كيف تعمل محركات البحث؟

من أهم ميزات الإنترنت وعنصره الأكثر وضوحًا وانتشارًا "الشبكة العالمية" هو أن هنالك مئات الملايين من الصفحات المتاحة، في انتظار تقديم المعلومات عن مجموعةٍ مذهلةٍ من المواضيع المتنوعة. إلا أن السيئة التي يعاني منها الإنترنت هي أن معظم مئات الملايين من الصفحات المتاحة تلك معنونةٌ وفقًا لرغبة المؤلف، وتقريبًا جميعها موجودةٌ على مخدماتٍ بأسماءٍ غامضةٍ. فعندما تحتاج إلى معرفة موضوعٍ معينٍ، كيف يمكنك أن تعرف أيّ الصفحات يجب أن تقرأ؟ إذا كنت مثل معظم الناس، فإنك تزور أحد محركات البحث على الإنترنت.

يقوم محرك البحث بأرشفة مئات الملايين من الصفحات في اليوم الواحد

محركات البحث على الإنترنت هي مواقعُ خاصةٌ على الويب مصممةٌ لمساعدة الأشخاص في العثور على المعلومات المخزنة على مواقع أخرى. وهناك اختلافاتٌ في الطرق التي تعمل بها محركات البحث المختلفة، ولكن جميعها تؤدي هذه المهام الأساسية الثلاث:

1. تبحث في الإنترنت، أو في أجزاء معينةٍ من الإنترنت، بالاعتماد على كلماتٍ هامةٍ.
2. تحتفظ بأرشيف **index** للكلمات التي تجدها ومواقعها.
3. تسمح للمستخدمين بالبحث عن كلماتٍ أو مجموعاتٍ من الكلمات الموجودة في ذلك الأرشيف.

امتلكت محركات البحث القديمة في أرشيفها بضع مئات الآلاف من الصفحات والوثائق، وتلقت ما يُقارب ألف أو ألفي استعلام (طلب بحث) كلّ يومٍ. أما اليوم، يملك محرك بحثٍ شهيرٍ في أرشيفه مئات الملايين من الصفحات، ويستجيب لعشرات الملايين من الاستعلامات يوميًا. في هذه المقالة، سنخبرك بكيفية تنفيذ هذه المهام الضخمة، وكيف تعمل محركات البحث لكي تمكنك من العثور على المعلومات التي تحتاجها على الويب.

زحف الويب Web Crawling

عندما يتحدث معظم الناس عن محركات بحث الإنترنت فإنهم يعنون في الحقيقة محركات بحث الشبكة العالمية. فقبل أن يصبح الويب العنصر الأكثر وضوحًا من الإنترنت، كان هناك بالفعل محركات بحثٍ لمساعدة الناس على العثور على المعلومات على الشبكة. كان هناك برامجٌ مثل غوفر **gopher** وآرتشي **Archie** تحتفظ بأرشيف للملفات المخزنة على خوادم متصلةٍ بالشبكة، فخفّضت بشكلٍ كبير من الوقت اللازم للعثور على البرامج والمستندات. وفي أواخر الثمانينيات ارتبط الحصول على نتائج هامةٍ من الإنترنت مع معرفة كيفية استخدام غوفر، آرتشي، فيرونيكا **Veronica**، وغيرها من البرامج المماثلة.

اليوم، يحدد معظم مستخدمي الإنترنت عمليات البحث التي يقومون بها على الويب، لذلك سنقتصر في هذه المقالة على محركات البحث التي تركز على محتويات صفحات الويب.

قبل أن يتمكن محرك البحث من إخبارك عن مكان ملفٍ أو مستندٍ ما، عليه في البداية أن يعثر على ذلك الملف. للعثور على معلوماتٍ في مئات الملايين من صفحات الويب الموجودة، يستخدم محرك البحث روبوتاتٍ برمجيةً خاصةً تُدعى "العناكب" لإنشاء قوائم بالكلمات الموجودة على مواقع الويب. عندما يقوم العنكبوت ببناء قوائمه، تُسمّى العملية بـ "زحف الويب" **Web Crawling** (إن تسمية جزء كبير من الإنترنت بالشبكة العالمية هو ما استُمدت منه تسمية تلك الروبوتات بالعناكب، وذلك لارتباط الشبكات بالعناكب). ومن أجل بناء والاحتفاظ بقائمةٍ مفيدةٍ من الكلمات، فإن عناكب محركات البحث يجب أن تبحث في الكثير من الصفحات.

