SEBHA UNIVERSITY CONFERENCE PROCEEDINGS, VOL.04 NO. 3 2025
DOI: 10.51984/SUCP.V4I3.4225

وقائع مؤتمرات جامعة سبها
Sebha University Conference Proceedings

Confrence Proceeeding homepage: http://www.sebhau.edu.ly/journal/CAS

تحسين دقة توصيات صفحات الويب عبر تجميع الجلسات والشبكات العصبية ذات الانتشار الخلفي على مجموعة بيانات CTI

مرام عمر ميلاد, ريم محمد دخيل, خالد على عثمان*, سعاد عبدالسلام التومي

كلية تقنية المعلومات، جامعة سبها، ليبيا

الملخص

فى الآونة الأخيرة، قدمت العديد من الأبحاث لتحسين جودة أنظمة توصية صفحات الويب Web page RS وباستخدام منهجية التنقيب في استخدام الويب Web Usage Mining (WUM). وهي تقنية ذكية تعتمد على إنتاج التوصية ببعض الصفحات من خلال المعالجة الممنهجة لبيانات التنقل لمستخدمي الويب المجهولين. ويقسم بها النظام لعدة مراحل إبتداء من مرحلة استخراج الأنماط ذات الصلة والتي تلعب فيها خوارزميات التنقيب دورا مهما. ومن ثم يتم تصنيفها وعرضها على المستخدم النشط, على هيئة روابط links, والتي قد تكون مصدر اهتمامه دون طلب صريح منه. ويتم تقييم هذه الأنظمة من خلال قياس دقة التنبؤ بقرارات المستخدم المستقبلية. الدقة هي مقياس لتقييم فعالية النظام بناءً على الحل المقترح والذي يعكس رضا المستخدم. ومع ذلك، فإن هذه الحلول الحالية لا زالت لم تحقق رغبة المستخدم الكلية. فى هذا العمل، نقدم دراسةً لنظام التنقيب لبيانات استخدام الويب بهدف تحسين دقة التنبؤ لنظام توصية صفحات الويب. لتحقيق ذلك، تم استخدام خوارزمية الانتشار العكسي Back-propagation التي تستند إلى الشبكة العصبية الاصطناعية لتحسين جودة مجموعة جلسات المستخدم المكتشفة. حيث تعتمد الدراسة إجراء العديد من التجارب على مجموعة بيانات سجلات مستخدمي الويب CTI dataset والتى أظهرت نتائجها تحسنًا فى جودة تقديم التوصيات.

الكلمات المفتاحية:

جلسات المستخدم المتداخلة
الشبكة العصبية ذات الانتشار العكسي
نظام توصية بصفحات الويب
نهج WUM

Enhancing Web Page Recommendation Accuracy via Session Clustering and Backpropagation Neural Networks on CTI Dataset

Maram Omar Melad, Reem Mohammed Dekeel, *Khaled Ali Othman, Suaad Abdelsalam Altomi

Faculty of Information Technology, Sebha University,

ABSTRACT

Recently, several studies have been conducted to improve the quality of web page recommendation systems (WPRs) using Web Usage Mining (WUM). Web page RS is an intelligent technique that generates page recommendations through systematic processing of anonymous web user navigation data. These are then categorized and presented to the active user, in the form of links, which may be of interest to the user without their explicit request. These systems are evaluated by measuring the accuracy of predicting future user decisions. Accuracy is a measure of the effectiveness of the system based on the proposed solution, which reflects user satisfaction. However, these current solutions still do not fully meet the user's desires. In this work, we present a study of a web usage data mining system to improve the prediction accuracy of a web page recommendation system. To achieve this, a back-propagation algorithm based on an artificial neural network was used to improve the quality of the detected user session set. The study relies on conducting several experiments on the CTI dataset, and the results showed an improvement in the quality of Web page recommendation.

Keywords:

Back-propagation neural network
Overlapping the user sessions
Web page recommendation system
The WUM approach

1. المقدمة

فى عصر البيانات الضخمة, أحدثت أنظمة التوصية Recommendation Systems(RSs) تغيرًا جذريًا في زيادة التفاعل بين المستخدمين ومواقع الويب. حيث أصبح تلقي التوصيات الآلية وبأشكال مختلفة جزءًا لايتجزأ من تجربة الإستخدام اليومي للإنترنت. فأنظمة التوصية RSs عبارة عن تقنية ذكية تساعد فى توجيه إنتباه مستخدمي الويب فى أوقات قليلة, وذلك عن طريق التنبؤ, إذا كان مستخدم معين يفضل عنصراً (مثل كتب, موسيقي, أفلام,

*Corresponding author:
E-mail addresses: kha.zidane@sebhau.edu.ly, (M. O. Melad) Mara.Melad@Sebhau.edu.ly,
(R. M. Dekeel) Reem.mohammed3@fsc.sebhau.edu.ly, (S. A. Altomi) su.altomi@sebhau.edu.ly

Article History : Received 20 February 2025 - Received in revised form 01 September 2025 - Accepted 07 October 2025

أخبار, صفحات ويب,, وغيرها) أم لا. حيث قللت هذه الأنظمة من مشكلة كثرة المعلومات المُقدمة, وساعدت مستخدمي الويب فى تحديد تفضيلاتهم للعناصر بكفاءة[1]. وبشكل عام، ينقسم نظام التوصيات إلى قسم لتنقيب البيانات, حيث تُجري عملية المعالجة استنادًا إلى البيانات المُجمعة حول العناصر والمستخدمين. وقسمٍ لتقديم التوصيات, حيث تخضع نماذج التوصية إلى العديد من عمليات البحث والتطوير لإستخراج تفضيلات المستخدمين[2].

