مقاربة لتكامل تقنيات الترشيح مع محركات البحث

م. خلدون مطر* د. مادلين عبود** د.صلاح الدوه جي***

المُلخص

تشكلُ محركاتُ البحث اليوم محوراً أساسيًا، وعاملاً مهماً للتعامل مع المحتوى الرقمي للمعلومات، الذي يزداد حجمه يوماً بعد يوم، وتَعودُ أهمية محركات البحث إلى أنها تكادُ تكون أهمُ وسيلةٍ للبحث على شبكة الإنترنت؛ ولكن نمو حجم المعلومات التي تديرها، وتسترجعها محركات البحث؛ أصابت المستخدم بحالة من طوفان المعلومات (information overloading)، وترافق ذلك مع ازدياد عدد المستخدمين، وتنوع اتجاهاتهم، فضلاً عن رغبتهم في الوصول إلى المعلومة الموائمة الأكثر دقة بأقل جهد- هذه العوامل كلها دَعت إلى استخدام تقنيات إضافية داعمة لتقنيات استرجاع المعلومات ومنها تقنيات الترشيح بهدف زيادة فعالية محركات البحث وتحسين أدائها.

عَملت عدة بحوث سابقة في هذا السياق، منها ما رَكز على الترشيح المرتكز على المحتوى (content based filtering)، ومنها ما رَكز على الترشيح التعاوني (collaborative based filtering)؛ لكن معظم هذه البحوث كانت تُعاني نقاط ضعف عدة كشدة الخصوصية وقصور المحتوى في الترشيح المرتكز على المحتوى، ومشكلة المستخدم الجديد في الترشيح التعاوني، فضلاً عن مشكلات تتعلق بأسلوب التعليم والتكيّف (adaptive) والنمذجة للاحة المستخدم.

نعرض في هذا البحث مقاربة لتكامل محركات البحث مع تقنيات الترشيح، وذلك من خلال علاقة ديناميكية للتهجين بين الترشيح التعاوني، والترشيح المرتكز على المحتوى؛ بهدف التخفيف من المحدوديات السابقة، وتحسين مقاييس الدقة والاستذكار للوثائق المسترجعة. تَستخدم المقاربة المقترحة نموذج أنطلوجي المجال (Domain ontology) في تمثيل لاحة المستخدم (user profile)؛ بهدف الحد من الأخطاء والتشويش الناتجة عن عدّ لاحة المستخدم ككيان واحد كما تَستفيد من تفاعل المستخدم ونشاطه، للقيام بعمليات التعليم والتكيّف المستمر للاحة المستخدم؛ لتعكس بشكل دائم شخصيته وميوله دون الاعتماد على أمثلة تدريبية فقط؛ بهدف تحسين الترشيح، وتلبية حاجة المستخدم بالحصول على المعلومات الموائمة بدقة أكبر.

الكلمات المفتاحية: ترشيح المعلومات، لاحة المستخدم، استرجاع المعلومات، الترشيح المرتكز على المحتوى، الترشيح التعاوني.

*أعد هذا البحث في سياق رسالة الماجستير للمهندس خلدون مطر، بإشراف الدكتورة مادلين عبود والدكتور صلاح الدوه جي.

**قسم هندسة البرمجيات ونظم المعلومات - كلية الهندسة المعلوماتية- جامعة دمشق.

***قسم هندسة البرمجيات ونظم المعلومات - كلية الهندسة المعلوماتية - جامعة دمشق.

1- تمهيد

تشكل فعاليةُ محرك البحث في استِعادة المعلومات الأكثر مواءمةً للمستخدم وترشِيحها؛ المشكلة العلمية الرئيسة في مجال نظم البحث عن المعلومات، وخصوصاً محركات البحث لذلك غالباً ما يُقيَّمُ أداء محركات البحث بقدرتها على استعادةِ، المعلومات الموائمة للمستخدم وترشِيحها بفعالية عالية. وتعاني أنظمة استرجاع المعلومات اليوم من عدة معوقات؛ ناتجة عن عوامل عدة منها:

1. ازديادِ حجم البيانات التي تتعامل معها محركات البحث [6].
2. ازدياد عدد المستخدمين المتفاعلين مع هذه الأنظمة وتنوع اهتماماتهم واحتياجاتهم، إذْ لكل واحد رغبته التي تقوده عند البحث؛ على الرغم من تطابق طلبات الاستعلام في كثيرٍ من الأحيان[10].
3. الاستعلام الذي يُزود به محرك البحث غالباً ما يكون قصيراً نوعاً ما: 77% كلمة واحدة، و32% كلمتان [5]، ولا يستطيع الاستعلام مهما كان أن يعبّر عن كامل وتمام رغبة المستخدم للمعلومة المطلوب استرجاعها.
4. تَسترجع محركات البحث اليوم عدداً كبيراً من النتائج مع أنه غالباً ما يكون عدد قليل منها موائمة لحاجات المستخدم [20]، فعلى سبيل المثال: من أجل أبسط استعلام "information filtering" لمحرك بحث مثل: Google يُعيد قرابة 263,000,000 نتيجة موائمة من وجهة نظر محرك البحث؛ فيصبح المستخدم في حالة غزارة من المعلومات المسترجعة. وذلك يَتطلب منه بذل جهد ووقت ليس بالقليل في تصفح النتائج المسترجعة ومعاينتها بشكل دقيق. تشير الإحصائيات أن نحو 94% من المستخدمين يقومون باستعراض النتائج ومعاينتها الثلاث الأولى فقط[5]؛ لأنّ الغالبيةَ من المستخدمين غير صبورين وليسو مستعدين لقضاء مزيد من الوقت وبذل مزيد من الجهد في عملية استكشاف موائمة النتائج المسترجعة.
5. تحوي الوثائق دائماً كمية أكبر من المعلومات والدلالة مقارنةً بالاستعلام، القصير نوعاً ما؛ لذلك يمكن القول: « المقارنة التي تتم بين الاستعلام والوثائق غير دقيقة بشكل كبير ».

لذلك يمكن أن نعدّ أنَّ ثَمةَ فجوة كبيرة بين المستخدم ومحرك البحث، والاستعلام بمفرده غير قادر على سد هذه الفجوة بشكل كامل؛ وهذا ناتج عن تجاهل محركات البحث التقليدية للاهتمامات والتفضيلات الشخصية للمستخدمين وتَجاهل الجهد المبذول من قبل المستخدمين في أثناء جَلسات البحث؛ للحصول على المعلومة المطلوبة، وبذلك يتم معاملة المستخدمين كلّهم بالطريقة والأسلوب نفسه، فأي استعلام يُزود به محرك البحث من قبل عدة مستخدمين يعيد النتائج المسترجعة نفسها للمستخدمين كلّهم [6].

وهكذا ونتيجة لكل ما سبق ظهرت الحاجة لأنظمة ومحركات بحث ذكية، وتطلّعَ كثير من الباحثين لمحركات بحث تستطيع أن تُوظّف المعلومات الضمنية والصريحة الناتجة عن تفاعل المستخدمين ونشاطهم، وتُعيد صياغة هذه المعلومات المُجمعة ونمذجتها فيما يُعرف بلاحة المستخدم للاستفادة منها في خدمةِ عمليةِ الاسترجاع؛ بترشيح النتائج الأكثر إرضاءً للمستخدم اعتماداً على لاحات المستخدمين وتعاونهم.

2- أنظمة ترشيح المعلومات

2-1- تعاريف الترشيح

لغوياً: لها عدة تعاريف أهمها:

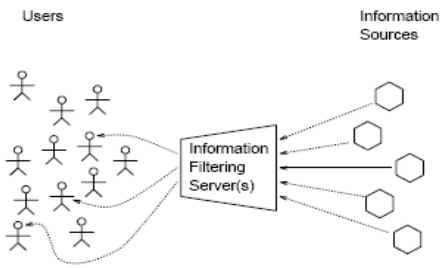
- عملية انتقاء وعرض المعلومات بحيث تلائم وتناسب مستقبل ما.
- عملية تنقيةَ وتهذيب لتدفق كبير من العناصر لتناسب مستقبل محدد.
-عملية تنقية أوتمرير عبر مرشح.

أمَّا في علوم تقانة المعلومات

- فهو عملية اكتشاف المعلومات التي تقابل حاجات المستخدمين من تدفق معلومات ديناميكي [1].
- وهو مجال بحثي يعرض أدوات لتميز بين المعلومات الموائمة وغير الموائمة، وذلك بإضفاء الطابع الشخصي على تدفق مستمر للمعلومات المُسترجعة[13].

أنظمة ترشيح المعلومات

- أنظمة ترشيح المعلومات هي وكلاء برمجية تقوم بمعالجة تدفق كبير من المعلومات المستمرة(stream) بهدف تَوزيع المعلومات وتَسليمها لمستخدم محدد عن طريق جذب المعلومات الأكثر تطابقاً مع اهتمامات وتفضيلات المستخدم واستبعاد المعلومات غير الموائمة [12][3].



