

تنقيب البيانات

مفردات المادة

الفصل الأول: مقدمة في تنقيب البيانات وعلم البيانات (مفاهيم أساسية، التطور من قواعد البيانات إلى التنقيب).

الفصل الثاني: مستودعات البيانات (Data Warehousing) بنية المستودعات، مكعبات البيانات، وتقنية المعالجة التحليلية عبر الإنترنت .

الفصل الثالث: معالجة البيانات المسبقة (Data Preprocessing) تنظيف البيانات، دمجها، اختزالها، وتحويلها لتهيئتها للتنقيب.

الفصل الرابع: تنقيب الأنماط المتكررة وقواعد الترابط (Association Rules) خوارزمية Apriori واكتشاف العلاقات المخفية

الفصل الخامس: التصنيف والتنبؤ (Classification & Prediction) أشجار القرار Decision Trees ، خوارزمية Naïve Bayes ، والشبكات العصبية.

الفصل السادس: التجميع والعنقدة خوارزمية K-Means ، والتجميع الهرمي.

الفصل السابع: اكتشاف القيم الشاذة (Outlier Detection) وتطبيقات التنقيب.



المرجع العالمي الأول في التدريس الجامعي (باللغة الإنجليزية)

• الكتاب (Data Mining: Concepts and Techniques (3rd Edition) :

• المؤلفون Jiawei Han & Micheline Kamber :

تنقيب البيانات وتعلم الآلة بدون برمجة 2023 د. علاء طعيمة

التحليل المتقدم وتنقيب البيانات 2026 د. مصطفى فؤاد عبيد

ما المقصود بالتنقيب في البيانات؟

التنقيب في البيانات هو تقنية بمساعدة الحاسوب تُستخدم في إجراء التحليلات لمعالجة مجموعات كبيرة من البيانات واستكشافها. باستخدام أدوات وأساليب التنقيب في البيانات، يمكن للمؤسسات اكتشاف الأنماط والعلاقات المخفية في بياناتها. التنقيب في البيانات يحوّل البيانات الأساسية إلى معرفة عملية. تستخدم الشركات هذه المعرفة بهدف حل المشاكل وتحليل التأثير المستقبلي لقرارات العمل وزيادة هوامش ربحها.

ما المقصود بمصطلح التنقيب في البيانات؟

يُعد مسمى "*التنقيب في البيانات*" تسميةً خاطئةً لأن الغرض منها ليس استخراج البيانات نفسها أو التنقيب فيها. وإنما المقصود منها أن هناك بالفعل قدر كبير من البيانات، وعملية التنقيب في البيانات تستخلص منها فقط المعنى أو المعرفة القيمة. وفيما يلي توضيح للعملية النمطية لجمع البيانات وتخزينها وتحليلها والتنقيب فيها.

- والمقصود من جميع البيانات هو جمعها من مصادر مختلفة مثل تعليقات العملاء والمدفوعات وأوامر الشراء.
- والمقصود بعملية وضع البيانات في مستودعات هو تخزين هذه البيانات في قاعدة بيانات كبيرة أو في [مستودع بيانات](#).
- تقوم تحليلات البيانات بإجراء المزيد من معالجة البيانات وتخزينها وتحليلها باستخدام برامج وخوارزميات معقدة.
- التنقيب في البيانات هو فرع من فروع تحليلات البيانات أو جزء من إستراتيجية تحليلات تُستخدم للعثور على الأنماط المخفية أو التي لم يسبق معرفتها في البيانات.

أهمية التنقيب في البيانات؟

يُعد التنقيب في البيانات جزءًا مهمًا من أي مبادرة تحليلية ناجحة. إذ يمكن للشركات استخدام عملية اكتساب المعرفة لزيادة ثقة العملاء، والعثور على مصادر جديدة للإيرادات، والحفاظ على عودة العملاء. ويساعد التنقيب الفعال في البيانات في مختلف جوانب تخطيط الأعمال وإدارة العمليات. وفيما يلي بعض الأمثلة حول كيفية استخدام التنقيب في البيانات في مختلف القطاعات.

الاتصالات والوسائط والتكنولوجيا

تستخدم القطاعات ذات التنافسية الشديدة مثل قطاع الاتصالات والوسائط والتكنولوجيا التنقيب في البيانات لتحسين خدمة العملاء بإيجاد الأنماط في سلوك العملاء. على سبيل المثال، يمكن لأي شركة تحليل أنماط استخدام النطاق الترددي وتقديم ترقيات أو توصيات مخصصة للخدمة.

