Study on Microblogging Marketing System Based on KNN Classification Algorithm

Qingqiang Meng[1], Xue Han[2]

1 North China Institute of Aerospace Engineering, Langfang, China
2 Langfang branch of Hebei University of Technology, Langfang, China

Corresponding author is Qingqiang Meng

Abstract
Combining with the status quo of micro-blog marketing, this article designs and implements the microblogging marketing system based on data mining technology. The smart microblogging system has two objectives: first, search for potential customers precisely; second, interact with customers. Owing to these two purposes, the system consists of two parts: the Fan-related Data Acquisition and Analysis Subsystem and Micro-blog Interactive Subsystem. The former mainly includes acquisition and analysis of micro-blog, micro-blog index analysis, fans attribute analysis, micro-blog propagation path analysis, and social network analysis. The latter mainly concerns about potential customers, recommends micro-blogs and friends to potential customers and comments or forwards the micro-blogs. Through transferring the API interface provided by Sina and Tencent micro-blog development platform, we get the raw data required by the system. Micro-blog short text preprocessing uses the CAS ICTCLAS analysis system, extracts improved mutual information feature from the text, classifies the short text by using KNN sort algorithm and achieves the purpose of micro-blog interaction with the target users. According to function testing and result interpretation, the overall system has reached the desired goals with a certain value in engineering applications.

Key words: DATA MINING, SHORT TEXT, CLASSIFICATION, INTERACTION

1. Introduction

Since 1995, Web technology has entered a stage of rapid development. The Internet Web page number and service site number increase exponentially [1]. In 2004, Internet PIW (publicly indexable Web) page number by an order of to“, magnitude, but also has daily added 8million new page speed. At the same time, the number of Web server can be doubled in 23 weeks. Web has become an open, dynamic, global information service center, and an important means of obtaining information. How to extract information from a large number of Web information that people interested in is an important subject in the study of modern information [2].

With the rapid development of the Internet, we have quickly stepped into the information age. Network information, however, let us too busy to attend it all. Just taking Sina micro-blog for example, there are more than half of netizens using micro-blog and generating hundreds of millions of micro-blogs. Confronted by such a huge market, how to mine data from potential customers has a unique attraction for the business.

With the rapid development and information technology, the text information increases exponentially. As an imporkant technology of managing large amount of information, text classification is able to solve the problem of chaotic information effectively.

Meanwhile, it's convenient for user to retrieve the required information accurately. Consequently, the text classification possesses high value of application value in the field of information retrieval, classification and filtering mails, tracking topics, etc, having been a hot research field in data mining. With the expanding of user number and the platform, it helps enterprise to create a good marketing environment. Micro-blog operators can be cooperated with enterprise in network marketing for new products, and new brand [3].

The abundant micro-blog information provides the opportunity and challenge to all the various trades and occupations. The enterprise can be communicated effectively with its users or potential users in the platform. The enterprise can more accurate positioning its own brand through the analysis of user data. However, with the surge in the number of users and instant information published, ordinary users get lost when they want to obtain the information which they are interested in. At the same time, the information which enterprise published will submerged in the vast amounts of data. As a result, users could not obtain the information.

2. Overview of KNN Classification Algorithm AND METHOD

Focusing on improving the performance of KNN classifier, this dissertation introduces the definition of text categorization, preprocessing procedure of text, definition and algorithms of feature selection, comparison of traditional and supervised term weighting, text classification algorithms, and performance measurement followed by depth studying and improving the method of feature selection, term weighting and classification.

(1) This dissertation put forward the improvement on feature selection on basis of ant colony optimization. By studying and design the fitness function, probabilistic transition rule and pheromone update rule, the improved method can exclude the associated features and redundant features, as well as reduce space and time of calculations effectively, then boost the calculation accuracy, as a result, making the classification performance better than before finally.

$$\mathrm{MI}(w, C_i) = \log\frac{P(x,y)}{P(x)P(y)} \approx \log\frac{F \times N}{(F + F_c) \times (F + F_w)} \tag{1}$$

Where F is the number of times

(2) This dissertation also proposes the improvement of supervised term weighting based on TF-RFIDR Based on the theory of supervised term weighting of TF-RF, this dissertation proposes the method of TF-RFIDF, combing the relevance frequency and inverse document frequency. It can take advantage of sample distribution and prior information of categories, thus improve the classification performance.

$$\begin{aligned}\mathrm{IG}(w) = -\sum_{i=1}^{M} P(C_i)P(C_i) + P(w)\sum_{i=1}^{M} P(C_i|w)\log P(C_i|w) \\ +P(\overline{w})\sum_{i=1}^{M} P(C_i|\overline{w})\log P(C_i|\overline{w})\end{aligned} \tag{2}$$

Where IG(w) is information gain.