الشكل(1) توزيع مصادر المعلومات بحسب المستخدمين [7]

أول ما طُبقت تقنيات الترشيح كانت في مواقع التجارة الالكترونية الضخمة، حيث عدد المنتجات المعروضة كبير جداً ولا يمكن بحال من الاحوال للمستخدم إمكانية استعراضها ومعاينتها للبحث عمَّا يناسبه؛ مما دعى إلى تطوير أنظمة للترشيح بحيث تقوم بانتخاب وفرز المنتجات المناسبة لكل مستخدم بحسب ميوله واهتماماته؛ هذا وفر كثيراً من الوقت والجهد على المستخدم، وساعده في العثور على المنتج المناسب بسرعة أكبر.

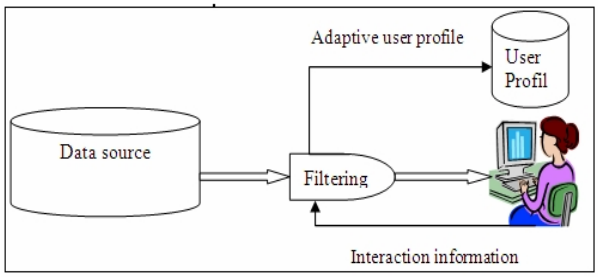
دُعيت هذه الأنظمة بأنظمة التوصية، ولذلك هنالك نوع من التكافؤ بين المصطلحات الآتية: أنظمة الترشيح، أنظمة التوصية، أنظمة إضفاء الطابع الشخصي.

2-2- مكونات أنظمة ترشيح المعلومات

يتألف نظام الترشيح من ثلاثة مكونات رئيسة [1] [19]

- لاحة المستخدم(user profile).
- وكيل برمجي لتعليم وتكيّف لاحة المستخدم.
- الترشيح.

يمكن تحقيق كل مكون من هذه المكونات السابقة بعدة نماذج، كما سنرى في فقرة البحوث المتعلقة لاحقاً، لكن هذه النماذج كلّها تتشابه في الهدف والغاية من كل مكون من مكوناتها.



الشكل (2) البنية العامة لأنظمة الترشيح

2-2-1- لاحة المستخدم

هي نموذج معرفي، أوقاعدة معرفية لتمثيل اهتمامات المستخدم، وتفضيلاته الحالية والماضية على العناصر المدروسة.

يَعكس هذا النموذج طبيعة المستخدم وشخصيته بالنسبة إلى النظام الذي يستخدمه، وكلما كانت لاحة المستخدم مضبوطة وصحيحة أكثر، كانت عملية الترشيح أنجح.

تُبنى لاحة المستخدم من خلال تفاعل المستخدم ونشاطه مع النظام، ومراقبة سلوكه وتفضيلاته على عناصر

المعلومات أوعلى السِمات والملامح المعرفية المستخرجة من هذه العناصر، ومن ثَم تطبيق تقنيات مختلفة(التنقيب في البيانات، التعلم التلقائي، وغيرها .....).

يَجري بناء، ونمذجة لاحة المستخدم بعدة طرائق أشهرها: النموذج الشعاعي حيث تُمثل لاحةالمستخدم [7] كشعاع من عناصر المعلومات أو كشعاع من الملامح والسِمات المستخرجة من عناصر المعلومات.

فعلى سبيل المثال: إذا كانت العناصر المدروسة وثائق نصية، يمكن تمثيل لاحة المستخدم كشعاع من الكلمات المفتاحية المُستخرجة من الوثائق المفضلة للمستخدم

$P=(w1،w2،w3،.....wn)$

يعبّر الوزن عن مدى أهمية الكلمة المفتاحية في لاحة المستخدم.

من أبرز الميزات التي يجب أن تتمتع بها لاحة المستخدم أياً كان النموذج المُتبع في بنائها، هي قُدرتها على التعلم والتكيّف.

2-2-2- وكيل برمجي لتعليم لاحة المستخدم وتكيّفها

للاستفادة من لاحة المستخدم في عمليات الترشيح يجب أن تكون مضبوطة ومعبرة بشكل حقيقي عن شخصية المستخدم المالك لها، وتُمثل اهتمامات المستخدم وميولاته التي يمكن أن تتغير وتتبدل بشكل كبير أوطفيف مع مرور الزمن؛ ولهذا لابدَّ للاحة المُستخدم من أن تَعكس التعديلات التي تطرأ على المستخدم، وهذا ما يسمى بتعليم لاحة المستخدم، وأهم ما يميّز هذه العملية في أنظمة الترشيح أنها مستمرة ولا تقتصر على مدة محددة، لأن اهتمامات المستخدم وتفضيلاته غير ثابتة.

غالباً ما تستفيد هذه العملية من التغذية الراجعة الناتجة عن تفاعل المستخدم ونشاطه مع النظام، بهدف معرفة العناصر الموائمة للمستخدم وتحديد درجة موائمة كل منها بشكل صريح أوضمني[5] .

الطريقة الصريحة: حيث يعبّر المستخدم صراحة عن مدى أهمية النتائج بالنسبة إليه. أهم ميزات التغذية الراجعة الصريحة أنها دقيقة في معرفة درجة الموائمة لكنها غير مرغوب فيها، لأنها مزعجة بالنسبة إلى المستخدم، فضلاً عنأن كثيراً من المستخدمين غير مهتمين بإعطاء تقييمات صحيحة.

أما الطريقة الثانية الضمنية: وهي الأكثر شيوعاً، فترتكز بشكل أساسي على تَتَبع، ومراقبة تفاعل، وسلوك المستخدم بشكل خفي وضمني، في محاولة للحصول على معلومات معينة من المستخدم، والاستفادة منها في معرفة مقدار اهتمام المستخدم وتفضيله للمعلومات المُقدمة له، تمهيداً لاستخدامها في عملية التعليم والتكيّف، بتطبيق تقنيات وأساليب مختلفة.

هنالك عدة عوامل وإشارات ضمنية تُستخدم في هذه الطريقة لمعرفة درجة الموائمة: كالوقت الذي يمضيه المستخدم على نتيجة معينة (time spent) وعدد الزيارات [10]، وعدد النقرات، والاستعلام[5]. أغلب هذه العوامل يتعلق بدراسة ومراقبة أفعال وتفاعل المستخدم مع نتائج النظام.

2-2-3- الترشيح

تقوم عملية الترشيح بغربلة، العناصر المقدمة من مصدرما تمحيصها ونخلها، لتناسب حاجات المستخدمين واهتماماتهم المتمثلة بلاحاتهم.

يتم في عملية الترشيح تطبيق علاقات وتوابع رياضية تختلف بحسب نمط الترشيح المُطبق وصنفه.

تصانيف أساليب ترشيح المعلومات

يمكن تصنيف ترشيح المعلومات بشكل عام في فئتين رئيسيتين بحسب نموذج الترشيح هما: الترشيح التعاوني والترشيح المرتكز على المحتوى[18] [4].

2-2-3-1- الترشيح المرتكز على المحتوى (content based filtering) [22]

لهذا النمط جذور وصفات مشتركة مع نظم استرجاع المعلومات، لأن هذه التقنية تعتمد على فكرة استخلاص السِمات والملامح من العناصر المهمة المفضلة بالنسبة إلى المستخدم والاستفادة منها في عملية بناء لاحة المستخدم. لاحة المستخدم في هذا النوع من الترشيح تعتمد في بنائها بشكل كامل على محتوى العناصر والأغراض المفضلة للمستخدم. والعملية الأساسية التي تتم في هذا النوع من الترشيح بغض النظر عن طريقة التحقيق هي قياس مدى التشابه بين المعلومة المتدفقة ولاحة المستخدم، وتُعدُّ علاقة التجيب بين شعاع لاحة المستخدم وشعاع المعلومة المُتدفقة، أحد أبرز العلاقات المُستخدمة في هذا النمط من الترشيح.

$$Sim(D,P)=\frac{\sum_{i=1}^{n} Wti * Wpi}{\sqrt{\sum_{i=1}^{n} Wti^{2}}\sqrt{\sum_{i=1}^{n} Wpi^{2}}}$$

P: شعاع لاحة المستخدم.

D: شعاع المعلومة المسترجعة.

Wti: وزن السِمة (i) المستخرجة من شعاع المعلومة.

Wpi: وزن السِمة(i) من شعاع لاحة المستخدم.

من أهم نقاط ضعف هذا النوع من الترشيح:

- شدة الخصوصية(over specialization): تعني تقديم النظام للنتائج ذات الارتباط بلاحة المستخدم؛ مما قد يحرم المستخدم من نتائج أخرى أكثر موائمة لحاجته[4].
- قصور المحتوى: قصور أنظمة الترشيح في فهم وتمثيل اهتمامات وتفضيلات المستخدمين في أثناء عملية بناء وتعليم لاحة المستخدمين[4].

