خوارزمية بحث متقدم لمحركات البحث في النصوص العربية

مروان البواب عضو مجمع اللغة العربية بدمشق

مقدمة

تعدَّدتْ طرائقُ محركات البحث في النصوص العربية، وتنوَّعتْ أساليبُها، وتَحسَّنَ أداؤها، وزادَ نُجُوعُها على نحوٍ ملحوظٍ بَّعلَّى في سرعتها وفي مضمون نتائجها. ومع هذا كلِّه، مازالت هذه المحركاتُ دون المستوى المنشود. ولا شكَّ في أن عددًا من العوائق والصعوبات يَحول دون الوصول إلى هذا المستوى، لعلَّ أهمَّها طبيعةُ الكتابة في النصوص العربية.

يتناول هذا المقالُ أثرَ طبيعة الكتابة العربية في عمليات البحث، ثم يَعرض خوارزمية بحثٍ متقدِّم يعمل ضمن محركات البحث في النصوص العربية، يُمكِّن من الحصولِ على نتائج بحثيةٍ أكثرَ تحديدًا من نتائج البحث العادية، ومن الحصولِ على نتائج لا تستطيع محركاتُ البحث الوصولَ إليها.

ولا بدَّ قبل البدء من الإشارة إلى النقاط الآتية:

- ليس الغرضُ من هذا المقترح تقديمَ محركِ بحثٍ عربي جديد، بل إضافةَ بحثٍ متقدمٍ إلى محركات البحث المتعدّمُ مكمّلُ لعمل محركات البحث، وليس بديلاً عنها.
- يُمْكن الاستفادةُ من هذا البحث المتقدم في وجوهٍ متعددة، غير أنَّ الاستفادةَ العظمى منه ستكون من نصيب الدراساتِ اللغويةِ المُنِيَّةِ على المدوَّنات العربية.
- لن أتعرَّض في هذا المقال إلى الجوانب الفنية المتعلِّقة بهذا البحث المتقدم، وإنما سأتناول جوانبَه اللغوية فحسب.

1 أثر طبيعة الكتابة العربية في عمليات البحث

فيما يلي أهمُّ خصائص الكتابة باللغة العربية، وأثرها في عمليات البحث في النصوص العربية.

1.1 غياب الشكل

يؤدِّي غيابُ الشكل في النصوص العربية إلى صعوباتٍ في عمليات البحث فيها تتجلَّى في تحديد معنى الكلمة المبحوث عنها ووظيفتها، وسأكتفى بذكر أمثلةٍ عليها:

- تغير المعنى نتيجة تغيُّر المبنى

مثال: (البَرّ): اليابسة (ضد البحر).

(البِرّ): الخير والإحسان.

(البُرّ): حبُّ القمح.

وعلى ذلك، فإن محرِّكاتِ البحث لا تستطيع تحديدَ المعنى المطلوب إذا لم تكن النصوصُ مشكولة.

وتجدر الإشارة إلى أنَّ شَكْلَ الكلمات شرطٌ لازمٌ لتحديد معانيها، ولكنه غيرُ كافٍ. فكثيرٌ من الكلمات المضبوطة بالشكل التام لا يمكن تحديدُ معناها إلا من سياق النص. فمثلاً كلمةُ (الْمُرَمُ) المشكولةُ كليًا لها معنيان؛ الأول: بلوغُ أقصى الكِبَر في السنِّ، والآخرُ الشكل الهندسيُّ المعروف.

- تغير وظيفة الكلمة

تُبيِّن الأمثلة الآتية أثر غياب الشكل في وظيفة الكلمة:

مثال أول: ملك: (مَلَكَ) فعل، (مَلِكٌ) اسم

مثال ثانٍ: عن: (عَنْ) حرف جر، (عَنَّ) فعلُ ماض

مثال ثالث: قتل: (قَتَلَ) مبنيٌّ للمعلوم، (قُتِلَ) مبنيٌّ للمجهول

مثال رابع: أسرة: (أُسرة) مفرد، (أُسِرَّة) جمع

2.1 اتصال أحرفِ سابقةِ الكلمةِ بها

مثال: لَسَعَتْ (من لَسَعَ يَلْسَعُ)، لَسَعَتْ = لَ + سَعَتْ (حرف اللام + الفعل سَعَى يَسْعَى)

