Using Data Mining in Recommendation Systems

استخدام تقانات التنقيب في البيانات في الأنظمة الناصحة

دراسة أعدت لنيل درجة الماجستير في علوم الويب

إعداد الطالب : عبد الظاهر ماهر السقا

Abdulzaher_81860@svuonline.org

بإشراف الدكتور : محمد مصطفى حجوز

t_mhajooz@svuonline.org

اختصاص : MWS

2021 M – 1442 H

استخدام تقانات التنقيب في البيانات في الأنظمة الناصحة

الملخص

الأنظمة الناصحة هي الأنظمة التي تساعد المستخدمين باتخاذ القرارات في المواضيع والأمور التي تحتاج إلى دعم للقرارات المراد اتخاذها بالإضافة إلى اختيار العناصر الأكثر ملائمة لأذواقهم ورغباتهم من مجموعة كبيرة من العناصر وذلك اعتماداً على تقانات التنقيب في البيانات . حيث أصبحت هذه الأنظمة إحدى الأدوات الأساسية المستخدمة في مواقع التجارة الإلكترونية ومواقع التواصل الاجتماعي ومحركات البحث. ومع ذلك فإن استخدام هذه الأنظمة وهذه التقانات في مختلف المواقع يواجه العديد من المعوقات على سبيل المثال: مشكلة الإقلاع البارد، قابلية التوسع والتشعب.

في هذه الرسالة قام الباحث ببناء نظام ناصح يساعد الطلاب على اتخاذ القرار المناسب في اختيار برنامج الماجستير المناسب لرغباتهم ومؤهلاتهم بالإضافة إلى تقديم مقترح حل لمشكلة الإقلاع البارد التي من الممكن أن توجد بالنظام، كما قمنا بالمقارنة بين عدة خوارزميات لقواعد الربط المعرفية وذلك من أجل اختيار الخوارزمية الأفضل لاستخدامها في حل مشكلتي قابلية التوسع والتشعب.

الكلمات المفتاحية: الأنظمة الناصحة، التنقيب في البيانات ، الأنظمة الناصحة في التعليم.

Using data mining in recommendation system

Abstract

Recommendation systems are systems that help users to make decisions on topics and matters that need support for the decisions to be taken, in addition to choosing the most appropriate elements for their tastes and desires from a wide range of elements, based on data mining techniques. These systems have become one of the primary tools used in e-commerce sites, social networking sites, and search engines. However, the use of these systems and technologies in different locations faces many obstacles, for example: the problem of cold boot, scalability and ramification.

In this thesis, the researcher has built a recommending system that helps students to make the appropriate decision in choosing the appropriate master's program for their desires and qualifications, in addition to presenting a proposal for a solution to the problem of cold boot that may exist in the system, and we have also compared several algorithms of cognitive association rules in order to choose the algorithm Best to use in solving my scalability and bifurcation problems.

Keywords: Recommendation systems, data mining, recommending systems in education.

المقدمة :

مع النمو السريع لتكنولوجيا المعلومات، يشهد العصر الحالي زيادة هائلة في توليد وجمع البيانات على شبكة الإنترنت وبالتالي أصبح تقديم المعلومات الصحيحة للشخص المحتاج لها أكثر صعوبة يوما بعد يوم، وهذا بدوره يزيد من تعقيد عملية اتخاذ القرار.

وتعد المنظمات التعليمية من أهم أجزاء المجتمع التي تلعب دوراً حيوياً في نمو وتنمية أي دولة. لذا فإن تطبيق التقنيات الحديثة هو عامل أساسي لمواكبة التطور ومتابعة النمو لتحقيق أفضل النتائج المستقبلية والذي بدوره هدف أي مجتمع لذلك توجهت العديد من الانظار نحو تقنيات استخراج المعرفة من البيانات الحالية وتحليلها و الاعتماد عليها كأداة استراتيجية للإدارة و الخطط مستقبلية .

ومن هنا تطرقت الرسالة إلى البحث في استخدام تقانات التنقيب في البيانات في الأنظمة الناصحة والتي أصبحت تستخدم بشكل كبير في العديد من المجالات سواء في مواقع الويب مثل مواقع التواصل الاجتماعي، أو مواقع التجارة الإلكترونية، أو في مجال المواقع الكبيرة جداً والتي تتميز ببيئة عمل معقدة ولديها قواعد بيانات ضخمة.

تعتبر الأنظمة الناصحة أداة مفيدة في التقليل من الحمل الزائد للمعلومات وتساعد إلى نحو كبير في المواقع التي تحتاج إلى نظم دعم القرار التي تتميز ببيئة عمل معقدة أيضاً. كما أنها تساعد في تقديم تنبؤات شخصية للمستخدم حول العناصر التي من الممكن أن يرغب بتصفحها أو انتقائها أكثر من غيرها.

أصبحت الأنظمة الناصحة الأداء الأساسية في معظم التطبيقات والمواقع الالكترونية، الأمر الذي دفع إلى البحث ملياً في تطوير مثل هذه الأنظمة وتحسين أدائها نظراً للفوائد التي تقدمها مثل هذه الأنظمة فهي تؤمن تنبؤات مسبقة لمستخدمي المواقع التي تحتاج إلى دعم القرار حول الخدمات أو المنتجات التي قد يرغبون باستخدامها أو شرائها أكثر من غيرها، وبالتالي أصبحت توفر على المستخدم الكثير من الوقت والجهد، كما أنها

تزيد من ثقة المستخدم بموقع الويب وذلك عندما تكون التنبؤات المقدمة للمستخدم دقيقة بحيث تنال قبوله وتطابق رغباته، كما أن الأنظمة الناصحة تساعد في زيادة الأرباح لدى المواقع التي تقدم خدمات أو منتجات للمستخدمين وذلك من خلال زيادة عدد المستخدمين في الموقع، والقيام بعملية تسويق للمنتجات والخدمات للمستخدمين المناسبين، كما أن الأنظمة الناصحة تعطي تصوراً واضحاً حول المناقلات التي تتم ما بين المستخدم والموقع وبالتالي تساعد إلى حد كبير في تنظيم وتطوير بنية الموقع الإلكتروني، حيث يشكل هذا قاعدة أساسية تعتمد عليها جميع عمليات التطوير التي سوف تتم على الموقع.

ولكن على الرغم من جميع الفوائد التي قدمتها الأنظمة الناصحة، وعلى الرغم من استخداماتها الواسعة في مجالات عديدة، إلا أنها لا تزال تعاني من العديد من المشاكل، والتي تم التطرق إليها في هذه الرسالة، حيث تناولت الرسالة أهم المشاكل التي قد تؤثر على أداء ودقة النظام الناصح، حيث تطرقنا في بحثنا حول مشكلة الإقلاع البارد التي تعترض الأنظمة الناصحة والتي تنتج عند دخول مستخدم جديد إلى الموقع أو عند إضافة اختصاص جديد إلى الموقع، حيث أن المستخدم الجديد للموقع لن يحصل على تنبؤات دقيقة، وذلك لأن النظام الناصح ليس لديه المعلومات الكافية عن سلوك المستخدم وميوله، لذلك سيمر المستخدم بفترة زمنية تكون فيها التنبؤات المقدمة له غير دقيقة بالقدر المطلوب، كما أن الاختصاص الجديد والذي ليس لديه معدلات بعد، فإنه لن يتم وضعه على قائمة التنبؤات المقدمة للمستخدمين (الطلاب) إلا بعد تجاوزه لحد العتبة (عدد المعدلات المقبول)، كما تطرق البحث في الرسالة إلى مشكلة قابلية التوسع ومشكلة التشعب وغيرها من المشاكل التي تظهر في الأنظمة الناصحة، حيث تم طرح حل لمشكلة الإقلاع البارد وكذلك لمشكلة قابلية التوسع، وتمت مناقشة دقة هذه الحلول ومقارنتها مع الحلول المقدمة سابقاً.

