要 旨

ラフ集合を用いたスパムブログ記事の特徴抽出と判別

中谷 浩輝

近年ブログの普及に伴い，ブログを対象としたサービスの必要性が高まっている．その中の一つとしてブログ検索のサービスが増加してきている．しかし，ブログの普及とともにスパムブログも増加してきており，検索結果にスパムブログが含まれてしまい効率的な検索の妨げとなっている．本研究では，検索結果からスパムブログを排除するために，非数値データからも知識獲得が可能なラフ集合を用いてスパムブログの特徴抽出と判別を行う．ブログ記事は，異なる 3 つのジャンルからキーワードを用いて検索を行い，検索結果の上位 100 件のブログ記事の本文のみを使用する．キーワードは,「東証」「リア・ディゾン」「秋華賞」の 3 つを使用する．ラフ集合の解析に必要な属性としては，句読点の数，文字数，行数などの 21 の特徴を用いる．これらの特徴から解析を行った結果，識別のための決定ルールには 21 属性から 10 以上の特徴を削減できることを確認した．また解析を行った結果，誤判定率が 30% 〜 50% と高くなった．今回は，ラフ集合の解析を粗く行ったため，誤判定率が高い結果となったと考えられる．

キーワード　ラフ集合

Abstract

Feature extraction of a SPAM blog report and distinction using a rough set

Koki NAKATANI

The need of the additional service for blog is increasing with the spread of blogs in recent years. Service of blog search is increasing as one of them. However, the SPAM blog is also increasing with the spread of blogs, SPAM blogs contained in search results, and it has become the disturbance of efficient search. In this thesis, in order to remove a SPAM blog from search results, the rough set, in which knowledge acquisition is possible is used even from the data whith take nonnumeric values. A blog article searches using a keyword from three different genres, and uses only the text of top 100 blog reports of search results. A keyword uses three, the "Tosyo"(Tokyo Stock Exchange), "Ria DIZON"(Japanese fashion model and singer), and "Syukasyo(Japanese house racing)". The features of 21, such as the number of punctuations, the number of characters, and the number of lines, are used for an attribute required for the analysis of a rough set. As a result of analysis from these features, ten or more features of the discriminant rule were reducible from 21 attributes. Moreover, as a result of analyzing, the misjudgment fixed rate became high with 30% $\sim$ 50%. Since the rough set was analyzed coarsely this time, it is thought that a result with a high misjudgment rate was brought.

key words rough set