Advances in Intelligent Systems and Computing 271

Keon Myung Lee
Seung-Jong Park
Jee-Hyong Lee *Editors*

Soft Computing in Big Data Processing

Springer

Advances in Intelligent Systems and Computing

Volume 271

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/11156

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

Keon Myung Lee · Seung-Jong Park
Jee-Hyong Lee
Editors

Soft Computing in Big Data Processing

Springer

Editors
Keon Myung Lee
Chungbuk National University
Cheongju
Korea

Jee-Hyong Lee
Sungkyunkwan University
Gyeonggi-do
Korea

Seung-Jong Park
Louisiana State University
Louisiana
USA

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-319-05526-8 ISBN 978-3-319-05527-5 (eBook)
DOI 10.1007/978-3-319-05527-5
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014933211

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Big data is an essential key to build a smart world as a meaning of the streaming, continuous integration of large volume and high velocity data covering from all sources to final destinations. The big data range from data mining, data analysis and decision making, by drawing statistical rules and mathematical patterns through systematical or automatically reasoning. The big data helps serve our life better, clarify our future and deliver greater value. We can discover how to capture and analyze data. Readers will be guided to processing system integrity and implementing intelligent systems. With intelligent systems, we deal with the fundamental data management and visualization challenges in effective management of dynamic and large-scale data, and efficient processing of real-time and spatio-temporal data. Advanced intelligent systems have led to managing the data monitoring, data processing and decision-making in realistic and effective way. Considering a big size of data, variety of data and frequent changes of data, the intelligent systems basically challenge new data management tasks for integration, visualization, querying and analysis. Connected with powerful data analysis, the intelligent systems will provide a paradigm shift from conventional store and process systems. This book focuses on taking a full advantage of big data and intelligent systems processing. It consists of 11 contributions that feature extraction of minority opinion, method for reusing an application, assessment of scientific and innovative projects, multi-voxel pattern analysis, exploiting No-SQL DB, materialized view, TF-IDF criterion, latent Dirichlet allocation, technology forecasting, small world network, and classification & regression tree structure. This edition is published in original, peer reviewed contributions covering from initial design to final prototypes and authorization.

To help readers understand articles, we describe the short introduction of each article as follows;

1. “Extraction of Minority Opinion Based on Peculiarity in a Semantic Space Constructed of Free Writing”: This article proposes a method for extracting minority opinions from a huge quantity of text data taken from free writing in user reviews of products and services. The extraction becomes outliers in a low-dimensional semantic space. Peculiarity Factor (PF) enables them to extract minority opinions for outlier detection.

2. “System and Its Method for Reusing an Application”: This article describes the system that can reuse the pre-registered applications in the open USN (Ubiquitous Sen-

sor Network) service platform. Cost and time can be reduced for implementing a duplicated application.

3. “Assessment of Scientific and Innovative Projects: Mathematical Methods and Models for Management Decisions”: This article is designed to improve the quality of assessment of scientific and innovative projects the mathematical methods and models. This methodology is applied to expert commissions who are responsible for venture funds, development institutes, and other potential investors required selecting appropriate scientific and innovative projects.

4. “Multi-voxel pattern analysis of fMRI based on deep learning methods”: This paper describes constructing a decoding process for fMRI data based on Multi-Voxel Pattern Analysis (MVPA) using deep learning method for online training process. The constructed process with Deep Brief Network (DBN) extracts the feature for classification on each ROI of input fMRI data.

5. “Exploiting No-SQL DB for Implementing Lifelog Mashup Platform”: To support effcient integration of heterogeneous lifelog service, this article states exploiting the lifelog mashup platform with the document-oriented No-SQL database MongoDB for the LLCDM repository. It develops an application of retrieving Twitter’s posts involving URLs.

6. “Using Materialized View as a Service of Scallop4SC for Smart City Application Services”: This paper proposes materialized view to be as service (MVaaS). A developer of an application can efficiently and dynamically use large-scale data from smart city by describing simple data specification without considering distributed processes and materialized views. It designs an architecture of MVaaS using MapReduce on Hadoop and HBase KVS.

7. “Noise Removal Using TF-IDF criterion for Extracting Patent Keyword”: This article proposes a new criteria for removing noises more effectively and visualizing the resulting keywords derived from patent data using social network analysis (SNA). It can quantitatively analyze patent data using text mining with TF-IDF used as weights and classify keywords and noises by using TF-IDF weighting.

8. “Technology Analysis from Patent Data using Latent Dirichlet Allocation”: This paper discusses how to apply latent Dirichlet allocation, a topic model, in a trend analysis methodology that exploits patent information. To perform this study, they use text mining for converting unstructured patent documents into structured data, and the term frequency-inverse document frequency (tfidf) value in the feature selection process.

9. “A Novel Method for Technology Forecasting Based on Patent Documents”: This paper proposes a quantitative emerging technology forecasting model. The contributors apply patent data with substantial technology information to a quantitative analysis. They can derive a Patent–Keyword matrix using text mining that leads to reducing its dimensionality and deriving a Patent–Principal Component matrix. It makes a group of the patents altogether based on their technology similarities using the K-medoids algorithm.