3.1 اتصالُ أحرفِ لاحقة الكلمة بها

مثال: دَلَّكَ (فعل ماضٍ مزيد)، دَلَّكَ = دَلَّ + كَ (فعل ماضٍ مجرد + ضمير متصل)

4.1 تَعَدُّدُ الصور المختلفة للكلمة

تبيُّن الأمثلةُ الآتية نماذج من الصور المختلفة التي يمكن أن تَردَ بما الكلماتُ في النصوص العربية:

مثال 1: غَزَا، غَزَتْ، غَزَوَا، غَزَوْا، غَزَوْتُ، غَزَوْنَا، غَرَوْنَا، غَرَانَا، غَرَوْنَا، غَرَا

مثال 2: مسطرة، مسطرته (قلب التاء المربوطة إلى مبسوطة)

مثال 3: أخوه (رفع)، أخاه (نصب)، أخيه (جر) (علامات الإعراب الفرعية)

مثال 4: كتابان، كتابين (المثنى في حالة: الرفع، والنصب والحر)

مثال 5: فتي، فتيان، عصا، عصوان (تثنية المقصور)

مثال 6: حمراء، حمراوان (تثنية الممدود)

مثال 7: القاضي، القاضون (جمع المنقوص)

مثال 8: يسعى، كَمْ يَسْعَ، يغزو، كَمْ يَغْزُ (المضارع المحزوم المعتل)

مثال 9: سماؤه، سماءه، سمائه (رسم الهمزة)

2 خوارزمية بحثٍ متقدِّمٍ لمحرِّكات البحث العربية

تقوم فكرة هذه الخوارزمية على مبدأ البحث في النصوص العربية اعتمادًا على ثلاثة عناصر؛ هي: جذر الكلمة (Root) P، ونوعُها (Kind) K، ووزنُها (Pattern) P. وهذا يقتضي أن نزوِّدَ كلمات النصوص بالثلاثية: (R, K, P). وبذلك يمكننا البحث:

- على مستوى: الجذر (-,-,R)، أو النوع (-,K,-)، أو الوزن (R,-,-)
 - أو على مستوى: الجذر والوزن $(\mathbf{R}\,,\,-\,,\,\mathbf{P})$ ، أو الجذر والنوع $(-\,,\,\mathbf{K}\,,\,\mathbf{R})$ ، أو الوزن والنوع $(\mathbf{R}\,,\,\mathbf{K}\,,\,\mathbf{P})$ ،
 - أو على مستوى الجذر والوزن والنوع (R, K, P) في آنٍ معًا.

وتحدر الإشارة إلى أنه **لا يجب** تزويدُ جميع كلمات النصوص بالثلاثية المشار إليها، بل يُستثنى منها الكلمات التي لا طائل في البحث عنها بواسطة جذورها أو أوزانها أو أنواعها. وفيما يلي أهمُّ أصناف هذه الكلمات:

- أسماء الأعلام؛ نحو: إبراهيم، مريم، حسان، فيليب، أحمد، هند...
 - أسماء الدول، والمدن، والجبال، والبحار، والأنهار...
 - أسماء الحيوانات، والنباتات، والمواد، والمنتجات، والمخترعات...
- الكلمات الدخيلة والمعرَّبة ؛ نحو: آزوت، بورصة، ألبوم، برتقال، بارود، صابون...
- الأدوات؛ نحو: إذا، لو، الذي، مهما، كي، حيث، لم، لن، في، لنا، له، عليه...
 - الضمائر المنفصلة؛ نحو: أنا، نحن، أنتم، هو، هي، هم، إياه، إياها، إياكم...
 - الأعداد؛ نحو: ثلاث، أربعة، أحد عشر، خمسون، مئة، ألف...

وقد بلغت النسبةُ المئوية لمجموع هذه الأسماء والأدوات قرابةً 35% من كلمات النصوص، استنادًا إلى إحصائياتٍ أجريناها على عددٍ من النصوص.

وفيما يلى عناصر هذه الخوارزمية:

1.2 جذر الكلمة

يتألف جذر الكلمة في الأعم الأغلب من ثلاثة أحرف، ومن أربعةِ أحرف في أقلِّ القليل.

