Wikipedia ページへの tfidf 法の適用

宮崎将隆、川端豪

本報告では tfidf 法に基づく話題キーワード選択法の改良を行う。ブログなどの限定された少数ページから tfidf を計算しようとすると、その基となる tf 及び idf の値が精度良く求められない。まず、idf については Web ページ全体から算出した idf で Wikipedia から算出した idf を近似できることが分かった。次に、tf については単語共起に基づくクラスタリング手法を導入し、キーワードのグループを構成した。少数ページから tf の計数を行う際に、グループに含まれるすべての単語の計算値の総和で代用する。実験によって、このようにして求めたグループ tf が真の tf と強い相関を持つことを確認した。

A Study of “tfidf” Measure using Wikipedia Statistics

Masataka Miyazaki and Takeshi Kawabata

This paper describes an improvement of the keyword selection criteria based on the “tfidf” measure. It is very difficult to estimate “tf (term frequency)” and “idf (inverse document frequency)” values from small amount of weblog pages. First, we investigate an approximation of the world wide idf value as the Wikipedia idf value. Experiments show that this idf approximation is promising. Secondly, we apply the clustering method to word co-occurrence and make several word groups. The tf value of a keyword is extrapolated as the sum of its group word frequency. Experiments show that the group-word based tf values counted in small amount of pages are strongly correlated to the true tf values.