2-2-3-2- الترشيح المرتكز على تعاون المستخدمين (collaborative based filtering)[20]

يعتمد على فكرة أن المستخدمين المتشابهين بالاهتمام والتفضيل على بعض العناصر، يمكن أن يتشاركوا أيضاً بالاهتمام والتفضيل على عناصر ووثائق أخرى، وهذا يتيح القيام بعملية ترشيح عناصر وأغراض لمستخدم ما اعتماداً على رأي مستخدمين آخرين مشابهين له ومشورتهم.

كان الترشيح التعاوني بداية" شائع الاستخدام في أنظمة التوصية، وطرح كبديل لأنظمة الترشيح المعتمدة على المحتوى على بعض العناصر التي يصعب تحليل محتواها وفهمها، كالصوت، والصورة، والملتيميديا، وقد أثبت دور فعَّال في محركات البحث من خلال ضبط دقة النتائج المسترجعة بشكل أكبر.

ولابدَّ لهذا النمط من الترشيح من عمليتين أساسيتين:

الأولى: تحديد مجموعة المستخدمين المشابهين لمستخدم ما ونرمز لها بـ (NBS_u)، العلاقة التالية أحد أبرز العلاقات المستخدمة في حساب التشابه بين المستخدمين في الترشيح التعاوني:

$$sim(u1,u2)=\frac{\sum_{i\in I_{u1u2}}[(R_{u1,i}-\bar{R}_{u1})(R_{u2,i}-\bar{R}_{u2})]}{\sqrt{\sum_{}(R_{u1,i}-\bar{R}_{u1})^{2}}\sqrt{\sum_{}(R_{u2,i}-\bar{R}_{u2})^{2}}}$$

$R_{u1,i}$: تقييم المستخدم u1 للعنصر i.

$\bar{R}_{u1}$: متوسط تقييمات المستخدم u1.

والثانية: التنبؤ بمقدار الاهتمام والتفضيل على عنصر ما اعتماداً على مجموعة المستخدمين المتشابهين.

أشهر علاقات الترشيح التعاوني:

$$Pu,i = \overline{Ru} + \frac{\sum_{n \in NBS_u} sim(u,n) * (Rn,i\_\overline{Rn})}{\sum_{n \in NBS_u} (sim(u,n))}$$

NBS_u : مجموعة المستخدمين المشابهين للمستخدمu.

Rn,i : تقييم المستخدم n للعنصر i.

$sim(u,n)$: درجة التشابه بين المستخدمينu و n.

$\overline{Rn}$: متوسط تقييمات المستخدم n.

Pu,i : درجة اهتمام وتفضيل المستخدم u للعنصر i.

يعاني هذا النمط من ضعف عملية الترشيح في حال كان المستخدم جديدً؛ ومن ثمَّ عدم توافر مستخدمين مشابهين ليتم الترشيح التعاوني اعتماداً.

3- محركات البحث وتقنيات الترشيح

3-1- الفرق بين نظم استرجاع المعلومات ونظم ترشيح المعلومات

في مقالة Nicholas بعنوان«أنظمة استرجاع المعلومات وأنظمة الترشيح وجهان لعملة واحدة» [3] أشار إلى أن هناك تطابقاً كبيراً نوعاً ما بين أنظمة ترشيح المعلومات وأنظمة استرجاع المعلومات في عدة نقاط أهمها:

- كلاهما يهدف في النهاية إلى تزويد المستخدم بالمعلومة الموائمة.
- كلاهما يَستخدمان النماذج نفسها في التصميم والبناء على سبيل المثال: النموذج الشعاعي الشائع الانتشار في أنظمة استرجاع المعلومات، يستخدم في كثير من الحالات لنمذجة لاحة المستخدم في أنظمة الترشيح عن المعلومات، وكذلك الأمر بالنسبة إلى النموذج الاحتمالي وغيره.
- كلاهما يَستخدمان المقاييس والمعايير نفسها في تقييم أداء هذه الأنظمة وجودتها، كمقياسي الدقة والاستذكار.

لكن مع هذا كلّه فمن الواضح أن الكثير من بحوث ونماذج نظم استرجاع المعلومات التقليدية تتجاهل بعض المشكلات والقضايا الأساسية التي تُعنى بها أنظمة الترشيح عن المعلومات.

لذلك لابدَّ أولاً من تحديد الفروق والاختلافات بين كل من أنظمة استرجاع المعلومات التقليدية وأنظمة الترشيح ومعرفة نقاط القوة ونقاط الضعف في كليهما، تمهيداً لتحقيق المكالمة والتوافق بينهما، بهدف تحسين أنظمة استرجاع المعلومات، الجدول الآتي يوضح الفروقات الأساسية بينهما مُستخلصة من أغلب المراجع :

أنظمة استرجاع المعلومات	أنظمة الترشيح
تستقبل الاستعلام من المستخدم، وتعيد النتائج الموائمة لهذا الاستعلام.	تعيد النتائج الموائمة لشخصية المستخدم وطبيعته.
لا تأخذ بالحسبان مستخدماً محدداً، تعامل المستخدمين كلّهم بالأسلوب نفسه.	تتنبأ بحاجة المستخدم اعتماداً على لاحة المستخدم.
يُوصف طلب المستخدم ومعلوماته المحتاج إليها بصيغة الاستعلام.	تُوصف معلومات المستخدم وحاجاته في لاحة المستخدم.
المطابقة تتم بقياس درجة تشابه استعلام المستخدم مع ممثل محتوى المعلومات.	الترشيح يتم بقياس درجة التشابه بين لاحة المستخدم مع ممثل محتوى المعلومات.
تقييم النتائج المسترجعة من قبل المستخدم تقود النظام للقيام بعملية التغذية الراجعة.	النظام يراقب سلوك المستخدم في تقييّم النتائج ويستفيد من هذه العملية في تعليم لاحة المستخدم بشكل مستمر.
حاجة المستخدم للمعلومات متغيرة ومتنوعة وتمثل بالاستعلام	حاجة المستخدم للمعلومات مستقرة نوعاً ما وتمثل بلاحة المستخدم.

لكن وبنظرة معمقة ودقيقة، نلاحظ أن نظم استرجاع المعلومات اليوم وخصوصاً محركات البحث أصبحت تشارك نظم الترشيح ببعض الخصائص والميزات، فسيل

المعلومات المتدفق للمستخدم لأجل أبسط استعلام، ونموحجم المعلومات التي تتعامل معه محركات البحث؛ يجعلنا نقول إنَّ مصدر المعلومات لم يعد ثابتاً ومستقراً، بالوقت نفسه دون تعارض، يمكننا أن نقول: «إنَّ حاجة المستخدم المتغيرة والمتمثل بالاستعلام في محركات البحث، يمكن أن تُفهم وتُدرك بشكل أفضل من قبل النظام، إذا فُسِّرَتْ بشكل حذر ودقيق في إطار اهتمامات المستخدم وتفضيلاته» خصوصاً مع ازدياد عدد المستخدمين وتنوع خلفياتهم؛ لذا كان لابد من فهم أكبر للمستخدم لتضييق مجال البحث بالاستعانة بنموذج لاحة المستخدم.

3-2- التكامل بين تقنيات الترشيح ومحركات البحث

ظهرت الحاجة إلى التكامل بين محركات البحث وتقنيات الترشيح بداية الأمر؛ نتيجة عَوز محركات البحث وافتقارها إلى نموذج لاحة المستخدم، فمحركات البحث اليوم مصممة اعتماداً على تقنيات استرجاع المعلومات التي تركز بشكل أساسي على التشابه المباشر بين الاستعلام والوثيقة تُوصف بأنها "one size fits all" [21]: وهو مصطلح يقصد به أن محركات البحث تسترجع النتائج نفسها لأجل استعلام محدد دون اعتبار لحالة وطبيعة المستخدم الذي طلبه، لذلكَ يُمكن اليوم لأكثر من مستخدم أن يكتبَ الاستعلام نفسه لمحرك البحث، وكلُ واحد منهم يقصد ويَنوي به الحصول على معلومات معينة مختلفة عن الآخر[2]، فالاستعلام يظل غامض ما دام لم يتحدد سياق البحث المتعلق بالمستخدم. فعلى سبيل المثال استعلام "jaguar" لمحرك بحث مثل جوجل يمكن أن يُعيد وثائق تتعلق بسيارات وثائق تتعلق بالحيوانات، دونما اعتبار لمجال اهتمامات المستخدم، ولهذا عنّون Sayaka Akioka [15] بحثه « ترشيح المعلومات لحل غموض استرجاع المعلومات »، فقد عدَّ أن بعض مشكلات الغموض في أنظمة استرجاع المعلومات يمكن أن يعالج باستخدام تقنيات الترشيح.

تَكمن أهمية لاحة المستخدم في محركات البحث، في المساعدة في ترشيح وتحسين ترتيب نتائج محرك بحث بما يتلاءم مع سياق البحث واهتمامات المستخدمين، وقد ازدادت الحاجة إلى هذا الجمع بين أنظمة الترشيح ومحركات البحث بسبب عدم استفادة مستخدمي محركات البحث من الجهد المبذول من قِبل مستخدمين آخرين مشابهين في أثناء جلسات بحث، وقيامهم بعملية المحاكمة البشرية لتحديد النتائج الموائمة من غير الموائمة لاستفساراتهم، على مبدأ التعاون والتشاور في الحصول على الأفضل؛ مما يوفر الجهد والوقت، وهذا ما يُعرف بالترشيح التعاوني.

