بسم الله الرحمن الرحيم

"قُل لَوْ كَانَ البحرُ مداداً لكلماتِ ربّي لنفدَ البحرُ قبلَ أن تنفدَ كلماتُ ربّي ولو جئنا بمثلهِ مدداً"

صدق الله العظيم

ملخـــــص

بدأت فكرة استرجاع المعلومات (Information Retrieval) خلال التحول الذي شهدناه في طرق تخزين المعلومات الثقافية والاجتماعية والعلمية من الأوراق والكتب إلى مكتبات رقمية منتشرة على الشبكات العالمية، وتأتي هذه المعلومات على شكل نصوص أو وثائق مصورة (كالصور والخرائط والرسومات التقنية وغيرها) أو صوت أو صور متحركة (الفيديو). في محاولة لجعل استرجاع هذه الكمية الهائلة من المعلومات بفاعلية كان لا بد من طريقة لاستنباط الكلمات المفتاحية آلياً (Automatic Term Extraction). هذه الطريقة الآلية تشكل حجر الأساس للعديد من التطبيقات ومنها محركات البحث، لأنها عملية شاقة ومكلفة أن يتم استنباط الكلمات المفتاحية يدوياً. وللتأكد من دقة المعلومات المسترجعة فإن الكلمات المفتاحية المستنبطة يجب أن تصف محتوى الوثائق المستنبطة منها وصفاً دقيقاً.

تقدم الباحثون في هذا الموضوع بالعديد من المقترحات والطرق المختلفة لاستنباط الكلمات المفتاحية، بعضها يعتمد على الإحصاء والاحتمالات وبعضها الآخر يعتمد على التحليل اللغوي وغيرها، تأتي هذه الأطروحة كمساهمة في كشف أوجه الاختلاف والتشابه وكذلك محاسن ومساوئ الطرق الإحصائية المستخدمة وهي: تكرار الكلمات في الوثيقة الواحدة (Term Frequency TF)، تكرار الوثائق المحتوية على كلمة معينة ضمن مجموعة من الوثائق (Inverse Document Frequency IDF)، الدمج بين تكرار الكلمات وتكرار الوثائق (TFxIDF)، والطريقة الرابعة هي نموذج قيم التمييز للكلمات (Term Discrimination Value Model) من خلال دراسة مقارنة بين هذه الطرق

الإحصائية وتطوير أداة محوسبة لاستنباط الكلمات المفتاحية آلياً (ATEWB) تستخدم للمقارنة من خلال ثلاث تجارب تقييمية تبرز العوامل المؤثرة على دقة النتائج وذلك لتحديد الظروف المناسبة لاستخدام تلك الطرق.

من ناحية أخرى، تهدف الدراسة إلى زيادة فاعلية الطرق الإحصائية وتسريعها من خلال استخدام محركات قواعد البيانات (Database Engines) والتي تختصر العمليات الحسابية في خوارزميات هذه الطرق بشكل ملحوظ، كما استخدمت في تسريع استرجاع الوثائق عن طريق تخزين نسخة منها في قاعدة البيانات.

الوثائق المستخدمة في التجربة الأولى هي مجموعة من الملخصات (Abstracts) لأبحاث في مجال الاستنباط الآلي للكلمات المفتاحية، والكلمات المفتاحية لها مستنبطة يدوياً لمقارنتها بالكلمات المفتاحية المستنبطة آلياً، أما تلك المستخدمة في التجربة الثانية والثالثة فهي وثائق أعدت خصيصاً لغرض البحث، وهي متوفرة على بعض مواقع الإنترنت المهتمة بموضوع الاستنباط الآلي للكلمات المفتاحية.