بسم الله الرحمن الرحيم

نماذج استرجاع المعلومات

Information Retrieval Models

د. فاتن سعيد بامفلح

تعد المضاهاة من العمليات الرئيسة في نظم استرجاع المعلومات، بل إنها تشكل أحد النظم الفرعية المكونة لتلك النظم. فمن خلال عملية المضاهاة تتم المطابقة بين التمثيلات المعبرة عن استفسارات المستفيدين، وبين التمثيلات المعبرة عن موضوعات الوثائق، بغرض استرجاع التسجيلات التي تلبي احتياجات المستفيدين من المعلومات.

والواقع أن نظم استرجاع المعلومات تسير في تطبيقها لعملية المضاهاة وفقاً لنماذج models معينة، يتم بناء عليها تحديد الأسلوب المتبع في النظام للتعرف على الوثائق ذات الصلة باستفسار المستفيد. وهناك نماذج متعددة متاحة للتطبيق من خلال نظم الاسترجاع. وتختلف الفكرة التي يقوم عليها كل نموذج من تلك النماذج، وبالتالي فإن نتائج الاسترجاع ودقتها، ودرجة ملاءمتها للرد على استفسار المستفيد تعتمد على طبيعة النموذج الذي يطبقه نظام استرجاع المعلومات.

ويشير Beaza–Yates و Ribeiro–Neto إلى أن نماذج الاسترجاع تعمل وفقاً لأربعة عناصر هي:

- الوثائق Documents : وتمثل مجموعة تتكون من تمثيلات الوثائق في المجموعة.
- الاستفسارات Queries : وتمثل مجموعة تتكون من تمثيلات احتياجات المستفيدين من المعلومات.
- الإطار Framework : وهو إطار لنمذجة تمثيلات الوثائق والاستفسارات والعلاقات بينهما.
- الترتيب Ranking : وهي وظيفة تشترك فيها تمثيلات الوثائق والاستفسارات حيث يتم تحديد ترتيب الوثائق في النتيجة حسب ما جاء في الاستفسار[1].

وقد تطورت تلك النماذج التي تعد بمثابة تقنيات أو أساليب متبعة للمضاهاة؛ فبعد أن كانت تعتمد فقط على مطابقة المصطلحات الواردة في الاستفسار بمصطلحات التكشيف الدالة على الوثائق لمعرفة مدى توافر المصطلحات نفسها في كل من تمثيلات الاستفسارات والوثائق، فقد أصبحت حالياً تعتمد على طرق أخرى قائمة على تطبيق أساليب إحصائية ورياضية لتحديد الوثائق الملائمة للرد على استفسارات المستفيدين، بل وتحديد درجة ملاءمتها، أي مدى وثاقة الصلة بين الوثائق وبين الاستفسار. وفيما يلي نوضح أبرز النماذج التي تستخدمها نظم استرجاع المعلومات لإجراء عملية المضاهاة.

أولاً: النموذج البوليني Boolean Model :

يعد النموذج البوليني أحد النماذج الكلاسيكية واسعة الانتشار في نظم استرجاع المعلومات. وعلى الرغم من مآخذ البعض على هذا النموذج، إلاّ أنه لازال مطبقاً ومنتشراً في العديد من نظم الاسترجاع التجارية. ويعتمد النموذج البوليني على تقسيم الوثائق إلى فئتين: فئة وردت فيها مصطلحات الاستفسار، وفئة لم ترد فيها المصطلحات.

وهذا التقسيم يعني أنه ليس هناك أي نوع من التدرج في تقييم مدى صلة الوثائق بالاستفسار، فهي إما وثائق ذات صلة، أو غير ذات صلة، الأمر الذي يجعل النتائج غير دقيقة تماماً لأنه عادة تكون بعض الوثائق ذات صلة وثيقة بالاستفسار، في حين أن البعض الآخر يكون أقل صلة به.

ولعل اعتماد النموذج البوليني على هذا النوع من التقسيم يجعل نتائج الاسترجاع غير دقيقة؛ على اعتبار أنها لا تبين أهمية كل وثيقة، ومدى ارتباطها بالاستفسار.