الخدمات المصرفية والتأمين

يمكن للخدمات المالية استخدام تطبيقات التنقيب في البيانات لحل المشكلات المعقدة التي تحدث بسبب الاحتيال والامثال وإدارة المخاطر واستنزاف العملاء. على سبيل المثال، يمكن لشركات التأمين معرفة السعر الأفضل للمنتجات من خلال مقارنة أداء المنتج السابق مع أسعار المنافسين.

التعليم

يمكن لمقدمي خدمات التعليم استخدام خوارزميات التنقيب في البيانات لاختبار الطلاب وتخصيص الدروس والتعلم من خلال اللعب. يمكن لطرق عرض تقدّم الطلاب الموحدة القائمة على البيانات أن تساعد المعلمين في معرفة ما يحتاجه الطلاب ودعمهم بصورة أفضل.

التصنيع

يمكن لخدمات التصنيع أن تستخدم تقنيات التنقيب في البيانات لتوفير تحليلات تنبؤية فورية للفعالية الكلية للمعدات ومستويات الخدمة وجودة المنتج وكفاءة سلسلة التوريد. على سبيل المثال، يمكن للمصنعين استخدام البيانات التاريخية القديمة في التنبؤ بتآكل آلات الإنتاج وتوقع الصيانة. ومن ثمّ، فإنهم سيتمكنون من تحسين جداول الإنتاج وتقليل وقت التعطل عن العمل.

البيع بالتجزئة

تمتلك شركات البيع بالتجزئة قواعد بيانات كبيرة للعملاء تضم بيانات أولية حول سلوك شراء العملاء. ويمكن للتنقيب في البيانات معالجة هذه البيانات للحصول على رؤى ذات صلة لحملات التسويق وتوقعات المبيعات. وبوجود نماذج بيانات أكثر دقة، يمكن لشركات البيع بالتجزئة تحسين المبيعات والخدمات اللوجستية لزيادة رضا العملاء. على سبيل المثال، يمكن للتنقيب في البيانات أن يكشف عن المنتجات الموسمية الشائعة التي يمكن تخزينها مسبقاً لتجنب نقصها في اللحظة الأخيرة.

كيفية عمل التنقيب في البيانات؟

تعد العملية القياسية عبر الصناعة للتنقيب في البيانات (CRISP-DM) بمثابة دليل توجيهي متميز يوضح البدء في عملية التنقيب في البيانات. إن العملية القياسية عبر الصناعة للتنقيب في البيانات (CRISP-DM) عبارة عن منهجية ونموذج عمليات حيث تتصف فيها الصناعة والأداة والتطبيق بالحيادية.

- بالنسبة للمنهجية، فهي تصف المراحل النموذجية في مشروع التنقيب في البيانات، ويحدد المهام المتضمنة في كل مرحلة، ويشرح العلاقات الموجودة بين هذه المهام.
- بالنسبة لنموذج العمليات، تقدم العملية القياسية عبر الصناعة للتنقيب في البيانات (CRISP-DM) نظرة عامة حول دورة حياة التنقيب في البيانات.

ما المراحل الست التي تمر بها عملية التنقيب في البيانات؟

بفضل المراحل المرنة الموجودة في العملية القياسية عبر الصناعة للتنقيب في البيانات (CRISP-DM) ، يمكن لفرق البيانات التنقل ذهاباً وإياباً بين المراحل حسب الحاجة. كما يمكن أيضاً للتقنيات البرمجية تنفيذ بعض هذه المهام أو دعمها.

1- فهم الأعمال

يبدأ عالم البيانات أو المُنقب في البيانات بتحديد أهداف المشروع ونطاقه. ويتعاون مع أصحاب المصلحة التجاريين لتحديد معلومات معينة.

- المشكلات التي تحتاج إلى حلول
- قيود المشروع أو حدوده
- تأثير الحلول المحتملة على الأعمال

ثم يستخدم هذه المعلومات في تحديد الغرض من التنقيب في البيانات وتحديد الموارد المطلوبة لاكتساب المعرفة.

2- فهم البيانات

بمجرد الوقوف على المشكلة الموجودة في العمل، يبدأ علماء البيانات في إجراء تحليل أولي للبيانات. ثم يقومون بجمع مجموعات البيانات من مصادر مختلفة، ويحصلون على حقوق إمكانية الوصول، ويقومون بإعداد تقرير حول وصف البيانات. يضم التقرير أنواع البيانات ومقدارها ومتطلبات الأجهزة والبرامج لمعالجة البيانات. وبمجرد موافقة الشركة على خطتهم، يبدأون في استكشاف البيانات والتحقق منها. ويعالجون البيانات باستخدام الأساليب الإحصائية الأساسية، ويقيمون جودة البيانات، ويختارون مجموعة البيانات النهائية للمرحلة التالية.

