تقليل أبعاد البيانات الضخمة



وجد علماء الحاسب في جامعة هارفارد **Harvard university** أنَّ نظرية جونسون-ليندينستراوس ليما **Johnson–Lindenstrauss lemma** الموضوعة قبل 30 عامًا هي أفضل طريقةٍ للمعالجة المسبقة للبيانات الضخمة **Big Data** لتصبح ذات بُعْدٍ أقل ليسهّل على الخوارزميات معالجتها والتعامل معها.

فعندما نفكر في المعلومات الرقمية، غالبًا ما نفكر في حجمها، على سبيل المثال، ربما يكون حجم الرسائل الإخبارية اليومية المرسلة بالبريد الإلكتروني بين 75 إلى 100 كيلوبايت، ولكن للبيانات أبعاد أيضًا تستند فيها إلى عداد المتغيرات في كل جزء منها، فمثلًا يمكن اعتبار الرسالة الإلكترونية على أنها متجه **Vector** ذو أبعاد كثيرة، حيث يوجد نَسَقٌ لكل كلمة في القاموس، والقيمة في ذلك النسق هي عدد المرات التي استُخدمت فيها هذه الكلمة في البريد الإلكتروني، لذلك فإن بُعد البريد الإلكتروني بحجم 75 كيلوبايت ويحتوي على

1000 كلمة من شأنه أن ينتج عنه متجه من ملايين الأبعاد.

هذه النظرة الهندسية للبيانات مفيدة في بعض التطبيقات كتعلّم تصنيف رسائل البريد الإلكتروني المزعجة **spam**، ولكن كلّما ازداد عدد الأبعاد طال وقت عمل الخوارزميّة وزادت الذاكرة التي تستخدمها.

مع تزايد تعقيد معالجة البيانات في منتصف وأواخر التسعينيات، لجأ علماء الحاسب إلى الرياضيات البحتة للمساعدة في تسريع المعالجة الخوارزمية للبيانات، وعلى وجه الخصوص وجد الباحثون الحل في نظريّةٍ أثبتها رياضيان يعملان في مجال التحليل الوظيفي في ثمانينيات القرن الماضي هما ويليام ب. جونسون **William B. Johnson** وَجورام ليندنستراوس **Joram Lindenstrauss**.

استخدم علماء الحاسب النظرية المعروفة باسم جونسون ليندينستراوس ليما **JL lemma** لتقليل أبعاد البيانات والمساعدة في تسريع جميع أنواع الخوارزميات في العديد من المجالات المختلفة من خوارزميّات التدفق والبحث إلى خوارزميات التقريب السريعة للجبر الخطّي والإحصائي وحتى خوارزميّات علم الأحياء الحسابيّ.

ومع تزايد أبعاد البيانات وتعقيدها، بدأ علماء الحاسب يتساءلون فيما إذا كانت **JL lemma** هي حقًا أفضل نهج متّبع للمعالجة المسبقة للبيانات الكبيرة لكي تصبح ذات بعد أقل تسهل معالجتها خوارزميًا.

أما الآن، فقد وضع جيلاني نيلسون **Jelani Nelson** الأستاذ المساعد للأستاذ جون لوب **John L. Loeb** في الهندسة والعلوم التطبيقية في كلية جون بولسون **John A. Paulson** للهندسة والعلوم التطبيقية في جامعة هارفارد نهايةً لهذا الجدل، حيث قُدِّمت ورقة بحثيةٌ في المؤتمر السنوي لمعهد الهندسة الميكانيكية والكهربائية **IEEE** حول أسس علم الحاسب في بيركلي **Berkeley** بولاية كاليفورنيا **California**، ووجد نيلسون المؤلف المشارك كاسبر غرين لارسن **Kasper Green Larsen** من جامعة آرهوس **Aarhus University** في الدنمارك أن نظرية **JL lemma** هي حقًا أفضل وسيلة لتقليل أبعاد البيانات.

ويقول نيلسون: "لقد أثبتنا أنَّ هناك مجموعة من البيانات الصلبة والتي من المستحيل التقليل من بعدها بما يتجاوز ما تنصّ عليه خوارزميّة **JL lemma**".

في الأساس، أظهرت نظرية **JL lemma** أنّ لأي مجموعة محدودة من النقاط ذات البعد الكبير، هناك مجموعة من النقاط بأبعادٍ أقل بكثير مما يحافظ على جميع المسافات بين النقاط بأقلّ قدرٍ من التشويه.