(3) This dissertation proposes the improvement of KNN classification algorithm based on association rules. Algorithm of Apriori is used to extract frequent feature set and its associated text of for each category for different types of training samples, so as to determine the appropriate number of neighbor k for unknown class of text, and then determine the category of text according to neighbors' category. The improved algorithm can determine the k value better, and reduce the time complexity.

$$\mathrm{CE}(w) = P(w)\sum_{i=1}^{M} P(C_i|w)\log\frac{P(C_i|w)}{P(C_i)} \tag{3}$$

Where CE(w) is expected cross entropy.

The experimental results at last show the three improved algorithms can improve classification accuracy for text classification, thus proving the effectiveness of the algorithms. (Fig. 1)

We introduces text classification research present situation and the patent classification background. Secondly, it systematically introduces the key technologies of text classification and various classification algorithms, and various classification algorithms in different fields of application. At present, KNN classifiers with respect to the other classifiers classify better in many classifiers, but it still has some shortcomings, such as the classification speed slow, the classification accuracy low. For overcoming these shortcomings of the KNN classification algorithm, we propose an idea that an optimized KNN algorithm classifier, the classification modular by training, classification and evaluation of three parts [4]. Optimized KNN algorithm is based on the cluster of the original space model the training set for processing, and the training set is similar to the text forms a cluster, each cluster as a common text, calculated for each cluster center vector, and then set a threshold, higher than the threshold of cluster management, and reformation of the training set. The classification algorithms to save the original text information based on the sparse char-

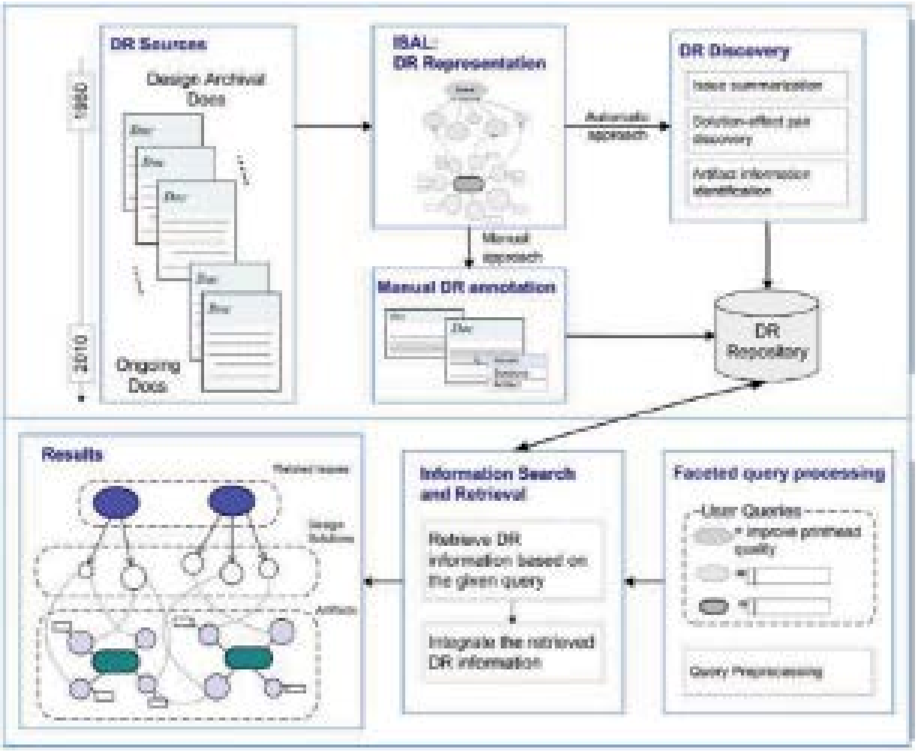


Figure 1. Framework and Management

acteristic, according to the text. This paper uses the compressed representation model, and then does the calculation of distance and the final will be the test texts which belong to the category of [5]. This algorithm not only reduces the amount of computation, but also improves the KNN classification speed and accuracy. (Fig.2)