3- تجهيز البيانات

يقضي المنقبين في البيانات معظم الوقت في هذه المرحلة، لأن برامج التنقيب في البيانات تتطلب بيانات عالية الجودة. وتقوم عمليات الأعمال بجمع البيانات وتخزينها لأسباب أخرى غير التنقيب، ويتوجب على المنقبين تنقيح هذه البيانات قبل استخدامها في النمذجة. ويدخل ضمن تجهيز البيانات العمليات التالية.

• تنظيف البيانات

على سبيل المثال، معالجة البيانات المفقودة ومعالجة أخطاء البيانات والقيم الافتراضية وتصحيحات البيانات.

تكامل البيانات ودمجها

على سبيل المثال، دمج مجموعتين مختلفتين من البيانات للحصول على مجموعة البيانات النهائية المستهدفة.

• تنسيق البيانات

على سبيل المثال، تحويل أنواع البيانات أو تكوين البيانات لتقنية التنقيب المستخدمة.

4- نمذجة البيانات

يتولى المنقبون في البيانات عملية إدخال البيانات التي تم تجهيزها في برامج التنقيب في البيانات ودراسة نتائجها. وللقيام بذلك، يمكنهم الاختيار من بين عدة تقنيات وأدوات خاصة بالتنقيب في البيانات. ويجب عليهم أيضاً إجراء اختبارات لتقييم جودة نتائج التنقيب في البيانات. ولأجل نمذجة البيانات، يمكن لعلماء البيانات القيام بما يلي:

- تدريب نماذج تعلم الآلة (ML) باستخدام مجموعات بيانات أصغر ذات نتائج معروفة
- استخدام النموذج لإجراء مزيد من التحليل لمجموعات البيانات غير المعروفة
- ضبط برامج التنقيب في البيانات وإعادة تكوينها حتى تصبح النتائج مرضية

5- التقييم

بعد إنشاء النماذج، يبدأ المنقبون في قياس البيانات مقارنةً بأهداف العمل الأصلية. ثم يتشاركون النتائج مع محلي الأعمال وجمعون التعليقات. وقد يجيب النموذج عن السؤال الأصلي إجابةً قاطعةً، أو يُظهر أنماطاً جديدة ولم تكن معروفة سابقاً. وبناءً على ملاحظات العمل، يمكن للمنقبين تغيير النموذج أو تعديل الهدف أو معاودة النظر في البيانات. ويُعد التعليق والتعديل والتقييم المستمر جزءاً من عملية اكتساب المعرفة.

6- النشر

أثناء النشر، يستخدم أصحاب المصلحة الآخرون نموذج العمل لإنشاء ذكاء الأعمال. ويتولى عالم البيانات التخطيط لعملية النشر، التي تتضمن تعليم الآخرين بماهية وظائف النموذج، والمراقبة المستمرة، وصيانة تطبيق التنقيب في البيانات. ويستخدم محللو الأعمال التطبيق في إنشاء تقارير للإدارة، ومشاركة النتائج مع العملاء، وتحسين عمليات الأعمال.

ما تقنيات التنقيب في البيانات؟

تُستمد تقنيات التنقيب في البيانات من مختلف مجالات التعلم المتداخلة، بما في ذلك التحليل الإحصائي، وتعلم الآلة (ML)، والرياضيات. وفيما يلي بعض الأمثلة على ذلك.

التنقيب في قواعد الارتباط

التنقيب في قواعد الارتباط هو عملية إيجاد العلاقات بين مجموعتي بيانات مختلفتين وغير مرتبطتين في الظاهر. توضح العبارات الشرطية احتمال وجود علاقة بين نقطتي بيانات. ويقوم علماء البيانات بقياس مدى دقة النتائج باستخدام معايير الدعم والثقة. الدعم يقيس عدد مرات ظهور العناصر ذات الصلة في مجموعة البيانات، بينما الثقة تُظهر عدد المرات التي تكون فيها العبارة الشرطية دقيقة.

على سبيل المثال، حينما يشتري أحد العملاء عنصرًا ما، فإنه في الغالب يشترى عنصرًا آخر له صلة بالعنصر الأول. ويمكن لبائعي التجزئة استخدام التنقيب في الارتباط في بيانات الشراء السابقة لتحديد اهتمام العميل الجديد. ويستخدم بائعو التجزئة نتائج التنقيب في البيانات لملء الأقسام الموصى بها بالمتاجر الإلكترونية.

التصنيف

التصنيف هو تقنية معقدة للتنقيب في البيانات تدرب خوارزمية تعلم الآلة على فرز البيانات إلى فئات متميزة. ولتحديد الفئة، يستخدم التصنيف طرقًا إحصائية مثل أشجار القرار وأقرب جار. وفي كل هذه الطرائق، تتم برمجة الخوارزمية مسبقًا بتصنيفات البيانات المعروفة لتخمين نوعية عنصر البيانات الجديد.