وقد جرت محاولات لتطوير هذا النموذج بشكل يحقق ترتيب النتائج حسب صلتها بالاستفسار، وذلك عن طريق تحديد الوثائق التي وردت فيها كل مصطلحات البحث الواردة في الاستفسار، ووضعها في مرتبة أفضل من تلك التي لم يرد فيها أحد المصطلحات. فلو اشتمل الاستفسار على ثلاث مصطلحات؛ فإن الوثائق التي وردت فيها جميع المصطلحات الثلاث تأتي في بداية نتيجة البحث، تليها الوثائق التي اشتملت على مصطلحين، ثم تلك التي اشتملت على مصطلح واحد فقط.

ولكن يظل الحكم على الوثائق معتمداً على ورود المصطلح فيها أو عدم وروده، دون الوضع في الاعتبار مدى تكرار وروده في الوثيقة الواحدة على سبيل المثال، وهو الأمر الذي تراعيه بعض النماذج الأخرى[٢].

ويعتمد النموذج البوليني على استخدام عوامل منطقية لصياغة الاستفسار، وتتمثل تلك العوامل في الآتي:

- عامل الإقران (و and) : ويعني أنه يجب اقتران المصطلحات الواردة في الاستفسار، ووجودها مجتمعة في الوثائق.
- عامل الفصل (أو or) : ويقصد به عدم اشتراط اقتران المصطلحات مع بعضها في الوثائق، حيث يكفي ورود أحد المصطلحات في الوثيقة ليتم استرجاعها.
- عامل الرفض (ماعدا– ليس not) : ويشير إلى اشتراط استبعاد المصطلحات التي تلي not بحيث يتم الاسترجاع فقط للوثائق التي لا تضم تلك المصطلحات[٣].

ثانياً: نموذج حيز المتجهات Vector Model :

يختلف هذا النموذج عن البوليني في أنه يعتمد في عملية المضاهاة على تحديد معدل range معين يمثل حيز الوثائق ذات الصلة، وفي هذا الحيز يتم تحديد قيمة كل وثيقة على أساس درجة صلتها بالاستفسار.

وهناك أكثر من طريقة يتبعها هذا النموذج مطبقاً أساليب الجبر لاحتساب قيمة value كل وثيقة. فقد يعتمد على مقاييس التشابه similarity measures سواء القائمة على فكرة احتساب البعد أو المسافة distance على اعتبار أن الوثائق القريبة من بعضها في الحيز يمكن أن تكون أكثر تشابهاً، أو اعتماداً على مقياس الزاوية angular measure القائم على فكرة أن الوثائق الموجودة في اتجاه واحد direction متشابهة مع بعضها.

وعادة تحدد قيمة الوثيقة ما بين الصفر والواحد (٠– ١) ، ففي مقاييس البعد أو المسافة تكون قيمة الوثائق الأكثر صلة بالاستفسار هي ١ ، في حين أن الوثائق غير ذات الصلة تأخذ القيمة ٠ ، وما بينهما قيم تمثل الوثائق الأخرى حسب درجة ارتباطها بالموضوع مثال: ٠.١ و ٠.٢ وهكذا...أما في مقياس الزاوية، فعلى العكس من ذلك، حيث تكون مسافة الوثيقة من نفسها صفر، وبالتالي فإن القيمة الأقل تعني درجة ارتباط أكبر بين الوثيقة والاستفسار، في حين أن القيمة الأكبر تعني درجة ارتباط أقل.

وبذلك يمكن القول بأنه وفقاً لهذا النموذج يتم الحكم على التشابه بين الاستفسارات والوثائق، وكذلك بين الوثائق وبعضها البعض وفقاً لمقاييس الأبعاد والزوايا بينها لتحديد درجة تشابه تلك الوثائق مع بعضها، وبالتالي درجة صلتها بالاستفسار[٤].

ونخلص مما سبق إلى أن استخدام نموذج المتجهات يتيح المضاهاة الجزئية، حيث يعتمد على وضع أوزان أو قيم للوثائق، تستخدم لاحتساب درجة التشابه بين كل وثيقة مخزنة في النظام وبين استفسارات المستفيدين. وقد يتم احتساب درجة التشابه على أساس احتساب معدل تكرار المصطلح داخل الوثيقة، حيث أن كثرة مرات وروده تدل على أن علاقة الوثيقة بالاستفسار أكبر، وعلى العكس من ذلك فإن قلة تكراره تشير إلى علاقة أقل بين الوثيقة والاستفسار[٥].

