جوانب الحياة اليومية المعاصرة من منظور بحثي أكاديمي وتطبيقي عملي في جميع المجالات.

ونظراً لأن علم تنقيب البيانات يتم تطبيقه بشكل أساسي على البيانات المخزنة في قواعد البيانات ومستودعات البيانات الضخمة، فقد تم تناول هذه الموضوعات بشيء من التفصيل وإضافتها كملاحق ضمن هذه الطبعة من الكتاب، "أساسيات قواعد البيانات" و"أساسيات مستودعات البيانات"، وذلك نظراً لأهميتها ولضرورة فهمها بالشكل الصحيح والمناسب للمتخصصين في تحليل وتنقيب البيانات ومن وجهة نظر تحليلية ربما تختلف بشكل أو بآخر عن وجهات نظر المتخصصين في مجالات أخرى.

نأمل أن يلبي هذا الكتاب احتياجات القرّاء من مختلف التخصصات والخبرات وأن يكون وسيلة سهلة وشيّقة ومبسّطة لهم، وأن يكون خطوة أولى من خطوات لاحقة في المستقبل لإصدار المزيد من الكتب والطبعات المنقحة الموسعة والمطورة بما يتناسب مع التطورات التي تحدث يومياً في هذا المجال.

أتمنى من الله العزيز القدير أن أكون قد وُفقت في هذا العمل، كما أتمنى أن يوفقنا جميعًا لما فيه الخير لهذه الأمة.

والله ولي التوفيق،

د. م. مصطفى فؤاد عبيد

إلى الباحثين والعلماء المتخصصين في البحث العلمي

يمكن اعتبار هذا الكتاب مدخلاً إلى ما يمكن أن يُطلق عليه "البحث العلمي العصري" الذي يعتمد في تحليل البيانات على أساليب تكنولوجية عصرية تلبي احتياجات الباحثين المهنيين المتخصصين في العديد من المجالات المتنوعة، العلوم والطب والصيدلة والمعلوماتية الحيوية والهندسة والتجارة والصناعة والزراعة والمجالات الإنتاجية والخدمية وغيرها من المجالات الحيوية في المجتمع، كما يمكن أن يكون وسيلة جيدة للباحثين الأكاديميين وطلاب الدراسات العليا في المعاهد والجامعات في جميع التخصصات، نظراً لما يحتويه من مفاهيم ونظريات وأفكار مفتاحية جديدة وعصرية لطرق وأساليب تحليل وتجزئة وتصنيف البيانات والتنبؤ واستكشاف واختبار الأنماط وقواعد التبعية والارتباط بين المتغيرات في البحوث الأكاديمية من منظور أكاديمي نظري ومهني تطبيقي، بطريقة تساهم في سد الفجوة الحادثة في المجتمع البحثي العربي بين النظرية والتطبيق.

إلى أصحاب المشروعات والمدراء التنفيذيين ومدراء الإدارة العليا

يغطي هذا الكتاب موضوعات متنوعة في التحليل المتقدم وتنقيب البيانات والتي تُعتبر العصب الرئيسي لعلم استخبارات الأعمال (ذكاء الأعمال) والمستخدمة بشكل أكثر في نظم الإدارة الذكية الحديثة في العديد من المجالات الخدمية والإنتاجية، ويمكن للقارئ أن يركز على الموضوعات الأكثر أهمية بالنسبة له من حيث مجال تخصصه في العمل، حيث يقدم الكتاب

الأفكار المفتاحية التي يمكن أن يتم تجييرها بشكل سلس لخدمة أهداف كل من الإدارة الإستراتيجية العليا والبحوث والتخطيط في المؤسسات والشركات والإدارة التنفيذية الوسطى فيها، والتي تتضمن إدارة الشئون المالية والإدارية والتسويق والمبيعات والعلاقات العامة وتحليل وإدارة المخاطر والإحصاء والمعلومات، سواء من خلال التطبيق المباشر في حال توفر المهارة اللازمة للتطبيق باستخدام الحاسوب أو بالطلب من المتخصصين من ذوي المهارة والكفاءة في استخدامه، كأن يطلب صاحب مطعم أو مديره التنفيذي مثلاً من الموظف المتخصص في الحاسوب أن يقوم بتحليل وتنقيب بيانات فواتير المطعم التاريخية من أجل استكشاف الطبق الإضافي المفضل للزبائن والذي يتناولونه بشكل أكثر مع الوجبات الأساسية والذي يطلق عليه "الطبق الجاذب".

إلى المحللين وخبراء تحليل وتنقيب البيانات

ربما يُشكل هذا الكتاب مجرد مدخلاً مبسطًا بالنسبة لعلماء الحاسوب والرياضيات وعلماء وخبراء المعلومات وتحليل وتنقيب البيانات وذكاء الأعمال بشكل عام، مقارنة بالمقررات الأكاديمية التي درسوها في المعاهد والجامعات باللغات الأجنبية المختلفة، وبالرغم من ذلك فإنه قد يُعتبر ذات أهمية كبيرة بالنسبة لهم كونه الكتاب الأول من نوعه باللغة العربية، بحيث يساهم في تبسيط وتوضيح طبيعة هذا العلم الحديث نسبيًا والأساليب والأدوات المستخدمة فيه والفوائد التي يوفرها في المجالات المتنوعة مع ربطها

بالمجتمع المهني العملي من خلال الأمثلة التوضيحية التطبيقية الشيقة، وهو ما يجعل منه وسيلة فعّالة تساهم في خلق التواصل والفهم المتبادل بين المتخصصين في هذا المجال وغير المتخصصين، سواء كان تواصلاً أفقياً بين الزملاء في نفس المستويات الإدارية بهدف تبادل الأفكار والرؤى حول أساليب التحليل والتنقيب المختلفة وأهدافها، أو تواصلاً رأسياً بينهم وبين رؤسائهم في المستويات الإدارية الأعلى بالشكل الذي يسهل لهم إمكانية تقديم أفكار جديدة والتعبير عنها بشكل يناسب بيئة العمل واحتياجاتها التطبيقية، في الواقع، إذا كان القارئ غير المتخصص يمكن أن يقتني نسخة واحدة من هذا الكتاب فإن القارئ المتخصص الذي يعمل في هذا المجال قد يفكر بشراء نسختين يقدم إحداها هدية لرئيسه في العمل من أجل تحقيق أفضل أساليب التواصل الإداري الفعّال المبني على أسس من الفهم المتبادل لطبيعة هذا التخصص والفوائد الكبرى التي يمكنه تحقيقها لصالح العمل.

إلى المدربين المتخصصين

يُغطي هذا الكتاب موضوع التحليل المتقدم وتنقيب البيانات بشكل عام، ويمكن أن يكون مادة تدريبية شاملة ومتكاملة لدورات تدريب متخصصة في هذا المجال، وبمستويات متعددة، كما يمكن تقسيمه إلى عدة أجزاء بحيث يمكن استخدام كل جزء منه كمادة تدريبية مصغّرة تستهدف فئات محددة من المتدربين، مثلاً كأن يتم استخدام فصل تنقيب واستكشاف الأنماط وقواعد التبعية والارتباط كمادة تدريبية منفصلة للمتخصصين في مجال

التسويق، أو استخدام فصل التنقيب باستخدام خوارزميات التصنيف والتنبؤ كمادة تدريبية للمتخصصين في مجال البحوث الطبية والمعلوماتية الحيوية أو للمتخصصين في إدارة وتحليل المخاطر في البنوك وشركات التأمين والاتصالات. ومع ذلك فإن التطبيق العملي باستخدام الحاسوب يمكن أن يتنوع بحسب تفضيلات كل مدرب، سواء كان يفضل استخدام برنامج الجداول الإلكترونية Excel، الذي يمكن اعتباره بيئة نموذجية لمعظم الأمثلة والتطبيقات الواردة في هذا الكتاب، أو باستخدام البرمجيات الأكثر تعقيدًا والمتخصصة في تنقيب البيانات واستخبارات الأعمال Data Mining & Business Intelligence Software، وبشكل عام يعتمد اختيار التطبيق المناسب، إضافة لتفضيلات كل مدرب، على طبيعة الدورة التدريبية وتخصص المتدربين ونوعية خوارزميات التنقيب التي سوف يتم التركيز عليها وشرحها بشكل تطبيقي في دورة التدريب.

إلى طلاب الجامعات والمعاهد المتخصصة

تم استخدام الأمثلة التوضيحية والرسوم والأشكال في هذا الكتاب بقدر الإمكان من أجل تبسيط فهم المادة العلمية وتقديمها من منظور أكاديمي نظري ومهني تطبيقي وبشكل مبسط مع الابتعاد بقدر الإمكان عن التعقيدات التي يمكن أن تتضمنها الكتب الجامعية كالمعادلات الرياضية المعقدة والخوارزميات المكتوبة بلغات البرمجة والتي لا تلزم إلاّ المتخصصين في علم الحاسوب، وهو ما يجعل هذا الكتاب مبسطًا وشيقًا ونافعًا للدارسين

في العديد من المجالات المرتبطة بتحليل البيانات بشكل عام، كالرياضيات والإحصاء وعلوم وهندسة المعلومات والبحث العلمي بكل فروعه وتخصصاته، وعلم الحاسوب بشكل خاص. بحيث يساهم في تبسيط فهم المادة العلمية وتوسيع مدارك الطالب وفتح آفاق جديدة له وتوضيح علم تنقيب البيانات بشكل تطبيقي عملي من خلال الأمثلة والتطبيقات العملية في مجالات متعددة، وبطريقة تساعده في اتخاذ القرار المناسب عند اختياره للتخصص الدقيق في مجال دراسته الأكاديمية أو حتى عند اختياره لمجال العمل بعد التخرج، فهذا الكتاب يُعتبر كالمرآة التي تعكس الواقع العملي غير المنظور في تلك المجالات بالنسبة للطالب والذي ينتظره بعد التخرج، بحيث يمكن أن يسهل له اختيار مجال العمل الذي يناسب قدراته واهتماماته وميوله من الناحية العملية.

والله ولي التوفيق،

مصطفى عبيد

الفصل الأول
مدخل إلى التحليل المتقدم وتنقيب البيانات

An Introduction to Advanced Analysis & Data Mining

المحتويات

- ما هو التحليل المتقدم وتنقيب البيانات وما هي إجراءاته وأدواته
- ما نوع البيانات التي يتم تنقيبها
- ما هية قواعد البيانات Database
- قاعدة البيانات العلائقية Relational Database
- لغة الاستعلام SQL Standard Query Language
- فوائد التنقيب في قواعد البيانات Database Mining
- أشهر تطبيقات تنقيب البيانات Data Mining Application

١. ذكاء الأعمال (استخبارات الأعمال) Business Intelligence

٢. محركات البحث على الإنترنت Web Search Engines

مدخل إلى التحليل المتقدم وتنقيب البيانات

ما هو التحليل المتقدم وتنقيب البيانات وما هي إجراءاته وأدواته

إن التقدم العلمي وانتشار استخدام التكنولوجيا في شتى مناحي الحياة اليومية أدى إلى رفع القدرة في توليد وجمع البيانات بشكل سريع في هذا العصر، وقد ساهم ذلك في حوسبة معظم الأعمال والعلوم والخدمات التي يتم تقديمها يومياً في كل مكان حول العالم، حتى أن معظم المنتجات بكافة أنواعها أصبح لها شيفرة رقمية تميزها عن بعضها البعض Bar Code، كما أن التقدم التكنولوجي تسبب في ظهور أنواع جديدة من البيانات كالنصوص والصور والفيديو وأنظمة متعددة المهام بالإضافة لشبكة الإنترنت التي تحتوي كميات هائلة من البيانات بكافة أشكالها، كل ذلك أدى إلى تضخم غير مسبوق في كميات البيانات التي يتم تخزينها يومياً، الأمر الذي أظهر حاجات ملحة لتقنيات جديدة وأدوات ذكية يمكن أن تساعد في تحويل هذا الكم الهائل من البيانات إلى معلومات مفيدة ومعرفة، والتي تمثلت في أدوات تنقيب البيانات.

إن تنقيب البيانات يهدف إلى استخلاص المعلومات المخبأة في كتل البيانات الكبيرة، وهي تكنولوجيا حديثة فرضت نفسها بقوة في عصر المعلوماتية، واستخدامها يوفر للشركات والمؤسسات في جميع المجالات، الأهلية والحكومية، القدرة على استكشاف، والتركيز على، أهم المعلومات في كتل

البيانات الكبيرة. كما تركز تقنيات التنقيب على الاستشعار وبناء التنبؤات المستقبلية واستكشاف الأنماط والارتباطات والسلوك والاتجاهات مما يسمح بتقدير القرارات الصحيحة واتخاذها في الوقت المناسب ووضع الحلول المناسبة للمشكلات والتخطيط والتطوير والتحديث في جميع المجالات.

وتجيب تقنيات التنقيب على العديد من الأسئلة، وفي وقت قياسي، بخاصة تلك النوعية من الأسئلة التي كان من الصعب الإجابة عليها، إن لم يكن مستحيلًا، باستخدام تقنيات التحليل الإحصائي الكلاسيكية، والتي كانت إن وجدت فإنها تستغرق وقتًا طويلاً والعديد من إجراءات التحليل.

إن علم التحليل وتنقيب البيانات يعتبر من العلوم الحديثة نسبيًا وهو يشكل امتدادًا لعلم التحليل الإحصائي والعصب الرئيسي لعلوم ذكاء الأعمال بكافة أشكالها، وقد نشأ علم التنقيب كنتيجة طبيعية للتطور الكبير الذي حدث في مجال نظم المعلومات والتضخم الكبير في المعلومات التي تنمو بشكل أسي، وخاصة بعد الانتشار الواسع لاستخدام أنظمة المعلومات وتراكم الكم الهائل من البيانات التي أصبحت متداولة يوميًا في العديد من المجالات، الأمر الذي أدى إلى الحاجة الملحة للإجابة على العديد من الأسئلة واستكشاف المعرفة والتقديرات والتنبؤات المستقبلية.

وتعتبر عمليات التحليل والتنقيب من أولويات عمل دوائر التخطيط في الشركات والمؤسسات في العالم في الوقت الراهن والأداة المثلى للمستويات الإدارية العليا التي تطمح للنجاح وتضمن استمراره بشكل استراتيجي، نظرًا لما

توفره من إمكانية لإنتاج المعرفة الحقيقية المخبأة في كتل البيانات الكبيرة الخاصة بأنشطة كل شركة أو مؤسسة والتي يتم تخزينها وتراكمها يومياً.

وتجدر الإشارة إلى أن تنقيب البيانات Data Mining كموضوع علمي متخصص يمكن تعريفه بعدة طرق، وتعبير "تنقيب البيانات" في الواقع لا يوضح بشكل دقيق كل المقصود منه، فبمقارنة هذا التعبير بعمليات استخراج الذهب من بين الصخور والرمال من باطن الأرض الذي يُعبر عنه بتعبير التنقيب عن الذهب وليس تنقيب الصخور والرمال، وبالمثل، فإن مفهوم

تنقيب البيانات كان لا بد له أن يكون "التنقيب عن المعرفة من البيانات"، وهو تعبير طويل نسبيًا كما أن اختصاره إلى "تنقيب المعرفة" فقط ربما لا يعبر بشكل دقيق عن أننا نتحدث عن التنقيب عن تلك المعرفة في كميات هائلة من البيانات، ومن هنا أصبح تعبير تنقيب البيانات أكثر رواجًا واستخدامًا للتعبير عن هذا العلم كوسيلة لاستكشاف المعرفة من كُتل البيانات الكبيرة.

ما نوع البيانات التي يتم تنقيبها ؟

باعتبارها تكنولوجيا عامة، فإن تقنيات تنقيب البيانات يمكن تطبيقها على أي نوع من البيانات، طالما أنها ذات معنى لطريقة وأسلوب التطبيق الذي يتم اتباعه، ومن أشهر البيئات التي يتم تطبيق تقنيات التنقيب فيها هي البيئات التي توفر أوعية مخصصة لحفظ وأرشفة البيانات مثل نظم الجداول الإلكترونية Excel ونظم قواعد البيانات وقواعد البيانات العلائقية Relational Database ومستودعات البيانات Data Warehouse، والتي يمكن أن تحتوي على بيانات المعاملات المختلفة كمبيعات محلات التجزئة مثلاً.

كما يمكن تطبيقها على سلاسل البيانات والأشكال وتدفقات البيانات والبيانات المكانية والنصوص وبيانات الوسائط المتعدد وبيانات الإنترنت.

ما هية قواعد البيانات ؟

إن نظم قواعد البيانات Database، وتسمى أيضًا نظم إدارة قواعد البيانات Database Management Systems، واختصارها هو DBMS، تتكون من مجموعة من البيانات المترابطة والتي تعرف بقاعدة البيانات ومجموعة من البرمجيات تكون مهمتها الإدارة والوصول إلى البيانات، وتوفر البرمجيات آلية لتعريف بنية قاعدة البيانات وطرق تخزين البيانات فيها والإدارة المتزامنة للبيانات وطرق الوصول إليها ومشاركتها وحفظها وتأمينها وضمان كفاءتها وتداولها بالطرق الصحيحة.

قاعدة البيانات العلائقية

قواعد البيانات العلائقية Relational Database هي نوع من قواعد البيانات التي تعتمد على إنشاء علاقات مميزة بين فئات البيانات فيها، بحيث تساهم تلك العلاقات في تيسير حفظها في أوعية مناسبة لها وإدارتها بالشكل الأمثل، من خلال تطبيق مزايا توفرها العلاقات. مثلا في قاعدة بيانات خاصة بالموظفين في أي مؤسسة يمكن إنشاء علاقة بين بيانات الموظف التي يتم الاحتفاظ بها في جدول خاص ببيانات الموظفين الشخصية مع بيانات أنشطة الموظف التي يتم حفظها في جدول مختلف، وبهذه الطريقة يتم اختزال الكثير من الوقت والجهد وحتى المساحة اللازمة لحفظ البيانات الشخصية بدلا من تكراراها في جدول الأنشطة مع كل نشاط جديد يتم تخزينه.

كما توفر قواعد البيانات العلائقية وسيلة لإيجاد روابط بين أنواع مختلفة من البيانات بطريقة تسهل من سبل إدارتها، فمثلاً إذا كنا بصدد طباعة جدول يحتوي بيانات الموظفين مرتب تنازليا بحسب الدرجة الوظيفية يكون بالإمكان تنفيذ هذا الأمر لو كانت الدرجات الوظيفية سهلة الترتيب، أما لو كانت تلك الدرجات من نوع البيانات غير الرتبية أو التي لا يمكن ترتيبها فإنه يلزم في تلك الحالة إضافة حقل جديد خاص بالترتيب وربطه بالحقل الخاص بالدرجة الوظيفية، بحيث يتم التعبير عن الدرجات بصورة قابلة للترتيب.

وتتكون قاعدة البيانات العلائقية من مجموعة من الجداول، لكل جدول أسم فريد، وكل جدول يتكون من مجموعة من السمات (الأعمدة أو الحقول)، ويتم فيه تخزين عدد كبير من السجلات (الصفوف)، وفي كل سجل في قاعدة البيانات يمثل كائن معرف بمفتاح فريد ووصف محدد لمجموعة من السمات ذات قيم محددة.

ففي قاعدة بيانات خاصة بمحل لبيع الأجهزة الإلكترونية يمكن أن تتألف من الجداول التالية بما فيها من حقول لكل جدول كما يلي:

- جدول الزبائن: ويحتوي على حقول مثل (رقم الزبون، الاسم، العنوان، العمر، المهنة، ... إلخ).
- جدول المنتجات: (رقم المنتج، الماركة، الفئة، النوع، السعر، بلد الصنع، المورد، ... إلخ).
- جدول الموظفين: (رقم الموظف، الفئة، المجموعة، الراتب، العمولة، ... إلخ).
- جدول الفروع: (رقم الفرع، الاسم، العنوان، ... إلخ).
- جدول المبيعات: (رقم الفاتورة، رقم الزبون، رقم الموظف، التاريخ، الوقت، طريقة الدفع، المبلغ، ... إلخ).

الجدول التالي يبين مقطع من أحد جداول حركة المبيعات في قاعدة بيانات أحد محلات الأجهزة الإلكترونية:

رقم الفرع	رقم الفاتورة	الصنف	عدد الوحدات	سعر الوحدة	طريقة الدفع
١	١٠٠١	كمبيوتر محمول	١	٢٠٠٠	نقدًا
١	١٠٠٢	طابعة ليزر	١	٥٠٠	نقدًا
٢	٢٠٠١	كمبيوتر محمول	٢	١٢٠٠	نقدًا
٢	٢٠٠٢	كاميرا رقمية	١	٨٠٠	بطاقة بنكية
....					

لغة الاستعلام

يمكن استخدام لغة الاستعلام الهيكلية Standard Query (SQL) Language من أجل التعامل مع البيانات بهدف الوصول إلى مجموعات جزئية منها تنطبق عليها شروط معينة، فمثلاً يمكن استخدام لغة الاستعلام للإجابة على أسئلة مثل: ما هي قائمة المنتجات التي تم بيعها في الشهر السابق؟

كما يمكن للغة الاستعلام تجميع البيانات وإجراء بعض الحسابات عليها، والتي يمكن باستخدامها الإجابة على أسئلة من نوع "كم كان مجموع المبيعات في السنة الماضية، موزعة بحسب الأفرع؟ أو مثلاً من هو موظف المبيعات الذي باع أكثر من بقية الموظفين في الشهر السابق؟ أو كم عدد حركات البيع التي تمت في السنة الماضية؟ وغيرها من الأسئلة التي تتكون إجاباتها من فئات جزئية مختلفة من المجموع الكلي للبيانات.

فوائد التنقيب في قواعد البيانات

أما باستخدام أدوات التنقيب في قواعد البيانات، فإننا نستطيع أن نذهب لأبعد من ذلك، وذلك من أجل البحث عن واستكشاف الاتجاهات والأنماط بالإضافة للتنبؤ بما يمكن أن يجري في المستقبل.

مثلاً كأن يتم استكشاف تفضيلات الزبائن في شراءهم لبعض المنتجات على غيرها من المنتجات الأخرى، والتنبؤ باحتمال إقبال الزبائن على شراء منتجات محددة وفقًا لتوفر معطيات يتم دراستها من البيانات التاريخية في قاعدة البيانات، أو أن يتم مثلاً استكشاف سلوك الزبائن في شراءهم لمنتجات معينة مع منتجات أخرى في نفس رحلة التسوق، وغيرها من الأنماط والاتجاهات التي تبين السلوك الشرائي للزبائن وتساعد في بناء التوقعات المستقبلية لسلوكهم الشرائي.

وتمثل قواعد البيانات وبيئات مستودعات البيانات بشكل عام بيئات متميزة لعلم تنقيب البيانات نظراً لأنها تعتبر البيئة المثالية التي يمكن تطبيق تقنيات التنقيب فيها، وذلك لأنها تعمل على تخزين البيانات بشكل منظم وضمن بنية وهيكلية تساعد بشكل كبير في تطبيق تقنيات التنقيب بالشكل الأمثل، وذلك باعتبار أن مستودعات البيانات يتم فيها تجميع البيانات من عدة مصادر والتي تشتمل على قواعد البيانات وغيرها من المصادر.

وتكمن أهمية تنقيب البيانات في الفوائد التي يمكن أن توفرها للشركات

والمؤسسات من خلال ما تقدمه من أساليب وأدوات وأفكار جديدة تساعد في رفع كفاءة الإنتاجية وتحسين وتطوير العمل، بالإضافة لمساهمتها في بنية التخطيط والإدارة الإستراتيجية في الشركة أو المؤسسة التي تستخدمه، وبناء التوقعات والتقديرات المستقبلية لسير العمل وفق المعطيات المتوفرة في كتل البيانات التي يتم تنقيبها. ويظهر ذلك من خلال أمثلة كثيرة لا حصر لها وبطرق متنوعة تهدف جميعها لاستكشاف السلوك والأنماط والاتجاهات التي تهم المدراء التنفيذيين ومدراء الإدارة العليا في الشركات والمؤسسات في جميع المجالات.

أشهر تطبيقات تنقيب البيانات

بشكل عام يمكن القول أنه "أينما توجد البيانات فإنه توجد معها تطبيقات تنقيب البيانات، ومن أشهر هذه التطبيقات في مجالات الأعمال هي:

- ذكاء الأعمال (استخبارات الأعمال)
- محركات البحث على الإنترنت

١. ذكاء الأعمال (استخبارات الأعمال)

من المهم جدًا في المؤسسات والشركات المتخصصة في عالم الأعمال الفهم الدقيق ومعرفة الخبايا المرتبطة بسير العمل، كمعرفة طبيعة الزبائن وسلوكهم الشرائي واتجاهاتهم وتفضيلاتهم، وفهم السوق وما يجري فيه، ومعرفة كميات العرض والطلب على المنتجات، وشدة المنافسة وكفاءة

منتجات المنافسين في السوق، وتوفر تكنولوجيا استخبارات الأعمال الرؤية الشاملة والمتكاملة للأعمال من خلال ما تقدمه من دراسة مفصلة للماضي وتوصيف دقيق للوضع الحاضر والتنبؤ بالمستقبل، وذلك من خلال ما توفره من تقارير وتحاليل وتقدير كفاءة إدارة الأعمال والاستخبارات التنافسية والمقارنات والتحاليل التنبؤية، والتي تعتمد جميعها على أدوات وخوارزميات تنقيب البيانات المتخصصة التي تهدف لدراسة وتحليل السوق وقياس سلوك ورضى الزبائن واستكشاف نقاط القوة ومكامن الضعف لدى المنافسين، وسبل الاحتفاظ بالزبائن المهمين والتركيز على شرائح وفئات محددة منهم، واتخاذ القرارات الذكية بشكل عام.

في الواقع فإن تنقيب البيانات يُعتبر النواة الأساسية لعمل استخبارات الأعمال Business Intelligence، فهو يعتمد عليه في كافة الإجراءات والأساليب المتبعة في تحليل البيانات، فمثلاً يتم استخدام أساليب تنقيب واستكشاف الأنماط وقواعد التبعية والارتباط من أجل دراسة وتحليل الأسواق وقياس سلوك ورضى الزبائن وتفضيلاتهم للمنتجات التي يشترونها معًا، أما تقنيات التصنيف والتنبؤ فإنها تُستخدم لدراسة الأسواق والمنتجات والعروض والطلبات عليها وتحليل المبيعات وتقدير المخاطر وتصنيف الزبائن والتنبؤ بمدى إقبالهم على شراء منتجات معينة، أما خوارزميات التجزئة العنقودية فإنها تساعد في إدارة العلاقات مع الزبائن، حيث يتم استخدام هذه التقنية من أجل تجزئة الزبائن إلى مجموعات يتشابه عناصر كل مجموعة

منها في خصائص معينة، ويتم التسويق لها بشكل محدد ومخصص وفقًا لتلك الخصائص المميزة لعناصر كل مجموعة من الزبائن، مع وضع أهداف وخطط وبرامج تسويق محددة لكل مجموعة بما يتناسب مع خصائصها المميزة ويحقق الأهداف التسويقية على المدى البعيد.

٢. محركات البحث على الإنترنت

محرك البحث على الإنترنت Web Search Engine هو خادم Server مخصص للبحث عن المعلومات على الإنترنت، ويقوم بتقديم نتائج لكل بحث يقوم به المستخدم، فتظهر له قائمة من النتائج التي يمكن أن تشتمل على صفحات الإنترنت أو الملفات النصية أو الصور أو أنواع أخرى من الملفات. ومحركات البحث تعمل بشكل آلي وتعتمد في عملها على خوارزميات تنقيب البيانات المصممة لتلبي احتياجات المستخدمين الذين يقومون بالبحث على الإنترنت باستخدام محركات البحث، فهي أساسًا عبارة عن تطبيقات مكبرة جدًا لتنقيب البيانات التي تستخدمها محركات البحث لأهداف متعددة، مثل ترتيب نتائج البحث بحسب حجم التدفق أو بحسب التكرار، والفهرسة وكيفية توزيع الفهارس وأسلوب البحث والإعلانات التي تظهر بجانب نتائج البحث وطرق انتقائها بحسب موضوع البحث وكيفية ظهور النتائج للمستخدم، سواء من حيث الترتيب أو من حيث المضمون بما يتناسب مع تفضيلات واهتمامات الشخص الذي يقوم بالبحث.

في الواقع فإن محركات البحث تُشكل تحديًا كبيرًا لخوارزميات تنقيب

البيانات، لثلاثة أسباب، فهي أولاً تحتوي على كميات هائلة من البيانات التي تتراكم يومياً بشكل غير مسبوق في أي مجال آخر، والذي يلزم لها الآلاف وربما عشرات الآلاف من أجهزة الحاسوب التي تتعاون معًا من أجل القيام بتنقيب الكميات الهائلة من البيانات، وهو مجال يتطور يومياً ويُعتبر من أوسع مجالات بحوث تنقيب البيانات المستقبلية، والسبب الثاني أن محركات البحث تتعامل مع المستخدمين بشكل حي ومباشر Online، وهو ما يتطلب أن يكون بمقدور محرك البحث الإجابة عن أسئلة واستفسارات المستخدمين بشكل فوري، وهذا يلزمه وجود خوارزميات تقييم وتصنيف تكون مهمتها تقييم وفهم طبيعة الأسئلة والاستفسارات التي يضعها المستخدم على محرك البحث، فمثلاً عندما يقوم أحد المستخدمين باستخدام محرك البحث للبحث عن كلمة "أمازون"، فإن محرك البحث لا يميّز ما إذا كان المستخدم يقصد بهذا الاستفسار عن نهر الأمازون أم عن موقع أمازون الشهير والمتخصص في بيع الكتب على الإنترنت، وبذلك فإنه يلزم وجود خوارزمية يعتمد عليها المحرك لمعرفة القصد الصحيح لهذا السؤال وذلك وفق ما يتوفر له من بيانات تاريخية سابقة لأسئلة شبيهة من مستخدمين آخرين حوال العالم وكيفية تعاملهم مع النتائج التي ظهرت لهم، لذا فإنه يتطلب الأمر التطوير الدوري والتحديث المستمر لمهام خوارزميات التنقيب طالما أن هناك المزيد من الاستفسارات وعمليات البحث من المستخدمين وتنوع ردود أفعالهم تجاه النتائج التي يحصلون عليها.

والسبب الثالث هو قيام بعض المستخدمين بالاستفسارات الفريدة التي ربما لا تظهر إلاّ مرة واحدة فقط أو بضع مرات، ورغم ذلك فإنها تتطلب وجود خوارزميات مخصصة مهمتها توفير الإجابة على تلك الاستفسارات، الأمر الذي يجعلها من التحديات المهمة لخوارزميات تنقيب البيانات بحكم أنه لا تتوفر بيانات تاريخية لعدد كبير من المستخدمين الآخرين في مثل هذا النوع من الاستفسارات، ويتم في هذه الحالة استخدام خوارزميات مطعّمة بالمعلومات التاريخية الخاصة بالمستخدم نفسه وأسلوبه في التصفح وتفضيلاته الشخصية والمواقع التي زارها مؤخرًا من أجل التنبؤ بالنتائج التي ينبغي إظهارها له وفقًا لتلك المعلومات.

الخلاصة

الجدول التالي يمكن تلخيص عملية تطور وتعاظم الإمكانيات التي وفرتها تكنولوجيا المعلومات خلال العقود الماضية، وذلك من خلال إمكانية الإجابة على التساؤلات التي توفرها التطبيقات المختلفة والتي تطورت تدريجياً منذ بدء استخدام الحاسوب وحتى الوصول إلى مستوى تنقيب البيانات كما يلي:

مرحلة التطور في استخدام تكنولوجيا المعلومات	السؤال الذي توفر إجابته تلك التكنولوجيا
تجميع البيانات Data Collection (1960s)	كم كان مجموع الأرباح في السنوات الخمس الأخيرة؟
الوصول للبيانات Data Access (1980s)	كم كان حجم المبيعات في مدينة القاهرة في شهر مارس الماضي؟
مستودعات البيانات ودعم القرار Data Warehousing & Decision Support (1990s)	كم كان حجم المبيعات في مدينة القاهرة في شهر مارس الماضي، مع مقارنة لكل المبيعات في المدن الأخرى؟
تنقيب البيانات Data Mining (Emerging Today)	ماذا يمكن أن يحدث لحجم المبيعات في مدينة القاهرة الشهر القادم، ولماذا؟

ويتضح من هذا الجدول كيف أن خوارزميات تنقيب البيانات أصبحت توفر الإمكانات اللازمة للتنبؤ بالسلوك المستقبلي ومن ثم وضع الحلول المناسبة للمشكلات قبل وقوعها في حال إمكان حدوثها، أو من باب التنبؤ بهدف التطوير والتحديث والتخطيط الجيد بشكل عام في شتى المجالات.

الفصل الثاني
التعرف على البيانات

Knowing the Data

المحتويات

- تمييز أنواع وخصائص وسمات البيانات Data Types & Attributes
- الوصف الإحصائي للبيانات Data Statistical Description
- التعرف على طرق التصوير المرئي للبيانات Data Visualization
- قياس تشابه واختلاف البيانات Measuring Data Similarity & Dissimilarity

أنواع وخصائص وسمات البيانات

حتى يتم تطبيق إجراءات تحليل وتنقيب البيانات ينبغي أولاً التعرف على ما هية البيانات التي سيتم تنقيبها، أنواعها وخصائصها أو سماتها التي تجعل منها بيانات من نوع معين، وذلك من أجل توضيح طرق التعامل معها في الدراسات والبحوث الهادفة لاستكشاف المعرفة المفيدة من كُتل البيانات الكبيرة والتي تعتمد في الأساس على تقنيات وخوارزميات التنقيب المختلفة.

في هذا الفصل سوف يتم التعرف على الأنواع المختلفة للبيانات Data Types وطبيعتها وخصائصها وسماتها Attributes وطرق التعامل معها، كما سوف يتم التطرق بالتفصيل المناسب لأساليب الوصف الإحصائي للبيانات، بالإضافة لطرق التصوير المرئي للبيانات، وأخيرًا لطرق قياس التشابه والاختلاف بين البيانات.

السمة أو المتغير

السمة هي حقل من حقول البيانات التي تمثل خاصية أو ميزة لكائن البيانات، واسم "سمة" أو ميزة أو متغير غالبًا ما تستخدم بشكل متنوع في المجالات المتعددة من مجالات التعامل مع البيانات، ففي الذكاء الاصطناعي جرت العادة على استخدام التعبير "ميزة" أما في مجال الإحصاء فجرت العادة على استخدام تعبير "متغير"، أما في مجال تنقيب البيانات فجرت العادة على

استخدام مصطلح "سمة" Attribute.

والسمة تصف كائن محدد في قاعدة البيانات، مثلاً في جدول بيانات الزبائن في أحد المحلات يمكن أن نطلق تعبير السمة على الحقول التالية (الاسم، العنوان، الدخل، ... إلخ)، وكل سجل من سجلات هذا الجدول سوف يمتلك مجموعة من السمات (المتغيرات) المميزة له.

وتوجد عدة أنواع مختلفة من السمات (المتغيرات) وتختلف هذه الأنواع عن بعضها البعض بحسب طريقة قياسها، فمنها السمات الاسمية والمنطقية والرتبية والرقمية، وفيما يلي وصفًا تفصيليًا لكل نوع.

١. البيانات الاسمية Nominal Data

إن البيانات ذات السمة الاسمية Nominal Attribute تعني أنها ترتبط بالأسماء وقيمها هي قيم اسمية أو رموز، مثل أسماء الأشياء أو الأشخاص، كما أنها لا تخضع للترتيب، ويمكن أن تمثل فئات أو تصنيفات معينة، مثلاً في قاعدة بيانات مبيعات أحد الشركات يمكن أن تأخذ بعض الحقول سمات مثل (الحالة الاجتماعية) والتي تكون القيم المحتملة لها هي (أعزب، متزوج، ...)، أو (المهنة) والتي يمكن أن تكون القيم المحتملة لها هي (مدرس، طبيب، مزارع، ...)، وهكذا فإن هذه القيم جميعها هي قيم اسمية، لذا فإنه يطلق على هذه السمات (الحالة الاجتماعية، المهنة، ...) أنها سمات اسمية.

وبالرغم من ذلك إلاّ أنه في بعض الأحيان قد تكون السمة اسمية

وتحتوي القيم فيها على أرقام، ولكن يتم التعامل معها كقيم اسمية وذلك مثلما يحدث في حقل (رقم الهاتف) أو (الرمز البريدي)، فالأرقام في هذه الحالة هي قيم اسمية وذلك لأنه لا يمكن جمعها أو طرحها أو مقارنتها حسابيًا مع بعضها البعض.

٢. البيانات المنطقية Boolean Data

إن البيانات ذات السمة المنطقية هي بيانات اسمية أيضًا، ولكن قيمها محصورة في قيمتين أو حالتين فقط، ويمكن التعبير عنها رقميًا باستخدام النظام الثنائي للأعداد بالقيمتين (صفر، ١)، حيث تعبّر القيمة (صفر) عن غياب السمة أو عدم تحققها، والقيمة (١) عن تحققها. وهي تناظر القيم (نعم، لا) المنطقية، فمثلاً في قاعد بيانات مرضى أحد أطباء الصدرية إذا كان هناك سمة (متغير) بعنوان "مدخن" التي تصف حالة المريض وما إذا كان مدخنًا أم لا، فإن القيمة (١) تعني أن المريض مدخن، والقيمة (٠) تعني أنه غير مدخن، كما يمكن استخدام القيم (نعم، لا) بدلاً من (١، صفر).

٣. البيانات الرتبية Ordinal Data

إن البيانات ذات السمة الرتبية هي البيانات التي يمكن أن تأخذ قيمًا لها ترتيب معين فيما بينها ويكون هذا الترتيب ذات معنى، ولكن دون الاهتمام أو حتى بدون الحاجة لمعرفة الفرق الفعلي بين القيم المتتالية في هذا الترتيب.

فمثلاً، في أحد مطاعم الوجبات السريعة يمكن أن تُستخدم عدة خيارات

لحجم المشروب الغازي الذي يتم اختياره مع وجبة الطعام، بحيث تأخذ القيم التالية: (صغير، متوسط، كبير)، وهذه القيم تكون في ترتيب واضح يبين تسلسل الحجم من الأصغر إلى الأكبر بالرغم من أننا لا نعلم الفرق بينها بشكل محدد، أو في استبانة تقييم الزبائن للمنتجات قد نجد أحد المتغيرات أو السمات التي تحدد رأي الزبون في منتج معين فتأخذ أحد القيم التالية: (سيء جدًا، سيء، جيد، جيد جدًا).

إن جميع البيانات من النوع الاسمي والمنطقي والرتبي هي بيانات نوعية Qualitative Data، أي أنها تصف ميزات كائن البيانات بدون تحديد كمياتها أو أحجامها، أما البيانات التي تحدد هذه القيم بطريقة كمية فإنه يطلق عليها بيانات كمية Quantitative Data أو ما يمكن تسميته بشكل أكثر تحديدًا بيانات رقمية Numeric Data.

٤. البيانات الرقمية Numeric Data

إن سمات البيانات الرقمية هي من النوع الكمي، أي أنها قيم قابلة للقياس، ويمكن التعبير عنها بأعداد صحيحة أو حقيقية، كما أنها يمكن أن تكون على شكل قياس الفترة Interval-Scaled أو قياس النسبة Ratio-Scaled.

أ. بيانات القياس الفتري Interval-Scaled

ويتم فيها تقسيم القيم إلى فترات متساوية، وقيم هذه الفترات لها ترتيب يُعتد به، ويمكن أن تكون قيم موجبة أو سالبة أو حتى صفر، كما يمكن

مقارنتها مع بعضها البعض وحساب الفرق فيما بينها.

مثلاً، قياس درجات الحرارة باستخدام التدريج المئوي في أيام مختلفة من الأسبوع، حيث يمكن الحصول على قياس محدد كل يوم، ويمكن ترتيب هذه القيم تنازليًا أو تصاعديًا لمعرفة الأيام الأكثر حرارة أو الأكثر بردًا، كما يمكن حساب الفرق في درجة الحرارة بين الأيام المختلفة، وكذلك في قياس التاريخ وغيرها من السمات الشبيهة، ولكن في هذا النوع من البيانات يتم التعامل مع القيم على أنها فترات وليس قيم رقمية بحتة، ولتوضيح هذا الأمر مثلاً فإنه لا يمكن القول بأن درجة الحرارة (٣٠) درجة مئوية هي ضعف درجة الحرارة (١٥) درجة مئوية، كما أنه لا يمكن القول بأنه عندما تكون درجة الحرارة صفر أنه لا توجد قيمة لها أو أنها صفر حقيقي، وكذلك في حالة سمة التاريخ فإن السنة صفر ميلادية لا تعني أنها بداية العالم.

ومع ذلك فإنه يمكن حساب الفرق بين درجات الحرارة المختلفة كما يمكن حساب القيم الإحصائية المختلفة لها كالوسط والوسيط.

ب. بيانات القياس النسبي Ratio-Scaled

هي البيانات ذات السمة الرقمية، وفيها تكون قيمة الصفر هي قيمة حقيقية، ويمكن مقارنتها معًا، كما يمكن ترتيبها وإجراء العمليات الحسابية عليها، وحساب القيم الإحصائية لها كالوسط والمنوال وغيرها، ومن أمثلتها (سعر المنتج، العمر، الدخل، ... إلخ) فكل هذه البيانات ذات سمة رقمية

من النوع النسبي ويمكن ترتيبها وإجراء العمليات الحسابية عليها ومقارنتها وحساب النسبة فيما بين كمياتها المختلفة، فإذا كان لدينا منتجين سعرهما على الترتيب (١٠٠) و (٥٠) فإننا يمكن أن نقول أن سعر المنتج الأول هو ضعف سعر المنتج الثاني.

من جهة أخرى، فإنه يمكن تقسيم البيانات من منظور آخر كما يلي:

البيانات المنفصلة والبيانات المتصلة

البيانات ذات السمة المنفصلة Discrete Attributes يمكن أن تأخذ قيم متعددة ولكنها محدودة بعدد معين، ومن أمثلتها سمة المهنة، فهناك عدد محدود من المهن، أي أنها تنتمي لفئة معدودة ومنتهية، كما أنها يمكن أن تكون قيم اسمية أو رقمية، مثلما يحدث عند استخدام النظام الثنائي، أو حتى في حقل العمر الذي يتم تحديده بالسنوات ضمن مجموعة محددة من الأرقام مثلاً (من ٠ إلى ١٠٠).

أما إذا لم تكن السمة منفصلة فتكون متصلة Continues Attributes، وهذا يعني أن القيم المحتملة لها هي قيم غير محدودة وغير معدودة، تمثل فئة الأعداد الحقيقية، ويمكن أن تكون قيم أعداد طبيعية ولكنها لا يمكن حصرها في مجموعة محددة ومنتهية من القيم، أي أن قيمها تنتمي لمجموعة أعداد غير منتهية وغير محددة. ومن أمثلتها حقول مثل (سعر المنتج).

الوصف والتحليل الإحصائي للبيانات

مفهوم التحليل الإحصائي للبيانات

من أجل إنجاح مرحلة تحضير البيانات للتحليل والتنقيب ومن ثم إنجاح عمليات تحليل وتنقيب البيانات نفسها وسائر إجراءات البحث العلمي بكافة أشكاله وبكل أنواعه، فإنه من الضروري أن يكون لدينا صورة وصفية عامة للبيانات التي يتم التخطيط لتحليلها أو للتنقيب فيها. ويتم ذلك من خلال التحليل الإحصائي الوصفي لهذه البيانات (Statistical Data Analysis).

ويمكن الحصول على الصورة الوصفية باستخدام المبادئ الأساسية للإحصاء الوصفي والأدوات المختلفة التي توفرها من أجل التحليل الإحصائي للبيانات ووصفها من منظور إحصائي، وذلك من أجل التعرف على سمات أو خصائص البيانات وإلقاء الضوء عليها عن قرب.

أهمية التحليل الإحصائي للبيانات

تكمن أهمية التحليل الإحصائي للبيانات في أنه يوفر وصفًا أوليًا للبيانات محل البحث أو الدراسة، بحيث يساعد الباحث أو المحلل على فهم طبيعة البيانات قبل البدء بإجراءات أخرى من إجراءات البحث العلمي أو تنقيب البيانات، كاختبار الفرضية كما في البحث العلمي، أو تطبيق الخوارزميات في حالة التنقيب في قواعد البيانات.

خطوات التحليل الإحصائي للبيانات

تشتمل خطوات التحليل الإحصائي للبيانات على مجموعة من المقاييس الإحصائية المتنوعة. ومنها التوزيع التكراري (Frequency Distribution) ومقاييس النزعة المركزية (Central Tendency) التي تحدد طريقة توزيع البيانات حول المركز، وتشتمل على المتوسط الحسابي والوسيط والمنوال. كما تشتمل على مقاييس التشتت (Dispersion)، والمدى، والمدى الربعي، والتباين والانحراف المعياري. وكذلك مقاييس الارتباط (Correlation) والانحدار (Regression). هذا بالإضافة لبعض أساليب الوصف والتحليل الإحصائي للبيانات باستخدام الأشكال البيانية (Graphics)، التمثيل البياني بالأعمدة والقطاعات الدائرية والرسوم الخطية والهندسية المتنوعة.

وفيما يلي شرحًا مفصلاً لكل المقاييس الإحصائية:

أولاً: التوزيع التكراري

مفهوم التوزيع التكراري

لكل مجموعة من القيم المعروضة بصورة عشوائية لا يمكن وصفها إحصائيًا بسهولة، وقد لا نتمكن من شرحها أو تفسيرها بوضوح. ولهذا نلجأ إلى مقياس التوزيع التكراري (Frequency Distribution)، وهو عملية يتم فيها تبويب البيانات في جداول تكرارية أو عرضها برسومات بيانية على شكل أعمدة أو مدرج أو مضلع تكراري بحسب طبيعة البيانات الإحصائية.

فإذا جاءت علامات مجموعة من الأفراد في اختبار تحصيلي، بعد أن تعلموا بطريقة معينة، بصورتها الأولية كما هي في التوزيع الافتراضي التالي:

٢١	١٩	٢٠	٢١	٢٦	١٩
٢٣	٢٠	٢٠	٨	٢٢	٢٠
٢٠	١٩	٢٠	٢٢	٢٠	٢١
٢١	١٩	١٩	٢٢	٢٤	

فإن جميع هذه القيم بالنسبة للمتغير مدار البحث مجموعة كلية أو فئة واحدة. إلا أنه يمكن تقسيم هذه الفئة إلى عدة فئات أو مجموعات جزئية بحيث:

- تنتمي كل قيمة لفئة واحدة.
- تتساوى طول الفئات لتسهيل الإجراءات أو العمليات الإحصائية.
- تبقى الفئات متصلة، أي لا يوجد إهمال للفئات التي تكرارها صفر.
- تستنفذ الفئات جميع القيم حتى لو كانت القيم متطرفة.
- يكون عدد الفئات مناسباً لمدى القيم، أي يكون هناك توازن بين سهولة عرض وتفسير البيانات.

طول الفئة في الجدول التكراري

إن طول الفئة في الجدول التكراري يعتمد على مدى القيم وعدد الفئات التي يقترحها الباحث.

في البيانات السابقة مثلا لا يُحتمل أن يتم تقسيمها بطول فئة أكبر من واحد. حيث يوفر هذا الطول عددًا مناسبًا من الفئات كما في الجدول التالي. حيث يستحسن ألا تقل عن خمس فئات ولا تزيد عن ١٥ فئة.

بناء جدول التوزيع التكراري في التحليل الإحصائي

فيما يلي جدول التوزيع التكراري الذي يمكن بنائه لمجموعة البيانات في المثال المذكور أعلاه:

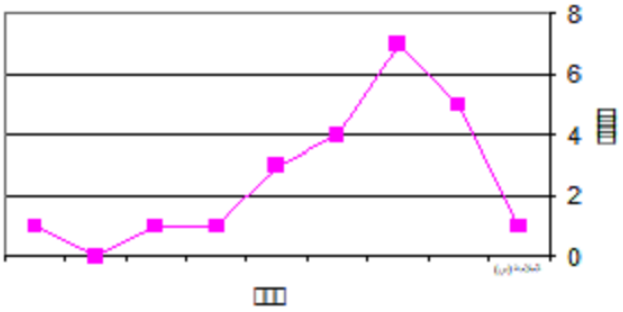
التكرارات (ت)	مركز الفئة (س)
١	٢٦
٠	٢٥
١	٢٤
١	٢٣
٣	٢٢
٤	٢١
٧	٢٠
٥	١٩
١	١٨

المضلع التكراري

الخطوة التي يسّهل فيها الباحث عرض البيانات الإحصائية بعد وضعها في جدول تكراري هي وضعها بصورة مضلع تكراري. ويوفر المضلع التكراري

إعطاء فكرة سريعة عن طبيعة البيانات وتوزيعها من حيث التفلطح (Kurtosis) والالتواء (Skewness).

يبين الشكل التالي عرضًا بيانيًا للتوزيع التكراري في الجدول السابق بصورة مضلع تكراري. حيث يفترض دائمًا أن توزيع القيم في الفئة توزيعًا منتظمًا وأن مركز الفئة هو المتوسط أو الوسط الحسابي للقيم في تلك الفئة:



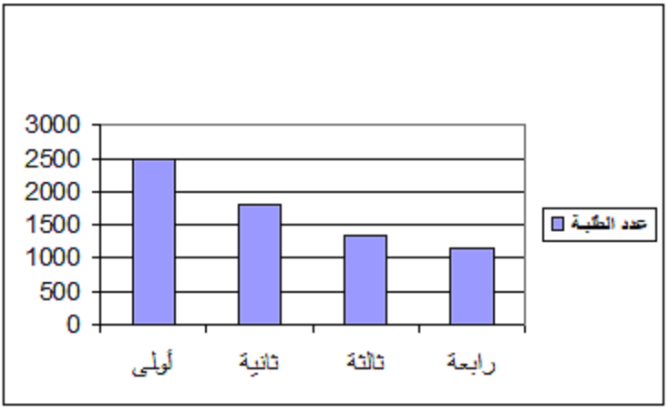
شكل (٢,١) المضلع التكراري

ثانيًا: التمثيل البياني بالأعمدة والقطاعات الدائرية

إذا كانت البيانات واقعة على متغير اسمي (تصنيفي) فلا تمثل بمضلع تكراري وإنما تمثل التكرارات أو نسبة التكرارات الواقعة ضمن كل فئة من فئات المتغير إما بالأعمدة (حيث يمثل ارتفاع العمود التكرار أو النسبة).

ويحدد موقع العمود بطريقة منطقية على الخط الأفقي أو الرأسي كما هو مبين في الشكل التالي، أو بالقطاعات الدائرية كما هو مبين في الشكل الذي

يليه، والذي يشير إلى نسبة الطلبة في السنوات الجامعية الأربعة في جامعة ما (حيث تمثل النسبة بمساحة القطاع في الدائرة).



شكل (٢,٢) تمثيل البيانات من النوع الاسمي بالأعمدة

شكل (٢,٣) تمثيل البيانات بالقطاعات الدائرية

ثالثًا: مقاييس النزعة المركزية

من أدوات الوصف والتحليل الإحصائي للبيانات مقاييس النزعة المركزية (Central Tendency). ومقاييس النزعة المركزية تخبرنا بطريقة توزيع البيانات وأكثرية القيم فيها. ومن أهم المقاييس وأكثرها انتشارًا هو الوسط أو المتوسط الحسابي الذي يحدد مركز مجموعة من البيانات بطريقة حسابية.

١. المتوسط الحسابي

المتوسط الحسابي أو الوسط الحسابي (Mean) لمجموعة من القيم هو حاصل قسة مجموع هذه القيم على عددها. أي أن:

المتوسط الحسابي لمجموعة من القيم = مجموع القيم ÷ عددها

فإذا كان لدينا بيانات رواتب مجموعة من الموظفين في إحدى الشركات، وكانت رواتبهم معرفة من خلال الفئة (س) التي تضم رواتب عدد (ن) من الموظفين كما يلي:

س = (س 1، س 2، س 3، س 4، .....، س ن)

فيكون المتوسط الحسابي لرواتب كل الموظفين هو:

المتوسط الحسابي = (س1 + س2 + س3 + س4 + ... + سن) ÷ ن

مثلا، إذا كانت لدينا هذه القيم للرواتب السنوية (بالآلاف) لمجموعة

من الموظفين مرتبة تصاعديًا كما يلي:

٣٠، ٣٦، ٤٧، ٥٠، ٥٢، ٥٢، ٥٦، ٦٠، ٦٣، ٧٠، ٧٠، ١١٠

فباستخدام تعريف المتوسط الحسابي يكون:

المتوسط الحسابي = (٣٠ + ٣٦ + ٤٧ + ٥٠ + ٥٢ + ٥٢ + ٥٦ + ٦٠ + ٦٣ + ٧٠ + ٧٠ + ١١٠) ÷ ١٢ = ٥٨

أي أن متوسط الرواتب لكل الموظفين هو (٥٨٠٠٠).

مشكلات المتوسط الحسابي

بالرغم من كثرة استخدام المتوسط الحسابي باعتباره كمية تساعد في التحليل الإحصائي للبيانات ووصفها إلاّ أنها لا تُعتبر الطريقة المثلى لقياس مركز البيانات.

والمشكلة الأساسية في المتوسط الحسابي أو الوسط الحسابي كأحد مقاييس النزعة المركزية هو أنه يتأثر بالقيم المتطرفة من مجموعة القيم الموجودة ضمن البيانات.

ويُقصد بـ القيم المتطرفة هي تلك القيم الكبيرة نسبيًا أو الصغيرة نسبيًا مقارنة مع بقية القيم، حيث يمكن أن يؤثر عدد قليل من مثل تلك القيم على قيمة المتوسط الحسابي الذي يتم احتسابه.

مثال توضيحي لتأثر المتوسط الحسابي بالقيم المتطرفة

في إحدى الشركات قد يتأثر متوسط الرواتب ويرتفع نتيجة لوجود عدد قليل من المدراء من ذوي الرواتب العالية نسبيًا. وبنفس الطريقة يمكن أن يتأثر متوسط درجات طلاب أحد الصفوف في أحد الامتحانات سلبًا نتيجة لوجود قيم أو درجات متدنية جدًا لبعض الطلاب.

ولإيقاف هذا التأثير الذي يتسبب به بعض القيم المتطرفة، العالية جدًا أو المنخفضة جدًا، يمكن استخدام طريقة المتوسط الحسابي المقلّم أو المقلّص (Trimmed Mean) والمتوسط الحسابي المقلّم أو المقلّص هو الذي يتم الحصول عليه بعد اقتطاع القيم المتطرفة العُليا والدنيا من مجموعة البيانات.

ففي المثال السابق، يتم استثناء عدد من القيم مثل الأولى والأخيرة (٣٠، ١١٠) باعتبارها قيم متطرفة ومن ثم يتم احتساب المتوسط الحسابي لبقية البيانات.

وبالرغم من إمكان تطبيق هذه الطريقة بسهولة إلاّ أنه ينبغي تجنب حذف عدد كبير من القيم المتطرفة من الاتجاهين لكيلا يؤدي ذلك إلى فقدان قيمة ودلالة البيانات.

حساب المتوسط الحسابي من الجدول التكراري

أما إذا كانت البيانات مبوبة في جدول تكراري فإن المتوسط يمكن حسابه باستخدام المعادلة التالية:

المتوسط الحسابي للبيانات المبوبة في جدول تكراري = مجموع حاصل

ضرب (مركز الفئة × تكرار الفئة) ÷ حجم العينة

مثلا في التوزيع التكراري التالي:

الفئة	مركز الفئة	التكرار
١٥ - ١٧	١٦	١
١٢ - ١٤	١٣	٢
٩ - ١١	١٠	٣
٦ - ٨	٧	٣
٣ - ٥	٤	١

وبتطبيق معادلة حساب المتوسط الحسابي:

المتوسط الحسابي = (١ × ١٦ + ٢ × ١٣ + ٣ × ١٠ + ٣ × ٧ + ١ × ٤) ÷ ١٠

أي أن المتوسط الحسابي في هذا المثال = ٩,٧

يُشكل الوسط أو المتوسط الحسابي نقطة الاتزان لأي توزيع. بمعنى أن مجموع انحرافات العلامات عن المتوسط = صفر.

ومن الخصائص الأخرى للوسط أنه الأقل تأثرًا بتقلبات العينة. ولذلك فهو مقياس النزعة المفضل في مجال الإحصاء الاستدلالي.

٢. الوسيط

الوسيط (بالإنجليزية: Median) هو من مقاييس النزعة المركزية ويُعبر بشكل أدق عن مركز البيانات لأنه القيمة التي تقع في منتصف مجموعة من القيم المرتبة. والوسيط هو القيمة التي تقسم البيانات إلى جزئين بحيث يقع الجزء الأول قبله والجزء الثاني بعده في الترتيب.

وفي مجال الإحصاء والاحتمالات فإن الوسيط يطبق بشكل عام على البيانات الرقية بالإضافة للبيانات من النوع الرتبي.

ولأي عدد (ن) من القيم لأحد المتغيرات (س) المرتبة ترتيباً تصاعديًا يكون الوسيط هو القيمة التي توجد في منتصف هذا الترتيب إذا كانت (ن) عدد فردي.

أي أن: الوسيط لعدد فردي ن من القيم = القيمة رقم (ن + ١) ÷ ٢

أما إذا كانت (ن) عدد زوجي فإن الوسيط يكون القيمتين اللتين تقعان في المنتصف.

أي أن: الوسيط لعدد زوجي ن من القيم هو القيمتين ذوات الترتيب (ن/٢) و (ن/٢)+١.

وإذا كانت البيانات من النوع الرقي يكون الوسيط هو المتوسط الحسابي للقيمتين.

مثال توضيحي لمقياس الوسيط في التحليل الإحصائي

في المثال السابق، لو كانت هذه هي قيم رواتب مجموعة الموظفين:

٣٠، ٣٦، ٤٧، ٥٠، ٥٢، ٥٢، ٥٦، ٦٠، ٦٣، ٧٠، ٧٠، ١١٠

فطالما أن عدد القيم هو عدد زوجي (١٢) قيمة، فإن الوسيط في هذه الحالة يكون ممثلا بالقيمتين السادسة والسابعة، وهما القيمتين (٥٢، ٥٦).

وبالتالي تكون قيمة الوسيط هي المتوسط الحسابي لهاتين القيمتين كما يلي:

الوسيط = (٥٢ + ٥٦) ÷ ٢ = ٥٤

أما لو قمنا بحذف القيمة الأخيرة فقط واعتبرنا أن البيانات هي:

٣٠، ٣٦، ٤٧، ٥٠، ٥٢، ٥٢، ٥٦، ٦٠، ٦٣، ٧٠، ٧٠

فيكون عدد القيم في هذه الحالة هو (١١) قيمة.

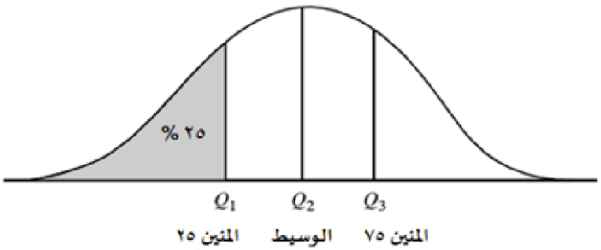
وبالتالي يكون الوسيط هو العدد ذو الترتيب = (١١+١) ÷ ٢ = ٦

أي أن الوسيط هو القيمة السادسة في الترتيب وهي القيمة ٥٢.

الوسيط والمئين

يُعرف الوسيط أيضًا بأنه "المئين ٥٠"، وبنفس الطريقة يمكن تعريف ما يُسمى بالقيم المئينية بشكل عام، على أنها القيم التي تقسم البيانات إلى قسمين بالنسبة المئوية المذكورة. مثلا المئين ٢٥ (بالإنجليزية: 25th Percentile) هو القيمة التي تقسم مجموعة من البيانات إلى قسمين بنسبة ٢٥٪ منها تقع

تحت تلك القيمة و٧٥٪ منها تقع فوقها. الشكل التالي هو رسم بياني يوضح الوسيط والمئين ٢٥ والمئين ٧٥:



شكل (٢،٤) رسم بياني يوضح الوسيط والمئين ٢٥ والمئين ٧٥

٣. المنوال

يمكن إعطاء وصفًا كميًا سريعًا للبيانات بعد تبويبها بالنظر إلى القيمة المنوالية. ويُعرف المنوال (بالإنجليزية: Mode) بأنه القيمة أو مركز الفئة أو الصفة المقابلة لأعلى تكرار في البيانات. والمنوال هو أبسط مقاييس النزعة المركزية من حيث إمكانية إيجاد قيمته، إلا أنه محدود الاستعمال. فهو أداة التحليل الإحصائي للبيانات الوحيدة التي يمكن استخدامها عندما تكون البيانات بمستوى القياس التصنيفي أو الاسمي.

ويتأثر المنوال كثيرًا بحجم العينة، وبتغير طول فئة البيانات ولذلك يعتبر قليل الثبات. كما أن المنوال لا يدخل كثيرًا في تحليلات إحصائية متقدمة في التحليل الإحصائي للبيانات. ففي المثال الخاص برواتب مجموعة من

الموظفين، الممثلة بالقيم التالية:

٣٠، ٣٦، ٤٧، ٥٠، ٥٢، ٥٢، ٥٦، ٦٠، ٦٣، ٧٠، ٧٠، ١١٠

يكون المنوال = ٥٢

أما في حالة التوزيع التكراري للفئات، فإذا كان في التوزيع فئتين متتاليتين أو أكثر وكانت متساوية في التكرار وبنفس الوقت أعلى التكرارات فإن المنوال هو معدل مراكز هاتين الفئتين.

مثال توضيحي لاستخدام المنوال في التحليل الإحصائي

التوزيع التكراري التالي ثنائي المنوال وهما القيمتين: ٢٢، ١٢:

الفئة	التكرار
٣٠ – ٣٤	١
٢٥ - ٢٩	٣
٢٠ - ٢٤	٨
١٥ - ١٩	٤
١٠ - ١٤	٨
٥ - ٩	٢

تمييز المنوال من الجدول التكراري

وُيلاحظ أنه يمكن أن تتشابه مجموعتين من البيانات وتتطابق قيم كل من الوسط أو المتوسط الحسابي والوسيط لهما. ومع ذلك يكون هناك اختلاف في طبيعة قيم كل مجموعة من حيث قربها أو بعدها عن المتوسط الحسابي أو

الوسيط، فمثلا المجموعتين التاليتين من القيم:

١٠، ٢٠، ٣٠، ٤٠، ٥٠، ٦٠

صفر، ١٠، ١٥، ٥٥، ٦٠، ٧٠

لهما نفس المتوسط الحسابي وهو ٣٥، ولكن البيانات في المجموعة الأولى تختلف عنها في المجموعة الثانية من حيث القيم من جهة ومن حيث البُعد عن المتوسط الحسابي من جهة أخرى. ومن هذا المنطلق ظهرت مقاييس إحصائية أخرى تختص بقياس مدى تشتت البيانات وبُعدها عن مقاييس النزعة المركزية.

رابعًا: مقاييس التشتت

مما سبق يتضح أنه لن يكون التحليل الإحصائي للبيانات أو وصفها كاملا بتحديد شكلها أو بتحديد مقياس النزعة المركزية الذي يناسبها وإنما قد يكتمل بتحديد درجة انتشار القيم المختلفة باستخدام مقياس مناسب من مقاييس التشتت (Dispersion).

وفيما يلي شرحًا مفصلاً لثلاثة مقاييس تشتت وهي المدى، التباين والانحراف المعياري.

١. المدى

المدى (بالإنجليزية: Range) يمثل المسافة بين أكبر قيمة وأقل قيمة في

مجموعة البيانات، والذي يُعتبر من أبسط مقاييس التشتت.

أي أن:

المدى لمجموعة من القيم = أكبر قيمة - أصغر قيمة

ففي المثال السابق، لو كانت هذه هي قيم رواتب مجموعة الموظفين:

٣٠، ٣٦، ٤٧، ٥٠، ٥٢، ٥٢، ٥٦، ٦٠، ٦٣، ٧٠، ٧٠، ١١٠

يكون المدى = ١١٠ - ٣٠ = ٨٠

كما يوجد نوعين آخرين للمدى وهما:

أ. المدى الربيعي

المدى الربيعي من مقاييس التشتت، وهو الفرق بين قيمتي المئين ٧٥ والمئين ٢٥. وعادة ما يُستخدم المدى الربيعي كمقياس تشتت عندما يُستخدم الوسيط كإحصائي نزعة مركزية .أي عندما تكون البيانات واقعة على مقياس رتبي أو عندما يكون في البيانات قيم متطرفة أو فئات مفتوحة.

ويمكن حساب المدى الربيعي بنفس طريقة حساب الوسيط مع استبدال القيمة ٥٠ بالقيم ٢٥ و٧٥ في المعادلة الخاصة بحساب الوسيط.

ب. المدى العشيري

وهو الفرق بين قيمتي المئين ٩٠ والمئين ١٠. والملاحظ أن الفرق بين

المدى العشيري والمدى الربيعي يكمن في النسبة المئوية للحالات المستبعدة في ذيلي التوزيع. ولذلك يُعتبر المدى العشيري بديلاً للمدى الربيعي إذا لوحظ بأن نسبة القيم التي تم اقتطاعها عالية نسبياً.

ويمكن حساب المدى العشيري بنفس طريقة حساب الوسيط مع استبدال القيمة ٥٠ بالقيم ١٠ و٩٠ في المعادلة الخاصة بحساب الوسيط.

وتنبع أهمية الحديث عن المدى الربيعي أو المدى العشيري كأحد مقاييس التشتت من ضرورة الانتباه إلى القيم المتطرفة وأثرها على النتائج وعلاقتها بحجم البيانات ومعالجتها بالطريقة المناسبة.

٢. التباين

التباين (Variance) من مقاييس التشتت وهو معدل مربعات انحرافات القيم الواردة في البيانات عن الوسط أو المتوسط الحسابي.

ويُرمز إلى التباين بالرمز σ^2 (ويُقرأ سيجما تربيع)، وهو يوضح مدى تشتت النواتج المحتملة عن قيمة المتوسط الحسابي (س).

معادلة حساب التباين:

$$\text{التباين} = \text{مج} \{(\text{س}_{\text{ن}} - \text{س})^2 \div \text{ن}\}$$

مثال على حساب التباين

نفرض أنه لدينا مجموعة من القيم التي نريد إيجاد قيمة الانحراف المعياري

لها، والقيم هي كما يلي:

٥، ٧، ٨، ١٠، ٢٠

أولاً: بتطبيق معادلة حساب الوسط أو المتوسط الحسابي، يكون لدينا:

المتوسط الحسابي = (٥ + ٧ + ٨ + ١٠ + ٢٠) ÷ ٥

أي أن:

المتوسط الحسابي = ٨

ثانيًا: بتطبيق معادلة حساب التباين، حيث لدينا:

التباين = مج {(س ن - س) ٢ ÷ ن}

أي أن:

التباين = ((س ١ – س)٢ + (س ٢ – س)٢ + (س ٣ – س)٢ + (س ٤ – س)٢ + (س ٥ – س)٢) ÷ ن

والآن بالتعويض عن القيم الخمسة وقيمة الوسط أو المتوسط الحسابي، نحصل على:

التباين = ((٥-٨)٢ + (٧-٨)٢ + (٨-٨)٢ + (١٠-٨)٢ + (٢٠-٨)٢) ÷ ٥

التباين = (٩ + ١ + ٠ + ٤ + ١٤٤) ÷ ٥

أي أن:

التباين = (١٥٨) ÷ ٥ = ٣١,٦

٣. الانحراف المعياري

الانحراف المعياري (بالإنجليزية: St. Deviation) هو من مقاييس التشتت، وهو يساوي الجذر التربيعي للتباين. ويتضح أن التباين والانحراف المعياري يعتمدان في قيمتيهما على الوسط أو المتوسط الحسابي.

معادلة حساب الانحراف المعياري:

الانحراف المعياري = الجذر التربيعي للتباين

الانحراف المعياري = الجذر التربيعي لـ (مجموع (س ن – س)^{٢} ÷ ن)

مثال على حساب الانحراف المعياري

في المثال السابق لحساب التباين، توصلنا إلى أن:

التباين = ٣١,٦

بأخذ الجذر التربيعي للطرفين:

الجذر التربيعي للتباين = الجذر التربيعي لـ (٣١,٦)

الانحراف المعياري = ٥,٦٢

خامسًا: التحليل الإحصائي باستخدام التوزيع

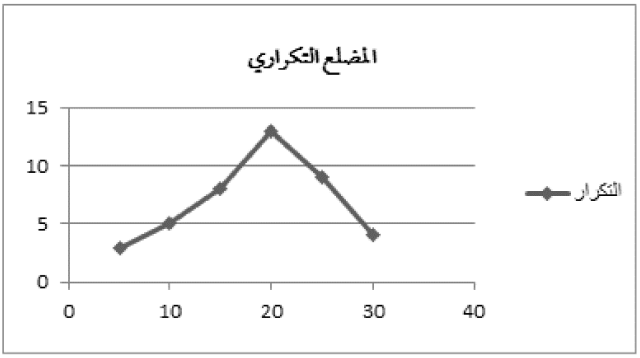
عند وضع البيانات في صورة مضلع تكراري فإنه يوفر إعطاء فكرة سريعة عن طبيعة توزيع (Distribution) البيانات من حيث التفلطح والالتواء.

