الدراسة المرجعية:

خوارزمية أقرب الجار (KNN):

لقد حاول العديد من الباحثين استخدام تصنيف أقرب الجار للتعرف على الأنماط والتصنيف حيث تتم مقارنة مجموعة اختبار معينة بمجموعة من مجموعات التدريب التي تشبهها. قام [35] في عملهم الخاص بتقديم نظرية المجموعة الضبابية إلى تقنية أقرب الجار لتطوير نسخة ضبابية(fuzzy set) من الخوارزمية. تظهر نتيجة مقارنة النسخة الضبابية بالنسخة الحادة أن الخوارزمية الضبابية تهيمن على نظيرتها من حيث معدل الخطأ المنخفض. في عمل [34] تم استخدام خوارزمية أقرب الجار إلى جانب خمس طرق تصنيف أخرى لدمج استخراج سجلات خادم الويب ومحتوى الويب لتصنيف نمط تصفح المستخدمين والتنبؤ بطلبات المستخدمين المستقبلية. تظهر النتيجة أن KNN تفوقت على ثلاث من الخوارزميات الأخرى ، بينما أدى اثنان منها أداءً موحدًا. كما تبنى [36] أيضًا تصنيف KNN للتنبؤ بموقع توطين الخلية البروتينية. تظهر نتيجة الاختبار باستخدام التحقق المتقاطع الطبقي أن تصنيف KNN يعمل بشكل أفضل من الطرق الأخرى التي تشمل تصنيف شجرة القرار الثنائي والتصنيفات الباييسية الساذجة(naïve Bayesian).

وبشكل اخص استخدم D.A. Adeniyi et al في [37] هذه الخوارزمية لانشاء نظام توصية خاص للتنبؤ **بالاخبار** التي تظهر للمستخدم حيث كانت **مشكلة البحث أن** المواقع الإلكترونية عادةً تقدم العديد من الخيارات أمام المستخدمين ، مما يجعل من الصعب و يستغرق وقتًا طويلاً العثور على المنتج أو المعلومات الصحيحة.

* **الهدف:** تطوير نظام توصية باستخدام أقرب جار (KNN) بناءً على بيانات نقرات المستخدم من موقع ويب لقارئ RSS ) Really Simple Syndication or Rich Site Summary.
* **الطريقة:** يتم تدريب طريقة تصنيف KNN لتحليل بيانات نقرات المستخدم في الوقت الفعلي ، وتصنيفها إلى مجموعات مستخدمين وتقديم توصيات مصممة خصيصًا لهم.
* **الخطوات:**
  + استخراج وتنظيف بيانات عناوين RSS للمستخدمين وتنسيقها.
  + تجميع البيانات في جلسات ذات مغزى وإنشاء مستود بيانات.
  + تصنيف المستخدمين بناءً على بيانات النقر الحالية ومطابقتها بمجموعات المستخدمين.
* **البيانات:**
  + تم استخدام قاعدة بيانات وصول مستخدمي RSS على مدار شهرين.
  + استخراج البيانات ومعالجتها مسبقًا وتجميعها في جلسات لإنشاء مستود بيانات.
* **الأسلوب:**
  + تم استخدام تصنيف أقرب جار لتحليل معلومات عنوان URL من قاعدة بيانات مستخدمي RSS.
  + تم التقييم باستخدام جلسات اختبار نموذجية ، وعرض النتائج والتحليل.
* **النتائج:**
  + يقدم نموذج KNN توصيات دقيقة على حسب ادعائهم حقق معدل دقة يزيد عن 70٪ ، مما يعني أن أكثر من 70٪ من الأخبار الموصى بها تتماشى مع متطلبات المستخدم الفورية.

استخدام K-mean مع KNN

توفر أنظمة توصية الأفلام آلية لمساعدة المستخدمين في تصنيف المستخدمين ذوي الاهتمامات المتشابهة. هذا النظام

(K-mean Cuckoo) لديه MAE 0.68 [39,38] ، في عام 2018 ، قام المؤلفون ببناء محرك توصية عن طريق تحليل مجموعات بيانات التصنيف التي تم جمعها من Twitter لتوصية الأفلام لمستخدم معين باستخدام [40].

لنتعمق بأحد الأبحاث [48] حيث قام الباحثون بانشاء نظام توصية للافلام المستخدمة معتددين اسلوب العنقدة وعلى نفس مجموعة البيانات التي تبنيناها في بحثنا.