دراسات سابقة:

1- قام الباحثون G Geetha, M Safa, C Fancy and D Saranya في عام 2018 باقتراح نظام توصية للأفلام لديه القدرة على التوصية بالأفلام لمستخدم جديد بالإضافة إلى المستخدمين الحاليين الآخرين بالاعتماد على عينة مأخوذة من قاعدة البيانات Movielens. حيث يقوم بالتنقيب في قواعد بيانات الأفلام لجمع جميع المعلومات المهمة ، مثل الشعبية والجاذبية ، المطلوبة للتوصية. نحن نستخدم التصفية القائمة على المحتوى والتعاونية وكذلك التصفية المختلطة ، والتي هي مزيج من نتائج هاتين التقنيتين ، لإنشاء نظام يوفر توصيات أكثر دقة فيما يتعلق بالأفلام.[1]

2- كما قام الباحثون M. Gupta, A. Thakkar, Aashish, V. Gupta and D. P. S. Rathore في عام 2020 بالاعتماد على عينة مأخوذة من قاعدة بيانات Youtube باستخدام خوارزميات K-NN والتصفية التعاونية لبناء نظام ناصح يقوم بتحسين الدقة والأداء لتقنية التصفية المنتظمة بحيث يتم التركيز بشكل أساسي على تعزيز دقة النتائج مقارنة بالتصفية القائمة على المحتوى. يعتمد هذا النهج على تشابه جيب التمام باستخدام خوارزمية الجار الأقرب بمساعدة تقنية التصفية التعاونية ، وفي نفس الوقت إزالة عيوب التصفية القائمة على المحتوى.[2]

هدف البحث:

هذه الدراسة جاءت لتبين أثر استخدام تقنيات التنقيب في البيانات وتطبيقها على بيانات السجل لموقع إحدى الجامعات والتحسينات التي من الممكن أن تضيفها من خلال إمكانية تحليل بيانات الطلاب وتصنيفهم إلى مجموعات متشابهة الأمر الذي يسهل على الجامعة معرفة سلوكيات فئات الطلاب المختلفة على حسب اهتماماتها، وبالتالي المساعدة على اتخاذ للقرارات الصائبة والملبية لحاجات التطوير وتحسين البرامج

التعليمية و الأداء وتطوير سير العمل. وذلك من خلال بناء نظام ناصح يساعد الطلاب في اتخاذ قراراتهم وخالٍ من مشكلتي الإقلاع البارد وقابلية التوسع والذي سيؤدي في النهاية إلى تسهيل عمليات الوصول إلى تنبؤات دقيقة.

مواد وطرق البحث:

- **الأنظمة الناصحة**

تقدم الأنظمة الناصحة خدماتها إلى ثلاث جهات أساسية **[3]** وهي:

1- **المستخدم** : من خلال تقديم اقتراحات تفيد في دعم قرار المستخدم أو تنبؤات له حول المنتجات التي قد يرغب بشرائها.
2- **البائع**(الجهة التي تعرض خدماتها أومنتجاتها إلكترونياً)
3- **الموقع** : (في حال كان الموقع يقدم خدماته باتجاهين للمستخدم والزبون) زيادة عدد زوار ومستخدمي الموقع الأمر الذي يحقق له شهرة كبيرة من جهة، و زيادة عدد الشركات التي تقوم بعرض منتجاتها وخدماتها فيه وبالتالي زيادة نسبة الأرباح في الموقع من جهة أخرى.

بشكل عام واعتماداً على ما سبق فإنه يمكن تقسيم طريقة جمع الموقع للمعلومات عن مستخدميه إلى قسمين رئيسيين هما: الشكل الصريح (explicitly) والشكل المضمن أو الضمني (Implicity)**[6]**:

المعايير	الشكل الصريح	الشكل الضمني
حجم البيانات	قليلة	كثيرة
نقاط القوة	تعطي بيانات مباشرة ودقيقة [3]	تقلل من الحمل على المستخدم [3]
نقاط الضعف	• عبء على المستخدم • مشكلة ندرة التقييم (ليس بالضرورة أن يتفاعل المستخدم مع الموقع ويعطي معدلات للمنتجات)	• تستهلك زمناً كبيراً في عمليات جمع المعلومات عن المستخدم ووضعها ضمن مسار خاص به في قاعدة البيانات • تكاليف acquisition عالية • عدم القدرة على التأكد من صحة البيانات
التحسينات التي يمكن أن تتم عليها	تحسين واجهات التقييم المعروضة للمستخدم بحيث تدفعه باتجاه إعطاء رأيه بالمنتجات أو الخدمات و بأداء النظام الناصح[7]	تحسين خوارزميات حساب التقييم بحيث تقلل زمن الحساب قدر الإمكان[7]

وعليه بناءً على هذا المنظور الشخصي يمكن تقسيم الأنظمة الناصحة إلى عدة أنواع وهي**[3]**:

1- الأنظمة الناصحة المعتمدة الفلترة المعتمدة على المحتوى.
Content-based Filterin RS

2- الأنظمة الناصحة المعتمدة على الفلترة التراكمية.
Collaborative-based Filtering RS

3- الأنظمة الناصحة المعتمدة على المعلومات السكانية.
Demographic RS

4- الأنظمة الناصحة المعتمدة على المعرفة.
Knowledge-based RS

5- الأنظمة الناصحة المعتمدة على المجتمع.
Community-based RS

6- الأنظمة الناصحة الهجينة (المركبة).
Hybrid RS

فوائد الأنظمة الناصحة

1. زيادة نسبة مبيعات الموقع [3]
2. تؤمن الأنظمة الناصحة تنبؤات مسبقة للمستخدم
3. تحافظ على وقت المستخدم [8]
4. تساعد المستخدم في اتخاذ القرار المناسب [4]
5. تزيد من ثقة المستخدم بالموقع [3]
6. تساعد على تحديث قاعدة البيانات الخاصة بالموقع
7. تساعد في تطوير بنية الموقع الإلكتروني

مشاكل الأنظمة الناصحة[3]:

1- التشعب Sparsity

2- قابلية التوسع Scalability

3- الإقلاع البارد Cold Start-Up

وتتضمن : مشكلة المستخدم الجديد – مشكلة المنتج الجديد – إقلاع النظام لأول مرة

4- حجم التخزين Storage Size

5- عرض التنبؤات Recommendations Presentation

خصائص النظام الناصح[3]:

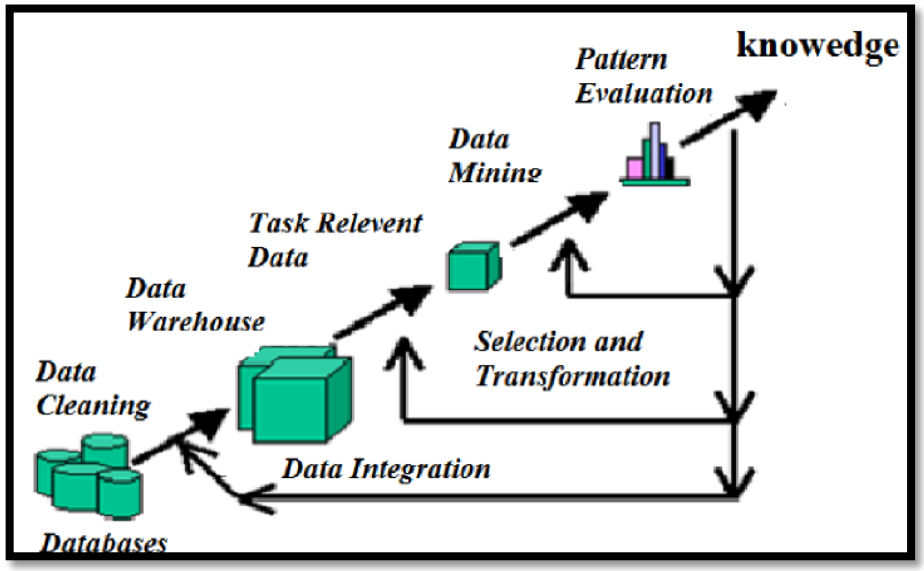
1- تفضيلات المستخدم	**User Preferences**
2- دقة التنبؤ	**PredictionAccuracy**
3- التغطية (الشمولية)	**Coverage**
4- التنوع	**Diversity**
5- المنفعة	**Utility**
6- الخطر	**Risk**
7- المتانة	**Robustness**
8- الملاءمة	**Adaptivity**
9- السرية أو الخصوصية	**Privacy**
10-الطرافة	**Novelty**

- **التنقيب في البيانات في قواعد البيانات**

تتألف عملية التنقيب في البيانات في قواعد البيانات من ثلاث مراحل رئيسية وهي **[11]:**

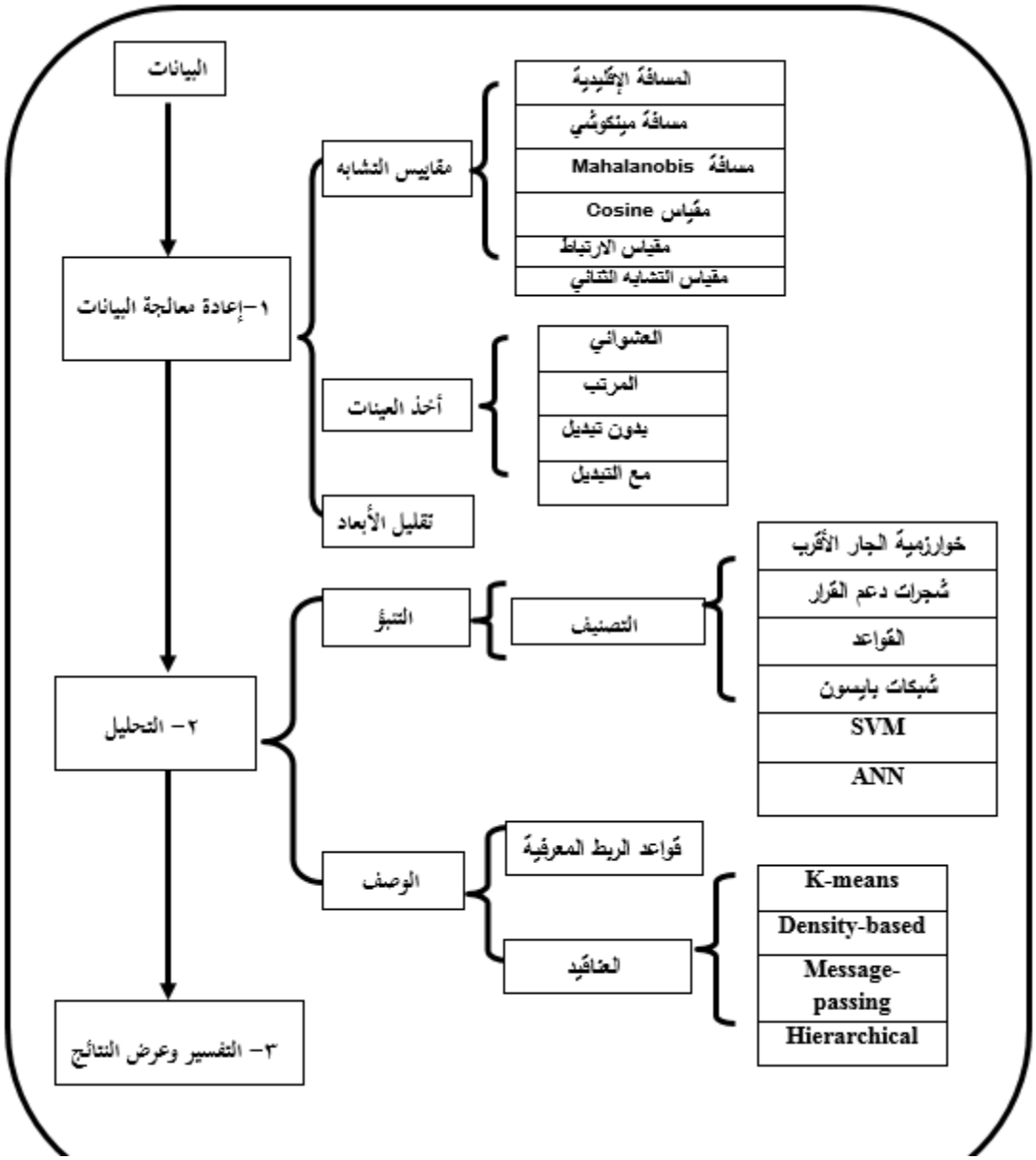
1) مرحلة المعالجة المسبقة للبيانات (data preprocessing)، وتتضمن المراحل التالية:
   1. فلترة البيانات (تنظيف البيانات) (data cleaning).
   2. مكاملة البيانات (data Integration).
2) مرحلة تحليل البيانات (data analysis)، وتشمل المراحل التالية:
   1. مرحلة اختيار وتحويل البيانات (data selection and transformation).
   2. التنقيب في البيانات (data mining).
3) مرحلة تفسير وعرض النتائج (result interpretation).

يوضح الشكل التالي البنية الأساسية لتقنية التنقيب في البيانات في قواعد البيانات**[12]**:



الشكل(1.2) البنية الأساسية لتقنية التنقيب في البيانات في قواعد البيانات

سنقوم باستعراض هذه المراحل، والتقنيات المستخدمة في كل مرحلة على حدا. يوضح الشكل(2) التقنيات المستخدمة في كل مرحلة من هذه المراحل **[13]**:



يجد بعض الباحثين أن هذه الخوارزميات تتشابه في أدائها في وقت التنفيذ إلى حد كبير ولكن تختلف في صعوبة تضمين هذه الخوارزميات **[3]**، يمكن القول أن هذا الكلام صحيح في حال كان حجم البيانات التي تتعامل معه هذه الخوارزميات صغير جداً، ولكن الفروقات بين هذه

الخوارزميات ونقاط الضعف والقوة في كل واحدة تظهر عند التعامل مع قواعد البيانات الضخمة حيث قمنا بتحديد بعض الميزات وخاصة فيما يتعلق بالخوارزميات المستخدمة في قواعد الربط المعرفية هذه الميزات هي:

1- البيانات المدعومة من قبل كل خوارزمية من هذه الخوارزميات (ويقصد بها قواعد البيانات).