أما الفائدة الأساسية التي بَحنيها من البحث على مستوى الجذر، فتكمن في الحصول على جميع

الصور التي يمكن أن ترد فيها الكلماتُ المشتقة من هذا الجذر. ومعلومٌ أن هذه الصور المتعدِّدةَ للكلمة تعود إلى عددٍ من الاعتبارات مثل: وجوه تصريف الكلمة، وحالاتها الإعرابية، واتصال السوابق واللواحق بحا... ويرتبط عددُ هذه الكلمات المشتقة (ومن ثَمَّ عددُ صورها) بخصوبة الجذر اللغوية؛ فكلما زادت خصوبة الجذر، زاد عددُ صور الكلمات المشتقةِ منه. وقد يصل عددُ هذه الصور إلى بضعة آلاف للجذر الواحد.

وهذا يعني أن البحث على مستوى الجذر يتناول جميع الكلمات المتعلّقة به من أفعال (مجردة ومزيدة، مبنية للمعلوم ومبنية للمحهول، مؤكّدة وغير مؤكّدة) في صيغة الماضي والمضارع والأمر، ومصادر (أصلية، وميمية، ومصدر المرة، ومصدر الهيئة)، وأسماء مشتقة (كاسم الفاعل، واسم المفعول، واسم الزمان والمكان، وأسماء الآلة، ومبالغة اسم الفاعل، والصفات المشبهة) في حالة الإفراد والتثنية والجمع، والرفع والنصب والجر، والتعريف والتنكير والإضافة... فمثلاً، إذا كان البحث على مستوى الجذر (حسب)، فإن نتائج البحث ستعطى الكلمات الآتية:

حَسَبَ يَحْسُبُ، حَسِبَ يَحْسَبُ، حَسُبَ يَحْسَبُ، حُسِبَ يَحْسَبُ، خُسِبَ يُحْسَبُ... (أفعال مجردة) حَاسَبَ يُحَاسِبُ، احْتَسَبَ يَحْتَسِبُ، تَحَسَّبَ يَتَحَسَّبُ، تَحَاسَبَ يَتَحَاسَبُ... (أفعال مزيدة) حَسْب، حِسَاب، حُسْبَان، حِسْبَة، مُحاسَبَة، احْتِساب، تَحَسُّب، تَحاسُب... (مصادر أصلية) حَسْبَة، حِسْبَة، مَحْسَب... (مصدر المرة، ومصدر الهيئة، والمصدر الميمي)

حاسِب، مَحْسُوب، حَسَّاب، مُحاسِب، مُحْتَسِب، مُتَحَسِّب، حَسِيب... (أسماء مشتقة) ولكلِّ من هذه الكلمات عشراتُ، بل مئاتُ الصور المختلفة التي يمكن أن ترد في النصوص.

2.2 نوع الكلمة

يفيد البحث عن طريق نوع الكلمة (فعل، اسم، مصدر، صفة...) في تضييق نطاق نتائج عمليات البحث وجَعْلِها أقربَ ما يكون إلى الاستعلام المطلوب. فإذا اقتصر الاستعلام على الأفعال مثلاً، فإن نتائج البحث تقتصر على الأفعال فقط.

مثال: إذا طلبنا البحث عن الأفعال المشتقة من الجذر (ح س ب)، حصلنا على النصوص التي ترد فيها أفعالُ هذا الجذر: المجردة والمزيدة، المبنية للمعلوم والمبنية للمجهول، المؤكّدة وغير المؤكّدة، في صيغة الماضي والمضارع والأمر، من مثل:

حَسَبَ، حَسَبْتُ، حَسَبْتُم، حُسِبَ، حَسِبَتَا، حَسِبْنَ... (أفعال ماضية مجردة) يَحْسُبُ، تَحْسُبِنَ، يَحْسُبُونَ، يَحْسُبْنَ... (أفعال مضارِعة مجردة) حاسَبَ، تَحاسَب، تَحَسَّب، احْتَسَبُوا... (أفعال ماضية مزيدة) يُحاسِبُ، يُحاسِبُون، تَحَسَّبان... (أفعال مضارِعة مزيدة) يُحاسِبُون، تَتَحَسَّبان... (أفعال مضارِعة مزيدة)

