

تطبيق الطرق الإحصائية في تحليل محتوى النصوص بلغتها العربية آلياً

مجلة العربية فصلية محكمة متخصصة في مجال المعلومات ، العام ٢٠٠٢ العدد الأول من النادي العربي
للمعلومات

د. عامرة حقي الفرغولي

تتضمن الدراسة التجريبية هذه عرضاً لتجربة اعتمدت تطبيق ثلاث طرق إحصائية لتحليل محتوى نصوص باللغة العربية، إضافة إلى المعالجة الحرفية. بهدف البرهنة على أن المعادلات الرياضية التي استخدمت على نصوص باللغة الإنكليزية يمكن استخدامها على نصوص باللغة العربية. أظهرت الدراسة أن الطريقة الإحصائية إضافة إلى الطريقة اللغوية في تحليل المحتوى (التكشيف) الآلي تفرز واصفات دالة على الموضوع الذي تمثله.

مقدمة

تحليل المحتوى: مصطلح ينضوي تحت مصطلح التكشيف، وما الخزن والاسترجاع إلا عمليتا التكشيف والبحث، ترتبط بتحليل المحتوى طريقتان:

(1) الطريقة الإحصائية

(2) الطريقة اللغوية (تحليل) إعرابي syntactical وتحليل دلالي (Symantical) والتعامل مع اللغة العربية الطبيعية آلياً فيه إشكالات وصعوبات، منها عامة مثل المرونة وتعدد وتفاوت المفردات وأخرى تتعلق بإشكالات صرفية أو نحوية وأهمية التقيد بالشفرة العربية الموحدة في تمييز أشكال الألف ومواقع الهمزة. إلخ، فقد تم تطوير برامج تشكيل الكلمات العربية تلقائياً في سياقاتها وفق دلالتها وتركيبها اللغوي، حيث اعتمد برنامج المحلل الصرفي في برنامج القرآن الكريم على المبدأ اللغوي وهو أساس بنية الكلمة العربية (فعل) لكونها الجذر، والاسترجاع يعتمد على المحلل الصرفي في ربط الكلمات المختلفة الصيغ بجذورها. وفي هذه الدراسة تم استخدام الأساليب الإحصائية إضافة إلى الأسلوب اللغوي في تحليل ومعالجة عناوين ومستخلصات (آلياً) لغرض الاسترجاع وتحديد واصفات تمثل محتواها، فقد تم إعداد البرامج اللازمة وعددها (١٥) برنامجاً، وجاءت النتائج مرضية لتثبت إمكانية تطبيق الأساليب الإحصائية التي طُبِّقها/ لون وارتائدي وسالتون/ في مجالات التكشيف والخزن والاسترجاع على نصوص باللغة الإنكليزية.

المشكلة

يحتاج الباحثون على اختلاف مستوياتهم إلى المعلومات ذات الصلة بموضوعات أبحاثهم باللغة العربية، وإحدى نقاط الوصول إلى تلك المعلومات هو الموضوع، حيث تتبع في مكتبتنا ومراكز المعلومات طرق تقليدية للوصول إلى المعلومات، أو طرق آلية يتم تحليل المحتوى فيها باللغة المقيدة، دون محاولة تطبيق الأساليب الإحصائية المطبقة من قبل متخصصي المعلومات منذ زمن بعيد في تحليل محتوى واسترجاع النصوص باللغة الإنكليزية.

الهدف

- 1- تجربة تحليل المحتوى والاسترجاع آلياً لنصوص علمية باللغة العربية.
- 2- الوصول إلى الطريقة الأفضل: التحليل اليدوي باللغة المقيدة أم الآلي باللغة الطبيعية.

الأهمية

تقديم نماذج جديدة في تحليل المحتوى باللغة العربية.
متخصص المعلومات هو من يؤدي هذا العمل لارتباطه التام باختصاصه.

الفرضيات

- 1- توليد واصفات مزدوجة آلياً يؤكد أهميتها.
- 2- استخدام اللغة الطبيعية آلياً يحسن الاسترجاع.

التجربة

أجريت التجربة آلياً على ٢٥٣ مستخلصاً مع عناوينها لتشكل نظاماً متكاملأ بعد تطبيق برامج القراءة وبرامج المعالجة (انظر الشكل ١) وفيها تم معالجة النصوص لغوياً وإحصائياً بتطبيق المعادلات الثلاث وتوليد ملف المعلومات الببليوغرافية لكل وثيقة، وملف بالواصفات مرتبة هجائياً مع أرقام الوثائق التي تشتمل عليها برامج تخص الاسترجاع.

البرامج

يتألف النظام من برامج فرعية بحدود ١٥ برنامجاً انطوت تحت برنامج رئيس واحد كوحدات (Units) وقد تم

2 - Borko, Harold and Charles L. Bernier. Indexing Concepts and Methods.-New York: Academic press, 1978.

3 - Hassan, A. M. H. Design and Implementation of a Hypertext Documents Structuring System. - (Master Thesis), University of Technology, 1994. 97p.

4 - Keen, E. M. and Hartly R.J. Phrase Processing in Text Retrieval Systems. - J. of Documents, 1994, 50(1), p.23.

5 - Salton, G. et al. The Theory of Term Importance in Automatic Text Analysis. - JASIS: 26 (1). Jan. Feb. 1975. pp. 33-44.

6- القواسمة، عبد الله. استخدام اللغة الطبيعية في استرجاع المعلومات ومشكلات ذلك في اللغة العربية. رسالة المكتبة. مج 37، ع 3، 1991. ص 53 - 93.

7- نزار محمد علي قاسم. الاشتقاق والمشتقات العربية وأثرها في تخزين المعلومات واسترجاعها رسالة المكتبة. مج 92، ع 1، 1991. ص 4 - 61.

شبكة اللغويات العربية