10. “A Small World Network for Technological Relationship in Patent Analysis”: Small world network consists of nodes and edges. Nodes are connected to small steps of edges. This article describes the technologies relationship based on small world network such as the human connection.

11. “Key IPC Codes Extraction Using Classification and Regression Tree Structure”: This article proposes a method to extract key IPC codes representing entire patents by using classification tree structure. To verify the proposed model to improve performance, a case study demonstrates patent data retrieved from patent databases in the world.

We would appreciate it if readers could get useful information from the articles and contribute to creating innovative and novel concept or theory. Thank you.

Keon Myung Lee
Seung-Jong Park
Jee-Hyong Lee

Contents

Extraction of Minority Opinion Based on Peculiarity in a Semantic Space Constructed of Free Writing
Analysis of Online Customer Reviews as an Example

Fumiaki Saitoh, Takuya Mogawa, and Syohei Ishuzu

Department of Industrial and Systems Engineering,
College of Science and Engineering,Aoyama Gakuin University
5-10-1 Fuchinobe,Chuo-ku,Sagamihara City, Kanagawa, Japan
saitoh@ise.aoyama.ac.jp

Abstract. Recently, the "voice of the customer (VOC)" such as exhibited by a Web user review has become easily collectable as text data. Based on large quantity of collected review data, we take the stance that minority review sentences buried in a majority review have high value. It is conceivable that hints of solution and discovery of new subjects are hidden in such minority opinions. The purpose of this research is to extract minority opinion from a huge quantity of text data taken from free writing in user reviews of products and services. In this study, we propose a method for extracting minority opinions that become outliers in a low-dimensional semantic space. Here, a low-dimensional semantic space of Web user reviews is constructed by latent semantic indexing (LSI). We were able to extract minority opinions using the Peculiarity Factor (PF) for outlier detection. We confirmed the validity of our proposal through an analysis using the user reviews of the EC site.

Keywords: Text Mining, Online Reviews, Peculiarity Factor (PF), Dimensionality Reduction, Voice of the Customer (VOC).

1 Introduction

The purpose of this research is to extract minority opinion from a huge quantity of text data taken from free writing in user reviews of products and services. Because of the significant progress in communication techniques in recent years, the "voice of the customer (VOC)" such as exhibited by Web user reviews has become easily collectable as text data. It is considered important to use VOC contributions to improve customer satisfaction and the quality of services and products[1]-[2].

In general, although text data of Web user reviews are readily available, they are typical concentrated into "satisfaction" or "dissatisfaction" categories. Existing text-mining tools can extract knowledge about such a majority user review sentence easily. Furthermore, because there is a strong possibility that

K.M. Lee, S.-J. Park, J.-H. Lee (eds.), *Soft Computing in Big Data Processing*,
Advances in Intelligent Systems and Computing 271,
DOI: 10.1007/978-3-319-05527-5_1,

knowledge about a majority review is obvious knowledge and an expected opinion, extraction of useful knowledge such as a new subject or hints of a solution can be difficult. For example, dissatisfied opinion concentrates on staff behavior in a restaurant where service is bad, and many satisfactory opinions about image quality are acquired for a product with a high-resolution display. It is easy to imagine concentrations of similar opinions occurring frequently as these examples.

Based on a large quantity of collected review data, we take the stance that minority review sentences buried in a majority review have high value. We aim at establishing a method for extracting minority Web review opinions.

Because Web review sentences consist of free writing, the variety of words appearing can be extensive. Word frequency vectors become very large sparse vectors that have many zero components. Therefore, an approach based on the frequency of appearance of words is not suitable for this task. However, approaches using TF-IDF as an index can be used for extracting characteristic words, but because these indexes function effectively only on data within large documents, they are unsuitable for Web review sentences, which appear within documents.

Therefore, we define a minority opinion Web review as an outlier in the semantic space of a Web review. In this study, we extract minority Web reviews through outlier detection on a semantic space constructed by using latent semantic indexing (LSI). The merit of our proposal is that it allaws us to extract minority opinion from free writing reviews accumulated in large quantities, without the need to read these texts.

2 Technical Elements of the Proposed Method

In this section, we outline the technical elements of our proposed method.

2.1 Morphological Analysis

To analyze a Web review comprising qualitative data quantitatively, it is necessary to extract the frequency of appearance of the words in each document as a feature quantity of the document. The number of lexemes appearing in all the data is a component of the word frequency vector. Morphological analysis is a technique for writing with segmentation markers between lexemes that is the unit of a vocabulary. This natural language processing is carried out based on information about grammar and dictionaries for every language[3][4].

Because the word frequency vector is measured easily, this processing is unnecessary for languages segmented at every word, such as English. However, it is a very useful tool for text mining of unsegmented languages such as Japanese, Korean and Chinese. We created a word frequency vector through a morphological analysis, because our subject was Japanese Web reviews.