كيف يبدأ أيّ عنكبوتٍ رحلاته عبر الويب؟ نقاط البداية المعتادة هي قوائم الخوادم التي تُستَخدَم بشكلٍ كبيرٍ وصفحات الويب الشعبية جدّاً. حيث يبدأ العنكبوت من موقعٍ شعبيٍّ، مُؤرشِفًا الكلمات على صفحاته ومُتبِعًا كلّ رابطٍ يجده داخل الموقع. وبهذه الطريقة، يبدأ نظام العنكبوت بالتنقل والتوسع بسرعة، وينتشر عبر الأجزاء الأكثر استخدامًا على نطاقٍ واسعٍ من الويب.

بدأ غوغل **Google** كمحرك بحثٍ أكاديميٍّ. وفي الورقة البحثية التي تصف كيفية بناء النظام، قدم الباحثان سيرجي برين **Sergey Brin** ولَورنس بيج **Lawrence Page** مثالًا عن مدى السرعة التي يمكن لعناكبهما أن تعمل بها، حيث بنيا نظامهما الأولي لاستخدام عناكب متعددة، عادةً ثلاثة عناكب في وقتٍ واحد. كان بإمكان كلّ عنكبوتٍ أن يبقي نحو **300** اتصالٍ مفتوحٍ بصفحات الويب في وقتٍ واحدٍ. وفي الأداء الأمثل وباستخدام أربعة عناكب، كان بإمكان نظامهم الزحف عبر **100** صفحةٍ في الثانية، وتوليد ما يُقارب **600** كيلو بايت من البيانات في كلّ ثانية.

لضمان استمرار عمل كلّ شيءٍ بسرعة يجب بناء نظام لتقديم المعلومات اللازمة للعناكب. كان لنظام غوغل المبكر خادم مخصص لتقديم عناوين **URL** للعناكب. بدلًا من الاعتماد على مزود خدمة إنترنت للوصول إلى مخدم أسماء النطاقات (**DNS**) الذي يترجم اسم خادمٍ ما إلى عنوان، كان لدى غوغل مخدم **DNS** خاص به، وذلك لتقليل التأخيرات قدر الإمكان.

عندما ينظر عنكبوت غوغل في صفحة أتش تي أم أل (**HTML**)، فإنه يأخذ علمًا بأمرين:

- الكلمات الموجودة داخل الصفحة.

- مواقع تواجد هذه الكلمات.

الكلمات الموجودة ضمن العناوين والعناوين الفرعية والوسوم الوصفية **Meta Tags** المخفية الدالة على المحتوى وغيرها من الأجزاء النصية التي تحمل أهميةً نسبيةً ضمن الصفحة تؤخذ جميعها في عين الاعتبار للعودة إليها أثناء عملية البحث التي يقوم بها المستخدم. بُني عنكبوت غوغل لأرشفة كلّ الكلمات المهمة في الصفحة، دون اعتبار أدوات التعريف مثل **a** و **an** و **the**. وهناك عناكب أخرى تعمل بمنهجياتٍ مختلفةٍ.

تهدف هذه النهج المختلفة عادةً إلى جعل العنكبوت يعمل بشكلٍ أسرع، أو لتسمح للمستخدمين بالبحث بشكلٍ أكثر كفاءةً، أو كليهما. على سبيل المثال، بعض العناكب تتبع الكلمات في العنوان، العناوين الفرعية والروابط، بالإضافة إلى أكثر 100 كلمةٍ مستخدمةٍ في الصفحة وكل كلمة في الأسطر العشرين الأولى من النص. ويقال إن محرك البحث ليكوس Lycos يستخدم هذه المنهجية في الزحف عبر النت.

هناك أنظمةٌ أخرى، مثل ألتافيستا **AltaVista**، تذهب في الاتجاه الآخر، حيث تقوم بأرشفة كلّ كلمةٍ في الصفحة، بما في ذلك "**a**" و"**an**" و"**the**" وغيرها من الكلمات غير الهامة، أي يهدف هذا النهج إلى الأرشفة الكلية في حين تقابله نهج أخرى تركز اهتمامها على الـ **meta tags** الموجودة في الجزء المخفي من صفحة الويب.