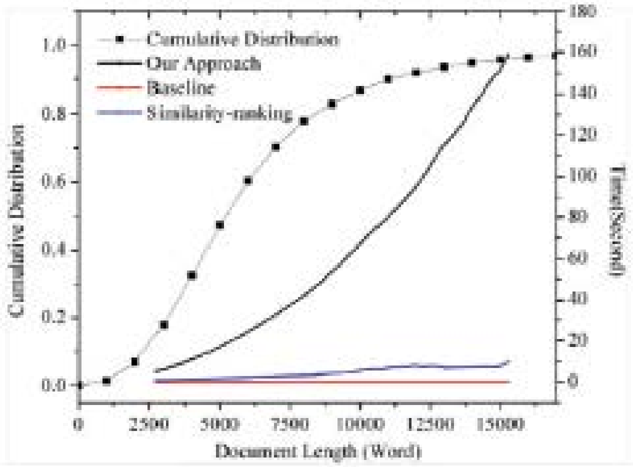


Figure 2. Simulation

3. Feature extraction technology in KNN text and model

In the data mining, text classification is an important area of research, KNN algorithm which is one of the best methods of text classifying in the vector space model (VSM) is a simple, example based and none-parameter method. The main steps are: text segmentation, feature extraction (feature weight calculation and characteristics of the word choice), building the feature model, training classifier. The feature extraction which is the core of the text classification system, the feature extraction method has a major impact on the result of text classification. The traditional feature extractions methods are based on statistical methods, commonly used are: DF (Document Frequency), ECE(Expected Cross Entropy), OR, IG (Information Gain), MI (Mutual Information), x} statistics (CHI) and so on. Above methods have many deficiencies: when categories and features have a high degree of uneven distribution, you cannot deal effectively with low-frequency words; for the mishandling of individual characteristics, resulting in the local optimal solution. In addition, KNN classification algorithm whether can select the appropriate K value will also affect the quality of classification results, the fixed K value ignores the influence of the category and the document number of training text. If the K value is too large, the text tends to belong to the class which contains more texts, classification performance is poor; If K value is too small, text has too few neighbors, this will reduce the classification accuracy.

$$w(t_i,d)=\frac{\left(\log\left(tf(t_i,d)\right)+1.0\right)\times\log(N/n_i)}{\sqrt{\sum_{t_i\in d}\left[\left(\log\left(tf(t_i,d)\right)+1.0\right)\times\log(N/n_i)\right]^2}}\qquad(4)$$

Where $w(t_i, d)$ is weight.

$$sim(x, d_i) = \frac{\sum_{n=1}^{m} w_n * w_{in}}{\sqrt{(\sum_{n=1}^{m} w_n^2)(\sum_{n=1}^{m} w_{in}^2)}} \quad (5)$$

Where x is The dimension of feature vector

Aiming at the problems at the feature extraction technology, this paper puts forward a new feature extraction technology which based on the genetic algorithm. In this method the statistical value of words that can identify the size of the correlation between words and category, which will be introduced to feature vector, as the initial population for genetic algorithm heuristic search, while the nature of the feature extraction. At the same time, this paper presents a new fitness function and crossover rules. This paper put forward a new fitness function and the cross-rule for the nature of feature extraction. Experiments have proved that the new feature extraction technology which based on the genetic algorithm can choose a category of accurate characterization of text feature. (Fig.3)

Adjacency matrix is the formula (6).

$$X=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad (6)$$

In order to solve the defect of the fixed K value, this paper proposes a kind of dynamic obtain k-valued for KNN classification algorithm, experimental results show that the dynamic obtain k-valued KNN classification algorithm with high performance.

4. Results and comparisons

To evaluate the classification accuracy of KNN classification system, we use Sohu news on the Internet to train and classify test. The corpus includes education, sports, environment, entertainment, technology, economy six categories, a total of 780 texts. At the same time, we test the improved KNN algorithm, and then analyze and compare the accuracy of the algorithm. Experimental data can be used for information retrieval, information filtering, digital libraries and web classification reference. (Fig.4)