احْسُبْ، احسُبِي، احسُبُوا، احْتَسِبُوا، حاسِبُوا، تَحَسَبُوا... (أفعال أمر) حُسِبَ، حُوسِبَتْ، حُوسِبُوا، حُوسِبْنَ... (أفعال ماضية مبنية للمجهول) يُحْسَبُ، يُحاسَبُونَ، تُحاسَبِينَ، تُحاسَبْنَ... (أفعال مضارعة مبنية للمجهول)

3.2 وزن الكلمة

الغرض من إضافة وزن الكلمة هو تضييقُ نطاق البحث أكثرَ فأكثر. فإذا اقتصر البحثُ على الأفعال مثلاً، ثم حُدِّد بوزنٍ معيَّن، فإن نتائج البحث تقتصر على هذا الوزن المحدَّد فحسب.

مثال: إذا طلبنا البحث عن الوزن (فُعِلَ) من بين أوزان أفعال الجذر (ح س ب)، حصلنا على النصوص التي ورد فيها الفعل (حُسِبَ) دون غيره من الأفعال.

وهكذا نجد أن نطاق البحث يكون واسعًا جدًّا على مستوى الجذر، ثم يَضيق بتحديد نوع الكلمة، ثم يضيق أكثر فأكثر بتحديد وزن الكلمة.

3 أهم مزايا الخوارزمية المقترحة

1.3 استبعاد نتائج البحث غير المطابقة للاستعلام

من مزايا خوارزمية البحث بطريقة (الجذر، النوع، الوزن) أنها تستبعد النتائج غير المطابقة للاستعلام. فمثلاً، إذا كان البحث عن كلمة تحتمل وجهين، دوران أحدهما أكبر بكثير من دوران الآخر، فإن نتائج بحث الوجه الأول تَطغى على نتائج بحث الوجه الثاني، ونكون أمام عدد هائل من النتائج يتعذّر معه، بل يستحيل أحيانًا، الوصول إلى النصوص التي تحتوي على الوجه الآخر لكلمة البحث. وفيما يلي مثالً يوضّع هذه الحالة:

مثال: إذا بحثنا عن كلمة (وَفِيّ) في محركات البحث المعتمِدة على مطابقة الكلمة حصلنا على أعدادٍ هائلة من النتائج التي وردت فيها كلمة (وفي) بصورة: (و + في: حرف عطف + حرف جر) [وهو الوجه الأكثر دورانًا]، وعددًا قليلاً من النتائج المطلوبة (وَفِيّ: من الوفاء)، يتعذّر على الباحث استخراجها من هذا الكم الهائل من النتائج.

أما في طريقة (الجذر، النوع، الوزن)، فيجري البحث بواسطة الجذر (و ف ي)، والنوع (صفة)، والوزن (فَعِيل). ونحصُل بالنتيجة على النصوص التي تحتوي على كلمة (وَفِيّ) حصرًا دون غيرها.

2.3 عدم تشعُّب نتائج البحث

ومن مزايا خوارزمية البحث بطريقة (الجذر، النوع، الوزن) أيضًا أنها تُقيِّد نتائج البحث بالوجه المراد من كلمة البحث. ذلك أنه عند البحث عن كلمةٍ تحتمل وجوهًا متعدِّدة، فإن نتائج البحث تشتمل على مجموع هذه الوجوه دونما ترتيب، أو فصلٍ فيما بينها. ونكون هنا أيضًا أمام عددٍ هائل من النتائج يتعذَّر معه الوصولُ إلى النصوص التي تحتوي على الوجه المطلوب من كلمة البحث. وفيما يلى مثالٌ

يوضِّح هذه الحالة:

مثال: إذا بحثنا عن كلمة (فَهُم)، أعطت محركاتُ البحث المعتمِدةُ على مطابقة الكلمةِ ملايينَ النصوص التي وردت فيها كلمة (فهم) بوجوهها المختلفة؛ مثل:

ولا تستطيع هذه المحرّكاتُ التمييزَ بين هذه الوجوه المختلفة للكلمة.