بشكل عام يمكن تَصنيف التكامل بين محركات البحث وتقنيات الترشيح في ثلاثة أنماط رئيسية:

3-2-1- تكامل محركات البحث مع الترشيح المرتكز على المحتوى[7]

لم تَعُد عملية الاسترجاع هنا تعتمد فقط الثلاثية (الوثيقة الاستعلام، تابع المطابقة) بسبب دخول عامل رابع وهو لاحة المستخدم، ويُمثَّلُ نموذج لاحة المستخدم هنا من الملاح والسِمات المستخرجة من الوثائق المفضلة للمستخدم.

يُقاسُ التشابه بين نموذج لاحة المستخدم ووثائق محركات البحث المسترجعة، تُستخدَمُ هذا التشابه في عملية ترتيب النتائج المسترجعة.

3-2-2- تكامل محركات البحث مع الترشيح التعاوني.

تَستثمر هنا محركات البحث جهود المحاكمة البشرية التاريخية المَبذولة من قبل مستخدمين مشابهين، للتمييز بين النتائج الموائمة وغير الموائمة للاستعلامات، مما يعطي الاسترجاع بُعداً دلالياً أكبرَ، ويجري هذا الترشيح

وفق علاقات الترشيح التعاوني المُستخدمة في أنظمة التوصية(الفقرة **2-2-3-2**) مع تعديلات تُراعي خصوصية محركات البحث كالاستعلام.

3-2-3- تكامل محركات البحث مع نماذج التهجين بين الترشيح المرتكز على المحتوى، والترشيح التعاوني.

حاول الباحثون من خلال هذا الأسلوب تَجاوز بعض مَحدوديات كل من النموذجين، واستغلال إيجابيات كلٍّ منهما.

هنالك عدة طرائق للدمج بين الترشيح التعاوني والترشيح المرتكز على المحتوى داخل محركات البحث أهمها:

- التوزين (weight)

يُعطي النظام هنا وزناً محدداً لكل من نموجي الترشيح

بحسب حالة فضاء المعلومات وطبيعة المستخدمين[25] [24].

- التبديل(switch)

يقوم النظام في هذه الحالة بالتبديل بين تقنيات الترشيح بحسب لاحة المستخدم وجلسة البحث[25].

- التتابع أو التتالي(cascade)

تُطبَّقُ تقنيات الترشيح هنا بشكل متتالٍ، فعلى سبيل المثال: يُطبق الترشيح المرتكز على المحتوى ثم يتم يُطبَّقُ الترشيح التعاوني على النتائج المسترجعة من ترشيح المحتوى[23] [25].

- المزج(mixed)

يُبدو نموذجيا الترشيح هنا كنموذج واحد لا يتجزأ حيث يتداخل نمطا الترشيح مع بعضهما لتشكيل لاحة المستخدم، ومن ثَم التشريح اعتماداً عليها[4].

نعرض في الفقرة اللاحقة بعض أهم الأعمال في كل صنف من الأصناف الثلاثة السابقة، الخاصة بتكامل محركات البحث مع تقنيات الترشيح.

3-4- الأعمال السابقة المتعلقة:

عُنيت كثير من البحوث السابقة بمسألة تَوظيف تقنيات الترشيح في استرجاع المعلومات، واتبع الباحثون في هذا الإطار سبلاً وطرائقَ مختلفة؛ كما استخدموا تقنيات متنوعة مستمدة من أنظمة استرجاع المعلومات وتقنيات التنقيب وغيرها.

كانت بدايات أعمال تَداخل أنظمة الترشيح مع محركات البحث في عام 1993 مع Tak W [7] الذي استخدم النموذج الشعاعي في تمثيل لاحة المستخدم وعلاقة التَجيب في حساب التشابه بين لاحة المستخدم والوثيقة المسترجعة وظل النموذج الشعاعي بدايةً حاضراً بقوة في كثير من بدايات تداخل أنظمة الترشيح مع أنظمة استرجاع المعلومات، وخصوصاً محركات البحث [7] [8]؛ ولعل ذلك يعود إلى شيوع استخدام النموذج الشعاعي في أنظمة استرجاع المعلومات والنتائج التي حققها هذا النموذج رغم وجود بعض العيوب فيه، فضلاً عن أنه قد يكون بسبب الرغبة في توحيد نموذج استرجاع المعلومات مع نموذج الترشيح.

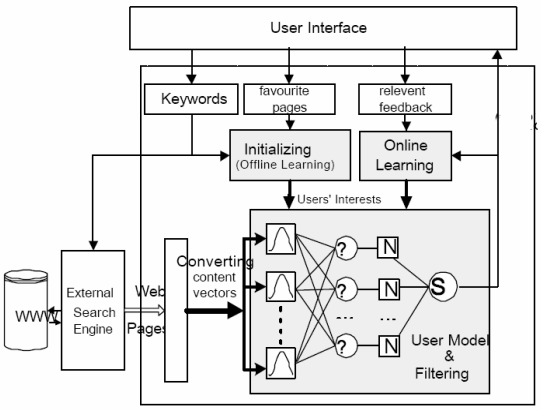
نعرض في الفقرة اللاحقة أهم البحوث في كل صنف من أصناف التكامل بين محركات البحث وتقنيات الترشيح

- اعتمد Rohini U [20] على المشاركة بين المستخدمين ضمن دائرة الاهتمام باستخدام الترشيح التعاوني لإعادة ترتيب النتائج المسترجعة، ويتم تعليم وتدريب لاحة المستخدم عن طريق أمثلة تدريبية ناتجة من التغذية الراجعة لتفاعل المستخدم مع محرك البحث.

- ارتكز Mirco Speretta [5] في بناء لاحة المستخدم على تصنيف المعلومات المجمعة من الاستعلامات وملخصات(snippets) الوثائق المسترجعة داخل هرمية مفاهيم معرفة مسبقاً مثل (Open Directory - ODP Project) ، ومن ثَم يُعطى لكل مفهوم في هرمية لاحة

المستخدم قيمة تعبّر عن مقدار اهتمام المستخدم بهذا المفهوم بحسب الوثائق الموائمة التي تصنف تحت هذا المفهوم. وتعتمد عملية ترشيح الوثائق على إعادة ترتيب الوثائق المسترجعة باستخدام علاقة تعتمد على قيمة الترتيب الأصلية للوثيقة الناتجة عن محرك بحث خارجي وقيمة الترشيح الناتجة عن التشابه بين لاحة المستخدم والوثيقة ويندرج هذا النمط من الترشيح تحت الترشيح المرتكز على المحتوى.

- استخدم Dai Xuewu [6] الشبكات العصبونية في بناء نظام الترشيح، والنظام عبارة عن شبكة عصبونية من ثلاث طبقات كما في الشكل (3).



الشكل (3) استخدام الشبكات العصبونية

طبقة الداخل: وهي المسؤولة عن استقبال أشعة الوثائق الداخلة إلى نظام الترشيح، كل عصبون في طبقة الداخل له شعاع وأوزان يمثلان لاحة المستخدم. تتم عملية تدريب وتعليم الشبكة بتغير قيم مركبات شعاع الأوزان.

والطبقة الخارجية: للشبكة تمثل عملية الترشيح النهائية حيث يُطبَّقُ تابع ترشيح خاص بها ويَنتج عنه قيمة الترشيح النهائية للوثيقة، تعاني هذه المقاربة من مشكلة شدة الخصوصية، وتحتاج إلى توافر أمثلة تدربية مسبقة لتدريب الشبكة بشكل مستمر.

- قام Ahu Sieg [10] بإضفاء الطابع الشخص على محرك البحث معتمدا" على نموذج الأنطلوجي لبناء لاحة المستخدم ووسم كل عقدة في الأنطلوجي بقيمتين عدديتين إحداهما تسمى درجة التفعيل وتعبر عن السياق الحالي، والثانية تسمى درجة الاهتمام وتعبّر عن السياق التاريخي لدرجة اهتمام المستخدم بهذه العقدة، وتتم عملية التعليم للاحة المستخدم بتعديل أوزان درجات الاهتمام من خلال تراكم تغيرات درجات التفعيل عبر جلسات البحث التي يقوم بها المستخدم، أمَّا عملية الترشيح فتعتمد على خوارزمية تقوم بالمرور على الوثائق المسترجعة كلّها وتحدّد المفهوم الذي ينتمي له، ثم تطبيق علاقة للترشيح تقوم على ثلاث معاملات درجة الاهتمام المتعلقة بهذا المفهوم والاستعلام والوثيقة.

[24] جمع بين الترشيح التعاوني والترشيح المرتكز على المحتوى بتحديد المُسبق لنسب الاعتماد على كل منهما بحسب القيمتين a و b

$$Phybrid = a \times Pcontent + b \times Pcollaborative$$

حيث a و b يحققان الشرط $a + b = 1$.