حاليا, توجد العديد من أنظمة التوصية الفعالة والمفيدة وفي مجالات تطبيقية متنوعة. ومن أبرز هذه الأنظمة, التوصية بصفحات الويب Web page RSs [3], [4], [5], [6], [7], والتي يستخدم بها التنقيب فى استخدام الويب Web Usage Mining- WUM بفعالية كنهج للتخصيص التلقائي, وكوسيلة للتغلب على أوجه القصور فى الأساليب أو المنهجيات التقليدية [5]. حيث تُجمع جلسات المستخدمين Clustering of users sessions اعتمادًا على الخوارزمية, وتقدم لهم توصيات وفقا لأختيار أفضل مجموعة إستقروا بها "الأكثر تشابهًا". وتعتبر جودة خوارزمية التجميع لتقسيم الجلسات partition of the user sessions إلى عدد معين من الأقسام لها تأثير كبير على دقة التنبؤ. ولكن تكمن المشكلة في صعوبةً التعامل مع الجلسات المتداخلة Overlapping of data sessions بين المجموعات المكتشفة. على وجه الخصوص, خوارزمية التجميع K-MEANS, تُعدّ الاختيار الأنسب ومعيارًا مهمًا فى تقسيم جلسات المستخدمين, كما هو الحال في أنظمة توصية صفحات الويب. ولكن فى ظل وجود جلسات متداخلة بين المجموعات المكتشفة تُنتج هذه الخوارزمية نتائج تجزئة غير كاملة, خاصة عند تطبيقها على مجموعات بيانات كبيرة. حيث يعدّ تحديد العدد الأمثل لمجموعة بيانات, والتى تتميز بكثافة مختلفة وأبعاد متعددة, تحديًا ومن القضايا المهمة فى مجال تنقيب البيانات[8].

فى هذه الورقة، نقدم نظاما لتحسين دقة التنبؤ بصفحات الويب عبر تجميع جلسات المستخدم, واستخدام الشبكات العصبية ذات الانتشار الخلفي Backpropagation algorithm (BP) للشبكات العصبية الاصطناعية. حيث تعمل خوارزمية BP على تدريب مجموعات الجلسات المجمعة "باستخدام خوارزمية التجميع K-MEANS", من خلال طريقة تُسمى قاعدة السلسلة. ببساطة، بعد كل مرور أمامي عبر الجلسات لمجموعة الصفحات المصنفة، يُجري الانتشار الخلفى مرورًا خلفيا مع ضبط معلمات النموذج (الأوزان وتحيزات الصفحات بالجلسات المكتشفة). ولتحقيق هذه الغاية، تصف الدراسة مجموعة من التجارب على مجموعة بيانات مستخدمي الويب الحقيقية CTI Dataset, والتي تعتمد فقط على سجلات بيانات الاستخدام المجهولة التي توفرها روابط النص التشعبي للموقع. أيضا تقوم الدراسة بمقارنة وتقييم جودة مجموعة الجلسات المحسنة، بالإضافة إلى مدى فعالية التنبؤ بالصفحات المكتشفة عند استخدامها كجزء من نظام التوصية لتخصيص الويب.

تم تنظيم هذه المقالة فى ستة أقسام رئيسية. القسم الأول هو المقدمة، التي تُحدد أهمية الموضوع والحاجة إلى تقييم مشكلة التنبؤ. القسم الثاني مناقشة الدراسات ذات الصلة واختيارها وتحليلها. ويُقدم القسم الثالث المنهجية المتبعة لعمل النظام. أما القسم الرابع فيمثل التجارب المُقدمة وفقا لمراحل عمل النظام. القسم الخامس, مناقشة وتقييم لنتائج النظام. يُقدم القسم السادس الخلاصة.

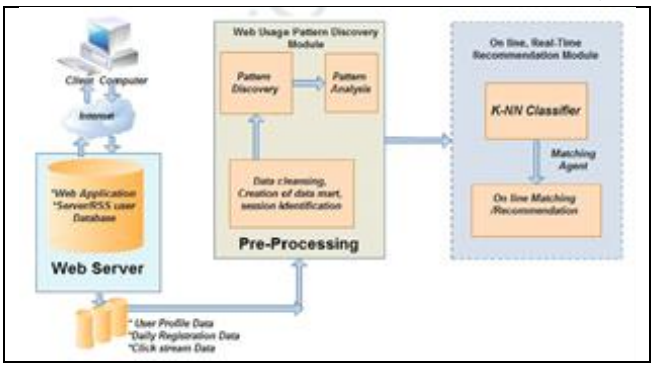
2. الدراسات السابقة

يتمثل مبدأ التوصية بالصفحات Web page RSs المستندة إلى الويب فى أنه يستند على رغبة المستخدم النشط، والتي قد يكون لها ارتباط بصفحات تم مشاهدتها مسبقا (من قبل مستخدم أخر) بحيث يمكن أن تقوم بحفظ وقت تصفح المستخدم مما يساهم فى تقليل التحميل الزائد للمعلومات [9]. حيث يوفر التنقيب فى استخدام الويب WUM منهجية متكاملةً لاستخراج أنماط والتي تُشفِّر نمودجا لسلوك المستخدمين واهتماماتهم[10], [11].

وفقا لدراسة أجراها Mishra وآخرون [12], قدمو بها نموذجًا جديدًا للتوصية بصفحات الويب يأخذ فى الاعتبار البيانات المتسلسلة المتوفرة فى نمط التنقل لمستخدمي الويب، إلى جانب بيانات المحتوى. حيث استخدم النظام طريقة محسنة وهى التجميع التقريبي العلوي للتشابه القائم على المجموعة التقريبية أثناء عملية التجميع overlapping similarity upper approximation during clustering, مما نتج عنه مجموعات لجلسات أكثر تشابها. وقد تم استخدام المجموعات الناتجة لإنشاء مصفوفة تغذية راجعة لتوليد التنبؤات. وللتحقق من النموذج, تم استخدام مجموعة بيانات CTI التي أظهرت أن دقة نموذج الدراسة تفوقت على دقة نماذج التنبؤ مثل التنبؤ العشوائي Random prediction ونماذج ماركوف Markov based models.

كما قدم Luu وآخرون نمذجة شخصية للتوصية بصفحات الويب تعتمد على قياس تشابه التسلسل الذي تم إنشاؤه هرميًا hierarchical session clustering by sequence similarity measure من خلال الاستفادة من وقت نشاط الوصول للمستخدم, وموقعه لتوليد التوصيات[13]. حيث تم اعتماده كمقياس يتعلق بتحسين تشابه الجلسات لتسهيل عملية التنبؤ. وتتميز الدراسة بأن التوصية تقوم بعملية استغلال مجموعة الجلسات لتوقع الحركات المستقبلية للزوار المطابقين مع مراعاة وقت النشاط الذي يقضيه الزوار على الصفحة, وموقعها فى الجلسات. ومع ذلك، فإن النموذج أنتج دقة تنبؤ منخفضة، والتي تقع حول 27.5٪ من التوصية.

