博 士（情報科学） 王 広為

学位論文題名

A Study on Weblog Reputation Information Analysis Using Text Mining

（テキストマイニングを用いたWeblog評判情報分析に関する研究）

学位論文内容の要旨

Recent years have seen rising interest in various language processing methods, text analysis methods, and ways to acquire information from text data, much due to the rise in the Internet and broadband usage. Notably, many large text sources of reputation type information recently exist, for instance consumer opinions collected in the form of free text answers through marketing research or on Web bulletin boards, call centers and so on. This reputation information is made up of an enormous amount of unstructured and/or semi-structured data which is very hard to manage and process in a short amount of time. As a result, the problem of "textual information analysis" has seen increasing attention over the last five years.

Text mining is a technology for analyzing these large amounts of text data from various viewpoints, to discover information, features, tendencies, correlations, "hidden information" and so on; to dig up information valuable to the business. This is especially important in the field of CRM (Customer Relationship Management), which tries to improve customer satisfaction by marketing. CRM grows rapidly at a growth rate of 20 percent per year. Opinion mining (OM) is a subdiscipline of text mining which tries to detect the opinions expressed in natural language texts. It aims to analyze the reputations of various products automatically by classifying them into positive or negative meaning from the "subjective" terms contained in a document. It can play an important role in providing a speedy search and analysis result to both consumers and manufacturers. For example, businesses always want to find public or consumer opinions regarding their products and services. Potential customers also want to know the opinions of existing users before they use a service or purchase a product. Moreover, opinion mining can also provide valuable information for placing advertisements in Web pages.

This dissertation presents our study on a Weblog reputation information analysis system using text mining technology. In this system, we propose an original opinion classification method that uses both supervised and unsupervised approaches to do Japanese opinion mining. Especially, our proposed approach effectively adapted SO-PMI (Semantic Orientation Using Pointwise Mutual Information) algorithm for Japanese. SO-PMI is an unsupervised approach proposed by Turney that has been shown to work well for English. When this algorithm was translated into Japanese naively, most phrases, whether positive or negative in meaning, received a negative SO. For dealing with this slanting phenomenon, we propose three original methods: to expand the reference words to sets of words, to

introduce a balancing factor and to detect neutral expressions. In our experiments, the proposed methods obtained a well-balanced result: both positive and negative accuracy exceeded 62 percent (the best accuracy: 69 percent), when evaluated on 1,200 opinion sentences sampled from three different domains (reviews of Electronic Products, Cars and Travels from Kakaku.com). In a comparative experiment on the same corpus, a supervised approach (SA-Demo) achieved a very similar accuracy to our method. This shows the validity of our proposed approach and the generality of SO-PMI.

Our proposed system also has another important function which is a novel graphic reputation analysis system for Japanese. Most related work only focuses on the information analysis step, and does not continue on to the step that makes the analysis results visible. However, in todays continuing growth of personal opinions on the Web, there is a great need for visualizing tools to help users master and browse the information more easily. Therefore, the proposed system presents an intuitive GUI using visual mining graphs aimed at inexperienced users who want to check opinions on the Weblog before purchasing something. The provided graphs can help a user to make a quick decision on which product is suitable with regards to some special feature (for example, design, price, battery and so on).

In the early days of the World Wide Web, web pages were predominantly written in English. As the web has become an indispensable resource to find information, nowadays many web documents can be found in other languages, for instance in Asian language such as Chinese, Japanese or Korean. Since most commonly used web search and mining algorithms were originally developed for English web documents, it is necessary to fine-tune or modify these algorithms for other languages. Most methods cannot be straight forwardly used on languages that differ a lot from English, for instance by having no white space between words, making heavy use of inflections, and making use of several alphabets simultaneously and so on. Data mining, artificial intelligence and information retrieval techniques used in Web mining must also consider the linguistic aspects, national culture, and the business practices of the respective countries.

In our studies, we found that it is not obvious that a method that works for English will be easy to use on other languages, especially when the languages differ a lot, like English and Japanese. For example, in Japanese there is no space between words, and there are several different alphabets that can be used to write the same word. We have examined what steps are useful to make SO-PMI work for Japanese. It not only shows that the SO-PMI algorithm or the proposed system can also work for other languages besides English (69 percent accuracy on Japanese Weblog of electronic products from our study; 74 percent accuracy on English data from Turney's study), but the suggested improvements are also likely useful for other Asian languages which have similar characteristics (for instance no space between words).

In this dissertation, we not only give detailed explanations of each step of the proposed algorithm and each process of the system, but also show the effectiveness of the proposed method and the consistency of the system in evaluation experiments. With improved analysis precision and an intuitive interface using visual mining graphs, we confirm that this kind of system can be helpful to users and become a very strong tool to help in making purchasing decisions.