أما في طريقة (الجذر، النوع، الوزن)، فيجري البحث بواسطة الجذر (ف ه م)، والنوع (مصدر)، والوزن (فَعْل). ونحصُل بالنتيجة على النصوص التي تحتوي على كلمة (فَهْم) دون غيرها.

3.3 ابتداع ضُرُوب متنوعة من البحث

ومن مزايا خوارزمية البحث بطريقة (الجذر، النوع، الوزن) أيضًا قدرتُها على البحث عن ضُروبٍ أخرى من البحث، لا تستطيع محرِّكاتُ البحث المعتمِدةُ على مطابقة الكلمة القيامَ بمعظمها. تَنتُج هذه الضُّروب من البحث عن طريق التحكُّم في أحد المكوِّنات الثلاثة: الجذر، والنوع، والوزن. وفيما يلي بعض الأمثلة:

مثال أول: إذا أجرينا بحثًا بطريقة (الجذر، النوع، الوزن) عن الجذر (ح س ب)، والنوع (اسم آلة) دون تحديد الوزن، حصلنا على النصوص التي تحتوي على أحد أسماء الآلة التي وردت من هذا الجذر؛ مثل:

حاسُوب، مِحْساب، مِحْسَب، مِحْسَبَة، حاسِبَة، حَسَّابَة...

وبهذه الطريقة لا نحتاج إلى إعادة البحث عددًا من المرات يساوي عدد الأوزان. إضافةً إلى أننا قد لا نعرف سلفًا أوزان أسماء الآلة القياسية والسماعية كي نُدْرِجَها في الاستعلام المطلوب.

مثال ثانٍ: إذا بحثنا عن الجذر (ء – –)، والنوع (فعل)، والوزن (فَعِلَ)، حصلنا على النصوص التي تحتوي على أحد الأفعال المهموزة الفاء التي تتصرَّف من باب (فَعِلَ يَفْعَلَ)؛ مثل:

أَبِقَ يَأْبَقُ، أَبِهَ يَأْبَهُ، أَذِنَ يَأْذَنُ، أَرِقَ يَأْرَقُ، أَرِمَ يَأْرَهُ، أَرِفَ يَأْرَفُ، أَرِفَ يَأْرَفُ، أَلِمَ يَأْلَمُ، أَنِفَ يَأْنَفُ، أَلِفَ يَأْلَفُ، أَلِمَ يَأْلَمُ، أَنِفَ يَأْنَفُ، أَلِفَ يَأْنَفُ، أَنِفَ يَأْنَفُ...

مثال ثالث: إذا بحثنا عن الجذر (- وء)، والنوع (فعل) دون تحديد الوزن، حصلنا على النصوص التي تحتوي على أحد الأفعال الجوفاء الواوية المهموزة اللام (المجردة والمزيدة)؛ مثل:

باءَ يَبُوءُ، ساءَ يَسُوءُ، ماءَ يَمُوءُ، ناءَ يَنُوءُ، أَساءَ يُسِيءُ، بَوَّأً يُتَبَوَّأً، بَوَّأً يَتَبَوَّأً، بَوَّأً يَتَبَوَّأً، اسْتَضاءَ يَسْتَظءَ يَسْتَظع

مثال رابع: إذا بحثنا عن الجذر (-- ي)، والنوع (اسم)، والوزن (فاعِل)، حصلنا على النصوص التي تحتوي على أحد الأسماء المنقوصة على هذا الوزن؛ مثل:

الباغِي، التالِي، الجابِي، الخالِي، الراقِي، الساقِي، الطاغِي، العاصِي، القاضِي، الماحِي، الماضِي، الناسِي، الناهِي، الهادِي...

مثال خامس: إذا بحثنا عن النوع (اسم)، والوزن (فاعُول) دون تحديد الجذر، حصلنا على النصوص التي تحتوي على أحد الأسماء التي وردت على هذا الوزن؛ مثل:

حاسُوب، ناقُوس، صارُوخ، ناسُوخ، ساطُور، ناطُور، ناقُور، جارُوف، خارُوق...

