20RM-2008-02
Method and System for Automatic Keyword Extraction

The present invention provides a method, a computer program, an apparatus, and a system for automatically generating one or more keywords from an electronic document. Examples of collections of keywords used for finding useful information include a back-of-the-book index for a book-length document, and keywords used as annotation links within electronic encyclopedias. The invention provides a method of automatically processing electronic documents in electronic form, in order to extract useful collections of keywords, which achieves a goal of more closely approaching a quality of output like that generated by human authors and/or professional indexers. Both unsupervised and supervised methods of automatic keyword extraction algorithms are provided, each with advantages in speed and performance. Novel features for use by machine learning algorithms in keyword extraction are introduced, that also have further applications in other areas of natural language processing using computer data processing systems. By combining keyword extraction with word sense disambiguation, a system for automatically annotating electronic documents with links to related information in an electronic encyclopedia is also described that can be used to enrich text for educational and other purposes.

For Additional Information, Please Contact:

The University of North Texas
Office of the Vice President for Research
and Economic Development
3940 North Elm, A160
Denton, TX 76207
Fax: 940-565-2944
Email: richard.croley@unt.edu