على سبيل المثال، يمكن للمحللين تدريب برنامج التنقيب في البيانات باستخدام صور مُسماة بالتفاح والمانجو. وبشيء من الدقة، يمكن للبرنامج التنبؤ بما إذا كانت صورة جديدة هي صورة لتفاحة أو مانجو أو فاكهة أخرى.

التجميع

المقصود بالتجميع هو تجميع نقاط بيانات متعددة معًا بناءً على أوجه التشابه بينها. وهو يختلف عن التصنيف لأنه غير قادر على تمييز البيانات حسب فئة معينة ولكن يمكنه العثور على أنماط في أوجه التشابه بينهما. وتكون نتيجة التنقيب في البيانات هي فئة من التجميعات التي فيها كل مجموعة تختلف عن غيرها من المجموعات الأخرى، رغم وجود بعض التشابه بين الكائنات الموجودة في كل مجموعة.

على سبيل المثال، يمكن لتحليل المجموعة أن يساعد في أبحاث السوق عند العمل مع بيانات متعددة المتغيرات من الاستبيانات. يستخدم باحثو السوق تحليل المجموعة لتقسيم المستهلكين إلى قطاعات سوق وفهم العلاقات بين المجموعات المختلفة بشكل أفضل.

تحليل التسلسل والمسار

يمكن لبرامج التنقيب في البيانات أيضًا البحث عن أنماط تحتوي على مجموعة معينة من الأحداث أو القيم التي تؤدي إلى أحداث لاحقة. ويمكن لبرامج التنقيب في البيانات التعرف على الاختلافات الموجودة في البيانات والتي تحدث على فترات منتظمة أو عند تغير نقاط البيانات (ارتفاعًا أو انخفاضًا) بمرور الوقت.

على سبيل المثال، يمكن لشركة ما استخدام تحليل المسار لمعرفة زيادات مبيعات بعض المنتجات قبل العطلات مباشرة، أو لملاحظة أن الطقس الأكثر دفئًا يجلب المزيد من الأشخاص إلى موقعها الإلكتروني.

ما أنواع التنقيب في البيانات؟

قد يكون لاستخراج البيانات فروعًا أو تخصصات مختلفة، بناءً على البيانات المطلوبة والغرض من التنقيب. ولنتلق نظرة على بعض منها فيما يلي.

التنقيب في العمليات

يُعد التنقيب في العمليات فرعاً من التنقيب في البيانات، وهو يهدف إلى اكتشاف العمليات التجارية ومراقبتها وتحسينها. وهو يستخرج المعرفة من سجلات الأحداث المتاحة في أنظمة المعلومات. ويساعد المؤسسات في رؤية وفهم ما يحدث في هذه العمليات من يوم لآخر.

على سبيل المثال، توجد لدى شركات التجارة الإلكترونية العديد من العمليات، مثل المشتريات والمبيعات والمدفوعات والتحويل والشحن. ويمكن لهذه الشركات، من خلال التنقيب في سجلات بيانات المشتريات الخاصة بها، معرفة أن موثوقية تسليم مورديها تبلغ 54٪، أو أن هناك نسبة 12٪ من الموردين يقومون بالتسليم في وقت مبكر بصورة مستمرة. كما يمكن لهذه الشركات استخدام هذه المعلومات في تحسين علاقات مورديها.

التنقيب في النصوص

التنقيب في النصوص أو التنقيب في البيانات النصية يستخدم برنامج التنقيب في البيانات لقراءة النص وفهمه. ويستخدم علماء البيانات التنقيب في النصوص لأتمتة طرق اكتساب المعرفة في الموارد المكتوبة مثل مواقع الويب والكتب ورسائل البريد الإلكتروني والمراجعات والمقالات.

على سبيل المثال، يمكن لشركة وسائط رقمية استخدام التنقيب في النصوص لقراءة التعليقات على مقاطع الفيديو الخاصة بها عبر الإنترنت تلقائياً وتصنيف آراء الجمهور على أنها إيجابية أو سلبية.

التنقيب التنبؤي

التنقيب في البيانات التنبؤي يستخدم ذكاء الأعمال للتنبؤ بالتوجهات. وهو يساعد قادة الأعمال في دراسة تأثير قراراتهم على مستقبل الشركة واتخاذ خيارات فعالة.

على سبيل المثال، قد تنتظر الشركة في بيانات مرتجعات المنتجات السابقة لتصميم نظام ضمان لا يسبب أي خسائر. باستخدام التنقيب التنبؤي، سيكون بمقدور هذه الشركات توقع العدد المحتمل للمرتجعات في العام المقبل ووضع خطة ضمان لمدة عام واحد تُراعي عند تحديد سعر المنتج.