سلك [23] أسلوب فريد في التهجين بين تقنيات الترشيح ارتكز على توسيع وتوزين الاستعلام وفق خوازمتيين متتاليتين

1. الأولى تقوم بتوسيع الاستعلام باستخدام لاحة المستخدم(الترشيح المرتكز على المحتوى).
2. والثانية تقوم بتوزين الكلمات المفتاحية للاستعلام باستخدام لاحات المستخدمين المتشابهيين(الترشيح التعاوني).

- اقترح Hsin-Chieh Huang **[4]** مزجاً بين الترشيح التعاوني والترشيح المرتكز على المحتوى في مرحلة فهرسة الوثائق؛ للحصول على لالحات المستخدمين وفق مرحلتين:

1. المرحلة الأولى: انطلق من مصفوفة تقييمات عناصر المعلومات الثنائية المُستخدمة في الترشيح التعاوني، ثم قام بعملية تمديد لهذه المصفوفة باستخدام علاقات الترشيح التعاوني، للتخلص من مشكلة التبعثر.
2. المرحلة الثانية: انطلق من مصفوفة أوزن ملاح وسمات عناصر المعلومات، ثم ضربها بالمصفوفة الناتجة من المرحلة الأولى للحصول على مصفوفة تُمثل العلاقة بين سِمات عناصر المعلومات والمستخدمين.

وبهدف تخفيض الأبعاد والحسابات، تُطبَّقُ تقنية latent semantic indexing على هذه المصفوفة الناتجة للحصول على المصفوفة النهائية، التي تُمثل لاحة المستخدم.

من خلال ما سبق نلاحظ أن معظم بحوث تكامل نظم الترشيح مع محركات البحث ارتكزت إمَّا على تطوير أحد نمطي الترشيح (المرتكز على المحتوى أو التعاوني)، أو التهجين بينهم.

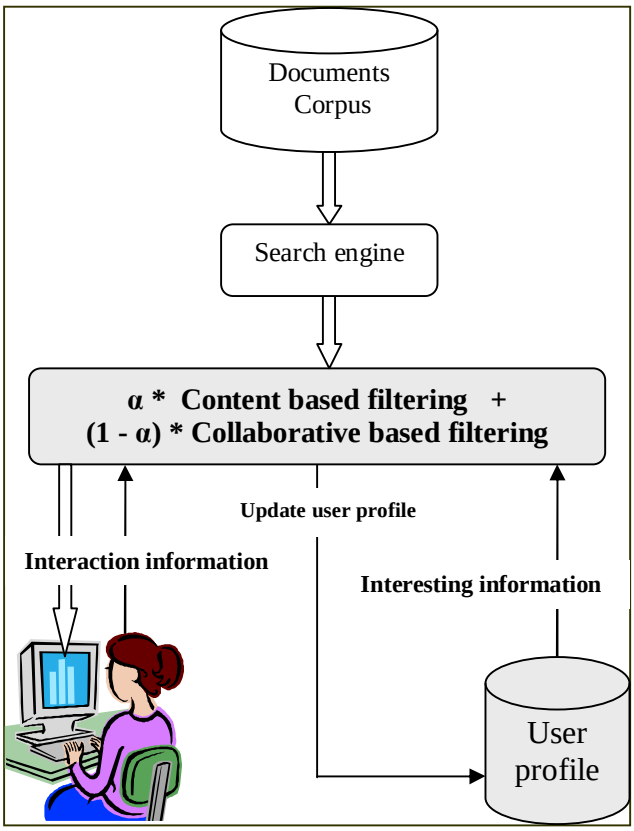
وقد أثبتت بحوث التهجين دوراً فعالاً، بالاستفادة من ميزات كلّ من نمطي الترشيح، ومحاولة التخلص من محدويات كل منهما، ولكن ظلت هنالك حاجة لإيجاد نموذج للتهجين يستطيع المُوازنة بين أهمية كل من نمطي الترشيح بشكل ألي بحسب طبيعة المستخدم وسياق البحث؛ وفي الوقت نفسه يستفيد من أفضل الأفكار والتقنيات بخصوص تمثيل وتعليم لاحة المستخدم، وهذا ما حاولنا أن نعمل عليه في المقاربة المقترحة.

4- المقاربة المقترحة للتكامل

4-1- النموذج العام للمقاربة المقترحة

تقوم الفكرة العامة لتكامل تقنيات الترشيح مع محركات البحث، على التهجين والجمع بين الترشيح المرتكز على المحتوى والترشيح المرتكز على تعاون المستخدمين؛ بهدف الاستفادة من النقاط الإيجابية في كلا النموذجين، وتجاوز بعض النقاط السلبية الناتجة من الاعتماد على نموذج وحيد على أن يتم التحديد الآلي لنسبة الاعتماد على كل من النموذجين في عملية الترشيح بحسب طبيعة المستخدم وسياق البحث.

والشكل التالي: يوضح البنية العامة للمقاربة المقترحة والعلاقة بين المكونات المختلفة بشكل مجرد، وبغض النظر عن النموذج التفصيلي لكل مكون من مكونات النظام التي سوف تشرح لاحقاً.



الشكل (4) البنية العامة للمقاربة المقترحة

α: مقدار بين الصفر والواحد، يحدد درجة الاعتماد على الترشيح المرتكز على المحتوى.

(α - 1): مقدار بين الصفر والواحد، يحدد درجة الاعتماد على الترشيح المرتكز على تعاون المستخدمين.

نلاحظ من الشكل (4) أن النظام يقوم بالحصول والتقاط نتائج محرك بحث الناتجة عن استفسار محدد، ثم إعادة تَرشيح هذه النتائج وتوليفها؛ بالاعتماد على المعلومات المُعرفة والمُضمنة في نماذج لاحات المستخدم. كما يقوم باستثمار المعلومات المُستخرجة من التغذية الراجعة، ومن تفاعل ورَدة فعل المستخدم على النتائج المسترجعة في عملية أساسية هي: تعليم وتكيّف لاحة المستخدم بشكل آني؛ بحيث تبقى وتظل لاحة المستخدم تعكس التمثيل الحقيقي لاهتمامات المستخدم وتفضيلاته، وتواكب التغيرات التي تطرأ على مختلف تصنيفات المجال المدروس.

4-2- مكونات نظام الترشيح للمقاربة المقترحة

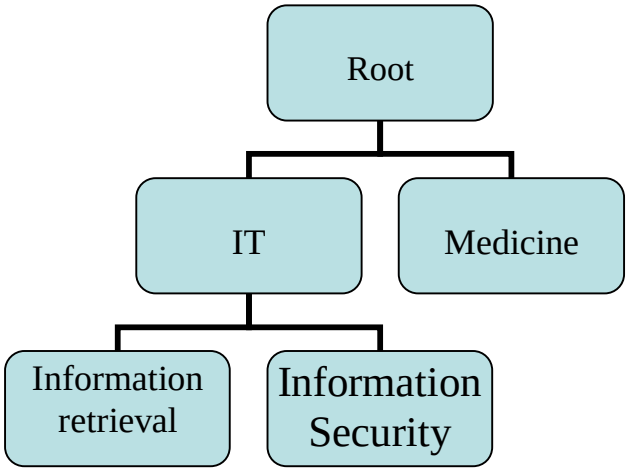
إن كل مكون من مكونات المقاربة المقترحة المبين في الشكل السابق يمكن أن يُنمذج ويبنى بعدة طرائق، وأن يعتمد على عدة تقنيات (تقنيات نظم استرجاع المعلومات –تقنيات التنقيب –تقنيات التعلم التلقائي –تقنيات الترشيح).

في القسم التالي سيوضَّحُ النموذج والتقنيات المقترحة لكل مكون من مكونات المقاربة المقترحة

(لاحة المستخدم، تقنيات تعليم وتدريب للاحة المستخدم تقنيات الترشيح المُستخدمة).

4-2-1 لاحة المستخدم

النموذج المقترح لتمثيل لاحة المستخدم في هذا النظام هونموذج أنطلوجي المجال(Domain ontology)، هذا النموذج الذي أثبت كفاءته في كثير من البحوث المتعلقة بهذا السياق [10]، [14]، [16]، [17] إذْ لاحة كل مستخدم في النظام، تُمثل ببنية هرمية من المفاهيم التي تُوصف المجال والفضاء، الذي يرتكز عليه النظام.



الشكل (5) تمثيل لاحة المستخدم باستخدام أنطلوجي المجالات

كل عقدة (مفهوم) داخل أنطلوجي لاحة المستخدم، تُمثل صفاً يحتوي على معلومات ومعرفة، تُعبر عن مقدار اهتمام المستخدم وتفضيله لهذا المفهوم في الأنطلوجي.

يتميّز نموذج الأنطلوجي لتمثيل لاحة المستخدم، بقدرته على تمثيل اهتمامات المستخدمين على مفاهيم مختلفة؛ مما يُسهل ضبط عملية الترشيح بشكل أكبر، ويساعد في الدلالي للكلمات المفتاحية بحسب المجال أو المفهوم الذي تنتمي إليه.