2- السرعة في المرحلة الأولى: يقصد بالمرحلة الأولى إما تشكيل المجموعات المرشحة في بعض الخوارزميات ومن ثم تشكيل المجموعات التكرارية أو مسح قاعدة البيانات لمرة واحدة كما في خوارزميات أخرى ومن ثم تشكيل المجموعات التكرارية.

3- السرعة في المرحلة الثانية: يقصد بالمرحلة الثانية بناء قواعد الربط بين العناصر.

4- الدقة: والتي تعبر عن مدى دقة قواعد الربط الناتجة عن تطبيق كل خوارزمية من الخوارزميات. وقد اعتمدنا على تطبيق عدة معايير من مقاييس الدقة مثل (....،Recall،F1)

5- عدد مرات مسح قاعدة البيانات: ويقصد بها عدد المرات التي تقوم بها كل خوارزمية بإجراء مسح كامل أو جزئي لقاعدة البيانات حتى الحصول على قواعد الربط.

6- يظهر الجدول التالي مثالاً توضيحياً لعينة أخذت من قاعدة بيانات معتمدة على المناقلات، حيث E،D،C،B،A هي منتجات موجودة في قاعدة البيانات، والمقصود بالمناقلات هو سجل المشتريات الخاص بزبون ما، فعلى سبيل المثال تدل المناقلة رقم /2/ الموجودة في العينة لدينا أن المستخدم قام بشراء المنتجات B،C،E.

الجدول (1) يتضمن عينة من قاعدة البيانات

الرقم	المناقلات
1	A C D
2	B C E
3	A B C E
4	B E

حيث تم استعراض آلية تعامل كل خوارزمية من هذه الخوارزميات مع نفس العينة من قاعدة البيانات وكانت النتائج على الشكل التالي:

الجدول(2) نتائج المقارنة بين الخوارزميات السابقة من ناحة الدعم للبيانات،السرعة

الخوارزمية	السنة	البيانات المدعومة	السرعة في المرحلة الأولى	السرعة في المرحلة الثانية
AIS	1993	منخفض جداً	بطيئة	بطيئة
SETM	1995	منخفض	بطيئة	بطيئة
Apriori	1996	محدود	سريعة	بطيئة
Apriori TID	1997	كبير	بطيئة	سريعة
Apriori-Hybrid	1999	كبيرة جداً	سريعة	سريعة
FP-Growth	2005	كبيرجداً	سريعة	سريعة

الجدول(3) نتائج المقارنة بين الخوارزميات السبقة من ناحة الدقة، عدد مرات مسح قاعدة البيانات وغيرها

الخوارزمية	الدقة	عدد مرات م قاعدة البيا	المجموعـــــــــ المرشحة لتكون تكرارية	المسحة التخزينية
AIS	منخفضة جداً	M+1	يوجد	كبيرة
SETM	منخفضة جداً	M+1	يوجد	كبيرة
Apriori	منخفضة	M+1	يوجد	كبيرة جداً
Apriori TID	أكثر دقة من Apriori	M	يوجد	كبيرة
Apriori – Hybrid	أكثر دقة من Aprioriوأقل دقة منFP–Growth	M	يوجد	كبيرة
FP–Growth	دقة عالية	2	لا يوجد	أقل من كل ما سبق

النتائج ومناقشتها:

من الحلول التي تمت مناقشتها لمشكلة التشعب نذكر:

1– الاعتماد على التشابه بين العناصر (المنتجات) في بداية الأمر بدلاً من التشابه بين مستخدمين، هذه الطريقة مفيدة في عدد كبير من الحالات ولكن ليس دائماً **[3]**

2- الطلب من المستخدم إعطاء معدلات للعديد من العناصر في كل مرة ، وخاصة عند بداية العمل بالنظام الناصح، وذلك حتى يتكون لدينا حد معين من المعلومات (المعدلات) يسمى **بالعتبة[10]**

من الحلول التي تمت مناقشتها لمشكلة قابلية التوسع نذكر:

شبكات بايسيان(Bayesian Network) ، تحليل الشعاع المفرد (SVD : Single Vector Decomposition) ، تخفيض البيانات، اختيار عينة عشوائية من المستخدمين وحساب التشابه معهم، تركيز البيانات، تصنيف البيانات إلى أنواع وأصناف (clustering)، .....إلخ [2] .

إن كل الطرق والتقانات السابقة تركز في حل مشكلة قابلية التوسع على فكرة تحسين أداء النظام الناصح بغض النظر عن دقة هذه التنبؤات. فمثلاً: شبكات بايسيان مناسبة للأنظمة الناصحة التي تعمل ضمن بيئات عمل يكون تغير اهتمامات الزبون بطيئاً، وهي غير مناسبة للعمل ضمن بيئات يتغير فيها حاجات الزبون واهتماماته بشكل دوري وسريع لأن ذلك يضعف أداء شبكات بايسيان بشكل كبير**[9]**.

أما الطرق المعتمدة على تصنيف المنتجات إلى أنواع (clustering) فهي تساعد بشكل كبير في تحسين أداء النظام إلا أنها تعاني من دقة منخفضة في التنبؤات من جهة، ومن جهة ثانية فإن اهتمامات الزبون لا تنحصر ضمن نوع واحد من المنتجات الأمر الذي يزيد من تعقيد هذه الطرق ويضعف أدائها. كما يمكن تحسين دقة هذه التنبؤات من خلال استخدام تحليل الشعاع المفرد SVD والذي يركز على دقة التنبؤات أكثر من أداء النظام و كفاءته. أما طرق تخفيض البيانات أو التركيز على البيانات فهي تضعف دقة التنبؤات المقدمة بشكل كبير، لأن حجماً كبيراً من المنتجات لن يكون موجوداً في عمليات حساب التشابه بين مستخدمين وبالتالي فإن التنبؤات المقدمة لن تكون دقيقة بالقدر المطلوب، كما أن بعض المنتجات لن تكون أبداً ضمن قائمة التنبؤات لأي مستخدم، بينما سرعة الحصول على التنبؤات كبيرة وبالتالي أداء النظام سيكون مقبولاً بشكل كبير.

من الحلول التي تمت مناقشتها لمشكلة الإقلاع البارد نذكر:

1- استخدام تقنية الأنظمة الناصحة الهجينة**[15]**، بمعنى أنه في حالة إضافة مستخدم جديد فإنه بدلاً من استخدام الفلترة التراكمية في البداية، يتم استخدام تقنية الفلترة المعتمدة على المحتوى لأنها لا تتطلب معلومات كثيرة عن المستخدم، حيث إنها تقوم على حساب التشابه بين العناصر (المنتجات) وليس بين المستخدمين

2- من الطرق الأخرى للدمج والتي استخدمت في حل هذه المشكلة هو الدمج بين المعلومات السكانية(الجغرافية) وبين الفلترة التراكمية كنموذج أول من جهة **[3]**، وبين المعلومات السكانية والفلترة المعتمدة على المحتوى كنموذج ثانٍ من جهة أخرى، وذلك على التفرع

3- من الحلول الشائعة المستخدمة في المواقع العالمية المشهورة هي استخدام شبكات التواصل الاجتماعي الخاصة بالمستخدم **[5]**(SNS : Social Network Services)، حيث تستخدم هذه الطريقة بغرض تحليل اهتمامات الزبون، واستخلاص بعض المعلومات المفيدة والتي تشكل قاعدة جيدة في حساب التشابه بين المستخدمين وخاصة عند استخدام الفلترة التراكمية.

