LNCS 3411

Sung Hyon Myaeng
Ming Zhou
Kam-Fai Wong
Hong-Jiang Zhang (Eds.)

Information Retrieval Technology

Asia Information Retrieval Symposium, AIRS 2004
Beijing, China, October 2004
Revised Selected Papers

Springer

Sung Hyon Myaeng Ming Zhou
Kam-Fai Wong Hong-Jiang Zhang (Eds.)

Information Retrieval Technology

Asia Information Retrieval Symposium, AIRS 2004
Beijing, China, October 18-20, 2004
Revised Selected Papers

Springer

Volume Editors

Sung Hyon Myaeng
Information and Communications University (ICU)
119 Munji-Ro, Yuseong-Gu, Daejeon, 305-714, South Korea
E-mail: myaeng@icu.ac.kr

Ming Zhou
Hong-Jiang Zhang
Microsoft Research Asia
5F, Beijing Sigma Center
No. 49 Zhichun Road Haidian District, Beijing 100080, China
E-mail: {mingzhou,hjzhang}@microsoft.com

Kam-Fai Wong
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong, China
E-mail: kfwong@se.cuhk.edu.hk

Library of Congress Control Number: 2005921104

CR Subject Classification (1998): H.3, H.4, F.2.2, E.1, E.2

ISSN 0302-9743
ISBN 3-540-25065-4 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11398479 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3411

Preface

The Asia Information Retrieval Symposium (AIRS) was established by the Asian information retrieval community after the successful series of Information Retrieval with Asian Languages (IRAL) workshops held in six different locations in Asia, starting from 1996. While the IRAL workshops had their focus on information retrieval problems involving Asian languages, AIRS covers a wider scope of applications, systems, technologies and theory aspects of information retrieval in text, audio, image, video and multimedia data. This extension of the scope reflects and fosters increasing research activities in information retrieval in this region and the growing need for collaborations across subdisciplines.

We are very pleased to report that we saw a sharp increase in the number of submissions and their quality, compared to the IRAL workshops. We received 106 papers from nine countries in Asia and North America, from which 28 papers (26%) were presented in oral sessions and 38 papers in poster sessions (36%). It was a great challenge for the Program Committee to select the best among the excellent papers. The low acceptance rates witness the success of this year's conference.

After a long discussion between the AIRS 2004 Steering Committee and Springer, the publisher agreed to publish our proceedings in the Lecture Notes in Computer Science (LNCS) series, which is SCI-indexed. We feel that this strongly attests to the excellent quality of the papers.

The attendees were cordially invited to participate in and take advantage of all the technical programs at this conference. A tutorial was given on the first day to introduce the state of the art in Web mining, an important application of Web document retrieval. Two keynote speeches covered two main areas of the conference: video retrieval and language issues. There were a total of eight oral sessions run, with two in parallel at a time, and two poster/demo sessions.

The technical and social programs, which we are proud of, were made possible by the hard-working people behind the scenes. In addition to the Program Committee members, we are thankful to the Organizing Committee (Shao-Ping Ma and Jianfeng Gao, Co-chairs), Interactive Posters/Demo Chair (Gary G. Lee), and the Special Session and Tutorials Chair (Wei-Ying Ma). We also thank the sponsoring organizations: Microsoft Research Asia, the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong, and LexisNexis for their financial support, the Department of Computer Science and Technology, Tsinghua University for local arrangements, the Chinese NewsML Community for website design and administration, Ling Huang for the logistics, Weiwei Sun for the conference webpage management, EONSOLUTION for the conference management, and Springer for the postconference

LNCS publication. We believe that this conference set a very high standard for a regionally oriented conference, especially in Asia, and we hope that it continues as a tradition in the upcoming years.

Sung Hyon Myaeng and Ming Zhou (PC Co-chairs)
Kam-Fai Wong and Hong-Jiang Zhang (Conference Co-chairs)

Organization

General Conference Co-chairs

Kam-Fai Wong, Chinese University of Hong Kong, China
Hong-Jiang Zhang, Microsoft Research Asia, China

Program Co-chairs

Sung Hyon Myaeng, Information and Communications University (ICU), South Korea
Ming Zhou, Microsoft Research Asia, China

Organization Committee Co-chairs

Jianfeng Gao, Microsoft Research Asia, China
Shao-Ping Ma, Tsinghua University, China