بناء على الدراسة التي أجراها Adeniyi وآخرون حول التنقيب فى استخدام الويب. حيث قامو بتطوير نظامًا للتوصية الذى تم اختباره فى الوقت الفعلى لتحديد نقرات العملاء أو الزائرين الجدد[4]. وتعتمد الدراسة على التصنيف بخوارزمية الجار الأقرب K-Nearest Neighbour K-NN والمُنفذ باستخدام طريقة المسافة الإقليدية. وأظهرت نتيجة الدراسة أن محرك التوصيات التلقائي (الشكل 1) قادر على إنتاج تصنيفات وتوصيات مفيدة وجيدة للعميل فى أي وقت, وذلك بناءً على متطلباته الفورية بدلاً من المعلومات المستندة إلى زيارته السابقة للموقع. وفى معظم الحالات، قدم النظام أقصى معدل دقة من التوصيات مساويًا لـ 70%.

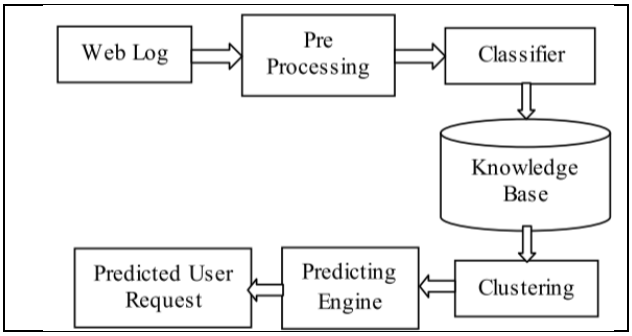


الشكل (1): محرك التوصيات التلقائي الذي يعمل بنموذج تصنيف K-NN

كما قدّم Forsati وآخرون [5] أداة فعالة لتوصية صفحات الويب من خلال استغلال بيانات جلسات المستخدمين, وباستخدام خوارزمية تجميع محسنة Harmony K-MEANS" (IHKSCR, لتقسيم بيانات الجلسة إلى عدد ثابت من المجموعات واستخدامها لتقديم التوصيات. حيث تم دمج طريقة تسمي HS optimization method and fine-tuning power بخوارزمية التجميع K-MEANS, لتحقيق جودة تجميع أفضل للجلسات, وذلك من خلال الجمع بين قدرتها الاستكشافية وقوة الضبط الدقيق لتحديد أعداد المجموعات لخوارزمية K-MEANS. أجريت الدراسة على مجموعة بيانات سجل CTI بجامعة DePaul وانتجت دقة تنبؤ والتي تقع حول 65٪ من التوصية.

أشار أيضا كل من Wagh و Patil ضرورة تحسين التنبؤ بإحتياجات المستخدم لتحسين تجربة تصفح الإنترنت وتزويده بما يريده في وقت أقل [7]. وقاما بإستخدم خوارزمية بحث أولية للعمق Longest Common (LCS) Subsequence على مجموعة بيانات CTI. وهي خوارزمية بحث مُحسّنة تعتمد على العمق أولاً وتعمل على تصنيف المستخدم النشط في إحدى المجموعات المكتشفة. أثناء التقسيم، تم استخدام قيم عتبة الحواف threshold values of edges، بالإضافة إلى قيم عتبة حجم المجموعة threshold values of cluster size، للحصول على أنماط تنقل أكثر فعالية. وقد تم التوصية فقط بصفحات الويب التي تفي بمعايير العتبة. وعند التوصية، أظهرت نتائج الدراسة تحسين دقة التنبؤ بعدد الصفحات. حيث سجلت الدقة القصوى للتنبؤ 61%.

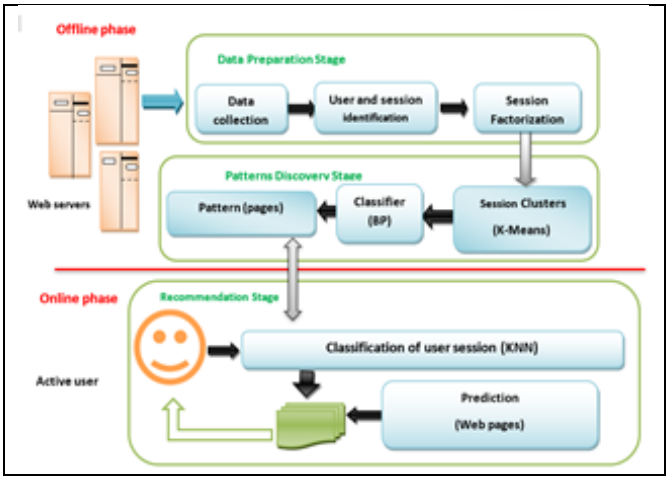
وفي دراسة, قامت om Prakash وأخرون بتحليل نمط تصفح المستخدمين باستخدام تقنية التصنيف تم التجميع [7]. في المرحلة الأولى، تُحدد هذه التقنية المستخدمين المحتملين من بيانات الاستخدام. وتُطبق تقنية تصنيف المستخدمين ذوي الاهتمامات المتشابهة بإستخدم خوارزمية النمط المتكرر (Maximum frequent pattern algorithm) وذلك لتحسين التنقل بناء على سلوك المستخدم. في المرحلة الثانية، استخدمت تقنية التجميع لتحديد طلبات المستخدمين والتنبؤ بها (الشكل 2). حيث تقوم تقنية التجميع بتحديث النمط التسلسلي المطلوب بناء على نقاط الثقل. وأيضا بناءً على المسافة التي يتم من خلالها تحديد نقطة الثقل وتحمل أقصى قيمة بناءً على طلب المستخدم. وقد تم استخدام هذه الأنماط للتنبؤ بطلب المستخدم النشط على مواقع الويب, وأظهرت الطريقة نتائج أفضل أثناء مقارنتها بأداء الخوارزمية الحالية مثل خوارزمية K-MEANS, وخوارزمية FCM. وحقق النموذج أقصى دقة للتنبؤ 77 %، على مجموعة بيانات CTI.



الشكل (2): نظام توصيات لتصنيف جلسات المسخدم باستخدام خوارزمية النمط المتكرر

3. منهجية الدراسة

تستخدم الدارسة المقترحة المنهجية الأساسية لتنقيب بيانات الاستخدام WUM. والمتبعة لإنتاج أنظمة توصية بصفحات الويب. حيث تهدف إلى استخراج معارف وأنماط مفيدة من بيانات مجمعة مسبقا من عدة مصادر خادم الويب. الشكل(3) يقدم خارطة التدفق للنظام المقترح من خلال الوحدات النمطية لمنهجية WUM.