* **الهدف:** بناء نظام توصية للأفلام باستخدام خوارزميتي K-Means و أقرب جار (KNN).
* **مجموعة البيانات:** مجموعة بيانات MovieLens من موقع Kaggle
* **التقييم:** يتم قياس أداء النظام باستخدام جذر متوسط مربع الخطأ (RMSE)، حيث تم تحقيق أفضل قيمة RMSE وهي 1.081648.
* **النهج المتبع:**

1. **إدخال المستخدم:** يوفر المستخدم معرف المستخدم والنوع الاجتماعي والعمر ورمز PIN.
2. **التجميع:** يتم استخدام طريقة مجموع المربعات المجمعة ضمن المجموعة لتحديد العدد الأمثل للمجموعات لتجميع K-means.
3. **مصفوفة الفائدة:** يتم بناء مصفوفة فائدة مجمعة، والتي تحدد متوسط التقييم الذي يمنحه المستخدمون لكل مجموعة:

**إنشاء مصفوفة الفائدة:** إنشاء مصفوفة فائدة تشير إلى المستخدم الذي قام بتقييم كل فيلم.

**تحديد العنقود:** استخدام طريقة WCSS لتحديد العدد الصحيح للمجموعات.

**مصفوفة الفائدة المجمعة:** تحديد مصفوفة الفائدة المجمعة بعد تطبيق تجميع K-means.

1. **حساب التشابه:** يتم استخدام معامل ارتباط بيرسون لحساب التشابه بين المستخدمين بناءً على مصفوفة الفائدة.
2. **التنبؤ:** يستخدم KNN مصفوفة الفائدة المجمعة والتشابه للتنبؤ بتوصيات الأفلام للمستخدم الذي أدخل بياناته.]
3. **المقارنة:** مقارنة التقييمات المتوقعة بالتقييمات الفعلية في مجموعة الاختبار.
4. **التقييم:** حساب جذر متوسط مربع الخطأ (RMSE) لتقييم دقة النموذج.

* **تعريف التعلم الآلي:** طريقة تحليل البيانات التي تعمل على أتمتة بناء النماذج التحليلية، مما يتيح للأنظمة التعلم من البيانات واتخاذ القرارات بأقل تدخل بشري.
* **الأداء:** يحقق جذر متوسط مربع خطأ (RMSE) أفضل مقارنة بالتقنيات الموجودة، مع عدد أقل من العناقيد.

تحلل القيمة المفردة (SVD)

تحلل القيمة المفردة (SVD) هي طريقة رياضية أساسية في استخراج البيانات. عادة ما يتم حساب تحلل القيمة المفردة دفعة واحدة، وتعقيد الوقت هو O (m2n + n3) [41] (حيث m و n هما حجم الصف وحجم العمود للمصفوفة على التوالي)، مما يعني أنه يجب معالجة جميع البيانات على الفور. لذلك، فهي ليست مجدية لمجموعات البيانات الضخمة جدًا. اقترح لانكزوس طريقة لتحليل القيمة المفردة يكون تعقيد الوقت لها O(mnr) [41]، حيث r هي رتبة تحليل القيمة المفردة. ومع ذلك، تتطلب طريقة لانكزوس معرفة قيمة r مسبقًا، وهي ليست دقيقة للقيم المفردة الصغيرة [42، 43]. تعتمد طرق تحديث تحليل القيمة المفردة بشكل أساسي على طريقة لانكزوس.

* في عمل [44]، قدم الباحثون خوارزمية تقريبية لتحليل القيمة المفردة تسمى ApproSVD، والتي تجمع بين تقنية تقليل الأبعاد الخاصة بتحلل القيمة المفردة والطريقة التقريبية. تكمن الحيلة وراء خوارزمية ApproSVD في أخذ عينات من بعض صفوف مصفوفة المستخدم-البند، وإعادة قياس كل صف بعامل مناسب لتكوين مصفوفة أصغر نسبيًا، ثم تقليل بعد المصفوفة الأصغر. و لكن تتطلب كمية المعلومات الهائلة على الإنترنت أنظمة توصية فعالة لتوفير محتوى مخصص فقد قد ردفو العمل السابق في .[45]
* **الحل:** اقتراح خوارزمية تدريجية تسمى Incremental ApproSVD، تجمع بين Incremental SVD و ApproSVD، للمعالجة القابلة للتطوير والفعالة.
* المساهمات

1. **خوارزمية تدريجية:** تقترح خوارزمية Incremental ApproSVD للتنبؤ الديناميكي بالتصنيفات غير المعروفة مع إضافة عناصر جديدة.
2. **تحليل الخطأ:** يوفر حداً أعلى للخطأ بين التصنيفات الفعلية والمتوقعة.
3. **التحقق التجريبي:** تظهر التجارب مزايا الخوارزمية المقترحة على مجموعات بيانات MovieLens تحت معايير مختلفة وهي نفس مجموعة البيانات المستخدمة في بحثنا

* المنهجية

1. **تحويل البيانات:** تحويل بيانات تصنيف المستخدم-الفيلم الثابتة إلى مصفوفة تصنيف المستخدم-الفيلم.
2. **عينة المصفوفة:** استخراج عدد ثابت من الأعمدة لتكوين مصفوفات أصغر.
3. **تحلل قيمة مفردة:** يحسب تحلل قيمة مفردة بناءً على تحلل قيمة مفردة للمصفوفة
4. **التنبؤ:** يستخدم متجهات تفضيلات المستخدم ومتجهات ميزات الفيلم للتنبؤ بالتصنيفات. يوصى بالأفلام ذات التصنيفات المتوقعة في النصف العلوي من النطاق.