١. خاصية الالتواء

تُشير خاصية الالتواء (Skewness) إلى درجة ابتعاد المنحنى التكراري عن التماثل. فقد تكون معظم القيم في الطرف الأدنى من التوزيع ويقل تكرار القيم كلما اقتربنا من الطرف الأعلى. وفي هذه الحالة يُوصف توزيع البيانات بأنه ملتوي التواء موجب. أما إذا كان العكس فيوصف بأنه ملتوي التواء سالب.

٢. خاصية التفلطح

تُشير خاصية التفلطح (بالإنجليزية: Kurtosis) إلى درجة تركيز التكرارات في منطقة الوسط للبيانات بالنسبة للتركيز في الطرفين مقارنة بالتوزيع الطبيعي القياسي. الشكل التالي يوضح التفلطح والالتواء في المضلع التكراري:



شكل (٢,٥) التفلطح والالتواء في المضلع التكراري

سادسًا: معاملات الارتباط

مفهوم الارتباط

من أساليب التحليل الإحصائي للبيانات ما يسمى بالارتباط، والارتباط هو مفهوم إحصائي يوضح العلاقة بين متغيرين أو أكثر. ونظرًا لتعدد أنواع البيانات أو المتغيرات وحتى وحدات القياس في البحث العلمي فقد تعددت أنواع معامل الارتباط وطرق حسابها.

والهدف من استخدام هذا المعامل يكون لإيجاد العلاقة بين متغيرين، وفحص ما إذا كانت علاقة إيجابية أو سلبية (علاقة طردية أو عكسية)، قوية أو ضعيفة.

كما تأتي أهمية دراسة الارتباط من دوره في التنبؤ كطريقة من طرق الحصول على المعرفة. فإذا كان الارتباط قويًا بين متغيرين فهذا يعني إمكانية

تقدير قيمة أحد المتغيرين عند معرفة القيمة المقابلة للمتغير الآخر بدقة أكبر مما لو كان الارتباط ضعيفًا.

الارتباط البسيط

يُقصد بالارتباط البسيط العلاقة بين متغيرين بصرف النظر عن نوع أي منهم من حيث نوع القياس، وأكثرها شيوعًا هو الارتباط بين متغيرين كل منهما من نوع القياس الفئوي أو من نوع القياس النسبي. ويحدد الارتباط عادة بالقوة والاتجاه (قيمة موجبة أو سالبة).

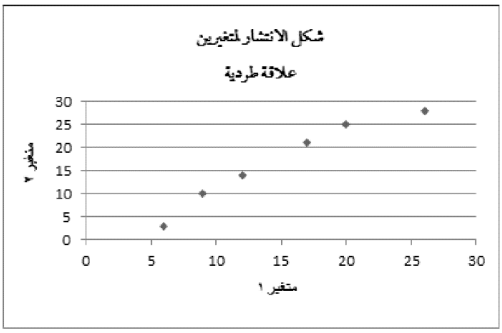
وتتلخص إجراءات تحليل الارتباط في الكشف عن قوة علاقة الارتباط واتجاهها من خلال كل من رسم شكل الانتشار وحساب قيمة معامل الارتباط، وفيما يلي تفصيل ذلك:

رسم شكل الانتشار

يعطي شكل الانتشار فكرة سريعة عن قوة واتجاه علاقة الارتباط بين متغيرين، فيتم تحديد قيم أحد المتغيرين على المحور الأفقي والمتغير الآخر على المحور الرأسي. وتحدد النقاط التي تشكل أزواج القيم إحداثياتها. والشكل الناتج بعد تحديد جميع النقاط هو شكل الانتشار.

وحتى يتضح معنى قوة العلاقة لا بد من مقارنة الشكل الناتج بشكل انتشار معين. فإذا كانت جميع النقاط واقعة على خط مستقيم فهذا يعني أن العلاقة تامة سواء كانت علاقة طردية أو عكسية. مثلاً، الشكل التالي يوضح

شكل الانتشار لمتغيرين مرتبطين بعلاقة طردية:



شكل (٢,٦) شكل الانتشار لمتغيرين مرتبطين طرديًا

حساب معامل الارتباط

يعتبر معامل الارتباط مؤشرًا كميًا على قوة العلاقة واتجاهها بين متغيرين، إذ يمكن أن يأخذ أي قيمة بين -١، ١. حيث تدل القيمة المحسوبة على قوة العلاقة وتدل الإشارة على اتجاهها، موجبة أو سالبة.

وتتعدد أنواع معاملات الارتباط حسب تعدد أنواع المتغيرات، فقد يكون الارتباط بين متغيرين كل منهما اسميًا، أو رتبيًا أو فئويًا، وربما كان خليطًا من هذه المتغيرات.

دلالة معامل الارتباط

في التحليل الإحصائي للبيانات يُشير معامل الارتباط إلى قوة واتجاه العلاقة بين متغيرين، ولكن هذه العلاقة لا تُفسَّر على أنها علاقة سببية، مع

أنها يمكن أن تكون كذلك.

ويمكن فحص معامل الارتباط بمقارنته بمعيار مُتفق عليه للعلاقة بين المتغيرات موضوع البحث. وقد جرى تصنيف قيم معامل الارتباط إلى (ضعيفة، متوسطة، قوية) إذا وقعت ضمن المدى (صفر إلى ٠,٣٩)، (٠,٤٠ إلى ٠,٦٩)، (٠,٧٠ إلى ١,٠٠) على الترتيب. ولكن هذه ليست قاعدة تُتبع دائمًا، فهذا أمر متروك للباحث على ضوء ما هو معروف عن العلاقة بين المتغيرات الواردة في البحث.

ومعامل الارتباط هو المؤشر الكمي على قوة العلاقة واتجاهها بين متغيرين، ويمكن أن يأخذ أي قيمة بين القيمتين (-١، ١)، حيث تدل قيمته المحسوبة على قوة العلاقة بين المتغيرين وتدل الإشارة على اتجاهها وما إذا كانت علاقة طردية (القيم الموجبة) أو علاقة عكسية (القيم السالبة).

أنواع معامل الارتباط

تتعدد أنواع معامل الارتباط بحسب تعدد أنواع البيانات أو المتغيرات التي يتم بحث أو تحليل الارتباط فيما بينها.

فقد يكون الارتباط بين متغيرين كل منهما اسميًا أو رتبيًا أو فئويًا، أو حتى خليطًا من هذه الأنواع، وبالتالي يختلف المعامل الذي يمكن استخدامه بحسب نوع وطبيعة تلك البيانات والمتغيرات.

وتوجد أربعة أنواع من معاملات الارتباط وهي:

١. معامل ارتباط بيرسون Pearson
٢. معامل ارتباط سبيرمان Spearman
٣. معامل ارتباط فاي φ
٤. معامل الارتباط الخطي الجزئي

وتُعتبر هذه الأنواع الأربعة هي الأكثر استخدامًا في مجالات البحث العلمي وتحليل البيانات أو تنقيب البيانات بشكل عام.

وفيما يلي وصفًا موجزًا لكل منها، مع شرح شروط استخدامها ومعادلة أو قانون حسابها مع التوضيح بالأمثلة التطبيقية:

معامل ارتباط بيرسون Person's Coeff

معامل ارتباط بيرسون أو معامل بيرسون هو معامل الارتباط بين متغيرين كل منهما من نوع البيانات المتصلة. وقد سُمي بهذا الاسم نسبة إلى العالم البريطاني كارل بيرسون الذي وضع أسس الإحصاء الرياضي.

وعند حساب معامل بيرسون فإنه يفترض أن العلاقة بين المتغيرين علاقة خطية، ويُفضل رسم شكل الارتباط للتأكد من ذلك قبل حساب هذا المعامل.

قانون حساب معامل بيرسون للارتباط

يمكن استخدام المعادلة التالية أو قانون حساب معامل بيرسون للارتباط لحساب قيمة المعامل كما يلي:

$$ر_{س ص} = \frac{\sqrt{ن \times مجـ (س \times ص) - (مجـ س \times مجـ ص)}}{\sqrt{(ن \times مجـ (س^{٢}) - (مجـ س)^{٢}) \times (ن \times مجـ (ص^{٢}) - (مجـ ص)^{٢})}}$$

قانون حساب معامل بيرسون للارتباط

مثال تطبيقي على معامل ارتباط بيرسون

المثال التالي يوضح خطوات حساب معامل بيرسون للارتباط، باستخدام القانون، بين عدد مرات شراء الزبون لمنتجات أحد المراكز التجارية (س) وتقييمه لهذه المنتجات (ص)، وعدد الزبائن في هذا المثال هو (ن).

بعد حساب كل س^{٢} وص^{٢} و(س × ص) نحصل على الجدول التالي:

س	ص	س ٢	ص ٢	س × ص
٤	٣	١٦	٩	١٢
١٥	١٠	٢٢٥	١٠٠	١٥٠
٨	٦	٦٤	٣٦	٤٨
٨	٧	٦٤	٤٩	٥٦
٦	٤	٣٦	١٦	٢٤

فيكون لدينا:

مجـ (س) = ٤١

مجـ (س^{٢}) = ٤٠٥

(مجـ س) ٢ = ١٦٨١

مجـ (س × ص) = ٢٩٠

مجـ (ص) = ٣٠

مجـ (ص ٢) = ٢١٠

(مجـ ص) ٢ = ٩٠٠

بالتعويض في قانون حساب معامل بيرسون، فإنه ينتج لدينا:

ر س ص = الجذر التربيعي لـ ((١٤٥٠ - ١٢٣٠) / (٢٠٢٥ - ١٦٨١) × (١٠٥٠ - ٩٠٠))

أي أن: ر س ص = ٠,٩٧ مقربًا لأقرب رقمين أو منزلتين عشريتين.

أي أن هناك علاقة ارتباط قوية بين عدد مرات شراء الزبون للمنتج وتقييمه له.

معامل ارتباط سبيرمان Spearman's Coeff

معامل ارتباط سبيرمان (أو معامل سبيرمان للارتباط) هو معامل الارتباط بين متغيرين كل منهما من بيانات النوع الرتبي.

ويُعتبر معامل سبيرمان للارتباط صورة أخرى من معامل بيرسون. فإذا كانت البيانات الإحصائية واقعة فعلا على مقياس رتبي أو أقرب إلى الرتبي منه إلى الفئوي فالمعامل المناسب للاستخدام هو معامل سبيرمان **ρ** (رو).

يصادف الباحث أحيانًا تشابهًا في رتب بعض البيانات للمتغير الواحد، وكلما زادت الرتب المشاركة كلما قلت دقة المعامل المحسوب بهذا المعامل.

قانون حساب معامل سبيرمان للارتباط

يمكن استخدام المعادلة التالية أو قانون حساب معامل سبيرمان للارتباط يدويًا لحساب قيمته كما يلي:

$$\rho \text{ (رو)} = ٦ \times (\text{مجـ ف}^{٢}) / \text{ن} \times (\text{ن}^{٢} - ١)$$

حيث أن: ف = فروق الرتب.

مثال على معامل سبيرمان للارتباط

المثال التالي يوضح حساب الارتباط بين رتب تقييم عدد (٦) زبائن للمنتج (س)، ورتب تقييمهم للمنتج (ص):

س	ص	ف ٢
١	٢	١
٦	٥	١
٥	٣,٥	٢,٢٥
٣	٣,٥	٠,٢٥
٢	١	١
٤	٦	٤
مجـ ف ٢		٩,٥

بالتعويض في قانون حساب معامل سبيرمان، يتم بالتالي الحصول على:

$$\rho = ٦ \times (\text{مجـ ف}^{٢}) / \text{ن} \times (\text{ن}^{٢} - ١)$$

وبالتعويض عن القيم يكون:

$$\rho = ٦ \times (٩,٥) / ٦ \times (٣٦ - ١)$$

$$\rho = ٠,٧٣ \text{ (مقربًا لرقمين أو لمنزلتين عشريتين).}$$

أي أن هناك علاقة ارتباط قوية بين تقدير الزبائن للمنتج (س) وتقديرهم للمنتج (ص).

معامل ارتباط فاي φ

معامل ارتباط فاي **φ** هو معامل الارتباط بين متغيرين كل منهما منفصل ثنائي، بمعنى أن كل منهما متغيرًا من النوع الاسمي ولكل متغير مستويين فقط. ولذلك لا يصلح هذا المعامل إذا كان لأحد المتغيرين أو لكليهما أكثر من مستويين.

قانون حساب معامل ارتباط فاي

المعادلة العامة أو قانون حساب معامل الارتباط فاي φ هو:

$$\varphi = \frac{(\text{[أ} \times \text{د]} - \text{[ب} \times \text{ج]})}{\sqrt{\text{[أ} + \text{ب]} \times \text{[ج} + \text{د]} \times \text{[أ} + \text{ج]} \times \text{[ب} + \text{د]}}}$$

قانون معامل ارتباط فاي

مثال على معامل ارتباط فاي φ

في المثال التالي عينة من عشرة أفراد من الجنسين تم اختيارهم لتقدير العلاقة بين جنس الأفراد ومدى قبولهم أو رضاهم عن الخدمات التي يقدمها أحد البنوك:

الخطوة الأولى: تمييز عناصر كل متغير بشيفرة معينة، كأن يعطي الجنس (ذكر، أنثى) أو الأرقام (١، ٠) على الترتيب، وتقييمهم للخدمات التي يقدمها البنك (قبول، رفض) أو الأرقام (١، ٠) على الترتيب.

والأرقام هنا ليس لها معنى كمي ولذلك يمكن أن يُصطلح على أي رقم أو رمز. ثم تبوب البيانات في جدول كالتالي:

رقم الفرد	1	2	3	4	5	6	7	8	9	10
الجنس	1	1	0	0	0	1	1	1	0	0
التقييم	1	1	1	0	0	1	1	0	1	1

تبويب البيانات تمهيدًا لحساب معامل ارتباط فاي

والآن لحساب معامل ارتباط فاي φ يلزم تجميع هذا الجدول على شكل اقتران ثنائي (أو دالة ثنائية) البعد كما يلي:

	قبول	رفض	المجموع
ذكر	أ = ٤	ب = ١	أ + ب = ٥
أنثى	ج = ٣	د = ٢	ج + د = ٥
المجموع	أ + ج = ٧	ب + د = ٣	أ + ب + ج + د = ١٠

تجميع البيانات المبوبة على شكل اقتران ثنائي

حيث تشير الرموز في الخلايا إلى عدد العناصر الناتجة من تقاطع فئات المتغيرين، فمثلا (أ = ٤)، وهذا يعني أن عدد الذكور الذين اعتبروا أن خدمات البنك مقبولة هو ٤.

وبتطبيق هذه المعادلة على المثال، يكون:

فاي (φ) = (٨ - ٣) / الجذر التربيعي لـ (٥ × ٥ × ٧ × ٣)

فاي (φ) = ٠,٢٢ مقربًا لمنزلتين عشريتين.

بمعنى أن هناك ارتباطًا ضعيفًا بين موقف الأفراد من الخدمات التي يقدمها البنك وجنسهم، إلا أن إشارة الارتباط موجبة، بمعنى أن الذكور يميلون إلى قبول هذه الخدمات أكثر من الإناث.

معامل الارتباط الخطي الجزئي

يُستخدم معامل الارتباط الخطي الجزئي في حالة وجود ثلاث متغيرات. وهو يقيس درجة العلاقة بين متغيرين اثنين بعد تثبيت أثر المتغير الثالث.

مثال على معامل الارتباط الخطي الجزئي

مثلا، لقياس درجة العلاقة بين الوزن والطول بعد تثبيت أثر اختلاف الأعمار على كل من الوزن والطول يوجد طريقتين:

الطريقة الأولى

نختار أفراد العينة من عمر واحد، وبذلك فإن الوزن سوف يتحدد على أساس الطول فقط.

الطريقة الثانية

نستخدم معامل الارتباط الخطي الجزئي والذي يُحسب باستخدام المعادلة أو القانون التالي:

قانون حساب معامل الارتباط الخطي الجزئي

$$ر_{س ص ع} = \frac{(ر_{س ص} - ر_{س ع} \times ر_{ص ع})}{\sqrt{(١ - ر^{٢}_{س ص}) \times (١ - ر^{٢}_{ص ع})}}$$

حيث $ر_{س ص}$ ، $ر_{س ع}$ ، $ر_{ص ع}$ هي معاملات الارتباط الخطية البسيطة بين (س، ص)، (س، ع)، (ص، ع) على الترتيب.

مع ملاحظة أنه إذا كان المتغيرين س، ص كلاهما متغير مستقل عن المتغير الثالث ع فإن:

ر س ع = ر ص ع = صفر.

اختيار المعامل المناسب

يتضح من شرح أنواع معاملات الارتباط واستخداماتها أنه يختلف باختلاف أنواع البيانات التي يهدف المعامل لإيجاد العلاقة بينها، ولذلك فإنه يلزم لاختيار وتحديد نوع المعامل المناسب الفهم المتعمق لطبيعة البيانات والمتغيرات التي يتم دراستها وتحليلها.

سابعًا: تحليل الانحدار

تحليل الانحدار (Regression) هو أسلوب يمكن بواسطته تقدير قيمة أحد متغيرين بمعلومية قيمة متغير آخر عن طريق معادلة رياضية تُسمى معادلة الانحدار وهي المعادلة التي تربط المتغيرين مع بعضهما البعض.

ومن خلال معادلة الانحدار يمكن عمل ما يلي:

- تحليل الانحدار: تحديد شكل العلاقة بين المتغيرين رياضيًا وبيانيًا.
- رسم خط الانحدار: توضيح اتجاه العلاقة بين المتغيرين على الرسم البياني.
- التنبؤ بقيمة أحد المتغيرين بدلالة المتغير الآخر، وذلك بالتعويض عن قيمة المتغير المعلوم في العلاقة أو المعادلة الرياضية وحساب قيمة المتغير الثاني فيها.

والانحدار له عدة أنواع، وهي:

١. الانحدار الخطي البسيط
٢. الانحدار المتعدد
٣. الانحدار غير الخطي

وفيما يلي وصفًا موجزًا لكل منها:

١. الانحدار الخطي البسيط

الانحدار الخطي البسيط بين متغيرين هو الانحدار الذي يكون فيه المتغير الأول (المتغير التابع) مرتبطًا بمتغير ثاني وحيد (المتغير المستقل) من خلال علاقة رياضية خطية، أي أنهما مرتبطان معًا بعلاقة رياضية من الدرجة الأولى.

مثلًا، العلاقة الرياضية التالية:

ص = أ + ب × س

تُمثل علاقة انحدار خطي بين المتغيرين: س، ص.

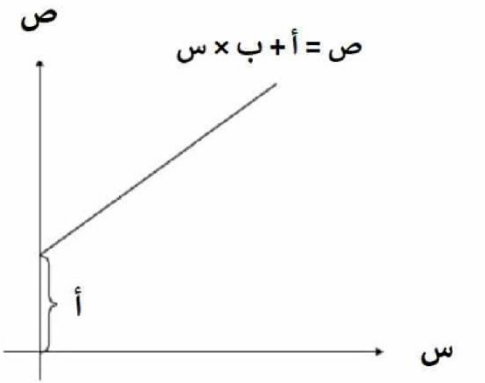
وذلك لأنها تُعبر عن معادلة أو علاقة خطية (خالية من $س^2$، $س^3$، ...، إلخ).

تمثيل الانحدار بالرسم البياني

من معادلة الانحدار الخطي:

ص = أ + ب × س

يمكن رسم خط بياني يمثل العلاقة الخطية بين المتغيرين (س) و(ص)، كما يلي:



شكل (٢,٧) رسم خط بياني يمثل العلاقة الخطية بين المتغيرين (س) و(ص)

حيث أن:

أ = ثابت الانحدار أو الجزء المقطوع من محور (ص)

ب = ميل الانحدار أو مُعامل انحدار (ص) على (س)

٢. الانحدار المتعدد

الانحدار المتعدد هو الانحدار الذي يرتبط فيه المتغير التابع بأكثر من متغير مستقل.

مثلًا، العلاقة الرياضية التالية تُمثل علاقة انحدار متعدد بين س، ص، ع:

ص = أ + ب × س + ج ×ع

٣. الانحدار غير الخطي

الانحدار غير الخطي هو الانحدار الذي يكون فيه المتغير التابع مرتبطًا بالمتغير المستقل بعلاقة رياضية غير خطية، كأن تكون مثلا علاقة رياضية من الدرجة الثانية أو من النوع الأسي (أي أنها تحتوي على س ن).

استخدام الانحدار في التنبؤ

نفرض أن لدينا متغيرين مرتبطين بعلاقة انحدار خطي من خلال العلاقة أو المعادلة الرياضية التالية:

ص = ٣ + ٤ × س

فإنه يمكن التنبؤ بقية المتغير (ص) عند معرفة قيمة المتغير (س) باستخدام العلاقة الرياضية التي تربط بينهما، وذلك كما يلي:

إذا علمنا أن (س = ٢٠)، نقوم بالتعويض عن قيمة (س) في معادلة الانحدار، فيكون لدينا:

ص = ٣ + ٤ × ٢٠ = ٨٣

وهكذا لكل قيمة للمتغير المستقل (س) يمكن معرفة القيمة المُناظرة للمتغير التابع (ص).

التصوير المرئي للبيانات
Data Visualization

١. التصوير البكسلي للبيانات

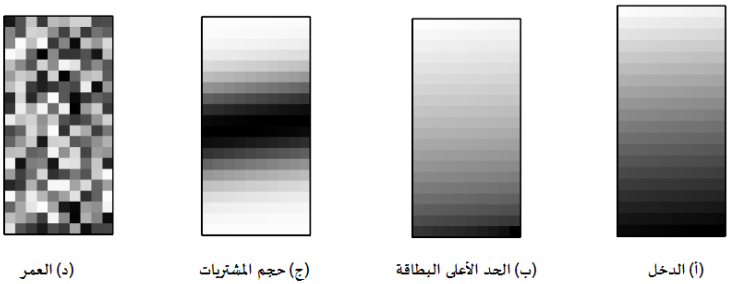
التصوير البكسلي للبيانات Pixel-Oriented Visualization هي طريقة سهلة للتعبير المرئي عن البيانات أو تصويرها مرئيًا، ويتم فيها استخدام التنقيط البكسلي Pixel الذي يخصص نقطة سوداء واحدة لكل وحدة من وحدات القياس للبيانات التي يتم التعبير عنها، فكلما زادت قيمتها يزداد عدد النقاط السوداء المستخدمة في التنقيط وبالتالي تزداد شدة التظليل في الصورة الناتجة، وبالعكس، كلما قلت القيمة التي يتم التعبير عنها بالتنقيط البكسلي تخف شدة التظليل في الصورة الناتجة.

مثلًا، لو افتراضنا أنه في أحد البنوك تم استخدام هذه الوسيلة لاستكشاف خصائص الزبائن الأكثر إنفاقًا باستخدام بطاقاتهم الائتمانية الصادرة عن البنك، فإنه يمكن للبنك بناء نماذج مخصصة لتصوير بيانات الزبائن الأساسية المرتبطة بتلك السمة والمراد بحثها واستكشاف علاقتها بزيادة الإنفاق باستخدام البطاقات، كالفئة العمرية، ومستوى الدخل وحجم المعاملات التي تمت باستخدام البطاقة.

ولكي يتم استكشاف العلاقة بين حجم الإنفاق ومستوى الدخل يتم ترتيب

البيانات تصاعديًا، أو تنازليًا، بحسب مستوى الدخل.

الشكل التالي يوضح مقارنة لتصوير بيانات زبائن أحد البنوك بالطريقة البكسلية مرتبة تنازليًا حسب مستوى الدخل:



شكل (٢،٨)، تصوير بيانات زبائن أحد البنوك بالطريقة البكسلية مرتبة تنازليًا حسب مستوى الدخل

الصورة الأولى من اليمين (أ) توضح تصويرًا لمستوى الدخل وقد تم ترتيبها تنازليًا (باتجاه الأعلى)، والصورة الثانية (ب) توضح الحد الأعلى للبطاقة المستخدمة، أما الصورة الثالثة (ج) فتبين حجم المشتريات باستخدام البطاقة، والصورة الرابعة (د) توضح عمر الزبون.

ويُلاحظ أنه عند مقارنة الصورة (ج) مع الصورة (أ) يتضح أن الزبائن من ذوي الدخل المتوسط نسبيًا هم الفئة الأكثر إنفاقًا باستخدام البطاقة، كما أنهم يتسمون بأنهم يمتلكون بطاقات شراء بحدود متوسطة عند المقارنة مع الصورة (ب)، ولكن الصورة (د) توضح أنهم من فئات عمرية مختلفة

ولا ينتمون لفئة عمرية محددة. وتتضح كل تلك الاستنتاجات من خلال ملاحظة شدة التظليل المرتفعة في منتصف الصورة (ج)، والمناظرة لمنتصف المسافة بين أكبر وأقل قيمة في الصور (أ)، (ب)، وتداخل المساحات المظللة وعدم انتظامها في الصورة (د).

٢. تصوير البيانات باستخدام تقنية الإسقاط الهندسي

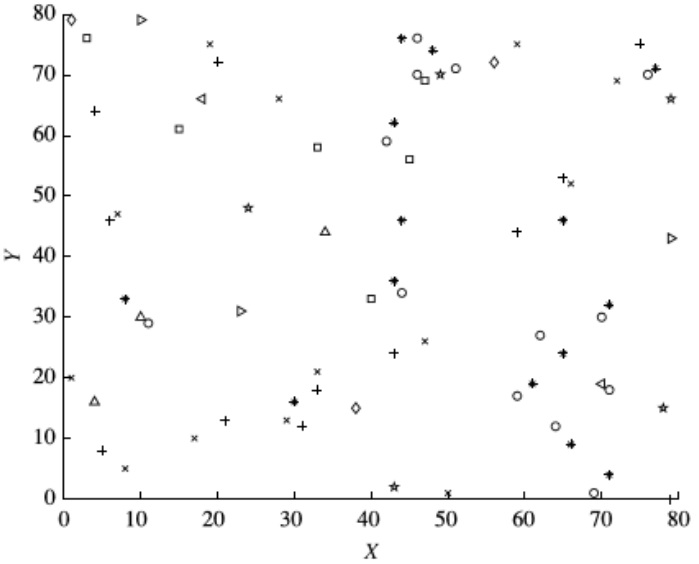
تفتقد تقنية التصوير البكسلي للبيانات للقدرة على المساعدة في توضيح وفهم كيفية توزيع البيانات التي تتقاطع في فضاءات متعددة الأبعاد، حيث أنه لا يمكنها أن تُظهر المساحات المظللة فيها مثلما تُظهره في حالة الأعمدة التي تعبر عن البيانات بشكل منفصل.

وتقنية الإسقاط الهندسي Geometric Projection Visualization تساعد في إيجاد الإسقاطات الشيقة لفئات البيانات متعددة الأبعاد، وذلك من خلال إظهارها على سطح مستوٍ في حالة البيانات ثنائية الأبعاد أو الأشكال المكعبة المجسمة في حالة البيانات ثلاثية الأبعاد، باستخدام أشكال القطع المبعثرة Scatter Plot.

الشكل التالي يمثل توزيعا بطريقة القطع المبعثرة ثنائية الأبعاد 2D-Scatter Plot، ويبين توزيع البيانات وفقًا لسمتين X، Y باستخدام نظام الإحداثيات الديكارتية ثنائي الأبعاد، ويتم إضافة البعد الثالث على الرسم الهندسي باستخدام رموز مختلفة لكل قيمة من القيم وتكون إحداثياتها ممثلة

بنقاط التقاطع لقيم كل من X وY التي يتم تحليلها.

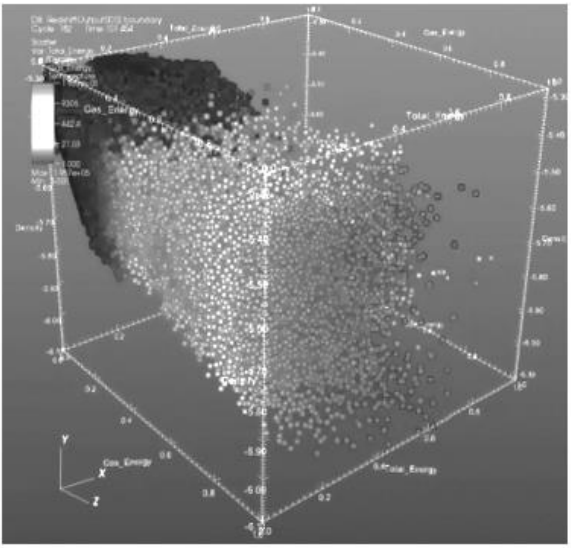
في الشكل التالي أحد الأمثلة على هذا النوع من التصوير، ويظهر فيه كيف يمكن ملاحظة وجود ترادف وتقارب مكاني للعلامتين + ,x حيث تظهر بشكل متكرر جنبًا إلى جنب في عدة أماكن على الرسم:



شكل (٢,٩)، تصوير البيانات بطريقة القطع المبعثرة ثنائية الأبعاد 2D-Scatter Plot

ويمكن لتقنية الإسقاط الهندسي أن تكون ثلاثية الأبعاد 3d-Scatter Plot، ويتم التعبير عنها باستخدام نظام الإحداثيات الديكارتية ثلاثي الأبعاد، بحيث تمثل الإحداثيات الثلاثة قيم كل من X، Y، Z، ويتم التعبير عن البعد الرابع بالتظليل في أماكن تقاطع القيم الثلاثة.

الشكل التالي أحد أمثلة تصوير البيانات باستخدام تقنية الإسقاط الهندسي ثلاثية الأبعاد، ويظهر فيه توزيع البيانات وفقًا لإحداثيات تقاطع قيم ثلاثة متغيرات x، y، z.



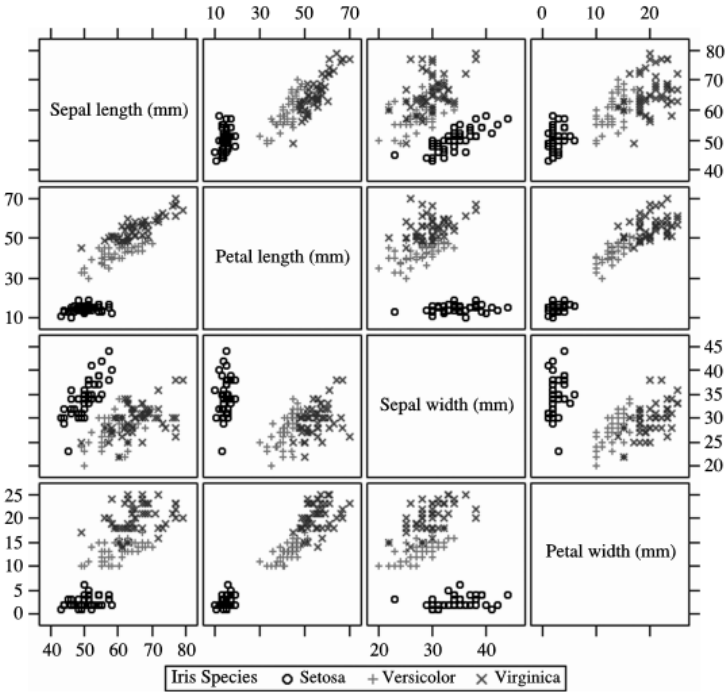
شكل (٢،١٠)، تصوير البيانات باستخدام تقنية الإسقاط الهندسي ثلاثية الأبعاد

٣. مصفوفة القطع المبعثرة

مصفوفة القطع المبعثرة Scatter Plot Matrix هي وسيلة لتوسيع الفائدة من طريقة الإسقاط الهندسي حتى يتم استخدامها في تصوير البيانات متعددة الأبعاد n×n، بحيث توضح تقاطعات كل بعد مع البعد الآخر وتوزيع

البيانات وفقًا للعلاقة البينية بين الأبعاد المختلفة.

في الشكل التالي مصفوفة قطع مبعثرة خماسية الأبعاد ٥×٥، تبين تصوير مجموعة من البيانات الخاصة بدراسة زهرة السوسن، والتي تتألف من ٤٥٠ نموذج لثلاثة أنواع منها، وفيه خمسة أبعاد للبيانات الخاصة بكل من الطول والعرض للجذع والفرع بالإضافة للنوع.



شكل (٢,١١)، مصفوفة قطع مبعثرة خماسية الأبعاد ٥×٥

٤. تصوير البيانات باستخدام تقنية الأيقونات

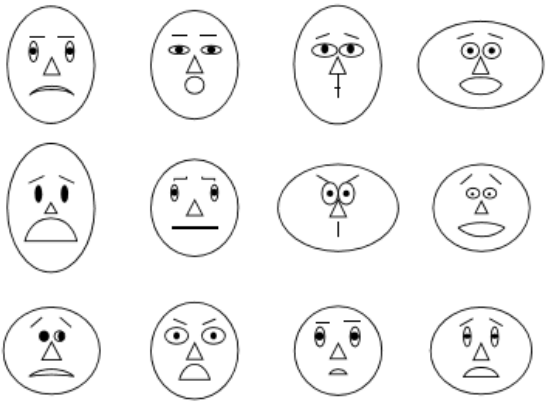
يمكن تصوير البيانات مرئيًا باستخدام تقنية الأيقونات Icon Based Visualization، وهي تقنية تُستخدم الرموز المصغرة من أجل تصوير البيانات ذات القيم متعددة الأبعاد، ومن أشهر أنواعها "وجوه شيرنوف" و"الأشكال النائتة".

أ. وجوه شيرنوف Chernoff Faces

ابتكرها العالم الإحصائي شيرنوف، وتُستخدم لاستعراض البيانات متعددة الأبعاد، والتي تصل إلى ١٨ متغير (بُعد)، باستخدام الوجوه الكرتونية، وتساعد وجوه شيرنوف في كشف الأنماط المختلفة للبيانات، وتُمثل مكونات الوجوه المستخدمة، كالعيون والآذان والفم والأنف، قيم المتغيرات من خلال الشكل والحجم والموضع والاتجاه، مثلاً يمكن ربط قيم المتغيرات مع خصائص الوجه المختلفة كحجم العين، تباعد العينين، طول الأنف، عرض الأنف، تقوس الفم، عرض الفم، حجم بؤبؤ العين، انحراف حاجب العين، وغيرها.

وتمكّن وجوه شيرنوف من استخدام القدرات العقلية للإنسان في ملاحظة الاختلافات الطفيفة في خصائص الوجوه المستخدمة في تمثيل عدة متغيرات في نفس الوقت. إن مشاهدة كميات كبيرة من جداول البيانات يُعتبر أمرًا مضجرًا، واستخدام وجوه شيرنوف يمكن أن يلخص تلك البيانات

ويجعلها أسهل للمستخدم حتى يستطيع فهمها واستكشافها.



شكل (٢,١٢)، تصوير البيانات باستخدام تقنية وجوه شيرنوف

وبهذه الطريقة يمكن تسهيل تصوير التناسق والتشابه أو الاختلاف والشذوذ بين كميات كبيرة من البيانات، بالرغم من محدودية هذه الطريقة في قدرتها على استكشاف العلاقات بين البيانات المختلفة.

كما تفتقد تقنية وجوه شيرنوف للقدرة على تصوير أنواع معينة من البيانات، وحتى في حالة التشابه بين وجهين من وجوه شيرنوف إذا كانا يمثلان مجموعتين من البيانات متعددة الأبعاد فإن البيانات نفسها قد تكون مختلفة وفقًا لترتيب الأبعاد والخصائص المستخدمة في الوجوه.

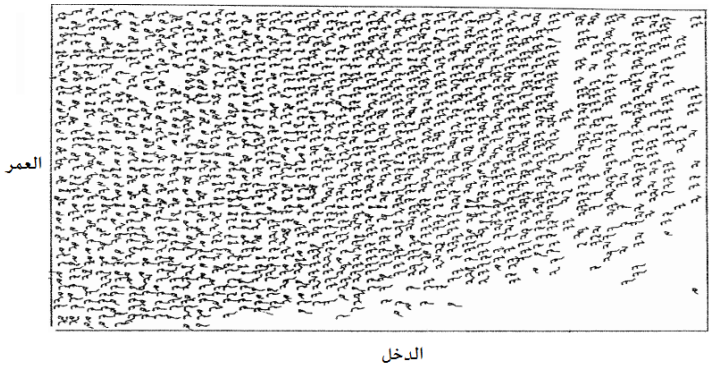
ب. الأشكال الناتئة Stick Figures

تصوير البيانات باستخدام تقنية الأشكال الناتئة هي عملية ربط مجموعة

من البيانات متعددة الأبعاد بأشكال ناتئة ذات خمسة أطراف، حيث يكون لكل شكل جسم وأربعة أطراف ويرتبط متغيرين منهم بمحوري س، ص، وترتبط بقية المتغيرات بخصائص الأطراف المتبقية، الطول والعرض وزاوية الميل.

الشكل التالي يوضح توزيع بيانات التعداد السكاني لأحد المناطق، بحيث يمثل كل من العمر والدخل على المحورين س، ص على الترتيب، وترتبط بقية المتغيرات (الجنس، التعليم، ...إلخ) بالأطراف المتبقية للأشكال الناتئة.

وبقدر ما تزداد كثافة الأشكال في توزيعها على إحداثيات المحورين س، ص فإنها تمثل زيادة في قيم تلك المتغيرات وطريقة توزيعها المناظرة على الرسم.



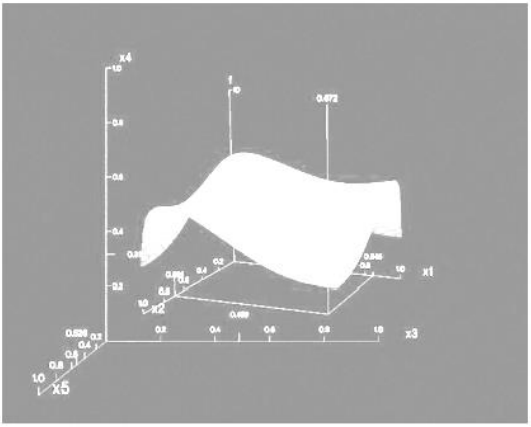
شكل (٢,١٣)، تصوير البيانات باستخدام تقنية الأشكال الناتئة

٥. تصوير البيانات باستخدام تقنية التصوير الهرمي

تشترك كل التقنيات السابقة في كونها تُستخدم لتصوير البيانات متعددة الأبعاد بشكل متزامن وبنفس الوقت على نفس الرسم، وهو أمر قد يكون صعبًا، وربما مستحيلاً، في حالة تصوير البيانات الضخمة متعددة الأبعاد بشكل كبير.

وتُعالج تقنية التصوير الهرمي Hierarchical Visualization هذا القصور من خلال استخدام أسلوب التجزئة، وذلك من خلال تجزئة كل الأبعاد إلى مجموعات جزئية ومن ثم تصوير تلك المجموعات بطريقة هرمية.

الشكل التالي يوضح تصويرًا باستخدام التقنية الهرمية لبيانات متعددة الأبعاد من خلال تجزئتها إلى مجموعتين من الأبعاد، لكل منها شكل ديكارتي مخصص لها:



شكل (٢,١٤)، تصوير البيانات باستخدام تقنية التصوير الهرمي

٦. تصوير البيانات والعلاقات المعقدة

قديمًا كانت تُستخدم تقنيات تصوير البيانات بشكل أساسي للبيانات الرقمية، أما هذه الأيام فإنها تُستخدم بشكل أكبر في تصوير البيانات غير الرقمية، مثل النصوص وبيانات شبكات التواصل الاجتماعي التي باتت متاحة للجميع ويكثر استخدامها بشكل كبير، بحيث أدى هذا الأمر لجعل تصوير وتحليل تلك البيانات أمرًا شيقًا، وأدى إلى ظهور تقنيات جديدة للتصوير تختص بهذا النوع من البيانات والعلاقات المعقّدة Complex Data Types And Relations.

مثلاً، يقوم الكثير من مستخدمي الإنترنت بوسم الصور والموضوعات والمدونات والمنتجات بوسوم مميزة Tags، وتقوم تقنية سحابة الوسوم Tag-Cloud بعملية تجميع إحصائي لكل الوسوم المستخدمة وترتيبها أبجديًا أو بحسب أي تفضيل آخر، ويتم تحديد أهمية كل وسم من الوسوم من خلال حجم الخط المستخدم أو اللون الممنوح له.

الشكل التالي يُظهر تقنية سحابة الوسوم المستخدمة في تصوير الوسوم الشائعة في أحد مواقع الإنترنت مرتبة ترتيبًا أبجديًا:

شكل (٢,١٥) تصوير مرئي للوسوم المستخدمة في أحد مواقع الإنترنت

ويتضح من الشكل أنه بالرغم من الترتيب الأبجدي للوسوم إلاّ أنه يُلاحظ أنه كلما ازداد عدد مرات استخدام الوسم إحصائيًا من قِبل المستخدمين في الموقع فإنه يزداد حجم الخط المكتوب به ذلك الوسم، سواء تكرر استخدامه من نفس الشخص أو من أشخاص متعددين، وبذلك تقوم عملية التصوير المرئي بإظهار الوسوم الأكثر استخدامًا إحصائيا بمجرد النظر إلى الشكل.

ومن الأمثلة المشهورة لتقنيات تصوير البيانات المعقدة الخريطة الشجرية التي تقسم البيانات إلى مجموعة من المستطيلات، وكما يظهر في الشكل التالي، والذي يبين طريقة تصوير شجرية لمحتوى الأخبار في جوجل، ويتم فيها تجميع كل محتويات الأخبار في فئات رئيسية تحتوي كل منها على مستطيل كبير

وبلون مميز، وفي كل فئة (أي بداخل كل مستطيل) يتم إعادة تجزئة محتويات الأخبار إلى فئات جزئية أصغر:

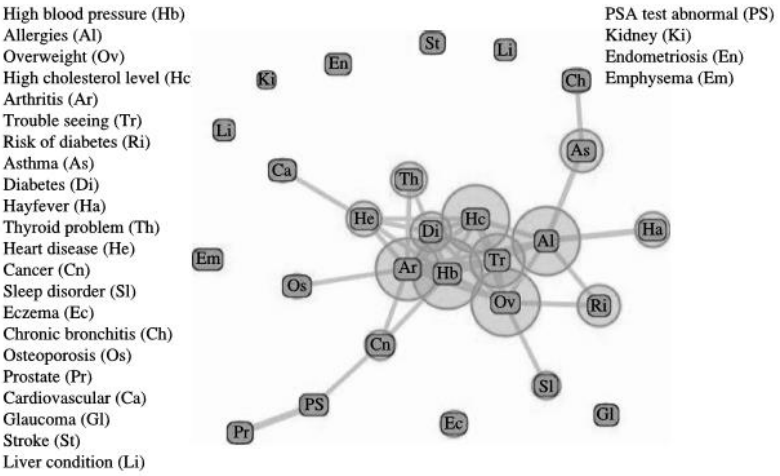
شكل (٢,١٦)، طريقة تصوير شجرية لمحتوى الأخبار في جوجل

٧. تصوير البيانات وعلاقات الارتباط

بالإضافة لتصوير البيانات المعقدة بهذه الطرق، فإنه يمكن تصوير العلاقات المعقدة Data and Correlations Visualization بين البيانات أيضًا، مثلاً، الشكل التالي يستخدم صورة رسم بياني لتأثير الإصابة بعدوى أحد الأمراض وعلاقته بالإصابة بأمراض أخرى، والدوائر المستخدمة في

الرسم تناظرها الأمراض المختلفة، وحجم الدائرة يتناسب طرديًا مع مدى انتشار المرض.

ويصل بين دائرتين خط مستقيم إذا كانت الأمراض المناظرة للدوائر مرتبطة مع بعضها البعض، كما يتناسب سمك الخط طرديًا مع قوة الارتباط فيما بينها:



شكل (٢,١٧)، تصوير الأمراض المختلفة ومدى انتشارها وعلاقات الارتباط فيما بينها

الخلاصة

توفر تقنيات تصوير البيانات أدوات فعّالة لتصوير البيانات، وقد قمنا بسرد مجموعة من طرق التصوير الشائعة والأفكار التي استندت عليها، وهناك العديد من الأدوات والطرق الشيقة التي تستخدمها تقنيات التصوير لأهداف مختلفة في تحليل وتنقيب البيانات.

وبالرغم من أن تقنيات التصوير المرئي للبيانات يمكن اعتبارها أحد أساليب التحليل والتنقيب القائمة بذاتها، إلاّ أنه يمكنها أيضًا أن تكون وسيلة إضافية تُستخدم لتحضير البيانات للتحليل والتنقيب، وذلك مثلما يحدث عند استخدامها في إظهار الأنماط المستكشفة في أدوات التنقيب الأخرى.

وبشكل عام، تُعتبر تقنيات التصوير المرئي للبيانات ذات أهمية كبيرة في هذا العصر، ومسار مهم من مسارات دراسات البحث العلمي العصري وتحليل وتنقيب البيانات، بخاصة مع انتشار استخدام التكنولوجيا في مختلف المجالات.

قياس تشابه واختلاف البيانات

في معظم تطبيقات تحليل وتنقيب البيانات، مثل التجزئة العنقودية وخوارزمية الجار الأقرب، تظهر الحاجة إلى طرق لتقييم مدى التشابه والاختلاف فيما بين البيانات، مثلاً قد يحتاج أحد المراكز التجارية إلى تجزئة زبائنه إلى مجموعات ذات خصائص مميزة، كأن يقوم بتجميع الزبائن المتشابهين في الدخل أو العمر، بحيث يمكن استخدام هذه المعلومات في وضع خطط واستراتيجيات التسويق المختلفة التي تستهدف فئات وشرائح معينة من الجمهور.

ففي خوارزميات التجزئة العنقودية يتم تصنيف البيانات وتقسيمها إلى فئات جزئية بحيث تنتمي العناصر المتشابهة لأحد الفئات وبقية العناصر التي تختلف عنها تنتمي للفئات الأخرى، وكذلك الأمر في خوارزمية الجار الأقرب فهي تبحث في التشابه بين العنصر المستكشف والعناصر الأقرب له، لذا فإنه يلزم معرفة تشابه البيانات من أجل تلك الأغراض.

مثلاً في أحد الخوارزميات التي تستكشف أنواع الأمراض وتشخيصها قد يتم اللجوء لبحث التشابه في الأعراض التي تظهر على أحد المرضى مع الأعراض التي تظهر على مرضى آخرين مصابين بالفعل بمرض معين، بحيث يتم استنتاج إصابته بنفس المرض، أو حتى استعداده للإصابة به، في حال وجود تشابه بالفعل بينه وبين المرضى الآخرين في سمات معينة، وبالتالي التوصل إلى التشخيص الصحيح لحالة المريض.

في هذا القسم سوف نتطرق لطرق قياس التشابه والاختلاف بين البيانات باستخدام أساليب تقريبية.

طرق قياس التشابه والاختلاف بين البيانات

إن مقياسي التشابه والاختلاف بين البيانات مرتبطان ببعضهما البعض. ومقياس تشابه عنصرين أ، ب يكون مساويًا للصفر، إذا لم يكن بينهما أي تشابه، فكلما زادت هذه القيمة يزداد التشابه بينهما وحتى الوصول إلى درجة التطابق عندما تصل هذه القيمة إلى الواحد الصحيح =١.

أما قياس الاختلاف بين عنصرين أ، ب، فيعمل بطريقة عكسية، فهو يبدأ بقيمة تساوي الصفر إذا كان العنصرين متشابهين لدرجة التطابق، وكلما زاد الاختلاف بينهما يزداد قياس الاختلاف حتى يصل لقيمة الوحد الصحيح، عندما يكونا مختلفين تمامًا. ويمكن الربط بين قيمة القياسين بالمعادلة الرياضية التالية:

قياس التشابه = ١ – قياس الاختلاف

كيفية قياس التشابه بين البيانات

يمكن بسهولة قياس التشابه أو الاختلاف بين البيانات عندما تكون رقمية، ويتم التعبير عن التشابه والاختلاف بطرق مختلفة، مثلاً يمكن قياس التشابه بين أعمار مجموعة من طلاب المدارس ومجموعة أخرى من طلاب الجامعات،

وبذلك تكون المقارنة من النوع الرقمي البحت التي يمكن التعبير عنها بسهولة باستخدام مقاييس النزعة المركزية، كالوسيط والمتوسط الحسابي مثلاً، ومع ذلك فهناك طرق أخرى لقياس التشابه والاختلاف بين البيانات الرقمية سيتم التطرق إليها في هذا الفصل، ولكن قياس تشابه واختلاف البيانات من النوع الاسمي تكون من خلال معرفة عدد السمات المشتركة فيما بينها وعدد السمات غير المشتركة.

تشابه واختلاف البيانات من النوع الاسمي

يمكن قياس الاختلاف بين البيانات الاسمية من خلال المعادلة الرياضية التالية:

قياس الاختلاف = (العدد الإجمالي للسمات – عدد السمات المشتركة) / العدد الإجمالي للسمات.

مثلاً، لو افترضنا أنه لدينا في إحدى قواعد البيانات الخاصة بمركز تجاري السجلات التالية لزبونين أ، ب، حيث:

الزبون	العمر	الدخل	المنطقة	رحلات التسوق	حجم المشتريات
أ	مسن	مرتفع	س	متوسط	متوسط
ب	شاب	متوسط	س	متوسط	متوسط

باعتبار أن جميع البيانات الواردة في الجدول هي من النوع الاسمي، يمكن

حساب الاختلاف بين الزبون أ، والزبون ب، ويكون:

قياس الاختلاف بين أ، ب = (إجمالي عدد السمات – عدد السمات التي يتشابهون بها) / إجمالي عدد السمات

الاختلاف = (٥ – ٣) / ٥ = ٢ / ٥ = ٠,٤

وذلك باعتبار أن كل سمة من السمات الخمسة الواردة في الجدول هي سمة اسمية تميز الزبون عن غيره من الزبائن.

استخدام نظام الأوزان في قياس التشابه والاختلاف بين البيانات

يمكن تخصيص أوزان أثقل نسبياً لبعض السمات حتى تكون مؤثرة بشكل أكبر في قياس التشابه، ففي المثال السابق يمكن مثلاً تخصيص وزن مقداره (١) فقط لكل السمات باستثناء سمة العمر يتم تخصيص الوزن (٢) لها، فيكون بالتالي إجمالي الوزن الكلي للسمات = ٦ بدلاً من ٥، ثم يتم استخدام معادلة حساب الاختلاف التالية:

قياس الاختلاف بين أ، ب = (إجمالي الوزن الكلي للسمات – مجموع أوزان السمات التي يتشابهون بها) / إجمالي الوزن الكلي للسمات

وتكون النتيجة أن يصبح:

قياس الاختلاف بين أ، ب = (٦ – ٣) / ٦ = ٠,٥

أي أن الاختلاف بين أ ، ب زاد عندما تم رفع الوزن النسبي الخاص

بالعمر.

كما يمكن تخصيص أوزان متنوعة للسمات بحسب احتياجات الجهة التي تقوم بالتحليل والتنقيب، وبحسب ما تستشعره من أهمية لتلك السمات والنتائج التي ترغب بالتوصل إليها من دراستها وتحليلها.

قياس تشابه واختلاف البيانات المنطقية أو في النظام الثنائي

يمكن قياس تشابه واختلاف البيانات المنطقية Boolean أو المستخدمة في النظام الثنائي Binary، أي تلك التي تأخذ قيمتين فقط بشكل دائم (١ أو صفر)، وقد تكون تلك القيم مناظرة لقيم ذات معنى في قائمة البيانات التي يتم بحث اختلافها وتشابهها، كأن تكون القيم المناظرة لها هي قيم منطقية (نعم و لا)، أو قيم (موافق وغير موافق)، ويتم قياس التشابه والاختلاف بين هذا النوع من البيانات من خلال المعادلة التالية:

الاختلاف بين المتغير (أ) والمتغير (ب) = (س + ص) / (ل + س + ص + ع)

حيث:

س = عدد السمات التي تكون قيمتها (١) للمتغير (أ) وقيمتها (صفر) للمتغير (ب)

ص = عدد السمات التي قيمتها (صفر) للمتغير (أ) وقيمتها (١) للمتغير (ب)

ل = عدد السمات التي قيمتها (١) للمتغير (أ) وقيمتها (١) للمتغير (ب)

ع = عدد السمات التي قيمتها (صفر) للمتغير (أ) وقيمتها (صفر) للمتغير (ب)

مع ملاحظة أن إجمالي عدد السمات هو (ن)، حيث:

ن = ل + س + ص + ع

ومن المتعارف عليه في بعض مجالات بحوث تشابه واختلاف البيانات من النوع المنطقي أو الثنائي أن يتم إهمال عدد السمات السلبية المشتركة (ع)، أي عدد السمات التي تأخذ القيمة (صفر) في كلا المتغيرين، وذلك باعتبار أنه ليس لها دلالة ذات تأثير حقيقي، في حين أنه لا يتم إهمال عدد السمات الإيجابية المشتركة (ل) أي عدد السمات التي قيمتها (١) في كلا المتغيرين، حيث يكون لهذا العدد دلالة ذات تأثير مهم لحساب الاختلاف بين البيانات. ويظهر هذا الأمر بشكل واضح في الاختبارات الطبية التي تتم على المرضى والهادفة لاستكشاف استجابتهم لأنواع مختلفة من الأدوية المخصصة للعلاج، حيث أن عدم الاستجابة من جميع المرضى هو أمر يتم إهماله واستبعاده من دراسة الاختلاف فيما بينهم ولا يغير من درجة تقييم فاعلية الدواء الذي يتم دراسته.

وبذلك تصبح المعادلة الخاصة بحساب الاختلاف بين البيانات الأكثر شيوعًا في البيانات المنطقية وبيانات النظام الثنائي كما يلي:

الاختلاف بين المتغير (أ) والمتغير (ب) = (س + ص) / (ل + س + ص)

مثال تطبيقي

نفرض أنه لدينا جدول السجلات التالية، والخاصة بمجموعة من المرضى في أحد المراكز الطبية، والذي يبين الإصابة بأعراض الحمى والسعال ونتائج أربعة اختبارات طبية محددة بقيمتين وهما القيمة (١) المناظرة للنتيجة الإيجابية Positive، والقيمة (صفر) المناظرة للنتيجة السلبية Negative، والجدول كما يلي:

المريض	الجنس	حمى	سعال	اختبار١	اختبار٢	اختبار٣	اختبار٤
أ	ذكر	نعم	لا	١	صفر	صفر	صفر
ب	أنثى	نعم	نعم	صفر	صفر	صفر	صفر
ج	ذكر	نعم	لا	١	صفر	١	صفر
....							

وبفرض أن حساب الاختلاف بين المرضى سوف يتم باستخدام الطريقة الثانية، التي يتم فيها إهمال السمات المشتركة ذات القيم السلبية حيث أنها تكون غير مهمة من المنظور الطبي، فإنه يكون:

الاختلاف بين المريض (أ) والمريض (ب) = (س + ص) / (ل + س + ص) = (١ + ١) / (١ + ١ + ١) = ٠,٦٧

الاختلاف بين المريض (أ) والمريض (ج) = (س + ص) / (ل

+ س + ص) = (٠ + ١) / (٢ + ٠ + ١) = ٠,٣٣

الاختلاف بين المريض (ب) والمريض (ج) = (س + ص) / (ل + س + ص) = (١ + ٢) / (١ + ١ + ٢) = ٠,٧٥

ويتضح من النتائج أعلاه أن المريض (ب) والمريض (ج) هما الأكثر اختلافًا، وبالتالي يكون احتمال إصابتهم بنفس المرض هو احتمال ضعيف، ومن بين المرضى الثلاثة فإن المريض (أ) والمريض (ج) هما الأكثر تشابهًا وبالتالي هما الأكثر احتمالاً بأن يُصابوا بنفس المرض.

قياس تشابه واختلاف البيانات من النوع الرقمي

يمكن قياس تشابه واختلاف البيانات من النوع الرقمي Numerical باستخدام معادلة المسافة الإقليدية، وهي المعادلة التي تحسب المسافة بين نقطتين في الفراغ متعدد الأبعاد والمنسوبة للعالم "إقليدس"، والتي يتم التعبير عنها رياضيًا كما يلي:

المسافة بين النقطتين (أ) و (ب) =

$$\sqrt{[(س_{أ} - س_{ب})^2 + (ص_{أ} - ص_{ب})^2 + (ع_{أ} - ع_{ب})^2 + \cdots\cdots]}$$

حيث أن كل من السجلات أ، ب يتم التعبير عنها باستخدام السمات س، ص، ع، ……، أي أن:

أ = (س أ، ص أ، ع أ، ……)

ب = (س ب، ص ب، ع ب، .....)

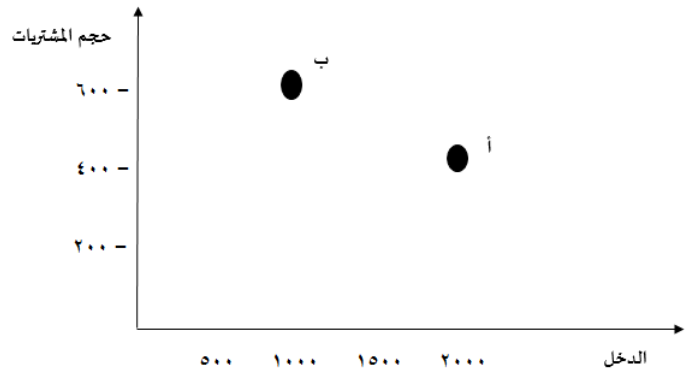
وحتى يتم قياس التشابه والاختلاف بين البيانات الرقمية فإنه يلزم تحويلها أولاً إلى قيم ضمن فترة محددة ما بين القيمتين (-١، +١)، أو بين القيمتين (صفر، ١) وذلك من أجل حساب مقدار التشابه أو الاختلاف فيما بينها بما يتناسب مع قيمة التشابه والاختلاف أصلاً والتي تقع ضمن تلك الحدود، ويتم ذلك باستخدام أسلوب التدريج الذي يناظر القيمة الحقيقية للبيانات، وسوف يتم التطرق لهذه الطرق بالتفصيل في الفصل الخاص بتحضير البيانات للتحليل والتنقيب، ومع ذلك سيتم استخدام هذه الطريقة في المثال التطبيقي التالي.

مثال تطبيقي

نفرض أنه لدينا في قاعدة بيانات أحد الشركات التجارية سجلات بعض الزبائن كما يلي:

الزبون	العمر	الدخل	المنطقة	حجم المشتريات
أ	٥٥	٢٠٠٠	م	٤٠٠
ب	٣٥	١٠٠٠	م	٦٠٠

ويمكن تمثيل بيانات كل من الدخل وحجم المشتريات للزبونين أ، ب بالرسم الديكارتي كما في الشكل التالي:



شكل (٢,١٨)، تمثيل بيانات الدخل وحجم المشتريات لزبونين أ، ب

حيث يمثل المحور (س) قيمة الدخل، ويمثل المحور (ص) حجم المشتريات. والمطلوب حساب الاختلاف بين السجل (أ) والسجل (ب) من حيث الدخل وحجم المشتريات، وهي القيم الرقمية التي وردت في الجدول.

وتكون قيمة الاختلاف بين النقطتين (أ)، (ب) من منظور هندسي هي المسافة التي تفصل بينهما على الرسم.

وحتى يمكن قياس الاختلاف بينهما من منظور نسبي تتراوح قيمته ما بين (صفر، ١) فإنه ينبغي أولاً تحويل البيانات إلى قيم تقع بداخل الفترة المحددة (صفر ، ١)، وذلك كما يلي:

بالنسبة لقيمة الدخل، فإن أعلى قيمة على التدريج = ٢٠٠٠، ولتكن هذه القيمة هي المناظرة للقيمة (١) على التدريج.

أما أصغر قيمة فتكون هي المناظرة للقيمة (صفر) على التدريج، ويمكن هنا اعتبار أن هذه القيمة هي التي تمثل الصفر الحقيقي فعلاً، فيكون بالتالي قيمة الدخل (١٠٠٠) الواردة في الجدول تناظر القيمة (٠,٥) على التدريج، حيث أنها تقع في منتصف المسافة بين صفر و٢٠٠٠.

وبالمثل يمكن تحديد قيم حجم المشتريات لكل من الزبون (أ)، (ب) بحيث تكون أعلى قيمة لحجم المشتريات وهي (٦٠٠) مناظرة للقيمة (١) على التدريج، وبالتالي تكون القيمة (٤٠٠) مناظرة للقيمة (٠,٦٧) على التدريج.

ولحساب الاختلاف بين (أ) و (ب) فإنه يكون مساوياً للمسافة بين النقطتين (أ) و (ب) حيث يتم التعبير عنهما بقيمة كل من الدخل وحجم المشتريات على التدريج كما يلي:

أ = (١، ٠,٦٧)

ب = (٠,٥، ١)

ويكون بحسب معادلة حساب قيمة الاختلاف التي تساوي المسافة بينهما:

$$\text{الاختلاف بين أ، ب} = \sqrt{[(١ - ٠,٥)^{٢} + (١ - ٠,٦٧)^{٢}]}$$

$$= \sqrt{[(٠,٥)^{٢} + (٠,٣٣)^{٢}]}$$

$$= ٠,٦$$

قياس تشابه واختلاف البيانات من النوع الرتبي

يمكن قياس تشابه واختلاف البيانات من النوع الرتبي Ordinal، وذلك باعتبار أن الترتيب فيها يكون له دلالة من حيث القيمة، مثلاً كما يحدث في حالة ترتيب درجات الحرارة لمدينة ما بأنها تأخذ القيم التالية (منخفضة، متوسطة، مرتفعة)، فبالرغم من عدم وجود قيمة رقمية لمثل تلك البيانات إلاّ أنها قابلة للترتيب تصاعديًا من الأصغر للأكبر أو تنازليًا من الأكبر للأصغر.

وحتى يتم قياس التشابه والاختلاف بين البيانات من النوع الرتبي فإنه يلزم تحويلها أولاً إلى بيانات من النوع الرقمي تقع ضمن تدريج محدد في الفترة (صفر، ١)، أي بنفس الطريقة التي تم فيها تحويل البيانات الرقمية، والفرق في هذه الحالة أنه في حالة البيانات الرتبية يلزم أولاً تحديد أوزان رقمية لكل قيمة من القيم حتى نستطيع أن نعطيها القيم المناظرة لها في التدريج على الفترة المحددة.

مثلا في حالة درجات الحرارة، يمكن افتراض أن القيم الحقيقية للدرجات كما يلي:

المنخفضة: هي التي تتراوح بين صفر و١٥ درجة مئوية
المتوسطة: هي التي تتراوح بين ١٦ درجة و٣٠ درجة مئوية
المرتفعة: هي التي تتراوح بين ٣١ درجة و٥٠ درجة مئوية

أما في الخطوة الثانية فيتم تحويل قيمة الدرجات إلى القيم المناظرة في التدريج، مثلاً لو افترضنا أن التدريج يبدأ بالقيمة (صفر) المناظرة للقيمة صفر درجة مئوية، وينتهي بالقيمة (١) المناظرة للقيمة (٥٠) درجة مئوية، فتكون قيمة (٣٠) درجة مئوية مناظرة للقيمة (٠,٦) على تدريج الفترة.

وفي الخطوة الأخيرة يتم حساب قيمة الاختلاف بين البيانات بنفس الطريقة التي تم اتباعها في حساب الاختلاف بين البيانات الرقمية، أي باستخدام معادلة المسافة الإقليدية بين نقطتين على الرسم الديكارتي، كما يلي:

الاختلاف بين المتغيرين (أ)، (ب) = المسافة بين النقطتين (أ) و (ب)

$$= \sqrt{[(س_{١} - س_{٢})^{٢} + (ص_{١} - ص_{٢})^{٢}]}$$

حيث أن: أ = (س١، ص١)، ب = (س٢، ص٢)

قياس تشابه واختلاف بيانات النصوص

يتم قياس تشابه واختلاف بيانات النصوص Text Similarity من خلال تحديد مجموعة من الكلمات الأساسية تُسمى الكلمات المفتاحية التي تظهر في تلك النصوص، والتي تُعتبر كلمات مميزة ولها دلالة معينة بالنسبة للهدف من المقارنة التي يجري تنفيذها، بعد ذلك تتم عملية إحصاء لعدد مرات تكرار ظهور تلك الكلمات في كل نص من النصوص التي يتم مقارنتها وبحث تشابهها واختلافها، ثم توضع النتائج في جدول مخصص كالمبين أدناه، وهو جدول يوضح عدد تكرار مجموعة من الكلمات المفتاحية في أربعة بحوث

متخصصة في تطوير التعليم بهدف استكشاف التشابه والاختلاف فيما بينها:

البحث	الكلمات المفتاحية									
	بحث	علوم	تربية	تعليم	تقنية	تطوير	تحديث	ابتكار	إبداع	تقدم
بحث (أ)	٥	٠	٣	٠	٢	٠	٠	٢	٠	٠
بحث (ب)	٣	٠	٢	٠	١	١	٠	١	٠	١
بحث (ج)	٠	٧	٠	٢	١	٠	٢	٣	٠	٠
بحث (د)	٠	١	٠	٠	١	٢	٠	٠	٣	٠

بعد ذلك يتم استخدام حساب التشابه بين النصوص من خلال تطبيق المعادلة الرياضية التالية:

التشابه بين النص أ والنص ب = (أ × ب) / || أ || × || ب ||

حيث:

أ × ب = مجموع حاصل الضرب الديكارتي لعدد تكرار الكلمات المفتاحية في كل نص

|| أ || = $\sqrt{[(\text{تكرار الكلمة الأولى})^2 + (\text{تكرار الكلمة الثانية})^2 + \dots]}$ في النص أ

|| ب || = $\sqrt{[(\text{تكرار الكلمة الأولى})^2 + (\text{تكرار الكلمة الثانية})^2 + \dots]}$ في النص ب

مثال تطبيقي

في الجدول المبين في الشكل أعلاه، يمكن حساب التشابه بين كل من

البحث الأول والبحث الثاني، حيث يتم التعبير عنهما بالمتغيرين (أ)، (ب) ويتم التعبير عن كل منهما من خلال عدد مرات ظهور الكلمات المفتاحية فيهما على الترتيب كما يلي:

أ = (٥، ٠، ٣، ٠، ٢، ٠، ٠، ٢، ٠، ٠)

ب = (٣، ٠، ٢، ٠، ١، ١، ٠، ١، ٠، ١)

ويكون:

أ × ب = (٥ × ٣ + ٠ × ٠ + ٣ × ٢ + ٠ × ٠ + ٢ × ١ + ٠ × ٠ + ١ × ٠ + ٠ × ٠ + ١ × ٠) = ٢٥

$$\| \text{أ} \| = \sqrt{(٥^٢ + ٠^٢ + ٣^٢ + ٠^٢ + ٢^٢ + ٠^٢ + ٠^٢ + ٢^٢ + ٠^٢ + ٠^٢)}$$

= ٦,٤٨

$$\| \text{ب} \| = \sqrt{(٣^٢ + ٠^٢ + ٢^٢ + ٠^٢ + ١^٢ + ١^٢ + ٠^٢ + ١^٢ + ٠^٢ + ١^٢)}$$

= ٤,١٢

وبتطبيق معادلة حساب التشابه بين البحث (أ) والبحث (ب)، فيكون:

التشابه بين البحث (أ) والبحث (ب) = ٢٥ / (٦,٤٨ × ٤,١٢)

= ٠,٩٤

أي أن البحث (أ) والبحث (ب) متشابهان بنسبة ٩٤ % تقريبًا وفق هذه الطريقة في القياس.

الفصل الثالث
تحضير البيانات للتحليل والتنقيب
Data Preprocessing

المحتويات

١. أهمية تحضير البيانات للتحليل والتنقيب

٢. تنظيف البيانات Data Cleaning

٣. دمج البيانات Data Integration

٤. اختزال البيانات Data Reduction

٥. تحويل البيانات وتفريد البيانات Data Transformation & Data Discretization

أهمية تحضير البيانات للتحليل والتنقيب

إن التضخم الكبير الذي أضحت عليه قواعد البيانات في هذا العصر يجعلها عرضة لاحتواء الكثير من البيانات المزعجة أو غير المتناسقة أو حتى فقدان بعض البيانات الموجودة فيها، وذلك بسبب ضخامتها وتدفقها من مصادر متعددة.

والبيانات ذات الجودة المنخفضة سوف تؤدي بطبيعة الحال إلى نتائج بجودة منخفضة أيضاَ عند تحليلها وتنقيبها، ومن أجل ذلك ينبغي رفع جودة البيانات أولاً ومن ثم يمكننا أن نتوقع ارتفاع كفاءة التحليل والتنقيب وتيسير عملياتها لتكون بالشكل الأمثل.