Special Session and Tutorials Chair

Wei-Ying Ma, Microsoft Research Asia, China

Interactive Posters/Demo Chair

Gary Geunbae Lee, POSTECH, South Korea

Steering Committee

Jun Adachi, National Institute of Informatics, Japan
Hsin-Hsi Chen, National Taiwan University, Taiwan
Lee-Feng Chien, Academia Sinica, Taiwan
Tetsuya Ishikawa, University of Tsukuba, Japan
Gary Geunbae Lee, POSTECH, South Korea
Mun-Kew Leong, Institute for Infocomm Research, Singapore
Helen Meng, Chinese University of Hong Kong, China
Sung-Hyon Myaeng, Information and Communications University, South Korea

Hwee Tou Ng, National University of Singapore, Singapore
Kam-Fai Wong, Chinese University of Hong Kong, China

Organizing Committee

Lee-Feng Chien, Academia Sinica, Taiwan (Publicity, Asia)
Susan Dumais, Microsoft, USA (Publicity, North America)
Jianfeng Gao, Microsoft Research Asia, China (Publication, Co-chair)
Mun-Kew Leong, Institute for Infocomm Research, Singapore (Finance)
Shao-Ping Ma, Tsinghua University, China (Local Organization, Co-chair)
Ricardo Baeza-Yates, University of Chile, Chile (Publicity, South America)
Shucai Shi, BITI, China (Local Organization)
Dawei Song, DSTC, Australia (Publicity, Australia)
Ulich Thiel, IPSI, Germany (Publicity, Europe)
Chuanfa Yuan, Tsinghua University, China (Local Organization)

Program Committee

Peter Anick, Yahoo, USA
Hsin-Hsi Chen, National Taiwan University, Taiwan
Aitao Chen, University of California, Berkeley, USA
Lee-Feng Chien, Academia Sinica, Taiwan
Fabio Crestani, University of Strathclyde, UK
Edward A. Fox, Virginia Tech, USA
Jianfeng Gao, Microsoft Research Asia, China
Hani Abu-Salem, DePaul University, USA
Tetsuya Ishikawa, University of Tsukuba, Japan
Christopher Khoo, Nanyang Technological University, Singapore
Jung Hee Kim, North Carolina A&T University, USA
Minkoo Kim, Ajou University, South Korea
Munchurl Kim, Information and Communication University, South Korea
Kazuaki Kishida, Surugadai University, Japan
Kui-Lam Kwok, Queens College, City University of New York, USA
Wai Lam, Chinese University of Hong Kong, China
Gary Geunbae Lee, POSTECH, South Korea
Mun-Kew Leong, Institute for Infocomm Research, Singapore
Gena-Anne Levow, University of Chicago, USA
Hang Li, Microsoft Research Asia, China
Robert Luk, Hong Kong Polytechnic University, China
Gay Marchionini, University of North Carolina, Chapel Hill, USA
Helen Meng, Chinese University of Hong Kong, China
Hiroshi Nakagawa, University of Tokyo, Japan
Hwee Tou Ng, National University of Singapore, Singapore
Jian-Yun Nie, University of Montreal, Canada
Jon Patrick, University of Sydney, Australia
Ricardo Baeza-Yates, University of Chile, Chile
Hae-Chang Rim, Korea University, South Korea
Tetsuya Sakai, Toshiba Corporate R&D Center, Japan
Padmini Srinivasan, University of Iowa, USA
Tomek Strzalkowski, State University of New York, Albany, USA
Maosong Sun, Tsinghua University, China
Ulrich Thiel, Fraunhofer IPSI, Germany
Takenobu Tokunaga, Tokyo Institute of Technology, Japan
Hsin-Min Wang, Academia Sinica, Taiwan
Ross Wilkinson, CSIRO, Australia
Lide Wu, Fudan University, China
Jinxi Xu, BBN Technologies, USA
ChengXiang Zhai, University of Illinois, Urbana Champaign, USA
Min Zhang, Tsinghua University, China

Reviewers

Peter Anick
Yunbo Cao
Yee Seng Chan
Hsin-Hsi Chen
Zheng Chen
Aitao Chen
Tee Kiah Chia
Lee-Feng Chien
Fabio Crestani
Edward A. Fox
Jianfeng Gao
Hani Abu-Salem
Xuanjing Huang
Tetsuya Ishikawa
Christopher Khoo
Jung Hee Kim
Minkoo Kim
Munchurl Kim
Kazuaki Kishida
Kui-Lam Kwok
Wai Lam
Gary Geunbae Lee
Mun-Kew Leong
Gena-Anne Levow
Hang Li
Mu Li
Hongqiao Li
Chin-Yew Lin
Robert Luk
Wei-Ying Ma
Gay Marchionini
Helen Meng
Sung Hyon Myaeng
Hiroshi Nakagawa
Hwee Tou Ng
Jian-Yun Nie
Jon Patrick
Ricardo Baeza-Yates
Hae-Chang Rim
Tetsuya Sakai
Tomek Strzalkowski
Maosong Sun
Ulrich Thiel
Takenobu Tokunaga
Hsin-Min Wang
Haifeng Wang
Ross Wilkinson
Kam-Fai Wong
Lide Wu
Jinxi Xu
Peng Yu
Chunfa Yuan
ChengXiang Zhai
Hong-Jiang Zhang
Min Zhang
Ming Zhou
Jian-lai Zhou