الوسوم الوصفية Meta Tags

تسمح العلامات الوصفية لمالك صفحةٍ بتحديد الكلمات الرئيسية (المفتاحية) والمفاهيم التي تُؤرشف الصفحة من خلالها. يمكن أن يكون ذلك مفيدًا، لا سيما في الحالات التي قد تكون فيها للكلمات الموجودة في الصفحة معانٍ مزدوجةٌ أو ثلاثيةٌ حيث تقوم هذه الوسوم بتوجيه محرك البحث في اختيار أيٍّ من المعاني الممكنة لهذه الكلمات هي الصحيحة. ومع ذلك، هناك خطرٌ في إفراط الاعتماد على العلامات الوصفية، لأن بعض ملاك الصفحات غير المدركين أو غير المبالين قد يضيفون علاماتٍ وصفيةٍ تناسب مواضيع شائعةً جدًّا ولكن لا علاقة لها بالمحتويات الفعلية للصفحة. ولتجنب ذلك، تعمل العناكب على ربط العلامات الوصفية بمحتوى الصفحة ورفض العلامات الوصفية التي لا تتطابق مع الكلمات الموجودة في الصفحة.

وكل هذا بافتراض أن مالك الصفحة يريد فعلًا تضمين صفحته في نتائج محرك البحث، ففي كثيرٍ من الأحيان قد لا يرغب مالك الصفحة بظهور صفحته على محرك بحثٍ شهيرٍ أو وصول نشاط عناكب البحث إليها. على سبيل المثال، لتكن لدينا لعبةٌ تقوم ببناء صفحاتٍ جديدةٍ ونشطةٍ في كلّ مرةٍ تُعرض فيها أجزاءٌ من الصفحة أو تُتّبع روابطٌ جديدةٌ. إذا وصل عنكبوت البحث إلى إحدى هذه الصفحات وبدأ بتتبع كل الروابط للصفحات الجديدة فقد تفشل عندها اللعبة بالتمييز بين نشاط العنكبوت وبين لاعبٍ بشريٍّ عالي السرعة وتدخل في حلقةٍ مفرغةٍ.

ولتجنب مثل هذه الحالات، طُوِّر بروتوكول استبعاد الروبوت (العنكبوت) **Robot Exclusion Protocol**. هذا البروتوكول (يُضمَّن في قسم الوسوم الوصفية في بداية صفحة الويب) يخبر العنكبوت بأن يترك الصفحة حيث لا يقوم بأرشفة الكلمات ولا يحاول تتبع الروابط الموجودة ضمن الصفحة.

بناء الأرشيف

بمجرد أن تنهي العناكب مهمة إيجاد المعلومات على صفحات الويب (علمًا أن هذه المهمة فعليًا لا تنتهي أبدًا نتيجةَ الطبيعةِ المتغيرة باستمرار لصفحات الويب والتي تعني أن العناكب في حالة زحفٍ مستمرٍ) يتوجب عندها على محرك البحث أن يخزن المعلومات بطريقةٍ

تجعلها مفيدةً. هناك عنصران أساسيان متعلقان بجعل البيانات المُجمعة في متناول المستخدمين:

- المعلومات المخزنة مع البيانات
- النهج المتّبع في أرشفة المعلومات

في أبسط الحالات، يخزن محرك البحث الكلمة وعنوان **URL** حيث يُعثر عليها، إلا أن هذا يجعل محرك البحث محدود الاستخدام حيث لا توجد طريقة لمعرفة أهمية استخدام الكلمة في الصفحة، أو عدد مرات استخدام هذه الكلمة، أو إذا كانت الصفحة تحتوي على روابط إلى صفحاتٍ أخرى تحتوي على الكلمة التي عُثر عليها. وبعبارةٍ أخرى، لن تكون هناك طريقةٌ لبناء قائمة الترتيب التي تحاول تقديم الصفحات الأكثر فائدةً في أعلى قائمة نتائج البحث.