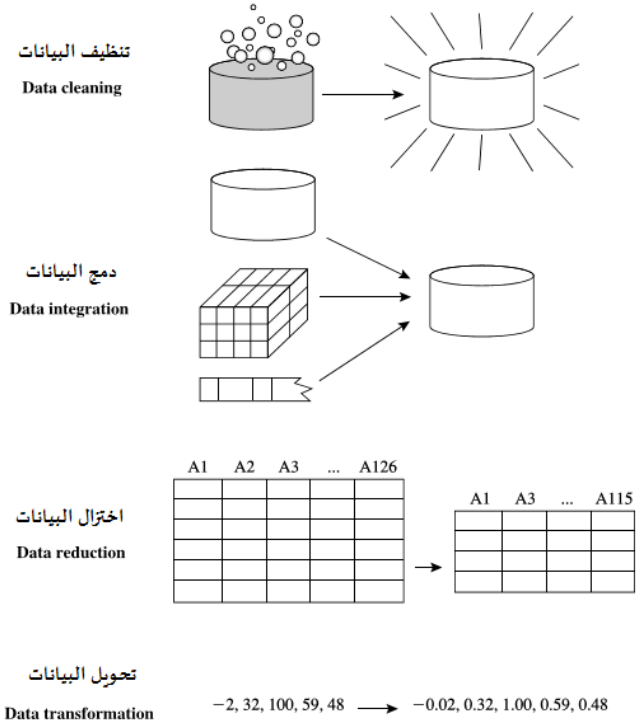
توجد عدة تقنيات مخصصة لتحضير البيانات للتحليل والتنقيب، وهي كما يلي:

١. تنظيف البيانات Data Cleaning: تعبئة البيانات المفقودة وإزالة الضوضاء والتعارض أو التنافر في البيانات.
٢. دمج البيانات Data Integration: دمج مجموعات مختلفة من البيانات من مصادر مختلفة في مخزن موحد للبيانات، مثل مستودعات البيانات Data Warehouse.
٣. اختزال البيانات Data Reduction: تقليص حجم البيانات من خلال عمليات التجميع أو اختصار التكرار في السمات أو بالتجزئة

العنقودية.

٤. تحويل البيانات وتفريد البيانات Data Transformation & Data Discretization: تحويل البيانات من صورة إلى صورة أو تفريدها وتسويتها وإعطائها قيم رقمية على تدريج معين عند الحاجة.

في هذا الفصل سوف يتم التطرق لكل أسلوب من هذه الأساليب بشيء من التفصيل.



شكل (٣,١)، أساليب مختلفة لتحضير البيانات للتحليل والتنقيب

تنظيف البيانات
Data Cleaning

تشتمل عمليات تنظيف البيانات على إجراءات متنوعة بحسب الحاجة، منها ما يهدف لاستكمال البيانات المفقودة ومنها ما يهدف لتنظيف البيانات المزعجة وغير المتناسقة. ويمكن توصيف تلك الإجراءات بالتفصيل كما يلي:

١. القيم المفقودة Missing Values

يمكن معالجة مشكلة القيم المفقودة في قاعدة البيانات بعدة طرق، ففي قاعدة بيانات زبائن أحد المراكز التجارية مثلاً قد يكون هناك فقدان لبعض بيانات الزبائن مثل عمر الزبون، ويمكن معالجة هذا الأمر بإحدى الطرق التالية:

أ. تجاهل الصفوف التي توجد فيها قيم مفقودة

يتم بهذه الطريقة تجاهل الصفوف التي تعاني من فقدان بعض البيانات فيها، وهي طريقة غير فعّالة، إلاّ إذا كان الصف به بيانات كثيرة مفقودة.

وقد تؤثر هذه الطريقة على مجمل نتائج التحليل والتنقيب المزمع تنفيذها إذا كانت هناك بيانات كثيرة مفقودة في عدد كبير من الصفوف (السجلات) في قاعدة البيانات.

ب. تعبئة البيانات المفقودة يدويًا

بشكل عام يمكن اعتبار هذه الطريقة أنها مضيعة للوقت والجهد، كما أنه قد يستحيل القيام بها في حالة الكم الهائل من البيانات التي بها عدد كبير من القيم المفقودة، وقد يتم استخدامها فقط في حالة فقدان عدد صغير من البيانات.

ت. استخدام ثابت موحد بدلًا من القيم المفقودة

يمكن بهذه الطريقة استبدال جميع القيم المفقودة في أحد الحقول بقيمة ثابتة وموحدة أو تسمية مثل "غير معروف"، ولكن في هذه الحالة عند إجراء عمليات التحليل وتنقيب البيانات سوف يعتقد برنامج التنقيب أن هذه التسمية الشائعة لها مدلول مهم بخاصة إذا كانت البيانات المفقودة كبيرة نسبيًا، وسوف تظهر هذه الدلالة في نتائج التحليل، وهو ما قد يؤدي إلى إضعافها.

ث. استخدام أحد قيم مقاييس النزعة المركزية بدلًا من القيم المفقودة

تُستخدم هذه الطريقة لملئ القيم المفقودة إذا كانت من النوع الرقمي، باستخدام أحد مقاييس النزعة المركزية، كالوسيط أو المتوسط الحسابي، فمثلاً إذا كان لدينا قاعدة بيانات زبائن أحد المراكز التجارية وفيها فقدان لبيانات بعض الزبائن كعمر الزبون، فيمكن حساب متوسط أعمار جميع الزبائن في قاعدة البيانات واستخدامه في ملئ القيم المفقودة في خانة العمر.

وبطبيعة الحال سوف تؤدي هذه الطريقة إلى تعزيز تلك القيمة في قاعدة

البيانات وزيادة تكرارها غير المرغوب لعدد أكبر من الزبائن، وهو ما قد يؤثر على نتائج التحليل والتنقيب.

ج. استخدام أحد قيم مقاييس النزعة المركزية لفئة البيانات التي تنتمي لها القيم المفقودة

كمحاولة للحصول على نتائج أكثر دقة وأقرب إلى الواقع، يمكن استخدام هذه الطريقة لملئ القيم المفقودة في قاعدة البيانات من خلال تصنيف الزبائن أولاً لفئات مختلفة، بحسب حجم مشترياتهم مثلاً، ثم ملئ خانة العمر المفقودة في كل فئة بالمتوسط الحسابي أو الوسيط الخاص بأعمار جميع الزبائن الذين ينتمون لتلك الفئة.

مثلاً، إذا كانت لدينا قيمة مفقودة في قاعدة البيانات لأحد الزبائن كالعمر مثلاً، وكان حجم المشتريات لديه محدد بقيمة معينة، فإنه يمكن ملئ خانة العمر بنفس قيمة الوسيط أو المتوسط الحسابي لأعمار الزبائن الذين يتشابهون معه في حجم المشتريات، أو بمعنى آخر ينتمون لهذه الفئة من الزبائن المميزين بهذا الحجم من المشتريات.

ح. استخدام القيمة الأكثر احتمالًا بالتنبؤ لتعبئة القيم المفقودة

ويتم ذلك من خلال أساليب أكثر تعقيدًا، ومن خلال تقنيات التنقيب المتخصصة، مثل شجرة القرار، التي تهدف للتنبؤ من خلال تنقيب البيانات الأخرى الموجودة والمتوفرة واكتشاف القيم المفقودة والتنبؤ بها بحسب نتائج

التحليل.

ويلاحظ في الطرق (ت، ث، جـ)، أنها قد تؤدي في الكثير من الأحيان إلى تعبئة القيم المفقودة بقيم غير صحيحة، أما الطريقة رقم (ح) فهي إستراتيجية شائعة الاستخدام، ومقارنة بالطرق الأخرى فهي تستخدم المعطيات الموجودة فعلاً من أجل التنبؤ بالقيم المفقودة، ونظراً لصحة كل المعلومات المستخدمة فإنه يمكن التنبؤ بقيم قريبة جدًا من القيم الحقيقية المفقودة كلما كان أسلوب التحليل والتنبؤ يعتمد على عدد أكبر من السمات والبيانات المتوفرة في قاعدة البيانات.

ملاحظات

قد لا يؤثر فقدان بعض البيانات على نتائج تحليلها وتنقيبها، ويحدث ذلك عندما تكون البيانات المفقودة غير ذات أهمية بالنسبة لأهداف التحليل والتنقيب، كأن يكون هناك فقدان لبيانات رخصة القيادة الخاصة بالزبون، سواء كان بسبب إهماله أو نسيانه لها أو حتى لأنه لا يمتلك رخصة قيادة أصلاً، وفي كل الأحوال سوف يترك هذا الحقل فارغًا، كما قد تلجأ بعض الشركات إلى وضع إجراءات محكمة لضبط عملية إدخال البيانات بحيث لا تسمح بوجود بيانات فارغة في حقول محددة في سجلات قاعدة البيانات إذا كانت ذات أهمية إستراتيجية ومهمة لعمليات التحليل والتنقيب.

وبالرغم من كل الجهود التي تُبذل من أجل تنظيف البيانات واستكمال

البيانات المفقودة في بعض حقول سجلات قواعد البيانات، إلاّ أن أفضل طريقة للحصول على البيانات الصحيحة والكاملة في قاعدة البيانات هي تصميم الإجراءات اللازمة لضمان صحة البيانات المُدخلة في البرامج أثناء عملية الإدخال وهي الوسيلة الأكثر كفاءة وفاعلية للحصول على بيانات صحيحة ومكتملة مهما تعددت طرق وأساليب إدخالها ومهما تنوعت قدرات ومهارات مدخلي البيانات، وذلك من خلال إظهار رسائل التنبيه التي تطلب ضرورة إكمال البيانات المفقودة أو تصحيح الخاطئ منها قبل الانتقال إلى السجل التالي أثناء عملية الإدخال.

٢. البيانات المزعجة Noisy data

الإزعاج في البيانات هو خطأ عشوائي أو تعدد قيم أحد المتغيرات التي يتم تحديدها بطرق القياس المختلفة بشكل يكون مزعجًا بشكل ما لعمليات التحليل والتنقيب، ويظهر مثل هذا الإزعاج عندما يتم استخدام تقنيات الإحصاء أو تقنيات تصوير البيانات، كما يمكن أن يُظهر قيمًا متطرفة تمثل أيضًا شكل من أشكال الإزعاج.

مثلاً، في قاعدة بيانات مبيعات أحد الشركات، فإن المتغير الخاص بسعر المنتج، وهو من النوع الرقمي، يكون له عدد كبير من القيم المختلفة نظرًا لاختلاف أسعار المنتجات المتعددة وندرة تطابق أسعارها، بحيث يُصعّب هذا الأمر من عملية التعبير عن هذا المتغير إحصائيًا كما أنه يؤثر سلبًا على

إجراءات التنقيب، لذا يتم اللجوء إلى ما يسمى عملية تنعيم للبيانات Data Smoothing، وذلك من أجل جعلها متجانسة وإزالة الإزعاج الموجود فيها.

فلو افترضنا مثلاً أن القيم التالية تمثل أسعار بعض المنتجات المتنوعة في أحد الشركات التجارية:

١١، ١٥، ٢٢، ٢٤، ٢٦، ٣١، ٣٦، ٤١، ٤٣

فعند النظر إلى هذه القيم من منظور إحصائي بهدف تطبيق إجراءات التحليل والتنقيب عليها يُلاحظ أن عددها كبير نسبياً وقد يرغب المحلل بإجراء تعديل عليها قبل بدء عمليات التحليل والتنقيب، بهدف تنعيمها وجعلها محدودة العدد بما يساهم في رفع كفاءة التحليل والتنقيب وتحقيق أهداف محددة، وتشبيه هذا الأمر ما يحدث عند استبدال هذه القيم بقيم جديدة تعبر عن مستوى سعر المنتج بأن يكون أحد المستويات الثلاثة (منخفض، متوسط، مرتفع) فقط كنوع من التنعيم. ومن ثم يتم استخدام هذه القيم الثلاثة في عمليات التحليل والتنقيب من أجل التوصل لحقائق مفيدة مبسطة وواضحة حول حجم المبيعات ومدى الإقبال على شراء هذه المنتجات، كما أن التنعيم بهذه الكيفية، التي نتج عنها تحول في نوع البيانات وإنشاء سمة جديدة، يندرج تحت أسلوب تحويل البيانات الذي سوف يتم التطرق إليه بالتفصيل في الأجزاء التالية من هذا الفصل.

وقد تكون عملية التنعيم محدودة بطريقة لا ينتج عنها تحول في نوعية البيانات أو إنشاء سمات جديدة، ويتم ذلك بأحد الطرق التالية:

أ. طريقة التكييس (التجميع في السلات) Binning

وهي طريقة لتنعيم القيم المتعددة لأحد المتغيرات من خلال تحديد قيم جديدة ومحدودة العدد، بحيث تكون من نفس النوع، وذلك من خلال استشارة القيم المجاورة لها أو المحيطة بها.

يتم في هذه الطريقة أولاً توزيع القيم المتوفرة في عدد من السلات، بشكل متساوٍ، يكون في كل سلة عدد متساوي من القيم، ثم يتم في الخطوة التالية تنعيم القيم بأن تأخذ جميعها قيم جديدة بأحد الطرق التالية:

١. يتم حساب قيمة المتوسط الحسابي أو الوسيط لكل القيم الموجودة في كل سلة ثم يتم استبدال كل القيم الواردة فيها بقيمة المتوسط الحسابي أو الوسيط.

٢. يتم تحديد القيمتين العليا والصغرى في كل سلة، ثم يتم استبدال كل قيمة وردت في السلة بالقيمة الأقرب لها من القيمتين العظمى أو الصغرى.

مثال توضيحي

بفرض أنه لدينا القيم التالية التي تمثل أسعار بعض المنتجات المتنوعة في أحد الشركات التجارية:

١١، ١٥، ٢٢، ٢٤، ٢٦، ٣١، ٣٦، ٤١، ٤٣

ولكي نقوم بإجراء عملية تنعيم لهذه البيانات، يتم أولاً توزيعها على ثلاثة سلات، كل سلة تحتوي على ثلاثة قيم فقط كما يلي:

السلة الأولى: ١١، ١٥، ٢٢

السلة الثانية: ٢٤، ٢٦، ٣١

السلة الثالثة: ٣٦، ٤١، ٤٣

وفي الخطوة الثانية يتم استبدال القيم الأصلية بالقيم الجديدة كما يلي وبحسب كل حالة:

١. استبدال القيم الأصلية بقيمة المتوسط الحسابي للقيم في كل سلة:

السلة الأولى: ١٦، ١٦، ١٦

السلة الثانية: ٢٧، ٢٧، ٢٧

السلة الثالثة: ٤٠، ٤٠، ٤٠

٢. استبدال القيم الأصلية بالقيمة الأقرب لها من القيم الصغرى والعظمى:

السلة الأولى: ١١، ١١، ٢٢

السلة الثانية: ٢٤، ٢٤، ٣١

السلة الثالثة: ٣٦، ٤٣، ٤٣

ويتضح من هذا المثال كيف تم تنعيم البيانات الذي تمثل في تقليل عدد القيم المختلفة لأسعار المنتجات، حيث أنها أصبحت (٣) قيم فقط في الحالة الأولى و(٦) قيم فقط في الحالة الثانية بعد أن كانت (٩) قيم في المعطيات الأصلية.

ويُلاحظ أنه كلما اتسع الفارق بين القيم الأصلية والقيم المستبدلة بها يزداد تأثير هذه العملية، وبالمقابل يمكن للسلات أن تحتوي على قيم متساوية إذا

كان معدل الفرق بين القيم فيها ثابت، وتجدر الإشارة إلى أن هذه الطرق تُستخدم أيضًا في تقنيات تفريد البيانات Data Discretization، وهو شكل من أشكال تحويل البيانات الذي سيتم التطرق إليه بالتفصيل لاحقًا.

ب. تقنية الانحدار Regression

إن إجراءات تنعيم البيانات Data Smoothing يمكن تطبيقها أيضًا باستخدام تقنية الانحدار، وهي تقنية مهمتها إسناد قيمة المتغير محل البحث إلى قيمة متغير آخر مستقل باستخدام دالة رياضية، وقد تكون هذه الدالة من النوع الخطي التي تعتمد على متغير مستقل واحد فقط إضافة للمتغير التابع الذي يتم تحديد قيمته الجديدة، ويسمى الانحدار عندئذٍ بالانحدار الخطي، وعند معرفة المتغير المستقل يمكن التنبؤ بقيمة المتغير التابع باستخدام الدالة الرياضية.

ويمكن تمثيل الدالة الرياضية بالمعادلة الشهيرة التالية:

س = أ + ب × ص

حيث:

س: هو المتغير التابع الذي يتم التنبؤ بقيمته.

ص: هو المتغير المستقل الذي تتم معرفة قيمته.

أ، ب: ثوابت تحقق شروط المعادلة.

أما الانحدار متعدد الأبعاد فهو عبارة عن دالة رياضية تشتمل على عدة متغيرات، ويتم تمثيلها بمساحة متعددة الأبعاد في الفراغ بدلاً من الخط

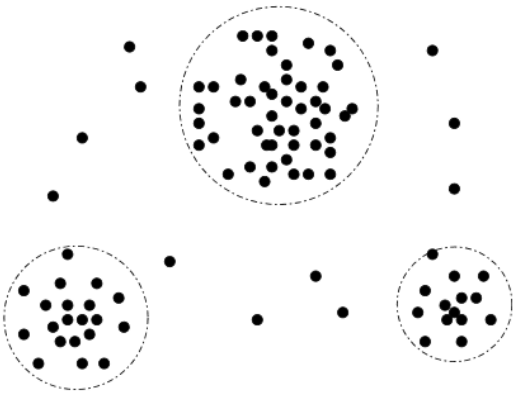
المستقيم، ويتم حساب قيمة المتغير التابع بمعرفة قيمة كل المتغيرات التي يعتمد عليها في الدالة الرياضية.

ج. تحليل القيم المتطرفة Outliers

عندما يتم تجميع البيانات في مجموعات متجانسة باستخدام تقنيات التجزئة العنقودية Clustering، والتي سيتم التطرق إليها بالتفصيل لاحقًا، يمكن ملاحظة وقوع بعض القيم خارج تلك المجموعات بحيث يتم اعتبارها قيمًا متطرفة Outliers.

وتتعدد أساليب التعامل مع مثل هذه القيم بحسب الهدف من كل أسلوب، ففي إجراءات تنظيف البيانات قد يلزم إهمال هذا النوع من البيانات وعدم أخذها بالاعتبار كونها بيانات مزعجة، وذلك حتى لا تؤثر على نتائج التحليل والتنقيب، ومع ذلك فقد يلزم أخذها بالاعتبار وعدم إهمالها حتى وإن كان لها تأثير على نتائج التحليل والتنقيب، وفي كل الأحوال فإنه يلزم تحليل هذه القيم وبحث أسباب تطرفها وما إذا كانت ناتجة عن أخطاء في إدخال البيانات أو أي أسباب أخرى، بحيث يتم تحديد طريقة التعامل معها وما إذا كان سيتم اعتبارها بيانات مزعجة ينبغي تنظيفها أو بيانات مهمة ينبغي أخذها بالاعتبار.

وسوف يتم التطرق بالتفصيل لهذا النوع من البيانات لاحقًا في فصل تحليل وتنقيب القيم المتطرفة.



شكل (٣,٢)، القيم المتطرفة

الخلاصة

كثيرًا من طرق تنعيم البيانات التي تهدف لجعلها بيانات متجانسة تُستخدم أيضًا في تفريد البيانات واختزال البيانات، مثلاً تقنية التكييس تقوم بتخفيض عدد القيم المختلفة لكل متغير، وهي عملية تشبه اختزال البيانات من أجل استخدامها في التنقيب، بخاصة المتغيرات ذات القيم المنطقية.

كذلك يمكن للأسلوب الهرمي في تفريد البيانات أن يعتمد على التجانس، مثلاً يمكن ربط قيم أسعار المنتجات بثلاثة قيم محددة بحيث تكون كما يلي: (سعر مرتفع، سعر متوسط، سعر منخفض)، ويتم بذلك استبدال جميع القيم

بالقيم الجديدة تمهيدًا لإجراء عمليات التحليل والتنقيب عليها.

من جهة أخرى، تُستخدم تقنيات كثيرة بهدف ضبط الإزعاج في البيانات قبل وقوعه، ومن أهم هذه التقنيات هي ما يتعلق بإجراءات وضوابط إدخال البيانات التي تُستخدم في التحكم بعملية الإدخال ومنع ارتكاب الأخطاء أثناء تلك العملية من المستخدمين مهما تنوعت مهاراتهم وقدراتهم ودقة وكفاءة أداءهم لعملية الإدخال، وبذلك تكون هي الوسيلة المثلى لمنع الإزعاج قبل وقوعه أصلاً بحسب مقولة أن الوقاية خير من العلاج.

ومن أمثلة هذه الإجراءات ما يتم وضعه من قيود أثناء عملية الإدخال على بعض المتغيرات التي تأخذ قيمًا محددة أو تقع ضمن نطاق معين، مثلاً في قاعدة بيانات زبائن أحد الشركات يمكن تحديد متغير "عمر الزبون" بأنه ينبغي أن يكون ضمن نطاق محدد من صفر إلى ١٠٠، ويتم ضبط إدخال هذه القيمة برسالة تنبيه تظهر لمدخل البيانات عندما يقوم بإدخال قيمة خاطئة لا تقع ضمن ذلك النطاق، وعندما يقوم المستخدم بإدخال القيمة (٤٤١) بدلاً من (٤١)، تظهر له رسالة تنبيه وتمنعه من استكمال العملية إلى أن يقوم بتصحيحها.

دمج البيانات
Data Integration

تتطلب عمليات تحليل وتنقيب البيانات في كثير من الأحيان إجراء عمليات دمج للبيانات من عدة مصادر، والدمج الجيد يساعد في اختصار البيانات وتجنب التكرارات الزائدة أو التناقضات، وهو ما يساهم في زيادة كفاءة وسرعة ودقة إجراءات التحليل والتنقيب.

إن عدم التجانس في بنية ودلالة البيانات يُشكل تحديًا كبيرًا لعمليات دمج البيانات، حيث أنها تتطلب الدمج بطريقة صحيحة مهما تعددت المصادر والبيئات الحاضنة لها بحيث يتم الحصول في النهاية على قاعدة بيانات كاملة ذات هيكلية موحدة وبيانات منظمة وموزعة بدقة في أماكنها المناسبة.

ومن أمثلة عمليات دمج البيانات أن يتم تجميع قواعد البيانات المخصصة لحفظ وتخزين حركة المبيعات في عدة فروع لأحد المراكز التجارية الكبرى، بحيث يتم في هذه العملية تجميع حركات الشراء بشكل صحيح وبطريقة تُظهر كل البيانات في أماكنها المناسبة، مع إضافة البيانات الخاصة بالفرع نفسه.

الجدول التالي يوضح عملية بسيطة من هذا النوع عند دمج بيانات من مصدرين:

قاعدة بيانات الفرع (ل):

م	الزبون	العمر	الدخل	المنطقة	حجم المشتريات
١	أ	٤٠	١٠٠٠	س	٢٠٠

٢	ب	٣٠	٨٠٠	س	٢٠٠
....	.....				

قاعدة بيانات الفرع (م):

م	الزبون	العمر	الدخل	المنطقة	حجم المشتريات
١	أ	٥٠	١٥٠٠	ص	٣٠٠
٢	ب	٤٥	١٢٠٠	ص	٢٥٠
....	.....				

دمج قاعدة بيانات الفرع (ل) والفرع (م):

م	الفرع	الزبون	العمر	الدخل	المنطقة	حجم المشتريات
١	ل	أ	٤٠	١٠٠٠	س	٢٠٠
٢	ل	ب	٣٠	٨٠٠	س	٢٠٠
٣	م	ج	٥٠	١٥٠٠	ص	٣٠٠
٤	م	د	٤٥	١٢٠٠	ص	٢٥٠
....						

اعتبارات مختلفة لضمان صحة تنفيذ عمليات الدمج

ومن التحديات التي تواجه عمليات دمج البيانات من عدة مصادر ما يتعلق بكل من طبيعة البيانات والمتغيرات وما تحتويه من قيم ينبغي أن تكون متماثلة ولا تتسبب في ظهور التناقضات، مثلاً في قاعدة بيانات أحد الفروع لأحد الشركات لو تم استخدام التعبير "رقم الزبون" لتخزين رقم بطاقة الزبون

الشخصية، وتم استخدام تعبير "رقم بطاقة الزبون" في فرع آخر فإنه سوف ينشأ تعارض عند دمج قاعدتي البيانات معًا.

كما أنه يمكن أن ينشأ التعارض من خلال اختلاف طبيعة القيم التي يتم تخزينها حتى وإن تطابقت أسماء المتغيرات، مثلاً كأن يتم تعريف القيم (١، صفر) لمتغير باسم "حالة التدخين" في قاعدة بيانات أحد فروع المراكز الطبية وبنفس الوقت يتم تخزين القيم (مدخن، غير مدخن) في فرع آخر لنفس المتغير.

كما قد تظهر التناقضات بين قيم المتغيرات من حيث نظم القياس المستخدمة فيها، كأن يتم استخدام النظام المتري لقياس المسافات بالمتر في قاعدة بيانات أحد الفروع واستخدام النظام الإنجليزي لقياس المسافات باليادرة في قاعدة بيانات فرع آخر، أو بنفس الطريقة عندما يتم استخدام مقاييس مختلفة لدرجات الحرارة كأن يتم استخدام نظام الدرجات المئوية في أحد الفروع ونظام درجات الفهرنهايت في فروع أخرى.

كذلك يتطلب الأمر التأكد من العمليات الحسابية التي تتم على بعض المتغيرات، مثل الخصومات والعروض، وما يمكن أن ينشأ عنها من تعارض بين قواعد البيانات المختلفة، مثلاً عند وجود خصم محدد بنسبة معينة على أسعار المنتجات في أحد الفروع وخصم بنسبة مختلفة في فروع أخرى فإن ذلك يمكن أن يؤثر في نتيجة الدمج وينتج عنه تعارض في الأسعار النهائية للمنتجات التي يتم تجميعها في قاعدة بيانات موحدة. ويحدث ذلك بشكل

أوضح في سلاسل الفنادق التي لها فروع في عدة بلدان، حيث تختلف أسعار الغرف لأسباب عديدة منها اختلاف العملة المستخدمة في كل بلد بالإضافة لاختلاف الخدمات التي يتم تقديمها مجانًا لكل غرفة، كأن يشمل سعر الغرفة تقديم وجبة فطور مجانية أو استخدام حمامات السباحة في أحد الفروع ولا يشملها في فروع أخرى.

وفي جميع الحالات قد ينشأ التناقض والتعارض في البيانات، سواء من حيث البنية والهيكلية التي يتم تصميمها لحفظ البيانات أو من حيث المحتوى والقيم الخاصة بكل متغير من المتغيرات التي يتم حفظها وتخزينها، وفي كل الأحوال فإنه يلزم في كثير من الأحيان القيام ببعض الإجراءات الإضافية قبل بدء تنفيذ عملية الدمج، وذلك من أجل ضمان تنفيذ عملية الدمج بالشكل الصحيح وتجنب الأخطاء التي يمكن أن تظهر وتؤدي إلى إضعاف نتائج التحليل والتنقيب المزمع تنفيذها، وتشتمل هذه الإجراءات على عمليات متعددة يتم الاختيار من بينها بما يتناسب مع الاحتياجات والأهداف في كل حالة، ويمكن تنفيذها باستخدام تقنيات تنظيف البيانات التي تم التطرق إليها سابقًا أو باستخدام تقنيات تحويل البيانات أو اختزالها والتي سوف يتم التطرق إليها لاحقًا في هذا الفصل.

اختزال البيانات

Data Reduction

إن تحليل وتنقيب كميات هائلة من البيانات يمكن أن يستلزم وقتًا وجهدًا كبيرًا، وهو ما يجعل منها عملية غير قابلة للتطبيق العملي أو غير ممكنة، وتساعد تقنيات اختزال أو اختصار البيانات في الحصول على تمثيل أقل لها يمكن أن يكون أصغر كثيرًا في الحجم مع الحفاظ على خصائص البيانات الأصلية.

وبذلك فإن تحليل وتنقيب البيانات المختصرة سيكون أكثر كفاءة ويعطي نتائج شبيهة إلى حد كبير بنتائج تحليل وتنقيب البيانات الأصلية. سوف نستعرض في هذا الجزء الاستراتيجيات المختلفة لاختزال البيانات مع شرح مفصل لتقنياتها.

الإستراتيجيات المتنوعة لاختزال البيانات

يمكن تقسيم هذه الاستراتيجيات إلى ثلاثة أنواع كما يلي:

١. اختصار الأبعاد (اختصار المتغيرات) Attributes Reduction

قد تحتوي قاعدة البيانات محل التحليل والتنقيب على عدد كبير من الحقول (المتغيرات أو السمات)، والتي قد يكون معظمها غير ذات أهمية بالنسبة للمسألة محل البحث أو الدراسة، مثلاً إذا كانت المهمة في أحد شركات بيع الأجهزة الإلكترونية هي تصنيف الزبائن من حيث ما إذا كانوا

سيُقبلون على شراء منتج جديد أو لا من خلال تحليل البيانات التاريخية لمشترياتهم من الشركة، فإن حقل مثل "رقم الهاتف" في قاعدة البيانات، وهو رقم الهاتف الخاص بالزبون، لن يكون ذا أهمية تُذكر في هذه المهمة، بينما يكون حقل "العمر" أو "مستوى الدخل" له أهمية كبيرة، وبالرغم من ذلك فإنه قد تكون هناك بعض الحقول المُبهمة التي يصعب تقدير مدى أهميتها بالنسبة لمسألة البحث، إلاّ أن تراكم الخبرة لدى المحللين يجعلهم قادرين على استبعاد الحقول غير المهمة والإبقاء على الحقول المهمة والمفيدة وذلك بحسب احتياجات التحليل والأهداف المرجوة منه.

مثلاً، في قاعدة بيانات أحد المراكز التجارية، إذا كان كل الزبائن في الفرع هم من سكان نفس المنطقة التي يقع فيها الفرع، فإن حقل الفرع وحقل المنطقة يمثلان بيانات متشابهة تقريباً، وبالتالي يمكن اختصار أحدهما والتركيز على الآخر عند القيام بتحليل وتنقيب البيانات.

مثلاً: من دمج قاعدة بيانات الفرع (ل) والفرع (م) لدينا:

م	الفرع	الزبون	العمر	الدخل	المنطقة	حجم المشتريات
١	ل	أ	٤٠	١٠٠٠	س	٢٠٠
٢	ل	ب	٣٠	٨٠٠	س	٢٠٠
٣	م	ج	٥٠	١٥٠٠	ص	٣٠٠

٤	م	د	٤٥	١٢٠٠	ص	٢٥٠
....						

وحيث أن ل تناظر س، م تناظر ص، فإنه يمكن اختصار البيانات لتصبح كما يلي:

م	الفرع	الزبون	العمر	الدخل	حجم المشتريات
١	ل	أ	٤٠	١٠٠٠	٢٠٠
٢	ل	ب	٣٠	٨٠٠	٢٠٠
٣	م	ج	٥٠	١٥٠٠	٣٠٠
٤	م	د	٤٥	١٢٠٠	٢٥٠
....					

٢. الاختصارات الرقمية بطريقة التوزيع التكراري للفئات

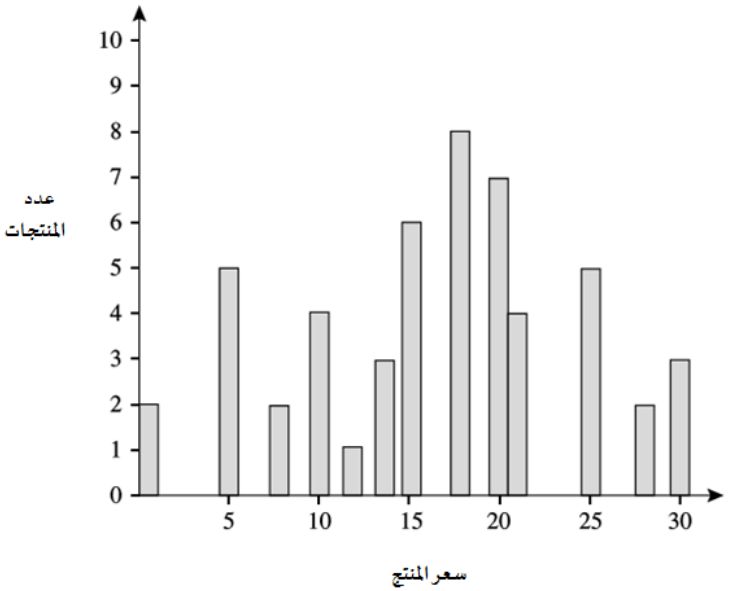
وهي تقنية يتم فيها التعبير عن قيم البيانات الأصلية بطريقة التوزيع التكراري للفئات المستخدمة في الوصف الإحصائي، بحيث يتم تمثيلها بطريقة مختصرة وأصغر حجمًا من طريقة التوزيع التكراري الاعتيادية، كما يتم التعبير عنها باستخدام الرسم البياني.

مثلًا، في قاعدة بيانات مبيعات أحد المراكز التجارية، نفرض أنه لدينا مجموعة من المنتجات المباعة خلال فترة زمنية معينة وكانت أسعارها كما يلي،

مرتبة ترتيبًا تصاعديًا:

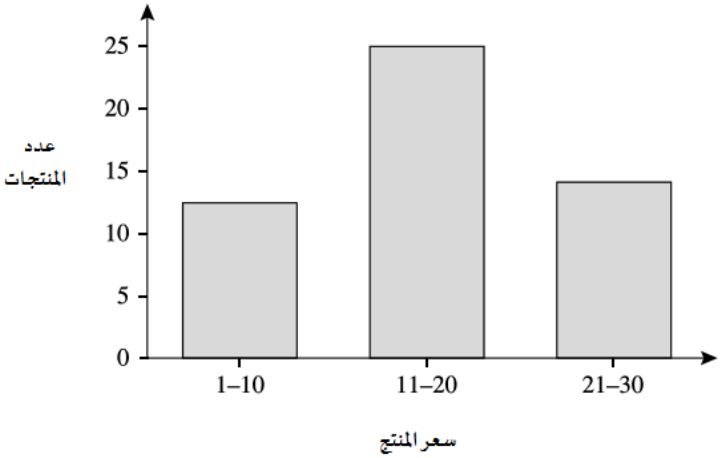
١، ١، ٥، ٥، ٥، ٥، ٥، ٨، ٨، ١٠، ١٠، ١٠، ١٠، ١٠، ١٢، ١٤، ١٤، ١٤، ١٤، ١٥، ١٥، ١٥، ١٥، ١٥، ١٥، ١٨، ١٨، ١٨، ١٨، ١٨، ١٨، ١٨، ٢٠، ٢٠، ٢٠، ٢٠، ٢٠، ٢٠، ٢١، ٢١، ٢١، ٢١، ٢١، ٢٥، ٢٥، ٢٥، ٢٥، ٢٥، ٢٨، ٢٨، ٣٠، ٣٠، ٣٠، ٣٠

فإنه يمكن تمثيل هذه البيانات بطريقة التوزيع التكراري بالشكل البياني التالي:



شكل (٣,٣)، التوزيع التكراري للمنتجات بحسب الأسعار

ثم تتم عملية الاختصار من خلال تجميع التكرارات في ثلاثة فئات لقيمة المتغير "سعر المنتج"، وهي على الترتيب (من١ إلى ١٠)، (من ١١ إلى ٢٠)، (من ٢١ إلى ٣٠)، بحيث تقع ضمنها كل القيم التي ظهرت في قاعدة البيانات، فيتم بذلك اختصار البيانات بدلاً من وجود عدد كبير من الأسعار التي تعبر عنها، ويصبح الشكل البياني الجديد والمبسط كما يلي:



شكل (٣,٤)، التوزيع التكراري للمنتجات بحسب فئات الأسعار

٣. تقنية التجزئة العنقودية للبيانات Clustering

وهي طريقة لتجزئة البيانات في حالة استهداف تحليل وتنقيب مجموعات منها أو من أجل اختزالها وتبسيطها وإظهار دلالات معينة لها من خلال التجزئة، ويتم في هذه التقنية تجزئة مجموعة البيانات الأصلية إلى مجموعات

فرعية تحتوي كل منها على عناصر متشابهة فيما بينها ومختلفة عن عناصر المجموعات الفرعية الأخرى. والتشابه بين العناصر في كل مجموعة يتم تعريفه بأنه مدى التقارب فيما بينها من حيث المسافة التي تفصلها عن بعضها البعض أو عن مركز المجموعة.

وتُستخدم تقنية التجزئة العنقودية كطريقة من طرق اختزال البيانات من أجل إتاحة الفرصة للتركيز على مجموعات فرعية منها ودراستها وتحليلها بشكل منفرد، أو من أجل مقارنتها معًا واستكشاف دلالات معينة قد تكون مهمة لأهداف متنوعة.

مثلاً يمكن تجزئة بيانات أحد الشركات إلى مجموعات من البيانات الخاصة بكل فرع من فروع الشركة، من أجل مقارنتها أو من أجل القيام بعمليات التحليل وتنقيب البيانات لكل فرع على حدة، فيتم تجزئة قاعدة البيانات الأصلية لهذا الغرض.

ويُلاحظ أن تقنية التجزئة تمثل عملية عكسية لعملية الدمج. وسوف يتم التطرق لخوارزميات التجزئة العنقودية لاحقًا.

٤. العينة الإحصائية العشوائية Sampling

وهي طريقة في اختزال البيانات تعتمد نفس الأسلوب المستخدم في البحوث العلمية، ويتم من خلالها اختيار عينة من البيانات لتحليلها وتنقيبها، وذلك بحسب الاحتياج، بحيث يتم اختزال كميات البيانات الضخمة بكمية

صغيرة ومحدودة منها والمتمثلة بالعينة الإحصائية.

ويكثر استخدام هذه الطريقة عند الرغبة في إجراء تحليل أولي للبيانات في المراحل الأولى من مراحل التحليل والتنقيب.

وتوجد أربعة طرق لاختيار العينة الإحصائية وهي كما يلي:

أ. أسلوب العينة العشوائية البسيطة بدون إحلال Simple Random Sample Without Replacement

وفيها يتم اختيار العينة العشوائية دون إحلال للسجلات التي يتم اختيارها في كل مرة، أي يتم استبعاد السجل الذي يتم اختياره من قاعدة البيانات الأصلية، ولكل عدد (ن) من السجلات في قاعدة البيانات يكون احتمال ظهور أي سجل في هذه العينة هو (١/ ن)، وجميع السجلات تتساوى في احتمال الظهور في هذا النوع من العينة.

ب. أسلوب العينة العشوائية البسيطة مع الإحلال Simple Random Sample With Replacement

وفيها يتم اختيار العينة العشوائية مع إرجاع السجلات المختارة إلى قاعدة البيانات الأصلية، بحيث يمكن أن يُعاد اختياره مرة أخرى.

ج. أسلوب العينة العشوائية التجميعية Cluster Sample

يتم في هذه الطريقة تجزئة سجلات قاعدة البيانات إلى مجموعات صغيرة

ومنفصلة عن بعضها البعض، ثم يتم اختيار بعض المجموعات بشكل عشوائي بدون إحلال بنفس طريقة اختيار السجلات، بحيث يتم اختيار عدد محدود من تلك المجموعات يكون أقل من العدد الكلي الذي تم الحصول عليه عن التجزئة.

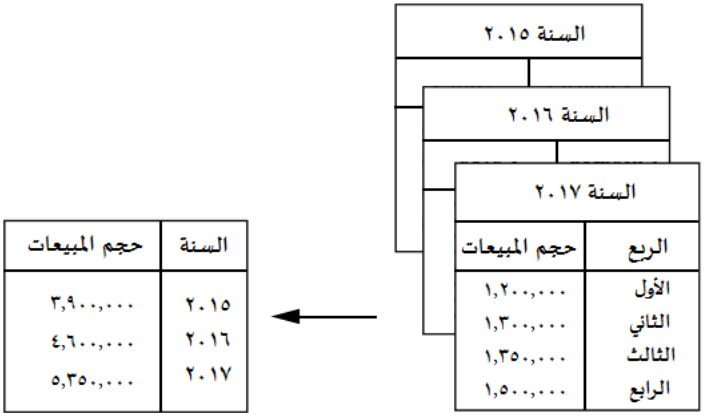
د. أسلوب العينة الطبقية Stratified Sample

وتُستخدم هذه الطريقة من أجل ضمان تمثيل كل البيانات في قاعدة البيانات وفق تجزئتها من منظور معين، مثلاً في قاعدة بيانات مبيعات أحد مطاعم الوجبات السريعة قد يلزم تجزئة الزبائن إلى مجموعات منفصلة بحسب الفئات العمرية ثم البدء باختيار العينة من كل مجموعة بطريقة طبقية بحيث يتم تمثيل الزبائن من جميع الفئات العمرية وبنسبة مماثلة لعدد سجلات كل فئة من الفئات.

٥. التجميع Aggregation

في بعض الأحيان يمكن أن تكون البيانات المتوفرة خاصة بفترات زمنية متفرقة، ويتم في هذه الحالة استخدام أسلوب التجميع من أجل تجميع بيانات كل الفترات ومن ثم تحليلها ودراستها من منظور جديد، مثلاً في قاعدة بيانات مبيعات إحدى الشركات قد تتوفر فيها جداول تبين بيانات المبيعات الربع سنوية لعدة سنوات متتالية، فتتم بهذه الطريقة تجميع الأربعة أرباع في كل سنة من السنوات من أجل الحصول على جدول جديد يوضح إجمالي

المبيعات السنوية، الشكل التالي يوضح هذا الأسلوب:

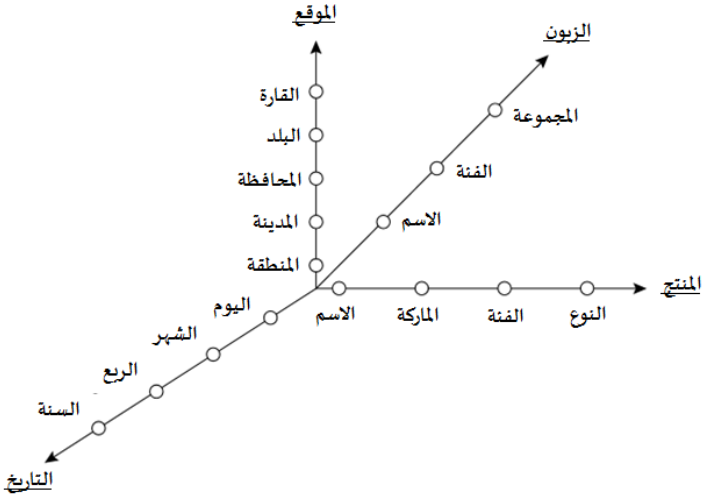


شكل (٣,٥)، تجميع بيانات المبيعات السنوية

٦. مكعب البيانات Data Cube

وهي طريقة للتعبير عن البيانات باستخدام الأشكال متعددة الأبعاد، ويتم فيها تجميع البيانات بحسب السمات المختلفة لها وبما يحقق احتياجات التحليل والتنقيب، مثلاً لو كان لدينا عدة قواعد بيانات لعدد من الفروع لأحد المراكز التجارية فإنه يمكن استخدام مكعب البيانات للتعبير عن مجمل البيانات التي تحتويها كل قواعد البيانات من أجل إظهار السمات المختلفة لها وتقاطعاتها مع بعضها البعض في رسم بياني واحد متعدد الأبعاد، وهو ما يطلق عليه مكعب البيانات Data Cube.

الشكل التالي يبين نموذج لمكعب بيانات، المحور الأول فيه يمثل سمات الزبائن أنفسهم، والمحور الثاني يمثل سمات الموقع، والمحور الثالث يمثل سمات المنتجات التي يتم شرائها، والمحور الرابع يمثل سمات عمليات الشراء من المركز مجمعة حسب الوقت والتاريخ.



شكل (٣,٦)، مكعب البيانات يُظهر مجموعة من البيانات متعددة الأبعاد

وتكمن الفائدة من استخدام مكعبات البيانات في أنها تلخصها وتظهرها كلها في شكل واحد تتقاطع فيه كل السمات لكل السجلات في قاعدة البيانات، كما يمكن تركيز عمليات التحليل والتنقيب على محورين من المحاور في كل مرة بحسب الهدف منها وبحسب طبيعة البيانات المتوفرة وبحسب ما يظهره الرسم البياني في المكعب من دلالات تخص جميع المتغيرات المتقاطعة معًا.

الخلاصة

توجد العديد من الأساليب والخوارزميات التي يتم استخدامها من أجل اختزال البيانات، وفي الواقع ينبغي ألا يزيد الوقت أو الجهد المستخدم في اختصار البيانات عن الوقت والجهد الذي نستهدف توفيره في مرحلة التحليل والتنقيب نفسها، وإلاّ فإنه من الأولى بدء تحليلها وتنقيبها بدون اختصار لأن اختصارها في هذه الحالة سوف يكون مضيعة للوقت والجهد.

من جهة أخرى، يتطلب الأمر الانتباه جيدًا لنوعية البيانات التي يتم اختزالها أو تجاهلها نتيجة القيام بأحد أساليب الاختزال، وذلك لأن استبعاد أو تجاهل البيانات المهمة سوف يؤدي إلى ضعف نتائج التحليل والتنقيب التي يتم تنفيذها فيما بعد، كما أن الإبقاء على البيانات التي لا تُعتبر مهمة بالنسبة للهدف من التحليل والتنقيب سوف يؤثر سلبًا على النتائج.

تحويل البيانات وتفريد البيانات
Data Transformation & Data Discretization

تُعتبر إجراءات كل من تحويل البيانات وتفريد البيانات من المراحل المهمة في تحضير البيانات للتحليل والتنقيب، حيث أنها تُساعد في الحصول على نتائج أفضل في التحليل والتنقيب وترفع من كفاءته، كما أنها تُساعد في تبسيط فهم الأنماط والارتباطات التي يتم استكشافها.

ويمكن حصر الاستراتيجيات المتنوعة لتحويل البيانات فيما يلي:

١. تنعيم البيانات Smoothing

وهي تقنية تُستخدم لإزالة الإزعاج من البيانات وتشتمل على كل من التكييس (تجميع البيانات في سلات) وتقنية الانحدار والتجزئة العنقودية، والتي تم استعراضها جميعًا في قسم تنظيف البيانات سابقًا.

٢. إنشاء السمات الجديدة Attribute Construction

وهي عملية إضافة سمة جديدة للبيانات لم تكن موجود سابقًا، وذلك للمساعدة في عمليات التحليل والتنقيب، مثلاً في قاعدة بيانات زبائن أحد المراكز التجارية يمكن إضافة سمة جديدة لكل الزبائن تكون وظيفتها التعبير عن قيمة مناظرة لحجم المشتريات لكل منهم، بحيث تأخذ قيمة واحدة من قيم الفئات التالية (منخفض – متوسط – مرتفع)، وذلك من واقع الأرقام

المحددة في قاعدة البيانات ومحددات يتم وضعها مسبقًا من قِبل القائمين على عمليات التحليل والتنقيب، الجدول التالي يوضح هذه الطريقة:

م	الفرع	الزبون	العمر	الدخل	حجم المشتريات	الفئة
١	ل	أ	٤٠	١٠٠٠	٢٠٠	منخفض
٢	ل	ب	٣٠	٨٠٠	٢٠٠	منخفض
٣	م	ج	٥٠	١٥٠٠	٣٠٠	مرتفع
٤	م	د	٤٥	١٢٠٠	٢٥٠	متوسط
....						

ويلاحظ من الجدول أعلاه كيف تم استحداث متغير جديد باسم "الفئة" ومنحه قيم مختلفة بحسب حجم مشتريات الزبون، بحيث تم اعتبار الفئات كما يلي:

منخفض: حجم المشتريات يقع في الفئة (صفر - ٢٠٠)

متوسط: حجم المشتريات يقع ضمن الفئة (٢٠١ - ٢٥٠)

مرتفع: حجم المشتريات يقع ضمن الفئة (٢٥١ - ٣٠٠)

٣. التجميع Aggregation

وهي طريقة تجميع لملخصات البيانات، مثلاً بيانات المبيعات اليومية في أحد المراكز التجارية يمكن تجميعها وحساب المبيعات الشهرية والسنوية

الإجمالية. وهذه الطريقة تشبه طريقة إنشاء مكعب البيانات Data Cube بهدف تحليلها في عدة مستويات.

٤. التطبيع Normalization

وهي طريقة لقياس البيانات والسمات باستخدام تدريج معين، بحيث يتم حصرها في مدى محدد، مثلاً من صفر إلى ١، أو من -١ إلى ١، ويتم حساب القيم المناظرة لهذه القيم باعتبارها حدود عظمى ودنيا من كل القيم المتوفرة في قاعدة البيانات.

مثلًا في قاعدة بيانات أحد المراكز التجارية إذا كان لدينا بيانات الزبائن كما يلي:

م	الفرع	الزبون	العمر	الدخل	حجم المشتريات
١	ل	أ	٤٠	١٠٠٠	٢٠٠
٢	ل	ب	٣٠	٨٠٠	٢٠٠
٣	م	ج	٥٠	١٥٠٠	٣٠٠
٤	م	د	٤٥	١٢٠٠	٢٥٠
....					

فإنه يمكن إضافة سمة جديدة بعنوان تدريج الدخل ليتم التعبير عنه بشكل نسبي لكل الزبائن على تدريج يقع حديه بين الصفر والواحد الصحيح، وذلك

كما يلي:

أقل قيمة للدخل في الجدول هي ٨٠٠ = الحد الأدنى للتدريج

أعلى قيمة للدخل في الجدول هي ١٥٠٠ = الحد الأعلى للتدريج

أي أن الحد الأدنى والحد الأعلى لقيمة الدخل = (٨٠٠ - ١٥٠٠)

فتكون هاتين القيمتين مناظرتين لقيم الحد الأدنى والأعلى على التدريج الذي يتم اختياره، وليكن هنا (صفر - ١).

ويتم حساب أي قيمة من قيم الدخل على هذا التدريج باستخدام المعادلة التالية:

قيمة الدخل على التدريج = (قيمة الدخل - الحد الأدنى) / (الحد الأعلى - الحد الأدنى)

وبحساب جميع القيم في الجدول يصبح لدينا سمة جديدة تكون قيمتها لكل سجل كما في الجدول التالي:

م	الفرع	الزبون	العمر	الدخل	حجم المشتريات	تدريج الدخل
١	ل	أ	٤٠	١٠٠٠	٢٠٠	٠,٢٩
٢	ل	ب	٣٠	٨٠٠	٢٠٠	٠,٠٠
٣	م	ج	٥٠	١٥٠٠	٣٠٠	١,٠٠

٤	م	د	٤٥	١٢٠٠	٢٥٠	٠,٥٧
....						

٥. تفريد البيانات Data Discretization

وهي طريقة يتم فيها استبدال السمات أو المتغيرات الرقمية بسمات أسمية أو تسميات فئوية، مثلما يحدث في سمة العمر، كأن يتم استبدال متغير العمر بفئات عمرية مختلفة مثلاً (من١٠ إلى ٢٠، من ٢١ إلى ٣٠، من ٣٢ إلى٤٠، ... إلخ)، أو أن يتم استبدالها بتسمية أسمية مثل (أطفال - بالغين - كبار السن).

كما يمكن إعادة التفريد مرة أخرى في مستويات أعلى بشكل هرمي.

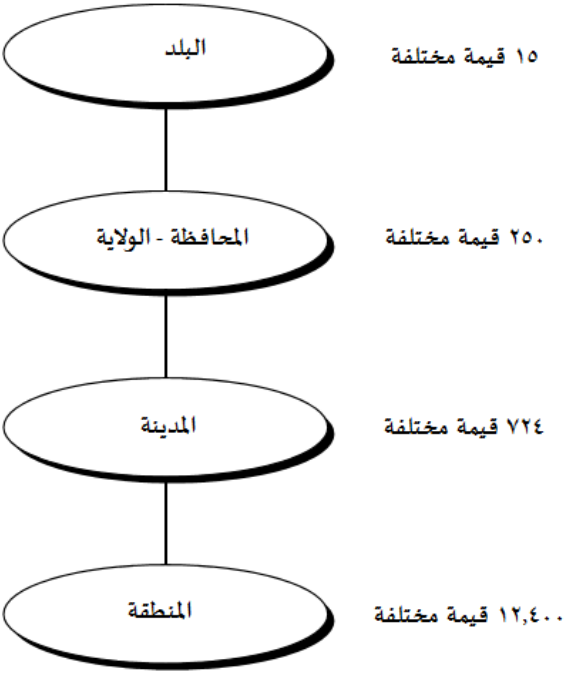
ويمكن تعريف أكثر من تسمية جديدة للمتغير نفسه، وذلك بحسب حاجات التحليل والتنقيب لكل متغير.

٦. توليد التسلسل الهرمي للبيانات الأسمية Hierarchy Generation of Nominal Data

وتُستخدم هذه الطريقة لتعميم البيانات الاسمية لمستويات أعلى، مثلاً في البيانات المتعلقة بالعنوان الخاص بالزبون الذي يظهر فيها اسم المنطقة يمكن توليد متغير جديد يتم تخصيصه لاسم المدينة، ويتم تحويل كل أسماء المناطق إلى اسم المدينة التي ينتمون لها.

وقد تحتوي قاعدة البيانات على عدة مستويات من المتغيرات الاسمية ضمنيًا، مثلاً عندما يكون فيها متغير يعبر عن المدينة ومتغير يعبر عن البلد، فطبيعي أن يكون متغير البلد في مستوى أعلى من مستوى المدينة، أو بمعنى آخر يشتمل على كل المدن التي تقع فيها بتسلسل هرمي.

الشكل التالي يوضح بيانات زبائن أحد سلاسل المطاعم العالمية تم تجميعها حسب البلد والمحافظة أو الولاية والمدينة والمنطقة بشكل هرمي:



شكل (٣,٧)، توليد التسلسل الهرمي لبيانات محل السكن لزبائن أحد المراكز التجارية

ويتضح أنه كلما يتم استحداث متغير جديد في المستوى الأعلى (مدينة، محافظة، بلد..) فإنه يحتوي على عدد من الفئات الأدنى مجتمعة بشكل تراكمي هرمي ويقل عدد السجلات المختلفة كلما تم الصعود في المستويات الأعلى، نظراً لاحتواء المجموعات في المستويات الأعلى على عدد من المجموعات الجزئية التي تتشابه فيها العناصر المشتركة.

ويتم التعبير عن هذا الأمر رقمياً بأن يُقال أن لدى شركة ما زبائن من خمسة عشر دولة مختلفة، أو أن نقول أن لديها زبائن من ٢٥٠ محافظة مختلفة تقع في ١٥ دولة، أو أن لديها زبائن من ٧٢٤ مدينة أو قرية مختلفة مثلاً، وأخيراً فإنه يمكن القول بأن لديها زبائن من ١٢٤٠٠ حي أو منطقة سكنية مختلفة.

ومع ذلك، فإن هذه العملية قد تبدو معكوسة إذا تم استخدامها في السلاسل الزمنية، مثلاً في قاعدة بيانات مبيعات إحدى الشركات يمكن أن يتم توليد سلاسل هرمية بحسب فترات زمنية مختلفة مثل السنوات والأشهر والأيام، يتوفر فيها سجلات لعدد كبير من السنوات، ولكن إذا تم تجميع السجلات حسب أسماء الأشهر فسوف ينتج عدد (١٢) قيمة مختلفة فقط وهي القيم المناظرة لأسماء الأشهر في السنة، أما لو تم تجميع المبيعات بحسب أسماء أيام الأسبوع فسوف نحصل على (٧) أسماء مختلفة فقط وهي المناظرة لأيام الأسبوع، وهو ما يجعل التسلسل الهرمي يظهر وكأنه ينتقل إلى مستويات أعلى بشكل تصاعدي، وهذا الأمر لا ينطبق في هذه الحالة، بالرغم

من أن أيام الأسبوع أقل من عدد الأشهر والسنوات. في الواقع فإن هذه الطريقة تفيد فقط عند دراسة وتحليل ومقارنة حجم المبيعات خلال أيام الأسبوع بهدف استكشاف أيام الذروة أو حتى ساعاتها في حال إضافة التوقيت بالساعة للبيانات محل الدراسة، أما توليد التسلسل الهرمي بهذا الشكل فإنه لا يشترط تشابه الأيام من حيث الاسم فقط، ولكنه يستلزم تطابقها أيضا، وهو ما يعني أن السنة الواحدة سوف تظل تحتوي على (٣٦٥) يوم مختلف، والشهر الواحد يحتوي على (٣٠) يوم مختلف.

الفصل الرابع
تنقيب واستكشاف الأنماط وقواعد التبعية والارتباط

Mining Patterns, Associations and Correlations

١. مفاهيم أساسية
٢. تحليل سلة المشتريات (مثال تشويقي) Market Basket Analysis
٣. تقييم قواعد التبعية والارتباط المستكشفة Rule Evaluation
٤. تنقيب قواعد التبعية والارتباط متعددة المستويات Mining Multi-Level Associations and Correlations
٥. تنقيب قواعد التبعية والارتباط متعددة الأبعاد Mining Multi-Dimensional Associations and Correlations
٦. قواعد التبعية والارتباط الاسمية والكمية Nominal & Quantitative Association Rules
٧. تنقيب واستكشاف الأنماط النادرة والأنماط السلبية Mining Rare Patterns and Negative Patterns
٨. تنقيب واستكشاف قواعد التبعية والارتباط المشروطة Mining Conditional Associations & Correlations
٩. تقييم قواعد التبعية والارتباط وتمييز القواعد المفيدة وغير المفيدة Rules Evaluation and Recognizing
١٠. قياس نوع وقوة الارتباط في قواعد التبعية والارتباط المستكشفة Correlation Type & Power Measurement

مفاهيم أساسية

لو تخيلنا أحد مدراء المبيعات في معرض لمحلات بيع الأجهزة الكهربائية يتحدث لأحد الزبائن ممن اشتروا مؤخرًا جهاز حاسوب وكاميرا رقمية على الترتيب من نفس المعرض، فما هو المنتج الذي يمكن أن يرشحه هذا المدير لذلك الزبون حتى يشتريه أثناء زيارته للمعرض؟

إن المعلومات التي تبين أي من المنتجات يتكرر شراؤها من زبائن المعرض بعد شرائهم جهاز حاسوب وكاميرا رقمية على الترتيب سوف تساعد مدير المبيعات في انتقاء المنتج التالي لعرضه على الزبون.

وفي هذا السيناريو تحديدًا فإن تنقيب واستكشاف الأنماط وقواعد التبعية والارتباط هي المعرفة التي تلزم لمتخذي القرار في مثل هذه الحالات.

والأنماط المتكررة يمكن أن تكون مجموعة من العناصر أو المنتجات Item-set أو الفئات المتسلسلة Subsequence أو البناءات الجزئية Substructure، تظهر بشكل متكرر في قواعد البيانات.

مثلاً في قاعد بيانات أحد محلات بيع التجزئة يمكن أن يتكرر ظهور منتجات الحليب والبيض كمجموعة من المنتجات التي يتم شراؤها معًا في نفس رحلة التسوق، ويصبح النمط في هذه الحالة هو تكرار مجموعة من العناصر Item-set.

أما تكرار التسلسلات فيكون بشراء الزبون لجهاز حاسوب أولاً في أول رحلة تسوق ثم قيامه بشراء كاميرا رقمية في الرحلة الثانية ثم شراء كرت ذاكرة في الرحلة الثالثة، بحيث تتم هذه السلسلة بشكل متكرر في قاعدة البيانات، ويطلق عليها نمط التسلسلات المتكررة Subsequence.

أما تسلسلات البناء الجزئي فيمكن أن تكون خليط من العناصر والفئات والتسلسلات المختلفة، وبهذه الحالة يطلق عليها نمط البناءات الجزئية المتكررة Substructure.

إن استكشاف الأنماط المتكررة بكل أنوعها يلعب دورًا مهمًا في تنقيب واستكشاف علاقات التبعية والارتباطات والعلاقات الأخرى الشيقة بداخل قاعدة البيانات، كما أنها تساعد في تصنيف وتجزئة البيانات وإجراءات التنقيب الأخرى للكميات الهائلة من البيانات التي يتم تجميعها وتخزينها بصفة مستمرة، الأمر الذي يساهم في تحسين الإدارة الإستراتيجية في كثير من المجالات كالصناعة والتجارة والزراعة والأعمال والخدمات بكافة أنواعها وأشكالها.

وباستخدام هذه التقنيات تستطيع المؤسسات والشركات بكافة أنواعها استكشاف العلاقات والارتباطات الخفية الشيقة بداخل قواعد البيانات وتبسيطها وتوضيحها لصنّاع القرار من أجل مساعدتهم في تحسين الإدارة واتخاذ القرارات ورفع كفاءة التخطيط وتحسين خطط التسويق المتعدد الأغراض وتحليل وإدارة المخاطر بطرق تؤدي لأفضل النتائج.

وقد أصبحت عمليات تنقيب الأنماط المتكررة من العمليات المهمة في تنقيب البيانات ويتم التركيز عليها كأحد مسارات بحوث تنقيب البيانات بشكل عام.

في هذا الفصل سوف نستعرض المفاهيم الأساسية وطرق تنقيب الأنماط المتكررة من أجل استكشاف علاقات الارتباط والتبعية الشيقة بين العناصر والفئات المختلفة بداخل قواعد البيانات العلائقية.

تحليل سلة المشتريات - مثال تشويقي
Market Basket Analysis

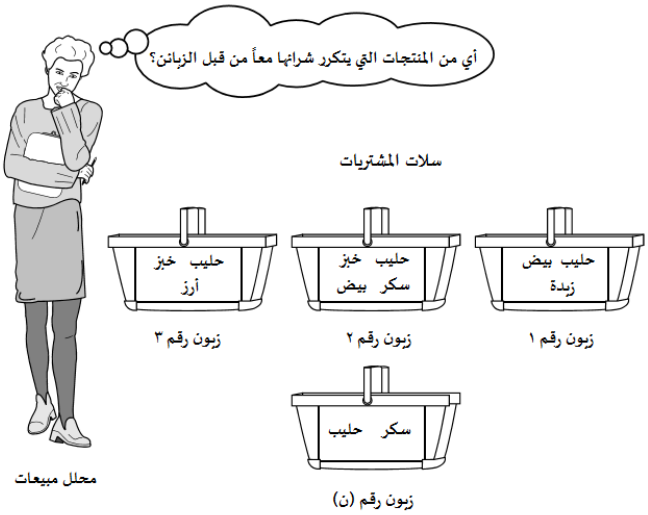
إن تنقيب الأنماط المتكررة يؤدي لاستكشاف علاقات التبعية والارتباط في سجلات قواعد البيانات العلائقية، ومع تراكم الكميات الهائلة من البيانات التي يتم جمعها وتخزينها بشكل مستمر تزداد الحاجة لاستكشاف مثل تلك الأنماط بداخل قواعد البيانات في كل مجالات الأعمال والصناعة والتجارة والأعمال والخدمات بكافة أنواعها.

واستكشاف علاقات التبعية والارتباط الشيقة في تلك السجلات تساهم في تعزيز قدرات متخذي القرار في المؤسسات والشركات في جميع المجالات، ففي مجال الأعمال تساعد في اختيار التصاميم المناسبة للعروض والكتالوجات ووضع خطط التسويق وتحليل سلوك التسوق للزبائن.

ومن أمثلة تنقيب الأنماط المتكررة "تحليل سلة المشتريات"، والذي يتم فيه تحليل السلوك والعادات الشرائية للزبائن من خلال إيجاد علاقات وقواعد التبعية والارتباط بين عناصر المنتجات المختلفة التي يضعها الزبائن في سلال مشترياتهم.

كذلك فإن استكشاف علاقات التبعية والارتباط يمكن أن يساعد تجار التجزئة في تطوير إستراتيجيات التسويق من خلال معرفتهم للمنتجات التي يتكرر شرائها معًا من الزبائن.

مثلاً، إذا قام الزبائن بشراء الحليب، فما هو احتمال أنهم يقومون بشراء الخبز (وأي نوع من الخبز) في نفس رحلة التسوق إلى المركز التجاري؟



شكل (٤,١)، سلات مشتريات مختلفة لعدد من الزبائن في أحد مراكز التسوق

وهذه المعلومات يمكن أن تؤدي إلى زيادة المبيعات لأنها تساعد تجار التجزئة في تخطيط وتنفيذ أساليب التسويق الانتقائية وهندسة وتوزيع المنتجات في الأماكن والرفوف المناسبة في مراكزهم التجارية.

ومن الأمثلة الشيقة التي تشرح تنقيب قواعد التبعية والارتباط هو مثال "تحليل سلة المشتريات" Market Basket Analysis، والذي يهدف لتحليل واستكشاف سلوك واتجاهات وعادات التسوّق للزبائن من خلال إيجاد الارتباطات الخفية بين كميات وأنواع المشتريات للمنتجات المختلفة لهم.

ففي مراكز التسوق الكبرى أو الهايبر ماركت Hyper Market، يمكن القيام بعمليات تحليل واستكشاف للارتباطات الخفية والعادات والاتجاهات والسلوك الشرائي للزبائن وتحديد المنتجات التي يتم شرائها معًا بشكل متكرر، وذلك من أجل المساعدة في تطوير إستراتيجيات التسويق في تلك المراكز.

فمثلًا، إذا قام الزبائن بشراء الحليب، فإلى أي درجة يُقبلون على شراء الخبز في نفس رحلة التسوق، وأي نوع من الخبز يشترون. ومثل تلك المعلومات تفيد الإدارة في إعادة ترتيب المنتجات واختيار وضعها على الرفوف بطريقة أفضل وتسهل الوصول إليها بحسب الارتباطات التي يتم استكشافها، كما تساعد في توجيه خطط التسويق لتكون بطرق فيها دعم وتعزيز للمنتجات المرتبطة ببعضها البعض.

إن تقنية تنقيب واستكشاف قواعد التبعية والارتباط من التقنيات الأساسية في التنقيب في البيانات وأكثرها شيوعا في مجال استكشاف المعرفة، وهي أقرب ما تكون إلى ما يسمى بعملية التنقيب بحد ذاتها، والذهب في هذه الحالة هو "قاعدة الارتباط أو التبعية"، وهذه القاعدة تبين ما يجري داخل قاعدة البيانات وتظهر لنا ما لم نكن نعرفه من قبل، وربما أيضًا ما لن نستطيع أن نعرفه.

وتُعتبر القاعدة التي يتم استكشافها من سجلات قاعدة البيانات ذات أهمية كبرى، هذا بالإضافة إلى التقديرات التي تصاحبها، والتي تساعد في تحديد احتمال وقوع حدث معين أو عدة أحداث متلازمة والمتمثلة بعملية شراء منتج أو مجموعة من المنتجات معًا.

وبشكل عام تعتبر هذه القواعد سهلة وبسيطة نسبيًا، فعلى سبيل المثال في تحليل قاعدة بيانات سلة مشتريات أحد المراكز التجارية يمكن أن نستكشف قاعدة الارتباط الشيقة التالية:

"إذا اشترى الزبون سلعة (أ) فإنه يشتري السلعة (ب) معها باحتمال ٨٠٪، وهذه الثنائية تحدث بإجمالي ٣٪ من كافة المشتريات".

والشكل العام للقاعدة هو : " إذا كان كذا وكذا فإن كذا "، أي أنه لها طرفين: الأيمن المستقل والأيسر التابع.

مثلًا: "إذا اشترى الزبون شرائح بطاطا فإنه يشتري كاتشب"

تقييم قواعد الارتباط والتبعية المستكشفة
Rules Evaluation

لكي تكون القاعدة مكتملة وذات فائدة، فإنه يلزم لها تقييم، وهو عبارة عن نوعين إضافيين من المعلومات التي يجب أن تلازمها، وهي:

الصحة Accuracy: هي نسبة صحة القاعدة (نسبة السجلات التي يتحقق فيها وقوع النتيجة في حال وقوع السبب).

التغطية Coverage: هي نسبة السجلات المحققة للقاعدة إلى كافة السجلات في قاعدة البيانات.

