Automatic Generation of Index Phrases for Document Retrieval Using Semantic and Syntactic Information

Martin van Bommel

Ernst J. Schuegraf

Department of Mathematics and Computing Sciences
St. Francis Xavier University
Antigonish, Nova Scotia. B2G 1C0

The aim of most present–day retrieval systems is to satisfy users' needs, by providing them with as many relevant documents as possible, while keeping the number of irrelevant ones down. Maximization of the standard measures of precision and recall is the goal. One way to improve precision and recall is to add keywords or phrases to the document. These keywords or phrases can be assigned manually by trained indexers or they can be generated automatically. Problems arising from manual indexing are well known: consistency of indexers, use of controlled versus uncontrolled vocabulary. Cost of manual indexing is very high and time consuming and a method of automatic indexing is to be preferred if the indexations are relevant and meaningful.

Keywords or phrases may also be generated automatically, a process that is based mainly on the use of word frequencies and term weights. It is well known that the best index terms are those that occur with approximately equal frequencies in the collection. The Zipfian distribution of terms limits the number of useful index terms but index phrases can extend this number.

Index phrases are normally generated using word frequencies, term weights and proximity information, but are often not very meaningful or descriptive. Semantic and syntactic information can be used to generate more concise index phrases.

A simple experimental indexing scheme that uses some semantic and syntactic information for the construction of index phrases is described. Results from experiments with the INSPEC data base are presented and some examples are given. A possible extension of this approach linking the derived indexations with a thesaurus is also mentioned.

Utilisation de l'information sémantique et syntaxique pour la génération automatique de syntagmes en recherche documentaire

Martin van Bommel
Ernst J. Schuegraf
Department of Mathematics and Computing Sciences
St. Francis Xavier University
Antigonish, NS

La plupart des systèmes de recherche documentaire présentement disponibles visent à satisfaire les besoins de leurs usagers en leur fournissant un nombre maximum de documents pertinents tout en réduisant au minimum le nombre de documents non pertinents. On cherche à maximiser les normes de précision et de rappel. Une méthode pour améliorer la précision et le rappel consiste à ajouter des expressions et des mots clef au document. Ces expressions et ces mots clef peuvent être attribués manuellement par des indexeurs, ou ils peuvent être générés automatiquement. Les problèmes liés à l'indexation manuelle sont bien connus: cohérence parmi les indexeurs et utilisation d'un vocabulaire dirigé par opposition à un vocabulaire non dirigé. Les coût d'indexation, en temps et en argent, sont très élevés, ce qui rend une méthode d'indexation automatique préférable, en autant que les termes générés soient pertinents et porteurs de signification.

Les expressions et les mots clef peuvent également être générés automatiquement. Cette opération se base avant tout sur la fréquence et la pondération des mots. Le fait que les termes d'index les plus utiles soient ceux apparaissant à fréquence approximativement égale dans la collection est bien connu. La loi de Zipf portant sur la distribution des termes limite le nombre de termes d'index utiles, mais l'utilisation d'expressions permet de l'accroître.

Des syntagmes sont normalement générés sur la base de la fréquence, de la pondération et des relations de voisinage des mots, mais il arrive souvent qu'ils ne soient très descriptifs ou significatifs. L'information sémantique et syntactique se révèle alors utile pour la génération de syntagmes plus concis.

Un procédé d'indexation simple faisant appel à de l'information sémantique et syntactique pour construire un index d'expressions est décrit. Les résultats d'essais menés sur la base de données INSPEC sont présentés, accompagnés de quelques examples. Un prolongement à cette méthode, qui relierait l'index ainsi généré à un thesaurus, est également mentionné.