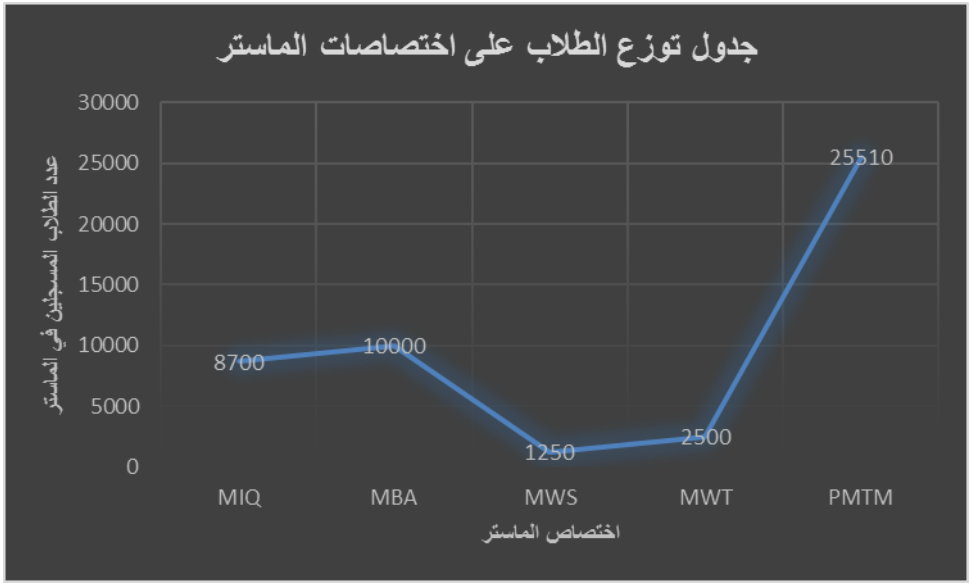
تحسين أداء الأنظمة الناصحة

قاعدة البيانات

سنتحدث عن قاعدة بيانات الجامعة الافتراضية نظراً لكوننا قمنا بتصميم قاعدة بيانات مبسطة عن موقع الجامعة تفيد في مجال الدراسة الخاصة بالبحث، هذه القاعدة تتضمن بيانات للطلاب مع بيانات خاصة بالمواد و الاختصاصات الجامعية واختصاصات الدراسات العليا، وانطلاقاً من مساعدة الطلاب في اتخاذ القرار المناسب في اختيار

الاختصاص المطلوب لابد من بناء نظام ناصح لهذا الموقع يساعد الطلاب على اتخاذ القرار المناسب لاختيار برنامج الماجستير الأنسب لهم والذين يرغبون بدراسته.

حتى نتمكن من معرفة مدى إمكانية الاستفادة من البيانات الخاصة بقاعدة بيانات الموقع لابد من معرفة نسبة الاستحسان لكل ماجستير من قبل المستخدمين في كافة الاختصاصات.



الشكل (1) يبين توزع الطلاب على اختصاصات الماستر

نلاحظ مما سبق أن الموقع يتضمن أراء مختلفة للطلاب وبنسب مختلفة في اختيار الاختصاص المطلوب والتي يمكن استخدامها كمعلومات مرجعية لتقديم تنبؤات مفيدة للطلاب الراغبين بالتسجيل، ولكن هذا المعيار يعتبر غير كافٍ في حال أردنا بناء نظام ناصح لهذا الموقع لا يعاني من مشاكل. فعلى سبيل المثال يمكن ان يتواجد في الموقع عدد كبير من الطلاب القدامى الذين أنهوا دراستهم الجامعية ولكن لم يقوموا بالتسجيل في أي اختصاص من اختصاصات الماستر الموجودة في الموقع، الأمر الذي سيؤثر على

أداء النظام الناصح وخاصة إذا كان هذا النظام يعتمد على خوارزمية الفلترة التراكمية لإعطاء التنبؤات.

كما أنه لابد من معرفة هل جميع اختصاصات الماجستير الموجودة ضمن الموقع لاقت استحسان الطلاب من كافة الاختصاصات الجامعية، لذا قمنا بدراسة توزع اختصاصات الماستر على الاختصاصات الجامعية الموجودة في الجامعة، وكانت على الشكل التالي:

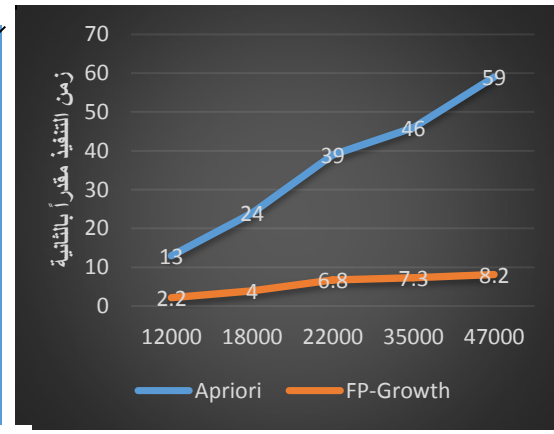
الجدول(4) جدول يوضح توزع المعدلات المعطاة للطلاب المسجلين في الماسترات

Master	Special	Student count	Student percentage
MIQ	BACT	960	2.002%
	BAIT	960	2.002%
	BSCE_Business Management	780	1.626%
	BSCE_Marketing	780	1.626%
	lnformation Technology Engineering AI	960	2.002%
	lnformation Technology Engineering SCN	960	2.002%
	lnformation Technology Engineering SE	960	2.002%
	Management Sciences_Finance and banking	780	1.626%
	Management Sciences_Human Resource Management	780	1.626%
	Management Sciences_Marketing	780	1.626%
MBA	BACT	550	1.147%
	BAIT	550	1.147%
	BSCE_Business Management	1450	3.023%
	BSCE_Marketing	1450	3.023%
	lnformation Technology Engineering AI	550	1.147%
	lnformation Technology Engineering SCN	550	1.147%
	lnformation Technology Engineering SE	550	1.147%
	Management Sciences_Finance and banking	1450	3.023%
	Management Sciences_Human Resource Management	1450	3.023%
	Management Sciences_Marketing	1450	3.023%
MWS	BACT	250	0.521%

	BAIT	250	0.521%
	lnformation Technology Engineering AI	250	0.521%
	lnformation Technology Engineering SCN	250	0.521%
	lnformation Technology Engineering SE	250	0.521%
MWT	BACT	500	1.043%
	BAIT	500	1.043%
	lnformation Technology Engineering AI	500	1.043%
	lnformation Technology Engineering SCN	500	1.043%
	lnformation Technology Engineering SE	500	1.043%
PMTM	BACT	2536	5.288%
	BAIT	2536	5.288%
	BSCE_Business Management	2566	5.350%
	BSCE_Marketing	2566	5.350%
	lnformation Technology Engineering AI	2536	5.288%
	lnformation Technology Engineering SCN	2536	5.288%
	lnformation Technology Engineering SE	2536	5.288%
	Management Sciences_Finance and banking	2566	5.350%
	Management Sciences_Human Resource Management	2566	5.350%
	Management Sciences_Marketing	2566	5.350%