أي أنه لا يعني تسمية القاعدة بهذا الاسم أنها صحيحة دومًا، فإنه من الضروري توضيح مدى صحة هذه القاعدة والشك فيها، وهذا يوضحه مصطلح (الصحة)، ومن جهة أخرى، فإنه أيضًا يتوجب توضيح مدى تغطية هذه القاعدة للبيانات في قاعدة البيانات، وهذا يبينه مصطلح (التغطية). والأمثلة التالية تبين بعض القواعد مع نسبة كل من الصحة والتغطية لها:

القاعدة	التغطية	الصحة
إذا اشترى الزبون بيض فإنه يشتري الحليب	20%	85%
إذا اشترى الزبون خبز فإنه يشتري الجبنة	6%	15%
إذا كان عمر الزبون ٣٣ عام واشترى البيض والجبنة فإنه يشتري عصير البرتقال	0.01%	95%

كيف يتم استثمار القاعدة

إن استخدامات القاعدة كثيرة، فمن الممكن تحديد الكثير من القرارات المبنية على قواعد يتم استكشافها في قواعد البيانات، مثلاً، يمكن لمركز تجاري أن يستكشف كافة القواعد الخاصة بمنتج معين، بان يكون طرفًا مستقلاً فيها (الأيمن)، ويتفحص مدى تأثير بيعه لهذا المنتج على بيعه للمنتجات الأخرى، وهل من الضروري اتخاذ خطوات معينة في خطته للتسويق لهذا المنتج مع تلك المنتجات، أو إذا أراد إلغاء بيعه لهذا المنتج فهل سيكون هناك تأثير على عمليات بيع المنتجات الأخرى.

والعكس صحيح، فباستكشاف القواعد الخاصة بمنتج معين بوضعه طرفًا تابعًا فيها (الأيسر)، يمكن استثمار تلك القواعد بأن يتم مثلاً استكشاف مدى تلازم بيع هذا المنتج مع منتجات أخرى معينة وبالتالي إعادة توزيع ترتيب عرض تلك المنتجات بطريقة أكثر فائدة من أجل زيادة المبيعات، أو تقديم عروض وخصومات جديدة تعتمد على مثل تلك العلاقات.

وهكذا، فإنه يوجد العديد من الأفكار التي يمكن أن تفيد فيها عمليات استقراء القاعدة، سواء كان الطرف المستكشف فيها طرفًا مستقلاً أو تابعًا ومن ثم استكشاف كافة العلاقات التي يكون طرفًا فيها.

من جهة أخرى، تفيد معلومات الصحة والتغطية المتلازمة مع القاعدة أيضًا في استكشاف مدى قوة القاعدة وماذا يجري داخل قاعدة البيانات

بشكل متكرر أكثر، فالصحة والتغطية تعزز من اتخاذ القرارات الأنسب عندما نريد استثمار القاعدة بشكل أفضل.

وبالمقابل، لا يعني ذلك أن كل قاعدة محققة بشكل دائم، فالطرف الأيمن ليس بالضرورة أن يؤدي للطرف الأيسر بشكل حتمي، والشكل التالي يوضح علاقة نسبتي الصحة والتغطية للقاعدة ومدى تأثيرهما عليها:

	الصحة عالية	الصحة منخفضة
التغطية عالية	القاعدة غالبًا صحيحة ويمكن استخدامها كثيرًا	القاعدة نادرًا ما تكون صحيحة ولكن يمكن استخدامها كثيرًا
التغطية منخفضة	القاعدة غالبًا صحيحة ولكنها نادرًا ما تُستخدم	القاعدة نادرًا ما تكون صحيحة وتُستخدم نادرًا

وتكمن الفائدة في عملية تحديد هاتين القيمتين بأنها تساعد في تقدير طريقة الاستفادة من القاعدة ومدى الحاجة لاستثمارها بشكل جيد. فعلى سبيل المثال إذا كان لدينا قاعدة ذات صحة عالية وتغطية منخفضة فإنها تشبه أن يكون لدينا فرس سباق يفوز دومًا في أي سباق يشترك به ولكنه لا يشترك في السباق إلا مرة واحدة فقط كل سنة، وحتى في مثل هذه الحالة سوف يكون بالإمكان تحقيق الجوائز الكبرى، ولكن في حالة اكتشاف قاعدة مثيلة في تحليل سلة مشتريات مركز تجاري فإنه لن تكون ذات فائدة فعلية يمكن تحقيق أية فوائد كبرى من خلالها.

طريقة تقييم القاعدة

في مثال تحليل سلة مشتريات أحد المراكز التجارية، ليكن لدينا القاعدة التالية:

"إذا اشترى الزبون حليب فإنه يشتري البيض"

والأعداد التالية كما يلي:

ن = ١٠٠: العدد الكلي للسجلات في قاعدة البيانات

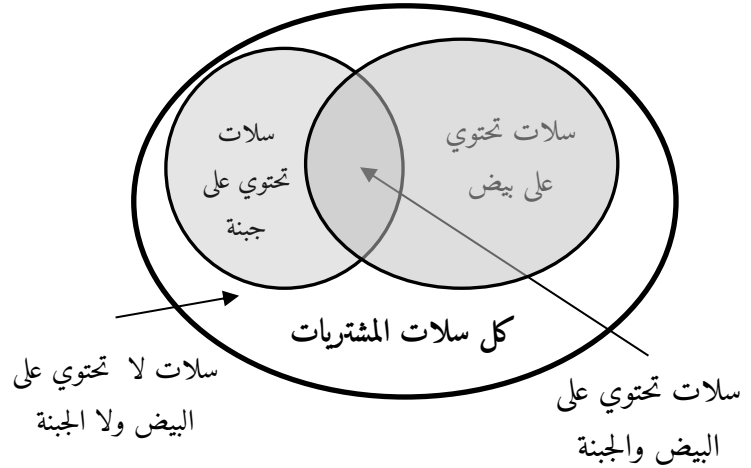
ب = ٣٠: عدد سجلات المشتريات التي تحتوي على البيض

ح = ٤٠: عدد سجلات المشتريات التي تحتوي على الحليب

م = ٢٠: عدد سجلات المشتريات التي تحتوي على البيض والحليب معًا

فتكون نسبة صحة القاعدة هي حاصل قسمة عدد سجلات المشتريات التي تحتوي على البيض والحليب معًا مقسومًا على عدد سجلات المشتريات التي تحتوي على الحليب. وتكون في هذه الحالة مساوية لـ ٢٠ / ٤٠ = ٥٠ %.

أما التغطية فتكون حاصل قسمة عدد سجلات المشتريات المحتوية على الحليب مقسومًا على العدد الكلي للسجلات في قاعدة البيانات. وهي هنا تساوي ٤٠ /١٠٠ = ٤٠ %. والشكل التالي يبين هذه القاعدة:



شكل (٤,٢)، شكل فن يوضح توزيع محتويات سلات المشتريات لأحد المراكز التجارية

مقومات القاعدة المفيدة

إن أكبر مشكلة يمكن مواجهتها في خوارزميات استقراء قواعد التبعية والارتباط هي كيف يمكن تمييز القاعدة المفيدة من القاعدة غير المفيدة، فمنها ما لا يكون لها فائدة عملية، ومنها ما تكون صحتها نادرة، ومنها ما تكون تغطيتها نادرة ولا يمكن تطبيقها، ومنها ما توضح معلومات بديهية لا حاجة لها.

من جهة أخرى، تعتبر القاعدة البسيطة والواضحة مفيدة أيضًا بعكس ما إذا كانت معقدة وصعبة الفهم، وكذلك يمكن اعتبار القاعدة مفيدة في حالة ما إذا كانت توضح علاقة فريدة. وسوف يتم التطرق لبعض أساليب تقييم قواعد التبعية والارتباط لاحقًا في هذا الفصل.

تنقيب واستكشاف قواعد التبعية والارتباط متعددة المستويات
Mining Multi-Level Associations and Correlations

في كثير من الأحيان قد يلزم لاستكشاف الأنماط المتكررة في قواعد البيانات توسيع مجال البحث في البيانات لمستويات أعلى من المستوى الذي تستهدفه الدراسة، مثلاً إذا افترضنا أن أحد الفروع في محلات بيع الأجهزة الإلكترونية يسعى لاستكشاف علاقات التبعية والارتباط في قواعد البيانات الخاصة بحركة مبيعات المنتجات لديه وفق تعريف كل منتج قائم بذاته فإنه يمكن اعتبار أن هذه العملية أحادية المستوى وتولد قواعد تبعية وارتباط أحادية المستوى أيضًا، أما إذا استهدفت الدراسة إيجاد علاقات التبعية والارتباط بين فئات مختلفة للمنتجات تتدرج في مستويات متعددة فإنه سيتولد عنها استكشاف قواعد تبعية وارتباط متعددة المستويات.

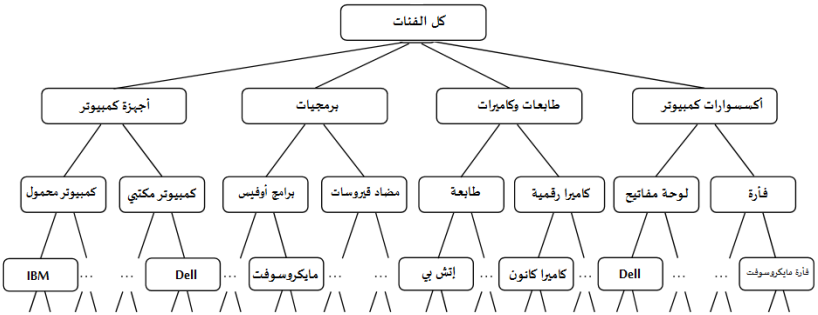
وفي كل الأحوال، سواء كانت الحاجة إلى توسيع الدراسة رأسيًا لمستويات أعلى للفئات الأكبر التي تحتوي العناصر الأصغر أو بالحفر نزولاً إلى مستويات أدنى للفروع والأقسام والأنواع، فإنه ينتج عنها تعدد للمستويات، وهو ما سوف يتم توضيحه في المثال التالي.

مثال توضيحي

لو افترضنا أنه في أحد محلات بيع الأجهزة الإلكترونية لدينا أحد جداول قاعدة بيانات حركة المبيعات في المعرض، والذي يظهر جزء منه كما يلي:

رقم الحركة	الأصناف المباعة
ح ١٠٠١	جهاز ماك بوك نوع آبل، كاميرا ذكية نوع برو
ح ١٠٠٢	برنامج مايكروسوفت أوفيس ٢٠١٣، فأرة لاسلكية ضوئية نوع مايكروسوفت
ح ١٠٠٣	فأرة ليزر نانو، لوحة مفاتيح كمبيوتر محمول
ح ١٠٠٤	نوتبوك ستوديو نوع دل، كاميرا رقمية نوع كانون
ح ١٠٠٥	تابلت نوع لينوفو، برنامج مضاد فايروسات نوع نورتون

كما لدينا خريطة التمثيل الهرمي التسلسلي لكل من العناصر الممثلة للمنتجات المباعة والفئات التي تنتمي لها كما يظهر في الشكل التالي:



شكل (٤,٣)، تمثيل هرمي متعدد المستويات بحسب الفئات للمنتجات المباعة في أحد محلات الأجهزة الإلكترونية

ويتضح من الشكل أنه يمكن تعميم بيانات المبيعات باستبدال المستويات الأدنى بما يناظرها من المستويات الأعلى، والتي تمثل الآباء والأجداد لها، وبشكل هرمي تسلسلي تصاعدي.

ففي الشكل السابق، لدينا خمسة مستويات يمثل المستوى الأول منها كل الفئات والمنتجات بدون تمييز بينها، والمستوى الثاني يمثل تصنيف عام للمنتجات المباعة تم التعبير عنها بمسميات مثل (أجهزة كمبيوتر، برامج، طابعات وكاميرات، مستلزمات وإكسسوارات كمبيوتر).

وفي المستوى الثالث تم إعادة تصنيف للفئات العامة مرة أخرى إلى فئات جزئية تمثل (كمبيوتر محمول، كمبيوتر مكتبي، برامج أوفيس، برامج مضادة للفيروسات، ... إلخ).

وفي المستوى الرابع تم تصنيف الفئات الجزئية لفئات أصغر بحيث تشتمل على (أجهزة كمبيوتر محمول من نوع آي بي أم، أجهزة كمبيوتر مكتبي من نوع دل، برامج أوفيس من نوع مايكروسوفت، .... إلخ).

والمستوى الخامس فهو يحدد كل منتج بعينه، كأن يكون (جهاز كمبيوتر مكتبي من نوع دل ومواصفات i7 ورقم تعريف xxxxx، ... إلخ).

إن التمثيل الهرمي للبيانات والسمات الاسمية غالبًا ما يظهر ضمنيًا في قواعد البيانات ويتم توليده بشكل تلقائي، وكما يظهر في الشكل السابق من خلال الانتماء التلقائي للعناصر في الفئات في المستويات الأدنى للفئات في المستويات الأعلى، ويتطلب هذا الأمر أن يتم بناء وتصميم قواعد البيانات منذ البداية بطريقة صحيحة تساهم في تحقيقه، كأن يتم استحداث الحقول والمتغيرات التي تعبر عن الفئات الأعلى.

وقد يلزم أحيانًا أن يتم توليد الفئات وبناء التسلسل الهرمي للمستويات الأعلى يدويًا، بخاصة إذا لم تكن متوفرة بقاعدة البيانات ولم تؤخذ بالاعتبار عند بدء تصميمها وبنائها، وذلك بأن يتم تطبيق أحد إجراءات تحضير البيانات للتحليل والتنقيب لتوليد حقول وسمات جديدة في قاعدة البيانات، كالتي تم التطرق إليها في فصل تحضير البيانات للتحليل والتنقيب.

فوائد استكشاف قواعد التبعية والارتباط متعددة المستويات

تكمن الفائدة من استكشاف قواعد التبعية والارتباط متعددة المستويات بأنها تساعد في معرفة المزيد من الارتباطات والعلاقات التي يمكن أن تكون أكثر قوة وفائدة للمحللين والباحثين، والتي يمكن استثمارها بشكل حقيقي وفعّال في التخطيط والإدارة بكل أنواعها ومستوياتها.

مثلًا، في قاعدة بيانات حركة مبيعات الأجهزة الإلكترونية، قد يكون من السهل استكشاف قاعدة تمثل علاقة تبعية وارتباط بين حركة بيع أجهزة الكمبيوتر من النوع دل Dell وبيع الكاميرات الرقمية من النوع كانون Canon، ولكنها سوف تكون قاعدة ضعيفة نظرًا لندرة تكرارها، حيث أن عدد قليل من الزبائن يمكن أن يشتروا هذين المنتجين معًا.

وبالتالي فإنه يكون من المتوقع استكشاف قواعد التبعية والارتباط القوية بين الفئات في المستويات الأعلى التي تحتوي على تلك المنتجات، كأن يتم استكشاف قاعدة ارتباط بين مبيعات أجهزة الكمبيوتر والطابعات بشكل

عام، وذلك نظرًا لكثرة تكرار هذا السلوك حيث أنه من الممكن أن يتكرر شراء منتجات من تلك الفئات معًا بغض النظر عن تحديد أنواعها.

طرق استكشاف قواعد التبعية والارتباط متعددة المستويات

إن طريقة استكشاف قواعد التبعية والارتباط متعددة المستويات هي نفس طريقة استكشاف قواعد التبعية والارتباط الاعتيادية، ولكن مع الأخذ بالاعتبار تكرار السجلات في الفئات الجزئية المنبثقة من الفئات في المستويات الأعلى التي يتم استكشاف علاقات التبعية والارتباط فيما بينها.

مثلًا، في قاعدة بيانات حركة مبيعات الأجهزة الإلكترونية، يمكن أن يتم استكشاف القاعدة التالية:

"إذا اشترى الزبون كمبيوتر محمول من نوع ديل فإنه يشتري طابعة من نوع إتش بي معها باحتمال ٦٠٪، وهذه الثنائية تحدث بإجمالي ٢٪ من كافة المشتريات". أما إذا تم استكشاف قاعدة ارتباط بين فئات في المستوى الأعلى فسوف تكون كما يلي:

"إذا اشترى الزبون كمبيوتر محمول فإنه يشتري طابعة معها باحتمال ٨٠٪، وهذه الثنائية تحدث بإجمالي ١٨٪ من كافة المشتريات".

ويلاحظ أنه لم يتم تحديد نوع الكمبيوتر أو الطابعة في هذه القاعدة، ومعنى ذلك أنه تم احتساب كل السجلات الخاصة بمبيعات أجهزة الكمبيوتر

والطابعات من جميع الأنواع، وهو ما يفسر ارتفاع نسبة التغطية لها.

تقييم قواعد التبعية والارتباط متعددة المستويات

يتم تقييم قواعد التبعية والارتباط متعددة المستويات بنفس الطريقة الاعتيادية، من خلال حساب كل من الصحة والتغطية:

- الصحة Accuracy: كم هي نسبة صحة القاعدة (وقوع النتيجة في حال وقوع السبب).
- التغطية Coverage: كم نسبة السجلات المحققة للقاعدة إلى كافة السجلات في قاعدة البيانات.

القواعد المفيدة والقواعد غير المفيدة

بطبيعة الحال قد تظهر بعض القواعد غير المفيدة نتيجة الإمعان في تعدد مستويات التصنيف نزولاً في المستويات الأدنى، بحيث تكون نسبة السجلات المحققة للقاعدة إلى كافة السجلات ضئيلة، كأن يتم استكشاف القاعدة التالية مثلًا:

"إذا اشترى الزبون كمبيوتر محمول من نوع ديل فإنه يشتري طابعة من نوع إتش بي معها باحتمال ٤٥٪، وهذه الثنائية تحدث بإجمالي ١٪ من كافة المشتريات".

أو أن يتم استكشاف قاعدة يمكن وصفها بأنها شخصية وفريدة ولا يمكن

الاستفادة منها بغرض التعميم مثل:

"إذا اشترى الزبون كمبيوتر محمول من نوع ديل بمواصفات (س) فإنه يشتري طابعة من نوع إتش بي بمواصفات (ص) معها باحتمال ٦٠٪، وهذه الثنائية تحدث بإجمالي ٠,١٪ من كافة المشتريات".

ويلاحظ في مثل هذا النوع من قواعد التعبية والارتباط ارتفاع نسبة صحة القاعدة حيث أنها فريدة من نوعها وتتحقق دائمًا فور ظهورها في سجلات قاعدة البيانات، ولكن ظهورها لا يُعتد به لأنه نادرًا ما يحدث، وبالتالي تصبح هذه القاعدة غير مفيدة.

تنقيب واستكشاف قواعد التبعية والارتباط متعددة الأبعاد

Mining Multi-Dimensional Associations and Correlations

في القسم السابق تم استعراض قواعد التبعية والارتباط البسيطة، والتي تعني بالتوقع الأحادي، أي التوقع بشراء منتج واحد وعلاقته بشراء منتج آخر كما في القاعدة التالية:

"إذا اشترى الزبون كاميرا رقمية فإنه يشتري طابعة"

حيث يتكرر في هذه القاعدة فعل الشراء باعتباره بُعدًا واحدًا.

وفي قواعد البيانات عادة ما يكون لدينا العديد من المتغيرات والسمات التي يمكن أن تكون محل الدراسة والبحث أو التحليل والتنقيب باعتبارها أبعادًا متعددة.

مثلاً، يمكن أن تحتوي قاعدة بيانات حركة مبيعات الأجهزة الإلكترونية على معلومات إضافية عن الزبون الذي يقوم بعملية الشراء مثل (عمر الزبون، الوظيفة، الدخل، العنوان، ... إلخ)، كل هذه البيانات يمكن أن يتم تخزينها في قاعدة بيانات حركة المبيعات، وتعتبر هذه السمات أبعادًا أخرى وبالتالي يمكن استكشاف قواعد التبعية والارتباط متعددة الأبعاد بحيث يكون شكلها كما يلي:

"إذا كانت الفئة العمرية للزبون هي (من ٢٠-٢٩ سنة) ومهنته (طالب)

فإنه (يشتري جهاز كمبيوتر محمول)"

أي أن قواعد التبعية والارتباط متعددة الأبعاد تشتمل على اثنين أو أكثر من المتغيرات، فالقاعدة السابقة اشتملت على ثلاثة أبعاد وهي (العمر، الوظيفة، فعل الشراء)، والتي تظهر كل منها مرة واحدة في القاعدة، لذا فهي لا تشتمل على أبعاد متكررة، وتسمى القواعد من هذا النوع متعددة الأبعاد بدون تكرارات.

أما إذا ظهرت تكرارات لبعض الأبعاد فإنها تسمى قواعد متعددة الأبعاد الهجينة، ويمكن توضيح هذا النوع من خلال المثال التالي:

"إذا كانت الفئة العمرية للزبون (من ٢٠-٢٩ سنة) واشترى جهاز كمبيوتر محمول، فإنه يشتري طابعة من نوع إتش بي."

أي أن فعل الشراء تكرر مرتين في القاعدة.

قواعد التبعية والارتباط الأسمية والكمية

Nominal & Quantitative Association Rules

يمكن للمتغيرات والسمات في قواعد البيانات أن تكون من النوع الاسمي أو الكمي، والقيم الاسمية كأسماء الأشياء والمسميات لها عدد محدود من القيم ولا يهم الترتيب فيها، مثلاً الوظيفة، اللون، نوع المنتج، ... إلخ).

أما المتغيرات الكمية فهي متغيرات رقمية وتتضمن ترتيب تلقائي لها وسط بقية القيم، مثل (العمر، الدخل، سعر المنتج، ... إلخ).

وعند استكشاف قواعد التبعية والارتباط متعددة الأبعاد قد يتم اللجوء إلى تحويل المتغيرات الرقمية إلى متغيرات اسمية باستخدام أسلوب التفريد الذي تطرقنا إليه سابقًا، كأن يتم استحداث متغير جديد ليعبر عن فئة الدخل حتى تكون أحد القيم التالية (من صفر إلى ٢٠٠٠، من ٢٠٠١ إلى ٣٠٠٠، من ٣٠٠١ إلى ٤٠٠٠، .....، أكثر من ١٠٠٠٠)، فبهذه الطريقة يتم التعامل مع المتغير الجديد باعتباره من النوع الاسمي وله قيم محددة ومعروفة وقابلة للعد.

وتكمن الفائدة في تحويل البيانات الرقمية غير المحدودة إلى بيانات اسمية محدودة بأسلوب التفريد في أنها سوف تفيد في استكشاف قواعد ارتباط قوية وذات دلالة ويمكن استثمارها، ويظهر هذا جلياً إذا ما نظرنا للمثال التالي الذي يعبر عن قاعدة ارتباط تستخدم متغير رقمي قبل إجراء عملية

التفريد:

"إذا كان (دخل الزبون = ٣٤٣٥) و(اشترى جهاز كمبيوتر محمول من النوع ديل) فإنه (يشتري طابعة من النوع إتش بي)".

ويُلاحظ من هذه القاعدة أنها حددت قيمة الدخل بشكل دقيق كقيمة رقمية، ونتيجة لهذا التحديد الرقمي فإنه قد تنطبق على عدد قليل جدًا من الزبائن في قاعدة البيانات، وربما لا تنطبق إلا على زبون واحد فقط، لتصبح حالة فريدة يستحيل البناء عليها واعتبارها قاعدة مفيدة حيث أن تغطيتها سوف تكون منخفضة جدًا.

أما إذا قمنا باستحداث متغير جديد وتسميته بعنوان (فئة الدخل) فإنه يمكن استخدامه في استكشاف القاعدة التالية:

"إذا كان دخل الزبون ينتمي للفئة (من ٣٠٠١ إلى ٤٠٠٠) و(اشترى جهاز كمبيوتر محمول من النوع ديل) فإنه (يشتري طابعة من النوع إتش بي)".

ويتضح من هذا المثال أن تكون التغطية فيه مرتفعة نسبيًا مقارنة بالمثال السابق، نظرًا لاتساع قيم المتغير الخاص بالدخل وشموله على عدد كبير من القيم، باعتبار أنه أصبح عبارة عن فئة من النوع الاسمي.

طرق استكشاف قواعد التبعية والارتباط الكمية

بالرغم من إمكانية تحويل البيانات الرقمية إلى بيانات اسمية إلاّ أنه يمكن

التعامل مع البيانات الرقمية بعدة طرق أخرى مع استمرار احتفاظها بالسمة الرقمية، كأن يتم استخدام نظريات الإحصاء وتطبيق المقاييس الإحصائية المختلفة على الأرقام كالمتوسط الحسابي، ويظهر ذلك كما في المثال التالي:

"إذا كان متوسط دخل الزبون (٤٠٠٠) واشترى جهاز كمبيوتر محمول فإنه يشتري طابعة".

أي أنه تم التعامل مع القيمة الرقمية التي تمثل الدخل باستخدام المقياس الإحصائي (المتوسط الحسابي) والذي يتم حسابه من واقع كل السجلات التي تحتوي على حركة شراء جهاز كمبيوتر محمول.

كما توجد العديد من الطرق الأخرى الأكثر تعقيدًا والتي يمكن استخدامها في توليد قواعد التبعية والارتباط الكمية، وهي تتنوع بتنوع أساليب التحليل والتنقيب التي يتم تطبيقها وبحسب احتياجات المحللين والباحثين والأهداف التي يسعون لاستكشافها من تلك القواعد.

تنقيب واستكشاف الأنماط النادرة والسلبية

Mining Rare Patterns and Negative Patterns

إن جميع قواعد التبعية والارتباط والأنماط المتكررة التي يتم استكشافها هي تلك التي ترتفع فيها معدلات التكرار في قواعد البيانات، ومع ذلك فقد يكون من الشيق أيضًا أن يتم استكشاف الأنماط نادرة الحدوث بدلاً من الأنماط كثيرة التكرار، أو استكشاف الأنماط التي تعكس علاقة ارتباط عكسية بين المتغيرات المختلفة، وهذه الأنواع من الأنماط تسمى الأنماط النادرة والأنماط السلبية على الترتيب.

أمثلة توضيحية

في قاعدة بيانات مبيعات محل للمجوهرات يمكن ملاحظة ندرة سجلات بيع الساعات المصنوعة من الألماس، ومع ذلك فإنه يمكن أن تمثل تلك السجلات النادرة قاعدة شيقة بالرغم من أنها نادرة الحدوث، حيث أنها تمثل فرصة استثمارية جيدة لصاحب المحل وفيها ربح وفير، الأمر الذي يجعله يهتم بها ويتفحص كل ما يخصها من بيانات، سواء من حيث خصائص الزبائن الذين قاموا بعمليات الشراء أو التوقيت والظروف والأسعار التي كانت سائدة لهذا المنتج وقت شرائه.

من جهة أخرى، ففي قاعدة بيانات مبيعات أحد المراكز التجارية إذا وجدنا أن الزبائن يشترون بشكل متكرر المشروبات الغازية العادية أو

المشروبات الغازية منزوعة السكريات Diet، ولكنهم لا يقومون بشرائها معًا، فإنه في هذه الحالة يمكن اعتبار أن شراء المياه الغازية العادية وشراء المياه الغازية منزوعة السكريات معًا هو نمط لعلاقة ارتباط عكسية.

تعريف

إذا كان لدينا مجموعتين من العناصر (أ)، (ب) متعددي التكرار في قاعدة البيانات ولكنهما نادرًا ما يحدثون معًا، أي أن:

نسبة تغطية [(أ) مع (ب)] أقل من [(نسبة تغطية (أ) × نسبة تغطية (ب)]

فإن المجموعة (أ) والمجموعة (ب) مرتبطتان عكسيًا، والمجموعة (أ U ب)، "أ اتحاد ب"، هي نمط لعلاقة ارتباط سلبية.

وإذا كانت نسبة تغطية (أ U ب) أقل بكثير من [(نسبة تغطية (أ) × نسبة تغطية (ب)]

فإن المجموعة (أ) والمجموعة (ب) مرتبطتان عكسيًا بقوة، والمجموعة (أ U ب) هي نمط لعلاقة ارتباط سلبية قوية.

مثال تطبيقي

في قاعدة بيانات أحد المراكز التجارية، نفرض أنه لدينا (٢٠٠) سجل وأنها تحتوي على تكرارات مبيعات المنتج من النوع (أ) بعدد (١٠٠) سجل،

ومبيعات المنتج (ب) تكرارات بعدد (١٠٠) سجل، بشكل منفرد لكل منها، ولكن هناك حالة سجل واحد يُظهر بيع المنتج (أ) والمنتج (ب) معًا، ففي هذه الحالة فإنه يكون:

نسبة تغطية مبيعات المنتج (أ) = ١٠٠ / ٢٠٠ = ٠,٥

نسبة تغطية مبيعات المنتج (ب) = ١٠٠ / ٢٠٠ = ٠,٥

ونسبة تغطية مبيعات المنتجين معًا (أ) و (ب) = ١/٢٠٠ = ٠,٠٠٥

ويكون: نسبة تغطية مبيعات المنتج (أ) × نسبة تغطية مبيعات المنتج (ب) = ٠,٥ × ٠,٥ = ٠,٢٥

أي أن:

[نسبة تغطية مبيعات المنتجين (أ) و (ب) معًا] أصغر بكثير من [نسبة تغطية مبيعات المنتج أ × نسبة تغطية مبيعات المنتج ب].

لذا فإنه يُقال أن المنتج (أ) مرتبط عكسيًا مع المنتج (ب)، حيث أن شراء أحدهما لا يُشجع على شراء المنتج الآخر.

تنقيب واستكشاف الأنماط وقواعد التبعية والارتباط المشروطة
Mining Conditional Patterns & Association Rules

إن عمليات التنقيب يمكن أن تؤدي لاستكشاف الآلاف من قواعد التبعية والارتباط بداخل قاعدة البيانات الواحدة، والتي يمكن أن يكون معظمها غير ذات أهمية للمحلل، وغالبًا ما يستخدم المحلل حدسه ورؤيته حتى يحدد الاتجاه الصحيح للتنقيب الذي يؤدي به لاستكشاف الأنماط والعلاقات الشيقة وأشكال قواعد الارتباط والتبعية الأكثر فائدة وأهمية والتي يرغب بالعثور عليها.

كما قد يلجأ المحلل إلى تحديد بعض الشروط أو القيود التي يمكن أن تساعد في استكشاف القواعد المهمة بحسب حدسه البحثي، بحيث تمنع هذه القيود من التوصل إلى القواعد غير المفيدة من وجهة نظره، ويكون من المستحسن أن يتم توظيف الحدس لوضع الاستثناءات لتكون بمثابة قيود تضبط ما يتم التوصل إليه من قواعد، ويمكن أن تشتمل تلك القيود على أي مما يلي:

١- قيود على أساليب التنقيب المستخدمة.
٢- قيود على أنواع البيانات.
٣- قيود على الأبعاد والمستويات.
٤- قيود على طرق التقييم (الصحة والتغطية).
٥- قيود على شكل القاعدة المستكشفة.

وقد تم التطرق إلى لمعظم تلك الطرق في الأبواب المختلفة في هذا الكتاب، باعتبار أنه يتناول الأساليب المتعددة للتنقيب والأنواع المختلفة للبيانات وتعدد الأبعاد والمستويات وطرق التقييم.

مثال توضيحي

في قاعدة بيانات مبيعات محلات الأجهزة الإلكترونية، نفرض أن فريق المحللين بصدد استكشاف علاقات الارتباط بين السمات الأساسية للزبائن والمنتجات التي يشترونها، وبدلاً من توصيف كل قواعد الارتباط التي تبين هذه العلاقة فيمكن الاهتمام باستكشاف سمات الزبائن الذين يقومون بشراء برامج أوفيس تحديدًا.

القاعدة التالية تمثل نموذجًا من هذا النوع:

"إذا كان الزبون بخصائص وسمات (أ، ب، ج، ..)، فإنه يشتري برنامج أوفيس".

ويتم استكشاف الخصائص والسمات (أ، ب، ج، ...) أثناء عملية التنقيب.

وبشكل عام فإنه يتم استكشاف وتوصيف القاعدة بوضوح من خلال كل السجلات التي تحتوي على مبيعات برامج الأوفيس، والموضحة في الطرف الأيسر من القاعدة. مثلاً، قد يؤدي التنقيب الهادف لتوصيف القاعدة السابقة إلى التوصل لما يلي:

"إذا كان عمر الزبون في الفئة العمرية (من٣٠ - ٣٩ سنة) والدخل (من ٤٠٠١ إلى ٦٠٠٠)، فإنه يشتري برنامج أوفيس.

مع تحديد كل من نسبة الصحة والتغطية لها.

أمثلة الأنماط الهائلة

على الرغم من أننا استعرضنا طرق تنقيب الأنماط المتكررة في مواقف مختلفة، إلا أن العديد من التطبيقات لديها أنماط مخفية يصعب تنقيبها، ويرجع ذلك أساسًا إلى طولها الهائل أو حجمها. ومن أمثلة ذلك تطبيقات المعلوماتية الحيوية، على سبيل المثال، حيث يكون النشاط الشائع هو تحليل بيانات الحمض النووي أو المصفوفة الدقيقة. يتضمن ذلك رسم خرائط وتحليل تسلسلات طويلة جدًا من الحمض النووي والبروتين. يهتم الباحثون بالعثور على أنماط كبيرة (مثل التسلسلات الطويلة) أكثر من البحث عن أنماط صغيرة لأن الأنماط الأكبر عادة ما تحمل معنى أكثر أهمية، ويُطلق على هذه الأنماط الكبيرة أنماطًا ضخمة، حيث تتميز عن الأنماط ذات مجموعات الدعم الكبيرة. يُعد العثور على أنماط ضخمة أمرًا صعبًا لأن التنقيب يميل إلى "الوقوع" في فخ العدد الهائل من الأنماط متوسطة الحجم قبل أن يصل إلى أنماط مرشحة ذات حجم كبير.

تقييم قواعد التبعية والارتباط، القواعد المفيدة وغير المفيدة
Rules Evaluating

إن السؤال المهم الذي يمكن طرحه عند استكشاف قواعد التبعية الارتباط هو: هل كل القواعد القوية المستكشفة تكون مفيدة؟ أي هل إذا كانت قيم نسبة التغطية والصحة للقاعدة مرتفعة تكون القاعدة مفيدة ؟

في الواقع، إن الإجابة على هذه الأسئلة هي (لا)، فالمحلل والمنقب هو الشخص الوحيد الذي يستطيع الحكم على القواعد المستكشفة وما إذا كانت مفيدة أم لا، وهذا الأمر قد يختلف من محلل إلى آخر، والطريقة المثلى لتقدير ما إذا كانت القاعدة مفيدة أم لا هي باللجوء لأدوات الإحصاء الكلاسيكية، التي يتم تطبيقها على البيانات المستكشفة ومقارنتها بما يتم الحصول عليه من قواعد ارتباط وتبعية ومن ثم الحكم على القاعدة بحسب نتيجة المقارنة.

مثال توضيحي

نفرض أننا مهتمون بتحليل حركة مبيعات أحد محلات الأجهزة الإلكترونية، وبشكل محدد حركة بيع كل من ألعاب الكمبيوتر والفيديو، وأنه لدينا عدد إجمالي من السجلات في جدول حركة المبيعات وهو (١٠٠٠٠) سجل، وأن البيانات تُشير إلى وجود (٦٠٠٠) سجل عملية بيع لألعاب الكمبيوتر، كما تُشير لوجود (٧٥٠٠) سجل لعملية بيع ألعاب الفيديو، وعدد (٤٠٠٠) سجل تحتوي على عمليات بيع ألعاب الكمبيوتر والفيديو معًا،

ونفرض أن الهدف من التنقيب هو استكشاف قواعد التبعية والارتباط بداخل قاعدة البيانات، بحيث يكون الحد الأدنى لنسبة التغطية فيها ٣٠٪ من السجلات، والحد الأدنى لنسبة الصحة لها ٦٠٪ من السجلات.

لذا تكون القاعدة التالية مقبولة للاستكشاف باعتبارها تحقق تلك الشروط، وشكل القاعدة كما يلي:

إذا اشترى الزبون (س) لعبة كمبيوتر >>> فإنه يشتري لعبة فيديو

حيث نسبة التغطية والصحة لهذه القاعدة كما يلي:

- التغطية = ٤٠٠٠ / ١٠٠٠٠ = ٠,٤٠
- الصحة = ٤٠٠٠ / ٦٠٠٠ = ٠,٦٦

وتعتبر هذه القاعدة، وفق هذه النسب، هي قاعدة قوية، حيث أن نسبة تغطيتها هي ٤٠٪ ونسبة صحتها هي ٦٦٪، ولكن الملفت للانتباه أن احتمال شراء ألعاب الفيديو هو ٧٥٪، وذلك نظراً لوجود (٧٥٠٠) سجل تحتوي على مبيعات ألعاب الفيديو من أصل (١٠٠٠٠)، وهذه النسبة (٧٥٪) هي في الواقع أكبر من نسبة صحة القاعدة نفسها والمقدرة بقيمة (٦٦٪)!

ووفق هذه المعطيات الجديدة فإنه يمكن التكهن بأن عمليات شراء ألعاب الكمبيوتر ترتبط عكسياً بعمليات شراء ألعاب الفيديو، أي أن من يشتري النوع الأول لا يشتري النوع الثاني.

ويتضح من هذا المثال أن عدم الدراية بمبادئ علم الإحصاء أو عدم الفهم الصحيح للظواهر الإحصائية سوف يؤدي إلى سوء تقدير قواعد التبعية

والارتباط التي يتم استكشافها، مما قد يؤدي إلى اتخاذ قرارات غير حكيمة في مجالات الأعمال وغيرها من المجالات إذا كانت تعتمد على مثل ذلك النوع من القواعد.

قياس نوع وقوة الارتباط في قواعد التبعية والارتباط المستكشفة
Correlation Type & Power Measurement

رأينا في المثال السابق كيف يمكن لقاعدة التبعية والارتباط أن تكون لها نسبة تغطية ونسبة صحة مرتفعتين، وبالرغم من ذلك فإنه ليس من الضروري أن حدوث طرفها الأيمن (الشرط) يؤدي إلى حدوث طرفها الأيسر (النتيجة).

وبدلاً من استخدام نسبة التغطية ونسبة الصحة من أجل تقدير العلاقة بين الحدثين في طرفي قاعدة الارتباط فإنه يمكن استخدام نظرية الاحتمالات من أجل معرفة طبيعة الارتباط بينهما وحساب قوته.

ونتلخص نظرية الاحتمالات أنه لأي حدثين (أ، ب) فإنهما يكونان مستقلين (أي أن حدوث أحدهما غير مرتبط بحدوث الآخر) فقط إذا كان:

احتمال وقوع الحدثين معًا (أ U ب) = [احتمال وقوع الحدث (أ)] × [احتمال وقوع الحدث (ب)].

وفيما عدا ذلك، فإن الحدثين (أ)، (ب)، هما مرتبطان كأحداث، أي يعتمد وقوع أحدهما على وقوع الآخر بشكل ما.

كما أن علاقة الارتباط بينهما يمكن قياسها باستخدام معامل الارتباط بالمعادلة التالية:

معامل الارتباط = احتمال وقوع الحدثين معًا (أ U ب) / [احتمال وقوع الحدث (أ) × احتمال وقوع الحدث (ب)].

وإذ كانت قيمة معامل الارتباط أقل من الواحد الصحيح فيكون وقوع الحدث (أ) مرتبط عكسيًا بوقوع الحدث (ب)، أما إذا كانت قيمة معامل الارتباط أكبر من الواحد الصحيح فيكون الحدثان أ، ب مرتبطين بشكل إيجابي، أي أن وقوع أحدهما يؤدي إلى وقوع الحدث الآخر. أما إذا كانت قيمة معامل الارتباط هي واحد صحيح فيكون الحدثان أ، ب مستقلين ولا يوجد أي ارتباط بينهما.

مثال توضيحي

بالعودة لنفس المثال السابق، والمتعلق بتحليل حركة مبيعات ألعاب الكمبيوتر والفيديو، ليكون الحدث (أ) يعبر عن حركة مبيعات ألعاب الكمبيوتر، والحدث (ب) يعبر عن حركة مبيعات ألعاب الفيديو، ومن البيانات المتوفرة يكون:

احتمال وقوع الحدث (أ) = احتمال شراء لعبة الكمبيوتر = ٦٠٠٠ / ١٠٠٠٠ = ٠,٦٠

احتمال وقوع الحدث (ب) = احتمال شراء لعبة الفيديو = ٧٥٠٠ / ١٠٠٠٠ = ٠,٧٥

احتمال وقوع الحدثين معًا (أ U ب) = احتمال شراء لعبة الكمبيوتر

ولعبة الفيديو

= ٤٠٠٠ / ١٠٠٠٠ =

٠,٤٠

ومن جميع ما سبق يكون:

معامل الارتباط = احتمال وقوع الحدثين معًا (أ U ب) / [احتمال وقوع الحدث (أ) × احتمال وقوع الحدث (ب)]

= ٠,٤٠ / [٠,٦٠ × ٠,٧٥]

= ٠,٨٩

وحيث أن قيمة معامل الارتباط هنا هي أقل من الواحد الصحيح، لذا فإنه يوجد علاقة ارتباط عكسية بين وقوع الحدث (أ) الذي يمثل شراء ألعاب الكمبيوتر ووقوع الحدث (ب) الذي يمثل شراء ألعاب الفيديو، أي أن شراء أحدهما لا يُشجع على شراء الآخر.

تطبيقات تنقيب الأنماط في الحياة العملية

لقد درسنا العديد من جوانب تنقيب الأنماط المتكررة، مع موضوعات متنوعة من خوارزميات التنقيب الفعالة وتنوع الأنماط التي يتم استكشافها، وفيما يلي نستعرض سبب جذب هذا المجال الكثير من الاهتمام. وما هي بعض مجالات التطبيق التي يكون فيها تنقيب الأنماط المتكررة مفيدًا؟

في الواقع، فقد تطرقنا إلى العديد من مجالات التطبيق بالفعل، مثل تحليل سلة المشتريات وتحليل الارتباط، ومع ذلك يمكن تطبيق تنقيب الأنماط المتكررة في العديد من المجالات الأخرى أيضًا. وتتراوح هذه من المعالجة المسبقة للبيانات وتصنيفها إلى التجميع وتحليل البيانات المعقدة.

للتلخيص، يُعد تنقيب الأنماط وقواعد الارتباط مهمة تنقيب عن البيانات تكتشف الأنماط التي تحدث بشكل متكرر معًا و/ أو لها بعض الخصائص المميزة التي تميزها عن غيرها، وغالبًا ما تكشف عن شيء متأصل وقيّم. قد تكون الأنماط عبارة عن مجموعات عناصر أو تكوينات فرعية أو قيم. تتضمن المهمة أيضًا اكتشاف الأنماط النادرة، وكشف العناصر التي نادرًا ما تحدث معًا ولكنها ذات أهمية.

يؤدي اكتشاف الأنماط المتكررة والأنماط النادرة إلى العديد من التطبيقات الواسعة والمثيرة للاهتمام، على النحو التالي:

يستخدم تنقيب الأنماط على نطاق واسع لتصفية الضوضاء وتنظيف

البيانات مثل المعالجة المسبقة في العديد من التطبيقات كثيفة البيانات. يمكننا استخدامه لتحليل بيانات المصفوفات الدقيقة، على سبيل المثال، والتي تتكون عادةً من عشرات الآلاف من الأبعاد (على سبيل المثال، تمثيل الجينات). يمكن أن تكون هذه البيانات صاخبة إلى حد ما. يمكن أن يساعدنا تنقيب الأنماط على التمييز بين ما هو مزعج وما هو ليس كذلك. قد نفترض أن العناصر التي تحدث بشكل متكرر معًا أقل احتمالًا أن تكون ضوضاء عشوائية ويجب عدم تصفيتها.

من ناحية أخرى، من المحتمل أن تكون الكلمات التي تحدث بشكل متكرر (على غرار كلمات الإيقاف في المستندات النصية) غير واضحة ويمكن تصفيتها. يمكن أن يساعد تنقيب الأنماط المتكررة في تحديد المعلومات الأساسية وتقليل الضوضاء.

غالبًا ما يساعد تنقيب الأنماط في اكتشاف الهياكل والمجموعات المتأصلة المخفية في البيانات. بالنظر إلى مجموعة بيانات الأبحاث، على سبيل المثال، يمكن لتنقيب الأنماط بسهولة العثور على مجموعات مثيرة للاهتمام مثل مجموعات المؤلفين المشتركين (عن طريق فحص المؤلفين الذين يتعاونون بشكل متكرر) ومجموعات المؤتمرات (عن طريق فحص مشاركة العديد من المؤلفين والمصطلحات الشائعة). يمكن استخدام هذا الهيكل أو اكتشاف الكتلة كعملية معالجة مسبقة لاستخراج البيانات الأكثر تعقيدًا.

على الرغم من وجود العديد من طرق التصنيف، فقد وجد البحث أنه

يمكن استخدام الأنماط المتكررة كأجزاء أساسية في بناء نماذج تصنيف عالية الجودة، ومن ثم يطلق عليها التصنيف المستند إلى الأنماط. النهج ناجح لأن (١) ظهور عنصر (عناصر) أو مجموعة (عناصر) نادرة جدًا يمكن أن يكون ناتجًا عن ضوضاء عشوائية وقد لا يكون موثوقًا به لبناء النموذج، ومع ذلك فإن النمط المتكرر نسبيًا غالبًا ما يحمل المزيد من المعلومات لبناء المزيد نماذج موثوقة (٢) الأنماط بشكل عام (أي، مجموعات العناصر التي تتكون من سمات متعددة) عادةً ما تحمل اكتساب معلومات أكثر من سمة واحدة (ميزة)؛ و(٣) الأنماط التي يتم إنشاؤها على هذا النحو غالبًا ما تكون مفهومة بشكل حدسي ويسهل شرحها. أفادت الأبحاث الحديثة بالعديد من الأساليب التي تميز الأنماط المثيرة للاهتمام والمتكررة والتمييزية واستخدامها في التصنيف الفعال.

يمكن أيضًا استخدام تنقيب الأنماط المتكررة بشكل فعال لتجميع الفضاء الجزئي في الفضاء عالي الأبعاد. يمثل التجميع تحديًا في الفضاء عالي الأبعاد، حيث يصعب غالبًا قياس المسافة بين جسمين. وذلك لأن هذه المسافة تهيمن عليها مجموعات الأبعاد المختلفة التي تسكن فيها الكائنات.

وبالتالي، بدلاً من تجميع الكائنات في مساحاتها الكاملة عالية الأبعاد، قد يكون من المفيد العثور على مجموعات في فضاءات فرعية معينة. في الآونة الأخيرة، طور الباحثون أساليب نمو نمطية تعتمد على الفضاء الجزئي والتي تجمع الكائنات بناءً على أنماطها المتكررة الشائعة. لقد أظهروا أن مثل هذه

الأساليب فعالة لتجميع بيانات التعبير الجيني القائمة على المصفوفات الميكروية.

يعد تحليل الأنماط مفيدًا في تحليل البيانات الزمانية المكانية وبيانات السلاسل الزمنية وبيانات الصور وبيانات الفيديو وبيانات الوسائط المتعددة. مجال تحليل البيانات الزمانية المكانية هو اكتشاف أنماط تحديد الموقع. يمكن أن تساعد هذه، على سبيل المثال، في تحديد ما إذا كان مرض معين ينتشر جغرافيًا مع أشياء معينة مثل بئر أو مستشفى أو نهر.

في تحليل بيانات السلاسل الزمنية، قام الباحثون بتقدير قيم السلاسل الزمنية في فترات (أو مستويات) متعددة بحيث يمكن تجاهل التقلبات الصغيرة والاختلافات في القيمة. يمكن بعد ذلك تلخيص البيانات في أنماط متسلسلة، والتي يمكن فهرستها لتسهيل البحث عن التشابه أو التحليل المقارن. في تحليل الصور والتعرف على الأنماط، حدد الباحثون أيضًا الأجزاء المرئية التي تحدث بشكل متكرر على أنها "كلمات مرئية"، والتي يمكن استخدامها للتجميع الفعال والتصنيف والتحليل المقارن.

كما تم استخدام تنقيب الأنماط لتحليل التسلسل أو البيانات الهيكلية مثل الأشجار والرسوم البيانية والتكرارات اللاحقة والشبكات. في هندسة البرمجيات، حدد الباحثون التكرارات المتتالية أو المتقطعة في تنفيذ البرنامج على أنها أنماط متسلسلة تساعد في تحديد الأخطاء البرمجية. يمكن التعرف على أخطاء النسخ واللصق في البرامج الكبيرة من خلال تحليل النمط المتسلسل الممتد لبرامج المصدر.

يمكن تحديد البرامج المسروقة بناءً على هياكل تدفق / حلقة البرنامج المتطابقة بشكل أساسي. يمكن تحديد التركيبات الفرعية للجملة شائعة الاستخدام لدى المؤلفين واستخدامها لتمييز المقالات التي كتبها مؤلفون مختلفون.

يمكن استخدام تنقيب الأنماط المتكررة والتمييزية كهياكل فهرسة بدائية (تُعرف باسم مؤشرات الرسم البياني) للمساعدة في البحث عن مجموعات وشبكات بيانات كبيرة ومعقدة ومنظمة. هذه تدعم البحث عن التشابه في البيانات المهيكلة بالرسوم البيانية مثل قواعد بيانات المركّبات الكيميائية أو قواعد البيانات المهيكلة XML. يمكن أيضًا استخدام هذه الأنماط لضغط البيانات وتلخيصها.

علاوة على ذلك، يتم استخدام تنقيب الأنماط المتكررة في أنظمة التوصية، حيث يمكن للناس العثور على الارتباطات ومجموعات سلوكيات العملاء ونماذج التصنيف القائمة على الأنماط الشائعة أو التمييزية.

أخيرًا، تعزز الدراسات حول طرق الحساب الفعالة في تنقيب الأنماط بشكل متبادل العديد من الدراسات الأخرى حول الحوسبة القابلة للتطوير.

الفصل الخامس
التحليل والتنقيب باستخدام خوارزميات التصنيف والتنبؤ

Classification & Predicting Data Mining Algorithms

١. مفاهيم أساسية
٢. التصنيف باستخدام استقراء شجرة القرار Decision Tree Induction
٣. التصنيف باستخدام نظرية الاحتمالات (النظرية الافتراضية) Bayes Theory Classification
٤. التصنيف باستخدام النظرية الافتراضية الشبكية Network Bayes Classification
٥. التصنيف باستخدام استقراء قواعد الارتباط Classification Using Rule Induction
٦. التصنيف باستخدام خوارزمية الشبكة العصبية Neural Network Algorithm
٧. التصنيف باستخدام خوارزمية الجار الأقرب Nearest Neighbor Algorithm
٨. خوارزميات التصنيف متعددة الفئات Multi-Class Classification Algorithms
٩. تقييم كفاءة واختيار خوارزميات التصنيف Evaluating and Selecting the Classification Algorithms

مفاهيم أساسية

إن خوارزميات التصنيف والتنبؤ هي شكل من أشكال تحليل البيانات والتي تستخلص نماذج تصف بشكل دقيق فئات وتصنيفات البيانات المهمة.

مثلًا، يمكن بناء نموذج تصنيف من أجل تقييم طلبات القروض في نظام منح القروض بأحد البنوك، بحيث يقسم مجموعة الطلبات إلى فئتين، آمنة وغير آمنة (أي أن فيها مخاطرة على البنك)، ومثل هذا التحليل يمكن أن يساعد في فهم البيانات وضبط نظام منح القروض.

وتستخدم خوارزميات التصنيف في العديد من المجالات، حيث أن لها تطبيقات عديدة، ومنها المستخدمة في نظم كشف عمليات الاحتيال، والتسويق المستهدف لفئات معينة، والتنبؤ بمستوى الأداء أو مدى الإقبال على شراء المنتجات، وميكنة الصناعات التحويلية، وتشخيص الأمراض.

مثلًا، إذا افترضنا أن مدير قسم القروض في أحد البنوك أو مدير إدارة وتحليل المخاطر فيه يرغب في تحليل البيانات المتوفرة لدى البنك حتى يعرف أي من طلبات القروض يمكن أن يكون آمنًا وأي منها يمكن أن يكون غير آمن أو فيه مخاطرة على البنك.

أو إذا افترضنا أن مدير التسويق في أحد محلات بيع الأجهزة الإلكترونية يرغب في تحليل بيانات المبيعات حتى يستطيع أن يخمن فيما إذا كان زبون بمواصفات وسمات معينة سوف يقوم بشراء جهاز كمبيوتر تم طرحه حديثًا

في المحل.

أو لو افترضنا أن باحثًا في المجال الطبي يرغب في تحليل البيانات المتوفرة لديه والمتعلقة بنتائج اختبارات علاجية لمرض السرطان حتى يستطيع التنبؤ بأفضل علاج للمرض من ثلاثة أنواع من الأدوية التي يقوم باختبارها على المرضى.

ففي كل مثال من الأمثلة المذكورة يتم استخدام طرق التصنيف من أجل تحليل البيانات، حيث يتم بناء نموذج التصنيف لتكون مهمته التنبؤ بالفئة أو السمات المحددة التي تصف الفئة التي يتم استكشافها، أو بتعبير أوضح الفئة التي ينتمي لها العنصر الذي يتم استكشافه، مثل (آمن أو خطر) في حالة طلبات القروض، و(نعم أو لا) في حالة التسويق المستهدف، والعلاج (أ أو ب أو ج) في حالة العلاج الطبي.

وفي كل الأمثلة المذكورة كانت النماذج التي تمثلها الفئات أو المسميات ذات قيم اسمية ولا يهم فيها الترتيب.

كما يمكن أن تكون القيم التي يتم التنبؤ بها رقمية، كأن يرغب مدير المبيعات في التنبؤ بالمبلغ الذي يمكن أن ينفقه زبون معين أثناء رحلة للتسوق داخل المركز التجاري، وهو ما يُطلق عليه التنبؤ الرقمي، وهو المجال الذي يهتم به أسلوب تحليل الانحدار كأحد أساليب التحليل الإحصائي.

في هذا الفصل سوف يتم التطرق بالتفصيل للتصنيف الهادف للتنبؤ بالقيم الاسمية.

طريقة عمل خوارزميات التصنيف والتنبؤ

إن التحليل باستخدام خوارزميات التصنيف هو عبارة عن عملية مكونة من خطوتين، تسمى الأولى خطوة التعلم Learn Step، حيث يتم فيها بناء نموذج التصنيف، والخطوة الثانية هي خطوة التصنيف Classification Step، حيث يتم فيها استخدام النموذج من أجل التنبؤ بالفئات أو السمات لبيانات محددة.

ويتم في خطوة التعلم تحليل البيانات باستخدام خوارزمية التصنيف ومن ثم استخلاص قواعد التصنيف Classification Rules، والخطوة الثانية هي خطوة التصنيف ويتم فيها استخدام بيانات السجل المستكشف من أجل تقدير مدى تحقق قاعدة أو قواعد التصنيف، فإذا تحققت قاعدة التصنيف فإنه يمكن تطبيقها على السجل المستكشف.

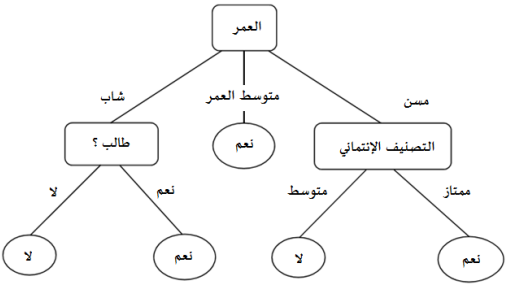
وعادة ما تكون نتيجة التقدير محددة مسبقًا، مما يجعل خطوة التعلم واضحة ومحددة، أما إذا كانت غير مقدّرة وبحاجة إلى تحديد فهذا يقودنا إلى الحاجة للتجزئة العنقودية حتى يمكن تجزئة البيانات إلى مجموعتين محددتين بالسمات المطلوب التصنيف وفقًا لها.

ويمكن تقييم خوارزميات التصنيف باستخدام مقياس نسبة صحة القاعدة Accuracy، وهي نسبة عدد السجلات التي يتم اختبارها وتصنيفها بشكل صحيح باستخدام الخوارزمية.

التصنيف باستخدام استقراء شجرة القرار
Decision Tree Induction

إن شجرة القرار هي نموذج استكشافي يظهر على شكل شجرة، كما يعبر اسمها، وبشكل دقيق يمثل كل فرع من فروعها سؤالاً تصنيفياً وتمثل أوراقها أجزاء من قاعدة البيانات تنتمي للتصنيفات التي تم بنائها.

فعلى سبيل المثال، إذا أردنا تصنيف الزبائن الذين يمكن أن يُقبلوا على شراء جهاز كمبيوتر في أحد محلات الأجهزة الإلكترونية، فإن تركيبة شجرة القرار الخاصة بهذا الشأن قد تظهر بالصورة التالية:



شكل (٥,١)، شجرة القرار الخاصة بتصنيف زبائن أحد محلات الأجهزة الإلكترونية

ففي الشكل أعلاه، يتم تصنيف الزبائن إلى ثلاثة أصناف من خلال السؤال الأول عن عمر الزبون، ثم تتم عملية التصنيف مرة أخرى من خلال المزيد من الأسئلة في الفروع، وهكذا إلى أن يتم الحصول على التصنيف النهائي بأن الزبون سيُقدم على شراء جهاز الكمبيوتر أم لا.

ومن الملاحظ في شجرة القرار أنها تقسم البيانات في كل فرع بدون إنقاص أي منها، أي أن عدد السجلات الكلي في الفرع الأم يساوي مجموع السجلات في الفرعين المنبثقين منه.

وتهدف تقنية شجرة القرار إلى تقسيم قاعدة البيانات بهدف معين سبق وأن تم تحديده، ويصبح وجود عنصر معين في إحدى المجموعات، وهي ممثلة هنا بالفروع، هو نتيجة لأنه حقق سلسلة الشروط الموضوعة وصولاً إلى هذا الفرع وليس فقط لأنه يشبه بقية عناصره، بالرغم من أنه لم يتم تعريف التشابه في هذه الحالة.

إن شجرة القرار والخوارزميات التي تستخدم لإنتاجها يمكن أن تكون معقدة ولكن النتائج التي تؤدي لها يمكن إظهارها بشكل مبسط وسهل الفهم وبفائدة عالية المستوى.

استخدام شجرة القرار في التنبؤ

بالرغم من أن شجرة القرار تستخدم في الاستكشاف وتحضير البيانات للعمليات الإحصائية إلا أنها أيضًا تستخدم وبشكل أكثر للتنبؤ. ومن المهم جدًا عند بناء خوارزمية شجرة القرار أن يؤخذ بعين الاعتبار أن تكون قابلة للتطبيق بقدر الإمكان وبشكل مثالي على كل البيانات المتوفرة، وبالرغم من ذلك سيظل دومًا ثغرات يمكن أن نصادفها عند التعامل مع نوعية معينة من المتغيرات.

والقاعدة الأساسية في بناء شجرة القرار هي إيجاد أفضل سؤال عند كل فرع من فروع الشجرة بحيث يقسم هذا السؤال البيانات إلى قسمين، القسم الأول منها ينطبق عليهم السؤال والقسم الثاني لا ينطبق، وهكذا يتم من خلال سلسلة من الأسئلة بناء شجرة القرار بفروعها المتسلسلة.

ومن الضروري أن تحقق الأسئلة الهدف من تقسيم البيانات والذي يكون دوما محاولة التنبؤ والتقدير. فالفرق بين السؤال الجيد والسؤال غير الجيد هو أنه إلى أي مدى يساعد السؤال في تنظيم البيانات وتقسيمها والتفريق بينها في فروع مختلفة بحيث يحتوي كل فرع على بيانات متجانسة نسبة للسؤال المطروح، وبعض خوارزميات شجر القرار تستخدم الاستكشاف لتحديد السؤال أو يتم تحديده بشكل عشوائي ومن ثم اختيار السؤال الأنسب والذي أدى للتنظيم الأفضل للبيانات.

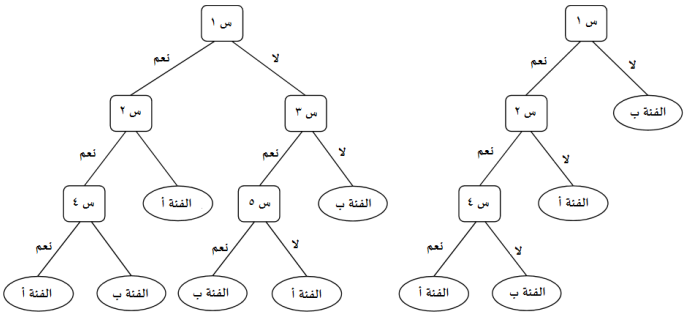
وتستمر عملية وضع الأسئلة وبناء الفروع للشجرة حتى تتوقف عن النمو ويحدث ذلك عندما نصل إلى فرع لا يحتوي إلا على سجل واحد أو على مجموعة من السجلات المتجانسة والتي لا يمكن التفريق بينها بشكل جوهري أو بطريقة تحقق الهدف من التقسيم.

تهذيب شجرة القرار Tree Pruning

عندما يتم بناء شجرة القرار فإن كثير من أفرعها قد تعكس بعض الشذوذ في البيانات محل الدراسة نتيجة الإزعاج أو وجود القيم المتطرفة، وإجراءات

تهذيب شجرة القرار تقوم بحل هذه المشكلة وذلك من خلال قص أو قطع الأفرع الشاذة باستخدام المقاييس الإحصائية المختلفة.

الشكل التالي يوضح شجرة قرار تم تهذيبها، ويظهر في الشكل أن شجرة القرار المهذبة أصغر وأقل تعقيدًا من الشجرة الأصلية، فهي أسهل للفهم وأسرع في تصنيف البيانات من الشجرة غير المهذبة.



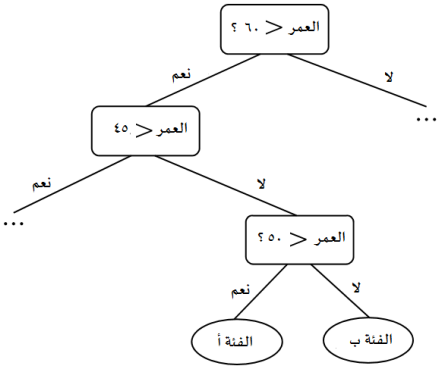
شكل (٥,٢)، تهذيب شجرة القرار

وتنقسم عملية التهذيب إلى مرحلتين، في المرحلة الأولى يتم اختيار الأفرع التي سوف يتم اقتطاعها من خلال اتخاذ قرار عدم الحاجة إلى مزيد من الأفرع والتجزئة عند نقطة معينة، والمرحلة الثانية يتم اعتبار هذه النقطة كنقطة نهاية وكأنها ورقة شجر ويتم منح هذه الورقة قيمة الفئة الأكثر تكرارًا للأوراق التي كانت متفرعة من كل الفرع. وهو ما نلاحظه في الشكل، حيث تم استبدال الفرع (س٣) كله واقتطاعه، واستبداله بورقة شجر واحدة وأعطيت القيمة (الفئة ب) حيث أنها الأكثر تكرارًا في أوراق الشجر في الفرع الذي تم اقتطاعه.

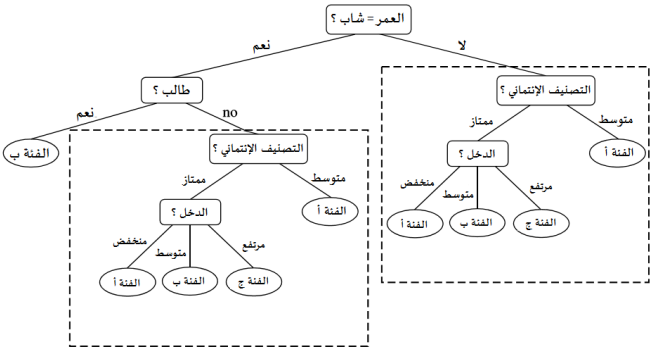
التكرار والاستنساخ في شجرة القرار

بالرغم من أن عملية تهذيب شجرة القرار ينتج عنها شجرة مختصرة وأصغر من الشجرة الأصلية، إلاّ أنها يمكن أن تظل كبيرة ومعقدة نسبيًا ولأسباب أخرى، كأن تُعاني من التكرار والاستنساخ، مما يجعلها صعبة التفسير.

الأشكال التالية توضح نموذج لكل حالة كما يلي:



شكل رقم (٥,٣)، شجرة قرار تحتوي على تكرار يتمثل في السؤال عن العمر



شكل رقم (٥,٤)، شجرة قرار تحتوي على استنساخ يتمثل في تكرار شجرة فرعية كاملة

فالتكرار Repetition يحدث عند إعادة اختبار سمة من السمات في نفس الفرع في الشجرة، وذلك كما يظهر في الشكل (٥-٣)، حيث تم السؤال عن العمر أكثر من مرة في نفس الفرع.

أما الاستنساخ Replication فإنه إعادة تكرار شجرة فرعية كاملة في أماكن متفرقة من الشجرة الرئيسية، وهو ما يظهر في الشكل (٥-٤)، حيث تكررت الشجرة الفرعية المعنونة بـ "التصنيف الائتماني".

وفي الحالتين تتأثر نسبة التغطية والفهم الصحيح لشجرة القرار، والحل الأمثل لهذه المشكلة هو استخدام التقسيم متعدد المتغيرات، أي التقسيم باستخدام تركيبة من السمات معًا، من خلال دمج الشرطين معًا.

ليصبح شكل قاعدة التصنيف كما يلي:

٤٥ < (العمر) < ٦٠

كما يمكن استخدام طرق أخرى في تمثيل البيانات مثل القواعد بدلاً من شجرة القرار، بحيث يتم استنباط القواعد الشرطية (التي تشتمل على عبارات من نوع إذا كان فإن) من شجرة القرار.

التصنيف باستخدام نظرية الاحتمالات (النظرية الافتراضية) Bayes Classification

التصنيف باستخدام نظرية الاحتمالات، أو النظرية الافتراضية، هي طريقة تصنيف إحصائية وتعتمد فكرتها على بناء الاحتمالات، وذلك من خلال التنبؤ باحتمال أن ينتمي سجل من سجلات قاعدة البيانات لفئة محددة، وتستند هذه الطريقة في التصنيف على النظرية الافتراضية ويُستفاد منها في تحليل البيانات الضخمة المخزنة في قواعد البيانات.

وسوف يتم بداية إلقاء نظرة على النظرية الافتراضية.

النظرية الافتراضية Bayes Theory

نظرية الاحتمالات، وتعرف باسم العالم بايز Bayes، وقد وضعها في القرن الثامن عشر، ويمكن توضيحها من خلال المثال التالي:

مثال توضيحي

لو افترضنا أنه لدينا قاعدة بيانات فيها عدد كبير من السجلات، وأن (أ) هو أحد السجلات فيها، وأن هذا السجل معرّف بعدد (ن) من السمات المحددة Attributes، وأن (ف) هي إحدى الفرضيات التي يمكن التعبير عنها رياضيًا أو بأشكال فن للمجموعات، مثل فرضية أن ينتمي السجل (أ) إلى فئة معينة.

ومن أجل تطبيق عمليات التصنيف فإن المطلوب حساب احتمال صحة هذه الفرضية، والذي يتم التعبير عنه رياضيًا بالرمز ح[(ف)/(أ)]، أي ما هو احتمال صحة فرضية أن ينتمي السجل (أ) للفئة (ف) إذا علمنا السمات التي تصف السجل (أ).

مثال تطبيقي

للتوضيح بشكل تطبيقي، ليكن لدينا قاعدة بيانات مبيعات محل الأجهزة الإلكترونية وفيها بيانات الزبائن التي تشتمل على (العمر، الدخل)، ونفرض أنه لدينا زبون (أ) عمره (٣٥ سنة) ودخله (٤٠٠٠)، وأن (ف) هي فرضية أن هذا الزبون سوف يقوم بشراء جهاز كمبيوتر، أو بمعنى آخر هي فرضية أن ينتمي هذا السجل للفئة المعرفة بأنها (شراء جهاز كمبيوتر=نعم).

أي أن ح [(ف)/(أ)] هي نسبة احتمال أن يقوم الزبون (أ) بشراء جهاز كمبيوتر عندما نعلم عمره ودخله.

كما يمكن حساب احتمال أن يقوم أي زبون بشراء جهاز كمبيوتر، بدون تحديد سماته أي بدون شروط، ويتم التعبير عن هذا الاحتمال رياضيًا بالرمز: ح (ف).

وبنفس الطريقة، يمكن حساب احتمال أن يكون الزبون الذي يشتري جهاز كمبيوتر هو بالفعل بعمر (٣٥) سنة ودخله (٤٠٠٠)، أي عكس الاحتمال الأول، ويتم التعبير عن هذا الاحتمال رياضيًا بالرمز:

ح[(أ)/(ف)].

أما احتمال أن يكون زبون (أ) في قاعدة البيانات هو بعمر (٣٥) سنة ودخل (٤٠٠٠)، فيتم التعبير عنه بالرمز: ح(أ).

من جميع ما سبق، يمكن صياغة النظرية الافتراضية بالمعادلة التالية:

$$ح[(ف)/(أ)] = ح[(أ)/(ف)] \times ح(ف) / ح(أ) \quad \text{..... (١)}$$

طريقة استخدام النظرية الافتراضية في التصنيف

إن طريقة استخدام النظرية الافتراضية في التصنيف يمكن تلخيصها في الخطوات التالية:

١. يتم حساب عدد السجلات في قاعدة البيانات محل الدراسة وليكن (ل)، مع تحديد السمات والمتغيرات محل الدراسة، بحيث يكون لكل سجل (أ) عدد (ن) من السمات.