الشكل (3): خارطة التدفق للنظام المقترح

يتكون النظام من وحدتين رئيسيتين؛ وحدة غير متصلة بالإنترنت (Offline Phase), وأخرى عبر الإنترنت (Online Phase). حيث تؤثر الوحدة غير المتصلة بشدة بالوحدة عبر الإنترنت. على وجه الخصوص، تعتبر الوحدة الغير متصلة مهمة لعملية اكتشاف الأنماط ومنها العبور إلى الوحدة عبر الإنترنت لإنتاج التوصية في الوقت الفعلي. وتحتوي الوحدة الغير متصلة Offline Phase على مرحلتين رئيسيتين ومنفصلتين لإجراء تحليلات WUM.

- المرحلة الأولى : إعداد البيانات (Data Preparation stage)

وتتضمن المعالجة المسبقة للبيانات المجمعة, لتنفيذ مجموعة من الخطوات على بيانات جلسة المستخدم. وتتضمن هذه الخطوات التنظيف data cleaning، وتعريف/تمييز المستخدم والجلسة User and session identification، وعولمة جلسات المستخدم User Sessions Factorization [14].

- المرحلة الثانية: اكتشاف الأنماط (Pattern Discovery stage)

وبها تقسيم جلسات المستخدمين إلى مجموعات clustering. وبشكل عام، فإن جودة تحديد مجموعات جلسات المستخدمين clustering of users sessions الفعلية يعتمد على قوة خوارزمية التنقيب المستخدمة. وخوارزمية التجميع K-MEANS هي الأنسب والأكثر استخداما لمهام تقسيم المجموعات [6], [7]. حيث تُستخدم المسافة كمقياس للتماثل، أي أن المسافة الأصغر بين العناصر ونقطة التجمع تظهر قدرًا أكبر من التشابه. ومع ذلك, فإن خوارزمية التجميع (K-MEANS) غير كافية لإكتشاف الجلسات المتداخلة بين المجموعات المكتشفة[6], [8], [15], [16]. وبناء على ذلك, تقوم الدراسة باستخدام خوارزمية الانتشار العكسي (BP) Back Propagation متعددة الطبقات لتدريب الجلسات المكتشفة في كل مجموعة عن طريق تطبيق وظيفة تحسين الأوزان المدخلة لصفحات الجلسة داخل المجموعات. وذلك لأجل الحصول نتيجة تنبوية عن طريق التدريب للتنبؤ بالجلسات المتداخلة. والهدف هو الحصول على أفضل أنماط مجمعة للاستخدام في وحدة الانترنت Online Phase.

أما وحدة الإنترنت Online Phase فهي لإنتاج التوصية في الوقت الفعلي. وتتعلق بتصنيف المستخدم النشط بناءً على اهتماماته في موقع ويب. تم استخدام خوارزمية الجار الاقرب K-Nearest Neighbours-KNN

للتصنيف. وهي خوارزمية تعلم خاضعة للإشراف وتتمز ببساطتها. حيث تصنف الحالات بناءً على تشابهها مع أقرب الحالات المدربة باستخدام مقاييس المسافة مثل المسافة الإقليدية Euclidean distance [17]. وفي هذا السياق، يُعد التنبؤ بقائمة صفحات الويب هدفًا آخر لهذه الوحدة.

4. التجارب The Experiments

في هذا القسم، أُجريت عدة تجارب على مجموعة بيانات مستخدمي الويب الحقيقيين CTI Dataset, مع استُخدام مقاييس مختلفة لتقييم الخوارزميات المقترحة.

1.4 مجموعة بيانات CTI Dataset

وتستند بيانات **CTI Dataset** الأصلية إلى سجلات الخادم لقسم علوم الكمبيوتر المضيف بجامعة[1]DePaul التي جمعت خلال فترة أسبوعين، "قصيرة المدى Short term". وتحتوي مجموعة البيانات على 13،745 جلسة مستخدم مجهول و 683 صفحة مميزة ومقسمة على ثلاثة ملفات رئيسية:

1. ملف CTI.cod: يحتوي هذا الملف على قائمة بالصفحات/الروابط (بين 0 و682صفحة).
2. CTI.tra: يحتوي هذا الملف على بيانات جلسة المستخدم (URL). حيث أن هناك تطابقًا مباشرًا بين الروابط في هذا الملف و الروابط الموجودة في الملف "CTI.cod". ولا تحتوي الملف على الزيارات المتكررة لنفس الصفحة في نفس الجلسة.
3. CTI.std: يمثل هذا الملف المصفوفة الزمنية لجلسات الاستخدام والتي تتوافق مع المشاهدات في ملف "cti.cod" . حيث يحسب الزمن المُستغرق (بالثواني) لمشاهدة الصفحة. وقد بلغ الحد الأقصى لمُدة المشاهدة 999 ثانية.

2.4 اكتشاف جلسات الاستخدام

يوضح هذا القسم مجموعة التجارب على الوحدة الغير متصلة Offline Phase ، والتي تقسم إلى مرحلتين رئيسيتين:

المرحلة الأولى، تتم فيها المعالجة المسبقة للبيانات المجمعة وإزالة جلسات المستخدم غير ذات الصلة. أيضا تحويل جلسات المستخدم إلى مصفوفة متعدد الأبعاد، مما يسمح لجزء تقنيات التنقيب عن البيانات بالبدء في عملية المعالجة. حيث ثم الاعتماد على حساب الوزن لكل صفحة (*p*-Page) عن طريق عاملين اهتمام رئيسين وهما: عدد النقرات "User clicks" و زمن التصفح "Duration time". وبالتالي، لكل مشاهدة وزمن تصفح لـ *p* ، يكون لدى *p* متجه متوسط يسمى *PageWeight* ، والذي يمكن حسابه من خلال إيجاد قيمة النسبة بناءً على عوامل الاهتمام.

وفقا لـ [14] اقترح مقياسًا إحصائيًا لحساب الوزن الإجمالي *PageWeight* للصفحات كما هو بالمعادلة رقم (1):

$$PageWeight(p_i) = \frac{2*F(p_i)*D(p_i)}{F(p_i)+D(p_i)}$$
$$Where\ i\ is\ number\ of\ VisitPage(1 \le i \le m) \quad \text{..... (1)}$$

حيث أن :

- " F " يمثل مجموع النقرات لصفحة الويب.
- " D " مجموع المدة التي استغرقها المستخدم بالصفحة.