بما أن خوارزميتي Apriroi و FP–Growth تعتبران من أشهر الخوارزميات المستخدمة مع هذا النمط من قواعد البيانات، فلا بد من اختيار الخوارزمية الأفضل والتي تعطي النتائج المرجوة، الأمر الذي دفعنا إلى المقارنة العملية بين أداء هاتين الخوارزميتين على العينة المأخوذة من قاعدة البيانات والتي قمنا بعملية تحليل لها. وذلك بالاعتماد على البرمجيتين **WEKA 3.7.1** و **SPMF 1.2.1**

وقد كانت النتائج على النحو التالي :

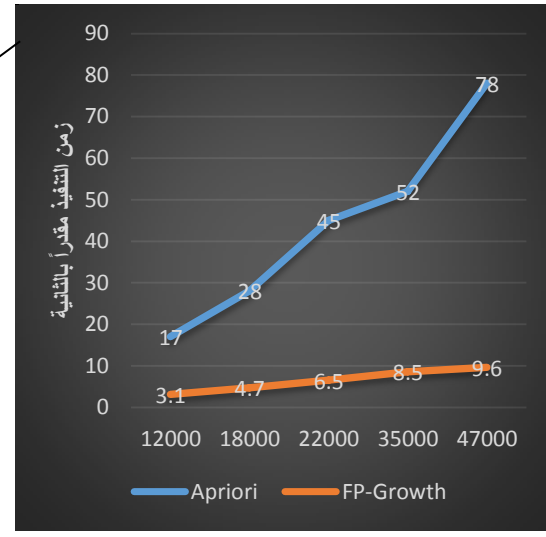


الشكل(2) مقارنة بين زمن تنفيذ كل من FP-Growth،Apriori باستخدام 3اختصاصات وأعداد مختلفة من السجلات

الخوارزمية / عدد السجلات	خوارزمية Apriori	خوارزمية Growth-
	زمن التنفيذ مقدراً بالثانية	
12000	13	2.2
18000	24	4
22000	39	6.8
35000	46	7.3

الجدول(5) مقارنة بين زمن تنفيذ كل من FP-Growth،Apriori باستخدام3 اختصاصات وأعداد مختلفة من السجلات

الجدول(6) مقارنة بين زمن تنفيذ كل من FP-Growth،Apriori باستخدام5 اختصاصات وأعداد مختلفة من السجلات



الخوارزمية / عدد السجلات	خوارزمية Apriori	خوارزمية Growth-
	زمن التنفيذ مقدراً بالثانية	
12000	17	3.1
18000	28	4.7
22000	45	6.5
35000	52	8.5
47000	78	9.6

الشكل(3) مقارنة بين زمن تنفيذ كل من Apriori،FP-Growth باستخدام 5 اختصاصات وأعداد مختلفة من السجلات

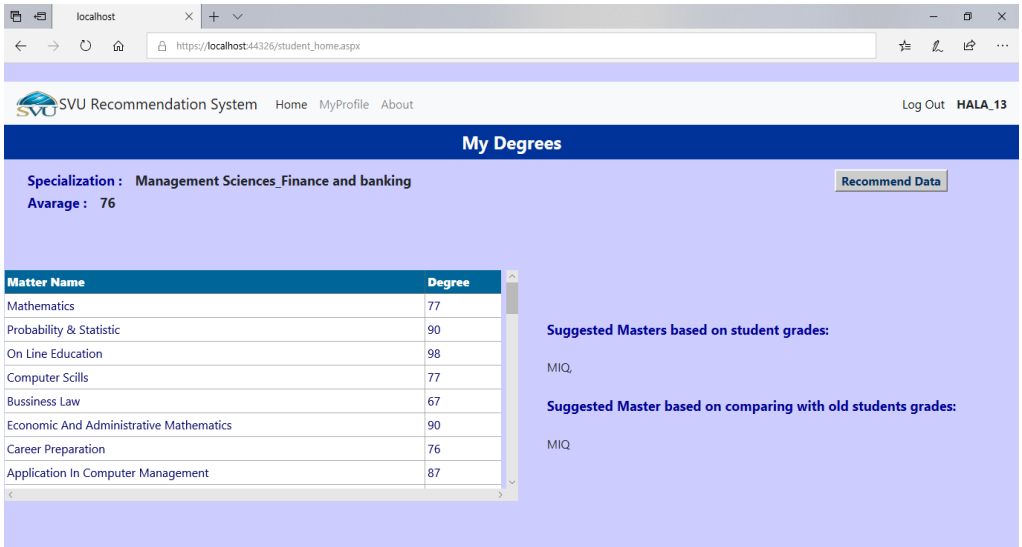
FP–G تعتبر أفضل من Apriori من ناحية التعامل مع حجم قاعدة البيانات ومن ناحية زمن التنفيذ لدينا، لذلك فتعتبر خوارزمية FP–Growth أفضل لحل مشكلتي قابلية التوسع والتشعب في حال التعامل مع مواقع الضخمة.

الحل المقترح لمشكلة الإقلاع البارد:

تعتمد فكرة الحل المقترح على متابعة حركة المستخدم (الطالب) ضمن الموقع وتسجيل تفاعلاته مع الاختصاصات الموجودة في موقع الجامعة وكذلك الأسئلة المطروحة، ويتم تسجيل هذه التفاعلات ضمن قاعدتي بيانات هما:old_student و New_student

في حالة الطالب من ذات الجامعة:

بعد أن يقوم الطالب بتسجيل الدخول إلى حسابه في موقع الجامعة وبعد أن يكون قد أنهى جميع مواده يظهر له المعدل الذي حصل عليه، فيتم بناء على قاعدة البيانات للطلاب القدامى اختيار المعدل الأعلى للماستر الذي تم تسجيله بناء على اختصاصه كتوصية أولى بالنسبة لاختصاصه، واختيار المعدل الأعلى للماستر الذي يستحسن التسجيل به بناء على معدله الذي حصل عليه. وذلك بمجرد النقر على زر التوصية في صفحته عند تسجيل الدخول.



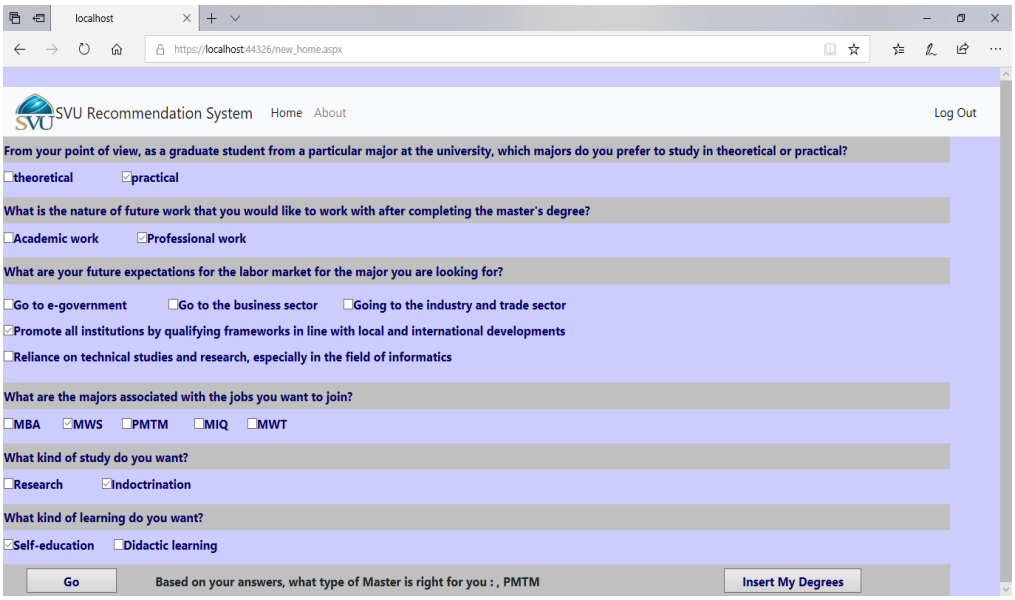
الشكل (4) يبين حالة الطالب من ذات الجامعة

وبعد أن يقوم هذا الطالب بالتسجيل على أحد الماسترات تضاف هذه المعلومات إلى قاعدة البيانات التي تمثل الطلاب القدامى والتي تزداد كلما قام الطلاب بالتسجيل على اختصاصات الماستر.