٢. بفرض أنه لدينا عدد من الفئات (م) كما يلي: ف١، ف٢، ف٣، ... ف م، التي تحدد تصنيف كل سجل من سجلات قاعدة البيانات، والمطلوب هو إيجاد الفئة (ف ر) التي ينتمي لها السجل (أ) الذي يتم استكشافه، فيكون هذا السجل المستكشف ينتمي للفئة (ف ر) فقط إذا كان:

ح [(ف ر) / (أ)] > ح [(ف و) / (أ)]، لكل ١ ≤ و ≤ م، و ≠ ر

.... (٢)

أي أن يكون احتمال انتماء السجل (أ) للفئة (ف ر) أكبر من احتمال انتماءه لبقية الفئات الأخرى.

ومن النظرية الافتراضية، وبالتعويض عن ح [(ف ر)/ (أ)] بقيمتها من النظرية وهي:

ح [(ف ر) / (أ)] = ح [(أ) / (ف ر)] × ح (ف ر) / ح (أ)

فيصبح الشرط الموضح في المتباينة رقم (٢) كما يلي:

ح [(أ) / (ف ر)] × ح (ف ر) / ح (أ) > ح [(أ) / (ف و)] × ح (ف و) / ح (أ)

وحيث أن قيمة ح (أ) ثابتة لجميع الفئات، فيمكن اختصار النتيجة إلى المتباينة التالية:

ح [(أ)/(ف ر)] × ح (ف ر) > ح [(أ)/(ف و)] × ح (ف و)

..... (٣)

لكل ١ ≤ و ≤ م، و≠ ر

٣. في قواعد البيانات التي تشتمل على العديد من السمات، قد يكون من الصعب حساب احتمالات ح [(أ)/(ف ر)] لكل السمات الخاصة بالسجل المستكشف (أ)، إلاّ أنه يتم في هذه الحالة افتراض استقلالية تلك

السمات عن بعضها البعض ويكون ناتج احتمالها معًا هو حاصل ضربها جميعًا بحسب نظرية الاحتمالات، أي أنه لحساب احتمال ح[(أ)/(ف ر)] يمكن استخدام المعادلة التالية:

ح [(أ)/(ف ر)] = ح [(أ ١)/(ف ر)] × ح [(أ ٢)/(ف ر)] × ح [(أ ٣)/(ف ر)] × .... × ح [(أ ن)/(ف ر)]

أي:

ح [(أ)/(ف ر)] = مجـ ك = ١ إلى ن ح [(أ ك) /(ف ر)]

حيث أ ١، أ ٢، أ ٣، ..... أ ن، هي كل السمات الخاصة بالسجل (أ).

٤ - في حالة السمات الاسمية تكون طريقة حساب احتمال كل سمة من السمات وتقاطعها مع أي فئة من خلال قسمة عدد مرات تكرار السجل بداخل الفئة.

أي أن ح [(أ ر)/(فر)] تساوي عدد السجلات التي تظهر في الفئة (ف ر) من إجمالي السجلات الموجودة فيها والتي تمتلك السمة (أ ر).

٥ - لكي يتم بناء التوقع الخاص بأحد السجلات (أ)، بمعلومية سماته الخاصة، يتم حساب احتمال ح [(أ)/(فر)] × ح (فر) لكل الفئات المتوفرة، ويقوم نموذج التصنيف بتوقع انتماء السجل للفئة (ف ر) فقط وفقط إذا تحققت المتباينة رقم (٣)، أي إذا كان:

ح [(أ) /(ف ر)] × ح (ف ر) > ح [(أ) / (ف و)] × ح (ف و)،
لكل ١ ≤ و ≤ م، و≠ ر

مثال تطبيقي

نفرض أنه لدينا في قاعدة بيانات مبيعات الأجهزة الإلكترونية الجدول التالي.

ويشتمل الجدول على وصف بعض السمات الخاصة بالزبائن، مثل العمر، الدخل، طالب (نعم أو لا)، التصنيف الائتماني، الفئة التي يتم استكشاف انتماء السجلات لها من عدمه، وتم التعبير عنها في هذا المثال باحتمال الإقبال على شراء جهاز كمبيوتر والقيم المتوقعة هي (نعم أو لا).

م	العمر	الدخل	طالب	التصنيف الائتماني	الفئة: شراء كمبيوتر
١	شاب صغير	مرتفع	لا	متوسط	لا
٢	شاب صغير	مرتفع	لا	ممتاز	لا
٣	متوسط العمر	مرتفع	لا	متوسط	نعم
٤	رجل كبير	متوسط	لا	متوسط	نعم
٥	رجل كبير	منخفض	نعم	متوسط	نعم
٦	رجل كبير	منخفض	نعم	ممتاز	لا
٧	متوسط العمر	منخفض	نعم	ممتاز	نعم
٨	شاب صغير	متوسط	لا	متوسط	لا

٩	شاب صغير	منخفض	نعم	متوسط	نعم
١٠	رجل كبير	متوسط	نعم	متوسط	نعم
١١	شاب صغير	متوسط	نعم	ممتاز	نعم
١٢	متوسط العمر	متوسط	لا	ممتاز	نعم
١٣	متوسط العمر	مرتفع	نعم	متوسط	نعم
١٤	رجل كبير	متوسط	لا	ممتاز	لا

سجلات من قاعدة بيانات المبيعات لمحل أجهزة إلكترونية

ونفرض أن السجل المراد تصنيفه (سجل لا ينتمي لقاعدة البيانات)، هو (أ)، حيث:

(أ): (العمر = شاب صغير، الدخل = متوسط، طالب = نعم، التصنيف الائتماني = متوسط).

ومن بيانات الجدول وأعداد تكرارات السجلات في كل فئة مقارنة بالعدد الإجمالي فيها، يمكن حساب القيم التالية:

احتمالات شراء جهاز الكمبيوتر أو عدم شراؤه:

- ح (شراء الكمبيوتر = نعم) = ٩ / ١٤ = ٠,٦٤٣
- ح (شراء الكمبيوتر = لا) = ٥ / ١٤ = ٠,٣٥٧

كما يمكن حساب الاحتمالات المشروطة الأخرى، وهي احتمالات شراء جهاز الكمبيوتر بحسب سمات الزبون المختلفة:

ح (العمر = شاب صغير، شراء الكمبيوتر = نعم) = ٢ / ٩ = ٠,٢٢٢
ح (العمر = شاب صغير، شراء الكمبيوتر = لا) = ٣ / ٥ = ٠,٦٠٠
ح (الدخل = متوسط، شراء الكمبيوتر = نعم) = ٤ / ٩ = ٠,٤٤٤
ح (الدخل = متوسط، شراء الكمبيوتر = لا) = ٢ / ٥ = ٠,٤٠٠
ح (طالب = نعم، شراء الكمبيوتر = نعم) = ٦ / ٩ = ٠,٦٦٧
ح (طالب = نعم، شراء الكمبيوتر = لا) = ١ / ٥ = ٠,٢٠٠
ح (التصنيف الائتماني = متوسط، شراء الكمبيوتر = نعم) = ٩/٦ = ٠,٦٦٧

ح (التصنيف الائتماني = متوسط، شراء الكمبيوتر = لا) = ٢ / ٥ = ٠,٤٠٠

ومن كل الاحتمالات السابقة يمكن حساب احتمال (شراء جهاز الكمبيوتر = نعم) لكل السمات المطلوبة معًا للسجل الذي يتم استكشافه كما يلي:

ح [(أ)/(شراء كمبيوتر = نعم)] = ح (العمر = شاب صغير، شراء الكمبيوتر = نعم)

× ح (الدخل = متوسط، شراء الكمبيوتر = نعم)
× ح (طالب = نعم، شراء الكمبيوتر = نعم)
× ح (التصنيف الائتماني= متوسط، شراء الكمبيوتر= نعم)

= ٠,٢٢٢ × ٠,٤٤٤ × ٠,٦٦٧ × ٠,٦٦٧

= ٠,٠٤٤

مع ملاحظة أنه تم ضرب جميع الاحتمالات للحصول على الاحتمال الإجمالي حيث أن جميع السمات منفصلة عن بعضها ولا تعتمد على بعضها البعض Independent، وذلك بحسب نظرية الاحتمالات.

وبنفس الطريقة يمكن حساب احتمال عدم شراء جهاز الكمبيوتر، والمعبر عنه باحتمال (شراء الكمبيوتر = لا)، للسجل المستكشف بنفس السمات، بحيث يكون:

ح [(أ) / (شراء كمبيوتر = لا)] = ٠,٦٠٠ × ٠,٤٠٠ × ٠,٢٠٠ × ٠,٤٠٠ = ٠,٠١٩

ويكون:

ح [(أ) / (شراء كمبيوتر = نعم)] × ح (شراء كمبيوتر = نعم) = ٠,٠٤٤ × ٠,٦٤٣ = ٠,٠٢٨

ويكون:

ح [(أ) /(شراء كمبيوتر = لا)] × ح (شراء كمبيوتر = لا) = ٠,٠١٩ × ٠,٣٥٧ = ٠,٠٠٧

وحيث أنه لا توجد إلاّ فئتين فقط (نعم، لا)، وأن احتمال الانتماء

للفئة (نعم) أكبر من احتمال الانتماء للفئة (لا) فتكون النتيجة أن نموذج التصنيف باستخدام النظرية الافتراضية يتوقع النتيجة (نعم) للسجل المستكشف.

أي أن الشخص (أ) بالسمات التالية: (العمر= شاب صغير، الدخل= متوسط، طالب= نعم، التصنيف الائتماني= متوسط)، يُتوقع له أن يقوم بشراء جهاز الكمبيوتر.

التصنيف باستخدام النظرية الافتراضية الشبكية

Network Bayes Classification

استعرضنا سابقًا استخدام النظرية الافتراضية البسيطة، وكانت فيها السمات في السجلات لا تعتمد على بعضها البعض، أما في النظرية الافتراضية الشبكية فإن السمات يمكن أن تكون مرتبطة ببعضها وتتأثر فيما بينها، أو بمعنى آخر فإن أي سجل من سجلات قواعد البيانات قد يتسم بسمة أو أكثر من تلك السمات في نفس الوقت، أو أن يتسم بها بنسب احتمالية معينة.

فإذا علمنا الفئة التي ينتمي لها سجل ما فإن سماته بالنظرية الافتراضية البسيطة كانت مشروطة بأن تكون منفصلة ومستقلة عن بعضها البعض ولا تخضع للاحتمالات، أما في النظرية الافتراضية الشبكية فإنه يمكن أن تكون المتغيرات غير منفصلة واحتمالية.

فالمتغيرات مثل (العمر، الدخل، الوظيفة، ... إلخ) تعتبر منفصلة وغير احتمالية، ويستحيل أن يحتمل أن يأخذ السجل الواحد قيمتين للسمة الواحدة، فالعمر مثلاً أما أن يكون (٣٥) أو (٣٦)، وهكذا لبقية السمات.

أما المتغيرات التي يمكن أن ترتبط ببعضها البعض هي كتلك التي تبين نسبة احتمال أو نسبة الإصابة بمرض مُعين عند معرفة سمات تتعلق بالوراثة،

ففي قاعدة بيانات أحد المراكز الطبية قد يتم تخصيص سمات معينة لتحديد نتائج اختبارات طبية أو سمات خاصة بالأعراض المرضية أو توقعات الإصابة بمرض معين والتي يتم التعبير عنها بالنسب الاحتمالية، ويؤدي ذلك لاستخدام نظريات الاحتمالات في كافة حسابات الاحتمالات الإجمالية لمجموع تلك السمات بشكل أكثر تعقيدًا من الأسلوب المتبع في النظرية الافتراضية البسيطة.

التصنيف باستخدام استقراء القاعدة
Classification Using Rule Induction

يتم في هذ النوع من التصنيف استخدام أسلوب بناء القاعدة، حيث يتم في خطوة التعلم تمثيل نموذج التصنيف باستخدام سلسلة من القواعد التي يُعبر عنها بصيغة (إذا كان .. فإن..).

والقواعد التي يتم استخدامها في التصنيف تُعتبر وسيلة جيدة لتمثيل البيانات والمعلومات وتبسيطها وفهمها، ونماذج التصنيف المبنية على القاعدة تستخدم مجموعة من القواعد على شكل (إذا كان .. فإن..) من أجل التصنيف، بحيث تكون على الشكل التالي:

إذا كان (شرط) فإن (نتيجة)

مثلاً، يمكن صياغة القاعدة التالية ق ١ حيث أن:

ق١ : "إذا كان (عمر الزبون= شاب صغير)، و(طالب= نعم)، فإن (شراء كمبيوتر= نعم)"

والعبارة الشرطية، في الطرف الأيمن من القاعدة، يمكن أن تحتوي على اختبار سمة واحدة أو أكثر (مثل العمر وغيرها من السمات)، والتي يتم جمعها معًا بشكل منطقي.

أما النتيجة فهي تحتوي على التوقع أو توقع الفئة المستهدفة في التصنيف،

ففي هذا المثال كان التوقع هو معرفة ما إذا كان الزبون سيقوم بشراء جهاز كمبيوتر.

ويمكن صياغة هذه القاعدة على الشكل التالي:

ق١ : (العمر= شاب صغير) و(طالب= نعم) >> (شراء كمبيوتر = نعم)

إذا تحقق الشرط، أو الشروط، في الطرف الأيمن للقاعدة، أي إذا تحققت كل اختبارات السمات المذكورة وكانت صحيحة لأحد السجلات، فإنه يُقال أن هذه القاعدة محققة، وأنها تغطي هذا السجل.

ويمكن تقييم القاعدة من خلال مقاييس كل من نسبة التغطية ونسبة الصحة.

فإذا كان لدينا قاعدة (ق)، ولدينا عدد من السجلات وليكن (ل)، وعدد السجلات التي تغطيها القاعدة هو (ن)، وعدد السجلات التي تحقق نتيجة التصنيف المذكور في القاعدة هو (م)، فإنه يمكن تعريف نسبة التغطية ونسبة الصحة للقاعدة (ق) من خلال المعادلتين التاليتين:

- نسبة التغطية للقاعدة (ق) = (ن) / (ل)
- نسبة الصحة للقاعدة (ق) = (م) / (ن)

أي أن التغطية هي نسبة عدد السجلات التي تغطيها القاعدة إلى العدد الكلي للسجلات في قاعدة البيانات.

أما الصحة فهي نسبة عدد السجلات التي ينطبق عليها نتيجة القاعدة إلى عدد السجلات التي تغطيها.

مثال توضيحي

بالعودة إلى الجدول (٥-١)، والذي يحتوي على سجلات من قاعدة بيانات المبيعات لمحل الأجهزة الإلكترونية، والمهمة المطلوبة هي توقع ما إذا كان أحد الزبائن سوف يقوم بشراء جهاز كمبيوتر.

نفرض أن (ق١) هي قاعدة تغطي عدد (٢) من السجلات الموجودة في هذا الجدول، وعددها الإجمالي (١٤)، وأن السجلين يحققان نتيجة القاعدة، فيكون:

نسبة التغطية للقاعدة (ق١) = ٢ / ١٤

= ١٤,٢٨ %

نسبة الصحة للقاعدة (ق١) = ٢ / ٢

= ١٠٠ ٪

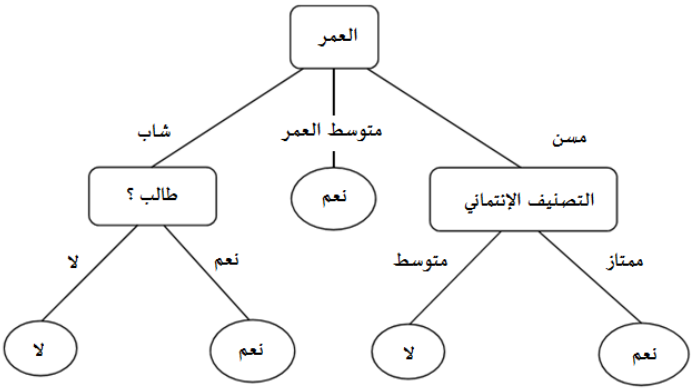
استخراج قواعد التصنيف من شجر القرار

تطرقنا سابقًا لطريقة بناء شجرة القرار من مجموعة من البيانات باعتبارها أحد أساليب التصنيف، وهي عملية شائعة من عمليات التصنيف ومن السهل فهم طريقة عملها، ولكن الشجرة التي يتم بنائها قد تكون صعبة وكبيرة

أحيانًا، ويتم اللجوء في هذه الحالات لاستخراج بعض قواعد التصنيف من شجرة القرار، من أجل تبسيطها. ومقارنة مع شجرة القرار فإن القواعد الشرطية ربما تكون أسهل للفهم عندما تكون شجرة القرار كبيرة جدًا.

ويمكن استخراج قواعد التصنيف من شجرة القرار من خلال توليد قاعدة واحدة فقط لكل مسار من المسارات التي تنبع من الجذر إلى كل ورقة من الأوراق، بحيث يكون كل تفرع من التفرعات الموجودة في المسار هو جزء من أجزاء الجملة الشرطية التي يتم تجميعها منطقيًا في الطرف الأيمن، وقيمة ورقة الشجرة التي تمثل قيمة الفئة المتوقعة تكون هي جواب الشرط أو نتيجة للقاعدة.

مثال توضيحي



شكل (٥,٥)، شجرة قرار يمكن تحويلها إلى مجموعة من قواعد التصنيف الشرطية

في الشكل أعلاه، لدينا شجرة قرار يمكن تحويلها إلى مجموعة من قواعد التصنيف الشرطية من نوع (إذا كان.. فإن..)، وذلك من خلال تقفي كل المسارات من نقطة الجذر إلى كل ورقة من أوراق الشجرة، ويمكن سرد كل القواعد المستخرجة من هذه الشجرة كما يلي:

ق١: إذا كان (العمر= شاب) و(طالب= لا) >>> فإن (شراء كمبيوتر=لا)

ق٢: إذا كان (العمر= شاب) و(طالب= نعم) >>> فإن (شراء كمبيوتر=نعم)

ق٣: إذا كان (العمر= متوسط العمر) >>> فإن (شراء كمبيوتر = نعم)

ق٤: إذا كان (العمر= مسن) و(التصنيف الائتماني= ممتاز) >>> فإن (شراء كمبيوتر= نعم)

ق٥: إذا كان (العمر= مسن) و(التصنيف الائتماني= متوسط) >>> فإن (شراء كمبيوتر= لا)

ويُلاحظ في هذا المثال أن السمات التي تظهر في الطرف الأيمن لكل القواعد هي سمات منفصلة أو متنافية منطقياً Independent، أي أنها لا تتقاطع، أي أنه لا توجد تعارضات بين القواعد التي يتم استخراجها، حيث أن كل سجل سوف يتم تغطيته من قاعدة واحدة فقط، وأن كل قاعدة تغطي

مجموعة من السجلات بشكل حصري دون أن تشترك مع غيرها من القواعد.

وفي هذا المثال، نظراً لوجود قاعدة واحدة لكل ورقة من أوراق الشجر فإنه يمكن القول بأن هذه العملية لا تُمثل تبسيطًا لشجرة القرار، حتى أنه أحيانًا ما يكون استخراج قواعد التصنيف من شجرة القرار يمثل زيادة في تعقيدها بدلاً من تبسيطها، وذلك مثلما يحدث عند استخراج قواعد التصنيف من شجرة القرار التي تُعاني من وجود التكرارات والاستنساخات، حيث تظهر مجموعة من القواعد الصعبة نظراً لتكرار بعض السمات.

كما يتم اللجوء في بعض الأحيان لتهذيب مجموعة القواعد التي يتم استخراجها من شجرة القرار، تمامًا مثلما يتم تهذيب شجرة القرار نفسها.

التصنيف باستخدام خوارزميات الشبكة العصبية
Neural Networks Algorithms

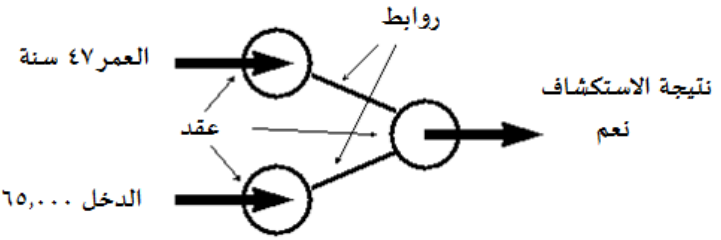
تعتبر الشبكات العصبية هي وأشجار القرار من أهم تقنيات التنقيب في البيانات، نظراً للنتائج الدقيقة التي يتم التوصل إليها باستخدام هذه الخوارزميات ولإمكانية تطبيقهما في حل العديد من المشاكل وبكافة الأنواع، هذا بالرغم من صعوبتهما والتي أدت لعدم الانتشار بشكل واسع لهما.

وخوارزمية الشبكة العصبية تشبه في تركيبتها تركيبة عقل الإنسان، فهي تعمل بنفس الطريقة كما يعمل العقل في نقل ومعالجة المعلومات والتوصل إلى الاستنتاجات واكتشاف الأنماط والتنبؤات، ونستطيع من خلالها تطبيق بعض ما يطبقه العقل الطبيعي، رغم أن العلماء لا يزالون حتى اليوم يكتشفون المزيد عنه ولم يلموا بكل تفاصيل عمله. ويتم بناء خوارزميات الشبكة العصبية بطرق معقدة جداً ومبنية على متغيرات رقمية، وعادة يتم استبدال قيم المتغيرات الاسمية بقيم رقمية لكي يتم استخدامها في الشبكة العصبية، وكثيراً ما تصب جهود المحللين باتجاه تبسيط وتسهيل خوارزميات من هذا النوع.

طريقة عمل خوارزمية الشبكة العصبية

إن خوارزميات الشبكات العصبية تعمل بنفس الطريقة التي يعمل بها

عقل الإنسان، فهناك العقد (التي تناظر الخلايا العصبية) والروابط التي تصل بينها (التي تناظر الوصلات العصبية)، والشكل التالي يوضح تركيب شبكة عصبية بسيطة:



شكل (٥,٦)، نموذج شبكة عصبية مبسطة تتنبأ بإقبال الزبون على شراء منتج معين

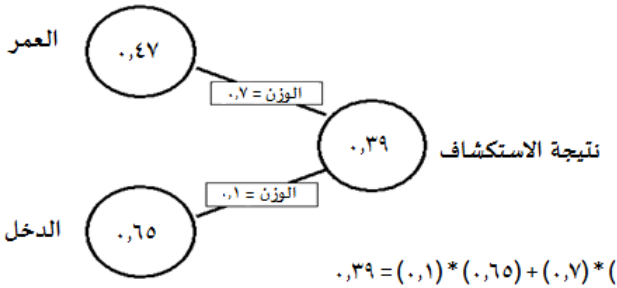
تمثل الدوائر المستديرة العقد، وتمثل الخطوط المستقيمة الروابط، وتعمل هذه الشبكة بأن يتم إدخال متغيرين بقيم معينة ويتم تطبيق دوال معينة بهدف حساب قيمة تنبؤية لمتغير ثالث، وتعتبر هذه القيمة قيمة تنبؤية باستخدام هذه الشبكة العصبية.

وفي هذا المثال، الشبكة العصبية تأخذ متغيري العمر والدخل وتعطي نتيجة تنبؤية هدفها استكشاف ما إذا كان الشخص سيُقبل على عرض معين لشراء أحد المنتجات.

ولكي يتم التنبؤ باستخدام الشبكة العصبية، يتم إدخال قيم المتغيرات المعلومة في العقد المخصصة للإدخال، ويصبح لكل عقدة قيمة المتغير الذي تم إدخاله، بعد ذلك يتم ضرب قيمة كل عقدة بقيمة الرابط المتصل بها

وتجميع كل النتائج بحسب المعادلة الرياضية المعرفة في الشبكة العصبية ومن ثم الحصول على نتيجة التنبؤ بوقوع الحدث الذي يتم استكشافه.

وفي هذا المثال، تم اعتبار أنه إذا كانت النتيجة (صفر) يكون من المتوقع استجابة الشخص للعرض المقدم له، وإذا كانت النتيجة مقدارها (١) فيكون من المتوقع عدم استجابته. ولتبسيط هذه العملية يمكن وصفها بالشكل التالي:



شكل (٥,٧)، شبكة عصبية تأخذ متغيري العمر والدخل للشخص وتتنبأ بإقباله على شراء منتج معين

وقد تم التعبير عن العمر والدخل بقيم تقع على التدريج بين (صفر، ١)، فالعمر في هذا المثال ٤٧ سنة وهو يناظر القيمة ٠,٤٧ على التدريج، والدخل في هذا المثال بقيمة ٦٥،٠٠٠ والمناظر للقيمة ٠,٦٥ على التدريج، أما الوصلات والتي تعبر عن الأوزان فقد تم تقديرها بالقيم (٠,٧) و(٠,١) على الترتيب، وبعد ضرب قيم العقد في قيم الوصلات وجمعها نحصل على

قيمة المتغير الذي نريد التوقع له فيكون الناتج هنا (٠,٣٩)، وهي قيمة أقرب للقيمة (صفر) منها للقيمة (١)، وبذلك يتم اعتبارها (صفر)، وتكون النتيجة هي توقع استجابة الشخص للعرض المقدم له لشراء جهاز الكمبيوتر.

طريقة بناء خوارزمية الشبكة العصبية

يتم بناء خوارزمية الشبكة العصبية عن طريق دراسة وتحليل أحداث وقعت بالفعل في سجلات سابقة في قاعدة البيانات، وبمقارنة التوقعات الناتجة من الشبكة العصبية مع السجلات الحقيقية السابقة يتم تطوير وتحديث قيم الأوزان في الشبكة بقيم أكثر دقة، وهذه يشبه إلى حد كبير عملية تصحيح أوراق إجابات الطلاب في المدرسة، فكلما كانت الأخطاء كبيرة فإنه يلزم لتصحيحها جهد أكبر، وكلما كانت الأخطاء قليلة يقل الجهد المطلوب لتصحيحها، وهكذا يتم بناء الشبكة العصبية بشكل تدريجي لتصبح أكثر صحة ودقة وكفاءة في تقدير التوقعات.

ويعتبر تحديد قيم الأوزان، المُعبر عنها بالوصلات، هي الأهم لبناء شبكة عصبية يتم باستخدامها التوقع بشكل جيد ودقيق.

وفي المثال السابق كانت الشبكة العصبية بسيطة جدًا وتهدف للتوضيح، وفي الواقع يمكن أن تكون خوارزمية الشبكة العصبية أكثر تعقيدًا وتحتوي على العديد من العُقد والوصلات التي قد تصل إلى عدة مئات.

العُقد المخفية في الشبكات العصبية

قد تحتوي خوارزمية الشبكة العصبية على نوعية أخرى من العقد والتي تسمى العقد المخفية، وتكون مهمة هذه العقد استشارية ولا يؤخذ بقيمها إلا بعد أن يتم اعتماد استشارتها في حالة صحتها وبعد التجربة الفعلية، فالعُقد المخفية هنا تلعب دورًا استشاريًا فقط، ومثلما يحدث في الجيش، فالقائد يستمع إلى العديد من الاستشارات ممن حوله من المستشارين قبل اتخاذ قرار معين، ولكنه بعد اتخاذ القرار واستكشاف نتائجه ومدى صحته، يصبح بإمكانه تمييز المستشارين الجيدين والذين كانت آرائهم أقرب للقرار الذي كان من المفترض أن يكون أنسب، وبالتالي سوف يعتمد آرائهم في المستقبل ويأخذ بها أكثر من آراء غيرهم، وهكذا، فالعُقد المخفية تلعب نفس هذا الدور، كلما تم تطبيق الخوارزمية يتم تطوير وتحديث العُقد الأصلية بأن تأخذ بالاعتبار قيم العُقد المخفية المناسبة والتي تدعم الحصول على نتائج أكثر دقة، وبالمقابل يتم إهمال قيم العُقد المخفية التي لم تحقق ذلك.

التصنيف باستخدام خوارزمية الجار الأقرب
Nearest Neighbor Algorithm

تُعتبر 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.

إن أحد الأمثلة الواقعية لخوارزمية الجار الأقرب هي أنه لو نظرت إلى جيرانك الذين يسكنون حولك تلاحظ أن لديهم دخل متشابه، لذا فإنه إذا كان جارك لديه دخل (١٠٠٠٠) سنويًا فإن احتمال أن يكون دخلك أنت أيضًا (١٠٠٠٠) سنويًا هو احتمال كبير، وربما يكون الاحتمال كبير جدًا أن دخلك بهذا الشكل إذا ما كان كل جيرانك كذلك، في حين يكون الاحتمال أقل بكثير عندما يكون دخلهم لا يتعدى (٥٠٠٠) فقط. فبهذه الطريقة أمكننا تخمين مقدار دخل شخص ما بمجرد النظر لجيرانه القريبين منه.

إن خوارزمية الجار الأقرب تُطبق بنفس الكيفية التي تم إظهارها بالمثال السابق، باستثناء أن طبيعة الجوار تختلف من مثال لآخر، ففي المثال المذكور كان الجوار المقصود هو الموقع الجغرافي. وفي قواعد البيانات يوجد العديد من العوامل التي يمكن أخذها بالاعتبار عند بحث الجوار الذي يمكن تطبيقه خلافًا لمكان السكن، فمثلاً قد يكون من المهم معرفة الجامعة التي درس فيها الشخص وما هي الشهادة التي حصل عليها عندما نريد التنبؤ بالدخل له،

عندئذٍ ستختلف رؤيتنا للجوار في خوارزمية الجار الأقرب ليصبح الجوار هنا متعلق باسم الجامعة والشهادة بدلاً من مكان السكن.

وكثيرًا ما تُستخدم خوارزمية الجار الأقرب في مجال الأعمال، ومن الأمثلة الشائعة الاستخدام تلك التي تساعد المستخدمين في تحديد احتياجاتهم عن طريق اختيار السلع الأقرب لتفضيلاتهم مقارنة بسلع قد تم شراؤها بالفعل وتحديدًا في مواقع شراء الكتب على الإنترنت، فقد يتم ترميز أحد الكتب بأنه كتاب جيد ومن ثم مقارنته بكتب قد تم شراؤها بالفعل من قِبل عدد كبير من المستخدمين وتكون قريبة منه من حيث المضمون فيتم ترشيحه لهم باعتبار أن احتمال شراؤه يكون مرتفعًا وفق طبيعة الجوار من حيث المضمون مع الكتب التي يفضلونها.

وهناك استخدامات أخرى كثيرة لهذه الخوارزمية في مجالات متعددة، وقد تكون العملية أكثر تعقيدًا كما هو الحال في مشكلات تحديد أسعار البضائع وفق السلاسل الزمنية السابقة، وتكمن المشكلة هنا في أنه لا يوجد شيء محدد نبحث له عن جوار إنما توجد سلسلة من المعلومات السابقة للأسعار والمطلوب هو تحديد السعر الجديد. والطريقة المثلى لحل مثل هذه المشاكل باستخدام خوارزمية الجار الأقرب هي أن يتم تحديد عدد من السجلات التجريبية، مثلًا عدد (١٠) سجلات، ثم استخدام القيم التسعة الأولى منها للتنبؤ بالقيمة العاشرة الأخيرة، وإذا كان لدينا (١٠٠) سجل يمكن تقسيمها لنحصل على (١٠) مجموعات وعشرة قيم يتم التنبؤ بها بنفس الطريقة.

ومن العمليات التي يمكن أن تساعد في تطوير وزيادة فاعلية خوارزمية الجار الأقرب هي الأخذ بعين الاعتبار عدد أكبر من الجوار في محيط السجل المراد استكشافه، ففي أحد الأمثلة قد يكون أقرب جار للسجل المراد استكشافه هو سجل بقيمة معينة، وكان هناك عدد آخر من السجلات المحيطة والمجاورة له ولكن بقيم مختلفة، عندئذٍ فإنه يُفضل تقدير قيمة السجل المطلوب بناءً على جميع السجلات وليس على أقرب جار للسجل فقط.

الثقة في نتائج الاستكشاف

إن الثقة في نتائج ما يتم التوصل له من استكشاف باستخدام خوارزمية الجار الأقرب لها أهمية أيضًا، ويُعبر عن ذلك بأن نقول أننا على ثقة بنسبة ٧٠٪ من قيمة معينة قمنا باستكشافها. وتتم عملية تقدير نسبة الثقة بعدة طرق، منها أن يؤخذ بعين الاعتبار المسافة بين السجل المستكشف وبين أقرب جار، فتلك المسافة تلعب دورًا مهما في تقدير نسبة الثقة، فإذا كانت المسافة قريبة جدًا فذلك يعني أننا نثق أكثر في النتيجة بعكس ما إذا كانت مسافة كبيرة.

ومن الطرق الأخرى التي يمكن الاعتماد عليها في تقدير الثقة هي بحث التجانس الذي يميز مجموعة الجوار حول السجل المراد استكشافه، فإذا كانت المجموعة كلها مشتركة وتؤدي لاستكشاف قيمة واحدة للسجل المستكشف فهذا يرفع نسبة الثقة في النتيجة، بعكس ما إذا كانت المجموعة منقسمة لقسمين ويؤديان لاستكشاف قيمتين مختلفتين للسجل المستكشف.

خوارزميات التصنيف متعددة الفئات
Multi-Class Classification Algorithms

معظم خوارزميات التصنيف التي تم التطرق إليها كانت مهمتها التنبؤ بانتماء أحد السجلات في قواعد البيانات إلى فئة واحدة من فئتين فقط، مثلاً في نموذج التصنيف الذي يتنبأ بإقبال الزبون على شراء منتج معين يتم التعبير عن ذلك بأحد النتيجتين (نعم أو لا)، وكذلك عندما يتم التنبؤ بأهلية الزبون الذي يتقدم بطلب للحصول على قرض من أحد البنوك، حيث يتم التعبير عن نتيجة التنبؤ بأحد الخيارين بأن يكون القرض الممنوح له (آمن أو خطر).

وهكذا فإن معظم نماذج التصنيف كانت تحدد التوقعات المحصورة بفئتين.

وفي خوارزميات التصنيف متعددة الفئات يمكن أن يؤدي نموذج التصنيف للتنبؤ بقيمة من مجموعة غير محددة من القيم، أو أن يتم التنبؤ بأن سجل ما من سجلات قواعد البيانات ينتمي لفئة من ضمن فئات عديدة غير محددة.

ومن أمثلة هذا النوع من خوارزميات التصنيف هو ما يتعلق بخوارزميات تصنيف المستندات Documents والبحوث العلمية، أو خوارزميات التعرف على الأصوات.

مثلاً، لو افترضنا أننا نرغب في بناء نموذج تصنيف آلي تكون مهمته تصنيف المستندات كالمقالات أو صفحات الإنترنت أو البحوث العلمية، وبشكل أكثر تحديدًا نريد من هذا النموذج أن يفرّق بين المستندات التي تتحدث عن رياضة كرة القدم أو رياضة التنس أو أي رياضة أخرى. فالمشكلة هنا أنه سيكون لدينا كم هائل من المستندات، كما أنها بلا أوصاف واضحة ومحددة يتم القياس وفقًا لها.

فمن الطرق المستخدمة في بناء نماذج تصنيف متعددة الفئات هي طريقة "واحد مقابل الكل" One-versus-all (OVA)، وبفرض وجود عدد (ف) من الفئات يتم استخدام أسلوب التصنيف الثنائي لكل فئة والممثل بالقيمتين (٠، ١)، بحيث يتم التعبير عن استكشاف انتماء السجل (أ) للفئة (ف١) بأن تعطي قيمة إيجابية (١) للفئة (ف١) مقابل قيم سلبية (٠) لبقية الفئات الأخرى. والعكس في حالة استكشاف عدم انتماء هذا السجل لتلك الفئة، وهكذا لكل سجل من السجلات، ويتم تجميع كل القيم بحيث تكون الفئة التي تحصل على عدد أكبر من القيم الموجبة هي الفئة التي ينتمي لها السجل المستكشف.

تقييم كفاءة واختيار خوارزميات التصنيف

Evaluating and Selecting the Classification Algorithms

بعد بناء نموذج تصنيف معين يمكن أن نبدأ بالتساؤل مثلاً عن مدى صحة التوقع أو التنبؤ الذي يقوم به نموذج التصنيف، مثلاً لو افترضنا أننا قمنا ببناء خوارزمية تصنيف تتنبأ بالسلوك الشرائي لأحد الزبائن، ولكننا نريد أن نعرف مدى صحة هذا التنبؤ الذي توفره تلك الخوارزمية، أي مدى صحة التنبؤ بإقبال أحد الزبائن على شراء منتج معين. أو نفرض أننا قمنا ببناء أكثر من نموذج تصنيف للتنبؤ بسلوك هذا الزبون، فكيف يمكننا أن نعرف أي من هذه النماذج هو أكثر صحة وكفاءة من غيره من النماذج ويمكن الاعتماد عليه في التنبؤ الصحيح عند مقارنتهم معًا من حيث مقياس الصحة والمقاييس الأخرى.

إن الإجابة على هذا السؤال يتمثل في توضيح الطرق المختلفة لتقدير مدى كفاءة خوارزميات التصنيف من خلال حساب مقياس الصحة ومقاييس أخرى مهمة تهدف لتقدير كفاءة الخوارزميات من حيث مدى صحتها وجودتها عند استخدامها من أجل التنبؤ.

مثال توضيحي

في قاعدة بيانات مبيعات الأجهزة الإلكترونية، نفرض أنه تم بناء نموذج تصنيف مخصص يهدف للتنبؤ بمدى إقبال زبون ما على شراء جهاز كمبيوتر،

وكانت مهمة نموذج التصنيف هي التنبؤ بتصنيف كل سجل ليكون منتمي لأحد الفئتين التاليتين:

فئة السجلات الموجبة Positives (م): وهي السجلات التي فيها (شراء كمبيوتر = نعم)

فئة السجلات السالبة Negatives (ن): وهي السجلات التي فيها (شراء كمبيوتر = لا)

وذلك بفرض أن الفئة الأساسية هنا هي فئة السجلات الموجبة، وهي الفئة التي تحتوي على كل السجلات التي يكون فيها التنبؤ بالتصنيف أمر إيجابي، وهذا الأمر هو عملية شراء الكمبيوتر، أما الفئة السلبية فهي التي تحتوي على كل السجلات الأخرى.

ونفرض أنه تم تطبيق هذه الخوارزمية على كل السجلات الموجودة، والتي نعرف بالفعل الفئة التي ينتمون لها، وذلك من أجل التنبؤ بتلك الفئة، بحيث يتم معرفة عدد السجلات التي يتم التنبؤ بفئتها بشكل صحيح True Positives وليكن هذا العدد يساوي (ت م)، وهو عدد السجلات الموجبة التي تم التنبؤ لها بشكل صحيح باستخدام نموذج التصنيف.

وبالمثل يكون عدد السجلات السالبة الصحيحة True Negatives وليكن (ت ن) هو عدد السجلات السالبة التي تم التنبؤ لها بشكل صحيح باستخدام نموذج التصنيف.

وعدد السجلات الموجبة الخاطئة False Positives (خ م) هو عدد السجلات السالبة التي تنبأ لها نموذج التصنيف بشكل خاطئ وألحقها بالفئة الموجبة، مثلاً التنبؤ بالإقبال على شراء الكمبيوتر لأحد السجلات الذي يبين أن الزبون لم يشترِ الكمبيوتر.

وعدد السجلات السالبة الخاطئة False Negatives (خ ن) هو عدد السجلات الموجبة التي تم اعتبارها بالخطأ سالبة مع أنها موجبة. أي أنها سجلات تنتمي لفئة (شراء الكمبيوتر = نعم) وقام نموذج التصنيف بالتنبؤ لها بأنها تنتمي للفئة (شراء كمبيوتر= لا).

ويمكن تلخيص كل هذه القيم في الجدول التالي:

		الفئة التي تم توقعها: نعم	الفئة التي تم توقعها: لا	المجموع
الفئة الصحيحة	نعم	ت م	خ ن	م
	لا	خ م	ت ن	ن
	المجموع	م "	ن"	م + ن

جدول تلخيص للقيم المتنبأ بها باستخدام نموذج تصنيف معين

حيث تخبرنا القيم (ت م) و (ت ن) متى كان نموذج التصنيف يعمل بشكل جيد من حيث التنبؤ بسلوك الزبون، سواء من حيث الإقبال أو عدم

الإقبال على الشراء. بينما تخبرنا القيم (خ م) و (خ ن) متى كان نموذج التصنيف يعمل بشكل خاطئ.

ويلاحظ من هذا النموذج بأنه يقوم بالتنبؤ بفئتين فقط وهي (شراء كمبيوتر = نعم) و(شراء كمبيوتر= لا)، ويمكن توسيعه ليغطي معطيات نموذج تصنيف يتنبأ بعدة فئات تصنيفية بنفس الطريقة.

نسبة صحة خوارزمية التصنيف Accuracy

من خلال معرفة جميع أعداد السجلات الواردة في الجدول السابق يمكن حساب نسبة صحة خوارزمية أو نموذج التصنيف وذلك باستخدام المعادلة التالية:

نسبة الصحة = [(ت م) + (ت ن)] / [(م) + (ن)]

أي أن نسبة صحة خوارزمية التصنيف Accuracy هي نسبة عدد السجلات التي تم تصنيفها بشكل صحيح إلى إجمالي عدد السجلات، وتسمى أحيانًا بنسبة إدراك الخوارزمية.

كما يمكن الحديث عن نسبة الخطأ Error Rate في نموذج التصنيف بنفس الطريقة، بحيث يكون:

نسبة الخطأ = [(خ م) + (خ ن)] / [(م) + (ن)]

أي أنها نسبة عدد السجلات التي تم التنبؤ بها بشكل خاطئ باستخدام

نموذج التصنيف إلى إجمالي عدد السجلات.

كما يلاحظ أن نسبة الصحة = ١ – نسبة الخطأ

مشكلة اختلال التوازن في خوارزميات التصنيف

تظهر هذه المشكلة عندما تكون الفئة المهمة التي يتم استكشافها هي الفئة النادرة، أي أن توزيع البيانات يعكس الأغلبية للفئة السالبة، والأقلية للفئة الموجبة، ومن أمثلة ذلك ما نجده في البيانات الطبية، كأن تكون الفئة النادرة مثل (الإصابة بمرض السرطان) في نموذج تصنيف تكون مهمته تحليل سجلات البيانات الطبية للمرضى، ويتم التنبؤ بأن أحد السجلات الخاص بأحد المرضى يحتمل إصابته بمرض السرطان وذلك بالإجابة بالقيمة (نعم أو لا)، وفي هذه الحالة إذا كانت مثلاً نسبة الصحة هي ٩٧٪ فإنها قد تجعل نموذج التصنيف يبدو وكأنه يتمتع بنسبة صحة عالية، ولكن ماذا إذا لم يكن في قاعدة البيانات غير ٣٪ فقط من السجلات المصابة بالفعل بمرض السرطان، فمن الواضح في هذه الحالة بأن نسبة الصحة ٩٧٪ قد لا تكون مقبولة بشكل كافٍ أو مُقنع، حيث أنه من الممكن أن يكون هذا النموذج، مثلاً، قادرًا على التصنيف الصحيح للسجلات التي لا تحتوي على الإصابة بمرض السرطان وبنفس الوقت يكون ضعيفًا أو غير قادر على تصيف السجلات التي تحتوي على الإصابة بالمرض فعلًا.

ومن هذه المنطلق تظهر لنا الحاجة لأنواع أخرى من مقاييس الكفاءة

التي تُبين لنا مدى جودة نموذج التصنيف وقدرته على إدراك القيم الموجبة (السرطان = نعم) ومدى قدرته على إدراك السجلات ذات القيم السالبة (السرطان = لا)، وهذه المقاييس الجديدة يتم التعبير عنها بمصطلحات كل من الحساسية والنوعيةSensitivity and Specificity، وهي مقاييس يتم استخدامها من أجل هذا الغرض على الترتيب.

قياس الحساسية والنوعية لخوارزميات التصنيف:

- الحساسية Sensitivity = (ت م) / (م)
- النوعية Specificity = (ت ن) / (ن)

كما أنه يمكن استنتاج أن نسبة الصحة يمكن التعبير عنها بدلالة الحساسية والنوعية من خلال المعادلة التالية:

نسبة الصحة = الحساسية × [(م) / (م) + (ن)] + النوعية × [(ن) / (م) + (ن)]

مثال تطبيقي

الجدول التالي يوضح ملخص تجميعي لنتائج تصنيف سجلات بيانات طبية، وقيم فئات التصنيف في النموذج المستخدم هي (نعم و لا) والتي تعبر عن المتغير أو السمة (الإصابة بمرض السرطان):

الفئات	نعم	لا	المجموع	التقدير %
نعم	٩٠	٢١٠	٣٠٠	٣٠,٠٠
لا	١٤٠	٩٥٦٠	٩٧٠٠	٩٨,٥٦
المجموع	٢٣٠	٩٧٧٠	١٠٠٠٠	

ومن هذا الجدول يمكن حساب القيم التالية:

الحساسية = ٩٠ / ٣٠٠ = ٣٠ %

النوعية = ٩٥٦٠ / ٩٧٠٠ = ٩٨,٥٦ %

ونسبة الصحة للنموذج = ٩٦٥٠ / ١٠٠٠٠ = ٩٦,٥ %

وبالتالي فإنه في هذا النموذج، وبالرغم من أنه يمتلك نسبة صحة عالية إلاّ أن قدرته على التنبؤ بالقيم الموجبة، والتي تناظر هنا الفئة النادرة، هي قدرة ضعيفة، وهو ما بيّنه قياس الحساسية بقيمة (٣٠٪)، ومع ذلك فإن هذا النموذج يمتلك قيمة نوعية مرتفعة، وهو ما يعني أنه يمكنه التنبؤ بكفاءة عالية بالقيم السلبية، وبمعنى آخر فإنه يمكن الوثوق بشكل أكبر في قدرة هذا النموذج على التنبؤ بعدم الإصابة بمرض السرطان من الثقة في قدرته على التنبؤ بالإصابة به.

قياس الدقة Precision

إن مقياس الدقة يُستخدم كثيرًا في نماذج التصنيف، والدقة هي نسبة

عدد السجلات الموجبة التي يتم التنبؤ لها وتصنيفها بشكل صحيح إلى إجمالي عدد السجلات الموجبة التي يتم التنبؤ لها بشكل صحيح أو خاطئ.

الدقة = (ت م) / [(ت م) + (خ م)]

قياس المثالية Recall

وهي نسبة السجلات الموجبة التي تم التنبؤ لها وتصنيفها بشكل صحيح إلى كل السجلات الموجبة، أي أنها نفس قيمة الحساسية.

المثالية = (ت م) / (م)

ففي المثال السابق يكون لدينا:

الدقة = ٩٠ / ٢٣٠ = ٣٩,١٣ %

المثالية = ٩٠ / ٣٠٠ = ٣٠ %

كما توجد بعض المقاييس الأخرى التي يمكن أن تُستخدم في تقييم خوارزميات التصنيف بطرق تقديرية من منظور كل مستخدم لها وبحسب البرمجيات المستخدمة في إنشاءها، وهي:

السرعة Speed

وهي تقدير لكمية وسرعة العمليات الحاسوبية التي اُستخدمت لتوليد وتطبيق خوارزمية التصنيف.

المتانة Robustness

وهي قدرة نموذج التصنيف على عمل التنبؤات الصحيحة عندما يكون هناك بيانات مزعجة أو بيانات مفقودة في قاعدة البيانات، ويتم تقييم هذا المقياس باستخدام معاملات تعبر عن درجة الإزعاج المتوفرة في البيانات وكمية البيانات المفقودة.

قابلية التوسع Scalability

وهي قدرة نموذج التصنيف على التعامل مع كميات أكبر من البيانات بنفس الكفاءة.

قابلية التفسير Interpretability

وهي مدى الفهم للتركيب الداخلي لبنية نموذج التصنيف أو التنبؤ، وهو مقياس موضوعي وليس مقياس عددي، ففي شجرة القرار مثلاً كلما كانت الشجرة أكثر تعقيدًا يكون مقياس قابلية التفسير لها بقيمة أقل.

الفصل السادس
التحليل والتنقيب باستخدام خوارزميات التجزئة العنقودية
Cluster Analysis

المحتويات

١. مفاهيم أساسية
٢. التجزئة العنقودية بالتقسيم Partition Clustering
٣. التجزئة العنقودية الهرمية Hierarchical Clustering
   أ- التجزئة الهرمية التشققية
   ب- التجزئة الهرمية التجميعية
٤. التجزئة العنقودية الاحتمالية Probabilistic Cluster Analysis
٥. التجزئة العنقودية عالية الأبعاد High Dimensional Clustering
٦. التجزئة العنقودية للأشكال البيانية والبيانات الشبكية Clustering Graph and Network Data
٧. التجزئة العنقودية المشروطة Clustering With Constraints
٨. تقييم التجزئة العنقودية Cluster Analysis Evaluation

مفاهيم أساسية

لو تخيلنا أن مدير عام علاقات الزبائن في أحد الشركات لديه خمسة مدراء يعملون تحت إمرته، وأنه يرغب في تنظيم كفاءة نظام العلاقات مع زبائن الشركة من خلال تجميعهم في خمسة مجموعات بحيث يتم إدارة كل مجموعة من المجموعات من قِبل أحد المدراء.

إستراتيجيًا، فإن مدير عام العلاقات يريد أن تحتوي كل مجموعة من المجموعات على الزبائن الذين يتشابهون بقدر الإمكان، بل والأكثر من ذلك فإنه لأي زبونين لديهما خصائص أو أنماط مختلفة فإنه ينبغي أن لا يتواجدا في نفس المجموعة.

والهدف الأساسي من هذه الإستراتيجية هو تطوير نظام العلاقات مع الزبائن وتطوير الحملات الترويجية التي تستهدف بشكل خاص كل مجموعة من الزبائن استنادًا إلى الملامح والسمات المشتركة لهم والتي جعلتهم ينتمون لنفس المجموعة.

فما هي تقنية التنقيب التي يمكن أن تساعد في إتمام هذه المهمة ؟

بخلاف تقنيات التصنيف، فالمجموعة أو الفئة التي يمكن أن ينتمي لها الزبون هنا غير معروفة، كما أنه بوجود عدد كبير من الزبائن وتعدد السمات التي تصفهم فإنه يصبح من الصعب والمكلف، أو ربما من المستحيل، تجزئتهم

يدوياً لمجموعات تخدم هذا الهدف، وتظهر الحاجة هنا لاستخدام تقنية التجزئة العنقودية Clustering.

إن التجزئة العنقودية هي عملية تجميع مجموعة من البيانات بداخل عدد من الفئات أو الأجزاء، بحيث تضم كل مجموعة أو فئة العناصر الأكثر تشابهًا ولكنهم أكثر اختلافًا عن العناصر التي تنتمي للمجموعات الأخرى.

كما يتم تقييم التشابه وعدم التشابه بناء على قيم السمات التي تصف البيانات، والتي غالبًا ما تشتمل على قياس المسافات فيما بينها، وقد تم التطرق لبحث تشابه واختلاف البيانات في فصل سابق من هذا الكتاب.

إن التجزئة العنقودية كأحد أدوات تنقيب البيانات لها تطبيقات في مجالات عديدة، في الطب والعلوم البيولوجية والمعلوماتية الحيوية، واستخبارات الأعمال وبحوث الشبكة العنكبوتية والمجالات الأمنية وغيرها من المجالات.

يستعرض هذا الفصل المفاهيم الأساسية وطرق وإجراءات التحليل بالتجزئة العنقودية وتقنياتها المختلفة.

أساليب التحليل بالتجزئة العنقودية
Cluster Analysis Procedures

إن التحليل بالتجزئة العنقودية هي عملية تجزئة مجموعة من البيانات إلى مجموعات جزئية، وكل مجموعة جزئية تمثل كتلة Cluster من البيانات، بحيث يكون عناصر كل مجموعة متشابهة وبنفس الوقت مختلفة عن بقية العناصر في المجموعات الأخرى.

وضمن هذا السياق، فإنه يمكن لطريقتين من طرق التجزئة أن ينتج عنهما مجموعات جزئية مختلفة لنفس البيانات، وتتم عملية التجزئة باستخدام خوارزميات تقود إلى توصيف المجموعات الجزئية التي لم تكن معروفة قبل التجزئة.

وكأحد أدوات تنقيب البيانات، فإن التجزئة العنقودية يمكن أن تكون أداة تنقيب قائمة بذاتها وتهدف لاكتساب المعرفة المخبأة بداخل قواعد البيانات وملاحظة خصائص كل مجموعة جزئية والتركيز على مجموعات محددة لمزيد من التحليل، ولكنها أيضًا يمكن أن تكون خطوة من خطوات تحضير البيانات للتحليل والتنقيب لكي تخدم غيرها من خوارزميات التنقيب.

ونظراً لازدياد استخدامات تقنيات التجزئة العنقودية في العديد من المجالات وبخاصة تلك التي تتعامل مع كميات هائلة من البيانات فقد أصبح هذا المسار من المسارات المهمة في بحوث التنقيب.

التجزئة العنقودية بطريقة التقسيم

Partition Clustering

تتلخص طريقة التقسيم في أنها عملية تجميع السجلات المتشابهة بقاعدة البيانات في مجموعات، ويتم ذلك بهدف الاستكشاف عالي المستوى لما يجري داخل قاعدة البيانات. ففي مجال الأعمال عادة ما يُستخدم التحليل العنقودي بطريقة التقسيم في تجزئة الزبائن، أو السكان بشكل عام، إلى مجموعات يمكن التسويق لها بشكل مباشر ومحدد. ولبناء هذه العملية التجميعية يتم استخدام المعلومات الأولية كالدخل السنوي، العمر، المهنة، وأية معلومات متوفرة في قواعد البيانات، ثم يتم تسمية كل مجموعة باسم بارز ومعبر عنها بشكل واضح. المثال التالي يوضح عملية من هذا النوع:

الرقم	الاسم	العمر	الدخل	مستوى التعليم	المهنة	المنطقة
١	نبيل	٤٥	متوسط	دبلوم	نجار	أ
٢	جهاد	٣٣	متوسط	دبلوم	موظف	ب
٣	منذر	٤٤	مرتفع	جامعي	تاجر	ب
٤	باسل	٣٨	متوسط	ثانوية عامة	سائق	أ
٥	محمود	٤٧	مرتفع	جامعي	عمل حر	ب
٦	إياد	٣٠	منخفض	ثانوية عامة	سائق	أ
٧	شوقي	٣٥	متوسط	جامعي	موظف	ب
٨	مراد	٣٦	مرتفع	دبلوم	تاجر	ب

٩	ياسر	٤٦	منخفض	دبلوم	حداد	أ
١٠	خالد	٤٢	منخفض	ثانوية عامة	عامل	أ

إذا كانت هذه بيانات بعض الزبائن في إحدى الشركات التجارية، فربما أردنا تجزئتهم إلى مجموعات بحيث يتوفر في كل مجموعة التوافق والانسجام البيني، فمثلاً يمكن تجزئة قاعدة البيانات إلى ٣ مجموعات وفق الدخل، لنحصل على المجموعات التالية:

الرقم	الاسم	العمر	الدخل	مستوى التعليم	المهنة	المنطقة

(١)

٣	منذر	٤٤	مرتفع	جامعي	تاجر	ب
٥	محمود	٤٧	مرتفع	جامعي	أعمال حرة	ب
٨	مراد	٣٦	مرتفع	دبلوم	تاجر	ب

(٢)

١	نبيل	٤٥	متوسط	دبلوم	نجار	أ
٢	جهاد	٣٣	متوسط	دبلوم	موظف	ب
٤	باسل	٣٨	متوسط	ثانوية عامة	سائق	أ
٧	شوقي	٣٥	متوسط	جامعي	موظف	ب

(٣)

٦	إياد	٣٠	منخفض	ثانوية عامة	سائق	أ
٩	ياسر	٤٦	منخفض	دبلوم	حداد	أ
١٠	خالد	٤٢	منخفض	ثانوية عامة	عامل	أ

من جهة أخرى، ربما قد نلجأ لإعادة التجزئة وفق اعتبارات أخرى كالمستوى التعليمي لنخلق مجموعات متوافقة بينيًا من منظور آخر، وفي هذا المثال سوف تتم التجزئة على حسب المستوى التعليمي كما هو مبين في الجدول التالي:

الرقم	الاسم	العمر	الدخل	مستوى التعليم	المهنة	المنطقة
(١)						
٣	منذر	٤٤	مرتفع	جامعي	تاجر	ب
٥	محمود	٤٧	مرتفع	جامعي	أعمال حرة	ب
٧	شوقي	٣٥	متوسط	جامعي	موظف	ب
(٢)						
١	نبيل	٤٥	متوسط	دبلوم	نجار	أ
٢	جهاد	٣٣	متوسط	دبلوم	موظف	ب
٨	مراد	٣٦	مرتفع	دبلوم	تاجر	ب
٩	ياسر	٤٦	منخفض	دبلوم	حداد	أ
(٣)						
٤	باسل	٣٨	متوسط	ثانوية عامة	سائق	أ
٦	إياد	٣٠	منخفض	ثانوية عامة	سائق	أ
١٠	خالد	٤٢	منخفض	ثانوية عامة	عامل	أ

ثم تستخدم هذه المعلومات التجميعية لترميز عناصر كل مجموعة بالطريقة المناسبة وفق قاعدة بياناتهم، ومن ثم استكشاف ردود الفعل للعروض

التسويقية التي يمكن تقديمها لكل منهم بحسب الخصائص والسمات التي تم استخدامها في التجزئة، مثلاً كأن يتم التسويق المستهدف لطلاب الجامعات من أجل حثهم على شراء منتج معين يهتمون به، وهكذا لبقية المجموعات، ومن ثم تقييم ودراسة سلوك واستجابة عناصر كل مجموعة وأخذ النتائج بالاعتبار.

لا توجد هناك أفضل طريقة للتجزئة بالتقسيم، فهناك دوما طرق متعددة وحسب الاحتياج، فالمثال السابق برغم بساطته إلا أنه وضح لنا أنه هناك دوما تجزئة تلبي الاحتياج ووفق رؤية معينة يتم تبنيها، فالتجزئة في كل الأحوال تقسم قاعدة البيانات وتبسطها وتسهل رؤيتها من منظور معين، وليس فقط بهدف استعراض البيانات أو تلخيصها وإنما بهدف استكشاف بعض الروابط أيضا والتوفيق بينها كما في المثال السابق.

التجزئة العنقودية الهرمية
Hierarchical Clustering

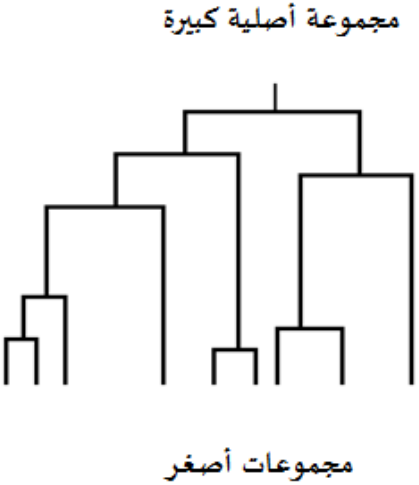
في بعض الأحيان قد نحتاج إلى تجزئة البيانات إلى فئات في مستويات مختلفة بشكل هرمي، وهو ما يُطلق عليه التجزئة العنقودية الهرمية، بحيث يتم تجميع البيانات في مجموعات بشكل هرمي أو شجري.

ويُعتبر تمثيل البيانات بشكل هرمي مفيد لأهداف تلخيص البيانات وتصويرها، مثلاً يمكن لأحد المدراء في أحد الشركات أن يقوم بتنظيم الموظفين لديه في فئات أساسية تحتوي المدراء التنفيذيين ومدراء الأقسام والموظفين على الترتيب، كما يمكن أن يتم تقسيم هذه الفئات إلى فئات جزئية أصغر منها، وهكذا، لتأخذ شكل التقسيم الهرمي، بحيث يتم استخدامها فيما بعد في تلخيص البيانات أو توصيفها، مثلاً في تجزئة من هذا النوع يمكن إيجاد متوسط رواتب مدراء الأقسام بسهولة.

ويُلاحظ في التجزئة الهرمية أنه بالرغم من أن تجزئة البيانات تمت بشكل هرمي إلاّ أن ذلك لا يعني أن البيانات لها هيكلية هرمية، أي أن المدراء والموظفين كانوا في نفس المستوى، فالتجزئة الهرمية هنا فقط من أجل تلخيص البيانات وإظهار خصائصها بشكل مختصر أو مضغوط، كنوع من أنواع تصوير البيانات.

ومع ذلك فإنه في بعض التطبيقات قد تعبر التجزئة الهرمية عن البناء

الهرمي لهيكلية البيانات، مثلاً، يمكن تقسيم مجموعة الحيوانات إلى مجموعات جزئية هرمية من حيث طبيعتها الفقارية واللاّفقارية، ومن ثم إظهار تفرعات كل نوع بشكل هرمي.



شكل (٦,١)، التجزئة العنقودية الهرمية

ويوجد نوعين من التجزئة الهرمية:

أ. التجزئة الهرمية التكتلية

ويتم في هذه التقنية بناء الأجزاء بشكل تصاعدي تكتلي بدءًا من القاعدة، فيكون لدينا مجموعات أولية تحتوي كل منها على عنصر واحد فقط ممثلاً بسجل من سجلات قاعدة البيانات، ثم يتم دمج كل مجموعتين قريبتين من

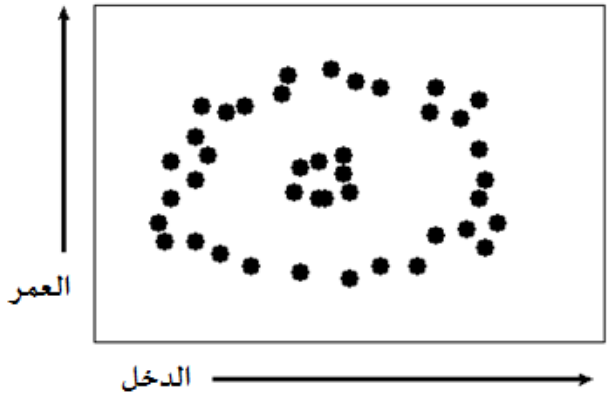
بعضهما في مجموعة جديدة أكبر منهما، ويستمر هذا الدمج بشكل هرمي تصاعدي حتى الوصول إلى رأس الهرم والممثل بمجموعة تحتوي كافة المجموعات الصغرى.

ب. التجزئة الهرمية التشققية

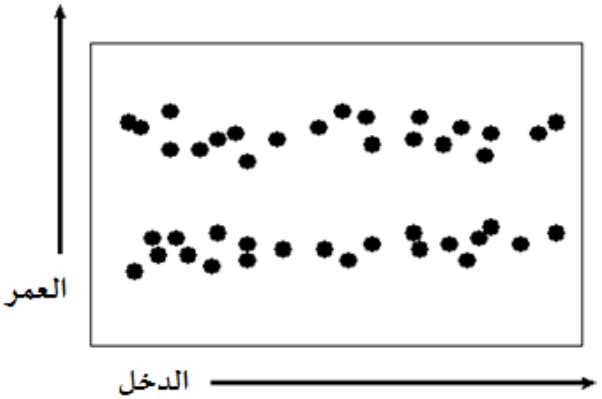
ويتم في هذه الطريقة عكس ما تم بالطريقة السابقة، حيث تبدأ هذه التقنية بوضع كافة سجلات قاعدة البيانات في مجموعة واحدة ومن ثم تجزئتها إلى أجزاء صغيرة تدريجيا أصغر فأصغر وحتى الحصول على أدنى مستوى من المجموعات والتي تحتوي كل منها على عنصر واحد ممثلا بسجل من قاعدة البيانات.

وبشكل عام، فإن عملية التجزئة تكون بهدف وضع العناصر المتجانسة بقدر الإمكان في مجموعات منفصلة، والقاعدة العامة لضم أي عنصر في مجموعة هي أن يكون العنصر مائلاً للتشابه بعنصر منها أكثر من أن يكون شبيهًا لعنصر من مجموعة أخرى.

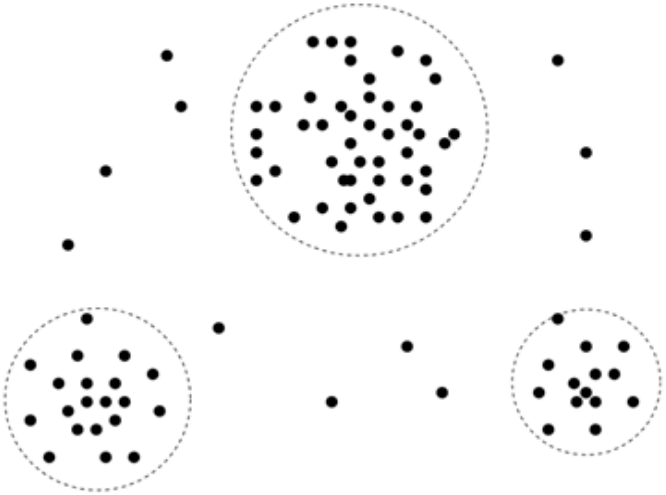
الأشكال التالية تبين كيف يمكن أن تقع العناصر المتجانسة معًا في مجموعة واحدة وبعدة طرق مختلفة تم استكشافها باستخدام خوارزميات التجزئة العنقودية:



شكل (٦,٢)، تجميع العناصر المتجانسة بالتجزئة العنقودية



شكل (٦,٣)، تجميع العناصر المتجانسة بالتجزئة العنقودية



شكل (٦,٤)، تجميع العناصر المتجانسة بالتجزئة العنقودية

شكل (٦,٥)، تجميع العناصر المتجانسة بالتجزئة العنقودية

التجزئة العنقودية الاحتمالية

Probabilistic Cluster Analysis

في أنواع التجزئة العنقودية التي تم التطرق إليها حتى الآن، كل عنصر من عناصر البيانات كان ينتمي لفئة جزئية واحدة من الفئات التي يتم تكوينها باستخدام خوارزميات التجزئة العنقودية، مثلاً كأن ينتمي أحد الزبائن لمجموعة يتولى إدارتها أحد مدراء التسويق، وفي تطبيقات أخرى قد يكون هذا الوضوح في توزيع العناصر على المجموعات الجزئية غير مرغوب، وقد يتطلب الأمر بعض المرونة أو الغموض في بعض التطبيقات.

مثال توضيحي

لو افترضنا أن أحد محلات الأجهزة الإلكترونية لديها معرض للبيع على الإنترنت، حيث يستطيع الزبون إجراء عملية الشراء عبر الإنترنت، كما يمكنه إضافة رأيه وتقييمه للمنتج بعد الشراء.

وقد تلقى بعض المنتجات تقييمات لها من قِبل بعض الزبائن، في حين أنه قد لا تلقَ منتجات أخرى أي تقييم، أو تتلقى عدد قليل منها، وربما يتعلق التقييم الواحد بعدد من المنتجات، وفي هذه الحالة إذا كانت المهمة دراسة كل التقييمات الواردة على الإنترنت فإنه من الصعب تحديد المجموعة التي ينتمي لها كل تقييم، حيث أنها تتعلق بالمنتجات المختلفة وبشكل متداخل.
مثلاً، إذا كان لدينا فئة جزئية مخصصة لمنتج (كاميرات الفيديو الرقمية)،

وفئة جزئية أخرى مخصصة لمنتج (أجهزة الكمبيوتر)، فما العمل إذا تعلق أحد التقييمات بالتوافق بين الكاميرا الرقمية وجهاز الكمبيوتر، حيث أن هذا التقييم سوف يتعلق بالمجموعتين، وأنه لا ينتمي لأحدها دون الأخرى.

ففي هذه الحالة يمكن استخدام طريقة التجزئة العنقودية التي تسمح بأن ينتمي التقييم لأكثر من فئة جزئية واحدة إذا كان يتعلق بأكثر من موضوع.

ويمكن استخدام الأوزان لكي تحدد مدى انتماء العنصر للفئة الجزئية مقارنة بانتمائه للفئات الجزئية الأخرى.

مثلاً، نفرض أن شركة الأجهزة الإلكترونية تسعى لدراسة سلوك زبائنها في التصفح وشراء المنتجات عبر الإنترنت، وذلك من خلال معرفة هل الزبون يبحث عن منتجات معينة أو يقارن بين المنتجات المختلفة أو يسأل عن الدعم الفني؟

التجزئة العنقودية تساعد في هذه الحالة حيث من الصعب أن يتم التحديد مسبقًا للأنماط المختلفة لسلوك المستخدم، بحيث تُعبر كل فئة جزئية من الفئات التي يتم تكوينها عن سلوك مميز للتصفح وينتمي لها فعل المستخدم الذي يسلك هذا السلوك، وبالتالي تصبح هناك احتمالية أن ينتمي فعل من الأفعال جزئيًا إلى عدة فئات، مثلاً كأن يقوم أحد الزبائن بشراء كاميرا رقمية ثم يقوم بمقارنة أنواع مختلفة من أجهزة الكمبيوتر في نفس جلسة التصفح، فإن السلوك في هذه الجلسة سوف ينتمي بطبيعة الحال للفئتين اللتين تعبران عن كل من شراء كاميرا رقمية ومقارنة أجهزة الكمبيوتر.