وبذلك فإن نتيجة الاسترجاع في النظام تضم وثائق مرتبة بدرجة أكثر دقة وتحديداً، على أساس أن الوثائق الأكثر مطابقة لاحتياجات المستفيدين تأتي في بداية النتيجة ثم تتدرج الوثائق الأقل صلة بالاستفسار.

ثالثاً: نموذج الاحتمالات Probabilistic Model :

يرى البعض أن كل من النموذج البوليني، ونموذج حيز المتجهات قائمان على معايير جامدة، ففي المضاهاة اعتماداً على النموذج البوليني إما أن تقابل الوثيقة الشرط المنطقي أو لا تقابله، أما في المضاهاة المعتمدة على نموذج حيز المتجهات، فإنه يتم وضع حدود للتشابه (الحيز) وإما أن تكون الوثيقة واقعة ضمن تلك الحدود أو لا تكون. لذا فقد رأي البعض أنه ينبغي العمل وفقاً لنماذج أخرى أكثر مرونة، ومن ذلك نموذج احتمالات الذي يعمل على تحليل المصطلحات في كل من الاستفسارات والوثائق من النواحي النحوية syntactic أو الدلالية semantic أو الواقعية pragmatic لتحديد العلاقة بين الاستفسارات والوثائق[٦].

ويذكر أن هذا النموذج يعتمد على استخدام نظرية الاحتمالات كأساس لعملية المعالجة، فبدلاً من مطابقة نفس المصطلحات الواردة في وصف الوثائق، فإنه يتم وفقاً للنموذج الاحتمالي إحصاء أو تقدير الاحتمالات التي يمكن أن تكون فيها الوثيقة ذات صلة بمستفيد معين، ومن ثم ترتب الوثائق المسترجعة ترتيباً تنازلياً وفقاً لاحتمالات صلتها بالاستفسار وفائدتها بالنسبة للمستفيد؛ فعلى سبيل المثال يتم تحديد الوثيقة ذات الصلة عن طريق توظيف المعلومات التاريخية لاستخدام تلك الوثيقة لاحتساب وإحصاء احتمالات صلتها بالاستفسار. والمقصود بذلك أن يتم تتبع عدد المرات التي حكم فيها المستفيدون على الوثيقة بأنها ذات صلة بالاستفسار في حالة استخدامهم لنفس مصطلح البحث، وبمعنى آخر إذا استخدم مستفيد مصطلح بحث لاسترجاع وثائق، وحكم على وثيقة من بينها على أنها ذات صلة بالموضوع، وتكرر هذا الحكم على الوثيقة من قبل أشخاص آخرين استخدموا نفس مصطلح البحث، فإنه يمكن الحكم على الوثيقة بأنها ذات صلة بالاستفسار. ويذكر

أن هذه طريقة واحدة من بين طرق متعددة تستخدم لتحديد احتمالات صلة الوثيقة بالاستفسار[٧].
وهناك طرق أخرى عديدة من بينها على سبيل المثال: مراعاة أن ورود مصطلحات البحث متبوعة بكلمات محددة يعني احتمالات أكثر بصلتها باستفسار معين، كما أن عدد المرات التي ترد فيها تلك المصطلحات متبوعة ببعضها في الوثيقة يعزز من احتمالات صلتها بالاستفسار[٨].

وعلى الرغم من أن البعض يرى أن نموذج الاحتمالات يعد أفضل من النموذج البوليني، ونموذج حيز المتجهات، إلاّ أن هناك من يرى أن النتائج التي يتم الحصول عليها باستخدام هذا النموذج ليست أفضل بكثير من النموذجين الآخرين، وهو الأمر الذي أدى من وجهة نظرهم إلى عدم اقتناع مطوري النظم بالتحول إلى هذا النموذج بدرجة كبيرة[٩].