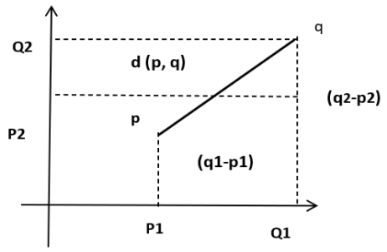
تم أيضا تقسيم الجلسات الغير مصنفة إلى فترتين زمنيتين غير متداخلتين لتشكيل مجموعتي بيانات التدريب Training Dat والاختبار Testing Data . اختير 70% من مجموعة البيانات عشوائيًا للتدريب. بينما استُخدم 30% أخرى للاختبار.

المرحلة الثانية، تم تجميع جلسات المستخدم User Session Clustering لاكتشاف أنماط تصفح المستخدمين ذات الصلة. باستخدام خوارزميات الدراسة المقترحة:

- خوارزمية k-Mean

تلعب خوارزمية K-MEANS دورًا حاسمًا في تحديد أفضل عدد K clusters بين مجموعات الجلسات المكتشفة. حيث تقوم خوارزمية التجميع K-MEANS بإنشاء مجموعات باستخدام متوسط قيمة جميع نقاط البيانات التي تنتمي إليها البيانات. وخطوات خوارزمية k-Mean كالتالي:

1. قم بتهيئة مركز المجموعات من n من نقاط البيانات xi، i= 1 ... n التي يجب تقسيمها في مجموعات k.
2. اختيار أقرب مجموعة إلى كل نقطة بيانات باستخدام المسافة الإقليدية.

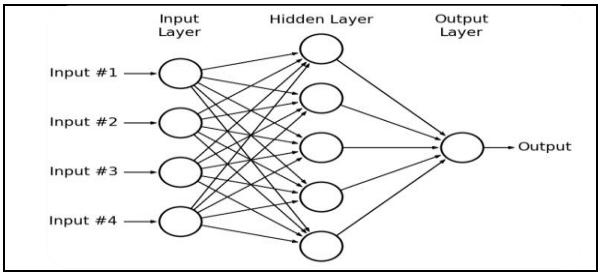


حيث قانون حساب المسافة الإقليدية $d(p,q) = (q1-p1)+(q2-p2)$

3. اضبط موضع كل مجموعة على متوسط جميع نقاط البيانات التي تنتمي إلى تلك المجموعة.

- خوارزمية الانتشار العكسي Back Propagation :

الانتشار العكسي Back Propagation، أو المعروف أيضًا باسم "الانتشار العكسي للأخطاء"، هو طريقة تُستخدم لتدريب الشبكات العصبية (الشكل4).



الشكل (4): بنية الشبكة العصبية الاصطناعية

وتعمل التجربة ببساطة على تقليل الفرق بين الناتج المتوقع لنموذج الشبكة والناتج الفعلي "للبيانات المجمعة". حيث تكتسب الشبكات العصبية تفضيلاتها من خلال الاستفادة من بيانات التدريب، وتحسين دقتها بمرور الوقت. ويعمل النموذج بشكل تكراري لضبط المدخلات (الأوزان w والتحيز **Bias**) لتقليل دالة التكلفة Cost function. حيث أنه في كل دورة، يُكيّف النموذج هذه المعلمات على صفحات الجلسات المكتشفة عن طريق تقليل الخسارة loss باتباع تدرج الخطأ Means Square Error-MSE وفقا للمعاددلة (2):

$$MSE = \frac{1}{n}\sum(yt-yp)^2 \text{.....(2)}$$

حيث أن

- n تمثل عدد الصفحات المدخة بالجلسة.
- Y تمثل الجلسة المراد التنبؤ بها.

- yt تمثل القيمة الحقيقية لتصنيف جلسة المستخدم.
- yp تمثل القيمة المتنبأ بها للجلسة وهي قيمة مخرجات الشبكة.

وبحساب انحدار التدرج باستخدام قاعدة السلسلة The chain rule، مما يسمح لها بالتنقل بفعالية بين الطبقات المعقدة في الشبكة العصبية.

1. نقوم باختيار عينة واحدة من مجموعة جلسات المستخدم فى كل محاولة.
2. نقوم بحساب المشتقات الجزئية للخسارة نسبة إلى أوزان الصفحات بالجلسة المختارة , وتمرير الناتج بدالة التنشيط " دالة السيجمويد " ومن تم إرسالها إلى طبقة الإخراج.
3. في حالة ارتفاع قيمة الخسارة, يتم تعديل القيم (الأوزان w والتحيز **Bias**).
4. نقوم باستخدام دالة التحديث من أجل تحديث كل وزن وانحياز . وبحساب المشتقات الجزئية اللأوزان من بين الطبقة المخفية وطبقة الإخراج. أخيرًا، تتم معالجة هذا الإخراج بواسطة وظيفة التنشيط حتي نحصل علي اقل خسارة.

مع تزايد عدد نماذج التدريب فى الجلسة الواحدة، تزداد قدرة الشبكة على التعلم، بالإضافة إلى القدرة على التعرف بشكل صحيح حتى على الجلسات المتداخلة والتي تحتوي على نسبة خسارة عاليةً. حيث يمكن استبعاد الجلسات المتداخلة من مرحلة التنبؤ وتصنيفها على أنها متطرفة outliers.

3.4 تجارب على التصنيف لإنتاج التنبؤ

ومن أجل اختبار أداء **نظام التوصية**، تم اجراء مجموعة التجارب على الوحدة المتصلة:

- خوارزمية الجار الاقرب K-Nearest Neighbours-KNN

وفقًا لتقنية الجار الأقرب KNN ، يتم تصنيف البيانات الجديدة من خلال تحديد أي تصنيف من مجموعة الصفحات ينتمي اليه المستخدم المجهولة. وتم تحديد قيمة معينة لـ k مما يساعدنا فى تصنيف المجموعة غير المعروفة [18]. وهناك فى التصنيف بشكل أساسي:

- الإدخال: مجموعة صفحات الغير مصنفة (بيانات المستخدم المجهول), تحديد قيمة لـ k.
- المخرجات: مجموعة صفحات مصنفة.

الخطوة 1: تخزين جميع مجموعات الجلسات المدربة.

الخطوة 2: لكل مجموعة جلسات مطلوب تصنيفها:

أ- احسب المسافة بينها وبين جميع مجموعات الجلسات المدربة باستخدام المسافة الإقليدية.

ب- أوجد أقرب k من مجموعات التدريب إلى المجموعة بيانات المستخدم.