في حالة الطالب من خارج الجامعة:

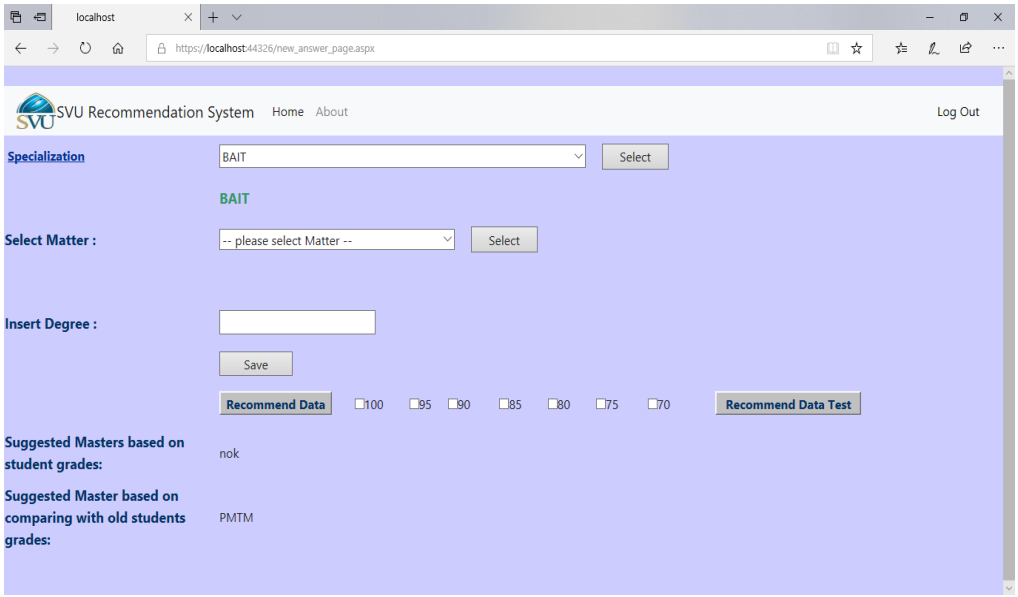
يقوم الطالب بالدخول إلى صفحة الزوار ومن ثم يقوم بالإجابة على مجموعة من الأسئلة وهي مرحلة قيام الطالب بالتفاعل مع ماستر ما (النقر على الإجابة الخاصة بالماستر والتي تعكس خلفية الطالب العلمية والمهنية حول الماستر).

في المرحلة الأولى يتم حساب المعدل النهائي للإجابات (الماسترات) التي قام الطالب بجمعها ومن ثم يتم اختيار الماستر الذي يملك أعلى معدل من بين باقي الماسترات وهو الماستر الذي يستوصى به كمرحلة أولى بناء على الأسئلة الموجهة.



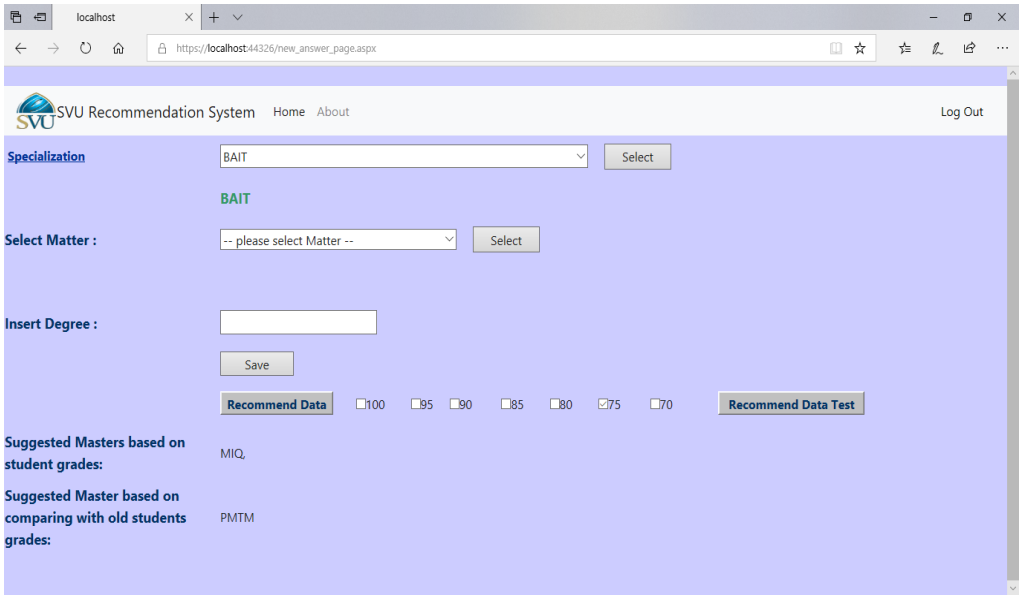
الشكل (5) يبين المرحلة الأولى في حالة الطالب من خارج الجامعة

في المرحلة الثانية يتم تحديد اختصاص الطالب ويتم تقديم التوصية بناء على قاعدة البيانات للطلاب القدامى اختيار المعدل الاعلى للماستر الذي تم تسجيله بناء على اختصاصه ومقارنة علاماته مع علامات الطلاب القدامى كتوصية ثانية بالنسبة لاختصاصه.



الشكل (6) يبين المرحلة الثانية في حالة الطالب من خارج الجامعة

في المرحلة الثالثة يتم وضع درجات الطالب في المواد التي تظهر له بناء على اختصاصه الذي وضعه ومن ثم يتم احتساب المتوسط الحسابي لعلاماته والذي بناء عليه يتم اختيار المعدل الأعلى للماستر الذي يستحسن التسجيل به كتوصية ثالثة بناء على معدله الذي حصل عليه.



الشكل (7) يبين المرحلة الثالثة في حالة الطالب من خارج الجامعة

مع العلم أنه في المرحلة الأولى والثالثة من الممكن أن يقوم الطالب بالتفاعل مع برنامج ماستر جديد لا يوجد ضمن قاعدة البيانات للطلاب القدامى وبالتالي التخلص من فترة الاقلاع البارد من ناحية ومن ناحية أخرى وجود التنوع في التنبؤات المقدمة.

