要 旨

単語出現頻度を用いたブログのフィルタリングと学習による適合記事検索手法

中谷 浩輝

本研究では, 大量のブログ記事から読者にとって有用な記事を選別・抽出する手法を提案する．ブログの普及により，多くの有用な情報がブログの形でウェブ上に蓄積されているが，ブログ記事は通常のウェブ文書と異なり，より手軽に情報の発信ができるため，有用なものだけでなく読者にとってあまり有用でない情報も数多く存在する．そこで本研究では，読者にとってより有用と感じられる記事を検索するために，単語の出現頻度などをスコア化する TF-IDF を用い，そのスコアのヒストグラムを特徴として利用することで，記事の有用性の判断を行う．ブログの記事の収集には， Hyper Estraier を使用し，芸能，IT 関連用語，経済などの多数のジャンルから取得を行う．取得した記事に対して，本研究では Yahoo! Japan が提供しているテキスト解析用の API を使用して，ブログ記事に対して形態素解析を行う．その後，単語ベクトル，単語ヒストグラムを用いて，ブログ記事 100 件の有用性をユーザからの教師信号を用いて学習をし，別の 100 件にて正答率を確認した．結果として，単語ヒストグラムを用いた場合の方が正答率が下がる結果となったが，不用なものを有用とする再現率を高める効果があることを確認した．今後，記事や単語数を多くした場合に，この再現率がどのような働きをするのか調べる必要がある．

キーワード　ブログ， TF-IDF， ニューラルネットワーク

Abstract

Filtering of using term frequency and Retrieving Relevant Blog Entries

NAKATANI Kouki

In this thesis, we propose a technique for the selection and extracting a useful blog entries for readers from a large amount of blog entries. The blog widely spread to the Internet, and a lot of useful information has been accumulated on the web in the form of the blog. However, not only the useful blog entries but also useless entries exist because the blog entries can send information more easily unlike a usual web document, then many casual writers upload their articles. TF-IDF, which express the occurrence rate etc. of the word, is used in this research to retrieve the useful article for the reader. The utility of the entries is judged by machine learning using the characterization of word histogram of the TF-IDF score. Many genres, for example entertainments, IT related terminologies, and economic, etc. are gathered to experiments by using Hyper Estraier for the collection of entries on the blog. The morphological analysis is performed to the blog entries by Yahoo! Japan API for the offered text analysis for the acquired entries. Afterwards, the utility of 100 blog entries is for used for learning by the neural-network by the word vector and the word histogram, and other 100 correct answer rates were examined. Consequently, the result of shows the word histogram the correct answer rate decreases. There was an effect of improving the recall ratio in which the useless one was assumed to be useful because most changes were not seen as for uselessness and the probability of mis-classifying. It will be necessary to examine what working this

recall ratio does when the article and the number of words are increased in the future.

keywords weblog, TF-IDF, neural-network