



الحل لمشاكل تقنيات البحث بالإنترنت

اعداد المدرس المساعد

فائز موسى لهماود

2017-2018

الحلول لمشاكل تقنيات البحث بالإنترنت

تكمّن الحلول المناسبة لمثل هذي المشاكل في المعالجة الطبيعية للغة

يجمع العاملون في مجال محركات البحث بأن الحل لكافة هذه المشاكل يكمن في تصميم محركات البحث، وبرمجيات البحث عموماً، بحيث تستفيد من برمجيات المعالجة الطبيعية للغة **Natural Language Processing** عموماً، والمعالجة الطبيعية للمعنى **Natural Meaning Processing**، والاستفادة من قواعد اللغة. النقطة الأساسية هنا هي أن اللغة عموماً مبنية على أساس قواعد ذات أساس شبيه بالمعادلات الرياضية، كما أن عبارات اللغة وكلماتها تكون مبنية حسب هيكلية معينة، فكل جملة مبتدأ وخبر، أو فعل وفاعل ومفعول به، كما أن الكلمات تأتي عادة من جذور وأصول. وهذه الهيكلية المبنية ضمن اللغة تتناسب وطبيعة عمل برمجيات الحاسوب، والتي تحتاج إلى هيكلية معينة تنفذ على أساسها عملياتها.

أما المعالجة الطبيعية للمعنى فهي أكثر صعوبة، فكلمة مثل "راحة" يمكن لها أن تعني عدة أشياء، كالارتياح بعد التعب، أو الراحة الأبدية والتي تعني الموت، أو راحة اليد. ولهذا فإن برمجيات البحث يجب أن تتمكن من فهم الكلمة ضمن سياق النص، ودور الكلمة في هذا السياق. ولكن حتى هذا الأسلوب لا يفلح دوماً في فهم المعنى. جرب أن تقرأ ديواناً شعرياً جيداً، وستجد أن هنالك الكثير من الكلمات الصعبة، والتي قد يحاول البعض فهمها من السياق، ولكننا عندما نعجز عن ذلك فإننا نتجه إلى القاموس. وهذا أسلوب يمكن لبرمجيات الحاسوب اتباعه، أي الاعتماد على قاموس أو فهرس بالكلمات والعبارات الشائعة الاستخدام، والبحث في السياق.

وفيما يلي بعض مفاهيم البحث التي تعتمد على المعالجة الطبيعية للغة:

1- دراسة الجُمْل

تتميز الجُمْل بأنها ذات هيكلية محددة وواضحة، مما يساعد في فهم المعنى بشكل سريع. ومن خلال تحديد نوع الجملة، يمكن لبرمجيات البحث أن تقوم بعملها بشكل أكثر دقة. ولعمل ذلك فإن هذه البرمجيات يجب أن تعتمد على كميات كبيرة من المصادر والمراجع اللغوية، والتي تحتوي على عبارات وجمل ذات علاقات محددة مسبقا. وباستخدام هذه المراجع يمكن لبرمجيات البحث أن تفهم بشكل أفضل طبيعة العلاقة بين الكلمات المختلفة ومواقعها لغير صحيحة ضمن الجمل. ورغم الفائدة العظيمة للمراجع ، فإن مدى فعاليتها يعتمد على تحديثها باستمرار، وإلا فإنها تفقد فاعليتها تدريجيا. وباستخدام المراجع ، وتصريف الجمل وإعرابها يمكن الحصول على نظم بحث قوية يمكنها فهم عبارات البحث ومعاني الكلمات ضمن السياق بشكل أفضل.

2- إزالة الالتباس: اختيار المعنى الصحيح

أوضحنا في النقطة أعلاه كيف يمكن للمراجع ونظم تصريف وإعراب الجمل أن تسهم بشكل أفضل في فهم المعنى من خلال السياق. وهذه التقنية مفيدة إذا كنا نبحث ضمن نص معين، ولكنها ليست مفيدة عند إدخال الاستفسارات والاستعلامات ضمن محركات البحث، حيث تتكون هذه عادة من عدد محدود من الكلمات، وبالتالي فإن حجم النص غير كاف لتحليل معنى الكلمات. والحل لهذه المشكلة واضح إلى درجة الإحراج ويتمثل في سؤال المستخدم عن المعنى الذي يقصده؛ فعندما يُدخل المستخدم كلمة "راحة" مثلاً ضمن مربع الاستعلام، فإن محرك البحث يسأله عن المعنى المقصود، أو المعنى المراد بالبحث عنه، قبل الشروع بعملية البحث. وتوجد اليوم العديد من محركات البحث التي تستخدم قواميس مضمنة تقوم بتقديم قوائم بالمعاني المختلفة التي تمثلها الكلمة الواحدة قبل الشروع في عملية البحث.