فتحديد سياق البحث من خلال تفاعل المستخدم مع النظام في جلسة البحث يمكننا من تحديد المفهوم الأكثر ارتباطاً وصلة من لاحة المستخدم، ومن ثمَّ الترشيح اعتماداً على هذا المفهوم.

كل عقدة في البنية الهرمية للاحة المستخدم تحتوي على مكونيين أساسيين:

1. جدول يحوي معلومات عن جلسات البحث التي قام بها المستخدم (الاستعلامات، الوثائق المرتبطة بكل استعلام درجة اهتمام وموائمة الوثائق للاستعلام من وجهة

نظر المستخدم، التاريخ،.....).

تُستخدم هذه المعلومات بشكل أساسي في الترشيح التعاوني، ويُبنى هذا الجدول من ملفات تسجيل التتبع (log fill)، بعد القيام بعمليات التنظيف والتجميع لهذه المعلومات.

2. شعاع من الكلمات المفتاحية المَوزُونة التي تعبّر عن اهتمامات وتفضيلات المستخدم للمفهوم التي تمثله هذه العقدة، ونرمز لهذا الشعاع بالرمز C، على سبيل المثال: لنفرض أن فضاء الوثائق مُكون من ثلاث كلمات مفتاحية هي [viruse، neural ، filtering] ومن ثمَّ يصبح كل مفهوم في أنطلوجي المجال يحتوي شعاعاً من ثلاث كلمات مفتاحية مَوزُونة، يعبّر الوزن عن اهتمام المستخدم بها ضمن إطار المفهوم الذي يحتوي الشعاع فقط، وهكذا يمكن للكلمة المفتاحية viruse أن تأخذ وزناً كبيراً ضمن المفهوم information Security، ووزناً صغيراً ضمن المفهوم Virus Diseases، بالنسبة إلى مستخدم مهتم بأنظمة أمن المعلومات، بعكس مستخدم يعمل كطبيب فتكون وزن الكلمة المفتاحية viruse أكبر في المفهوم الطبي، مقارنة بالمفهوم الذي يتعلق بأمن المعلومات. يتم توزين أوزان مركبات هذا الشعاع وتعديلها اعتماداً على عملية التعليم للاحة المستخدم [الفقرة 4-2-2].

4-2-2 تعليم لاحة المستخدم وتكيّفها

تتميز الطريقة المُتبعة في هذا المقاربة لتعليم وتكيّف لاحة المستخدم، بالتعليم الأوتوماتيكي الضمني الدائم والمستمر بحيث يواكب التغيرات التي تطرأ على لاحة المستخدم (عدم الاعتماد على أمثلة تدريبية فقط)، من خلال جلسات البحث التي يقوم بها.

تقوم فكرة التعليم، على الاستفادة من التغذية الراجعة الموائمة من قبل المستخدم على الوثائق، وتحديد أهمية الوثيقة المسترجعة بناء على مقدار الوقت الذي يمضيه المستخدم على تلك الوثيقة (spent time)، متناسباً عكساً مع طول الوثيقة [9] وفق العلاقة الآتية:

$$I(D) = \frac{ST}{len(D)}$$

$I(D)$: مقدار يعبّر عن مدى اهتمام الوثيقة المُسترجعة وتفضيلها للمستخدم.

ST : كمية الوقت التي أمضاها المستخدم على الوثيقة المسترجعة.

$len(D)$: طول الوثيقة المسترجعة.

يُقاس التشابه بين تلك الوثيقة ومفاهيم لاحة المستخدم لتحديد المفهوم الأكثر تشابهاً مع الوثيقة، ثم القيام بعملية التعليم والتكيّف للمفهوم المحدد، وفق العلاقة الآتية :

$$Cnew = \frac{\text{Cold} + \text{I(D)}}{\text{Max(C)}}$$

C new : شعاع يمثل المفهوم الأكثر تشابهاً مع الوثيقة المسترجعة، من مفاهيم أنطلوجي لاحة المستخدم، بعد إجراء عملية التعليم.

C old : شعاع يمثل المفهوم الأكثر تشابهاً مع الوثيقة المسترجعة، من مفاهيم أنطلوجي لاحة المستخدم، قبل إجراء عملية التعليم.

Max(C): يُمثل وزن أعظم كلمة مفاحية في شعاع المفهوم

(c) من مفاهيم أنطلوجي لاحة المستخدم.

4-2-3 الترشيح

الأسلوب المتبع في ترشيح الوثائق المسترجعة الناتجة عن استفسار محدد مبين في العلاقة الآتية:

F(U,D,Q) = α * Content based filtering + (1 - α) * collaborative based filtering (1)

U: المستخدم الذي يقوم بجلسة البحث.

D: الوثيقة المسترجعة.

Q: الاستعلام المُدخل.

α: مقدار بين الصفر والواحد يُحدَّدُ درجة الاعتماد على الترشيح المرتكز على المحتوى ويعطى بالعلاقة:

$$\alpha = \frac{1}{1+\sum_{n\in N} sim\ (u,n)} \qquad (2)$$

إذْ:

N: مجموعة المستخدمين المشابهين للمستخدم الحالي(u).

$sim(u,n)$: التشابه بين المستخدم الحالي u والمستخدم n.

(α - 1) : المقدار المتمم لـα، ويحدد درجة الاعتماد على الترشيح التعاوني.

وقد اعتُمِدَ على العلاقات (1) و (2) بناءً على التجارب التي أجريت (الفقرة **5-2**)، إذْ تبيّن أن تأثير الترشيح التعاوني يكون أفضل في تحسين مقاييس الدقة والاستذكار إذا تحقق شرطان:

1. ازدياد عدد المستخدمين المشابهين للمستخدم الحالي.
2. ازدياد درجة التشابه بين المستخدم الحالي والمستخدمين المشابهين.

ومن ثمَّ كان لابد من زيادة الاعتماد على الترشيح التعاوني على حساب الترشيح المرتكز على المحتوى، كلما زاد عدد المستخدمين المشابهين، وزاد تشابههم مع المستخدم الحالي.

في بداية عملية الترشيح يُحدَّدُ أولاً: المفهوم C، الذي تنتمي له الوثيقة D، ضمن المفاهيم التي تُوّصف فضاء الوثائق. وعادةً تتم عملية تصنيف الوثائق في مفاهيم وصفوف بشكل مسبق في أنظمة استرجاع المعلومات؛ لذا تكون لدينا معرفة مسبقة عن المفهوم الذي تنتمي إليه الوثيقة المُسترجعة.

الترشيح التعاوني يحسب بالعلاقة:

$$coll(u,d,q)=\overline{Ru}+\frac{\sum_{n\in NBS_u} sim(u,n)*(R(n,d,q)-\overline{Rn})}{\sum_{n\in NBS_u}(sim(u,n))} \qquad (3)$$

d: الوثيقة المسترجعة.

q: الاستعلام المُدخل من قبل المستخدم.

NBS_u: مجموعة المستخدمين المشابهين للمستخدم الحالي.

$sim(u,n)$: التشابه بين المستخدمين u وn ضمن المفهوم C، الذي تنتمي له الوثيقة D.

$R(n,d,q)$: درجة موائمة الوثيقة d للاستعلام q من وجهة نظر المستخدم n.

$\overline{R\ n}$: متوسط تقييمات المستخدم n.

$\overline{R\ u}$: متوسط تقييمات المستخدم الحالي u.

الترشيح المرتكز على المحتوى يحسب بالعلاقة:

حساب التشابه بين الوثيقة المسترجعة ولاحة المستخدم

$$Content\_based\_filtering\ (u,d)=sim\ (u,d) \qquad (4)$$

u: شعاع لاحة المستخدم ضمن المفهوم C، الذي تنتمي له الوثيقة.

d: شعاع الوثيقة المسترجعة من محرك البحث.

يُحسَبُ التشابه بين شعاع الوثيقة، وشعاع لاحة المستخدم باستخدام علاقة التجيب في النموذج الشعاعي.

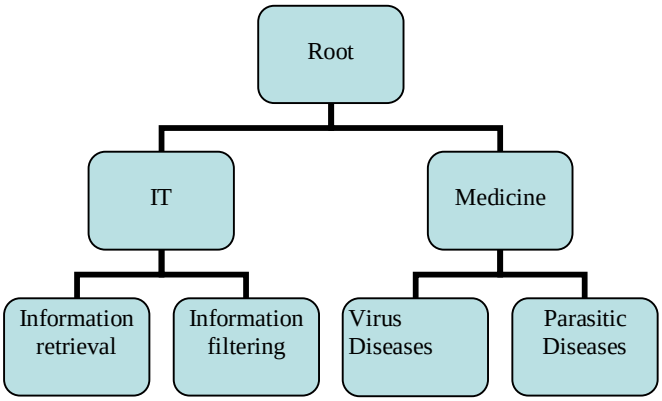
وهكذا وبحسب العلاقة (1)، يتم الحد من مشكلتي شدة الخصوصية، وحدودية المحتوى في الترشيح المرتكز على المحتوى؛ بزيادة نسبة الاعتماد على الترشيح التعاوني وتخفيف نسبة الاعتماد على الترشيح المرتكز على المحتوى كلما زاد عدد المستخدمين المشابهين و زاد تشابههم مع المستخدم الحالي؛ وهذا يُتيح للمستخدم تعرّف آراء مستخدين آخرين وتوصياتهم، والتحرر من الارتباط الكبير بلاحته.