ج- عيّن الفئة الأكثر شيوعًا في أقرب k من مجموعات الجلسات المدربة إلى المجموعة الغير مصنفة

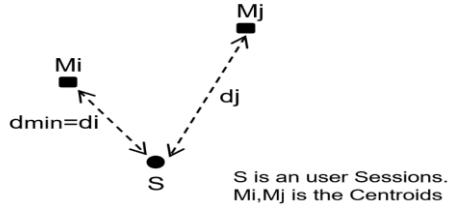
4.4 إجراء التجارب ومقاييس التقييم

تم تطبيق النظام التجريبي على مجموعة بيانات CTI وتنفيذه بشكل أساسي باستخدام معالج Intel® Core™ i5 Processor بتردد 3.7 جيجاهرتز، وذاكرة وصول عشوائي (RAM) سعة 8جيجابايت، ونظام تشغيل Windows 10 Pro. أيضا تم تنفيد جميع الخوارزميات المقترحة بالنظام واختبارها باستخدام Java™ SE Development Kit 8، باستثناء خوارزمية التجميع (K-MEANS) التي شُغّلت واختبرت تجميع الجلسات في بيئة Waikato", برنامج Weka الإصدار 3.6.9". حيث يُعد Weka مجموعة شائعة من برامج التعلم الآلي، مكتوبة بلغة Java والموظفة لمهام التنقيب عن البيانات.

المقياس Davies–Bouldin Index (DBI). وهو مقياس لحساب نسبة مجموع التبعثر داخل الكتلة إلى الفصل بين الكتلة [15], [19]. و يتم حساب DBI (المعادلة 3) بناءً على حالة خوارزمية K-MEANS كما يتم حسابه بعد استخدام خوارزمية التدريب BP .

$$\mathrm{DBI}(k) = \frac{1}{k}\sum_{k=1}^{n} \max_{k\neq l} \left\{ \frac{S(R_k)+S(R_l)}{d(R_k,R_l)} \right\} \quad \text{.....(3)}$$

حيث أن S(Rk)، المسافة الكلية لجلسات المستخدمين فى المجموعة k إلى مركز المجموعة (وفقًا لـ S(Rl), كدرجة للمسافات داخل المجموعة. وللتوضيح, نفترض أن Mi و Mj هما نقطا تجمع لمجموعتين بيانات (الشكل 7). وبالتالي، فإن أدنى مسافة (dmin) بين الجلسة S وأقربها تعني Mi. في حين أن dj هي المسافة بين الجلسة S ونقطا تجمع الأخرى Mj.



حيث و $d(Rk,Rl) = \|Mi - Mj\|$..

تستخدم المسافة النسبية لتحديد ما إذا كانت جلسة المستخدم متداخلة أم لا. تُحسب المسافة النسبية باستخدام المعادلة (4).

$$\mathbf{T'} = \left(\frac{\mathbf{dj}}{\mathbf{di}}\right) \geq \mathbf{T} \quad \text{.....(4)}$$

حيث $(i \neq j)$، T' هو إرجاع منطقي، وT يمثل عتبة التداخل.

ويعتبر DBI معيارًا عامًا للتحقق من صحة الكتلة لخوارزميات الكتلة الجزئية نظرًا لأن نتائجه مستقلة عن عدد المجموعات. حيث أن القيمة الأقل لمؤشر DBI تعني ان النتيجة الأفضل لمجموعات الجلسات.

من أجل فعالية التوصية, قمنا بقياس أداء كل طريقة باستخدام ثلاثة مقاييس قياسية مختلفة، وهي الدقة **Precision** والتغطية **Coverage** ومقياس **F-measure** .

- **Precision :** وهو قياس الدقة لعدد الصفحات الموصى بها والمناسبة بالنسبة لإجمالي التوصيات R1 . كما في المعادلة رقم (5):

$$\mathrm{Precision(Rl,A)} = \frac{|\mathrm{Rl}\cap(\mathrm{A\text{-}SW})|}{|\mathrm{Rl}|} \text{.....(5)}$$

حيث أن تقسَّم كل جلسة مستخدم نشطة User active sessions (A) إلى جزأين باستخدام Sliding Window (SW). يُمثِّل أحد الجزأين عمليات التنقل للمستخدم الحالي، والتي تُستخدم لمطابقة الملفات الشخصية في مرحلة التصنيف. بينما يُمثِّل الجزء الآخر الصفحات المتبقية، والتي تُستخدم لتقييم الدقة .

- **Coverage** : وهو هي مقياس عدد الصفحات الموصى بها المناسبة بالنسبة إلى العدد الإجمالي لصفحات الويب فيما الصفحات يزورها المستخدم. كما فى المعادلة رقم (6).

$$Coverage(Rl,A) = \frac{|Rl\cap(A-SW)|}{|A-SW|||} \text{.....(6)}$$

- **F-measure** وهو المتوسط التوافقي لكل من مقياس Precision ومقياس Coverage (المعادلة 7) يصل مقياس F إلى أقصى قيمته عند تعظيم

. Coverage و Precision

$$F\text{-measure}(Rl,A) = \frac{2*precision(Rl,A)*Coverage(Rl,A)}{precision(Rl,A)+Coverage(Rl,A)} \quad \ldots\ldots\ldots(7)$$

4 النتائج

- نتائج مرحلة إعداد البيانات

في مرحلة اعداد البيانات، "Weight". ومن تم تسوية جميع قيم Page Weight بين 0 و1. حيث يوضح الشكل (5) ملف صفحة الويب المجمع، بما في ذلك قيم Page Weight لكل صفحة.