وبعد سنوات من تأثيرها الرائع في التحليل الوظيفي، وجد علماء الحاسب أن خوارزمية **JL lemma** يمكن أن تكون بمثابة خطوة المعالجة المسبقة مما يتيح تقليل أبعاد البيانات إلى حد كبير قبل تشغيل الخوارزميات.

وبدلًا من المرور بكل بعدٍ من أبعاد البيانات مثل مئات الأبعاد من رسائل البريد الإلكتروني، تستخدم خوارزمية **JL lemma** نظام التصنيف الهندسي لتسريع العملية، حيث أنّه في علم الهندسة لا تهم الأبعاد الفردية بقدر أوجه التشابه بينها فمن خلال تعيين أوجه التشابه هذه والربط بينها تُحفَظ هندسة البيانات والزوايا بين نقاط البيانات وذلك في أبعاد أقل.

وبطبيعة الحال، فإن لخوارزمية **JL lemma** مجموعة واسعة من التطبيقات تتعدّى بكثير فلاتر البريد الإلكتروني المزعج، فهي تستخدم الحساسات المستوية لإعادة تجميع الإشارات البديلة باستعمال بضعة قياسات خطية، وتجميع البيانات ذات الأبعاد الكبيرة جدًا، والعثور على الحمض النووي الأساسي في مجال البيولوجيا الحسابية.

يضيف نيلسون: "لا يزال أمامنا طريق طويل لفهم أفضل تقليل ممكن لأبعاد مجموعات البيانات المحددة بدلًا من مقارنتها بأسوأ الحالات. أعتقد أن هذا الاتجاه مهم للغاية للعمل في المستقبل، ولا تزال هناك بعض الأسئلة المفتوحة المثيرة للاهتمام المتعلقة بالسرعة الممكن تنفيذها للتقليل من أبعاد البيانات، وخصوصًا عندما نصادف متجهات متفرّقة أبعادها كبيرة جدًا، أي أن العديد من الإحداثيات تساوي الصفر، وهذه الحالة النادرة ذات أهميّة كبيرة في العديد من التطبيقات العملية، فعلى سبيل المثال، المتجهات المنبثقة عن رسائل البريد الإلكتروني قليلة للغاية لأن البريد الإلكتروني النموذجي لا يحتوي على كل كلمة في القاموس".

تُعد نظرية جونسون ليندينستراوس ليما نتيجةً أساسية في الهندسة متعددة الأبعاد، ولكن لا تزال هناك ثغرة لوغاريتمية مزعجة بين الحدود العليا والدنيا للحد الأدنى الممكن للبعد المطلوب كدالةٍ لعدد النقاط والتشويه المسموح به.

يقول أستاذ الرياضيات نوغا ألون **Noga Alon** الذي أثبت أفضل حدّ أدنى متعلق بالمشكلة: "إنّ العمل الذي أنجزه جيلاني نيلسون وكاسبر غرين لارسن في الآونة الأخيرة قد حل المشكلة، وهو إثبات قوي ورائع لقوة الدمج الذكية بين المنطق التوافقي مع الأدوات الهندسية في حل المشكلة التقليدية".

مصطلحات

البيانات الصلبة **Hard Data**: هي بيانات على شكل أرقام ورسوم بيانية، وفي عالم البيانات الضخمة **Big Data** وعالم إنترنت الأشياء **IOT** تُوصف البيانات الصلبة بأنها البيانات التي تولدها الأجهزة والتطبيقات كالأجهزة الذكية والحواسيب والحساسات وأجهزة القياس الذكية وسجلات المناقلات البنكية وغيرها.

- التاريخ: 2018-09-03
- التصنيف: تكنولوجيا

#نظرية جونسون–ليندينستراوس ليما #خوارزميّات التدفق والبحث #نظام التصنيف الهندسي #مجال البيولوجيا الحسابية #البيانات الصلبة

المصطلحات

- **البيانات الضخمة (Big data)**: مصطلح متطور يصف كمية ضخمة جداً من البيانات المُهيكلة وغير المُهيكلة، يمكن تحليلها حسابياً للحصول على الأساليب والنزعات والعلاقات والروابط، خصوصاً تلك التي تتعلق بالسلوك والتفاعل البشري.

المصادر

- Science Daily
- الصورة

المساهمون

- ترجمة
  - ولاء الحايك
- مُراجعة

- علي مرعي
- تحرير
  - رأفت فياض
  - أحمد كنينة
- تصميم
  - عمرو سليمان
- نشر
  - بيان فيصل