3- التصنيف التلقائي

هذا الأسلوب هو الأقدم، حيث يتم تصنيف الوثائق حسب تصنيفات وفروع معينة، ومن ثم البحث بشكل منفصل ضمن كل تصنيف عن المعلومات المطلوبة. فمثلاً، عند البحث عن كلمة "نواة" فإن بحثك قد يقودك إلى تصنيف يتعلق بعلوم الزراعة، وأنواع الحبوب، وما إلى ذلك، ولكنه في الوقت نفسه قد يقودك إلى تصنيف يقع ضمن علوم الفيزياء النووية. والحل هنا يكمن في تصنيف الوثائق المتعلقة بأنوية الحبوب والمزروعات في قسم الزراعة مثلاً، والوثائق المتعلقة بأنوية الذرات في قسم الفيزياء الذرية. ولكن ماذا لو كانت لدينا وثيقة تتعلق بتأثير التجارب النووية على أنوية الحبوب والمزروعات؟ هل يجب عندها وضع الوثيقة في القسمين، أم ما هو الحل؟ وهنا تنشأ لدينا مشكلة المعنى المزدوج. ومن هنا تأتي أهمية التصنيف التلقائي، حيث يتم استخدام علوم النحو من تصنيف وإعراب، واستخدام المراجع والقواميس، بحيث يتمكن النظام من "فهم" المواضيع الرئيسية في وثيقة ما.

ويتم ذلك باستخدام أساليب إحصائية تقوم بدراسة تكرار الكلمات ضمن وثيقة ما، ومن ثم تحديد السياق، والذي يساعد في عملية البحث. وكمثال على ذلك لنأخذ كلمة، أو اسم مثل "فهد" أو "ليث" وهي أسماء عربية دارجة. ولنفترض أننا كتبنا موضوعا في مجلتنا عن شخص اسمه "فهد"، ولنفترض أن محرك بحث أراد تصنيف مقالنا هذا. في هذه الحال، وإذا كان محرك البحث يستخدم أسلوب التصنيف التلقائي، فإنه سيقوم من خلال دراسة النص ونوعية الكلمات الموجودة فيه وعلاقاتها وتكرارها، بتحديد أن المقال ينتمي إلى تصنيف علوم الحاسوب والإنترنت وليس الحيوانات والوحوش البرية. وتتبع برمجيات التصنيف التلقائي قواعد معينة يحددها المبرمجون، أو يمكن للآلة نفسها أن تتعلم ذاتيا كيف تقوم بتصنيف الكلمات. أو يمكن استخدام الأسلوبين، بحيث يتم تصنيف الوثائق التي تتبع نمطا معيناً بشكل تلقائي، في حين يتم تحويل تلك التي يستحيل تصنيفها إلى عامل بشري كي يقوم بذلك .

4- زيادة دقة الاستعلامات

من يستعمل محركات البحث باستمرار يعلم أنه من المحتوم الحصول على مئات الألوف من النتائج على الأقل عند البحث عن موضوع معين بشكل عام. أما عند تضيق نطاق البحث، فإننا نحصل على مقدار أقل من النتائج الأكثر دقة. ولزيادة دقة الاستعلامات، فإن بعض محركات البحث يقوم بتعديل الاستعلام، وذلك بتقديم معان مرادفة لكلمات البحث إلى المستخدم، لمساعدته على تحديد موضوع بحثه بدقة أكبر. ففي مثل هذه النظم، وإذا كان موضوع استعلامك هو "الرياضة" مثلاً، فإن النظام يقدم لك عدداً من البدائل التي تساعد في تحديد موضوع البحث مثل "التربية البدنية"، "اللياقة البدنية" أو "اللياقة" وغير ذلك من المرادفات التي تساعد على توجيه العملية بحيث يحصل المستخدم على أكبر كم من النتائج الدقيقة التي تعبّر عن موضوع البحث.

5- ربط المفاهيم

إذا جربت اليوم أن تبحث عن العبارة "اسطوانات الليزر" (وهي العبارة العامية والقديمة للأقراص المدمجة) فإنك لن تحصل على الكثير من النتائج المفيدة من محركات البحث، حيث أن عبارة "الأقراص المدمجة" هي الكلمة الشائعة والمستخدم في الغالبية العظمى من المطبوعات. ولهذا فإن العديد من محركات البحث هذه الأيام تستخدم أسلوب الربط بين المفاهيم، بحيث أنك إذا قمت بكتابة العبارة "أقراص الليزر" فإنك ستحصل على معلومات عن "الأقراص المدمجة". ويعتمد هذا الأسلوب على تحديد العلاقة بين الكلمات والعبارات في قاعدة البيانات بشكل مسبق، كما أنه مفيد عند البحث في وثائق متعددة اللغات، فالبحث عن كلمة "أقراص الليزر" يمكن أن يعطي نتائج لوثائق باللغة الإنجليزية عن "Compact Disks أو CDs وما إلى ذلك، وهذا أسلوب مفيد جدا في عالم إنترنت لم تعد فيه اللغة الإنجليزية هي السائدة .