التجزئة العنقودية عالية الأبعاد

High Dimensional Clustering

في جميع خوارزميات التجزئة العنقودية غالبًا ما يكون عدد السمات والخصائص التي تتم التجزئة باستخدامها قليل نسبيًا، وقد لا يتجاوز عشرة سمات في كثير من الأحيان.

ففي التجزئة العنقودية لبيانات زبائن أحد محلات الأجهزة الإلكترونية قد يتم استخدام بعض السمات الخاصة بالزبون، مثل (العمر، الدخل، حجم المشتريات، الوظيفة، ... إلخ)، وقد لا تتعدى جميعها (١٠) سمات.

أما إذا كانت السمات كثيرة جدًا فإنه يطلق على التجزئة العنقودية التي تستند على عدد كبير من السمات تجزئة عنقودية عالية أو عديدة الأبعاد.

ومن أمثلة هذا النوع من التجزئة العنقودية هو ما يمكن أن يقوم به مدير المبيعات في محلات الأجهزة الإلكترونية عندما يرغب بتجزئة الزبائن إلى مجموعات بحسب ما قاموا بشرائه من منتجات من المعرض.

مثال توضيحي

الجدول التالي يوضح توزيع للزبائن في صفوف مختلفة وتقاطعها مع مشتريات كل زبون من المنتجات المختلفة الموضحة في الأعمدة:

الزبون	المنتج ١	المنتج ٢	المنتج ٣	المنتج ٤	المنتج ٥	المنتج ٦	المنتج ٧	المنتج ٨	المنتج ٩	المنتج ١٠
محمد	١	٠	٠	٠	٠	٠	٠	٠	٠	٠
أحمد	٠	٠	٠	٠	٠	٠	٠	٠	٠	١
محمود	١	٠	٠	٠	١	٠	٠	٠	٠	١

ويتضح من هذا المثال أن السمات سوف تكون كثيرة، وذلك نظرًا لوجود الآلاف من المنتجات، وهو ما يعني الآلاف من الأبعاد، الأمر الذي يترتب عليه الحاجة إلى إنشاء عدد مماثل من المجموعات الجزئية التي تُشكل عناقيد ضخمة وفوضوية.

ويتم اللجوء في هذه الحالة لأحد طريقتين من أجل حل هذه المشكلة.

الطريقة الأولى: هي التجزئة العنقودية لعدد محدود من السمات

الطريقة الثانية: تخفيض عدد الأبعاد للحصول على عدد مناسب نستطيع تطبيق التجزئة العنقودية عليه، وغالبًا ما يتم في هذه الطريقة إنشاء أبعاد جديدة بتركيب عدة أبعاد معًا من الأبعاد الأصلية.

مثال تطبيقي

لو قامت إحدى شركات الأجهزة الإلكترونية بجمع بيانات تقييم الزبائن

للمنتجات لتحليلها من أجل التوصية بالمنتجات المناسبة لهم، فإنه يمكنها أن تضع كل التقييمات في مصفوفة حيث يمثل كل صف فيها زبون من الزبائن، وكل عمود فيها يمثل أحد المنتجات، وكل قيمة (عنصر) بداخل المصفوفة يمثل تقييم الزبون للمنتج، والتي يمكن أن تكون إحدى القيم التالية: (جيد، متوسط، سيء)، أو أن تكون سلوك معين، مثل: (شراء، عدم شراء)، كما في الجدول التالي:

	المنتجات			
الزبائن	ص ١١	ص ٢١	....	ص ١ن
	ص ١٢	ص ٢٢	....	ص ٢ن
	.....	.....	.....	.....
	ص م ١	ص م ٢	....	ص م ن

حيث العنصر (ص ١١) هو رأي الزبون رقم(١) في المنتج رقم(١)، أو أنه يمثل عملية شراء الزبون رقم (١) للمنتج رقم (١)، والعنصر (ص ٢١) هو رأي الزبون رقم (١) في المنتج رقم (٢)، وهكذا، ويمكن تحليل هذه المصفوفة باتجاهين، اتجاه الزبائن واتجاه المنتجات، حيث يتم تكوين مجموعات الزبائن الذين يتشابهون في تفضيلاتهم أو نمط شرائهم عندما يتم التعامل مع المنتجات باعتبارها سمات، والعكس صحيح إذا قامت الشركة باعتبار الزبائن

هي السمات، فإنها تستطيع تنقيب واستكشاف مجموعات المنتجات التي تتشابه من حيث اهتمامات الزبائن وتقييمهاتهم.

والأكثر من ذلك أن الشركة تستطيع تكوين مجموعات جزئية من واقع سمات الزبائن والمنتجات معًا بحيث تحتوي كل منها على مجموعة من الزبائن، وتتضمن مجموعة من المنتجات في نفس الوقت.

مثلاً، يمكن تكوين مجموعة من الزبائن الذين يتشابهون في تفضيلهم لمجموعة من المنتجات، وهذه المجموعة سوف تكون مجموعة جزئية من مصفوفة (الزبائن - المنتجات) الكبيرة.

ويمكن استخدام هذه المجموعة الجزئية في تقديم التوصيات في اتجاهين:

الاتجاه الأول: هو أن تقوم الشركة بالتوصية بالمنتجات للزبائن الجدد الذين يشبهون الزبائن في المجموعة الجزئية، بحيث تكون من نفس أنواع المنتجات التي يفضلونها.

الاتجاه الثاني: هو أن توصي الشركة بمنتجات جديدة تشبه المنتجات الموجودة في المجموعة الجزئية، لهم ولغيرهم من الزبائن الجدد الذين يشبهونهم في السمات.

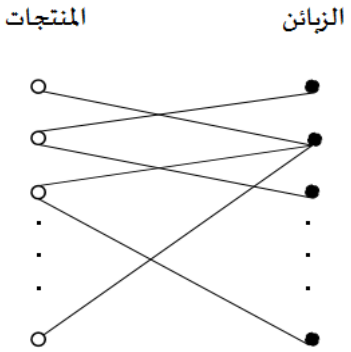
خوارزميات التجزئة العنقودية للأشكال البيانية والبيانات الشبكية
Clustering Graph and Network Data

إن التحليل بالتجزئة العنقودية للأشكال البيانية يمكن أن يساعد في استخراج المعرفة والمعلومات القابلة للقياس منها، وقد انتشر هذا النوع من التنقيب في العديد من المجالات.

في نفس المثال السابق المتعلق بسلوك الزبائن في شراء منتجات الأجهزة الإلكترونية يمكن تمثيل البيانات باستخدام شكل بياني ينقسم إلى قسمين:

- القسم الأول: يمثل الزبائن
- القسم الثاني: يمثل المنتجات

وكل نقطة في القسم الأول تمثل أحد الزبائن، وكل نقطة في القسم الثاني تمثل منتج معين.



شكل (٦,٦)، شكل بياني يمثل حركة شراء الزبائن للمنتجات

وفي الشكل، يصل خط بين أحد الزبائن وأحد المنتجات إذا كان الزبون قد قام بشراء المنتج.

وباستخدام التجزئة العنقودية للشكل البياني الناتج يمكن مثلاً تكوين فئة الزبائن الذين قاموا بشراء مجموعة من المنتجات المتشابهة، وهو ما يساعد في التوصية بشراء المنتجات للزبائن الذين ينتمون لتلك المجموعة، كأن تقوم الشركة مثلاً بالتوصية لأحد الزبائن بشراء كاميرا رقمية قام بشرائها معظم الزبائن في تلك المجموعة، حيث يكون من المتوقع أن يُقبل على شرائها في حال لم يقم بذلك بعد، وذلك من منطلق التشابه القائم فيما بينهم جميعًا.

وعلى العكس من ذلك، فإنه يمكن تكوين مجموعة من المنتجات التي قاموا بشرائها مجموعة متشابهة من الزبائن، ويمكن التوصية بمنتج معين في حالات معينة، مثلاً إذا كانت الكاميرا الرقمية وكارت الذاكرة يشكلون معًا مجموعة جزئية من هذا النوع، أي يتكرر شراؤهما معًا من مجموعة متشابهة من الزبائن، فعندما يقوم أحد الزبائن بشراء الكاميرا الرقمية فقط فإنه يمكن للشركة أن تقوم بالتوصية له بشراء كارت الذاكرة.

التجزئة العنقودية المشروطة
Clustering With Constraints

قد نلجأ أحيانًا لاستخدام المعرفة المسبقة التي تساهم في توصيف التجزئة العنقودية المطلوبة، بحيث تُشكل هذه المعرفة قيودًا على التجزئة المزمع القيام بها.

ويمكن تقسيم أنواع هذه القيود من حيث مكان تطبيقها إلى ما يلي:

١. قيود على العناصر بداخل المجموعات

وتنطبق هذه القيود على العناصر بداخل كل مجموعة وكيفية تجميعها معًا، سواء من حيث مدى التشابه فيما بينها أو الحد الأقصى للاختلاف، وذلك من خلال مقارنة وحساب السمات المختلفة لكل عنصر مع العناصر الأخرى.

٢. قيود على عملية التجزئة نفسها

وهي قيود يمكن وضعها على المجموعات التي يتم تكوينها وتبين متطلبات معينة أو إمكان استخدام سمات محددة، أو حتى الحد الأدنى لعدد عناصرها أو شكلها أو إجمالي عدد المجموعات الذي ينتج عن التجزئة العنقودية.

٣. قيود على مقاييس التشابه

يمكن تطبيق بعض القيود على المقاييس المستخدمة في التشابه بين العناصر

في نفس المجموعة، وطريقة حسابها.

مثال توضيحي

عند تجزئة زبائن محل الأجهزة الإلكترونية يمكن اشتراط أن يكون كل الزبائن المتشابهين في العنوان أو منطقة السكن في نفس المجموعة.

وإذا كانت الشركة بصدد توزيع الأدوار على (١٠) موظفين لمتابعة العلاقات مع الزبائن فإنه يمكن تجزئة الزبائن إلى (١٠) مجموعات جزئية بشرط أن تحتوي كل مجموعة على (١٠٪) من إجمالي عدد الزبائن، وذلك بهدف ضمان التوزيع العادل لحجم العمل على كل من الموظفين العشرة.

تقييم التجزئة العنقودية
Cluster Analysis Evaluation

لا توجد تجزئة عنقودية نموذجية لكل أنواع وفئات البيانات، ولكن يمكن أن تكون هناك طريقة نموذجية لكل نوع أو فئة من البيانات المحددة، وبشكل عام فإن هذه الطريقة تقع في منتصف المسافة بين اعتبار فئة البيانات كلها مجموعة واحدة يتم الحصول عليها بالتجزئة العنقودية، وبين تجزئتها إلى مجموعات تحتوي كل منها على عنصر واحد فقط.

عدد المجموعات التي يتم إنشاءها

من أشهر طرق التجزئة العنقودية التي يمكن اعتبارها مناسبة لعدد كبير من فئات وأنواع البيانات وينتج عنها تجزئة جيدة هي تلك التي تقسم فئة البيانات التي تحتوي على عدد (ن) من العناصر إلى عدد من المجموعات مقداره = $\sqrt{(ن / ٢)}$، وهو ما يؤدي لتكوين مجموعات جزئية تحتوي كل منها على عدد من العناصر = ٢ن .

كما يمكن تقييم كفاءة التجزئة العنقودية من خلال تقدير مدى التجانس بين المجموعات التي يتم إنشاءها.

تجانس المجموعات التي يتم إنشاءها

عند تجزئة قاعدة البيانات إلى مجموعات بطريقة التجزئة العنقودية، فإن

أهم خصائص المجموعات الناتجة هي تجانس عناصرها، وسواء كانت التجزئة مبنية على متغير واحد أو عدة متغيرات فإنه من المفيد أن يكون كل عناصر المجموعة متجانسة بالنسبة لقيمة المتغير الذي تم استخدامه، ويعتمد عدد المجموعات الناتجة على مقدار التجانس الموجود بين عناصر كل مجموعة. إن أفضل مثال للتجزئة بهدف الحصول على المجموعات الأكثر تجانسًا على الإطلاق هو تجزئة قاعدة البيانات إلى مجموعات يكون فيها عدد عناصر كل مجموعة هو عنصر واحد فقط، وبطبيعة الحال لن تؤدي هذه التجزئة إلى أية نتائج إضافية تُذكر، بل سنحصل على نفس البيانات الأصلية مجزأة إلى العناصر المكونة لها.

إذن فالتجزئة الجيدة هي تجزئة تقع بين حالتين وهما تجزئة بأكبر عدد من المجموعات وتجزئة بأقل عدد من المجموعات وبحسب الحالة والمتغير الذي يُستخدم في التجزئة.

الفصل السابع
تحليل وتنقيب القيم المتطرفة وأنواع البيانات المعقدة

Analyzing and Mining Outliers and Complex Data Types

المحتويات

١. مفاهيم أساسية

٢. أنواع القيم المتطرفة Outliers Types

٣. طرق استكشاف القيم المتطرفة Outliers Mining Procedures

٤. تحليل وتنقيب البيانات المعقدة Mining Complex Data Types

مفاهيم أساسية

إن القيم المتطرفة يمكن أن تتواجد وسط مجموعة من البيانات بحيث أنها لا تتوافق أو لا تشبه في سلوكها السلوك العام لبقية البيانات، وغالبًا ما تكون مختلفة عنها بشكل واضح وملفت للانتباه عن بقية البيانات الموجودة معها.

ويمكن أن تظهر القيم المتطرفة عن طريق الخطأ، كأن يتم التعبير عن عمر أحد الأشخاص بالقيمة (٩٩٩) نتيجة خطأ في إدخال البيانات أو نتيجة استخدام القيم الافتراضية في برنامج الإدخال عن طريق الخطأ، وخلاف ذلك فقد تكون قيم متطرفة بالفعل مقارنة بمثيلاتها من القيم الأخرى. مثلاً، يمكن أن يكون راتب المدير العام في أحد الشركات يمثل قيمة متطرفة وسط بقية القيم الخاصة برواتب بقية الموظفين.

وكثيرًا من خوارزميات تنقيب البيانات تحاول التقليل من التأثير المزعج للقيم المتطرفة أو حتى إزالتها نهائيًا، وهو ما يؤدي أحيانًا إلى فقدان المعلومات المهمة المخبأة أو الخفية. وحيث أن الإزعاج بالنسبة لشخص ما قد يكون إشارة مهمة لشخص آخر، فإن القيم المتطرفة نفسها التي قد تكون مزعجة من منظور ما يمكن أن يكون لها دلالات شيّقة ومهمة في بعض الأحيان من منظور آخر.

وهناك تطبيقات عديدة لتنقيب البيانات المتطرفة، ومن أشهر تطبيقاتها تلك المستخدمة في كشف عمليات الاحتيال Fraud Detection، عندما يتم استخدامها في كشف الاستعمال غير الطبيعي وغير المألوف لبطاقات

الائتمان أو خدمات الاتصالات.

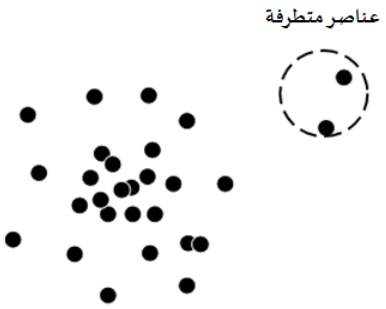
كما يمكن استخدامها في وضع خطط التسويق المستهدف والمخصص لتعريف السلوك الشرائي للزبائن من ذوي الدخل المنخفض جدًا أو المرتفع جدًا، باعتبارها قيم متطرفة.

كما يمكن استخدامها في المجالات الطبية وتحليل الحالات المرضية من أجل إيجاد الاستجابات غير الطبيعية للعلاجات الطبية المتنوعة.

ويمكن وصف عملية تنقيب القيم المتطرفة كما يلي:

إذا كان لدينا عدد من البيانات أو العناصر (ن) وكان لدينا عدد (م) من القيم المتطرفة المتوقع تواجدها وسط تلك العناصر، فيكون المطلوب هو إيجاد عناصر الفئة (م) والتي يمكن اعتبارها العناصر الأكثر تطرفًا ولا تتناسق مع بقية البيانات.

الشكل التالي يوضح وضعية القيم المتطرفة بالنسبة لمجموعة من البيانات:



شكل (٧,١)، توصيف القيم المتطرفة

أنواع القيم المتطرفة

١. القيم المتطرفة العامة (الاعتيادية)

وتكون فيها القيم متطرفة بطبيعتها، أي بدون الاقتران بسمات أخرى التي قد تجعلها مشروطة، مثلاً، راتب المدير العام في الشركة هو قيمة متطرفة من هذا النوع، لأنه ثابت لكل الأشهر وطوال العام ولا يقترن بسمات أخرى.

٢. القيم المتطرفة الخاصة (الفردية)

وهي القيم التي تكون متطرفة في أحوال معينة، مثلاً درجة الحرارة في مدينة ما يمكن أن تكون قيمة متطرفة من هذا النوع، فإذا كانت المدينة في منطقة باردة شتاءً وحارة صيفًا، فإن درجة الحرارة (١٢) درجة مئوية هي قيمة غير متطرفة في فصل الشتاء، أما إذا كانت في فصل الصيف فتكون عندئذٍ قيمة متطرفة فردية، حيث أنها اعتمدت على الموقع الجغرافي والتاريخ، أو أيةً معطيات أخرى.

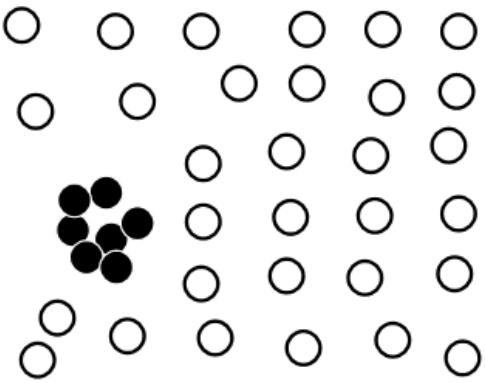
مثلاً، يمكن لنظام مراقبة حركة مشتريات بطاقات الائتمان في أحد البنوك، والذي يهدف لكشف عمليات الاحتيال، أن يقوم بالكشف عن حركة شراء غير اعتيادية، باعتبارها قيمة متطرفة، مقارنة بالسلوك الاعتيادي للزبون صاحب البطاقة، سواء من حيث القيمة الشرائية أو مكان تنفيذ عملية الشراء، ويحدث ذلك عند ملاحظة إجراء عملية شراء بمبلغ كبير نسبياً مقارنة

بمتوسط مبالغ الشراء الذي اعتاد عليه الزبون بحيث يتم اعتبارها قيمة متطرفة.

وإذا اعتبرنا أن مكان تنفيذ عملية الشراء معطيات يمكن الاستناد عليها فإن هذه العملية يمكن ألا يتم اعتبارها قيمة متطرفة طالما أنه تم تنفيذها من نفس المكان الذي يستخدمه صاحب البطاقة، وبالتالي فإنها تتحول إلى فرصة، باعتبار أن هذا الزبون ينتقل إلى شريحة أعلى من شرائح الزبائن بحسب حجم المشتريات باستخدام تلك البطاقات.

٣. القيم المتطرفة الجماعية

وهي حالة من حالات ظهور مجموعة من القيم المتطرفة معًا، ومن أمثلتها ما يمكن أن يحدث في أحد مطاعم الوجبات السريعة من تأخير في توصيل الطلبات للمنازل.



شكل (٧,٢)، القيم المتطرفة الجماعية

فالتأخير يمكن أن يحدث في أي وقت وفي أي يوم، ولا يمكن أن يتم احتسابه كقيمة متطرفة، فهو يحدث من حين إلى آخر، ولكن عندما يحدث تأخير في توصيل (١٠٠) طلب في يوم واحد فإن كل تلك الطلبات كمجموعة كاملة يشكلون معًا قيمة متطرفة تسمى قيمة متطرفة جماعية، بالرغم من أن كل واحد منها لا يمكن اعتباره قيمة متطرفة بشكل منفرد، وبالتالي يتطلب الأمر أن تقوم إدارة المطعم بحث هذه المشكلة في ذلك اليوم لمعرفة أسبابها ومنع تكرارها.

وبشكل عام، فإن أي مجموعة من البيانات يمكن أن يكون فيها أنواع مختلفة من القيم المتطرفة، كما أنه يمكن لعنصر ما أن ينتمي لأكثر من نوع من أنواع القيم المتطرفة.

ويمكن استخدام تطبيقات تنقيب القيم المتطرفة في العديد من المجالات ولأهداف متنوعة، فالقيم المتطرفة العامة هي أبسط الأنواع، والقيم المتطرفة الخاصة تستلزم مزيدًا من المعلومات لتحديد الخصائص والسمات التي تجعلها قيمًا متطرفة، أما القيم المتطرفة الجماعية فإنها تتطلب معلومات أكثر عن العلاقات البينية بداخل مجموعة معينة حتى يمكن الكشف عن إمكانية اعتبارها متطرفة بشكل جماعي.

٤. القيم المتطرفة في البيانات عالية الأبعاد

يُشكل استكشاف القيم المتطرفة في بيئة البيانات عالية الأبعاد تحديًا

كبيرًا، وذلك لأن اعتبار القيمة متطرفة في مثل هذه الحالات يكون وفقًا لمعايير أكثر تعددًا وتشعبًا بحكم تعدد الأبعاد التي يتم القياس بالنسبة لها، وبالتالي إمكانية معرفة العلاقة الحقيقية التي تربط بين العناصر المختلفة للبيانات فيها.

وفي حالة استكشاف القيم المتطرفة في بيئات البيانات عالية الأبعاد فإنه يلزم تحديد المزيد من المعلومات المرتبطة بتلك القيم حتى يتم توضيح طبيعة تطرفها بشكل محدد، مثلاً ينبغي تفسير كيفية اعتبار القيمة متطرفة، وتوضيح السياق الذي تم اعتبارها متطرفة وفقًا له، بالإضافة لتحديد الفئة الجزئية التي تم اعتبار هذه القيمة متطرفة بالنسبة لها، وأخيرًا يجب توضيح مدى قابلية التوسع في اعتبار هذه القيمة متطرفة بحسب وضعها بالنسبة لمجموعات جزئية أخرى من بيئة البيانات.

طرق استكشاف القيم المتطرفة

١. استكشاف القيم المتطرفة باستخدام خوارزميات التصنيف

ويتم ذلك من خلال تحديد بعض السمات ومقاييسها الطبيعية أو الاعتيادية مع تحديد معدلات الحد الأقصى لتلك القيم بحيث يمكن اعتبار القيم التي تتخطاها هي قيم متطرفة، وتكون مهمة خوارزمية التصنيف هي فحص كل عنصر من العناصر واكتشاف الفئة التي ينتمي إليها، بحيث تكون أحد الفئتين التاليتين (فئة العناصر الطبيعية أو غير المتطرفة، وفئة العناصر المتطرفة)، ومن ثم يتم استكشاف القيمة المتطرفة من خلال التنبؤ بالفئة التي تنتمي لها عند تطبيق خوارزمية التصنيف عليها.

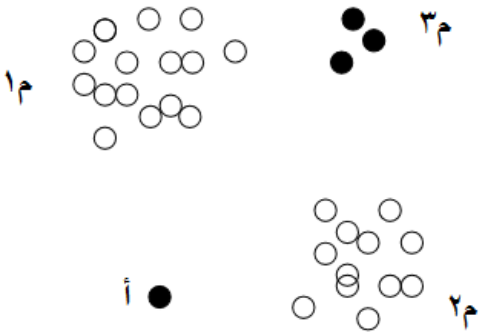
٢. استكشاف القيم المتطرفة باستخدام خوارزميات التجزئة العنقودية

وتتم هذه الطريقة على مرحلتين، المرحلة الأولى تكوين المجموعات الجزئية باستخدام التحليل بالتجزئة العنقودية، ومن ثم البحث عن العناصر التي لا تنتمي لأي مجموعة من تلك المجموعات فتكون هي القيم المتطرفة.

والفرق بين طريقة استكشاف القيم المتطرفة بالتصنيف وطريقة استكشافها بالتجزئة العنقودية أن الأولى تُستخدم عندما يكون لدينا معلومات نستند عليها من أجل تصنيف البيانات واكتشاف ما هو شاذ أو متطرف قياسًا مع تلك المعلومات، أما في الحالة الثانية فيتم استخدامها عندما لا يكون

لدينا معلومات أو خصائص وسمات نستند عليها في القياس.

في الواقع، إن الاختلاف في استخدام الطريقتين يوضح بالفعل الاختلاف القائم بين أساليب التحليل والتنقيب باستخدام خوارزميات التصنيف والتجزئة العنقودية بشكل عام.



شكل (٧,٣)، استكشاف القيم المتطرفة باستخدام التجزئة العنقودية

في الشكل أعلاه، العنصر (أ) قيمة متطرفة، لأنه لا ينتمي لأي من المجموعات الأخرى، كما يمكن اعتبار الفئة (م٣) بأكملها قيمة متطرفة أيضًا.

٣. استكشاف القيم المتطرفة باستخدام المقاييس الإحصائية الكلاسيكية

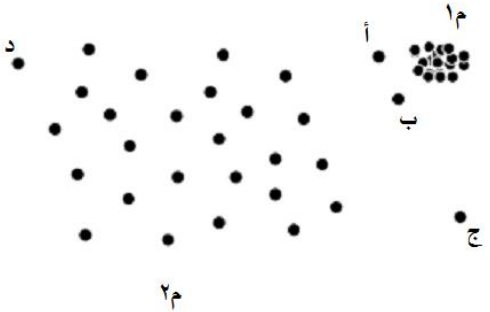
وهي الطريقة المتبعة في التحليل الإحصائي الكلاسيكي، وباستخدام المقاييس الإحصائية المتنوعة، كالمتوسط والوسيط والانحراف المعياري،

بحيث يتم الحكم على القيم التي تبتعد كثيرًا عن تلك المقاييس بأنها قيم متطرفة طالما أنها تبتعد كثيرًا عن مراكز تجمع البيانات.

٤. استكشاف القيم المتطرفة بقياس القرب من الجيران

يمكن استكشاف القيم المتطرفة من خلال قياس المسافة بينها وبين اقرب جيرانها، فإذا كانت هذه المسافة كبيرة جدًا مقارنة بالمسافة بين العناصر الأخرى وجيرانها فإنها تمثل قيمًا متطرفة في نفس الفئة من البيانات، وعادة ما يتم قياس المسافة إلى عدد محدد من الجيران، (٣) مثلاً.

كما تختلف النظرة للقيم المتطرفة بحسب موقع الفئة التي يتم القياس بالنسبة لها أو لعناصرها، مثلاً في الشكل التالي يظهر كيف تكون القيم (أ)، (ب) قيمًا متطرفة بالنسبة للفئة (م١) ولكنها قيم غير متطرفة بالنسبة للفئة (م٢). أما القيمة (ج) فهي قيمة متطرفة بالنسبة للفئتين (م١) و(م٢)، والقيمة (د) قيمة ليست متطرفة بالنسبة للفئة (م٢).



شكل (٧,٤)، استكشاف القيم المتطرفة بقياس القرب من الجيران

تحليل وتنقيب أنواع البيانات المعقدة
Mining Complex Data Types

تُستخدم تقنيات تحليل وتنقيب البيانات في مجالات عديدة تزداد اتساعًا يوميًا وتزداد معها التحديات في مواجهة أنواع متعددة من البيانات الأكثر تعقيدًا، كما تتنوع اتجاهات بحوث التنقيب وفقًا لتنوع وتعدد تلك المجالات.

وتشتمل أنواع البيانات المعقدة على السلاسل الزمنية والسلاسل الرمزية والسلاسل البيولوجية، بالإضافة للأشكال البيانية وشبكات الإنترنت، والبيانات المكانية والبيانات الزمكانية، وبيانات الأشياء المتحركة وبيانات النظم الإلكترونية، وبيانات الوسائط المتعددة، والنصوص وبيانات الإنترنت وتدفقات البيانات.

وسيتم إلقاء الضوء على كل نوع من هذه الأنواع في هذا القسم.

١. تحليل وتنقيب بيانات السلاسل Mining Sequence Data

(السلاسل الزمنية – السلاسل الرمزية – السلاسل البيولوجية)

إن السلسلة هي قائمة من الأحداث المرتبة، سواء كانت مرتبة زمنيًا أو من حيث الرمز أو الطبيعة البيولوجية.

السلاسل الزمنية تتكون من سلسلة طويلة من البيانات الرقمية التي يتم تسجيلها في فترات زمنية متساوية، كالدقيقة أو الساعة أو اليوم، ويمكن

استخدام السلاسل من هذا النوع في الكثير من المجالات، كالأعمال والعلوم والطب ودراسات البيئة والمناخ. مثلاً يمكن استخدامها في التنبؤ بتوقعات المبيعات المستقبلية وتحليل حركة أسعار الأسهم وتحليل الظواهر الطبيعية، وتحليل درجات الحرارة والرياح، ودراسات توقعات الزلازل والأعاصير.

وفي سلاسل البيانات الرمزية، فإنها تتكون من سلاسل طويلة من الأحداث أو البيانات الاسمية، والتي لا يمكن ملاحظتها في فترات زمنية متساوية، ففي كثير من الأحيان تكون الفترات بين الأحداث غير مهمة، ومن أمثلتها سلاسل رحلات تسوق الزبائن أو سلاسل التفاعل مع صفحات الإنترنت، وكذلك في سلاسل الأحداث المتعددة في مجالات العلوم والهندسة والطب والطبيعة والبيئة وتنمية المجتمع.

أما في السلاسل البيولوجية فهي تشتمل على سلاسل تحليل الـ دي أن إيه DNA وتحليل البروتين، ومثل هذه السلاسل طويلة جدًا وتحمل بين طياتها دلالات خفية، منطقية مهمة وشديدة التعقيد ، كما أن الفترات بين عناصر تلك السلاسل له أهمية أيضًا.

٢. تحليل وتنقيب الأشكال البيانية وبيانات الشبكات Mining Graph And Network Data

إن الأشكال البيانية تمثل بنية أكثر تعميمًا وتشعبًا في تركيبها من البيانات العادية، بحيث أنها يمكن أن تشتمل على العناصر والمجموعات والسلاسل أو

الشبكات أو الأشجار. وهناك تشكيلة واسعة من أنواع الأشكال البيانية التي تمثل بيانات الإنترنت وشبكات التواصل الاجتماعي والشبكات البيولوجية والمعلوماتية الكيميائية والمعلوماتية الحيوية والوسائط المتعددة والنصوص.

وقد ازدادت أهمية تنقيب الأشكال البيانية والشبكات وأصبحت تشكل مسارًا مهمًا من مسارات بحوث التنقيب نظرًا للتقدم التكنولوجي واتساع انتشار التكنولوجيا في هذا العصر.

وتنقيب أنماط الأشكال البيانية يهدف عادة لاستكشاف تكرارات الأشكال البيانية الجزئية في شكل بياني واحد أو عدة أشكال، وتنقيب أنماط الأشكال البيانية له عدة تطبيقات شيقة في العديد من المجالات.

مثلاً، يمكن استكشاف تكرارات الأشكال الجزئية المميزة التي تعبر عن بنية متشابهة من البيانات الخاصة بتركيب أنواع متعددة من العلاجات الطبية واستجابات المرضى لها.

٣. تحليل وتنقيب البيانات المكانية Mining Spatial Data

إن تحليل وتنقيب البيانات المكانية يهدف إلى استكشاف الأنماط والمعرفة وقواعد الارتباط المكانية وأنماط التشابه المكاني للبيانات المكانية، وهذه البيانات يمكن أن تكون على شكل متجهات أو نقاط أو صور نقطية أو الوسائط المتعددة الجغرافية.

ومن أمثلتها دراسات التطور التاريخي للمدن والشواطئ والأراضي والتضاريس وأنماط المناخ، والتنبؤ بالزلازل والأعاصير، واتجاهات الاحترار المستقبلية.

وقد تطور هذا المجال كمسار من مسارات بحوث التنقيب وزاد من أهميته وكفاءته بعد التطور الملحوظ في استخدام أجهزة الهاتف المحمول ونظام تحديد المواقع الجغرافية العالمي (جي بي إس) Global Positioning System، (GPS)، وخدمات خرائط الإنترنت وخدمات المناخ والأقمار الصناعية وتكنولوجيا الفيديو.

مثلاً، علماء الحيوان يستخدمون أدوات القياس عن بُعد والتصوير والمراقبة الحية للحيوانات البرية من أجل تحليل سلوكهم البيئي.

وفي أنظمة المرور، تم تزويد المركبات بأنظمة تحديد المواقع الجغرافية لكي تقوم بإرشاد السائقين للمواقع المختلفة وتقدير المسافات التي تفصلهم عنها.

وفي الأرصاد الجوية، يتم استخدام الأقمار الصناعية لملاحظة واستكشاف الأعاصير وتقديرات المناخ ودرجات الحرارة اليومية.

٤. تحليل وتنقيب بيانات نظم المراقبة Mining Observation System Data

تشتمل بيانات نظم المراقبة على عدد كبير من المعلومات والبيانات

الفيزيائية المركبة والمتداخلة، ومن أمثلتها نظم مراقبة المرضى في أقسام العناية المركزة في المستشفيات، والتي تربط نظام مراقبة المريض مرئيًا بالكاميرات وإلكترونيًا بأجهزة القياس المختلفة بشبكة الطوارئ المركزية.

ومن أمثلتها أيضًا أنظمة مراقبة شبكة المواصلات والطرق العامة التي ترتبط بأماكن التحكم والإشراف وإدارات المرور، والتي تتكون من مجموعة من أجهزة الاستشعار والرادارات وكاميرات الفيديو التي ترتبط بمركز المعلومات والتحكم بإدارات المرور.

ومعظم البيانات من هذا النوع ذات طبيعة متحركة ومتقلبة وغير متناسقة أو مترابطة، وهي معلومات مهمة جدًا لصناعة القرار اللحظي، فهي بحاجة دائمًا لوجود ارتباط وثيق بين ما يحدث في الموقف الحالي وقاعدة كبيرة من المعلومات والمعرفة تقوم بعمل الحسابات والتقديرات اللحظية وتعطي استجابة فورية، مثلاً عند حدوث أمر طارئ لأحد المرضى في غرفة العناية المركّزة يقوم نظام التنقيب بتقييم هذا الحدث ومعرفة مدى خطورته ومن ثم إرسال إشارة للأجهزة الإلكترونية التي تطلق الإنذار وتقوم بتنبيه المختصين في المستشفى للأمر الطارئ الذي حدث للمريض.

وأبحاث التنقيب في هذا المجال تشتمل على العديد من الأهداف، كاستكشاف الأحداث النادرة والشاذة وتقدير الدقة والفاعلية في سير العمل وتقدير كفاءة ومدى تكامل الأنظمة المستخدمة.

٥. تحليل وتنقيب بيانات الوسائط المتعددة Mining Multimedia Data

وهي عملية تهدف لاستكشاف الأنماط الشيّقة في قواعد بيانات الوسائط المتعددة التي تخزن وتدير تشكيلة واسعة من كائنات الوسائط المتعددة، والتي تشتمل على بيانات الصور والفيديو والصوت بالإضافة إلى سلاسل البيانات والنصوص والعلامات والعلاقات.

وتستهدف عمليات الاستكشاف بحث التشابه أو التطابق في الصور والخطوط والأصوات وغيرها من البحوث، ففي البحوث المتخصصة في شبكة الإنترنت تهدف عمليات التحليل والتنقيب لاستكشاف سلوك المستخدمين للشبكة ومعرفة تفضيلاتهم واهتماماتهم تجاه المحتويات المتشابهة في مواقع الإنترنت المختلفة، ويتم استخدام نتائج التحليل والتنقيب في وضع الإعلانات التي تناسب كل مستخدم بحسب تفضيلاته واهتماماته، كأن يتم توجيه الإعلانات الخاصة بالكاميرات الرقمية لهواة التصوير من المستخدمين.

٦. تحليل وتنقيب النصوص Text Mining

هي عمليات تحليل وتنقيب النصوص مثل الأخبار والمقالات والأوراق والبحوث العلمية والفنية والكتب ومحتويات المكتبات الإلكترونية والمدونات وصفحات الإنترنت.

ومن أهم أهدافها استخراج المعلومات المهمة وذات الجودة العالية من

كميات النصوص الكبيرة، وذلك من خلال استكشاف الأنماط والاتجاهات الفريدة، ومدى أصالة الأفكار الواردة في النص، كما يمكن أن تهدف إلى تصنيف المقالات بحسب المجالات المختلفة أو تلخيصها بملخصات وافية. ومن أمثلتها عمليات تحليل آراء الزبائن وتقييمهم للمنتجات على مواقع الشركات المُنتجة لها على الإنترنت.

ويكثر استخدام هذا النوع من التنقيب في المؤسسات الأكاديمية والمنتديات العلمية بهدف تقييم البحوث وتحكيمها وقياس مدى تميّزها مقارنة بالبحوث الأخرى الشبيهة بها، بالإضافة لتصنيفها وفهرستها بحسب الكلمات المفتاحية وتحليلها من حيث المحتوى.

٧. تحليل وتنقيب بيانات الإنترنت Mining www

إن شبكة الإنترنت العالمية ضخمة ومنتشرة على مستوى العالم، وهي مركز معلومات عالمي، يتضمن الأخبار والإعلانات والمعلومات العامة في شتى المجالات، كالصحة والتعليم والتجارة والصناعة والزراعة وسائر المجالات الأخرى، وهي تحتوي على تشكيلة ديناميكية وغنية من المعلومات وتُشكل مصدرًا خصبًا للبيانات القابلة للتحليل والتنقيب.

وتطبيقات تنقيب بيانات الإنترنت تهدف لاستكشاف الأنماط والتركيب الهيكلي والمعرفة بشكل عام من الإنترنت، ووفقًا لأهدف تحليل وتنقيب بيانات الإنترنت فإنه يمكن تقسيمها إلى ثلاثة أنواع:

أ. تنقيب محتويات الإنترنت Mining Web Content

وفيها يتم تحليل المحتوى، كالنصوص والوسائط المتعددة، من أجل فهم محتوى صفحات الإنترنت وفهرستها بحسب المضمون والكلمات المفتاحية وترتيبها العالمي وملخصاتها. وقد اهتمت شركات البحث على الإنترنت بهذا النوع من أنواع التنقيب من أجل زيادة كفاءة الخدمات التي تقدمها لمستخدمي الإنترنت.

ب. تنقيب بنية وهيكلية الإنترنت Mining Web Structure

وتتم باستخدام تقنية تنقيب الأشكال البيانية من أجل تحليل بنية وهيكلية الشبكة والصلات بين المواقع والارتباطات التشعيبية التي تساعد على التنقل بين المواقع المختلفة، أو بهدف مقارنة المواقع المختلفة وتحليل البنية الشجرية والهيكلية للمحتوى في كل موقع.

ج. تنقيب استخدام الإنترنت Mining Web Usage

وهي عملية استخراج المعلومات المفيدة التي تبين سلوك المستخدمين في التصفح واستكشاف أنماط الاستخدام لمجموعات من المستخدمين، من أجل فهم أساليب واتجاهات البحث على الإنترنت والارتباطات فيما بينها، والتنبؤ بما يبحث عنه المستخدمون على الإنترنت.

ويساعد هذا النوع من التنقيب في تحسين كفاءة وفاعلية البحث وتحسين أو ترقية المنتجات وإظهارها بشكل أفضل لمجموعات مختلفة من المستخدمين

وبالوقت المناسب.

وشركات البحث على الإنترنت تستخدم هذا النوع من التنقيب بشكل روتيني من أجل تحسين جودة الخدمات التي تقدمها للمستخدمين.

٨. تنقيب التدفق المعلوماتي (تدفق البيانات) Mining Data Streams

ويقصد به تحليل وتنقيب البيانات التي تتدفق بكثافة عالية وبأشكال متغيرة، وبطريقة لانهائية، وسمات متعددة الأبعاد.

ومثل هذه البيانات لا يمكن تخزينها في نظم قواعد البيانات التقليدية، كما أن معظم النظم المخصصة لهذا النوع من البيانات يمكن أن تقرأ ما يتدفق فيها لمرة واحدة فقط في ترتيب تسلسلي معين، وهو ما يشكل تحديًا كبيرًا لتقنيات التنقيب فيها.

وقد أدت بعض البحوث المهمة إلى تقدم في تطوير طرق فعّالة لتنقيب تدفق البيانات وخاصة في مجال استكشاف الأنماط المتكررة والمتسلسلة والتحليل متعدد الأبعاد والتصنيف والتجزئة العنقودية.

وكانت الفلسفة العامة لهذه التقنيات هي تطوير خوارزميات تهدف إلى تنقيب مجموعات من اللقطات المنفردة لهذا التدفق باستخدام قدرات محدودة في التخزين والحوسبة.

ومن أمثلة تطبيقات هذا النوع من أنواع التنقيب، المراقبة الحية للحيوانات البرية، لدراسة سلوكهم اليومي والبيئي، ودراسات التفاوض والإقناع، وتدفق النصوص والفيديو وتدفق شبكات الكهرباء وأجهزة الاستشعار وبحوث الإنترنت وأنظمة المراقبة.

الفصل الثامن
تخطيط عمليات تنقيب البيانات
وتطبيقاتها المختلفة في المجتمع

المحتويات

١. تخطيط عمليات تنقيب البيانات Data Mining Planning
٢. تنقيب البيانات في المجتمع Data Mining in Society
٣. تطبيقات تنقيب البيانات في المجالات الحيوية المختلفة في المجتمع
- الطب والصيدلة والعلوم الطبية والصيدلانية
- العلوم البيولوجية والهندسة الوراثية
- الفيزياء والجيولوجيا وعلوم والفضاء والفلك
- الإدارة وإدارة الأعمال والعلوم الإدارية
- الاقتصاد والاستثمار والخدمات العامة
- البنوك وشركات التأمين والاتصالات
- التعليم والتدريب والتنمية البشرية
- الصحة العامة والتأمين الصحي
- العلوم الأمنية والشرطية
- تطبيقات عامة بشكل تكاملي في المجالات الحيوية في المجتمع
٤. تطبيق عملي: سيناريو استخدام نظام التوصية

تخطيط عمليات تنقيب البيانات

Data Mining Planning

يُعتبر تخطيط عمليات التنقيب في البيانات من الأمور المهمة للحصول على أفضل النتائج، فالتخطيط الجيد يؤدي للنتائج الجيدة.

ويمكن تلخيص الخطوات الأولية للتنقيب في البيانات فيما يلي:

١. تحديد المشكلة المراد بحثها وإيجاد الحلول لها

٢. بناء قاعدة بيانات التنقيب

٣. التعرف على واستكشاف البيانات

٤. تحضير البيانات للتنقيب

٥. بناء نموذج التنقيب المناسب

٦. تطبيق النموذج

٧. استخراج النتائج

وفيما يلي شرحًا مختصرًا لكل خطوة:

١. تحديد المشكلة المراد بحثها وإيجاد الحلول لها

من الضروري لبحث حل مشكلة ما أن يتم تحديدها بشكل دقيق، وجمع

كافة المعلومات المتعلقة بها، وتبدأ هذه العملية بمراجعة وافية لطبيعة عمل الشركة أو المؤسسة المعنية وأهدافها والبدء بتحديد طبيعة المشكلة وأسبابها وجذورها وتفرعاتها. وليس بالضرورة أن تكون هناك مشكلة بمعنى الكلمة ولكن قد تكون المشكلة مُصاغة على شكل هدف معين، ومن أحد الأمثلة على مشاكل من هذا النوع ما يمكن صياغته على شكل هدف مثل: "زيادة نسبة المبيعات في السنة القادمة بنسبة ٣٠٪".

٢. بناء قاعدة بيانات التنقيب

وتعتبر من ضمن العمليات المهمة في مراحل التحضير، وهي من المراحل الطويلة والتي قد تأخذ وقتًا طويلاً، وحتى إذا كانت هناك قاعدة بيانات موجودة بالفعل فإنه يُستحسن بناء قاعدة بيانات أخرى جديدة بهدف التحليل والتنقيب، بحيث تشتمل على أية تعديلات أو عمليات تحضير بكل الطرق التي يمكن تطبيقها والتي تلزم لأغراض التنقيب بشكل خاص، مثلاً كأن يتم إضافة حقول جديدة أو إجراء عمليات تحويل لبعض الحقول الموجودة بالفعل.

٣. التعرف على واستكشاف البيانات

البدء باستكشاف البيانات وتحديد الحقول الأكثر أهمية والقيم التي يمكن من خلالها تحقيق أفضل النتائج، كذلك يمكن في هذه المرحلة إعداد بعض الأشكال البيانية الأولية التي تلخص البيانات أو تقدم وصفًا إحصائيًا بسيطًا

لها بهدف التعرف على طبيعة البيانات عن قرب واستكشاف ما يمكن أن يكون له دلالة تساهم في تحديد مسارات وملامح التحليل المتقدم والتنقيب المزمع تنفيذه.

٤. تحضير البيانات للتنقيب

وتشمل جميع عمليات التحضير التي يمكن تطبيقها على البيانات، والتي تشتمل على عمليات التنظيف والدمج والاختزال والتحويل التي يمكن أن يلزم تطبيقها على البيانات، ومن أمثلتها تحديد وانتقاء المتغيرات والسجلات وتخليق متغيرات جديدة بحسب الحاجة، كإنشاء متغيرات قابلة للقياس بدلاً من بعض المتغيرات غير الرقمية في قاعدة البيانات، وكذلك حساب متغيرات جديدة مبنية على أكثر من متغير إذا لزم الأمر، وغيرها من عمليات التحضير المختلفة.

٥. بناء نموذج التنقيب المناسب

اختيار خوارزمية التنقيب المناسبة والبدء بإنشائها، وذلك وفق العديد من الاعتبارات التي يمكن الاستناد عليها، كالغرض من الخوارزمية وإمكانية تطبيقها وتحقيقها لهذا الغرض ومدى إمكانية تحقيقها للأهداف المرجوة منها.

٦. تطبيق النموذج

تطبيق نموذج التنقيب وإجراء كافة التعديلات المناسبة بعد التطبيق

الأولي في حالة وجود أخطاء، مع مراعاة أهداف المشكلة وبحث إمكانية حلها بالنموذج الذي تم بناءه.

٧. استخرج النتائج

استخراج نتائج التحليل والتنقيب وتوضيح الحلول للمشكلة محل البحث وكيفية تطبيقها وفق ما توفر من نتائج تم التوصل إليها.

اختيار التقنية المناسبة

لا توجد قاعدة محددة يتم البناء عليها لاختيار تقنية من تقنيات التنقيب، ويتم الاختيار عادة بناءً على الخبرة في هذا المجال والتجربة الفعلية للتقنيات والخوارزميات المختلفة ومدى فاعليتها، ومن جهة أخرى قد تكون المفاضلة أيضًا بين التقنيات بقدر ما يكون هناك توفراً للأدوات المناسبة والخبرة في استخدام التطبيقات الجاهزة، ومع تراكم الخبرة والتجربة يتم تقييم الخيارات وتحديد المناسب منها وتطبيقه بالشكل الأفضل الذي يتناسب مع الإمكانات والأدوات المتوفرة من جهة ويناسب الاحتياجات والأهداف المرجوة من التحليل والتنقيب من جهة أخرى.

تنقيب البيانات في المجتمع
Data Mining in Society

إن تنقيب البيانات موجود في كل مكان من حولنا، وفي معظم أوجه حياتنا اليومية، سواء كنا نلاحظه وندركه أو لا نشعر به، إنه يؤثر في طريقة تسوقنا وعملنا وبحثنا عن المعلومات، كما أنه يؤثر على أساليب وأوقات الترفيه، والصحة والرفاهية التي نتمتع بها، ومنها ما هو تطبيقات ذكية وخفية، مثل تلك الموجودة في محركات البحث على الإنترنت التي تقوم بترتيب نتائج البحث بحسب أهميتها لنا دون أن نشعر، ومثل أنظمة حجز تذاكر السفر وحجوزات الفنادق والمطاعم، وعروض المنتجات المختلفة في المواسم والأعياد أو حتى في المناسبات الخاصة.

فكل تلك النظم تعتمد على خوارزميات التنقيب الذكية الخفية التي قد لا يشعر بوجودها المستخدم، وتُعتبر مجهولة بالنسبة له.

فعندما نقوم بالتسوق في أحد المراكز التجارية ونستلم فاتورة الحساب، يتم تخزين بياناتها في قواعد بيانات مخصصة، ومع تكرار رحلات التسوق وتراكم البيانات يتم استخدام تطبيقات التحليل والتنقيب على تلك البيانات واستكشاف التفضيلات المتنوعة للزبائن، ماذا يشترون وما هو سلوكهم في التسوق، حتى في بعض المحلات الكبرى والشهيرة، التي تعتمد على تقنيات التنقيب في عملها، فإنها تسمح بدخول موردي الأصناف المختلفة، أو حتى

المُصنعين لها، إلى تطبيقات التنقيب من أجل استخدامها في استكشاف أنماط وسلوك الشراء المتنوعة للزبائن في الفروع المختلفة واستثمار هذه المعرفة في التحكم في المخزون وتقدير أماكن وضع المنتجات على الأرفف وكمياتها المناسبة، وتطوير استراتيجيات التسويق والترويج لمنتجاتهم.

كما ينتشر استخدام تطبيقات وخوارزميات التحليل والتنقيب في مواقع التسوق على الإنترنت، كمواقع بيع الكتب والموسيقى والأفلام والألعاب، ويتم تقديم توصيات بالمنتجات لكل زبون بحسب آراء الزبائن والتفضيلات الشخصية، وهو الأسلوب المستخدم في شركة أمازون Amazon، التي تسخّر تقنيات التنقيب لخدمة استراتيجياتها التسويقية.

ويختلف التسوق على أرض الواقع عن التسوق على الإنترنت اختلاف جوهري، حيث أنه عندما يدخل الزبون إلى المتجر ويبدأ التسوق على أرض الواقع فيتم اعتباره زبونًا بالفعل ويصعب خسارته بسهولة، أو بمعنى آخر يكون قرار انتقاله إلى متجر آخر مكلفًا ويندر حدوثه، أما في التسوق على الإنترنت فإن انتقال الزبون إلى متجر آخر يتم بسهولة وبضغطة زر، الأمر الذي يُشكل تحديًا كبيرًا للشركات التي تتيح التسوق على الإنترنت، ويجعلها تكثّف من استخدامها لتقنيات التنقيب من أجل ضمان الاحتفاظ بالزبون لأطول فترة ممكنة، وتقديم أفضل الخدمات له بما يتناسب مع سماته وتفضيلاته الشخصية، وتقديم ما أمكن من معلومات يمكن أن تفيده وتساعده في اتخاذ قراره بالشراء ومحاولة الاحتفاظ به كزبون دائم.

وكثيرًا من الشركات زادت من استخدامها لأدوات التنقيب لأغراض تطوير إدارة علاقاتها بالزبائن، والتي تساعد في تقديم خدمات تناسب احتياجات كل زبون بحسب تفضيلاته واهتماماته بدلاً من التسويق العشوائي، فمن خلال دراسة سلوك وأنماط الشراء مثلاً تستطيع الشركات تطوير آلية الإعلان والترويج بطرق ذكية بحيث لا تسبب الإزعاج للزبائن نتيجة تراكم الإعلانات للمنتجات التي لا يهتمون بها، الأمر الذي يؤدي إلى التوفير الملموس لهذه الشركات في الوقت والجهد والتكاليف، كما أنه يوفر للزبائن المهتمين بالفعل القدرة على متابعة الإعلانات التي تهمهم وعدم إضاعة الوقت في متابعة ما لا يهمهم.

أما في الشركات المتخصصة في البحث على الإنترنت، مثل شركة جوجل Google، والتي توفر بلايين المعلومات عن المواقع والصفحات المفهرسة على الإنترنت فإنها تعتمد في عملها على خوارزميات التنقيب بشكل أكبر، فعند البحث في جوجل عن أي كلمة أو تعبير تقوم جوجل بعرض قائمة من المواقع التي يرتبط محتواها بهذا التعبير، مرتبة وفقًا لأهميتها، وذلك باستخدام خوارزميات تنقيب مختلفة، تشمل ترتيب المواقع بحسب أهميتها وبحسب الأنماط الأكثر تكرارًا التي تم اكتشافها بالتنقيب والتي تبين سلوك المستخدمين تجاه تلك النتائج بالضغط عليها، مثلاً، عند البحث عن تعبير أو حتى كلمتين مثل (القاهرة الإسكندرية)، فإن جوجل يقوم بعرض المواقع الأكثر أهمية والأكثر زيارة من المستخدمين، كخدمة القطارات ومواعيدها بين القاهرة

والإسكندرية، أي أنها تعرض المعلومات المهمة بطريقة ذكية بعد اكتشاف أنماط الضغطات التي قام بها المستخدمون في فترات سابقة.

كما تقوم شركة جوجل بعرض الإعلانات التجارية المدفوعة التكاليف لشركات أخرى، ترتبط منتجاتها وخدماتها بموضوع البحث الذي يقوم به المستخدم، وهو الأمر الذي يجعله سعيدًا ويقلل من انزعاجه من الإعلانات التي لا يهتم بها.

ويمكن القول بأن معظم تطبيقات وخوارزميات التنقيب في قواعد البيانات غير منظورة للمستخدم، وقد لا يُدرك أن ما يحصل عليه من نتائج يعتمد على خوارزميات التنقيب، أو أن سلوكه يمكن أن يُستخدم في تغذيتها، ومع ذلك فإن التحديات المستقبلية لخوارزميات التنقيب في العديد من المجالات لا تزال كبيرة، وتتمتع بآفاق واسعة ومتعددة، والتي لا تزال تستلزم المزيد من البحوث والدراسات لتطوير خوارزميات من شأنها المساهمة في زيادة فاعلية وكفاءة النظم المستخدمة في تلك المجالات.

ومن أمثلة تلك التحديات ما يتعلق بدراسة الاستجابة للإعلانات التجارية، والعوامل المؤثرة التي ترفع من معدلات واحتمالات الإقبال على شراء منتج معين عند مشاهدة الإعلان الخاص به، ويعتبر هذا المسار من مسارات بحوث التنقيب الحديثة، والأكثر أهمية في مجال الأعمال في هذا العصر، وقد أضحى من أولويات عمل شركات الاستشارات والتسويق العالمية في الوقت الراهن.

من جهة أخرى، يُشكل مسار بحوث التنقيب في مجال الطب والصحة العامة تحديًا ذات أبعاد إستراتيجية عالمية بخاصة في الدول الكبرى التي قامت بإنشاء مؤسسات متخصصة تُنظم المنافسات والمسابقات الدولية الكبرى للمحللين في كل أنحاء العالم للتنافس من أجل تحليل وتنقيب البيانات الطبية والصحية في تلك الدول وتقديم النتائج والتوصيات التي تؤدي إلى تحسين الأوضاع الصحية فيها والحد من انتشار الأمراض والتوفير في ميزانية منظومة الصحة بشكل عام.

تطبيقات تنقيب البيانات في المجالات الحيوية المختلفة في المجتمع

١. الطب والصيدلة والعلوم الطبية والصيدلانية

تُعتبر تقنيات التحليل المتقدم وتنقيب البيانات من أهم الأدوات التي تُستخدم في المجالات الطبية والصيدلانية، بخاصة في مجال الاستكشاف والتقييم للأوضاع الصحية السائدة وبحث أسباب الإصابة بالإمراض واستكشاف السلوك المرضي في المجتمع، وذلك من أجل المساهمة في وضع الخطط والسياسات الطبية والصحية المناسبة والحد من انتشار الأمراض، وأينما وُجدت البيانات الطبية والصحية فإنه يمكن استخدام تقنيات التحليل والتنقيب من أجل دراستها وتحليلها واستكشاف كل ما من شأنه المساهمة في تحسين الوضع الصحي بشكل عام في المجتمع وتطوير أداء المؤسسات الصحية والتقليل من مخاطر التعرض للأمراض، وتطوير سياسات وإجراءات التوعية الصحية للفرد والأسرة، كما تساعد تقنيات التنقيب في استكشاف وتوصيف الأمراض الأكثر شيوعًا في مناطق محددة أو أزمنة معينة أو ظروف وأحوال بعينها، بهدف وضع الحلول المناسبة واتخاذ سبل الوقاية اللازمة للحد من انتشار الأمراض، بالإضافة لإعداد البحوث المتخصصة في دراسة الأدوية والعلاجات الطبية وسبل تطويرها وتحديثها ورفع كفاءتها وفاعليتها وصلاحيتها وقدرتها على العلاج.

٢. الهندسة الوراثية والمعلوماتية الحيوية

يعتمد علم الهندسة الوراثية والمعلوماتية الحيوية بشكل أساسي على تقنيات وخوارزميات تنقيب البيانات، حتى أن بعض خوارزميات التنقيب تم تطويرها خصيصًا لأغراض بحوث المعلوماتية الحيوية بطريقة جعلت من علم التنقيب والمعلوماتية الحيوية مسارين من مسارات العلوم المتكاملة ويطور كل منهما الآخر بشكل تبادلي. كما تُعتبر بيئة المعلوماتية الحيوية خصبة لتقنيات التنقيب نظرًا لما تحتويه من كم هائل من البيانات التي تكاد تكون لانهائية والمتمثلة في سلاسل الأحماض الأمينية والسلاسل الجينية التي تُعد بالمليارات، وقد شكلت أدوات التنقيب وسيلة فعّالة لدراسة وتحليل السلاسل بهدف استكشاف الأنماط وفك شيفرات الهندسة الوراثية للإنسان، الأمر الذي ساهم في فهم طبيعة الأمراض واكتشاف عقاقير جديدة وفعّالة لعدد كبير منها وتطوير العقاقير الموجودة ورفع كفاءتها وفاعليتها. وبشكل عام تُساهم تقنيات تنقيب البيانات في رفع كفاءة البحوث والدراسات الحيوية بكل أشكالها وتعزز من قدرة الباحثين على التحليل المتقدم والمعمق للمعلومات الحيوية بطرق جديدة وغير مسبوقة.

٣. الفيزياء والجيولوجيا وعلوم والفضاء والفلك والبيئة والمناخ

في ظل التطور التكنولوجي الكبير في هذا العصر أصبح بمقدور الباحثين والعلماء المتخصصين في بحوث الفضاء والفلك والطبيعة الحصول على المزيد

من البيانات، سواء التطور الحادث في تقنيات التصوير الرقمي أو التلسكوبات والعدسات المستخدمة فيها وأدوات الرصد الأخرى، وقد ساهم هذا التطور في تراكم كم هائل من البيانات الجديدة التي لم يكن بالإمكان الحصول عليها سابقًا، وكل ذلك أدى إلى الحاجة إلى تقنيات جديدة تواكب هذه النقلة وتساهم في تحليل هذا الكم الجديد من البيانات، ما جعل من أدوات تنقيب البيانات الوسيلة المثلى للاستفادة من تلك النقلة وتحويل الكم الهائل من البيانات الجديدة إلى معرفة تساهم في فهم الطبيعة وتعزز من قدرة الإنسان على مواجهة التغيرات والتقلبات الطبيعية والمناخية، واستكشاف خفايا وأسرار الطبيعة والاستفادة من هذه المعرفة لصالح البشرية في كل مكان.

٤. الإدارة وإدارة الأعمال والاقتصاد والصناعة والاستثمار

تعتمد الشركات والمؤسسات الاقتصادية والصناعية والاستثمارية على تقنيات تنقيب البيانات في معظم أركان العمل، سواء من حيث الإدارة التنفيذية في الأقسام والإدارات المتخصصة أو من حيث الإدارة الإستراتيجية العليا فيها، حتى أنه يمكن القول بدون مبالغة بأن بعض نظم الإدارة الحديثة تم تطويرها وفقًا لما وفرته تقنيات تنقيب البيانات من معرفة جديدة وبطرق لم تكن متوفرة من قبل، وتُساعد تقنيات تنقيب البيانات في تطوير الأعمال في الشركات والمؤسسات من منظور إداري وتنظيمي وبدقة عالية تتشابه مع دقة وتنظيم تقنيات التنقيب سواء من حيث تطوير الأداء وزيادة الإنتاجية

ورفع معدلات الكفاءة وتحسين جودة المنتجات والأعمال وتطوير آلية العمل وخطوط الإنتاج وتحسين خطط التسويق، أو من حيث تعزيز قدرة متخذي القرار وواضعي السياسات والإستراتيجيات في الشركة أو المؤسسة، نظراً لما توفره من معرفة مفيدة لصنّاع القرار بطرق جديدة تتطور يومياً وتتنوع بتنوع مجالات الأعمال المختلفة.

٥. البنوك وشركات التأمين والاتصالات

إن تقنيات التحليل المتقدم وتنقيب البيانات هي العصب الرئيسي لمعظم المؤسسات المالية بشكل عام والبنوك وشركات التأمين بشكل خاص، وذلك نظراً لما توفره من قدرة على الاستكشاف والتنبؤ الذي يُعتبر من أهم الاحتياجات المعرفية لهذا النوع من المؤسسات، سواء من أجل تقدير وتحليل المخاطر واتخاذ ما يلزم لتجنبها أو من أجل استكشاف الفرص الاستثمارية واقتناصها قبل فوات الأوان، إضافة لما توفره من قدرة معرفية تساهم في إدارة هذه المؤسسات بطرق حديثة تعتمد بشكل أساسي على المعرفة المستكشفة والتي يتم التنبؤ بها من خلال دراسة وتحليل وتنقيب ما يتوفر من بيانات تاريخية في مجال العمل أو حتى في مجالات أخرى، ففي شركات التأمين يتم الاعتماد بشكل أساسي على دراسة وتحليل بيانات حوادث المركبات من كل الأنواع ومن ثم الاستكشاف والتنبؤ بنسبة احتمال التعرض لها من أجل احتساب أقساط التأمين المناسبة لكل نوع، وكذلك ما

يحدث في نظام منح القروض في البنوك الذي يتأسس على وجود خوارزميات تصنيف وتنبؤ تقوم بفحص بيانات وسمات الشخص المتقدم للقرض وتقدير المخاطرة المترتبة على منحه إياه وفق ما يتوفر من بيانات تاريخية عن المقترضين السابقين الذين يتشابهون معه في السمات، كالعمر ومستوى الدخل وطبيعة العمل أو المهنة وغيرها، ومدى التزامهم في السداد. وتعتبر مجالات عمل المؤسسات المالية والبنوك وشركات التأمين وحتى شركات الاتصالات من المجالات الحيوية المهمة لتقنيات تنقيب البيانات وبيئة خصبة لتطبيق العديد من تقنيات التنقيب الاستكشافية والتنبؤية التي تلزم مثل هذا النوع من المؤسسات أكثر من غيرها.

٦. التعليم والتدريب والتنمية البشرية

يمكن استخدام تقنيات التحليل المتقدم وتنقيب البيانات من أجل تطوير منظومة التعليم والتعليم العالي ومنظومة التدريب والتنمية البشرية بشكل عام في أي مجتمع، وذلك بإجراء البحوث المتخصصة التي تستهدف تطوير المناهج والتخصصات العلمية والتدريبية وربطها باحتياجات المجتمع، باستخدام تقنيات تنقيب تقوم بتحليل وتنقيب بيانات الطلاب ومحتوى المناهج الدراسية ونتائج الامتحانات بهدف استكشاف تفضيلات الطلاب وميولهم العلمية وقدراتهم الفكرية والعملية ومدى توافقها مع محتوى المناهج والمقررات الدراسية، كما تفيد تقنيات التنقيب في دراسة وتحليل نتائج امتحانات قبول

الطلاب في الجامعات والمعاهد في الكليات والأقسام المختلفة لاختيار التخصصات الأكاديمية التي تتناسب مع ميولهم وتفضيلاتهم وقدراتهم، بالإضافة لدراسة وتحليل منظومة الإدارة التعليمية بشكل عام واستكشاف سبل تطوير وتحديث آلية العمل وكيفية رفع مستوى أداء وكفاءة المدرسين وسبل حل المشكلات التي تعيق مسيرة التنمية بشكل عام في المجتمع والتي ترتبط بشكل مباشر أو غير مباشر بمنظومة التعليم.

٧. الصحة العامة والتأمين الصحي والرياضة

تُعتبر مجالات الصحة العامة والتأمين الصحي والرياضة من المجالات التي تعتمد على تقنيات تنقيب البيانات بشكل كبير، ويزداد استخدام تقنيات تنقيب البيانات في هذه المجالات بشكل خاص، وذلك نظراً للزيادة في عدد السكان وزيادة المشكلات الصحية في المجتمعات وتراكم الكم الهائل من البيانات الأمر الذي أدى لحاجة ملحة لتحليلها وتنقيبها واستكشاف المعرفة التي يمكن أن تقدم حلولاً للمشكلات الصحية وإنشاء نظام تأمين صحي ذكي يعتمد على تقنيات التنقيب التنبؤية من أجل وضع أفضل خطة لنظام تأمين صحي يستفيد منه كل أفراد المجتمع وبأقل التكاليف، وهو أمر يمكن تحقيقه بسهولة باستخدام تقنيات تنقيب البيانات في هذا العصر.

من جهة أخرى فإن استخدام التكنولوجيا في المجال الرياضي أدى إلى ظهور استخدامات جديدة لتقنيات التنقيب في هذا المجال والتي تهدف

لدراسة وتحليل الأداء واستكشاف سلوك اللعب لدى اللاعبين والفرق الرياضية ومدى الانسجام فيما بينهم كما تهدف لاستكشاف علاقات الارتباط بين حالات الفوز أو الخسارة وبين الأداء بطريقة معينة أو بمشاركة لاعبين معينين أو في ظروف معينة، وذلك بالاعتماد على البيانات التي أصبحت متوفرة في عصر التكنولوجيا والمتمثلة بالإحصاءات والبيانات المتنوعة التي لم يكن بالإمكان توفيرها في عقود سابقة، وهو ما أدى لمزيد من الاستخدامات لتقنيات التنقيب في المجال الرياضي، حتى أننا نستطيع أن نقول بدون مبالغة بأن النجاح الرياضي في هذا العصر بات يعتمد على ما يتوفر من معرفة لدى المدربين يتم استخلاصها من تحليل وتنقيب البيانات ربما بشكل أكبر من اعتماده على ما يتوفر من قدرات لدى اللاعبين أنفسهم.

٨. العلوم الأمنية والشرطية والدفاع المدني

من أهم التحديات الحديثة لتقنيات تنقيب البيانات في هذا العصر هي ما يتعلق بمسألة الأمن والعلوم الأمنية والشرطية بشكل عام، بخاصة بعد تعاظم المخاطر الأمنية وتعدد أشكالها وأنواعها في ظل تطور تكنولوجي ذا حدين ساهم في ظهور أشكال جديدة للجريمة وتعدد مصادرها وأسبابها ودوافعها بشكل أدى إلى الحاجة لاستخدام تقنيات جديدة تساعد في درء المخاطر الأمنية، وتُستخدم تقنيات التنقيب في تحليل أسباب ارتكاب الجريمة أو أسباب الحوادث والحرائق والوقوف على الأسباب الأكثر شيوعًا وعوامل

انتشارها في مناطق محددة أو أزمنة معينة والتطور الذي يطرأ عليها، كما يتم بناء خوارزميات أمنية متخصصة تهدف للاستكشاف والتنبؤ بوقوع الجرائم أو الحوادث المرورية والحرائق قبل وقوعها، وتتلخص مهمة مثل هذه الخوارزميات في دراسة وتحليل سمات وخصائص من يُقبلون على ارتكاب جرائم من نوع معين أو يتسببون في حوادث الطرق ومن ثم استخدام النتائج في التنبؤ وتصنيف أشخاص آخرين وبحث مدى استعدادهم لارتكاب مثل هذه الجرائم أو التسبب بمثل تلك الحوادث. وتتنوع أساليب بناء الخوارزميات الأمنية بحسب طبيعة المخاطر الأمنية في المجتمع والاحتياجات الملحّة فيه، كما تساهم النتائج التي يتم الحصول عليها من تقنيات وخوارزميات التنقيب الأمنية في تطوير القوانين والتشريعات التي يتم تحديثها بما يتناسب مع نتائج تلك التحليلات، كأن يتم رفع سن الحصول على رخصة قيادة المركبات إلى سن (٢٠) سنة مثلاً عندما تستكشف إحدى الخوارزميات الأمنية أن احتمال أن يتسبب بحوادث المرور من هم دون سن العشرين سنة أكبر بكثير من احتمال أن يتسبب بها من تخطوا سن العشرين.