لتحقيق نتائج أكثر فائدةً، تخزن معظم محركات البحث أكثر من مجرد الكلمة وعنوان **URL** الخاص بها. قد يقوم محرك البحث بتخزين عدد المرات التي تظهر فيها الكلمة على الصفحة، أو تعيين وزنٍ لكلّ إدخالٍ بحيث يكون له قيمٌ متزايدةٌ تُخصص للكلمات عند ظهورها بالقرب من أعلى المستند أو في عناوين فرعية أو في روابط أو في الوسوم الوصفية أو في عنوان الصفحة. لكلّ محرك بحثٍ تجاريٍّ معادلةٌ مختلفةٌ لتعيين الوزن للكلمات في أرشيفه، وهو أحد الأسباب التي تؤدي إلى ظهور قوائمَ مختلفة بترتيب مختلفٍ للصفحات عند البحث عن نفس الكلمة على محركات البحث المختلفة.

بغض النظر عن المجموعة الدقيقة من المعلومات الإضافية التي يخزنها محرك البحث ستَرمّز البيانات بحيث تُوفَّر مساحة التخزين. على سبيل المثال، يظهر بحث غوغل الأصلي باستخدام **2** بايت (**8** بت لكلٍّ منهما) لتخزين المعلومات حول الوزن بما في ذلك ما إذا كانت الكلمة قد كُتِبَت بحرفٍ كبيرٍ وحجم خطها وموقعها ومعلوماتٍ أخرى للمساعدة في ترتيب النتيجة. قد يستهلك كلٌّ من تلك العوامل **2** أو **3** بتات داخل التجميع ذي الحجم **2** بايت (**8** بتات = **1** بايت). ونتيجةً لذلك، يمكن تخزين قدرٍ كبيرٍ من المعلومات في شكلٍ مضغوطٍ جدًا. بعد ضغط المعلومات، تكون جاهزةً للأرشفة.

للأرشيف غرضٌ واحدٌ هو السماح بإيجاد المعلومات بأسرع ما يمكن. هناك العديد من الطرق لبناء الأرشيف، ولكنّ واحدةً من أكثر الطرق فعاليةً هي بناء جدول التجزئة **Hash Table**. في التجزئة تُطبَّق معادلةٌ لإرفاق قيمةٍ عدديةٍ لكلّ كلمةٍ. صُمِّمت هذه المعادلة بحيث توزع المدخلات بالتساوي عبر عددٍ مُحدد مسبقًا من الأقسام. هذا التوزيع العددي يختلف عن توزيع الكلمات عبر الأبجدية، وهذا هو مفتاح فعالية جدول التجزئة.

في اللغة الإنجليزية هناك بعض الحروف التي يكثر تواجدها في بداية الكلمات، بينما يقل ابتداء الكلمات ببعض الحروف الأخرى. على سبيل المثال، ستجد أن قسم الحرف **M** من القاموس هو أضخم بكثير من قسم الحرف**X** وهذا التفاوت يعني أن إيجاد كلمة تبدأ بحرفٍ شائعٍ جدًّا قد يستغرق وقتًا أطول بكثير من العثور على كلمةٍ تبدأ بحرفٍ أقل شيوعاً. عملية التجزئة تلغي هذا الفرق وتقلل من الوقت المتوسط الذي يستغرقه لإيجاد المدخل كما تفصل الأرشيف عن المدخل. يحتوي جدول التجزئة على رقم التجزئة **hash number** بالإضافة إلى مؤشر إلى البيانات الفعلية، والتي يمكن فرزها في أي طريقةٍ تسمح بتخزينها بأكبر قدرٍ من الكفاءة. إن الجمع بين كفاءة الأرشيف والتخزين الفعال يجعل من الممكن الحصول على نتائج بسرعةٍ، حتى عندما يقوم المستخدم بإنشاء بحثٍ معقد.

إنشاء البحث

يتضمن البحث من خلال الأرشيف مستخدمًا يقوم بإنشاء الاستعلام وإرساله من خلال محرك البحث. يمكن أن يكون الاستعلام بسيطًا جدًّا، كلمةً واحدةً على الأقل. يتطلب إنشاء استعلامٍ أكثر تعقيداً استخدام المعاملات المنطقية التي تسمح لك بتنقيح وتوسيع مصطلحات

البحث. والمعاملات المنطقية الأكثر شيوعاً هي:

- **And**: يجب أن تظهر جميع المصطلحات التي نجمع بينها بكلمة **AND** في الصفحات أو المستندات، وبعض محركات البحث تستخدم المعامل "+" بدلًا من الكلمة **AND**.
- **OR**: على الأقل واحدٌ من المصطلحات التي نجمع بينها بكلمة **OR** يجب أن تظهر في الصفحات أو المستندات.
- **NOT**: يجب ألا يظهر المصطلح أو المصطلحات التالية لكلمة **NOT** في الصفحات أو المستندات، وبعض محركات البحث استبدلت الكلمة **NOT** بالمعامل "-".
- **FOLLOWED BY**: يجب أن تتبع أحد المصطلحات مباشرةً بمصطلح آخر.
- **NEAR**: يجب أن يكون أحد المصطلحات على بعد عددٍ محددٍ من الكلمات عن المصطلح الآخر.
- **QUOTATION MARKS** (علامات الاقتباس): يتم التعامل مع الكلمات بين علامتي الاقتباس على أنها عبارةٌ واحدةٌ، ويجب العثور على هذه العبارة داخل المستند أو الملف.

مستقبل البحث

عمليات البحث التي تحددها المعاملات المنطقية هي عمليات بحثٍ حرفيةٍ، حيث يبحث المحرك عن الكلمات أو العبارات تمامًا كما تُدخل ويمكن أن يؤدي ذلك إلى مشكلة عندما يكون للكلمات المُدخلة معانٍ متعددة. على سبيل المثال، كلمة "**Bed**" باللغة الإنجليزية يمكن أن تعني مكان النوم أو المكان الذي تُزرع فيه الزهور أو مساحة التخزين في شاحنة أو المكان حيث تضع الأسماك بيوضها. إذا كنت مهتمًا بمعنىً واحدٍ فقط من هذه المعاني فقد لا ترغب بمشاهدة صفحات تعرض جميع المعاني الأخرى. يمكنك في هذه الحالة إنشاء عملية بحثٍ حرفيٍّ يحاول تجاهل المعاني غير المرغوب فيها، ولكن من الجميل أن يساعدك محرك البحث نفسه في ذلك.

أحد مجالات البحث الخاصة بمحركات البحث هو البحث القائم على المفهوم. بعض هذه الأبحاث ينطوي على استخدام التحليل الإحصائي للصفحات التي تحتوي على الكلمات أو العبارات التي تبحث عنها من أجل العثور على صفحاتٍ أخرى قد تكون مهتمًا بها. من الواضح أن المعلومات المخزنة حول كلّ صفحةٍ تُعتبر أكبر بالنسبة لمحرك البحث القائم على المفهوم مقارنةً بمحركات البحث القائمة على عمليات البحث الحرفية ولذلك يلزم إجراء المزيد من عمليات المعالجة لكلّ عملية بحثٍ. ومع ذلك، فإن العديد من المجموعات تعمل على تحسين كل من النتائج والأداء لهذا النوع من محركات البحث، بينما انتقل آخرون إلى مجالٍ آخر من البحوث، يدعى استعلامات اللغة الطبيعية **Natural-Language Queries**.

الفكرة وراء استعلامات اللغة الطبيعية هي أنه يمكنك كتابة سؤالك بنفس الطريقة التي تسأل بها إنسان يجلس بجانبك حيث لا حاجة لتتبع المعاملات المنطقية أو هياكل الاستعلامات المعقدة. يُعتبر موقع **AskJeeves.com** من المواقع الأكثر شعبيةً والذي يعتمد استعلامات اللغة الطبيعية، وهو يقسم الاستعلام إلى كلماتٍ مفتاحيةٍ ويطبقها بعد ذلك على أرشيف المواقع الذي قام ببنائه، لكن هذه الطريقة تنجح فقط مع استفساراتٍ بسيطةٍ، ولكن المنافسة شديدةٌ لتطوير محرك استعلام اللغة الطبيعية يستطيع استقبال استعلامات أكثر تعقيدًا.

- التاريخ: 2018-07-12
- التصنيف: كيف تعمل الأشياء؟

#المعلومات #الإنترنت #محركات البحث #زحف الويب #الأرشيف

المصادر

- howstuffworks
- الصورة

المساهمون

- ترجمة
  - زين الهوشي
- مُراجعة
  - فرح درويش
- تحرير
  - حنان مشقوق
  - رأفت فياض
- تصميم
  - رنيم ديب
- صوت
  - أهلة عبيد
- نشر
  - أمل أحمد