بالمقابل يتم التركيز على الترشيح المرتكز المحتوى بشكل أكبر في حالة المستخدم الجديد، الذي يملك عدداً قليلاً من المستخدمين المشابهين، ريثما يزداد عددهم بمرور الزمن.

5- التجارب والتقييمات

أُجريتِ التجارب على فضاء وثائق مكون من نحو 11000 وثيقة مصنفة بشكل مسبق في بنية هرمية.

المستوى الأول من الهرمية مقسم إلى صنفين: مجال طبي ومجال مرتبط بعلوم المعلومات، المجال الطبي مقسم إلى 11 صنفاً فرعياً مرتبطاً بأمراض طبية، والمجال المتعلق بعلوم المعلومات مقسم إلى صفين: الأول يتعلق باسترجاع المعلومات والثاني بترشيح المعلومات، الشكل (6) يوضح جزءاً من الأصناف الأساسية التي تنتمي إليها وثائق الفضاء.



الشكل (6) أصناف وثائق الفضاء

أشهر مقاييس محركات البحث هما الدقة والاستذكار اللذان اعتُمِدَ في التقييم في هذا البحث.

مقياس الدقة(Precision): نسبة عدد النتائج المسترجعة الموائمة إلى النتائج المسترجعة [3].

$$precision = \frac{number\ of\ relevent\ documents\ retrieved}{total\ number\ of\ documents\ retrieved}$$

وهو يعبّر عن قدرة النظام على جعل النتائج المسترجعة موائمة.

مقياس الاستذكار: نسبة عدد النتائج المسترجعة الموائمة إلى النتائج الموائمة كلّها في الفضاء[3].

$$precision = \frac{number\ of\ relevent\ documents\ retrieved}{number\ of\ relevant\ documents\ in\ collection}$$

وهو يعبّر عن قدرة النظام على استرجاع النتائج الموائمة من كامل الفضاء.

أُجريت التجارب على نحو 6 استعلامات وفي كل استعلام أُجريت ثلاث جلسات بحث من قبل 3 مستخدمين بشكل متتالٍ لمراقبة عملية تعليم لاحة المستخدم، وتحسن النتائج تدريجياً؛ الاستعلامات المتعلقة بالمجال الطبي أُجريت من قبل أطباء.

في كل تجربة تجري المقارنة بين أربع حالات: النموذج التقليدي، ونموذج الترشيح المرتكز على المحتوى، نموذج الترشيح التعاوني، وأخيراً النموذج المقترح.

تقوم مخططات التجارب اللاحقة، بعرض مخططات الدقة والاستذكار لأول عشر نتائج مسترجعة من محرك البحث إذ لأجل كل استعلام تُقاسُ الدقة والاستذكار عند كل وثيقة مسترجعة من الوثائق العشر الأولى، ثم بعد ذلك يُؤخَذُ متوسط الدقة والاستذكار لأجل الاستعلامات كلّها عند ترتيب كل نتيجة من واحد إلى عشر.

الملحق رقم(1) يوضح المعلومات التفصيلة للتجارب والتقييمات، التي أجريت في المرحلة الأخيرة من كل استعلام.

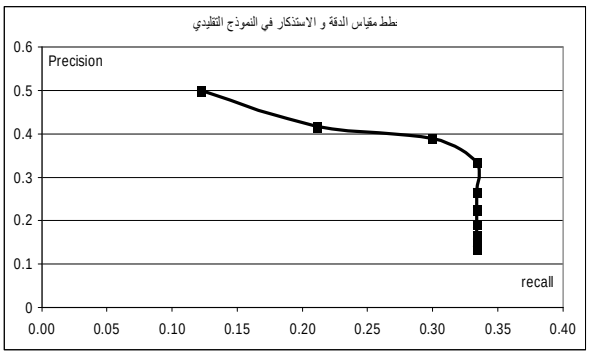
المحور العامودي في مخططات التجارب كلّها يدلّ على مقياس الدقة، أمَّا المحور الأفقي فيدل على مقياس الاستذكار.

5-1- تأثير النموذج المقترح في مقاييس الدقة والاستذكار

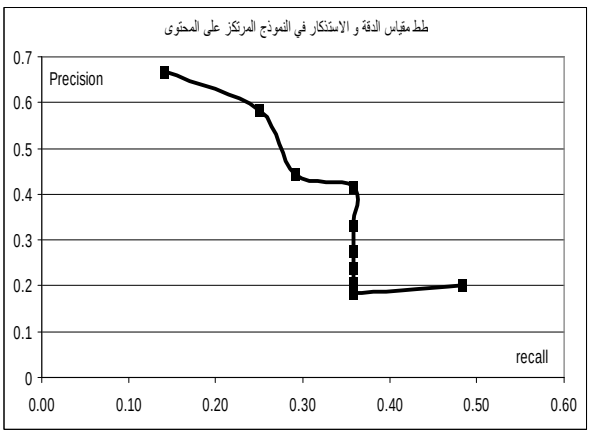
أُجريتْ أربع تجارب أساسية، للتأكد من فعالية النموذج المقترح: التجربة الأولى أجريت على محرك بحث تقليدي دون أي عملية ترشيح بحسب المستخدم، وكان الهدف منها فقط المقارنة بالتجارب الأخرى اللاحقة المتكاملة مع

نماذج الترشيح، التجربة الثانية كانت لتكامل محرك البحث مع الترشيح المرتكز على المحتوى فقط، التجربة الثالثة كانت لتكامل محرك البحث مع الترشيح المرتكز على تعاون المستخدمين فقط، أمَّا التجربة الرابعة فكانت عملياً لاختبار النموذج المقترح القائم على التهجين بين النموذج المرتكز على المحتوى والنموذج التعاوني.

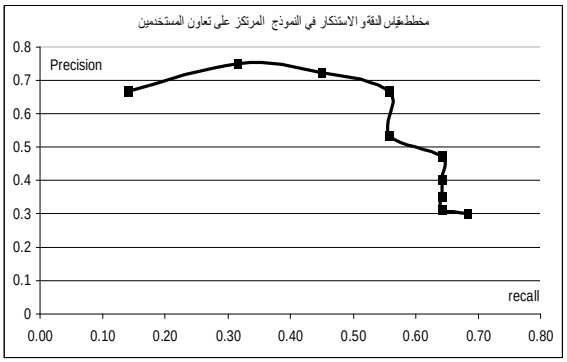
5-1-1 مقياس الدقة والاستذكار في النموذج التقليدي



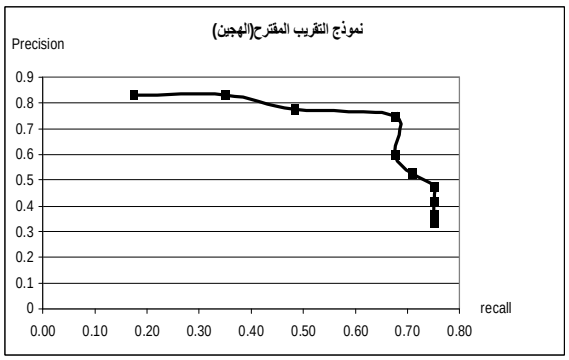
5-1-2- مقاييس الدقة والاستذكار للتكامل مع الترشيح المرتكز على المحتوى



5-1-3 مقاييس الدقة والاستذكار للتكامل مع الترشيح التعاوني



5-1-4- مقاييس الدقة والاستذكار للمقاربة المقترحة



5-1-5- مقارنة مقاييس الدقة والاستذكار

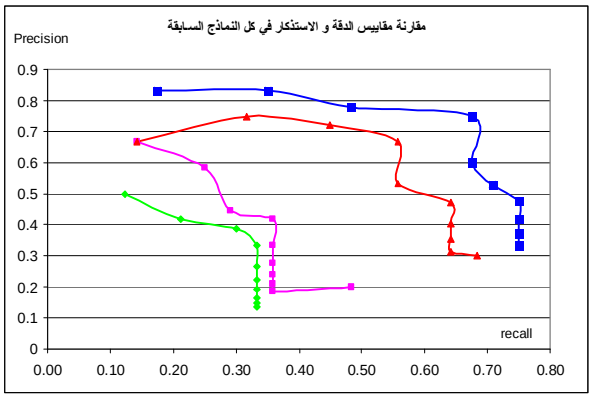
تبيّن المخططات هنا بشكل أوضح، المقارنة بين النماذج المختلفة حيث يتبين أن الترشيح التعاوني (اللون الأحمر) له تأثير واضح وجلي مقارنة بالترشيح المرتكز على المحتوى، كما أن النموذج الهجين يتفوق على كلا النموذجين السابقين في تحسين مقاييس الدقة والاستذكار.