مثال سادس: إذا بحثنا عن الجذر (د ف ع)، دون تحديد النوع والوزن، حصلنا على النصوص التي تحتوي على أحد الأفعال أو الأسماء أو الصفات أو المصادر... التي وردت من هذا الجذر؛ مثل:

انْدِفاع، انْدَفَعَ، تَدافَع، تَدَافُع، تَدَفُع، دافِع، دافِع، دِفاع، دِفاعِي، دِفاعِيَّة، دَفَعَ، دُفْع، دُفْع، دُفْع، دُفْع، دُفْع، مُدافِع، مُدافِع، مُدافِع، مُدافَع، مِدْفَع، مِدْفَع، مَدْفُوع، مُنْدَفِع...

مثال سابع: إذا بحثنا عن الجذر (و - ي)، والنوع (فعل)، والوزن (افْعِلْ)، حصلنا على النصوص التي تحتوي على أحد أفعال الأمر من اللفيف المفروق؛ مثل:

قِ (من وَقَى يَقِي)، عِ (من وَعَى يَعِي)، فِ (من وَفَى يَفِي)، لِ (من وَلَى يَلِي)، شِ (من وَشَى يَشِي)...

مثال ثامن: إذا بحثنا عن النوع (جمع)، والوزن (فَعاعِيل) دون تحديد الجذر، حصلنا على النصوص التي تحتوي على أحد جموع التكسير التي وردت على هذا الوزن؛ مثل:

دَنانِير (جمع دينار)، سَكاكِين (جمع سكِّين)، دَجاجِيل (جمع دَجَّال)،

قَرارِيط (جمع قيراط)، أَبابِيل (جمع لا واحد له)...

مثال تاسع: إذا بحثنا عن الوزن (فَعْلان) دون تحديد الجذر والنوع، حصلنا على النصوص التي تحتوي على أحد الأسماء أو الصفات التي وردت على وزن هذا الوزن؛ مثل:

شَعْبان، سَعْدان، مَرْجان، حَوْران، حَسَّان، عَمَّان... (أسماء) عَطْشَان، غَضْبان، سَكْران، شُبْعان، عجلان، ملآن، ظَمْآن، نَدْمان، كَسْلان... (صفات) وغيرها كثير.

بقي أن نشير إلى نقطة مهمَّة وهي أن البحثَ بطريقة (الجذر، النوع، الوزن) يُعَدُّ مُكمَّلاً للبحث بطريقة مطابقة الكلمة وتطويرًا له. وبذلك نستفيد من مزايا البحث العادي والبحث المتقدم معًا للحصول على نتائج بحثٍ فعَّالة، كما هو موضَّح في المثال الآتي:

إذا أردنا البحث عن كلمة (مَلِك) بجميع صورها، فإننا نطلب البحث عن الجذر (م ل ك)، والنوع (اسم)، والوزن (فَعِل)، كما مرَّ آنفًا. ونختار عدم مطابقة الكلمة. فنحصُل على النصوص التي تحتوي على كلمة (مَلِك)، وكذلك على النصوص التي تحتوي على الكلمات التي تشترك معها في أحرفها وحركاتما؛ مثل:

وَمَلِك، الْمَلِك، والْمَلِك، فالْمَلِك، للْمَلِك، كالْمَلِك... (الكلمة مع سوابقها) مَلِكُهُ، مَلِكُها، مَلِكُنا، مَلِكُهُم، مَلِكِي، مَلِكُهُما... (الكلمة مع لواحقها) وَمَلِكُهُ، فَمَلِكُها، بمَلِكِها، كَمَلِكِنا، ومَلِكُهُما... (الكلمة مع سوابقها ولواحقها)

أما إذا أردنا البحث عن كلمة (مَلِك) بمذه الصورة فقط، فإننا نطلب البحث عن الجذر (م ل ك)، والنوع (اسم)، والوزن (فَعِل). ونختار مطابقة الكلمة. فنحصل على النصوص التي تحتوي على كلمة (مَلِك) فقط.

المراجع

- "محركات البحث في النصوص العربية وصفحات الإنترنت"، مروان البواب، مجلة الحياة الفكرية، العدد 2، 2009.
- "أسلوب معالجة اللغة العربية في المعلوماتية (الكلمة-الجملة)"، مروان البواب ومحمد حسان الطيان، كتاب "استخدام اللغة العربية في المعلوماتية"، إصدار المنظمة العربية للتربية والثقافة والعلوم، تونس، 1996.