Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

Constructing Intelligent Recommendation System for e-Books

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Doctor of Philosophy in
Computer Science**

By

Abdullah Mohammed Saleh Sultan

Supervised by

Asst. Prof. Dr. Alaa Yaseen Taha Mohammed

2024 A.D. **1446 A.H.**

Abstract

Developing recommendation systems is essential and effective as they contribute to enhancing user experience and satisfaction in light of the rapid growth witnessed by digital libraries and e-commerce platforms. This thesis aims to study and develop recommendation models dedicated to e-books that suggest related books based on user preferences and book features.

This thesis proposed four models for e-book recommendation, three collaborative models and one content-based model: the first model is User-based Weighted KNN (UWKNN), the second model is called Item-based Weighted KNN (IWKNN), and the third model is model-based collaborative filtering model (MCFM), while the fourth proposed model is Content-based Hybrid Embedding Model (CBHEM).

The design of both models UWKNN and IWKNN consists of three stages. The first stage includes preprocessing on the Book-Crossing dataset. The second stage includes using the proposed UWKNN algorithm in the first model to add exponential growth weights to users based on their interactions, and the proposed IWKNN algorithm in the second model to add exponential growth weights to books based on the ratings they received. Then, the similarity matrix is created by using different similarity measures, including cosine similarity, mean squared difference (MSD), and Pearson correlation. The third stage includes providing recommendations.

In addition to the above, a third collaborative filtering model based on matrix factorization (MCFM) has been proposed, which uses machine learning algorithms to analyze matrix factors, specifically SVD, NMF, and SVDplus+. The grid search algorithm was used to find the best possible hyperparameters for the model.

The fourth proposed CBHEM model includes five stages, the first is the stage of preprocessing the data (Goodreads dataset), the second stage is the feature engineering stage which includes analyzing book descriptions and other textual data to extract keywords using the KeyBERT model. Then comes the third stage which is the linguistic embedding process where three embedding methods were applied: TF-IDF, Glove, Word2Vec, in addition to proposing a hybrid embedding between Glove and Word2Vec. The fourth stage includes building the similarity matrix, and finally the stage of providing recommendations.

The test results showed that using the two proposed algorithms UWKNN and IWKNN achieved the best performance of the traditional KNN algorithms, as the RMSE values were equal to 1.291 when using the UWKNN algorithm and 1.218 when using the IWKNN algorithm. The test results presented the best accuracy of the proposed models was when using the PearonBaseline similarity measure. As for the MCFM model, it achieved the highest performance when using SVD, as the RMSE was equal to 1.15.

In the last model CBHEM, the highest accuracy was obtained when using the proposed hybrid embedding technique, as the accuracy reached 94.17%, while the accuracy was 78.33% when using the Glove embedding technique and 88.17% when using the Word2Vec technique.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

بناء نظام توصية ذكائي للكتب الالكترونية

اطروحة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة دكتوراه فلسفة في
علوم الحاسوب

من قبل

عبدالله محمد صالح سلطان

بإشراف

أ.م.د. الاء ياسين طه محمد

١٤٤٦هـ ٢٠٢٤ م

المستخلص

يعد تطوير انظمة التوصية امرا ضروريا وفعالا اذ تساهم في تعزيز تجربة المستخدم ورضاه في ظل النمو المتسارع الذي تشهده المكتبات الرقمية اضافة الى المنصات التابعة للتجارة الإلكترونية. الهدف من هذه الأطروحة دراسة وتطوير نماذج توصية مخصصة للكتب الالكترونية تقوم باقتراح الكتب ذات الصلة بناءً على تفضيلات المستخدم وسمات الكتاب.

اقترحت هذه الاطروحة اربعة نماذج لتوصية الكتب الالكترونية، حيث تم تقديم ثلاثة نماذج تعاونية هي: النموذج الاول (UWKNN) المعتمد على المستخدم الموزون وخوارزمية الجار الاقرب k، النموذج الثاني يسمى (IWKNN) المعتمد على العنصر الموزون وخوارزمية الجار الاقرب k, والنموذج الثالث (MCFM) نموذج التصفية التعاونية القائم على النموذج, اما النموذج الرابع المقترح فهو نموذج قائم على المحتوى والتضمين الهجين (CBHEM).

النموذج الاول UWKNN هو نموذج تعاوني قائم على المستخدم UWKNN اما **النموذج الثاني IWKNN** فهو نموذج تعاوني قائم على العنصر (الكتاب). يتكون تصميم كلا النموذجين من ثلاثة مراحل اساسية, تتضمن المرحلة الاولى اجراء المعالجة المسبقة على مجموعة البياناتBook-Crossing تليها المرحلة الثانية والتي يتم خلالها استخدام الخوارزمية المقترحة UWKNN في النموذج الأول لإضافة اوزان نمو اسي للمستخدمين بناءا على تفاعلاتهم، والخوارزمية المقترحة IWKNN في النموذج الثاني لاضافة اوزان نمو اسي للكتب بناءا على التقييمات التي حصلت عليها. ومن ثم إنشاء مصفوفة التشابه من خلال استخدام مقاييس مختلفة للتشابه، بما في ذلك تشابه جيب التمام، ومتوسط الفرق التربيعي (MSD)، وارتباط بيرسون, اما المرحلة الثالثة فتضمنت تقديم التوصيات.

النموذج الثالث الذي تم اقتراحه هو (MCFM)، حيث يستخدم هذا النموذج خوارزميات التعلم الآلي لتحليل العوامل المصفوفية، وتحديدًا خوارزمية تحليل القيمة المفردة (SVD) ، خوارزمية تحليل المصفوفة غير السالبة (NMF) وخوارزمية +SVDplus. لتحقيق أقصى استفادة من هذا النهج، تم استخدام خوارزمية البحث الشبكي للبحث عن أفضل المعلمات الممكنة للنموذج.

النموذج الرابع CBHEM الذي تم اقتراحه يتضمن خمسة مراحل، الأولى هي مرحلة اجراء المعالجة الأولية على البيانات (Goodreads dataset), المرحلة الثانية هي مرحلة هندسة السمات وتشمل تحليل مواصفات الكتب والنصوص لاستخراج كلمات مفتاحية باستخدام نموذج KeyBert. ثم بعد ذلك تأتي المرحلة الثالثة وهي عملية التضمين اللغوي حيث تم تطبيق ثلاث طرق للتضمين: TF-IDF، GloVe، Word2Vec , بالاضافة الى اقتراح تضمين هجين بين GloVe وWord2Vec . المرحلة الرابعة تتضمن بناء مصفوفة التشابه, واخيرا مرحلة لتقديم التوصيات.

أظهرت نتائج الاختبار ان استخدام الخوارزميتين المقترحتين UWKNN وIWKNN حققا أداء افضل خوارزمياتKNN التقليدية حيث كانت قيم RMSE تساوي ١,٢٩١ عند استخدام خوارزمية UWKNN و١,٢١٨ عند استخدام خوارزمية IWKNN , كما أظهرت نتائج الاختبارات ان افضل أداء للنماذج كان عند استخدام مقياس التشابه خط اساس بيرسون,. اما النموذج MCFM فقد حقق اعلى اداءعند استخدام SVD حيث كانت RMSE يساوي ١,١٥.

في النموذج الاخير CBHEM تم الحصول على اعلى دقة عند استخدام تقنية التضمين المهجنة المقترحة اذ بلغت الدقة ٩٤,١٧% بينما كانت الدقة ٧٨,٣٣% عند استخدام تقنية التضمين Glove و ٨٨,١٧% عند استخدام تقنية Word2Vec.