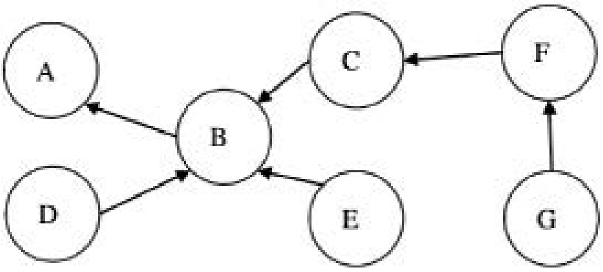
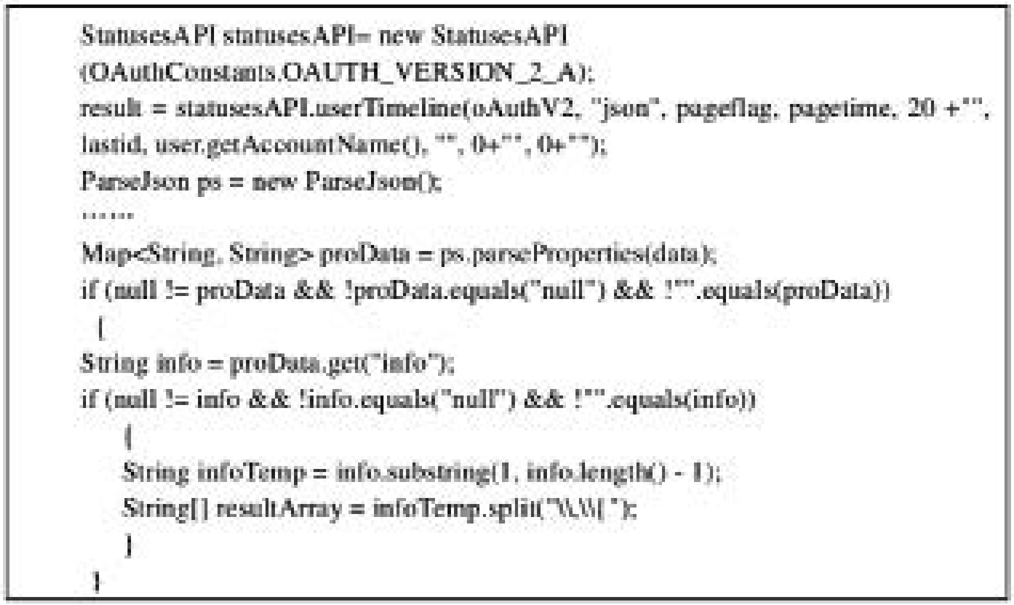


Figure 3. Micro-blog Propagation Sociogram



```
StatusesAPI statusesAPI= new StatusesAPI
(OAuthConstants.OAUTH_VERSION_2_A);
result = statusesAPI.userTimeline(oAuthV2, "json", pageflag, pagetime, 20 +"",
lastid, user.getAccountName(), "", 0+"", 0+"");
ParseJson ps = new ParseJson();
……
Map<String, String> proData = ps.parseProperties(data);
if (null != proData && !proData.equals("null") && !"".equals(proData))
  {
String info = proData.get("info");
if (null != info && !info.equals("null") && !"".equals(info))
    {
    String infoTemp = info.substring(1, info.length() - 1);
    String[] resultArray = infoTemp.split("\\,\\{ ");
    }
  }
```

Figure 4. Micro-blog core code

Conclusion

This paper introduces the relative theories of Chinese text classification, such as the vector space model, the Chinese word segmentation, the feature selection, the classification method, evaluation indicator, weight calculation method and similarity calculation method.

Through analyzing the TFIDF in details, contrapose it only consider the shortage of word frequency and the feature distribution in the training text set, proposed an improved scheme add the feature distribution in each class and all texts within class into the original formula.

Contrapose the shortcomings in calculating the text similarity, put out one improved scheme based on the in-depth analysis of KNN classification method. The new scheme introduces the idea of central vector classification method, and taking into account the number of common feature between the text to be classified and the training text is important to the dassification.

Based on the theoretical research, construct a Chinese text categorization system including four functional modules, which are pretreatment module, feature selection module, classification module and evaluation and display module. This system uses SQL Server 2000 as its back-end database, and is realized through C# language.

Finally using the realized Chinese text categorization system as the testing platform, verify the validity and feasibility of improvement TFIDF weight calculation method and KNN classification method through experiment.

References

1. Christel, M. G., Smith, M. A., Taylor, C. R., & Winkler, D. B., January. Evolving video skims into useful multimedia abstractions. In Proceedings of the SIGCHI conference on Human factors in computing systems, 1998, pp.171-178.
2. De Medeiros, A. P., & Schwabe, D., Kuaba approach: Integrating formal semantics and design rationale representation to support design reuse. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 2008, 22(04), pp. 399-419.
3. Fu, Y. H., Campioli, M., Van Oijen, M., Deckmyn, G., & Janssens, I. A., Bayesian comparison of six different temperature-based budburst models for four temperate tree species. Ecological Modelling, 2012,230, pp. 92-100.
4. Wang, J., & Tao, Q., Machine learning: The state of the art. Intelligent Systems, IEEE, 2008, 23(6), pp. 49-55.
5. Lim, S. C. J., Liu, Y., & Lee, W. B., Multifacet product information search and retrieval using semantically annotated product family ontology. Information Processing & Management, 2010,46(4), pp. 479-493.