رابعاً: النموذج الضبابي Fuzzy set Model :

يتم تمثيل الوثائق والاستفسارات وفقاً للنموذج الضبابي بواصفات متصلة جزئياً بدلالات المحتوى، بمعنى أن المطابقة هنا تتم بشكل تقريبي وليس كلي.
ويعمل النموذج على أساس الوضع في الاعتبار أن كل مصطلح في الاستفسار يحدد مجموعة ضبابية fuzzy set ، ولكل وثيقة درجة عضوية a degree of membership ضمن تلك المجموعة. وعادة تكون درجة العضوية أقل من ١ وأكثر من صفر ، مما يعني أن الوثائق التي تحمل قيمة صفر ليست ضمن عضوية تلك المجموعة[١٠].
ويوضح Korfhage مثال للنتائج التي يمكن الحصول عليها باستخدام النموذج الضبابي؛ حيث يشير إلى أنه في حالة توجيه استفسار حول نوع من الكلاب الصغيرة يطلق عليه cocker spaniel فوفقاً للنموذج الضبابي فإن مصطلح سلالة الكلب breeds of dog يعد ذو صلة بالموضوع، لأنه من المتوقع أن تتضمن الوثائق الخاصة بالمصطلح الأخير معلومات مفيدة حول الموضوع المطلوب. كذلك فإن الوثائق التي تندرج تحت مصطلح الكلاب dogs عموماً يتم استرجاعها أيضاً لأنها قد تتضمن معلومات مفيدة؛ وإن كانت أقل فائدة من سابقتها[١١].
وكما هو واضح من المثال؛ فإن عملية المضاهاة لم تقتصر فقط على المصطلح المطابق تماماً لمصطلح البحث، ولكن تجاوزته إلى مصطلحات أخرى مطابقة له جزئياً وذلك على اعتبار أن هناك علاقة دلالية semantically تربط بين المصطلحات.

ويذكر أن العمليات الحسابية التي يتطلبها النموذج الضبابي تعد أقل تعقيداً من تلك التي يقوم النظام بإجرائها عند تطبيق نموذج الاحتمالات.

وهناك اتجاه يسمح بتحديد واصفات للوثيقة بحيث تعطي تلك الواصفات مؤشرات كمية أو نوعية تحدد قيمة المعلومات فيها؛ ومن ذلك على سبيل المثال: مهمة إلى حد ما fairly important ، ومهمة جداً very significant ، أو ذات صلة جزئية partially relevant ، ومن ثم تتم ترجمة تلك المصطلحات إلى قيم العضوية الخاصة بالوثيقة[١٢].

الخلاصة:

أوردت المقالة بعض نماذج استرجاع المعلومات المستخدمة في نظم استرجاع المعلومات لتحقيق المضاهاة اللازمة بين استفسارات المستفيدين وبين الوثائق المختزنة في نظام المعلومات، وقد تبين أن كل نموذج يعمل بطريقة مختلفة اعتماداً على معادلات حسابية خاصة به، وبالتالي فإن كل نموذج يحقق نتائج تختلف في دقتها وملاءمتها عن النموذج الآخر.

---

[1] Beaza–Yates, Ricardo & Ribeiro–Neto , Berthier . Modern Information Retrieval .- New York : ACM press , 1999 .- p23-24

[2] Korfhage , Robert R . Information Storage and Retrieval .- New York .- Wiley computer publishing , 1997 .- p84

[٣] بامفلح ، فاتن سعيد . أساسيات نظم استرجاع المعلومات الإلكترونية .- الرياض: مكتبة الملك فهد الوطنية ، ٢٠٠٦م .- ص١٦١

[4] Chowdhury , G. G. Introduction to Modern Information Retrieval . 2nd ed .- London: facet publishing, 2004 .- p176-180

[5] Beaza–Yates, Ricardo & Ribeiro–Neto , Berthier . op. cit .- p27-29

[6] Korfhage , Robert R .- op. cit .- p88-92

[٧] نقلاً عن: بامفلح ، فاتن سعيد . استرجاع المعلومات في المكتبات الرقمية.- مجلة المكتبات والمعلومات العربية.- س٢٧،ع٣ (يوليو ٢٠٠٧).

[8] Korfhage , Robert R .- op. cit .- p88-92

[9] Ibid

[10] Beaza–Yates, Ricardo & Ribeiro–Neto , Berthier . op. cit .- p34-35

[11] Korfhage , Robert R .- op. cit .- op. cit .- p92-93

[12] Ibid