学位論文審査の要旨

主　査　教　授　荒　木　健　治
副　査　教　授　山　本　　強
副　査　教　授　長谷川　美　紀

学位論文題名

A Study on Weblog Reputation Information Analysis Using Text Mining

(テキストマイニングを用いたWeblog評判情報分析に関する研究)

本論文は、教師あり学習と教師なし学習手法を融合した独創性の高い日本語評判情報分析手法を提案し、直感的な視覚インタフェース用いた商品購買の意思決定支援システムを構築したものである。

近年インターネットがブロードバンド化し、コンピュータの処理能力や言語情報処理分野の諸研究が進展したことから、テキスト型データの取得方法や解析手法への関心が高まっている。また、市場調査の自由回答型データ、Web掲示板の発言、企業内の営業報告書、コールセンターで収集した消費者の意見など、多種多様かつ膨大なテキスト型評判情報が存在している。

テキストマイニングはこれら膨大なテキストデータを様々な観点から分析し、「隠れた」情報や特徴、傾向、相関関係などを発見し、ビジネスに有効な価値ある情報を掘り起こすための技術である。その目的はマーケティングによって顧客の満足を向上させようとするCRM(Customer Relationship Management)の分野において特に重要となっている。CRMは年間20%の伸び率で急成長している。テキストマイニングとは人工知能技術と自然言語処理技術を融合した技術であり、ここ数年で実用化が進展し、現在も急速に発展しつつある新しい技術分野である。この技術の最も重要な適用分野の一つに「顧客の声」の分析がある。これはコールセンター及びWebの掲示板に寄せられた膨大な問い合わせや不具合情報を分析し、主要な要望・不満を把握することで、業務改善や経営戦略の立案に役立てようとするものである。テキストマイニングは「大量の文書情報の中から新たな知見を迅速に抽出する作業を支援する」という点から、個人レベルから企業活動全体に至る品質と効率の向上に寄与することを最終目的としている。

本論文は、テキストマイニングを用いたWeblog評判情報分析に関する研究について述べたものである。著者は日本語の意見分析を行うために、教師あり学習と教師なし学習の各々の利点を十分に利用した上で融合する独自の意見分類手法の提案を行った。意見フレーズは日本語の文法構造に基づいて属性部と評判部から構成されている。提案手法では、1つの意見フレーズを属性表現と評判表現に分離させた後、属性表現を教師あり学習によって分類し、評判表現を教師なし学習によって分類する。

教師あり学習を用いた手法は通常トレーニングコーパスを利用して分類器に学習させているた

め、コーパス作成の手間をかけると比較的高い精度が得られるという特徴を有する。一方、教師なし学習手法はデータの背後に存在する本質的な構造を抽出するために用いられる。この手法はコーパスが不要のためデータ作成のコストがかからないと精度が低くなってしまう。調査によると、属性表現は類似表現が多く表現の数が限られているため、コーパスの作成が容易であるので、教師あり学習の手法 (Naive Bayes) を用い属性表現を価格、デザイン、機能、バッテリーなどに分類することができる。また、教師なし学習の手法 (SO-PMI:Semantic Orientation Using Pointwise Mutual Information) は単語の意味オリエンテーション (肯定/否定) を分析できることが他の関連研究で報告されているので、この手法を用いて日本語の評判表現を肯定的または否定的な意味に分類する手法の開発を行った。

SO-PMI は Turney によって提案された教師なし学習の手法で、すでに英語の意味分析に有効であることが報告されている。このアルゴリズムを単純に日本語に用いた場合、肯定の意味を持つフレーズも否定の意味を持つフレーズもほとんどのフレーズが否定的な意味と判断されてしまう。この問題に対処するために、著者は独自の 3 つの手法の提案を行った。その 3 つの方法とは、参照単語の単語セットへの拡張、バランスファクターの導入、中立表現の検出である。実験により、提案手法は肯定的な精度と否定的な精度ともに 62% を超えるというバランスの良い結果が得られた。評価対象は 3 つの異なったドメイン (価格.com より電子商品、車、および旅行に関するレビュー) である。これらのドメインから 1,200 意見文の抽出を行った。同じコーパスを用いた比較実験では、教師あり学習の手法を利用しているシステムとほぼ同程度の精度を達成している。この比較実験の結果はコストの面を考えると教師なし学習を導入した提案手法は有効であり、また SO-PMI アルゴリズムの固有の言語によらない汎用性も示していると考えられる。

また、本研究では完全な日本語評判分析システムの構築を行った。本システムは属性ごとに視覚マイニンググラフで分析結果を表示できるという重要な機能がある。ほとんどの関連研究は情報分析にのみ焦点をあて、分析結果をわかりやすく表示するというステップまで進んでいない。しかし、現在ウェブ上の個人的な意見が猛スピードで増大している。このような情報洪水時代には、ユーザがより容易に各種情報を獲得できること及びわかりやすい視覚的分析結果表示機能を提供できるツールが必須となる。したがって、著者は何かを購入する前に Weblog で関連する意見をチェックしようとしている一般的なユーザを対象に視覚マイニンググラフで直感的な GUI を提供しているシステムの提案を行った。表示されたグラフは、ユーザがデザイン重視、価格重視、バッテリー重視などの各自の好みでどの製品が適しているかの購買意思決定を支援することができる。

本論文では、提案されたアルゴリズムとシステムのそれぞれの過程を詳述し、評価実験によって提案手法と実験システムの有効性を示した。今後、より一層分析精度の向上を行う予定である。また、本システムは視覚マイニンググラフを用いた直感的なインタフェースも有している。このような評判情報分析システムは一般ユーザが利用する有効な購買意思決定支援ツールとなるものと考えられる。

以上を要約すると、著者は日本語の Weblog 評判情報の分析とその結果を提示するための画期的な手法とシステムを提案し、その性能評価実験により提案手法の有効性の確認を行った。本研究により、テキストマイニング技術を用いた Weblog の評判情報の活用技術の向上に貢献するところ大なるものがある。よって、著者は北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。