٩. تطبيقات عامة بشكل تكاملي في المجالات الحيوية المختلفة في المجتمع

إن الفوائد التي توفرها تقنيات التحليل المتقدم وتنقيب البيانات لا تقتصر فقط على تلك الفوائد التي يتم تحقيقها في مختلف المجالات الحيوية في المجتمع بشكل منفرد، بل إنها تمتد لتصل إلى حد إمكان تحقيق الفائدة الأهم والتي

تنتج عن تحليل كميات هائلة من البيانات التي تخص عدة مجالات معًا بشكل تكاملي، فمثلاً يمكن استخدام تقنيات التنقيب من أجل دراسة كل ما يختص بمسألة التعليم مثلاً وبشكل تكاملي مع مجالات أخرى كالفنون والآداب والرياضة والصحة، وذلك من أجل بحث واستكشاف العلاقات البينية والارتباطات والتبعية ومدى التأثير والتأثر المتبادل بين كل من تلك المجالات بحيث يتم التوصل إلى استنتاجات من شأنها أن تساعد صنّاع القرار في وضع السياسات والاستراتيجيات والخطط اللازمة للتنمية والتطوير في كل تلك المجالات بشكل تكاملي يعزز كل منها للآخر.

وبشكل عام يمكن الاستفادة من تقنيات تنقيب البيانات في كل المجالات الحيوية في المجتمع وبشكل تكاملي، وصولاً إلى مجتمع المعرفة الذي يتم بناءه على أسس علمية ذكية من أجل تحقيق التنمية الإستراتيجية المستدامة والتقدم والازدهار بأقل التكاليف من جهة، وضمان إيجاد الحلول السريعة والذكية لأية مشكلات أو أزمات قد تطرأ في المستقبل من جهة أخرى.

تطبيق عملي: سيناريو استخدام نظام التوصية

لنفترض أنك قمت بزيارة موقع الويب الخاص بمحل لبيع الكتب عبر الإنترنت (مثل Amazon) بهدف شراء كتاب كنت ترغب في قراءته. تكتب اسم الكتاب. هذه ليست المرة الأولى التي تزور فيها موقع الويب. لقد قمت بتصفحه من قبل وقمت بعمليات شراء منه في عيد الميلاد الماضي. يتذكر متجر الويب زياراتك السابقة، حيث قام بتخزين معلومات تدفق النقرات والمعلومات المتعلقة بمشترياتك السابقة. يعرض النظام وصف وسعر الكتاب الذي حددته للتو. إنه يقارن اهتماماتك مع العملاء الآخرين الذين لديهم اهتمامات مماثلة ويوصي بعناوين كتب إضافية، قائلاً "العملاء الذين اشتروا الكتاب الذي حددته اشتروا أيضًا هذه العناوين الأخرى أيضًا". من خلال مسح القائمة، ترى عنوانًا آخر يثير اهتمامك وتقرر شراءه أيضًا.

لنفترض الآن أنك تذهب إلى متجر آخر عبر الإنترنت بنية شراء كاميرا رقمية. يقترح النظام عناصر إضافية يجب مراعاتها استنادًا إلى الأنماط التسلسلية المحددة مسبقًا، مثل "العملاء الذين يشترون هذا النوع من الكاميرات الرقمية من المرجح أن يشتروا علامة تجارية معينة من الطابعة أو بطاقة الذاكرة أو برنامج تحرير الصور في غضون ثلاثة أشهر". عليك أن تقرر شراء الكاميرا فقط، دون أي عناصر إضافية. بعد أسبوع، تتلقى كوبونات من المتجر بخصوص العناصر الإضافية.

تتمثل إحدى ميزات أنظمة التوصيات في أنها توفر التخصيص لعملاء

التجارة الإلكترونية، وتروج للتسويق الفردي. تقدم أمازون، وهي شركة رائدة في استخدام أنظمة التوصية التعاونية، "متجرًا مخصصًا لكل عميل" كجزء من إستراتيجيتها التسويقية. يمكن أن يفيد التخصيص كل من المستهلكين والشركة المعنية. من خلال امتلاك نماذج أكثر دقة لعملائها، تكتسب الشركات فهمًا أفضل لاحتياجات العملاء. يمكن أن تؤدي تلبية هذه الاحتياجات إلى نجاح أكبر فيما يتعلق بالبيع المتقاطع للمنتجات ذات الصلة، والارتقاء بالمبيعات، والارتباطات بالمنتجات، والعروض الترويجية الفردية، والسلال الأكبر، والاحتفاظ بالعملاء.

تحدد مسألة التوصية مجموعة، C ، من المستخدمين ومجموعة، S، من العناصر. لنفترض أن وظيفة الأداة المساعدة تقيس فائدة عنصر، أو عناصر، للمستخدم، ج. يتم تمثيل الأداة بشكل عام بواسطة خوارزمية تصنيف ويتم تعريفها مبدئيًا فقط للعناصر التي سبق تصنيفها بواسطة المستخدمين. على سبيل المثال، عند الانضمام إلى نظام التوصية بالأفلام، يُطلب من المستخدمين عادةً تقييم عدة أفلام. المساحة C, S لجميع المستخدمين والعناصر المحتملة ضخمة. يجب أن يكون نظام التوصية قادرًا على الاستقراء من التصنيفات المعروفة إلى غير المعروفة وذلك للتنبؤ بتوليفات العناصر والمستخدمين. يُنصح المستخدم بالعناصر ذات أعلى تصنيف / فائدة متوقعة له.

كيف يتم تقدير فائدة عنصر ما للمستخدم؟

- **الطرق المستندة إلى المحتوى**

في الطرق المستندة إلى المحتوى، يتم تقديرها بناءً على الأدوات المساعدة التي يخصصها نفس المستخدم لعناصر أخرى متشابهة. تركز العديد من هذه الأنظمة على التوصية بالعناصر التي تحتوي على معلومات نصية، مثل مواقع الويب والمقالات والرسائل الإخبارية. يبحثون عن القواسم المشتركة بين العناصر. بالنسبة للأفلام ، قد يبحثون عن أنواع أو مخرجين أو ممثلين متشابهين. بالنسبة للمقالات، قد يبحثون عن مصطلحات مماثلة. الأساليب القائمة على المحتوى متجذرة في نظرية المعلومات. يستخدمون الكلمات الرئيسية (التي تصف العناصر) وملفات تعريف المستخدمين التي تحتوي على معلومات حول أذواق المستخدمين واحتياجاتهم. يمكن الحصول على هذه الملفات الشخصية بشكل صريح (على سبيل المثال، من خلال الاستبيانات) أو تعلمها من سلوك معاملات المستخدمين بمرور الوقت.

- **الطرق المستندة إلى التعاون**

يحاول نظام التوصية التعاوني التنبؤ بفائدة العناصر للمستخدم، u، بناءً على العناصر التي تم تصنيفها مسبقًا بواسطة مستخدمين آخرين يشبهون u. على سبيل المثال، عند التوصية بالكتب، يحاول نظام التوصيات التعاوني العثور على مستخدمين آخرين لديهم تاريخ في الاتفاق معك (على سبيل المثال،

يميلون إلى شراء كتب مماثلة، أو إعطاء تقييمات مماثلة للكتب). يمكن أن تكون أنظمة التوصية التعاونية إما قائمة على الذاكرة (أو الكشف عن مجريات الأمور) أو قائمة على النموذج.

- **الطرق المستندة على الذاكرة**

تستخدم الأساليب القائمة على الذاكرة بشكل أساسي الأساليب البحثية لإجراء التصنيف التنبؤي بناءً على المجموعة الكاملة للعناصر التي تم تصنيفها مسبقًا من قبل المستخدمين. بمعنى، يمكن تقدير التصنيف غير المعروف لمجموعة العناصر والمستخدمين كمجموع لتصنيفات المستخدمين الأكثر تشابهًا لنفس العنصر. عادةً ما يتم استخدام نهج الجار الأقرب، أي أننا نجد المستخدمين الآخرين (أو الجيران) الأكثر تشابهًا مع مستخدمنا المستهدف، u. يمكن استخدام طرق مختلفة لحساب التشابه بين المستخدمين. تستخدم الأساليب الأكثر شيوعًا إما معامل ارتباط بيرسون أو التشابه. يمكن استخدام التحليل العنقودي، والذي يعدل مع حقيقة أن المستخدمين المختلفين قد يستخدمون مقياس التصنيف بشكل مختلف. تستخدم أنظمة التوصية التعاونية القائمة على النموذج مجموعة من التقييمات لتعلم نموذج، والذي يتم استخدامه بعد ذلك لعمل تنبؤات التقييم. على سبيل المثال، يتم استخدام النماذج الاحتمالية والتحليل العنقودي (الذي يجد مجموعات من العملاء ذوي التفكير المماثل) والتصنيف باستخدام نظرية الاحتمالات Bayesian وتقنيات التعلم الآلي الأخرى.

قابلية التوسع وضمان الجودة للمستهلك

تواجه أنظمة التوصية تحديات كبيرة مثل قابلية التوسع وضمان توصيات الجودة للمستهلك. على سبيل المثال، فيما يتعلق بقابلية التوسع، يجب أن تكون أنظمة التوصية التعاونية قادرة على البحث من خلال الملايين من الجيران المحتملين في الوقت الفعلي. إذا كان الموقع يستخدم أنماط تصفح كمؤشرات لتفضيل المنتج، فقد يحتوي على آلاف نقاط البيانات لبعض عملائه. يعد ضمان توصيات الجودة أمرًا ضروريًا لكسب ثقة المستهلكين. إذا اتبع المستهلكون توصية من النظام ولكنهم لم يعجبهم المنتج بعد ذلك، فمن غير المرجح أن يستخدموا نظام التوصية مرة أخرى.

كما هو الحال مع أنظمة التصنيف، يمكن لأنظمة التوصية أن ترتكب نوعين من الأخطاء: السلبيات الكاذبة والإيجابيات الخاطئة. هنا، السلبيات الكاذبة هي المنتجات التي يفشل النظام في التوصية بها، على الرغم من أن المستهلك سيحبها. الإيجابيات الكاذبة هي المنتجات التي يوصى بها ولكن المستهلك لا يحبها. الإيجابيات الكاذبة أقل استحسانًا لأنها يمكن أن تزعج المستهلكين أو تغضبهم.

أنظمة التوصيات المستندة إلى المحتوى محدودة بالميزات المستخدمة لوصف العناصر التي يوصون بها. والتحدي الآخر لكل من أنظمة التوصية القائمة على المحتوى والتعاون هو كيفية التعامل مع المستخدمين الجدد الذين لم يتوفر لهم سجل شراء حتى الآن.

خوارزميات هجينة

تدمج المناهج الهجينة كلاً من الأساليب القائمة على المحتوى والأساليب التعاونية لتحقيق مزيد من التوصيات المحسّنة. كانت جائزة Netflix عبارة عن مسابقة مفتوحة أقامتها خدمة تأجير أقراص DVD عبر الإنترنت، بدفع تعويضات قدرها مليون دولار لأفضل خوارزمية للتوصية بتقييم تقييمات المستخدمين للأفلام، بناءً على التصنيفات السابقة.

أظهرت المنافسة والدراسات الأخرى أنه يمكن تحسين الدقة التنبؤية لنظام التوصية بشكل كبير عند مزج العديد من المتنبئين، خاصةً باستخدام مجموعة من الطرق المختلفة إلى حد كبير، بدلاً من تنقيح تقنية واحدة.

أنظمة التوصية التعاونية هي شكل من أشكال الإجابة الذكية على الاستفسارات، والتي تتكون من تحليل الغرض من الاستعلام وتقديم معلومات عامة أو مجاورة أو مرتبطة بالاستعلام. على سبيل المثال، بدلاً من مجرد إعادة وصف الكتاب والسعر استجابةً لاستعلام العميل، يقوم نظام التوصية بإعادة معلومات إضافية ذات صلة بالاستعلام ولكن لم يتم طلبها صراحةً (على سبيل المثال، تعليقات تقييم الكتاب أو توصيات الكتب الأخرى أو إحصاءات المبيعات) كإجابة ذكية لنفس الاستعلام.

تم بحمد الله،

د. م. مصطفى فؤاد عبيد

قائمة المراجع

1. Alex Freitas and Simon Lavington, "Mining Very Large Databases with Parallel Processing", Kluwer Academic Publishers, 1998.

2. A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 1988.

3. Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for Knowledge Discovery in Databases", IEEE Transactions on Knowledge and Data Engineering, 1993.

4. David Freedman, Robert Pisany, Roger Purves, "Statistics", USA.1980

5. David J. Hand, Heikki Mannila, Padhraik Smyth, "Principles of Data Mining, Adaptive Computation and Machine Learning", Thomas Dietterich Editor, USA. 2001,

6. James T. McClave, P. George Benson, Terry Sincich, "Statististics For Business & Economic", 7th Ed, International Edetion, Prentice-Hall. Inc., New Jersey, USA. 1998,

7. Jiawei Han, Micheline Lamber, "Data Mining Concepts And Techniques", Academic Press, CA, USA. 2001,

8. J. Ross Quinlan, "Programs for Machine Learning",

Morgan Kaufmann Publishers, 1993.

9. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

10. M-S. Chen, Jiawei Han, and Philip S. Yu, "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, 1996.

11. R. Lyman Ott, "An Introduction to Statistical Methods and Data Analysis", Fourth Edition, Duxbury Press, California, USA. 1992.

12. Sholom M. Weiss and Nitin Indurkhya, "Predictive Data Mining: A Practical Guide", Morgan Kaufmann Publishers, 1998.

13. Surajit Chaudhuri, "Data Mining and Database Systems: Where is the Intersection?", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 21, no. 1, March 1998.

14. Tom M. Mitchell, "Does machine learning really work?", AI Magazine, vol. 18, no. 3, pp. 11-20, Fall 1997.

15. V. Cherkassky and F. Mulier, "Learning From Data", John Wiley & Sons, 1998.

الملحق (١)
أساسيات قواعد البيانات Databases

تعريف قواعد البيانات

قواعد البيانات ومفردها قاعدة البيانات أو قاعدة المعلومات أو قاعدة المعطيات Database، هي مجموعة مشتركة من عناصر البيانات ذات الصلة والمرتبطة مع بعضها البعض بعلاقة منطقية أو رياضية. وتُستخدم لدعم أنشطة مؤسسة أو شركة معينة.

ويمكن النظر إلى قاعدة البيانات كمستودع للبيانات التي يتم تعريفها مرة واحدة ثم الوصول إليها من قبل العديد من المستخدمين.

وتتكون قاعدة البيانات من جدول واحد أو أكثر.
ويتكون الجدول من سجل (صف) أو أكثر.
ويتكون السجل من حقل أو أكثر.

ومن أمثلة السجلات، السجل الخاص بموظف معين في إحدى الشركات، فهو يتكون من عدة حقول مثل: رقم الموظف، اسم الموظف، الدرجة الوظيفية، تاريخ التعيين، الراتب الأساسي، القسم التابع له، ... إلخ.

وغير ذلك من بيانات الموظفين التي تُخزن في جهاز الحاسوب بشكل منظّم ضمن برنامج حاسوب يسمى نظام إدارة قواعد البيانات Database

Management System يقوم بتسهيل التعامل مع البيانات والبحث فيها وتمكين المستخدم من الإضافة والتعديل عليها.

يتم استرجاع البيانات باستخدام أوامر مخصصة في لغة تُسمى لغة الاستعلام الهيكلية Structured Query Language وُيرمز لها بالاختصار SQL، حيث تساعد هذه اللغة في الاستعلام عن بيانات محددة أو تلخيص البيانات واسترجاع المعلومات المفيدة التي تساعد الإدارة في عملية اتخاذ القرار.

أهداف قواعد البيانات

إن الهدف الأساسي من تصميم قواعد البيانات هو إنشاء بيئة متكاملة لحفظ وتخزين كميات البيانات التي تخص جهة أو مؤسسة أو شركة معينة، مع التركيز على طريقة تنظيم البيانات بحيث تكون نموذجية.

أي أن الهدف الرئيسي لمصمم قواعد البيانات هو هيكلة البيانات بحيث تستوعب كميات كبيرة من البيانات بشكل منظم وتكون خالية من التكرار ويمكن استرجاعها وتعديلها والإضافة عليها دون المشاكل التي يمكن أن تحدث مع وجود التكرار فيها. مع الأخذ بالاعتبار جعل تركيبة البيانات أقرب للطبيعة التصنيفية لتحقيق الاستفادة القصوى منها.

وتحقق قواعد البيانات الكثير من الأهداف لجهات الإدارة الوسطى والإدارة الاستراتيجية.

كما توفر إمكانية القيام بعمليات التحليل وتنقيب البيانات والتوصل إلى الاستنتاجات اللازمة لمتخذي القرار في الشركة أو المؤسسة.

أنواع قواعد البيانات

هناك تركيبات لقواعد البيانات حسب نوع العلاقة المنطقية أو الرياضية بين أنواع البيانات المختلفة فيها.

يمكن تعريف ثلاثة أنواع من قواعد البيانات من حيث طبيعة التركيب وهي كما يلي:

١. قواعد البيانات العلائقية
٢. قواعد البيانات الهيكلية
٣. قواعد البيانات الهرمية

أما من حيث حجم قواعد البيانات وطبيعة استخدامها يوجد أيضًا:

١. قواعد البيانات الضخمة
٢. مستودعات البيانات

وفيما يلي وصفًا موجزًا لكل منها:

- **قواعد البيانات العلائقية**

قواعد البيانات العلائقية هي التي تعتمد التركيب العلائقي بين عناصر البيانات، وهو اعتماد علاقة محددة بين عناصر البيانات، مثل أن تكون قيمة

عنصر معتمدة على حاصل جمع عنصرين. وهذا التركيب هو أنجح التراكيب المطبقة في عالم قواعد البيانات المعلوماتية، وذلك بسبب إعطائه تنوع في نوع العلاقة بين البيانات، لأن احتمالية تنفيذ العلاقات فيه أكبر من أي تركيب آخر.

- **قواعد البيانات الهيكلية**

قواعد البيانات الهيكلية هي التي تعتمد التركيب الهيكلي بين عناصر البيانات. وهو اعتماد علاقة الهيكل التنظيمي بين عناصر البيانات.

مثال ذلك أن يكون عنصرين مصنفين تحت عنصر واحد أو تابعين له.

- **قواعد البيانات الهرمية**

وهي قواعد البيانات التي تعتمد علاقة التركيب الهرمي بين عناصر البيانات.

مثال ذلك أن يكون كل عنصر مسؤول عن عنصر واحد فقط وليس أكثر، بالترتيب من أعلى إلى أسفل، وهكذا.

- **قواعد البيانات الضخمة**

عند النظر إلى أنواع قواعد البيانات من منظور الحجم، يمكن تقسيم قواعد البيانات إلى:

- قواعد بيانات بسيطة تخدم مؤسسة أو شركة محددة

- قواعد بيانات ضخمة أو مراكز البيانات

وقواعد البيانات الضخمة أو مراكز البيانات هي عبارة عن مباني ضخمة تتميز بأنها مجهزة بالإعدادات المطلوبة لقواعد البيانات من ناحية المساحة التشغيلية والمصادر البشرية ومتخصصي الصيانة والطقس الملائم من خلال المكيفات، ولها عدة مولدات وقائية للطاقة ويتم استخدامها لتخزين وإدارة الكميات الضخمة من البيانات.

ومن أمثلة مراكز البيانات تلك التي تملكها شركات التكنولوجيا العملاقة مثل شركة جوجل وفيسبوك وغيرها.

- **مستودعات البيانات**

يوجد أيضًا نوع من أنواع قواعد البيانات التي يتم تجميعها من عدة مصادر وهي مستودعات البيانات Data Warehouses.

ومستودع البيانات عبارة عن تجميع للبيانات التاريخية من قواعد البيانات التقليدية من عدة مصادر لاستخراج تقارير وتحاليل تفيد في أداء عمل المؤسسات والشركات.

ومن أمثلة هذا النوع تلك المستخدمة في المقر الرئيسي لشركة لها عدة أفرع متواجدة في عدة دول أو مدن مختلفة.

خصائص قاعدة البيانات

تحتوي قاعدة البيانات على الخصائص التالية:

- إنه تمثيل لبعض جوانب العالم الحقيقي أو مجموعة من عناصر البيانات (الحقائق) التي تمثل معلومات العالم الحقيقي.
- قاعدة البيانات منطقية ومتماسكة ومتسقة داخلياً.
- تم تصميم قاعدة البيانات وبناءها وملؤها بالبيانات لغرض معين.
- يتم تخزين كل عنصر بيانات في حقل.
- تشكل مجموعة الحقول جدولاً. على سبيل المثال، يحتوي كل حقل في جدول الموظف على بيانات حول الموظف الفردي.
- يمكن أن تحتوي قاعدة البيانات على العديد من الجداول. على سبيل المثال، قد يحتوي نظام العضوية على جدول عنوان وجدول عضو فردي.

نظام إدارة قواعد البيانات

نظام إدارة قواعد البيانات Database Management System ويُرمز له بالاختصار DBMS، هو عبارة عن مجموعة من البرامج التي تُمكّن المستخدمين من إنشاء قواعد البيانات وصيانتها والتحكم في الوصول إليها. الهدف الأساسي لنظام إدارة قواعد البيانات DBMS هو توفير بيئة ملائمة وفعّالة للمستخدمين لاسترداد المعلومات وتخزينها أو استرجاعها أو الإضافة

أو التعديل عليها أو حذفها، حيث يقوم البرنامج بالربط بين المستخدم وبين محرك قاعدة البيانات، لأداء تلك المهمة.

وفى حال وجود علاقة بين جداول قاعدة البيانات يسمى هذا بنظام إدارة قواعد البيانات العلائقية أو نظام إدارة قواعد البيانات الارتباطية.

باستخدام نهج قاعدة البيانات، مثلاً، يمكننا الحصول على نظام إدارة شئون الموظفين في أي شركة أو مؤسسة. يتم استخدام نظام إدارة قواعد البيانات DBMS من قبل قسم شؤون الموظفين والإدارة العليا وقسم الحسابات للوصول إلى قاعدة البيانات المشتركة وتسهيل عمل كل قسم.

بيئات قواعد البيانات

بيئة قواعد البيانات هي البيئة البرمجية أو النظام الذي يوفر الأدوات والبرمجيات اللازمة لتصميم وإدارة قواعد البيانات.

وتوجد العديد من بيئات قواعد البيانات التي يتم تطويرها من قبل شركات عالمية متخصصة تتنافس فيما بينها لتوفير أفضل الأدوات والأساليب اللازمة لإدارة قواعد البيانات.

ومن أشهر بيئات قواعد البيانات:

- نظام أوراكل لإدارة قواعد البيانات العلائقيةOracle
- ماي إس كيو إلMy SQL

- فيربيرد Firebird
- مايكروسوفت أكسس Microsoft Access
- بوستجرس Postgress
- قاعدة بيانات بيركلي Birckly
- بورلاند إنتربيز Borland Interbase
- ميكروسوفت إس كيو إل سيرفر Microsoft SQL Server
- إنفورميكس Informix
- بي تريف B-trieve
- آي بي إم دي بي٢ IBMDB2
- سايبيز Sybase

فوائد قواعد البيانات

إدارة البيانات تعني الاهتمام بها حتى تعمل من أجلنا وتكون مفيدة للمهام التي نقوم بها. باستخدام نظام إدارة قواعد البيانات (DBMS) ، لم تعد المعلومات التي يتم جمعها وإضافتها إلى قواعد البيانات عرضة للتشويش العَرَضي، كما أنها تصبح أكثر سهولة وتكاملًا مع بقية العمل أيًا كان نوعه.

تسمح لنا إدارة البيانات باستخدام قواعد البيانات بأن نصبح مستخدمين

استراتيجيين للبيانات التي لدينا. وغالبًا ما نحتاج إلى الوصول إلى البيانات وإعادة ترتيبها للاستخدامات المختلفة. والتي قد تشمل:

- إنشاء قوائم بريدية
- كتابة تقارير الإدارة
- توليد قوائم قصص إخبارية مختارة
- تحديد احتياجات العملاء المختلفة

كما تسمح قوة المعالجة التي توفرها قواعد البيانات بمعالجة البيانات التي تحتويها، بحيث يمكنها أن تقوم بعمليات مثل:

- الفرز Sort
- الترتيب Arrange
- إنشاء حلقات الوصل Match
- إنشاء ارتباطات Link
- تخطي حقول معينة Skip fields
- إجراء الحسابات على الحقول Calculate

ونظراً لتعدد استخدامات قواعد البيانات، نجدها تعمل على تشغيل جميع أنواع المشاريع.

أمثلة استخدامات قواعد البيانات

- تطوير موقع على شبكة الإنترنت لحفظ المستخدمين المسجلين في موقع متخصص للبيع عبر الإنترنت
- تطوير تطبيق لأرشفة ومتابعة المستفيدين من الخدمات الاجتماعية
- تطوير نظام السجلات الطبية لمرافق الرعاية الصحية
- إنشاء دفتر عناوين البريد الإلكتروني للزبائن
- تجميع مجموعة من الوثائق المرتبطة بكلمات معينة
- تطوير نظام يصدر حجوزات الطيران في شركة طيران
- تطوير نظام مصرفي في البنوك
- تطوير نظام إدارة مدرسية
- وغيرها من النظم الأخرى.

وتوفر قواعد البيانات المصممة بشكل صحيح إمكانية الوصول إلى المعلومات المحدثة الدقيقة، لأن التصميم الصحيح يُعد ضروريًا لتحقيق الأهداف في العمل مع قواعد البيانات. كما أن التصميم المنطقي الجيد لقواعد البيانات يساهم في أن يلبي الاحتياجات في الشركة أو المؤسسة ويساعد الإدارة في اتخاذ القرارات المناسبة في الوقت المناسب.

ففي إحدى شركات بيع التجزئة، لو افترضنا أن إدارة الشركة تسعى إلى

تطوير آلية التسويق لمنتج معين، فإنها تلجأ من أجل تحقيق هذا الهدف إلى تحليل البيانات الخاصة بالمبيعات التاريخية لهذا المنتج، والتي يتم تخزينها في قاعدة بيانات المبيعات. وبهذه الطريقة يمكن للإدارة أن تقوم بتحليل تلك البيانات واستنتاج الطرق والأساليب المناسبة لتطوير خطط التسويق لهذا المنتج. أي أن الإدارة تقوم بدراسة البيانات التاريخية لحركة مبيعات هذا المنتج المتوفرة في قاعدة بيانات المبيعات من أجل وضع الخطط المستقبلية للتسويق له.

مميزات قواعد البيانات

هناك عدد من الخصائص التي تميّز نهج قواعد البيانات عن نظم إدارة الملفات أو النهج القائم على الملفات.

وفيما يلي وصفًا موجزًا لأهم مميزات قواعد البيانات:

الوصف الذاتي لنظام قواعد البيانات

يُشار إلى نظم قواعد البيانات بالوصف الذاتي لأنها لا تحتوي فقط على قاعدة البيانات نفسها، ولكن أيضًا البيانات الوصفية التي تحدد وتصف البيانات والعلاقات بين الجداول في قاعدة البيانات. يتم استخدام هذه المعلومات بواسطة برنامج نظام إدارة قواعد البيانات DBMS أو مستخدمي قاعدة البيانات إذا لزم الأمر. هذا الفصل بين البيانات والمعلومات حول

البيانات يجعل نظام قاعدة البيانات مختلفًا تمامًا عن النظام التقليدي القائم على الملفات حيث يكون تعريف البيانات جزءًا من برامج التطبيق.

العزل بين البرنامج والبيانات

في النظام القائم على إدارة الملفات، يتم تعريف بنية ملفات البيانات في برامج التطبيق، لذلك إذا أراد المستخدم تغيير بنية الملف، فقد تحتاج جميع البرامج التي تصل إلى هذا الملف إلى تغيير أيضًا.

أما في نهج قواعد البيانات، يتم تخزين بنية البيانات في كتالوج النظام وليس في البرامج. لذلك، هناك تغيير واحد هو كل ما يلزم لتغيير هيكل الملف. هذا العزل بين البرامج والبيانات يسمى أيضًا استقلالية بيانات البرنامج.

دعم لطرق عرض متعددة

تدعم قواعد البيانات طرق عرض متعددة للبيانات. العرض عبارة عن مجموعة فرعية من قاعدة البيانات، والتي يتم تعريفها وتخصيصها لمستخدمين معينين للنظام. قد يكون للعديد من المستخدمين في النظام وجهات نظر مختلفة للنظام. قد يحتوي كل عرض فقط على البيانات التي تهم مستخدمًا أو مجموعة من المستخدمين.

تقاسم البيانات ونظام متعدد المستخدمين

تم تصميم أنظمة قواعد البيانات الحالية بطريقة تدعم تعدد المستخدمين.

أي أنها تسمح للعديد من المستخدمين بالوصول إلى نفس قاعدة البيانات في نفس الوقت. يتم تحقيق هذا الوصول من خلال ميزات تسمى استراتيجيات التحكم في التزامن. تضمن هذه الاستراتيجيات أن البيانات التي يتم الوصول إليها صحيحة دائمًا وأن تكامل البيانات يتم الحفاظ عليه.

يُعد تصميم أنظمة قواعد البيانات الحديثة متعددة المستخدمين تحسينًا كبيرًا عن تلك الموجودة في الماضي والتي اقتصر استخدامها على شخص واحد في كل مرة.

التحكم في تكرار البيانات

في نهج قواعد البيانات، من الناحية المثالية، يتم تخزين كل عنصر بيانات في مكان واحد فقط في قاعدة البيانات. في بعض الحالات، لا يزال تكرار البيانات موجودًا لتحسين أداء النظام، ولكن يتم التحكم في هذه التكرارات من خلال برمجة التطبيقات ويتم الاحتفاظ بها إلى الحد الأدنى من خلال تقديم أقل تكرار ممكن عند تصميم قاعدة البيانات.

تبادل البيانات

إن دمج جميع البيانات، بالنسبة للمؤسسة، داخل نظام قاعدة بيانات له مزايا عديدة، ومنها:

- أولاً، يسمح بمشاركة البيانات بين الموظفين وغيرهم ممن لديهم حق

الوصول إلى النظام.

- ثانيًا، يمنح المستخدمين القدرة على توليد معلومات أكثر من كمية معينة من البيانات أكثر مما يمكن تحقيقه بدون التكامل.

تطبيق قيود صحة إدخال البيانات

يجب أن توفر أنظمة إدارة قواعد البيانات القدرة على تحديد وإنفاذ قيود معينة لضمان صحة البيانات التي يتم إدخالها عند قيام المستخدمين بإدخال بيانات بحيث تكون صالحة وللحفاظ على تكامل البيانات. والقيد في قواعد البيانات هو تقييد أو قاعدة تُملي ما يمكن إدخاله أو تحريره في جدول مثل الرمز البريدي باستخدام تنسيق معين، أو إضافة مدينة صالحة في حقل المدينة.

هناك العديد من أنواع قيود قواعد البيانات. تُحدد أنواع البيانات، مثلاً، نوع البيانات المسموح بها في الحقل، على سبيل المثال القيد الذي يسمح بإدخال الأرقام فقط في حقل مخصص للأرقام. يضمن تفريد البيانات مثل المفتاح الأساسي عدم إدخال أي نسخ مكررة من السجلات. يمكن أن تكون القيود بسيطة (قائمة على قاعدة بسيطة داخل تصميم الجدول) أو معقدة (قائمة على جمل برمجية معقدة).

منح الصلاحيات وتقييد الوصول غير المصرح به

لن يكون لجميع مستخدمي نظام قاعدة بيانات معين نفس امتيازات الوصول. على سبيل المثال، قد يمتلك مستخدم واحد حق الوصول للقراءة فقط (أي القدرة على قراءة ملف دون إجراء تغييرات)، بينما قد يمتلك مستخدم آخر امتيازات القراءة والكتابة، وهي القدرة على قراءة وتعديل ملف. لهذا السبب، يجب أن يوفر نظام إدارة قاعدة البيانات نظامًا فرعيًا للأمان لإنشاء أنواع مختلفة من حسابات المستخدمين والتحكم فيها ومنح الصلاحيات المختلفة وتقييد الوصول غير المصرح به.

استقلالية البيانات

ميزة أخرى لنظم إدارة قواعد البيانات هي القدرة على السماح باستقلالية البيانات. بمعنى آخر، يتم فصل أوصاف بيانات النظام أو البيانات التي تصف البيانات (البيانات الوصفية) عن البرامج التطبيقية. هذا ممكن لأن التغييرات في هيكل البيانات يتم معالجتها بواسطة نظام إدارة قاعدة البيانات وليست مدمجة في البرنامج نفسه.

تزامن المعالجة

يجب أن يشتمل نظام إدارة قاعدة البيانات على أنظمة فرعية للتحكم في التزامن. تضمن هذه الميزة بقاء البيانات متسقة وصالحة أثناء معالجة المعاملات حتى إذا قام العديد من المستخدمين بتحديث نفس المعلومات.

توفير لوجهات نظر متعددة للبيانات

يسمح نظام إدارة قواعد البيانات (DBMS) بحكم طبيعته للعديد من المستخدمين بالوصول إلى قاعدة البيانات الخاصة به إما بشكل فردي أو في وقت واحد. ليس من المهم أن يكون المستخدمون على دراية بكيفية ومكان تخزين البيانات التي يصلون إليها.

وظائف النسخ الاحتياطي والاسترداد

النسخ الاحتياطي والاسترداد هي طرق تسمح بحماية البيانات من الضياع. توفر نظم قواعد البيانات عملية منفصلة، عن عملية النسخ الاحتياطي للشبكة، لنسخ البيانات احتياطيًا واستردادها. في حالة فشل محرك الأقراص الثابتة وتعذر الوصول إلى قاعدة البيانات المخزنة على محرك الأقراص الثابتة، فإن الطريقة الوحيدة لاسترداد قاعدة البيانات هي من نسخة احتياطية. إذا فشل نظام الكمبيوتر في منتصف عملية تحديث معقدة، فإن النظام الفرعي للاسترداد مسؤول عن التأكد من استعادة قاعدة البيانات إلى حالتها الأصلية. هاتان فائدتان إضافيتان لنظام إدارة قاعدة البيانات.

تطوير قاعدة البيانات

أحد الجوانب الأساسية لهندسة البرمجيات بشكل عام هو التقسيم الفرعي لعملية التطوير إلى سلسلة من المراحل أو الخطوات، يُركز كل منها على جانب

واحد من التطوير. يُشار أحيانًا إلى مجموعة هذه الخطوات باسم دورة حياة تطوير البرامج .(SDLC) ينتقل منتج البرنامج خلال دورة الحياة هذه (بشكل متكرر أحيانًا، حيث يتم تحسينه أو إعادة تطويره) حتى يتم إيقافه نهائيًا عن الاستخدام.

من الناحية المثالية، يمكن التحقق من صحة كل مرحلة في دورة حياة التطوير قبل الانتقال إلى المرحلة التالية.

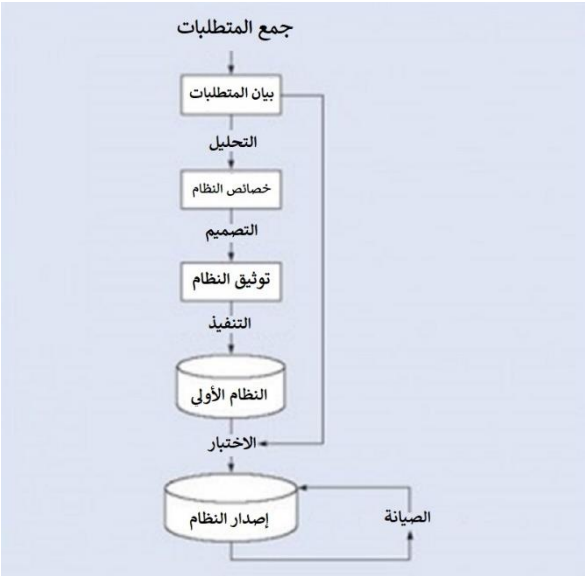
دورة حياة تطوير البرمجيات - نموذج الشلال

نموذج الشلال، هو النموذج المستخدم في معظم كتب هندسة البرمجيات.

ونموذج الشلال بشكل عام يمكن تطبيقه على أي عملية من عمليات تطوير نظام حاسوب. تظهر العملية كتسلسل صارم للخطوات حيث يكون ناتج خطوة واحدة هو المدخل إلى الخطوة التالية.

ويجب إكمال كل خطوة في النموذج قبل الانتقال إلى الخطوة التالية.

يمكن استخدام نموذج الشلال كوسيلة لتحديد المهام المطلوبة، جنبًا إلى جنب مع المدخلات والمخرجات لكل نشاط. المهم هو نطاق الأنشطة.



نموذج الشلال - تطوير قواعد البيانات

ويمكن تلخيص الخطوات التي يشتمل عليها نموذج الشلال على النحو التالي:

يتضمن تحديد المتطلبات التشاور مع أصحاب المصلحة والاتفاق معهم حول ما يريدون من النظام، معبرًا عنه كبيان للمتطلبات.

يبدأ التحليل من خلال النظر في بيان المتطلبات والانتهاء من خلال إنتاج مواصفات النظام. المواصفات هي تمثيل رسمي لما يجب أن يفعله النظام، ويتم التعبير عنه بعبارات مستقلة عن كيفية تحقيقه.

يبدأ التصميم بمواصفات النظام، وينتج وثائق التصميم ويقدم وصفا

مفصلا لكيفية إنشاء النظام.

التنفيذ هو بناء نظام كمبيوتر وفقًا لوثيقة تصميم معينة مع مراعاة البيئة التي سيعمل فيها النظام (على سبيل المثال، أجهزة أو برامج محددة متاحة للتطوير). قد يتم التنفيذ عادةً باستخدام نظام أولي يمكن التحقق من صحته واختباره قبل طرح نظام نهائي للاستخدام.

يقارن الاختبار بين النظام المُطبّق ومواصفات وثائق ومتطلبات التصميم ويُصدر تقرير قبول أو، في الغالب، قائمة بالأخطاء والأخطاء التي تتطلب مراجعة عمليات التحليل والتصميم والتنفيذ لتصحيحها (الاختبار هو عادةً المهمة التي تؤدي لنموذج الشلال الذي يتكرر خلال دورة الحياة).

تتضمن الصيانة التعامل مع التغييرات في المتطلبات أو بيئة التنفيذ، أو إصلاح الأخطاء أو نقل النظام إلى بيئات جديدة) على سبيل المثال، ترحيل نظام من حاسوب مستقل إلى محطة عمل UNIX أو بيئة شبكية .(نظراً لأن الصيانة تتضمن تحليل التغييرات المطلوبة، وتصميم الحل، وتنفيذ واختبار هذا الحل على مدى عمر نظام برمجيات تمت صيانته، بالتالي فإنه سيتم إعادة النظر في دورة حياة الشلال بشكل متكرر.

دورة حياة قاعدة البيانات

يمكننا استخدام نموذج الشلال كأساس لنموذج تطوير قاعدة البيانات، والذي يتضمن ثلاث افتراضات:

يمكننا فصل تطوير قاعدة البيانات - أي تحديد وإنشاء مخطط لتعريف البيانات في قاعدة البيانات - عن عمليات المستخدم الذي يستخدم قاعدة البيانات.

يمكننا استخدام بنية المخططات الثلاث كأساس لتمييز الأنشطة المرتبطة بالمخطط.

يمكننا تمثيل القيود المحددة على فرض دلالات البيانات مرة واحدة داخل قاعدة البيانات، بدلاً من تمثيلها بشكل منفرد لكل عملية من عمليات المستخدم الذي يستخدم البيانات.

باستخدام هذه الافتراضات مع نموذج الشلال، يمكننا أن نرى أن هذا النموذج التخطيطي يمثل نموذجًا للأنشطة ونواتجها لتطوير قاعدة البيانات. ينطبق على أي فئة من نظم إدارة قواعد البياناتDBMS ، وليس فقط قواعد البيانات العلائقية (أو قواعد البيانات ذات النهج العلائقي).

من الناحية العملية، فإن تطوير قواعد البيانات هو عملية الحصول على متطلبات العالم الحقيقي، وتحليل المتطلبات، وتصميم بيانات ووظائف النظام، ثم تنفيذ العمليات في النظام.

مراحل تطوير قواعد البيانات

تتألف عملية تطوير قواعد البيانات من المراحل التالية:

١. جمع المتطلبات
٢. التحليل
٣. التصميم المنطقي
٤. التنفيذ
٥. تحقيق التصميم
٦. نشر قاعدة البيانات

وفيما يلي شرحًا مفصلاً لكل مرحلة من هذه المراحل:

المرحلة الأولى: جمع المتطلبات

الخطوة الأولى هي جمع المتطلبات. خلال هذه الخطوة، يجب على مصممي قواعد البيانات مقابلة العملاء (مستخدمي قواعد البيانات) لفهم النظام المُقترح والحصول على البيانات والمتطلبات الوظيفية وتوثيقها. ونتيجة هذه الخطوة هي وثيقة تتضمن المتطلبات التفصيلية المقدمة من قبل المستخدمين.

يتضمن تحديد المتطلبات التشاور والاتفاق بين جميع المستخدمين بشأن البيانات الثابتة التي يريدون تخزينها جنبًا إلى جنب مع اتفاقية حول معنى وتفسير عناصر البيانات. يلعب مدير البيانات دورًا رئيسيًا في هذه العملية حيث يقدم نظرة عامة على الأعمال والقضايا القانونية والأخلاقية داخل المنظمة التي تؤثر على متطلبات البيانات.

يتم استخدام مستند متطلبات البيانات لتأكيد فهم المتطلبات مع المستخدمين. للتأكد من فهمها بسهولة، لا ينبغي أن تكون رسمية بشكل مفرط أو مشفرة للغاية. يجب أن يقدم المستند ملخصًا موجزًا لمتطلبات جميع المستخدمين - وليس مجرد مجموعة من متطلبات الأفراد - لأن الهدف هو تطوير قاعدة بيانات مشتركة واحدة.

يجب ألا تصف المتطلبات كيفية معالجة البيانات، بل ما هي عناصر البيانات، وما هي السمات التي تتمتع بها، والقيود التي تنطبق والعلاقات التي تربط بين عناصر البيانات.

المرحلة الثانية: التحليل

يبدأ تحليل البيانات ببيان متطلبات البيانات ثم يتم إنتاج نموذج بيانات مفاهيمي. الهدف من التحليل هو الحصول على وصف تفصيلي للبيانات التي تتناسب مع متطلبات المستخدم بحيث يتم التعامل مع كل من الخصائص العالية والمنخفضة للبيانات واستخدامها. وتشمل هذه الخصائص مثل النطاق المحتمل للقيم التي يمكن السماح بها للسمات (على سبيل المثال، في قاعدة بيانات المدرسة، رمز المقرر، عنوان المقرر والنقاط المقررة له).

يوفر نموذج البيانات المفاهيمية تمثيلًا رسميًا مشتركًا لما يتم توصيله بين العملاء والمطورين أثناء تطوير قاعدة البيانات، فهو يركز على البيانات في قاعدة البيانات، بغض النظر عن الاستخدام النهائي لتلك البيانات في عمليات

المستخدم أو تنفيذ البيانات في بيئات كمبيوتر محددة. لذلك، يهتم نموذج البيانات المفاهيمية بمعنى البيانات وهيكلها، ولكن ليس بالتفاصيل التي تؤثر على كيفية تنفيذها.

نموذج البيانات المفاهيمي هو تمثيل رسمي للبيانات التي يجب أن تحتوي عليها قاعدة البيانات والقيود التي يجب أن تفي بها البيانات. يجب التعبير عن ذلك بعبارات مستقلة عن كيفية تطبيق النموذج. ونتيجة لذلك، يركز التحليل على الأسئلة التي من النوع: "ما هو المطلوب؟" وليس التي من النوع: "كيف يتحقق ذلك؟."

المرحلة الثالثة: التصميم المنطقي

يبدأ تصميم قواعد البيانات بنموذج بيانات مفاهيمي وينتج مواصفات لمخطط منطقي؛ سيحدد هذا نوع نظام قاعدة البيانات المطلوب وما إذا كانت قاعدة بيانات الشبكية أو علائقية أو كائنية.

يمكننا استخدام تمثيل علائقي لنموذج البيانات المفاهيمية كمدخل لعملية التصميم المنطقي.

مخرجات هذه المرحلة هي مواصفات علائقية مفصلة، مخطط منطقي، لجميع الجداول والقيود اللازمة لتلبية وصف البيانات في نموذج البيانات المفاهيمي. خلال هذا النشاط التصميمي، يتم تحديد الخيارات الأكثر ملاءمة لتمثيل البيانات في قواعد البيانات. يجب أن تأخذ هذه الاختيارات

في الاعتبار معايير التصميم المختلفة بما في ذلك، على سبيل المثال، مرونة التغيير، والتحكم في الازدواجية وأفضل طريقة لتمثيل القيود.

إن الجداول التي يحددها المخطط المنطقي هي التي تحدد البيانات المخزنة وكيف يمكن معالجتها في قاعدة البيانات.

قد يميل مصممو قواعد البيانات المطلعون على قواعد البيانات العلائقية ولغة الاستعلام الهيكلية SQL إلى الذهاب مباشرة إلى التنفيذ بعد إنتاج نموذج بيانات مفاهيمي. ومع ذلك، فإن مثل هذا التحول المباشر للتمثيل العلائقي إلى جداول لغة الاستعلام الهيكلية SQL لا يؤدي بالضرورة إلى قاعدة بيانات تحتوي على جميع الخصائص المطلوبة وهي:

خصائص التصميم المنطقي لقاعدة البيانات

- الاكتمال
- التكامل
- المرونة
- الكفاءة
- قابلية الاستخدام

يُعد نموذج البيانات المفاهيمي الجيد خطوة أولى أساسية نحو قاعدة بيانات بها هذه الخصائص، ولكن هذا لا يعني أن التحويل المباشر إلى جداول SQL ينتج تلقائيًا قواعد بيانات جيدة. ستمثل هذه الخطوة الأولى

بدقة الجداول والقيود اللازمة لاستيفاء وصف نموذج البيانات المفاهيمي، وبالتالي ستلبي متطلبات الاكتمال والتكامل، ولكنها قد تكون غير مرنة أو توفر قابلية استخدام ضعيفة. ثم يتم ثني التصميم الأول لتحسين جودة تصميم قاعدة البيانات.

والثناء هو مصطلح يهدف إلى التقاط الأفكار المتزامنة لثني شيء ما لغرض مختلف وإضعاف جوانبه أثناء ثنيه.

والخطوات التكرارية (المتكررة) التي ينطوي عليها تصميم قاعدة البيانات الغرض الرئيسي منها هو التمييز بين المهمة العامة للجداول التي يجب استخدامها عن المهام التفصيلية للأجزاء المكونة لكل جدول، وتعتبر هذه الجداول واحدة في كل مرة، على الرغم من أنها ليست مستقلة عن بعضها البعض. كل تكرار يتضمن مراجعة لجداول سيؤدي إلى تصميم جديد؛ يُشار إلى هذه التصميمات الجديدة إجمالاً باسم تصميمات القطع الثاني، حتى إذا تكررت العملية لأكثر من حلقة واحدة.

استراتيجيات التصميم المنطقي لقاعدة البيانات

أولاً: بالنسبة لنموذج بيانات مفاهيمي معين، ليس من الضروري تلبية جميع متطلبات المستخدم التي تمثلها من خلال قاعدة بيانات واحدة. يمكن أن تكون هناك أسباب مختلفة لتطوير أكثر من قاعدة بيانات واحدة، مثل الحاجة إلى التشغيل المستقل في مواقع مختلفة أو سيطرة الإدارات على

البيانات "الخاصة بهم". ومع ذلك، إذا كانت مجموعة قواعد البيانات تحتوي على بيانات مكررة، ويحتاج المستخدمون إلى الوصول إلى البيانات في أكثر من قاعدة بيانات واحدة، فهناك أسباب محتملة لإمكانية تلبية قاعدة بيانات واحدة لمتطلبات متعددة، أو الحاجة إلى فحص المشكلات المتعلقة بتكرار البيانات وتوزيعها.

ثانيًا: أحد الافتراضات حول تطوير قواعد البيانات هو أنه يمكننا فصل تطوير قاعدة البيانات عن تطوير عمليات المستخدم التي تستخدمها. ويستند هذا إلى توقع أنه بمجرد تنفيذ قاعدة البيانات، يتم تحديد جميع البيانات التي تتطلبها عمليات المستخدم المحددة حاليًا ويمكن الوصول إليها؛ لكننا نحتاج أيضًا إلى المرونة للسماح بتلبية التغييرات المستقبلية للمتطلبات. عند تطوير قاعدة بيانات لبعض التطبيقات، قد يكون من الممكن التنبؤ بالطلبات الشائعة التي سيتم تقديمها إلى قاعدة البيانات حتى نتمكن من تحسين تصميمنا للطلبات الأكثر شيوعًا.

ثالثًا: على المستوى التفصيلي، تعتمد العديد من جوانب تصميم قواعد البيانات وتنفيذها على نظام إدارة قواعد البيانات (DBMS) المستخدم. إذا تم إصلاح اختيار DBMS أو تم إجراؤه قبل مهمة التصميم فيمكن استخدام هذا الاختيار لتحديد معايير التصميم بدلاً من الانتظار حتى التنفيذ. أي أنه من الممكن دمج قرارات التصميم لنظام إدارة قواعد بيانات DBMS معين بدلاً من إنتاج تصميم عام ثم تخصيصه لنظام إدارة قواعد البيانات DBMS

أثناء التنفيذ.

أولويات التصميم المنطقي لقاعدة البيانات

ليس من غير المألوف أن تجد أن تصميمًا واحدًا لا يمكنه تلبية جميع خصائص قاعدة البيانات الجيدة في وقت واحد.

لذلك من المهم أن المصمم قد أعطى الأولوية لهذه الخصائص (عادة باستخدام معلومات من مواصفات المتطلبات).

على سبيل المثال، لتحديد ما إذا كانت النزاهة أكثر أهمية من الكفاءة وما إذا كانت قابلية الاستخدام أكثر أهمية من المرونة في تطوير نظام قاعدة بيانات معين. في نهاية مرحلة التصميم، سيتم تحديد المخطط المنطقي من خلال عبارات لغة تعريف بيانات SQL DDL، والتي تصف قاعدة البيانات التي يجب تنفيذها لتلبية متطلبات المستخدم.

المرحلة الرابعة: التنفيذ

ينطوي التنفيذ على بناء قاعدة بيانات وفقًا لمواصفات المخطط المنطقي.

يشمل ذلك تحديد مخطط تخزين مناسب وإنفاذ الأمان ومخطط خارجي وما إلى ذلك. يتأثر التنفيذ بشدة باختيار نظم إدارة قواعد البيانات المتاحة وأدوات قواعد البيانات وبيئة التشغيل. هناك مهام إضافية بخلاف مجرد إنشاء مخطط قاعدة بيانات وتنفيذ القيود - يجب إدخال البيانات في الجداول،

ويجب معالجة المشكلات المتعلقة بالمستخدمين وعمليات المستخدمين، ويجب دعم أنشطة الإدارة المرتبطة بالجوانب الأوسع لإدارة بيانات الشركة. وتمشيا مع نهج نظم إدارة قواعد البيانات DBMS، ينبغي معالجة أكبر عدد ممكن من هذه المخاوف داخل النظام.

وسوف يتم التطرق إلى بعض هذه المخاوف بإيجاز فيما يلي:

عمليًا، يتطلب تنفيذ المخطط المنطقي في نظام إدارة قواعد بيانات معين معرفة تفصيلية جدًا بالميزات والتسهيلات المحددة التي يجب أن يقدمها نظام إدارة قواعد البيانات. في عالم مثالي، وتماشيًا مع الممارسة الجيدة لهندسة البرمجيات، فإن المرحلة الأولى من التنفيذ ستشمل مطابقة متطلبات التصميم مع أفضل أدوات التنفيذ المتاحة ثم استخدام تلك الأدوات للتنفيذ. من حيث قواعد البيانات، قد يتضمن هذا اختيار منتجات البائعين مع متغيرات نظم قواعد البيانات DBMS ولغة الاستعلام SQL الأكثر ملاءمة لقاعدة البيانات التي نحتاج إلى تنفيذها. ومع ذلك، نحن لا نعيش في عالم مثالي، وفي أغلب الأحيان سيتم اختيار الأجهزة والقرارات المتعلقة بنظام إدارة قواعد البيانات DBMS قبل النظر في تصميم قاعدة البيانات بوقت طويل. وبالتالي، يمكن أن ينطوي التنفيذ على مرونة إضافية للتصميم للتغلب على أي قيود على البرامج أو الأجهزة.

المرحلة الخامسة: تحقيق التصميم

بعد إنشاء التصميم المنطقي، نحتاج إلى إنشاء قاعدة بياناتنا وفقًا للتعريفات التي قمنا بإنتاجها. بالنسبة للتطبيق باستخدام نظام إدارة قواعد البيانات العلائقية أو الارتباطية، من المحتمل أن يتضمن هذا استخدام لغة الاستعلام الهيكلية SQL لإنشاء الجداول والقيود التي تُلبي وصف المخطط المنطقي واختيار مخطط التخزين المناسب (إذا كان نظام إدارة قواعد البيانات يسمح بهذا المستوى من التحكم).

إحدى طرق تحقيق ذلك هي كتابة عبارات SQL DDL المناسبة في ملف يمكن تنفيذه بواسطة نظام إجارة قواعد البياناتDBMS ، بحيث يكون هناك سجل مستقل، ملف نصي، لعبارات لغة الاستعلام الهيكلية SQL التي تحدد قاعدة البيانات. طريقة أخرى هي العمل بشكل تفاعلي باستخدام أداة قاعدة بيانات مثل SQL Server Management Studio أو Microsoft Access ومهما كانت الآلية المستخدمة لتنفيذ المخطط المنطقي، فإن النتيجة هي تحديد قاعدة البيانات، مع الجداول والقيود، ولكنها لن تحتوي على بيانات لعمليات المستخدم.

المرحلة السادسة: نشر قاعدة البيانات

بعد إنشاء قاعدة البيانات، هناك طريقتان لملء الجداول:

- إما: من البيانات الموجودة.

- أو: من خلال استخدام تطبيقات المستخدم المطورة لقاعدة البيانات.

بالنسبة لبعض الجداول، قد تكون هناك بيانات موجودة من قاعدة بيانات أو ملفات بيانات أخرى.

على سبيل المثال، عند إنشاء قاعدة بيانات لمستشفى، يمكن توقع أن هناك بالفعل بعض السجلات لجميع الموظفين التي يجب إدراجها في قاعدة البيانات. يمكن أيضًا إحضار البيانات من وكالة خارجية (غالبًا ما يتم إحضار قوائم العناوين من الشركات الخارجية) أو يتم إنتاجها أثناء مهمة إدخال بيانات كبيرة (يمكن أن يتم تحويل السجلات اليدوية إلى ملفات الكمبيوتر بواسطة وكالة متخصصة في إدخال البيانات). في مثل هذه الحالات، فإن أبسط طريقة لتعبئة قاعدة البيانات هي استخدام وظائف الاستيراد والتصدير الموجودة في نظام إدارة قواعد البيانات.

تتوفر عادةً تسهيلات لاستيراد البيانات وتصديرها بتنسيقات قياسية مختلفة (تُعرف هذه الوظائف أيضًا في بعض الأنظمة باسم تحميل البيانات وتفريغها). يتيح الاستيراد إمكانية نسخ ملف البيانات مباشرة إلى جدول. عندما يتم الاحتفاظ بالبيانات بتنسيق ملف غير مناسب لاستخدام وظيفة الاستيراد، فمن الضروري إعداد برنامج تطبيقي يقرأ في البيانات القديمة، ويحولها حسب الضرورة، ثم يدرجها في قاعدة البيانات باستخدام شفرة أو كود لغة الاستعلام الهيكلية SQL الذي تم إنتاجه خصيصًا من اجل ذلك الهدف. يُشار إلى نقل كميات كبيرة من البيانات الحالية إلى قاعدة بيانات

على أنها حمل ضخم. قد يتضمن التحميل المجمَّع للبيانات كميات كبيرة جدًا من البيانات التي يتم تحميلها، جدول واحد في كل مرة حتى تجد أن هناك تسهيلات نظام إدارة قاعدة البيانات DBMS لتأجيل التحقق من القيد حتى نهاية التحميل المجمع.

إرشادات تطوير مخطط تصميم قاعدة البيانات

ملاحظة: هذه إرشادات عامة ستساعد في تطوير أساس قوي لتصميم قواعد البيانات الفعلية (النموذج المنطقي).

- توثيق جميع الكيانات المكتشفة خلال مرحلة جمع المعلومات.
- توثيق جميع السمات التي تنتمي إلى كل كيان. تحديد المرشح والمفاتيح الأساسية. التأكد من أن جميع السمات غير الرئيسية لكل كيان تعتمد بشكل كامل على المفتاح الأساسي.
- تطوير مخطط أولي لمراجعة التقارير الإلكترونية ومراجعته مع الموظفين المناسبين. (هذه عملية تكرارية).
- إنشاء كيانات (جداول) جديدة للسمات متعددة القيم والمجموعات المتكررة. ودمج هذه الكيانات (الجداول) الجديدة في مخطط التقارير الإلكترونية. والمراجعة مع الموظفين المناسبين.
- التحقق من نمذجة التقارير الإلكترونية من خلال تطبيع الجداول.

مصطلحات قاعدة البيانات

الجداول

تُنظم قاعدة البيانات المعلومات في أوعية مخصصة تُسمى الجداول، وهي قوائم تحتوي على صفوف وأعمدة تشبه جدول البيانات.

وقد يكون لدينا جدول واحد فقط في قاعدة بيانات بسيطة، ولكن يلزم عادة أكثر من جدول واحد في معظم قواعد البيانات.

مثلاً في قاعدة بيانات إحدى الشركات، قد يلزم جدول يحتوي على معلومات حول المنتجات وجدول آخر يحتوي على معلومات عن أوامر الشراء وجدول آخر يحتوي على معلومات حول العملاء.

السجلات والحقول

في الجداول، يُسمى كل صف سجل ويُسمى كل عمود حقل.

ويُعتبر السجل طريقة متناسقة ذات معنى لجمع معلومات حول شيء ما.

بينما يُعتبر الحقل عنصرًا واحدًا من المعلومات، هو نوع العنصر الذي يظهر في كل سجل.

مثلاً، في جدول المنتجات، يمكن أن يحتوي كل صف أو سجل على معلومات حول منتج واحد، ويحتوي كل عمود أو حقل في هذا السجل على نوع من المعلومات حول هذا المنتج، مثل الاسم أو السعر.

المفتاح الأساسي

المفتاح الأساسي في أي جدول هو عمود يتم استخدامه في تعريف كل صف بشكل فريد، أي بشكل لا يقبل التكرار.

ومن أمثلة المفتاح الأساسي: رقم تعريف الموظف، أو رقم بطاقة الهوية الشخصية، أو في جدول المنتجات ممكن أن يكون "معرّف المنتج."

التصميم الجيد لقاعدة البيانات

توجد بعض المبادئ الخاصة بتوجيه عملية تصميم قاعدة البيانات حتى تتم بشكل جيد. ومن هذه المبادئ:

المبدأ الأول: تجنب تكرار البيانات

ويعني هذا المبدأ تجنب تكرار البيانات حتى لا تحتوي قاعدة البيانات على بيانات زائدة، فالبيانات الزائدة غير صالحة لأنها تهدر المساحة وتزيد من احتمالية الأخطاء وحالات عدم التناسق.

المبدأ الثاني: صحة ودقة المعلومات وتكاملها

وهذا مبدأ مهم لتصميم قاعدة بيانات جيدة، فإذا كانت قاعدة البيانات تحتوي على معلومات غير صحيحة، فإن أي تقارير تقوم بسحب المعلومات من قاعدة البيانات ستحتوي أيضًا على معلومات غير صحيحة. كنتيجة لذلك، فإن أي قرارات تتخذها الإدارة تستند إلى هذه التقارير سوف تكون مضللة، أي

أنها مبنية على معلومات خاطئة.

وبالتالي فإن من سمات التصميم الجيد لقاعدة البيانات يستلزم:

- تقسيم المعلومات في جداول لها عناوين لتقليل البيانات المكررة.
- ضم المعلومات في الجداول معًا، كلما لزم الأمر.
- ضمان دقة وصحة البيانات وتكاملها.

خطوات تصميم قاعدة البيانات

تتألف خطوات تصميم قاعدة البيانات مما يلي:

١. تحديد الغرض من قاعدة البيانات
٢. تحديد المعلومات المطلوبة وتنظيمها
٣. تقسيم المعلومات في جداول
٤. تحويل عناصر المعلومات إلى أعمدة
٥. تحديد المفاتيح الأساسية
٦. إنشاء العلاقات بين الجداول
   - علاقة واحد إلى متعدد
   - علاقة متعدد إلى متعدد
   - علاقة واحد إلى واحد

وفيما يلي شرحًا مفصلاً لكل خطوة:

١. تحديد الغرض من قاعدة البيانات

إن تحديد الغرض من قاعدة البيانات يساعد في التحضير لعملية تصميم قاعدة البيانات بالطريقة المثلى.

ويُستحسن قبل بدء تصميم قاعدة البيانات أن نكتب الغرض من قاعدة البيانات على الورق.

فالغرض من الكتابة هو توضيح طريقة استخدام قاعدة البيانات وتحديد الخطوط العريضة لمستلزمات هذا الاستخدام.

بالنسبة لقاعدة بيانات صغيرة لشركة تستند إلى العمل من المنزل، مثلاً، قد نكتب شيئًا بسيطًا مثل "تحتفظ قاعدة بيانات العملاء بقائمة معلومات العميل لغرض يتعلق بإنتاج مراسلات البريد والتقارير". إذا كانت قاعدة البيانات أكثر تعقيدًا أو تُستخدم بواسطة العديد من الأشخاص، ويحدث ذلك غالبًا في الشركات والمؤسسات الكبرى، يمكن أن يكون الغرض فقرة أو أكثر ويجب أن يتضمن طريقة وتوقيت استخدام كل شخص لقاعدة البيانات. الفكرة هي أن يكون لدينا بيان بالمهمة المطلوب تنفيذها ويتم تطويره بشكل جيد. ويمكن أن يُشار إليه من خلال عملية التصميم. في الواقع، إن وجود بيان مثل هذا يساعد في التركيز على الأهداف عندما نقوم باتخاذ القرارات.

٢. تحديد المعلومات المطلوبة وتنظيمها

يتم في هذه الخطوة جمع كل أنواع المعلومات التي قد يلزم تسجيلها في قاعدة البيانات، مثل:

اسم المنتج، رقم الطلب، ... ، إلخ.

وحتى يتم البحث عن المعلومات المطلوبة وتنظيمها، نبدأ بالمعلومات الحالية.

على سبيل المثال، يمكن تسجيل أوامر الشراء في دفتر الأستاذ أو الاحتفاظ بمعلومات العميل في نماذج ورقية في خزانة الملفات. ثم نقوم بجمع المستندات وسرد كل نوع من أنواع المعلومات التي تُظهر، مثلاً، كل مربع نقوم بتعبئته في نموذج، إذا لم يكن لدينا أي نماذج موجودة. نفرض أنه يجب تصميم نموذج لتسجيل معلومات العميل، فما هي المعلومات التي نريد وضعها في النموذج؟ ما هي مربعات التعبئة التي نود إنشاءها؟ نقوم بتحديد كل عنصر من هذه العناصر وسرده. مثلاً، نفترض أننا نحتفظ حاليًا بقائمة العملاء على بطاقات الفهرس. قد يُظهر فحص هذه البطاقات أن كل بطاقة تحتوي على اسم العميل والعنوان والمدينة والدولة والرمز البريدي ورقم الهاتف. ويمثل كل من هذه العناصر عمودًا محتملاً في جدول.

أثناء إعدادنا لهذه القائمة، لا يعني ذلك ضرورة إنجازها على أكمل وجه من المرة الأولى. بدلاً من ذلك، يمكن سرد كل عنصر يتبادر إلى الذهن.

في حالة استخدام عدة أشخاص لقاعدة البيانات، يمكن طلب معرفة أفكارهم أيضًا. كما يمكن ضبط القائمة في وقت لاحق.

تنظيم البيانات وفقًا لاحتياجات المخرجات

بعد ذلك، يجب مراعاة أنواع التقارير أو المراسلات البريدية المطلوب إنشاؤها من قاعدة البيانات.

على سبيل المثال، قد نريد تقرير مبيعات المنتج لعرض المبيعات حسب المنطقة أو تقرير ملخص عن المخزون يعرض مستويات مخزون المنتج.

قد نرغب أيضًا في إنشاء رسائل نموذجية لإرسالها إلى العملاء للإعلان عن حدث بيع أو تقديم مكافأة. يتم تصميم التقرير وتخيل ما سيبدو عليه. ما هي المعلومات التي نريد وضعها في التقرير؟ نقوم بسرد كل عنصر. ثم ننفذ نفس الإجراء للرسالة النموذجية وأي تقرير آخر نرغب في إنشاءه.

يساعد تكوين فكرة حول التقارير ومراسلات البريد التي قد نريد إنشاءها على تحديد العناصر التي سنحتاج إليها في قاعدة البيانات.

على سبيل المثال، نفرض أننا منحنا فرصة للعملاء للاشتراك (أو إلغاء الاشتراك) في تحديثات البريد الإلكتروني الدورية، ونريد طباعة قائمة بأسماء الأشخاص الذين اشتركوا فيها. لتسجيل هذه المعلومات، يمكن إضافة عمود "إرسال بريد إلكتروني" إلى جدول العملاء. ويمكن تعيين الحقل إلى "نعم" أو "لا" لكل عميل.

يقترح إجراء إرسال رسائل البريد الإلكتروني إلى العملاء عنصرًا آخر يجب تسجيله. عندما نعلم أن العميل يريد تلقي رسائل البريد الإلكتروني، سنحتاج أيضًا إلى معرفة عنوان البريد الإلكتروني الخاص به لإرسال الرسائل عليه. وبالتالي فإننا بحاجة إلى تسجيل عنوان البريد الإلكتروني لكل عميل.

من المناسب إنشاء نموذج أولي لكل تقرير أو قائمة مخرجات، مع مراعاة العناصر التي سنحتاج إليها لإنشاء التقرير.

على سبيل المثال، عندما نقوم بفحص رسالة نموذجية، قد يتبادر إلى الذهن بعض الأمور. إذا كنا نريد تضمين تحية مناسبة، مثل الديباجة التي تبدأ بالتحية "السيد" أو "السيدة" أو "الآنسة"، يجب إنشاء عنصر تحية. ويمكن أيضًا أن نبدأ الرسالة بـ "عزيزي السيد/محمد" بدلاً من "عزيزي السيد/ محمد مصطفى". يوضح هذا أننا نريد إبقاء اسم العائلة منفصلاً عن الاسم الأول، حتى يكون بالإمكان استخدامه بشكل منفرد.

التقسيم إلى أجزاء أصغر

النقطة الأساسية التي يجب تذكرها هي تقسيم كل معلومة إلى أجزاء أصغر مفيدة.

مثلًا، فيما يخص الاسم، لتوفير اسم العائلة بسهولة، يمكن تقسيم الاسم إلى قسمين، الاسم الأول واسم العائلة. لفرز تقرير حسب اسم العائلة، على سبيل المثال، فمن المساعد أن يكون اسم عائلة العميل مخزن بشكل منفصل.

بشكل عام، إذا أردنا فرز تقرير أو البحث عنه أو حسابه أو تقديمه استنادًا إلى عنصر المعلومات، يجب وضع هذا العنصر في الحقل الخاص به.

إن التفكير في الأسئلة التي نريد أن نجد لها إجابات في قاعدة البيانات، مثل: كم عدد مبيعات المنتج المتميز التي قمنا بتنفيذها الشهر الماضي؟ أين يقيم أفضل عملائنا؟ من هو مورد المنتج الأكثر مبيعًا؟ فإن هذه الأسئلة ستساعدنا من البداية في تعريف المزيد من العناصر في قاعدة البيانات أو الحاجة إلى التقسيم إلى أجزاء أصغر.

بعد تجميع هذه المعلومات، سنكون جاهزين للخطوة التالية.

٣. تقسيم المعلومات في جداول

يتم في هذه الخطوة تقسيم عناصر المعلومات إلى عناوين أو وحدات رئيسية، تمثل كل وحدة عدد من المعلومات المرتبطة ببعضها البعض والمتعلقة بشأن معين، مثل وحدة أو عنوان خاص بالموظفين وعنوان خاص بالمنتجات وعنوان خاص بالطلبات، وهكذا، حيث يصبح كل عنوان جدولاً فيما بعد.

ولتقسيم المعلومات إلى جداول، يتم اختيار الوحدات أو العناوين الرئيسية.

مثلاً، إذا اشتملت عناصر المعلومات لدينا على بيانات تخص المنتجات والموردين والعملاء والطلبات. فإنه من المنطقي أن يتم البدء بهذه الجداول

الأربعة:

- الجدول الأول: معلومات حول المنتجات
- الجدول الثاني: معلومات حول الموردين
- الجدول الثالث: معلومات حول العملاء
- الجدول الرابع: معلومات حول الطلبات

على الرغم من أن هذا لا يُعبر عن كل قائمة الجداول التي يمكن إنشاءها، لكنه نقطة بداية جيدة.

ويمكن متابعة تحسين هذه الجداول حتى الحصول على تصميم يعمل بشكل جيد.