Table of Contents

Web Search

Linguistic Issues in IR

Document/Query Models

Enabling Technology

Mobile Applications

Automatic Word Clustering for Text Categorization Using Global Information

Chen Wenliang, Chang Xingzhi, Wang Huizhen, Zhu Jingbo, and Yao Tianshun

Natural Language Processing Lab
Northeastern University, Shenyang, China 110004
chenwl@mail.neu.edu.cn

Abstract. High dimensionality of feature space and short of training documents are the crucial obstacles for text categorization. In order to overcome these obstacles, this paper presents a cluster-based text categorization system which uses class distributional clustering of words. We propose a new clustering model which considers the global information over all the clusters. The model can be understood as the balance of all the clusters according to the number of words in them. It can group words into clusters based on the distribution of class labels associated with each word. Using these learned clusters as features, we develop a cluster-based classifier. We present several experimental results to show that our proposed method performs better than the other three text classifiers. The proposed model has better results than the model which only considers the information of the two related clusters. Specially, it can maintain good performance when the number of features is small and the size of training corpus is small.

1 Introduction

The goal of text categorization is to classify documents into a certain number of predefined categories. A variety of techniques for supervised learning algorithms have demonstrated reasonable performance for text categorization[5][11][12]. A common and overwhelming characteristic of text data is its extremely high dimensionality. Typically the document vectors are formed using bag-of-words model. It is well known, however, that such count matrices tend to be highly sparse and noisy, especially when the training data is relatively small. So when the text categorization systems are applied, there are two problems to be counted:

- High-dimensional feature space: Documents are usually represented in a high-dimensional sparse feature space, which is far from optimal for classification algorithms.
- Short of training documents: Many applications can't provide so many training documents.

A standard procedure to reduce feature dimensionality is feature selection, such as Document Frequency, χ^2 statistic, Information Gain, Term Strength, and

S. H. Myaeng et al. (Eds.): AIRS 2004, LNCS 3411, pp. 1–11, 2005.

Mutual Information[13]. But feature selection is better at removing detrimental, noisy features. The second procedure is cluster-based text categorization[1][2][3] [10]. Word clustering methods can reduce feature spaces by joining similar words into clusters. First they grouped words into the clusters according to their distributions. Then they used these clusters as features for text categorization.

In this paper, we cluster the words according to their class distributions. Based on class distributions of words, Baker[1] proposes a clustering model. In clustering processing, we will select two most similar clusters by comparing the similarities directly. But Baker's model only considers two related clusters, when computing the similarity between the clusters without taking into account the information of other clusters. In order to provide better performance, we should take into account the information of all the clusters when computing the similarities between the clusters. This paper proposes a clustering model which considers the global information over all the clusters. The model can be understood as the balance of all the clusters according to the number of words in them.

Using these learned clusters as features, we develop a cluster-based Classifier. We present experimental results on a Chinese text corpus. We compare our text classifier with the other three classifiers. The results show that the proposed clustering model provides better performance than Baker's model. The results also show that it can perform better than the feature selection based classifiers. It can maintain high performance when the number of features is small and the size of training corpus is small.

In the rest of this paper: Section 2 reviews previous works. Section 3 proposes a global Clustering Model (globalCM). Section 4 describes a globalCM-based text categorization system. Section 5 shows the experimental results. Finally, we draw our conclusions at section 6.

2 Related Work

Distributional Clustering has been used to address the problem of sparse data in building statistical language models for natural language processing[7][10]. There are many works[1][2] related with using distributional clustering for text categorization.

Baker and McCallum[1] proposed an approach for text categorization based on word-clusters. First, find word-clusters that preserve the information about the categories as much as possible. Then use these learned clusters to represent the documents in a new feature space. Final, use a supervised classification algorithm to predict the categories of new documents. Specifically, it was shown there that word-clustering can be used to significantly reduce the feature dimensionality with only a small change in classification performance.