النموذج التقليدي

النموذج المرتكز على المحتوى

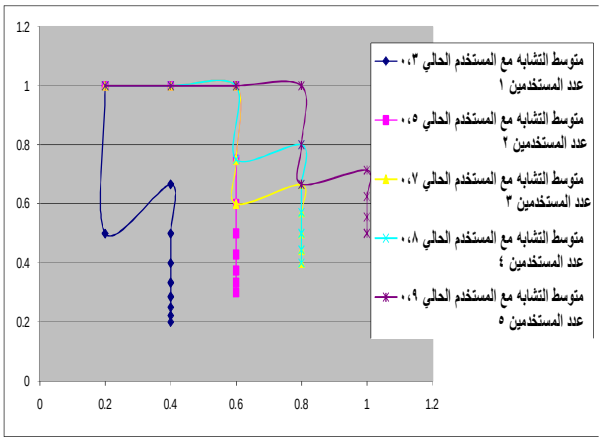
النموذج المرتكز على تعاون المستخدمين

نموذج المقاربة المقترحة(الهجين)



5-2- تأثير المستخدمين المشابهين، في الترشيح التعاوني

نلاحظ من خلال المخططات في الفقرة(5-1) التفوق الواضح للترشيح التعاوني على الترشيح المرتكز على المحتوى، ولكن هذا الأمر مرتبط بعدد المستخدمين المشابهين، ومقدار تشابههم مع المستخدم الحالي، وقد أظهرت التجارب أن زيادة عدد المستخدمين المشابهين في علاقة الترشيح التعاوني، وزيادة تشابههم مع المستخدم الحالي؛ يحسن أداء محركات البحث كما تُوضح المخططات التالية، أجريت التجارب هنا فقط، في حالة التكامل مع الترشيح التعاوني.



ويمكن أن نفسر هذه الحالة بأنَّ: زيادة عدد المستخدمين المشابهين، وزيادة متوسط تشابههم مع المستخدم الحالي يتيح الفرصة للاستفادة من مستخدمين آخرين مشابهين له فتكون النتائج أفضل وأدق، وهذا أمر منطقي ما دام الترشيح التعاوني يعتمد اعتماداً أساسياً على مشورة آخرين وخبرتهم مروا بحالة البحث نفسها، وقد ساعدت هذه النتيجة بشكل أساسي في صياغة العلاقة (1).

Q) = α * Content based filtering +،D،F(U

(1 - α) * collaborative based filtering (1)

$$\alpha = \frac{1}{1+\sum_{n \in N} sim\ (u,n)} \qquad (2)$$

تؤمن العلاقة (1) زيادة الاعتماد على الترشيح التعاوني بزيادة عدد المستخدمين المشابهين، وزيادة تشابههم مع المستخدم الحالي، على حساب الترشيح المرتكز على المحتوى.

5-3- الخلاصة والآفاق المستقبلية :

من خلال التجارب التي أجريت على المقاربة المقترحة تبين أن: يؤدي الترشيح المرتكز على المحتوى، دوراً مهماً في تحديد مجال البحث وسياقه في معظم الحالات، فعلى سبيل المثال: في حالة الاستفسار neural "network" تختلف النتائج المسترجعة بحسب طبيعة المستخدم واهتماماته، فيما إذا كان طبيباً أو له علاقة بالذكاء الاصطناعي. أمَّا الترشيح المرتكز على تعاون المستخدمين فيؤدي دوراً محورياً في تقديم النتائج الأكثر دقة للمستخدمين، وتزداد هذه الدقة بازدياد عدد المستخدمين المشابهين وعلاقة الترشيح التعاوني الأساسية وازدياد تشابههم مع المستخدم الحالي؛ ولذلك أُخذَتْ هذه الملاحظة بالحسبان في علاقة الترشيح النهائية للمقاربة المقترحة.

أمَّا النقطة الأخرى في المقاربة المقترحة فهي: اعتماد نموذج أنطلوجي المجال في نمذجة لاحة المستخدم، حيث تتم المقارنات والتعليم فقط على بعض المفاهيم والأصناف مما خفف الكثير من الأخطاء والتشويش، الناتجة عن

اعتبار لاحة المستخدم كتلة واحدة كما في بعض البحوث السابقة.

هنالك مجالات مستقبلية عديدة للمقاربة المقترحة تتمثل في عدة محاور منها ما يتعلق بالعمل على تسريع زمن استجابة النظام وتحسينه نظراً إلى التكلفة الإضافية المترتبة على حسابات الترشيح، ومنها ما يتعلق بتطوير مكونات نظام الترشيح الأساسية، فعلى صعيد الأنطلوجي يكون بإغنائها بعلاقات دلالية وطرائق استدلال تتيح الاستفادة أكثر من الميزات التي تقدمها هذه البنية، كما يمكن العمل على تطوير طرائق وأساليب جديدة لتعليم لاحة المستخدم، لما لهذه العملية من أهمية في جودة عملية الترشيح.

المراجع:

1- Shang, F . Wang, l and Shi, l .(2010). An Approach of Web Text Information Filtering Based on Domain Ontology to Expand Users' Requirements literature: (IEEE) Proceedings 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE).

2- Liu، Fang، Clement Yu، and Weiyi Meng. "Personalized Web search for improving retrieval effectiveness." IEEE Transactions on Knowledge and Data Engineering 16.1 (2004): 28+. Academic OneFile. Web. 26 Oct. 2010.

3- Nicholas J. Belkin and W. Bruce Croft."Information filtering and information retrieval: Two sides of the same coin? ".G A L E G R O U P .Dec 1992 v35 n12.

4- Hsin-Chieh Huang A Content via Collaboration Approach to Text Filtering Recommender Systems 2006

5- Mirco Speretta. Personalizing Search Based on User Search Histories .B.Sc., Udine University, Udine, Italy 2000.

6- Dai Xuewu1, Vic Grout2، Tang Haokun3 and Li Jianguo1 NEURAL NETWORKS -BASED MULTI-INTEREST INFORMATION FILTERING. University of Wales، NEWI, Wrexham، UK.

7- Tak W. Yan and Hector Garcia-Molina Index Structures for Information Filtering Under the Vector Space Model Department of Computer Science، Stanford University, Stanford CA94305 November 8، 1993.

8- Wang Shuda; Liu Jiangping; Wang Riu Research of Information Filtering Based on Vector Space Model Harbin Univ. of Commerce, Harbin، China 2009

9- Masahiro MORITA Yoichi SHINODA Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval.

10- Ahu Sieg، Bamshad Mobasher, Robin Burke Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search IEEE Intelligent Informatics Bulletin November 2007 Vol.8 No.1.

11- Denis Lemongew Nkweteyim A collaborative filtering Approach to predict web pages of interest from navigation patterns of past users within An academic website.

12- Xujuan Zhou، Yuefeng Li، Peter Bruza، Yue Two-Stage Model for Information Filtering Faculty of Information Technology Queensland University of Technology Brisbane، QLD 4001، Australia. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

13- Kuflik and Peretz Shoval Generation of User Profiles for Information F i l t e r i n g 2000.

14- Ahu Sieg، Bamshad Mobasher، Robin Burke Ontological User Profiles for Representing Context in Web Search 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

15- Sayaka Akioka Hideo Fukumori Yoichi Muraoka Information Filter for Ambiguous Information Retrieval 2008.

16- C.F. So, Chapmann C.L. Lai، and Raymond Y.K. Lau Ontological User Profiling and Language Modeling for Personalized Information Services 2009 IEEE International Conference on e-Business Engineering.

17- Paul Alexandru. Chirita Wolfgang Nejdl .Using ODP Metadata to Personalize Search.

18- Rickard C¨oster Learning and Scalability in Personalized Information Retrieval and Filtering November 28, 2002.

19- Wang Shuda Yang Jing Research on the Information Filtering of OWL Text Based on Semantic Analysis.
20- Rohini U Vasudeva Varma A Novel Approach for Re-Ranking of Search results using Collaborative Filtering 2006.
21- Vamshi Ambati Rohini Uppuluri Improving Re-ranking of Search Results using Collaborative Filtering 2002.
22- Robin van Meteren Maarten van Someren Using Content-Based Filtering for Recommendation1.
23- Mittal, N.; Nayak, R.; Govil، M.C.; Jain, K.C.;, "A Hybrid Approach of Personalized Web Information Retrieval," Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on , vol.1, no., pp.308-313, Aug. 31 2010-Sept. 3 2010 doi: 10.1109/WI-IAT.2010.270 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5616572&isnumber=5614551.
24- Ling Yun Wang Xun Gu Huamao A Hybrid Information Filtering Algorithm Based on Distributed Web log Mining Third 2008 International Conference on Convergence and Hybrid Information Technology, College of Computer & Information Engineering، Zhejiang Gongshang University.
25- Siles H, Castillo D, Hybrid Content-Based Collaborative-Filtering Music Recommendations, University of Twente, The Netherlands 2007.

.

تاريخ ورود البحث إلى مجلة جامعة دمشق 2011/10/5.