* **مجموعات البيانات:**
  + **MovieLens:** 943 مستخدمًا و 1682 فيلمًا بـ 100000 تصنيف (وهي نفس مجموعة البيانات المستخدمة في بحثنا).
  + **Flixster:** 8465 مستخدمًا و 9602 فيلمًا بـ 5326788 تصنيفًا.
* **مقياس التقييم:** متوسط الخطأ المطلق (MAE) وجذر متوسط مربع الخطأ (RMSE).
* **التحقق المتقاطع الخماسي:** تقسم التصنيفات إلى خمسة أقسام للتدريب والاختبار.

Slope One

تُعد خوارزمية Slope One من خوارزميات التصفية التعاونية القائمة على العناصر والتي يتم نشرها على نطاق واسع في الصناعة. تعتمد خوارزميات Slope One على المتنبئين على شكل

f(x) = D(x, C) / b

بناءً على تقييمات المستخدمين لل عناصر ، يمكننا الحصول على خط الانحدار بين أي زوج من العناصر. إنها خوارزمية بسيطة للغاية لأنها تستخدم فقط متوسط الفرق بين تقييمات العنصرين كمعلمة حرة وحيدة [46]

. بالإضافة إلى ذلك ، تم أيضًا دراسة نوع اخر ، وهو Slope One ثنائية القطب. يحاول حل مشكلة أخرى وهي أن الثناء وردود الفعل السلبية على تأثير المستخدم في اتخاذ القرار مختلف. أولاً ، يقسم العناصر إلى تلك التي قام المستخدم بتقييمها بشكل إيجابي وتلك التي تم تقييمها بشكل سلبي ، ثم يطبق مخطط Slope One المرجح بشكل منفصل ويستخدم المتوسط المرجح كنتيجة نهائية [46،47].

في مجال الويب الدلالي ، يحاول الباحثون تحسين أنظمة التوصية باستخدام التقنيات الدلالية بشكل رئيسي بناءً على سحابة البيانات المفتوحة المرتبطة (LOD) ونموذج نظام التوصية القائم على المحتوى (يختلف عن النموذج القائم على العناصر ، عادةً لا يستخدم نموذج قائم على المحتوى سوى ميزات العناصر لحساب التوصيات). حسب زعم [47] يمكن جلب الكثير من معلومات السمات المفيدة حول العناصر من LOD ، وفي بعض الأحيان لا تكون هذه المعلومات متاحة بسهولة خاصةً عند البدأ في بناء أنظمة التوصية.فإن أنظمة التوصية القائمة على التصفية التعاونية بشكل عام تعمل بشكل أفضل بكثير من تلك القائمة على المحتوى. ويعزى ذلك أساسًا إلى أن خوارزميات قائمة على المحتوى تتجاهل سلوك المستخدم ، وبالتالي تتجاهل القانون الوارد في شعبية العناصر وسلوك المستخدم. لذلك ، فإن دقته منخفضة نسبيا.

لاحظ الباحثون في [49] أن خوارزميات Slope One التقليدية للترشيح التعاوني ، على الرغم من كونها فعالة وسهلة الاستخدام ، إلا أنها تفتقر أحيانًا إلى دقة التنبؤ.

* **الحل:** اقتراح نهجًا محسنًا لـ Slope One باستخدام التقنيات الدلالية للاستفادة من العلاقات الضمنية بين العناصر بناءً على البيانات المرتبطة.

**المساهمات**

* **خوارزمية محسنة:** تقدم خوارزمية Slope One محسنة تتضمن مسافات دلالية مستمدة من البيانات المرتبطة.
* **التكامل الدلالي:** يستخدم البيانات المرتبطة لالتقاط العلاقات الضمنية للعناصر ، مع ضبط الأوزان في تنبؤ Slope One.