3 Global Clustering Model Based on Class Distributions of Words

In this section, we simply introduce the class distribution of words[1]. Then we propose the Global Clustering Model, here we name it as globalCM. In our clustering model, we define a similarity measure between the clusters, and add the candidate word into the most similar cluster that no longer distinguishes among the words different.

3.1 Class Distribution of Words

Firstly, we define the distribution $P(C|w_t)$ as the random variable over classes C, and its distribution given a particular word w_t. When we have two words w_t and w_s, they will be put into the same cluster f. The distribution of the cluster f is defined

$$\begin{aligned} P(C|f) &= P(C|w_t \vee w_s) \\ &= \frac{P(w_t)}{P(w_t)+P(w_s)} \times P(C|w_t) \\ &+ \frac{P(w_s)}{P(w_t)+P(w_s)} \times P(C|w_s) \ . \end{aligned} \tag{1}$$

Now we consider the case that a word w_t and a cluster f will be put into a new cluster f_{new}. The distribution of f_{new} is defined

$$\begin{aligned} P(C|f_{new}) &= P(C|w_t \vee f) \\ &= \frac{P(w_t)}{P(w_t)+P(f)} \times P(C|w_t) \\ &+ \frac{P(f)}{P(w_t)+P(f)} \times P(C|f) \ . \end{aligned} \tag{2}$$

3.2 Similarity Measures

Secondly, we turn to the question of how to measure the difference between two probability distributions. Kullback-Leibler divergence is used to do this. The KL divergence between the class distributions induced by w_t and w_s is written $D(P(C|w_t)||P(C|w_s))$, and is defined

$$-\sum_{j=1}^{|C|} P(c_j|w_t) \log \frac{P(c_j|w_t)}{P(c_j|w_s)} \ . \tag{3}$$

But KL divergence has some odd properties: It is not symmetric, and it is infinite when $p(w_s)$ is zero. In order to resolve these problems, Baker[1] proposes a measure named "KL divergence to the mean" to measure the similarity of two distributions(Here we name it as S_{mean}). It is defined

$$\frac{P(w_t)}{P(w_t)+P(w_s)} \times D(P(C|w_t)||P(C|w_s \vee w_t))$$
$$+\frac{P(w_s)}{P(w_t)+P(w_s)} \times D(P(C|w_s)||P(C|w_s \vee w_t)) . \quad (4)$$

S_{mean} uses a weighted average and resolves the problems of KL divergence. But it only considers the two related clusters without thinking about other clusters. Our experimental results show that the numbers of words in learned clusters, which are generated by Baker's clustering model, are very different. Several clusters include so many words while most clusters include only one or two words.

We study the reasons of these results. When Equation 4 is applied in the clustering algorithm, it can't work well if the numbers of words in the clusters are very different at iterations.

For example, we have a cluster f which include only a word(In Baker's clustering model, a new candidate word will be put into an empty cluster). We will compute the similarities between f and the other two clusters(f_i and f_j) using Equation 4. Let f_i has many words(ie. 1000 words) and f_j has one or two words. We define:

$$\begin{aligned} S_i = & \frac{P(f)}{P(f)+P(f_i)} \times D(P(C|f)||P(C|f \vee f_i)) \\ & + \frac{P(f_i)}{P(f)+P(f_i)} \times D(P(C|f_i)||P(C|f \vee f_i)) \\ & = (1-\alpha_i) \times D_{i1} + \alpha_i \times D_{i2} . \end{aligned} \quad (5)$$

$$\begin{aligned} S_j = & \frac{P(f)}{P(f)+P(f_j)} \times D(P(C|f)||P(C|f \vee f_j)) \\ & + \frac{P(f_j)}{P(f)+P(f_j)} \times D(P(C|f_j)||P(C|f \vee f_j)) \\ & = (1-\alpha_j) \times D_{j1} + \alpha_j \times D_{j2} . \end{aligned} \quad (6)$$

According to Equation 2, if a word is added to a cluster, the word will affect tiny to the cluster which includes many words and affect remarkable to the cluster which includes few words. So the distribution of $f \vee f_i$ is very similar to f_i because f_i has many words and f has only one word. And then D_{i2} is near zero. α_i is near 1 and $(1-\alpha_i)$ is near zero because the number of f_i is very large than f. We know:

$$S_i \approx D_{i2} \approx 0 . \quad (7)$$

So when we compute the similarities between f and the other clusters using Equation 4, f will be more similar to the cluster which includes more words.