يتميز الحل المقترح بما يلي:

1- الحل المقترح يعطي نتائج لا تقتصر على نوع واحد، إنما تتضمن برامج ماستر من اختصاصات جديدة أخرى فإما أن يكون برنامج الماستر جديداً أو الاختصاص الجامعي المسجل جديداً (وهذه ميزة مهمة جداً)حيث نرى أن هذا الحل يقدم حلاً لمشكلة الاقلاع البارد بالنسبة للمستخدم الجديد وبالنسبة لبرنامج الماستر الجديد، حيث إن عدداً كبيراً من الحلول المقدمة لمشكلة الإقلاع البارد تعطي حلولاً تقدم قائمة تنبؤات من نمط واحد.

2- يشكل الحل المقترح حل لمشكلة تسمى Cold Start-up Continous User(COCOU) ، والتي تظهر عندما يتوجه الطالب إلى برنامج ماستر يمكن التسجيل عليه من خلال مجموعة من الاختصاصات الجامعية يختلف عن برنامج الماستر الذي

تتوجه إليه ذات الاختصاصات والتي يكون اختصاص الطالب من ضمنها ولكن لم يتم تسجيل برنامج الماستر المذكور من قبل إحدى هذه الاختصاصات، في هذه الحالة تقدم خوارزمية CF نتائج غير دقيقة لأنها سوف تعتمد على المعدلات السابقة التي أعطاها المستخدم لمنتجات من أنماط تختلف عن النمط الجديد الذي قام المستخدم باختياره.

3- قائمة التنبؤات المقدمة للمستخدم تتميز بالدقة المتزايدة بمرور الزمن، حيث إن عدد المستخدمين المشابهين للمستخدم الحالي سوف يقل بمرور الزمن نتيجة الزيادة في عدد ظهورات الأنماط التي قام المستخدم باختيارها، أي سنحصل على مستخدمين يشابهون المستخدم الحالي بنسب متزايدة.

الاستنتاجات والتوصيات:

تعرفنا في هذه الدراسة على الأنظمة الناصحة وعلى بعض المشاكل التي تعاني منها وعلى الأخص مشكلة الإقلاع البارد ،التشعب وقابلية التوسع، كما تطرقنا إلى تقنيات التنقيب في البيانات في الأنظمة الناصحة ومراحلها، إضافة إلى المقارنة التي أجريناها بين خوارزميات قواعد الربط المعرفية من خلال استخدام عدة معايير مثل الدقة والسرعة وحجم قاعدة البيانات التي تدعمها كل خوارزمية، بالإضافة إلى استخدام برمجيتين مفتوحتي المصدر وهما SPMF ،WEKA حتى وصلنا إلى وضع حل لمشكلة الإقلاع البارد والتأكد من دقة النتائج. وتوصلنا إلى النتائج التالية:

1- طبيعة النظام الناصح ومتطلباته وحجم قاعدة البيانات الخاصة به تلعب الدور الأساسي والمهم في اختيار الخوارزمية الأفضل من خوارزميات قواعد الربط المعرفية.

2- تعتبر خوارزمية FP-Growth أفضل خوارزمية من خوارزميات قواعد الربط المعرفية سواء أكان من ناحية دقتها أو دعمها الكبير لقواعد البيانات الضخمة أو من خلال السرعة، لذلك تعتبر هذه الخوارزمية هي الأفضل للاستخدام في مجال الأنظمة الناصحة الخاصة بمواقع التجارة الإلكترونية.

3- يعتبر تطبيق خوارزميةFP-Growth على الأنظمة الناصحة ذو فعالية كبيرة جداً، فهي في البداية تعطي نتائج دقيقة من ناحية ربط المستخدمين والمنتجات بقواعد ربط ذات دقة عالية، الأمر الذي يمكن من استخدامها في إعطاء تنبؤات دقيقة، كما أن تطبيق هذه الخوارزمية يعالج مشكلة التشعب التي تعاني منها الأنظمة الناصحة وذلك من خلال القيام بعملية التخلص من المنتجات التي لها معدلات =0، وأيضاً تشكل هذه الخوارزمية حل لمشكلة قابلية التوسع التي تعاني منها الأنظمة الناصحة وذلك لأن أداء هذه الخوارزمية لا يتأثر بشكل كبير نتيجة زيادة حجم قاعدة البيانات هذا من جهة، ومن جهة أخرى فإنه مهما كبر حجم قاعدة البيانات فإن التعامل معها سوف يكون سهلاً كون هذه الخوارزمية تقوم بترتيب قاعدة البيانات لدينا ضمن بنية الشجرة.

4- إن تطبيق الحل المقترح لحل مشكلة الإقلاع البارد يساعد بشكل كبير في تقديم قائمة تنبؤات للطلاب تتميز بالدقة التي تتزايد بمرور الزمن، كما أن هذا الحل يحسن الأداء من ناحية عدد مرات مسح قاعدة البيانات، إضافة إلى أن التنبؤات المقدمة تتميز بارتباطها بالواقع العملي وهذا الأمر الذي يميز هذا الحل.

5- من الممكن أن يتم تطوير النظام ليشمل المرحلة الجامعية وذلك بإضافة بعض المعايير التي تعتمد على علامات الاختصاص للمواد التخصصية في المفاضلة على الاختصاص الجامعي المطلوب.

المراجع:

[1]GEETHA G, SAFA M, FANCY C and SARANYA D , 2018- **A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System**

[2] GUPTA M, THAKKAR A, AASHISH V, Gupta and D. P. S. Rathore,2020- **"Movie Recommender System Using Collaborative Filtering**," *(ICESC).* pp. 415-420

[3] RICCI، L. ROKACH، SHAPIRA، R 2011- **Recommender Systems Handbook.**SpringerNewYork، 845P.

[4]Mahmood، T.، Ricci، F.،2013 -**Learning and adaptivity in interactive recommender systems**. ICEC '07: Proceedings of the ninth international conference on Electronic commerce،vol 1.75–84.

[5]Panagiotis Symeonidis،Alexandros Nanopoulos، and Yannis Manolopoulos.2010-**Aunified framework for providing recommendations in social tagging systems based on ternary semantic analysis**. IEEE Transactions on Knowledge and Data Engineering،vol 2.301-315.

[6]Massa، P.،&Avesani، P.2012-**Trust-aware recommender systems**.ACM Conference on Recommender Systems vol 15.90-113.

[7]Adomavicius، G.،Tuzhilin، A.2012-**Toward the next generation of recommendersystems:a survey of the state-of-the-art and possible extensions**. IEEE Transactions on Knowledge and Data Engineering 17، 734–749 (2012)

[8]Koychev، I.، Schwab، I.2010-**Adaptation to drifting user's interests**. Proceedings of ECML 2010 Workshop: Machine Learning in NewInformationvol 16.39–46

[9] Miyahara، K.،and Pazzani، M.J.2010- **Collaborative filtering with the simple Bayesian classifier**. In Pacific Rim International Conference on Artificial Intelligencepp200-201.

[10] B.Ricci S. Heckerman،2014-**Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering**. In Proceedings of the Fifth International Conference on Computer and Information Technology.vol 4-13-29.

[11]HanJiaweiC،KamperMichelineR،PeiJian،K 2011- **Data Mining Concepts and Techniques**.Morgan Kaufman، 3rdUSA، 740p.

[12] MARGARET،T 2007- **Data mining Introductory and advanced topics.** Pearson Education،2rd. London، 488 p
[13] Hand D، MannilaHeikki P، Smyth ،P 2009- **Principles of Data Mining.** MIT Press، USA، 564p