عند مراجعة قائمة العناصر الأولية المتوفرة في كل جدول، قد نفكر في وضعها في جدول واحد بدلاً من الأربعة جداول الموضحة في الرسم التوضيحي السابق.

ولكن سوف نتعلم في المثال التالي لماذا تُعد هذه فكرة سيئة.

مثال توضيحي

نفرض أنه تم تصميم جدول خاص بالطلبات، وفيه كافة البيانات الأولية من أكثر من جدول، كأن يكون كل صف فيه يحتوي على معلومات حول المنتج والمورد الخاص به. وحيث أنه يمكن الحصول على عدد كبير من

المنتجات من المورد نفسه وبالتالي يترتب على ذلك وجوب تكرار اسم المورد ومعلومات عن عنوانه عدة مرات. وهذا أمر مُهدر للوقت والجهد وكذلك مهدر لمساحة التخزين في القرص. لذا فإن تسجيل معلومات المورد مرة واحدة فقط في جدول موردين منفصل، ومن ثم ربط هذا الجدول بجدول "المنتجات"، يعد حلاً أفضل بكثير.

احتياجات التعديل

تتمثل المشكلة الثانية في عدم تقسيم المعلومات في جداول ووضعها في جدول واحد في نشوء مشكلة عندما نحتاج إلى تعديل معلومات معينة. مثلاً، نفرض أننا نحتاج إلى تغيير العنوان الخاص بمورد، ولأنه يظهر في عدة أماكن في نفس الجدول المجمّع، سوف نضطر إلى تغيير العنوان في كل تلك الأماكن، وقد نقوم بتغيير العنوان في مكان عن طريق الخطأ وننسى تغييره في الأماكن الأخرى، لذا فتسجيل عنوان المورد في مكان واحد فقط يقوم بحل هذه المشكلة.

عند تصميم قاعدة البيانات، ينبغي دائمًا تسجيل كل معلومة مرة واحدة فقط. إذا وجدنا أنفسنا نكرر المعلومات نفسها في أكثر من مكان، مثل عنوان مورد معين، فإنه يمكن القول بأن تصميم قاعدة البيانات غير جيد وبالتالي فإنه يتوجب وضع هذه المعلومات في جدول منفصل حتى يُصبح التصميم أكثر جودة وكفاءة.

احتياجات الحذف

وأخيرًا، لنفترض وجود منتج واحد فقط تقوم شركة معينة بتوفيره، ونريد حذفه مع الاحتفاظ بمعلومات العنوان واسم المورد. كيف نريد حذف سجل المنتج دون فقدان معلومات المورد؟ بالطبع لا يمكننا إجراء ذلك. لأن كل سجل يحتوي على معلومات حول المنتجات، وكذلك معلومات حول المورد، فلا يمكننا حذف واحدة دون حذف الأخرى. لإبقاء المعلومات التالية منفصلة، يجب أن نقوم بتقسيم الجدول إلى اثنين: جدول لمعلومات المنتج وآخر لمعلومات المورد. يجب أن يؤدي حذف سجل المنتج إلى حذف المعلومات المتعلقة بالمنتج فقط وليس المعلومات المتعلقة بالمورد.

عند اختيار العنوان الممثل بالجدول، يجب أن تحتوي الأعمدة على المعلومات ذات الصلة بالعنوان فقط.

مثلًا، يجب على جدول المنتجات تخزين المعلومات حول المنتجات فقط. لأن عنوان المورد هو معلومة تتعلق بالمورد، وليس معلومة حول المنتج، فهو ينتمي إلى جدول الموردين.

٤. تحويل عناصر المعلومات إلى أعمدة

في هذه الخطوة، يتم تحديد ما هي المعلومات التي نريد تخزينها في كل جدول. حيث يصبح كل عنصر حقلاً ويتم عرضه كعمود في الجدول.

مثلاً، قد يتضمن جدول الموظفين حقولاً مثل "اسم الموظف، "اسم العائلة"، "تاريخ التوظيف"، وهكذا.

لتحديد الأعمدة في جدول، يتم تحديد المعلومات التي نحتاجها لتفاصيل العنوان المسجل في الجدول.

مثلاً، بالنسبة لجدول "العملاء"، يمكن أن تكون أعمدة الجدول معبرة عن البيانات التفصيلية التالية:

- الاسم
- العنوان
- المدينة
- الولاية
- الرمز البريدي
- إرسال بريد إلكتروني
- التحية
- عنوان البريد الإلكتروني
- ..إلخ

يحتوي كل سجل في الجدول على مجموعة الأعمدة نفسها، حيث يمكن تخزين معلومات الاسم، العنوان، المدينة، الولاية، الرمز البريدي، وإرسال بريد إلكتروني، التحية، لكل سجل. على سبيل المثال، يتضمن عمود العنوان

عناوين العملاء. ويحتوي كل سجل على بيانات حول عميل واحد ويحتوي حقل "العنوان" على عنوان هذا العميل.

بعد تحديد مجموعة الأعمدة الأولى لكل جدول، يمكن تحسين الأعمدة.

مثلاً، يُفضّل تخزين اسم العميل كعمودين منفصلين: الاسم الأول واسم العائلة، حيث يمكن القيام بالفرز والبحث والفهرسة لتلك الأعمدة فقط. وبشكل مماثل، يتكون العنوان فعليًا من خمس مكونات منفصلة هي: العنوان والمدينة والولاية والرمز البريدي والبلد/المنطقة، حيث يعد ذلك مفيدًا لتخزينها في أعمدة منفصلة. إذا كنا نريد إجراء بحث أو عملية فرز أو تصفية حسب الولاية، على سبيل المثال، فإننا نحتاج إلى معلومات الولاية المخزنة في عمود منفصل.

يجب أيضًا مراعاة ما إذا كانت قاعدة البيانات تتضمن معلومات محلية فقط أو الدولية أيضًا.

مثلاً، إذا كنا نخطط لتخزين العناوين الدولية، يُفضّل أن يكون لدينا عمود المنطقة بدلاً من الولاية، حيث يمكن لمثل هذا العمود احتواء الولايات المحلية ومناطق البلدان/المناطق الأخرى.

وبشكل مماثل، قد يكون "الرمز البريدي" أفضل بشكل أكبر إذا كنا نريد تخزين العناوين الدولية.

فيما يلي بعض التلميحات المهمة لتحديد الأعمدة الخاصة بالجدول:

عدم تضمين البيانات المحسوبة

في معظم الحالات، يجب عدم تخزين نتيجة العمليات الحسابية في الجداول.

بدلًا من ذلك، يمكن الحصول على نتائج إجراء العمليات الحسابية عندما نريد رؤية نتيجة العملية.

مثلًا، نفترض وجود تقرير "منتجات تحت الطلب" الذي يعرض الإجمالي الفرعي للوحدات تحت الطلب لكل فئة من المنتج في قاعدة البيانات. ولكن لا توجد أي أعمدة للإجمالي الفرعي للوحدات تحت الطلب في أي جدول. بدلاً من ذلك، يتضمن جدول المنتجات عمود "وحدات تحت الطلب" الذي يقوم بتخزين الوحدات تحت الطلب لكل منتج. وباستخدام تلك البيانات، تقوم قاعدة البيانات بحساب الإجمالي الفرعي في كل مرة نقوم فيها بطباعة التقرير.

بمعنى أنه يجب عدم تخزين الإجمالي الفرعي نفسه في جدول.

تخزين المعلومات في أجزاء منطقية أصغر

قد نحتاج إلى حقل واحد للأسماء الكاملة أو لأسماء المنتجات إلى جانب أوصاف المنتج.

إذا قمنا بدمج أكثر من نوع واحد من المعلومات في الحقل فسوف يكون

من الصعب استرداد المعلومات الفردية في وقت لاحق.

والتصميم الأفضل هو تقسيم المعلومات إلى أجزاء منطقية.

مثلاً، يمكن إنشاء حقول منفصلة للاسم الأول واسم العائلة أو لاسم المنتج وفئته ووصفه.

عندما نقوم بتحسين أعمدة البيانات في كل جدول يُصبح الجدول جاهزاً للخطوة التالية وهي اختيار المفتاح الأساسي له.

٥. تحديد المفاتيح الأساسية

يتم في هذه الخطوة اختيار المفتاح الأساسي لكل جدول.

ويعتبر المفتاح الأساسي عموداً يتم استخدامه لتعريف كل صف بشكل فريد لا يقبل التكرار.

مثل:

- "معرف المنتج" أو
- "معرف الطلب"

ويجب أن يتضمن كل جدول عموداً أو مجموعة أعمدة تقوم بتعريف كل صف مخزن في الجدول بشكل فريد. يُعد هذا غالباً هو رقم المعرف الفريد، مثل رقم معرف موظف أو رقم تسلسلي. في مصطلحات قاعدة البيانات، تسمى هذه المعلومات المفتاح الأساسي للجدول. تستخدم بيئة قواعد البيانات

حقول المفتاح الأساسي لإقران البيانات إلكترونيًا بشكل سريع بين الجداول المتعددة وجمع البيانات معًا بدون الحاجة إلى تجميعها يدويًا.

إذا كان لدينا معرف فريد لجدول، مثل رقم المنتج الذي يحدد كل منتج في الكتالوج بشكل فريد، يمكن استخدام هذا المعرف كمفتاح أساسي للجدول، لكن فقط إذا كانت القيم في هذا العمود دائمًا مختلفة لكل سجل أي أنها لا تقبل التكرار لأنه لا يمكن وجود قيم مكررة في المفتاح الأساسي. على سبيل المثال، لا يمكن استخدام أسماء الأشخاص كمفتاح أساسي لأن الأسماء غير فريدة. يمكن بسهولة أن يكون لدينا شخصان بنفس الاسم في نفس الجدول.

قيمة المفاتيح الأساسية

يجب توفير قيمة دائمًا للمفتاح الأساسي.

إذا كانت قيمة العمود غير معينة أو غير معروفة (قيمة مفقودة) في بعض الأحيان، فلا يمكن استخدامها كمكون في المفتاح الأساسي.

كما يجب أن نقوم دائمًا باختيار مفتاح أساسي لن تتغير قيمته.

يمكن استخدام المفتاح الأساسي للجدول كمرجع في الجداول الأخرى في قاعدة بيانات التي تستخدم أكثر من جدول. إذا تم تغيير المفتاح الأساسي، يجب تطبيق التغيير أيضًا في أي مكان تتم فيه الإشارة إلى المفتاح. يؤدي استخدام مفتاح أساسي لن يتغير إلى تقليل احتمالية عدم مزامنته مع جداول

أخرى تشير إلي، أي مرتبطة معه بعلاقة.

غالبًا ما يتم استخدام رقم فريد عشوائي كمفتاح أساسي. على سبيل المثال، قد تقوم بتعيين رقم طلب فريد لكل طلب. الغرض الوحيد لرقم الطلب هو تحديد الطلب. وبمجرد تعيين هذا الرقم لا يمكن تغييره.

الترقيم التلقائي للمفاتيح الأساسية

إذا لم نضع في الاعتبار أي عمود أو مجموعة أعمدة لتحديده كمفتاح أساسي مناسب، فيجب استخدام عمود يحتوي على نوع البيانات "ترقيم تلقائي". عند استخدام نوع البيانات "ترقيم تلقائي"، يتم تلقائيًا توليد قيمة لكل سجل عبارة عن مسلسل تلقائي. هذا النوع من المعرف غير صحيح؛ فهو لا يحتوي على معلومات حقيقية تصف الصف الذي تمثله. تعد المعرفات غير الحقيقية مثالية للاستخدام كمفتاح أساسي لإنها لن تغيير. المفتاح الأساسي الذي يحتوي على معلومات حول الصف، على سبيل المثال رقم الهاتف أو اسم العميل، من المحتمل أن يتغير لأن المعلومات الحقيقية نفسها يمكن أن تتغير.

غالبًا ما يمكن استخدام العمود المعيّن لنوع البيانات "ترقيم تلقائي" كمفتاح أساسي جيد لأنه لا يوجد معرفان متماثلان للمنتج.

في بعض الحالات، قد نحتاج إلى استخدام حقلين أو أكثر معًا لتوفير المفتاح الأساسي للجدول.

مثلاً، يمكن لجدول "تفاصيل الطلب" الذي يقوم بتخزين عناصر السجل

للطلبات استخدام عمودين في المفتاح الأساسي الخاص به: معرف الطلب ومعرف المنتج.

عندما يستخدم المفتاح الأساسي أكثر من عمود، يسمى ذلك أيضًا بالمفتاح المركب.

بالنسبة لقاعدة بيانات مبيعات المنتج، يمكن إنشاء عمود "ترقيم تلقائي" لكل جدول ليعمل كمفتاح أساسي.

مثلاً، يمكن إنشاء معرف المنتج لجدول المنتجات ومعرف الطلب لجدول الطلبات ومعرف العميل لجدول العملاء ومعرف المورد لجدول الموردين.

٦. إنشاء العلاقات بين الجداول

يتم في هذه الخطوة مراجعة كل جدول وتحديد كيفية ارتباط البيانات في جدول ما بالبيانات الموجودة في الجداول الأخرى.

ويمكن إضافة الحقول إلى الجداول أو إنشاء جداول جديدة لتوضيح العلاقات إذا لزم الأمر.

مثلاً، يمكن الربط بين جدول المنتجات وجدول المبيعات من خلال حقل "معرف المنتج"، بحيث يظهر في جدول المبيعات كل حركات البيع التي تمت للمنتج بدلالة "معرف المنتج"، ولكن بدون تكرار بيانات المنتج كاملة في جدول المبيعات.

بعد أن قمنا بتقسيم المعلومات في جداول، نحتاج إلى طريقة لتجميع المعلومات مرة أخرى بطرق ذات معني.

على سبيل المثال، يتضمن أحد النماذج معلومات من عدة جداول، بحيث تأتي المعلومات الواردة فيه من الجدول التالية:

- جدول العملاء
- جدول الموظفين
- جدول الطلبات
- جدول المنتجات
- جدول تفاصيل الطلب

في نظام إدارة قواعد البيانات العلائقية أو الارتباطية، يمكن تقسيم المعلومات في جداول منفصلة، جداول قائمة على عناوين. ثم يمكن بعد ذلك استخدام علاقات الجداول لتربيط وجمع المعلومات معًا، إذا لزم الأمر.

وهناك ثلاثة أنواع من العلاقات في قواعد البيانات، وهي:

- علاقة واحد إلى متعدد
- علاقة متعدد إلى متعدد
- علاقة واحد إلى واحد

وفيما يلي شرحًا مفصلاً لطريقة إنشاء كل نوع من هذه الأنواع من العلاقات:

إنشاء علاقة واحد إلى متعدد

علاقة واحد إلى متعدد يكون فيها لكل سجل في الجدول الأول يوجد عدة سجلات مرتبطة به في الجدول الثاني.

مثلاً، في جداول المنتجات والموردين في قاعدة بيانات طلبات المنتج، من المعروف أنه يمكن للمورد توفير أي عدد من المنتجات.

ويترتب على ذلك وجود العديد من المنتجات في جدول المنتجات لأي مورد في جدول الموردين.

وبالتالي، تُعتبر العلاقة بين جدول الموردين وجدول المنتجات من نوع علاقة واحد إلى متعدد.

لتمثيل علاقة واحد إلى متعدد في تصميم قاعدة البيانات، نستخدم المفتاح الأساسي الموجود في الجانب "واحد" من العلاقة ثم إضافته كعمود إضافي أو أعمدة إضافية إلى الجدول الموجود في الجانب "متعدد" من العلاقة. في هذه الحالة، نضيف عمود معرف المورد من جدول الموردين إلى جدول المنتجات. بعد ذلك سيتم استخدام رقم معرّف العميل في جدول المنتجات لتحديد موقع المورد الصحيح لكل منتج بشكل آلي.

يسمى عمود معرف المورد في جدول المنتجات بالمفتاح الخارجي.

المفتاح الخارجي هو مفتاح أساسي لجدول آخر.

ويعتبر العمود "معرف المورد" في جدول المنتجات مفتاح خارجي لإنه يُعد أيضًا المفتاح الأساسي في جدول الموردين.

يمكن توفير الأساس للانضمام إلى الجداول المرتبطة عن طريق تأسيس مفاتيح أساسية ومفاتيح خارجية.

إذا لم نكن متأكدين من الجداول التي يجب أن تتشارك عمودًا عامًا، يضمن تحديد علاقة واحد إلى متعدد أن الجدولين المضمنين سيتطلبان عمودًا مشتركًا.

إنشاء علاقة متعدد إلى متعدد

علاقة متعدد إلى متعدد يكون فيها لكل سجل في الجدول الأول قد يوجد عدة سجلات مرتبطة به في الجدول الثاني.

والعكس صحيح، لكل سجل في الجدول الثاني قد يوجد عدة سجلات مرتبطة به في الجدول الأول.

مثلاً، العلاقة بين جدول المنتجات وجدول الطلبات. فقد يتضمن طلب واحد أكثر من منتج واحد.

وعلى الجانب الآخر، يمكن أن يظهر منتج واحد في عدة طلبات.

لذلك، يمكن أن يكون لكل سجل في جدول "الطلبات" عدة سجلات في جدول "المنتجات."

ويمكن أن يكون لكل سجل في جدول "المنتجات" عدة سجلات في جدول "الطلبات."

يسمى هذا النوع من العلاقة علاقة متعدد إلى متعدد لأنه يمكن أن يكون لأي منتج العديد من الطلبات ويمكن أن يكون لأي طلب العديد من المنتجات.

تجدر الإشارة إلى أنه من المهم أخذ جانبي العلاقة بالاعتبار، لاكتشاف علاقات متعدد إلى متعدد بين الجداول.

تتضمن عناوين الجدولين، الطلبات والمنتجات، علاقة متعدد إلى متعدد. يمثل ذلك مشكلة.

ولفهم المشكلة، لنتخيل ماذا سيحدث إذا حاولنا إنشاء علاقة بين جدولين بإضافة حقل معرف المنتج إلى جدول الطلبات. للحصول على أكثر من منتج لكل طلب، سنحتاج إلى أكثر من سجل في جدول الطلبات لكل طلب. يمكننا تكرار معلومات الطلب لكل صف مرتبط بطلب واحد، ما ينتج عنه تصميم غير كافٍ قد يؤدي إلى بيانات غير صحيحة. سنواجه نفس المشكلة إذا قمنا بوضع حقل معرف الطلب في جدول المنتجات، حيث سيكون لدينا أكثر من سجل في جدول المنتجات لكل منتج. كيف يمكن حل هذه المشكلة؟

حل مشكلة إنشاء علاقة متعدد إلى متعدد

الحل هو إنشاء جدول ثالث، يسمى غالبًا جدول "تفاصيل الطلب" حيث نقوم بتقسيم علاقة متعدد إلى متعدد إلى علاقتيّ واحد إلى متعدد.

ثم يمكن إدراج المفتاح الأساسي من كل من الجدولين في الجدول الثالث.

كنتيجة لذلك، يقوم الجدول الثالث بتسجيل كل تكرارات العلاقة أو مثيلاتها.

يمثل كل سجل في جدول "تفاصيل الطلب" عنصر سجلاً واحدًا في الطلب.

يتألف المفتاح الأساسي لجدول "تفاصيل الطلب" من حقلين المفاتيح الخارجية من جداول "الطلبات" و"المنتجات."

استخدام حقل معرف الطلب وحده لا يعمل كمفتاح أساسي لهذا الجدول لأنه يمكن أن يكون للطلب الواحد العديد من عناصر السجل. يتم تكرار معرف الطلب لكل عنصر سجل في الطلب، حيث لا يحتوي الحقل على قيم فريدة. واستخدام حقل معرف المنتج وحده لا يعمل أيضًا لأنه يمكن أن يظهر المنتج الواحد في العديد من الطلبات المختلفة.

ولكن عند دمج الحقلين معًا، ينتج عن ذلك دائمًا قيمة فريدة لكل سجل.

في قاعدة بيانات مبيعات المنتج، لا يرتبط جدول الطلبات وجدول

المنتجات ببعضها البعض مباشرة.

لكنهما يرتبطا بشكل غير مباشر عن طريق جدول تفاصيل الطلب.

يتم تمثيل علاقة متعدد إلى متعدد بين الطلبات والمنتجات في قاعدة البيانات باستخدام علاقات واحد إلى متعدد بحيث يكون:

يحتوي جدول "الطلبات" وجدول "تفاصيل الطلب" على علاقة واحد إلى متعدد. يمكن أن يكون لكل طلب أكثر من عنصر سجل، لكن كل عنصر سجل يتصل بطلب واحد فقط.

يحتوي جدول "المنتجات" وجدول "تفاصيل الطلب" على علاقة واحد إلى متعدد. قد يكون لكل منتج العديد من عناصر السجل المرتبطة به، لكن يشير كل عنصر سجل إلى منتج واحد فقط.

من جدول "تفاصيل الطلب"، يمكن تحديد كل المنتجات الموجودة في طلب محدد. ويمكن أيضًا تحديد كل الطلبات الخاصة بمنتج محدد.

إنشاء علاقة واحد إلى واحد

علاقة واحد إلى واحد يكون فيها لكل سجل مرتبط من الجدول الأول يوجد سجل مطابق واحد في الجدول الثاني.

مثلاً، نفرض أننا بحاجة إلى تسجل بعض معلومات المنتج الإضافية الخاصة التي قد نحتاج إليها نادرًا أو التي تنطبق فقط على منتجات قليلة. يمكن

وضعها في جدول منفصل لأننا لا نحتاج المعلومات بشكل متكرر ولأنه يمكن أن ينتج عن تخزين المعلومات في جدول المنتجات مساحة فارغة لكل منتج لا تنطبق عليه. مثلاً، يمكن إنشاء جدول المنتجات، ويمكن استخدام معرف المنتج كمفتاح أساسي. وتعتبر العلاقة بين هذا الجدول التكميلي وجدول المنتج علاقة واحد إلى واحد. لكل سجل في جدول المنتجات، يوجد سجل مطابق واحد في الجدول الإضافي.

عند تحديد علاقة من هذا النوع، يجب أن يشارك الجدولان لحقل مشترك.

عند الكشف عن الحاجة إلى علاقة واحد إلى واحد في قاعدة البيانات، فإنه يمكن وضع معلومات من الجدولين معًا في جدول واحد.

أما إذا لم نرغب في القيام بذلك لسبب ما يتعلق بطبيعة البيانات ومدى توفرها لكل السجلات فنستخدم طريقة إنشاء جدول تكميلي.

توضح القائمة التالية كيفية تمثيل العلاقة في التصميم:

إذا كان للجدولين العنوان نفسه، يمكن إعداد العلاقة باستخدام المفتاح الأساسي نفسه في الجدولين.

إذا كان للجدولين عناوين مختلفة بمفاتيح أساسية مختلفة، فيتم اختيار أحد الجدولين وإدراج المفتاح الأساسي الخاص به في الجدول الآخر كمفتاح خارجي.

يساعد تحديد العلاقات بين الجداول على ضمان الحصول على الجداول والأعمدة الصحيحة.

وعند وجود علاقة واحد إلى واحد أو واحد إلى متعدد، يجب مشاركة الجداول المضمنة لعمود أو أعمدة مشتركة.

وعند وجود علاقة متعدد إلى متعدد، يجب توفير جدول ثالث لتمثيل العلاقة بالشكل الصحيح.

الملحق (٢)
أساسيات مستودعات البيانات Data Warehouse

تقوم مستودعات البيانات بتعميم البيانات ودمجها في مساحة متعددة الأبعاد. يتضمن بناء مستودعات البيانات تنظيف البيانات، وتكامل البيانات، وتحويل البيانات، ويمكن اعتباره خطوة مهمة في المعالجة المسبقة لتنقيب البيانات. علاوة على ذلك، توفر مستودعات البيانات أدوات معالجة تحليلية عبر الإنترنت (OLAP) للتحليل التفاعلي للبيانات متعددة الأبعاد من درجات الدقة المتنوعة، مما يسهل التعميم الفعَّال للبيانات وتنقيب البيانات. يمكن دمج العديد من وظائف تنقيب البيانات الأخرى، مثل الارتباط والتصنيف والتنبؤ والتحليل العنقودي، مع عمليات OLAP لتعزيز التنقيب التفاعلي للمعرفة عند مستويات متعددة من التجريد. وبالتالي، أصبح مستودع البيانات منصة متزايدة الأهمية لتحليل البيانات وOLAP وسيوفر منصة فعالة لتنقيب البيانات. لذلك، يشكل تخزين البيانات وOLAP خطوة أساسية في عملية استكشاف المعرفة. يقدم هذا الفصل نظرة عامة على مستودع البيانات وتكنولوجيا OLAP. تعد هذه النظرة العامة ضرورية لفهم عملية تنقيب البيانات واستكشاف المعرفة بشكل عام.

في هذا الفصل، ندرس تعريفًا مقبولًا لمستودع البيانات ونرى لماذا يقوم المزيد والمزيد من المؤسسات ببناء مستودعات البيانات لتحليل بياناتها (القسم

٤,١). على وجه الخصوص، نقوم بدراسة مكعب البيانات، وهو نموذج بيانات متعدد الأبعاد لمستودعات البيانات وOLAP، بالإضافة إلى عمليات OLAP مثل العرض الإجمالي، والتنقل، والتشريح والتقطيع. كما نلقي نظرة على تصميم واستخدام مستودع البيانات. بالإضافة إلى ذلك، نناقش تنقيب البيانات متعدد الأبعاد، وهو نموذج قوي يدمج مستودع البيانات وتكنولوجيا OLAP مع تنقيب البيانات. نظرة عامة على تنفيذ مستودع البيانات تفحص الاستراتيجيات العامة لحساب مكعب البيانات بكفاءة، وفهرسة بيانات OLAP، ومعالجة استعلام OLAP. أخيرًا، ندرس تعميم البيانات عن طريق الاستقراء الموجه للسمة. تستخدم هذه الطريقة التسلسل الهرمي للمفاهيم لتعميم البيانات على مستويات متعددة من التجريد.

المفاهيم الأساسية

يقدم هذا القسم مقدمة عن مستودعات البيانات. نبدأ بتعريف مستودع البيانات. نوضح الاختلافات بين أنظمة قواعد البيانات التشغيلية ومستودعات البيانات، ثم نوضح الحاجة إلى استخدام مستودعات البيانات لتحليل البيانات، بدلاً من إجراء التحليل مباشرة على قواعد البيانات التقليدية. ويتبع ذلك عرض معماري لمخزن البيانات. بعد ذلك، ندرس ثلاثة نماذج لمستودعات البيانات - نموذج مؤسسة، وسوق بيانات، ومستودع افتراضي. يصف قسم الأدوات المساعدة الخلفية لتخزين البيانات، مثل الاستخراج والتحويل

والتحميل. وأخيرًا ، يعرض القسم مستودع البيانات الوصفية، الذي يخزن البيانات حول البيانات.

ما هو مستودع البيانات ؟

يوفر مستودع البيانات الهياكل والأدوات لمديري الأعمال لتنظيم بياناتهم بشكل منهجي وفهمها واستخدامها لاتخاذ قرارات استراتيجية. تُعد أنظمة مستودعات البيانات أدوات قيّمة في عالم اليوم التنافسي سريع التطور. في السنوات العديدة الماضية، أنفقت العديد من الشركات ملايين الدولارات في بناء مستودعات بيانات على مستوى المؤسسة. يشعر الكثير من الناس أنه مع تزايد المنافسة في كل صناعة، فإن تخزين البيانات هو أحدث سلاح تسويقي لا بد منه كوسيلة للاحتفاظ بالعملاء من خلال معرفة المزيد عن احتياجاتهم.

ولكن ما هو بالضبط مستودع البيانات ؟

تم تعريف مستودعات البيانات بطرق عديدة، مما يجعل من الصعب صياغة تعريف دقيق. بشكل عام، يُشير مستودع البيانات إلى مخزن بيانات يتم الاحتفاظ به بشكل منفصل عن قواعد البيانات التشغيلية للمؤسسة. تسمح أنظمة تخزين البيانات بدمج مجموعة متنوعة من أنظمة التطبيقات. وهي تدعم معالجة المعلومات من خلال توفير منصة صلبة من البيانات التاريخية الموحدة للتحليل.

وفقًا لـ William H. Inmon، أحد المهندسين الرائدين في بناء أنظمة تخزين البيانات، فإن:

"مستودع البيانات عبارة عن مجموعة بيانات موجهة نحو الموضوع ومتكاملة وزمنية وغير متقلبة لدعم عملية اتخاذ قرارات الإدارة" [Inm96].

يقدم هذا التعريف القصير والشامل السمات الرئيسية لمستودع البيانات. تميز الكلمات الرئيسية الأربع (الموجهة نحو الموضوع، والمتكاملة، والزمنية، وغير المتقلبة) مخازن البيانات عن أنظمة مستودع البيانات الأخرى، مثل أنظمة قواعد البيانات العلائقية، وأنظمة معالجة المعاملات، وأنظمة الملفات.

دعونا نُلقي نظرة فاحصة على كل من هذه الميزات الرئيسية:

بيانات موجهة نحو الموضوع

يتم تنظيم مستودع البيانات حول الموضوعات الرئيسية مثل العملاء والموردين والمنتج والمبيعات. بدلاً من التركيز على العمليات اليومية ومعالجة المعاملات في أي مؤسسة، يُركز مستودع البيانات على نمذجة وتحليل البيانات لصانعي القرار. ومن ثم، توفر مستودعات البيانات عادةً عرضًا بسيطًا وموجزًا لقضايا موضوع معين من خلال استبعاد البيانات غير المفيدة في عملية دعم القرار.

بيانات مدمجة أو متكاملة

عادةً ما يتم إنشاء مستودع بيانات من خلال دمج مصادر متعددة غير

متجانسة، مثل قواعد البيانات العلائقية والملفات الثابتة وسجلات المعاملات عبر الإنترنت. يتم تطبيق تقنيات تنظيف البيانات ودمج البيانات وتكاملها لضمان الاتساق في اصطلاحات التسمية وهياكل الترميز ومقاييس السمات وما إلى ذلك.

بيانات زمنية

يتم تخزين البيانات لتوفير المعلومات من منظور تاريخي (على سبيل المثال، السنوات ٥-١٠ الماضية). تحتوي كل بنية رئيسية في مستودع البيانات، إما بشكل ضمني أو صريح، على عنصر زمني.

بيانات غير متقلبة

مستودع البيانات هو دائمًا مخزن منفصل للبيانات يتم تحويله من بيانات التطبيق الموجودة في بيئة التشغيل. بسبب هذا الفصل، لا يتطلب مستودع البيانات معالجة المعاملات واستردادها وآليات التحكم في التزامن. عادة ما يتطلب عمليتين فقط في الوصول إلى البيانات: التحميل الأولي للبيانات والوصول إلى البيانات.

وباختصار، فإن مستودع البيانات هو مخزن بيانات متناسق لغويًا يعمل بمثابة تنفيذ فعلي لنموذج بيانات دعم القرار. يقوم بتخزين المعلومات التي تحتاجها المؤسسة لاتخاذ قرارات استراتيجية. غالبًا ما يُنظر إلى مستودع البيانات على أنه بنية يتم إنشاؤها من خلال دمج البيانات من مصادر متعددة

غير متجانسة لدعم الاستعلامات المنظمة و/ أو الاستعلامات المؤقتة وإعداد التقارير التحليلية وصنع القرار.

استنادًا إلى هذه المعلومات، فإننا نعتبر مستودعات البيانات بمثابة عملية بناء واستخدام مستودعات البيانات. يتطلب بناء مستودع البيانات تنظيف البيانات ودمج البيانات وتماسكها. غالبًا ما يتطلب استخدام مستودع بيانات مجموعة من تقنيات دعم القرار. يسمح هذا "للعاملين في مجال المعرفة" (مثل المديرين والمحللين والمديرين التنفيذيين) باستخدام المستودع للحصول على نظرة عامة سريعة على البيانات بشكل ملائم واتخاذ قرارات سليمة بناءً على المعلومات الموجودة في المستودع. يستخدم بعض المؤلفين مصطلح مستودعات البيانات للإشارة فقط إلى عملية بناء مستودع البيانات، بينما يستخدم مصطلح نظام إدارة قواعد بيانات مستودع البيانات DBMS للإشارة إلى إدارة مستودعات البيانات واستخدامها. لن يتم التمييز هنا.

كيف تستخدم المؤسسات المعلومات من مستودعات البيانات؟

تستخدم العديد من المؤسسات هذه المعلومات لدعم أنشطة اتخاذ القرارات التجارية، بما في ذلك:

- زيادة التركيز على العملاء، والذي يتضمن تحليل أنماط الشراء لدى العملاء (مثل تفضيل الشراء، ووقت الشراء، ودورات الميزانية، وشهية الإنفاق)

- إعادة تنظيم المنتجات وإدارة محافظ المنتجات عن طريق مقارنة أداء المبيعات حسب ربع السنة والعام وحسب المناطق الجغرافية لتحسين استراتيجيات الإنتاج
- تحليل العمليات والبحث عن مصادر الربح
- إدارة علاقات العملاء وإجراء التصحيحات البيئية وإدارة تكلفة أصول الشركة.

تخزين البيانات مفيد أيضًا من وجهة نظر تكامل قاعدة البيانات غير المتجانسة. تجمع المؤسسات عادةً أنواعًا متنوعة من البيانات وتحتفظ بقواعد بيانات كبيرة من مصادر معلومات متعددة وغير متجانسة ومستقلة وموزعة. من المرغوب فيه للغاية، ولكن من الصعب، دمج هذه البيانات وتوفير الوصول إليها بسهولة وكفاءة. تم بذل الكثير من الجهد في صناعة قواعد البيانات ومجتمع البحث العلمي من أجل تحقيق هذا الهدف.

نهج قاعدة البيانات التقليدية التي تهدف إلى دمج قواعد البيانات غير المتجانسة هو بناء مغلفات وموحدات (أو وسطاء) على رأس قواعد بيانات متعددة غير متجانسة. عندما يتم طرح استعلام على موقع عميل، يتم استخدام قاموس بيانات التعريف لترجمة الاستعلام إلى استعلامات مناسبة للمواقع الفردية غير المتجانسة المعنية. ثم يتم تعيين الاستعلامات وإرسالها إلى معالجات الاستعلام المحلية. يتم دمج النتائج التي تم إرجاعها من مواقع مختلفة في مجموعة إجابات عامة. يتطلب هذا النهج القائم على الاستعلام عمليات تصفية ودمج

معلومات معقدة، ويتنافس مع المواقع المحلية لمعالجة الموارد. وهي غير فعَّالة ويمكن أن تكون مكلفة للاستعلامات المتكررة، وخاصة الاستعلامات التي تتطلب تجميعات.

توفر مستودعات البيانات بديلاً مثيرًا للاهتمام لهذا النهج التقليدي. بدلاً من استخدام نهج قائم على الاستعلام، تستخدم مستودعات البيانات نهجًا مستحدثًا يتم فيه دمج المعلومات من مصادر متعددة غير متجانسة مسبقًا وتخزينها في مستودع للاستعلام والتحليل المباشرين. على عكس قواعد بيانات معالجة المعاملات عبر الإنترنت، لا تحتوي مستودعات البيانات على أحدث المعلومات. ومع ذلك، فإن مستودع البيانات يجلب أداءً عاليًا لنظام قاعدة البيانات غير المتجانسة المتكاملة لأنه يتم نسخ البيانات ومعالجتها مسبقًا ودمجها وإضافة تعليقات توضيحية لها وتلخيصها وإعادة هيكلتها في مخزن بيانات دلالي واحد. علاوة على ذلك، لا تتداخل معالجة الاستعلام في مستودعات البيانات مع المعالجة في المصادر المحلية. علاوة على ذلك، يمكن لمستودعات البيانات تخزين ودمج المعلومات التاريخية ودعم الاستعلامات المعقدة متعددة الأبعاد. ونتيجة لذلك، أصبحت مستودعات البيانات شائعًة في الصناعة.

الاختلافات بين أنظمة قواعد البيانات التشغيلية ومستودعات البيانات

نظرًا لأن معظم الأشخاص على دراية بأنظمة قواعد البيانات العلائقية التجارية، فمن السهل فهم ماهية مستودع البيانات من خلال مقارنة هذين

النوعين من الأنظمة.

المهمة الرئيسية لأنظمة قواعد البيانات التشغيلية عبر الإنترنت هي إجراء المعاملات عبر الإنترنت ومعالجة الاستعلام. تسمى هذه الأنظمة أنظمة معالجة المعاملات عبر الإنترنت (OLTP). وهي تغطي معظم العمليات اليومية لمؤسسة مثل الشراء والمخزون والتصنيع والخدمات المصرفية والرواتب والتسجيل والمحاسبة. من ناحية أخرى، تخدم أنظمة مستودعات البيانات المستخدمين أو العاملين في مجال المعرفة ومدراء تحليل البيانات وصنع القرار. يمكن لهذه الأنظمة تنظيم وتقديم البيانات بتنسيقات مختلفة من أجل استيعاب الاحتياجات المتنوعة للمستخدمين المختلفين. وتُعرف هذه الأنظمة بأنظمة المعالجة التحليلية عبر الإنترنت (OLAP).

يتم تلخيص الفرق بين الميزات الرئيسية لـ OLTP وOLAP على النحو التالي:

توجيه المستخدمين والنظام: نظام OLTP موجه للعملاء ويستخدم للمعالجة والاستعلام من قبل كتبة وعملاء ومتخصصين في تكنولوجيا المعلومات. نظام OLAP موجه نحو السوق ويستخدم لتحليل البيانات من قبل العاملين في مجال المعرفة، بما في ذلك المديرين والمديرين التنفيذيين والمحللين.

محتويات البيانات: يدير نظام OLTP البيانات الحالية التي عادة ما تكون

مفصَّلة للغاية بحيث لا يمكن استخدامها بسهولة لاتخاذ القرار. يدير نظام OLAP كميات كبيرة من البيانات التاريخية، ويوفر تسهيلات للتلخيص والتجميع، ويخزن ويدير المعلومات على مستويات مختلفة من الدقة. تعمل هذه الميزات على تسهيل استخدام البيانات لاتخاذ قرارات مستنيرة.

تصميم قاعدة البيانات: عادة ما يعتمد نظام OLTP نموذج بيانات علاقة الكيان (ER) وتصميم قاعدة بيانات موجهة نحو التطبيق. عادة ما يعتمد نظام OLAP إما نموذج النجمة أو ندفة الثلج (انظر القسم ٤,٢,٢) وتصميم قاعدة بيانات موجهة نحو الموضوع.

العرض: يركز نظام OLTP بشكل أساسي على البيانات الحالية داخل مؤسسة أو قسم معين فيها، دون الرجوع إلى البيانات التاريخية أو البيانات في المنظمات المختلفة. في المقابل، غالبًا ما يمتد نظام OLAP إلى إصدارات متعددة من مخطط قاعدة البيانات، بسبب العملية التطورية للمؤسسة. تتعامل أنظمة OLAP أيضًا مع المعلومات التي تنشأ من منظمات مختلفة، وتدمج المعلومات من العديد من مخازن البيانات. بسبب حجمها الضخم، يتم تخزين بيانات OLAP على وسائط تخزين متعددة.

أنماط الوصول: تتكون أنماط الوصول لنظام OLTP بشكل أساسي من المعاملات الفردية الدقيقة القصيرة. مثل هذا النظام يتطلب آليات تحكم واسترداد التزامن. ومع ذلك، فإن الوصول إلى أنظمة OLAP هي في الغالب عمليات للقراءة فقط (لأن معظم مستودعات البيانات تقوم بتخزين المعلومات

التاريخية بدلاً من المعلومات الحديثة)، على الرغم من أن العديد منها يمكن أن يكون استعلامات معقدة.

وتشمل الميزات الأخرى التي تميز بين أنظمة OLTP وOLAP حجم قاعدة البيانات، وتكرار العمليات، ومقاييس الأداء.

ولكن، لماذا يكون لديك مستودع بيانات منفصل؟

نظراً لأن قواعد البيانات التشغيلية تخزن كميات هائلة من البيانات قد تتساءل: "لماذا لا تجري معالجة تحليلية عبر الإنترنت مباشرة على قواعد البيانات هذه بدلاً من قضاء وقت وموارد إضافية لإنشاء مستودع بيانات منفصل؟" أحد الأسباب الرئيسية لهذا الفصل هو المساعدة في تعزيز الأداء العالي لكلا النظامين. تم تصميم قاعدة بيانات تشغيلية وضبطها من المهام وأحمال العمل المعروفة مثل الفهرسة والتجزئة باستخدام المفاتيح الأساسية، والبحث عن السجلات المفصلية، وتحسين الاستعلامات "المقولبة". من ناحية أخرى، غالبًا ما تكون استعلامات مستودع البيانات معقدة. أنها تنطوي على حساب مجموعات البيانات الكبيرة على مستويات ملخصة، وقد تتطلب استخدام تنظيم البيانات الخاصة، والوصول، وأساليب التنفيذ على أساس وجهات نظر متعددة الأبعاد. معالجة استعلامات OLAP في قواعد البيانات التشغيلية من شأنه أن يؤدي إلى تدهور كبير في أداء المهام التشغيلية.

علاوة على ذلك، تدعم قاعدة البيانات التشغيلية المعالجة المتزامنة

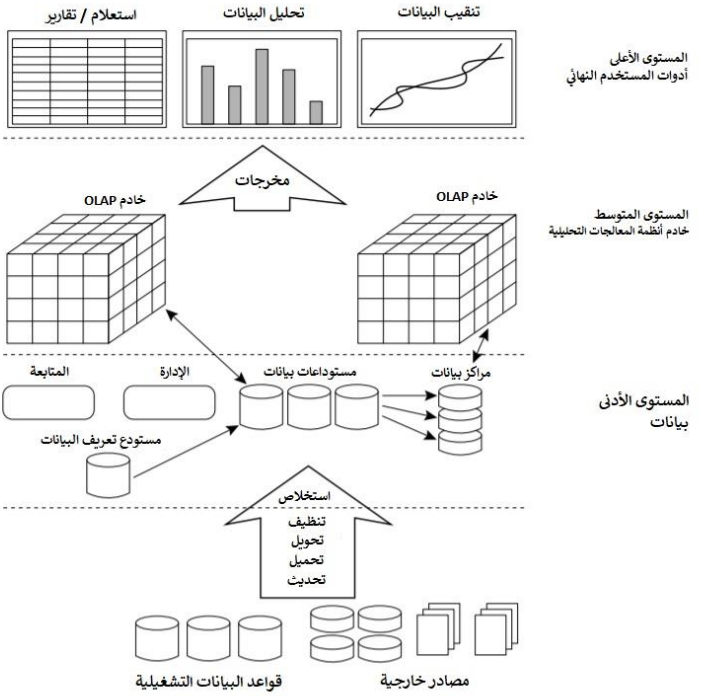
لمعاملات متعددة. هناك حاجة إلى آليات التزامن والاسترداد (مثل القفل والتسجيل) لضمان اتساق المعاملات وقوتها. غالبًا ما يحتاج استعلام OLAP إلى الوصول للقراءة فقط لسجلات البيانات من أجل التلخيص والتجميع. قد تؤدي آليات التحكم في التزامن والاسترداد، إذا تم تطبيقها على عمليات OLAP هذه، إلى تعريض تنفيذ المعاملات المتزامنة للخطر وبالتالي تقليل إنتاجية نظام OLTP بشكل كبير.

وأخيرًا، يعتمد فصل قواعد البيانات التشغيلية عن مستودعات البيانات على الهياكل والمحتويات واستخدامات البيانات المختلفة في هذين النظامين. يتطلب دعم القرار بيانات تاريخية، في حين أن قواعد البيانات التشغيلية لا تحتفظ عادة ببيانات تاريخية. في هذا السياق، فإن البيانات الموجودة في قواعد البيانات التشغيلية، على الرغم من وفرتها، عادة ما تكون بعيدة عن الاكتمال لصنع القرار. يتطلب دعم القرار الدمج (مثل التجميع والتلخيص) للبيانات من مصادر غير متجانسة، مما يؤدي إلى بيانات عالية الجودة ونظيفة ومتكاملة. على النقيض من ذلك، تحتوي قواعد البيانات التشغيلية فقط على بيانات أولية تفصيلية، مثل المعاملات، التي يجب دمجها قبل التحليل. نظرًا لأن النظامين يوفران وظائف مختلفة تمامًا ويتطلبان أنواعًا مختلفة من البيانات، فمن الضروري حاليًا الاحتفاظ بقواعد بيانات منفصلة. ومع ذلك، بدأ العديد من موردي أنظمة إدارة قواعد البيانات العلائقية التشغيلية في تحسين هذه الأنظمة لدعم استعلامات OLAP. مع استمرار هذا الاتجاه، من المتوقع أن

يخفض الفصل بين أنظمة OLTP وOLAP.

تخزين البيانات: بنية متعددة المستويات

غالبًا ما تعتمد مستودعات البيانات على بنية من ثلاث طبقات، كما هو موضح في الشكل التالي:



١. الطبقة السفلى

هو خادم قاعدة بيانات المستودعات والذي يكون دائمًا تقريبًا نظام قاعدة

بيانات علائقية.

تُستخدم الأدوات والمرافق الخلفية لتغذية البيانات في المستوى السفلي من قواعد البيانات التشغيلية أو مصادر خارجية أخرى (على سبيل المثال، معلومات الملف الشخصي للعميل المقدمة من المستشارين الخارجيين). تقوم هذه الأدوات والأدوات المساعدة في استخلاص البيانات والتنظيف والتحويل (على سبيل المثال، لدمج البيانات المماثلة من مصادر مختلفة في تنسيق موحد)، بالإضافة إلى وظائف التحميل والتحديث لتحديث مستودع البيانات (انظر القسم ٤,١,٦). يتم استخلاص البيانات باستخدام واجهات برنامج التطبيق المعروفة بالبوابات. البوابة مدعومة من قبل DBMS الأساسي وتسمح لبرامج العميل بإنشاء كود لغة استعلام هيكلية SQL ليتم تنفيذه على الخادم. تتضمن أمثلة العبّارات ODBC (اتصال قاعدة البيانات المفتوح) وOLEDB (Object Linking وتضمين قاعدة البيانات) بواسطة Microsoft وJDBC (اتصال قاعدة بيانات Java). يحتوي هذا المستوى أيضًا على مستودع بيانات التعريف، الذي يقوم بتخزين معلومات حول مستودع البيانات ومحتوياته. وسوف يتم التطرق لوصف مستودع البيانات الوصفية بشكل أكبر بعد قليل.

٢. الطبقة الوسطى

عبارة عن خادم OLAP يتم تطبيقه عادةً باستخدام إما (١) نموذج OLAP علائقي (ROLAP) (أي، نظام DBMS علائقي موسّع يقوم بتعيين

العمليات على بيانات متعددة الأبعاد إلى عمليات علائقية قياسية)؛ أو (٢) نموذج OLAP متعدد الأبعاد (MOLAP) (أي خادم لأغراض خاصة ينفذ البيانات والعمليات متعددة الأبعاد بشكل مباشر). وسوف يتم التطرق لخوادم OLAP بعد قليل.

٣. الطبقة العليا

هي طبقة المستخدم النهائي، والتي تحتوي على أدوات الاستعلام وإعداد التقارير، وأدوات التحليل، و/ أو أدوات تنقيب البيانات (على سبيل المثال، تحليل الاتجاهات، والتنبؤ، وما إلى ذلك).

نماذج مستودع البيانات: مستودع المؤسسة وسوق البيانات والمستودع الافتراضي

من وجهة النظر الهندسية، هناك ثلاثة نماذج لمستودعات البيانات:

- مستودع المؤسسة
- مركز البيانات
- المستودع الافتراضي

مستودع المؤسسة

يجمع مستودع المؤسسة جميع المعلومات حول الموضوعات التي تغطي المؤسسة بأكملها. يوفر تكامل البيانات على مستوى الشركة، عادة من نظام

تشغيل أو أكثر أو من موفري المعلومات الخارجيين، وهو متعدد الوظائف في النطاق. عادةً ما يحتوي على بيانات تفصيلية بالإضافة إلى بيانات ملخصة، ويمكن أن يتراوح في الحجم من بضعة غيغابايت إلى مئات غيغابايت أو تيرابايت أو أكثر. قد يتم تنفيذ مستودع بيانات المؤسسة على أجهزة الحاسوب المركزية التقليدية، أو أجهزة الخادم الفائقة، أو منصات العمارة المتوازية. يتطلب نمذجة تجارية واسعة النطاق وقد يستغرق سنوات في التصميم والبناء.

مركز بيانات

يحتوي مركز البيانات Data Mart على مجموعة فرعية من البيانات على مستوى الشركة ذات قيمة لمجموعة معينة من المستخدمين. يقتصر النطاق على مواضيع محددة مختارة. على سبيل المثال، قد يقتصر سوق بيانات إدارة التسويق على بيانات للعملاء والمنتجات والمبيعات. تميل البيانات الواردة في مخططات البيانات إلى تلخيصها.

عادةً ما يتم تنفيذ مخططات البيانات على خوادم الأقسام منخفضة التكلفة التي تستند إلى نظام تشغيل Unix / Linux أو Windows. من المرجح أن يتم قياس دورة تنفيذ سوق البيانات بالأسابيع بدلاً من الأشهر أو السنوات. ومع ذلك، فقد ينطوي على تكامل معقد على المدى الطويل إذا لم يكن تصميمه وتخطيطه على مستوى المؤسسة.

اعتمادًا على مصدر البيانات، يمكن تصنيف مراكز البيانات على أنها

مستقلة أو تابعة. يتم الحصول على بيانات مراكز البيانات المستقلة من البيانات الملتقطة من واحد أو أكثر من أنظمة التشغيل أو موفري المعلومات الخارجية، أو من البيانات التي يتم إنشاؤها محليًا داخل إدارة معينة أو منطقة جغرافية معينة. مصادر البيانات في مراكز البيانات التابعة يتم الحصول عليها مباشرة من مستودعات بيانات المؤسسة.

المستودع الافتراضي

المستودع الافتراضي عبارة عن مجموعة من معاينات قواعد البيانات التشغيلية.

من أجل معالجة استعلام فعَّالة، قد تتحقق فقط بعض طرق عرض التلخيص المحتملة. من السهل بناء مستودع افتراضي ولكنه يتطلب سعة إضافية على خوادم قواعد البيانات التشغيلية.

"ما هي إيجابيات وسلبيات نهج التطوير من أعلى إلى أسفل ومن أسفل إلى أعلى لتطوير مستودع البيانات؟" يعمل التطوير من أعلى إلى أسفل لمستودع المؤسسة كحل منظم ويقلل من مشاكل التكامل. ومع ذلك، فهو مُكلف ويستغرق وقتًا طويلاً للتطوير ويفتقر إلى المرونة نظرًا لصعوبة تحقيق الاتساق والتوافق على نموذج بيانات مشترك للمؤسسة بأكملها. يوفر النهج التصاعدي لتصميم وتطوير ونشر مراكز البيانات المستقلة المرونة والتكلفة المنخفضة والعائد السريع للاستثمار. ومع ذلك، يمكن أن يؤدي إلى مشاكل عند دمج

مختلف بيانات مراكز البيانات المتباينة في مستودع بيانات المؤسسة المتسق.

تتمثل إحدى الطرق المُوصى بها لتطوير أنظمة مستودعات البيانات في تنفيذ المستودع بطريقة تدريجية وتطورية، كما هو موضح في الشكل ٤,٢.

أولاً، يتم تعريف نموذج بيانات الشركة عالي المستوى خلال فترة زمنية قصيرة إلى حد معقول (مثل شهر أو شهرين) يوفر نظرة متكاملة ومتسقة ومتكاملة للبيانات بين الموضوعات المختلفة والاستخدامات المحتملة. هذا النموذج رفيع المستوى، على الرغم من أنه سوف يحتاج إلى تحسين في تطوير مستودعات بيانات المؤسسة وسلاسل بيانات الأقسام، لإلاّ أنه سوف يقلل إلى حد كبير من مشاكل التكامل في المستقبل.

ثانيًا، يمكن تنفيذ مخططات البيانات المستقلة بالتوازي مع مستودع بيانات المؤسسة بناءً على نفس مجموعة بيانات الشركة المذكورة سابقًا. ثالثًا، يمكن إنشاء مخططات البيانات الموزعة لدمج مراكز البيانات المختلفة عبر خوادم الموزع. أخيرًا، يتم إنشاء مستودع بيانات متعدد المستويات حيث يكون مستودع المؤسسة هو الوصي الوحيد على جميع بيانات المستودع، ثم يتم توزيعه على مختلف مراكز البيانات التابعة.

الاستخلاص والتحويل والتحميل

تستخدم أنظمة مستودع البيانات الأدوات والمرافق الخلفية لتعبئة وتحديث بياناتها، كما يظهر في الشكل السابق، تتضمن هذه الأدوات والمرافق

الوظائف التالية:

- استخلاص البيانات، الذي يجمع البيانات عادة من مصادر خارجية متعددة وغير متجانسة.
- تنظيف البيانات، الذي يكشف الأخطاء في البيانات ويصححها قدر الإمكان.
- تحويل البيانات، الذي يحول البيانات من تنسيق قديم أو مضيف إلى تنسيق المستودعات.
- تحميل، الذي يفرز، يلخص، يدمج، يحسب المشاهدات، يتحقق من التكامل، ويبني المؤشرات والأقسام.
- تحديث، الذي ينشر التحديثات من مصادر البيانات إلى المستودع.

إلى جانب أدوات التنظيف والتحميل والتحديث وأدوات تعريف البيانات، توفر أنظمة مستودع البيانات عادة مجموعة جيدة من أدوات إدارة مستودع البيانات.

يعد تنظيف البيانات وتحويل البيانات من الخطوات المهمة في تحسين جودة البيانات، وبالتالي نتائج تنقيب البيانات (انظر الفصل ٣). نظراً لأننا مهتمون في الغالب بجوانب تكنولوجيا تخزين البيانات المتعلقة بتنقيب البيانات، فلن ندخل في تفاصيل الأدوات المتبقية، ونوصي القرّاء المهتمين بمراجعة الكتب المخصصة لتقنية تخزين البيانات.

مستودع البيانات الوصفية

البيانات الوصفية هي بيانات حول البيانات. عند استخدامها في مستودع بيانات، فإن البيانات الوصفية هي البيانات التي تحدد كائنات المستودع. يُظهر الشكل السابق مستودع البيانات الوصفية داخل الطبقة السفلى من بنية تخزين البيانات. يتم إنشاء بيانات التعريف لأسماء البيانات وتعاريف المستودع المعطى. يتم إنشاء بيانات التعريف الإضافية والتقاطها من أجل الطابع الزمني لأي بيانات مستخرجة، ومصدر البيانات المستخرجة، والحقول المفقودة التي تمت إضافتها بواسطة عمليات تنظيف البيانات أو عمليات الدمج.

يجب أن يحتوي مستودع البيانات الوصفية على ما يلي:

وصف لبنية مستودع البيانات، والذي يتضمن مخطط المستودع، والعرض، والأبعاد، والتسلسل الهرمي، وتعريفات البيانات المشتقة، بالإضافة إلى مواقع ومحتويات مراكز البيانات.

البيانات الوصفية التشغيلية، والتي تشمل نسب البيانات (تاريخ البيانات التي تم ترحيلها وتسلسل التحولات المطبقة عليها)، وعملة البيانات (نشطة، أو مؤرشفة، أو منظفة)، ومعلومات المراقبة (إحصاءات استخدام المستودعات، وتقارير الأخطاء، ومسارات المراجعة).

الخوارزميات المستخدمة للتلخيص، والتي تتضمن خوارزميات تعريف القياس والأبعاد، وبيانات عن التفاصيل، والأقسام، ومجالات الموضوع،

والتجميع، والتلخيص، والاستعلامات والتقارير المحددة مسبقًا.

رسم الخرائط من البيئة التشغيلية إلى مستودع البيانات، والذي يتضمن قواعد بيانات المصدر ومحتوياتها، وأوصاف البوابة، وأقسام البيانات، واستخلاص البيانات، والتنظيف، وقواعد التحويل والافتراضات، وقواعد تحديث البيانات وتنظيفها، والأمن (إذن المستخدم والتحكم في الوصول).

يحتوي مستودع البيانات على مستويات تلخيص مختلفة، منها البيانات الوصفية.

البيانات المتعلقة بأداء النظام، والتي تتضمن مؤشرات وملفات تعريف تعمل على تحسين الوصول إلى البيانات وأداء الاسترداد، بالإضافة إلى قواعد توقيت وجدولة عمليات التحديث والنسخ الدوري.

البيانات الوصفية للأعمال، والتي تشمل مصطلحات وتعريفات الأعمال، ومعلومات ملكية البيانات ، وسياسات الرسوم.

تشمل الأنواع الأخرى البيانات التفصيلية الحالية (والتي تكون دائمًا على القرص)، والبيانات التفصيلية القديمة (والتي عادة ما تكون على القرص الثانوي)، والبيانات الملخصة بشكل أولي، والبيانات الملخصة بشكل نهائي (التي قد تكون أو لا تكون مخزنة بالفعل).

تلعب البيانات الوصفية دورًا مختلفًا تمامًا عن بيانات مستودع البيانات الأخرى وهي مهمة لأسباب عديدة. على سبيل المثال، يتم استخدام البيانات

الوصفية كدليل لمساعدة محلل نظام دعم القرار في تحديد محتويات مستودع البيانات، وكدليل لرسم الخرائط عند تحويل البيانات من بيئة التشغيل إلى بيئة مستودع البيانات. تعمل البيانات الوصفية أيضًا كدليل للخوارزميات المستخدمة للتلخيص بين البيانات التفصيلية الحالية والبيانات الملخصة بشكل أولي، وبين البيانات الملخصة بشكل أولي والبيانات الملخصة بشكل نهائي. يجب تخزين البيانات الوصفية وإدارتها باستمرار (أي على القرص).

الملحق (٣)
قائمة مصطلحات تنقيب البيانات
Data Mining Index

قائمة مصطلحات تنقيب البيانات التي تم اعتمادها في مجمع اللغة العربية بالقاهرة والذي تزامن مع إصدار الطبعة الأولى من كتاب التحليل المتقدم وتنقيب البيانات الذي أصدرته دار الفكر العربي للطباعة والنشر والتوزيع بالقاهرة عام ٢٠١٦، وهي كما يلي:

الصحة Accuracy: مقياس خاص بقواعد الارتباط والنماذج الاستكشافية التنبؤية، يبين نسبة صحة القاعدة أو نسبة صحة توقع النموذج الاستكشافي.

الذكاء الاصطناعي Artificial Intelligence: مجال علمي يهتم بتخليق السلوك الذكي في الآلات والمعدات حتى تصبح قابلة للتعلّم والاستفادة من الأخطاء.

قواعد التبعية Association Rules: قاعدة على شكل "إذا كان فإن"، تقوم بربط الأحداث ببعضها البعض في قاعدة البيانات، مثل ربط عمليات شراء المنتجات المختلفة من محلات بيع التجزئة.

التكييس، تجميع العناصر في سلات Binning: أحد طرق تحضير البيانات للتحليل والتنقيب، من أمثلتها تجميع زبائن أحد الشركات بحسب

عدة فئات عمرية.

ذكاء الأعمال (استخبارات الأعمال) Business Intelligence: مجموعة متكاملة من التقنيات والأدوات التي تهدف لتحليل البيانات واستكشاف المعرفة التي تساهم في تعزيز قدرة متخذي القرار في المؤسسات والشركات وتعزز من قدرات التخطيط والإدارة الإستراتيجية فيها.

التجزئة العنقودية Clustering: تقنية تهدف لتجميع سجلات قاعدة البيانات في مجموعات منفصلة على أساس التشابه بين العناصر في كل مجموعة والاختلاف مع عناصر بقية المجموعات.

الاحتمالات المشروطة Conditional Probability: احتمال وقوع حدث معين بشرط وقوع حدث آخر.

التغطية Coverage: نسبة عدد السجلات التي تحقق قواعد التبعية أو الارتباط التي يتم استكشافها في قاعدة البيانات إلى إجمالي عدد سجلات قاعدة البيانات.

نظم إدارة قواعد البيانات Database Management System (DBMS): نظم برمجية تهدف للتحكم وإدارة مجموعات البيانات، حفظها وتخزينها وتبويبها والاستعلام عنها وتلخيصها وتجميعها وتطبيق الحسابات المختلفة عليها والتنقيب فيها.

تنقيب البيانات Data mining: تقنيات تهدف لاستكشاف المعرفة والمعلومات المفيدة المخبأة في قواعد البيانات وكل البيانات الكبيرة، ومن

أمثلتها خوارزميات تهدف لاستكشاف السلوك والاتجاهات والأنماط وقواعد التبعية والارتباط وخوارزميات التصنيف والتنبؤ والتجزئة العنقودية واستكشاف القيم المتطرفة.

مستودعات البيانات Data Warehouse: وهي تكتل مركزي لكمية ضخمة من البيانات التي يتم تجميعها من عدة مصادر مثل قواعد البيانات المتعددة والمتوفرة في عدة فروع لشركة ما، والبيانات الأخرى التي تلزم لمتخذي القرار في الشركة كالبيانات الإحصائية العامة المحلية والعالمية وبيانات السوق وغيرها.

جودة البيانات Data Quality: ويتم تقديرها بحسب وجود الأخطاء أو الحاجة للتنظيف والتحضير، وكلما كانت جودة البيانات مرتفعة كلما كانت عمليات التحليل والتنقيب أكثر كفاءة ودقة ومن ثم الحصول على نتائج أفضل.

شجرة القرار Decision Trees: تقنية من تقنيات تنقيب البيانات والتحليل الإحصائي المتقدم التي تقوم بإنشاء شجرة لها فروع متعددة يمثل كل منها تجزئة بطريقة محددة للبيانات التي يتم استكشافها وتحليلها.

الحقل Field: البنية الأولية الأساسية لقاعدة البيانات، وكل جدول يتكون من مجموعة من الحقول، والحقل يعبر عن متغير أو ميزة أو سمة من سمات وخصائص البيانات التي تحتويها قاعدة البيانات في صفوف متعددة تأخذ جميعها قيم مختلفة لكل سمة من السمات الموجودة في كل حقل أو

عمود.

اختبار الفرضيات Hypothesis Testing: طريقة إحصائية في اختبار علاقات محددة بين المتغيرات في البحوث العلمية بحيث يتم التعبير عن العلاقات باستخدام فرضيات غير مؤكدة ويتطلب اختبار صدقها من أجل تأكيد صحتها.

الأحداث المستقلة Independence Events: من منظور إحصائي ومن منظور نظرية الاحتمالات هي الأحداث التي لا يعتمد وقوع أحدها على وقوع الآخر.

استكشاف المعرفة Knowledge Discovery: هو الهدف الأساسي لعلم تنقيب البيانات، ويتم استكشاف المعرفة الموجودة في قواعد البيانات وكل البيانات الكبيرة من خلال تقنيات التحليل والتنقيب.

تعلم الآلة Machine Learning: مجال في العلوم يهتم بتطوير قدرات الآلات وجعلها قادرة على التعلم من الأخطاء، وهو يعتبر فرع من فروع الذكاء الاصطناعي الذي يشتمل على مسارات أخرى عديدة.

تحليل سلة المشتريات Market Basket Analysis: أحد الأمثلة الشهيرة التي توضح تقنية استقراء قواعد الارتباط من خلال تحليل سلة مشتريات زبائن محلات بيع الأغذية الكبرى واستكشاف الأنماط والارتباطات بين عمليات الشراء للمنتجات المختلفة بشكل تجميعي.

النموذج Model: عبارة عن تصميم يتم إعداده من أجل تطبيق خوارزمية

من خوارزميات تنقيب البيانات، ويمكن أن يكون نموذج استكشافي أو تصنيفي أو نموذج تجزئة بحسب طبيعة تقنية التنقيب المستخدمة في إنشائه.

الجار الأقرب Nearest Neighbor: تقنية من تقنيات تنقيب البيانات تقوم باستكشاف قيمة لمتغير مجهول من خلال المقارنة مع قيم جيرانه الذين يشبهونه في قاعدة البيانات.

الشبكة العصبية Neural Network: تقنية من تقنيات تنقيب البيانات وهي نموذج يتم بنائه بهدف التنبؤ، ويعمل بنفس الطريقة التي يعمل بها عقل الإنسان.

تحليل القيم المتطرفة Outlier Analysis: طريقة في تحليل البيانات للكشف عن القيم المتطرفة الشاذة بشكل ملفت في قواعد البيانات، ويمكن استخدامها من أجل تنظيف البيانات واستكشاف دلالاتها الخفية.

الاستكشاف أو التنبؤ Prediction: طريقة تهدف للتكهن بالقيم المجهولة للمتغيرات من خلال تطبيق نماذج الاستكشاف والتنبؤ عليها، ويتم تصميم نماذج الاستكشاف والتنبؤ باستخدام تقنيات تنقيب البيانات المختلفة.

التحليل التنبؤي Predictive Analytics: وهو تقنية من تقنيات التحليل المتقدم وتنقيب البيانات ويهدف للتنبؤ بالأحداث المستقبلية أو السلوك المستقبلي، كالتحليل الهادف للتنبؤ باحتمال إقبال الزبائن على شراء منتج جديد.

نموذج استكشافي Predictive Model: نموذج يتم استخدامه في عملية

الاستكشاف والتنبؤ في تقنيات تنقيب البيانات.

السجل Record: وهو سجل قاعدة البيانات الذي يتم التعبير عنه بمجموعة من السمات، كأن يكون سجل موظف في أحد الشركات، والذي يبين سمات الموظف والبيانات الخاصة به كالعمر والمهنة والراتب وغيرها من البيانات.

الانحدار Regression: تقنية متقدمة في الإحصاء تهدف لاستكشاف قيمة متغير مجهول باستخدام قيمة متغيرات أخرى معلومة مرتبطة بالمتغير المجهول من خلال معادلات رياضية ممكن أن تكون معادلات خطية بمتغيرين.

قواعد البيانات العلائقية Relational Database (RDB): نوع من أنواع قواعد البيانات التي يتم فيها توصيف علاقات فريدة بين الجداول والحقول المختلفة لضبط تدفق البيانات في قاعدة البيانات وضمان صحة البيانات ودقتها وتكاملها.

العينة العشوائية Sampling: تقنية من تقنيات الإحصاء تهدف لانتقاء عدد معين من البيانات حتى يتم دراستها وتحليلها إحصائيًا ومن ثم تعميم النتائج على مجموعة البيانات الأصلية.

لغة الاستعلام الهيكلية Structured Query Language (SQL): لغة قياسية تُستخدم في قواعد البيانات من أجل الاستعلام عن والوصول إلى بيانات محددة وفق معايير متنوعة يتم استخدامها في الاستعلام.

التسويق الاستهدافي Targeted Marketing: طريقة في التسويق

تعتمد على استهداف مجموعات وشرائح وفئات معينة من الزبائن تم تعريفهم بواسطة أحد تقنيات تنقيب البيانات مثل التجزئة العنقودية.

تنقيب النصوص Text Mining: تقنية تستخدم في دراسة وتحليل النصوص بكافة أشكالها من خلال تحليل المحتوى وبحث التشابه فيما بين النصوص المختلفة، ويكثر استخدامها في تقييم البحوث العلمية وبحث مدى أصالتها والابتكار فيها.

التصوير المرئي Visualization: تقنية تستخدم لتصوير البيانات وتلخيصها وتوضيحها وإظهار خصائصها المميزة باستخدام الأشكال والرسومات وبطرق مختلفة.

تنقيب شبكة الإنترنت Web Mining: تقنية لدراسة وتحليل بيانات الإنترنت، سواء من حيث المحتوى أو الهيكلية أو طريقة التصفح والاستخدام والتنقل بين المواقع.

المحتويات

نبذة عن المؤلف

الدكتور مهندس نظم معلومات / مصطفى فؤاد عبيد

عالم وباحث متخصص في القانون والرياضيات وعلوم الحاسوب

مؤسس ورئيس تحرير مجلة البحث العلمي العصري الإلكترونية

الترقيم المعياري الدولي:

ISNN: 2520-5250

مؤسس ومدير عام مركز البحوث والدراسات متعدد التخصصات

الموقع الإلكتروني:
www.mdrscenter.com

الاقتباس من الكتاب

في حالة الاقتباس من هذا الكتاب يمكن الإشارة إليه في قائمة المراجع كما يلي:

عبيد، مصطفى فؤاد، التحليل المتقدم وتنقيب البيانات، النسخة الإلكترونية، غزة، فلسطين، ٢٠٢٦م.

حقوق النشر

تم إتاحة هذه النسخة الإلكترونية للاستخدام الشخصي مجانًا على شبكة الإنترنت في ملف بصيغة PDF

إن إعادة نسخ أو طبع أو نشر أو توزيع لهذا الكتاب أو لأي جزء منه لأغراض تجارية دون الحصول على الموافقة الكتابية من المؤلف يعرض صاحبه للمساءلة القانونية.

جميع الحقوق محفوظة للمؤلف © مصطفى فؤاد عبيد

All Rights Reserved © Mustafa Ebaid

غزة – فلسطين – ٢٠٢٦