ID	عدد النقرات	الزمن	Page Weight	URL
0	530	24023	0.131592644	/admissions/
1	26	3936	0.006471572	/admissions/career.asp
2	11	1530	0.002738028	/admissions/checklist.asp
3	96	16300	0.023893002	/admissions/costs.asp
4	41	3167	0.010204379	/admissions/default.asp
5	89	8229	0.022148901	/admissions/general.asp
6	17	792	0.004231236	/admissions/helloworld/arabic.asp
7	21	934	0.005226699	/admissions/helloworld/chinese.asp
8	3	53	0.000746720	/admissions/helloworld/italian.asp
9	3	197	0.000746740	/admissions/helloworld/portugese.asp
10	7	363	0.001742360	/admissions/helloworld/russian.asp
11	7	197	0.001742317	/admissions/helloworld/thai.asp
12	4	516	0.000995657	/admissions/i20_support.doc
13	29	5652	0.007218328	/admissions/i20visa.asp
14	20	902	0.004977840	/admissions/inqinsert.asp
15	52	10399	0.012942897	/admissions/international.asp
16	37	4015	0.009209208	/admissions/mailrequest.asp
17	19	733	0.004728898	/admissions/orientation.asp
18	21	821	0.005226626	/admissions/phdfaq.asp
19	160	21570	0.039816319	/admissions/requirements.asp
20	89	3515	0.022142721	/admissions/statuscheck.asp
21	518	14259	0.128415817	/advising/
22	6	210	0.001493440	/advising/curriculumyear.asp

الشكل (5): نتيجة تحويل البيانات

تم توزيع نتيجة *Page Weight* في مصفوفة n × m حيث يمثل n و m على التوالي, بحيث يتم تمثيل كل جلسة من ملف التنقل الأصلي لمجموعة بيانات CTI. الشكل (6) مثالاً على نتيجة تحويل البيانات.

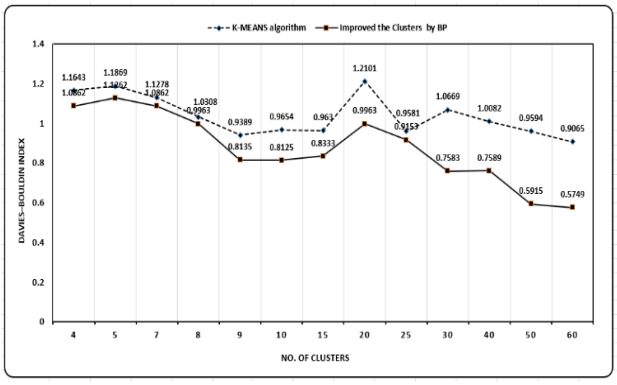
0.131592644	0.0	0.002738028	....	0.002937973
0.0	0.006471572	0.002738028	....	0.0
0.0	0.0	0.0	....	0.0
0.131592644	0.006471572	0.0	....	0.002937973
0.0	0.006471572	0.002738028	....	0.0
0.131592644	0.0	0.0	....	0.002937973
0.0	0.006471572	0.0	....	0.002937973
....	....	....	....	....
0.131592644	0.0	0.002738028	....	0.002937973

الشكل (6): مصفوفة جلسات المستخدمين

بهذه الطريقة، يمكن استخدام السجلات المعدة لتنقيب جلسات المستخدم بشكل صحيح في مرحلة اكتشاف الأنماط.

- نتائج مرحلة اكتشاف الأنماط

في هذا القسم، يتم تقديم النتائج بناءً على الطريقة المقترحة لهذه المرحلة. حيث تم تجميع جلسة المستخدم الفعالة من خلال المرحلة السابقة. حيث تم تشغيل خوارزميات التجميع (K-MEANS) على مصفوفة جلسات المستخدمين الإجمالية. أيضا, تم استخدام خوارزمية الانتشار العكسي BP لإعادة تدريب مجموعة جلسات المستخدمين المكتشفة. الشكل (7) يوضح نتائج عملية تقييم جلسات الاستخدام.



الشكل (7): نتائج خوارزميات تجميع الجلسات المتداخلة.

نلاحظ من الشكل السابق انخفاض في قيمة مقياس DBI بعد عملية التدريب مقارنة بمقياس خوارزمية K-MEANS وهذا يعني تحسين في مستوى التكتل داخل المجموعة الواحدة. كما نلاحظ من الشكل أن مجموعات الجلسات من 9-10 أكثر استقرارا والتي ينخفض بها مقياس DBI إلى 0.8125. في هذا العمل، تم اختيار عدد 10 مجموعات من جلسات المستخدم لأنها تعتبر الأفضل جودة لانتاج التنبؤ.

- نتائج مرحلة التصنيف والتنبؤ

وهي بداية الوحدة النمطية عبر الإنترنت (Online Phase) للتنبؤ بصفحات الويب RS. حيث تم استخدام استراتيجية الجار الاقرب (kNN) يمكن تصنيف المستخدم تلقائيًا إلى أحد الفئات المكتشفة مسبقًا. وللتنبؤ، يتم تقييم جودة التوصية. تم مطابقة ملفات تعريف المستخدم المختارة مع الجزء المتبقي من جلسة المستخدم النشطة (الجلسات المعدة للاختبار) لقياس دقة التنبؤ للنظام. وباستخدام مقياس وزن الصفحة (threshold) تتراوح من 0.1 إلى 1 للتحكم في قائمة صفحات التوصيات.

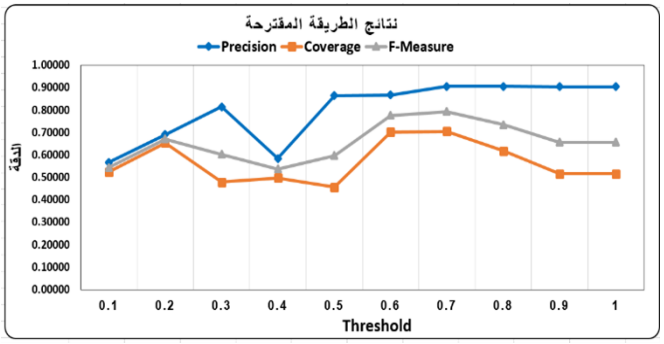
حيث يظهر الجدول (1) نتائج تجربة التنبؤ عبر الإنترنت باستخدام مقاييس precision , coverage , F-measure

الجدول (1) : نتائج تجربة التنبؤ للمقاييس عبر الانترنت:

precision , coverage , F-measure

Threshold (*t*)	Precision	Coverage	F-Measure
0.1	0.56730	0.52477	0.54521
0.2	0.69101	0.65346	0.67171
0.3	0.81520	0.47919	0.60358
0.4	0.58511	0.49746	0.53773
0.5	0.86399	0.45600	0.59694
0.6	**0.86734**	**0.70214**	**0.77604**
0.7	**0.90596**	**0.70560**	**0.79332**
0.8	0.90583	0.61879	0.73529
0.9	0.90445	0.51565	0.65682
1	0.90448	0.51565	0.65683

الشكل (8) مقياس الدقة F-measure المتوسطة بين الدقة لـ precision و coverage. ويوضح أن نظام صفحات الويب المقترحة أكثر فاعلية. حيث تحقق الدقة القصوى التنبؤ بنسبة 0.79 عند يكون الحد t = 0.7 .