**منهجية**

1. **تحديد البيانات المرتبطة ذات الصلة:** يختار مجموعات فرعية من البيانات المفتوحة المرتبطة بتنسيق RDF ثلاثي الشكل ذات صلة بالمجال (على سبيل المثال ، الأفلام).
2. **مضاهاة معرفات الموارد العالمية إلى مجموعات البيانات التقليدية:** ينشئ تناسقًا بين موارد البيانات المرتبطة ومجموعات البيانات التقليدية باستخدام مطابقة السلاسل.
3. **حساب المسافات الدلالية:** يستخدم خوارزمية LDSD لتحديد المسافات الدلالية ، ودمجها على أنها تشابهات عنصر إلى عنصر في مجموعات البيانات التقليدية.
4. **الدمج في Slope One:** يدمج المسافات الدلالية في خوارزمية Slope One المرجحة ، مع ضبط حسابات الوزن لتحسين دقة التوصية.

**المنصة التجريبية**

* **مجموعة البيانات:**
  + **MovieLens:** توفر البيانات الأساسية للتجربة ، بما في ذلك تقييمات المستخدمين ومعلومات أساسية عن العناصر.
* **معايير التقييم:**
  + **دقة التنبؤ:** تقاس باستخدام جذر خطأ التربيع المتوسط (RMSE) والخطأ المطلق المتوسط (MAE).
  + **دقة التوصية لأعلى N:** يتم تقييمها باستخدام معايير الدقة والاستدعاء.
* **كفاءة التنفيذ:** تحافظ على بساطة وكفاءة خوارزمية Slope One التقليدية مع تحقيق دقة أفضل.
* **التحقق التجريبي:** يؤكد مزايا النهج المقترح من خلال التجارب ، مما يوضح إمكاناته في التطبيقات العملية.

.

.

[34]L. Habin, K. Vlado, Combining mining of web server logs and web content for classifying users’ navigation pattern and predicting users future request, J. Data Knowledge Eng. 61 (2007) (2006) 304–330, http://dx.doi.org/10.1016/j.datak.2006.06.001.

[35] M.K. James, R.G. Michael, A.G. James, A fuzzy K-Nearest Neighbor Algorithm. IEEE Transactions on System Man and Cybernetics, vol. SMC-15 No4.[0018-9472/85/0700-0580$01.00], 1985.

[36] H. Paul, N. Kenta, Better Prediction of Protein Cellular Localization Sites with the K-Nearest Neighbor Classifier, ISMB-97, Proceeding of America Association for Artificial Intelligence, USA, 1997, pp. 147–152.

[37] Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, *12*(1), 90-108.

[38] D.A. Adeniyi, Z. Wei, Y. Yongquan,Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method,Saudi Computer Society, King Saud University,October 2014

[39] Rahul Kataria , Om Prakash Verma,An effective collaborative movie recommender system with cuckoo search,Egyptian Informatics Journal,2016,Volume 18, Issue 2, July 2017, Pages 105-112 <https://doi.org/10.1016/j.eij.2016.10.002>

[40] Das D., Chidananda H.T., Sahoo L. (2018) Personalized Movie Recommendation System Using Twitter Data. In: Pattnaik P., Rautaray S., Das H., Nayak J. (eds) Progress in Computing, Analytics and Networking. Advances in Intelligent Systems and Computing, vol 710. Springer, Singapore

[41] ] G. Chen, F. Wang, C. Zhang, Collaborative filtering using orthogonal nonnegative matrix tri-factorization, Inf. Process. Manag. 45 (3) (2009) 368–379

[42] ] M.W. Berry, Large-scale sparse singular value computations, Int. J. Supercomput. Appl. 6 (1) (1992) 13–49.

[43] H. Nakayama, A. Hattori, Incremental learning and forgetting in RBF networks and SVMs with applications to financial problems, in: Knowledge-Based Intelligent Information and Engineering Systems, Springer, 2003, pp. 1109–1115.

[44] X. Zhou, J. He, G. Huang, Y. Zhang, A personalized recommendation algorithm based on approximating the singular value decomposition (ApproSVD), in: Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, vol. 02, IEEE Computer Society, 2012, pp. 458–464.

[45] Zhou, X., He, J., Huang, G., & Zhang, Y. (2015). SVD-based incremental approaches for recommender systems. *Journal of Computer and System Sciences*, *81*(4), 717-733.

[46] Lemire, D., Maclachlan, A.: Slope One predictors for online rating-based collaborative filtering. In: Proceedings of SDM (2005)

[47] Cacheda, F., Carneiro, V., Fernandez, D., Formoso, V.: Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. IEEE Trans. Web 5(1), 2:1–2:33 (2011)

[48] Ahuja, R., Solanki, A., & Nayyar, A. (2019, January). Movie recommender system using k-means clustering and k-nearest neighbor. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 263-268). IEEE.

[49] Yang, R., Hu, W., & Qu, Y. (2013). Using semantic technology to improve recommender systems based on slope one. In *Semantic web and web science* (pp. 11-23). Springer New York.