الشكل (8): مقياس التنبؤ بدقة التوصية للنظام

5 المناقشة وتقييم النظام

وفقًا للنتائج التجريبية، فإنها الدراسة تشير إلى أن الطريقة المقترحة للنظام تحسن من جودة مجموعة جلسات المستخدمين عن طريق خوارزمية الانتشار العكسي BP بالوحدة النمطية غير المتصلة بالإنترنت. ومن أجل تقييم ما إذا كان النظام المنفذ يوفر خيارًا أفضل للتنبؤ، تم مقارنة نظام التوصية بحلول سابقة. حيث يوضح الجدول (2) نتائج المقارنة لنسبة الدقة.

الجدول (2) : مقارنة نسبة الدقة لانتاج Web Page RS.

Dataset	الخوارزمية	Precision	Coverage	F-Measure
CTI data	K-MEANS	0.5728	0.5071	0.6202
	FCM + Rec. Probability	0.611	0.6520	0.6462
	FCM KNN	0.7902	0.7195	0.6983
	Maximum Frequent	0.8741	0.9133	0.7774
	الطريقة المقترحة	0.80596	0.70560	**0.79332**

6 الخلاصة

فى هذه الورقة، تناولنا مشكلة تقسيم جلسات المستخدمين باستخدام التجميع. واقترحنا تحسين دقة التنبؤ بتوصيات صفحات الويب وفقا لطلبات المستخدمين المستقبلية. حيث تم تطبيق خوارزمية الانتشار العكسي BP لتحسين جودة اكتشاف مجموعات جلسات المستخدمين والمجمعة بخوارزمية K-MEANS. حيث تم اجراء تجارب مكثفة على مجموعة بيانات الاستخدام CTI dataset لتقييم جودة النظام المقترح. أظهرت النتائج تحسين دقة التنبؤ مقارنةً بدراسات سابقة. يترك هذا العمل اتجاهًا للعمل المستقبلي الذي يستحق التحقيق وهو البحث فى الطرق المحسنة لتصنيف المستخدمين الجدد كخوارزميات k-NN والتى يتم فيها تدريب مجموعة من المصنفات ودمجها لتحسين التنبؤ.

المراجع

[1] H. Zhou, F. Xiong, and H. Chen, "A Comprehensive Survey of Recommender Systems Based on Deep Learning," Oct. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app132011378.

[2] H. Ko, S. Lee, Y. Park, and A. Choi, "A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields," Jan. 01, 2022, *MDPI*. doi: 10.3390/electronics11010141.

[3] J. Chanda and B. Annappa, "An improved web page recommendation system using partitioning and web usage mining," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2015. doi: 10.1145/2816839.2816910.

[4] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Applied Computing and Informatics*, vol. 12, no. 1, pp. 90–108, Jan. 2016, doi: 10.1016/j.aci.2014.10.001.

[5] R. Forsati, A. Moayedikia, and M. Shamsfard, "An effective Web page recommender using binary data clustering," *Inf Retr Boston*, vol. 18, no. 3, pp. 167–214, Jun. 2015, doi: 10.1007/s10791-015-9252-4.

[6] P. P. G. Om, S. Ananthakumaran, M. Sathishkumar, and R. Ganeshan, "Analyzing the User Navigation Pattern from Web Logs Using Maximum Frequent Pattern Approach," in *Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021, pp. 877–883. doi: 10.1109/ICICT50816.2021.9358751.

[7] R. Wagh, "A Novel Web Page Recommender System for Anonymous Users Based on Clustering of Web Pages," *Asian Journal of Convergence in Technology*, vol. V Issue I, 2019, [Online]. Available: www.asianssr.org

[8] A. Nowak-Brzezinska and C. Horyn, "Outliers in rules - The comparision of LOF, COF and KMEANS algorithms.," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 1420–1429. doi: 10.1016/j.procs.2020.09.152.

[9] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," 2005.

[10] B. Mobasher, "Data mining for Web personalization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2007, pp. 90–135. doi: 10.1007/978-3-540-72079-9_3.

[11] S. Sakarkar, V. Chaudhari, T. Gaurkar, A. Veer, and M. K. Scet, "Web personalisation based on user interaction : Web Personalisation," in *Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021*, Institute of Electrical and Electronics Engineers Inc., Feb. 2021, pp. 234–238. doi: 10.1109/ICICV50876.2021.9388384.

[12] R. Mishra, P. Kumar, and B. Bhasker, "A web recommendation system considering sequential information," *Decis Support Syst*, vol. 75, pp. 1–10, Apr. 2015, doi: 10.1016/j.dss.2015.04.004.

[13] Vinh-Trung Luu; Germain Forestier; Mathis Ripken; Frédéric Fondement; Pierre-Alain Muller, "Web usage prediction and recommendation using web session clustering," *2016 Eleventh International Conference on Digital Information Management (ICDIM)*, no. 107–113, Sep. 2016, doi: 10.1109/ICDIM.2016.7829779.

[14] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization," 2002. [Online]. Available: www.netperceptions.com

[15] K. Ali Othman, M. Nasir Sulaiman, N. Mustapha, and N. Mohd Sharef, "Local Outlier Factor in Rough K-Means Clustering," *Pertanika J. Sci. & Technol*, vol. 25, pp. 211–222, 2017, [Online]. Available: http://www.pertanika.upm.edu.my/

[16] Z. Cheng, C. Zou, and J. Dong, "Outlier detection using isolation forest and local outlier," in *Proceedings of the 2019 Research in Adaptive and Convergent Systems, RACS 2019*, Association for Computing Machinery, Inc, Sep. 2019, pp. 161–168. doi: 10.1145/3338840.3355641.

[17] A. X. Wang, S. S. Chukova, and B. P. Nguyen, "Ensemble k-nearest neighbors based on centroid displacement," *Inf Sci (N Y)*, vol. 629, pp. 313–323, Jun. 2023, doi: 10.1016/j.ins.2023.02.004.

[18] M. Kumari and S. Soni, "A Review of classification in Web Usage Mining using K-Nearest Neighbour," 2017. [Online]. Available: http://www.ripublication.com

[19] , Matthew James , Melanie Davies , Kamlesh Khunti, Mike Catte , Tom Yates , Alex Rowlands , Evgeny Mirkesg Petra Jones, "a new outlier detection mothod for k-means clustering algorithm," *J Biomed Inform*, vol. 104, no. 103397, Apr. 2020.