Research Journal of Applied Sciences, Engineering and Technology 4(11): 1500-1503, 2012
ISSN: 2040-7467

Submitted: December 13, 2011 Accepted: January 21, 2012 Published: June 01, 2012

# An Improved Classification Algorithm for Structured Data Based on Secondary Data Processing

Yubo Jia, Hongdan Fan, Qian Zhang, Xu Li and Guanghu Xia
College of Information, Zhejiang Sci-Tech University, Hangzhou, 310018, China

**Abstract:** Secondary Data Processing deals the information further by re-crawling and categories based on the basic of structured data. It is the key researching module of Vertical Search Engines. This paper introduces an application model of vertical search engine briefly and proposes an improved classification algorithm for the categories to enhance the advantage of vertical search engines. The algorithm achieves the responsiveness and the accuracy of vertical search by reducing the time complexity and accelerating the speed of classification. This paper proved the improved algorithm has the better feasibility and robustness when it's used in secondary structured data processing based on vertical search engines.

**Key words:** Algorithm, categories, classification, data processing, search

## INTRODUCTION

According to the International Federation of With the development of the Internet, search engines continue to meet the demand of great information resources, but cannot take into account the accuracy and responsiveness (Jincheng and Peiliang, 2009) of information search, so vertical search engines have emerged to meet the needs of users at this time. But how to classify web pages and texts in the searching process is critical. In this paper, we propose an improved algorithm to reach the aim of categories.

Secondary Data Processing and Participle Module is the key modules of the model. It is a process of structured analysis that the mode separation of the text content on web pages stored in the database, adjustment of data and analysis of related links. It will deal the information further by re-crawling and categories based on the basic of structured data. Secondary processed information is stored into Index Database to provide specific service for users. Relevance Ranking Module sorts the data from Index Database according to user requirements and sends to the User Interface Module. So a successful search results back to the user.

This study proposed an improved classification algorithm to deal the structured data based on secondary data processing in order to enhance the advantages of vertical search engines. The algorithm achieves the responsiveness and the accuracy of vertical search by reducing the time complexity and accelerating the speed of classification. This paper proved the improved algorithm which has the better feasibility and robustness when it's used in secondary structured data processing based on vertical search engines.

## COMMON CLASSIFICATION ALGORITHM

**The basic idea of categories algorithm:** To represent the texts that need to be classified and web pages as vectors and calculate the similarity of vectors between web pages and samples from the space consisted by the training samples. Then we can obtain k pieces of nearest and most similar texts or pages. According to the type of these web pages to determine the category of the new ones, then compute the seniorities of classes in the k neighbors from the new version in turn. Assign the pages and texts to the class with the most powerful seniority.

Texts are represented into a vector in the vector space model (Yoshida *et al*., 2010), so calculating the similarity of texts can be transformed into computing the cosine of the angle between vectors. Assume two pages or texts $d_i = (w_{i1}, w_{i2}, \ldots, w_{in})$, $d_i = (w_{i1}, w_{i2}, \ldots, w_{jn})$, The formula of similarity between $d_i$ and $d_j$ is $sim(d_i, d_j)$:

$$sim(d_i, d_j) = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{\sum_{i=1}^{n} w_{ik}^2} \sqrt{\sum_{j=1}^{n} W_{jk}^2}}$$

The greater value of $sim(d_i, d_j)$ means the smaller angle between $d_i$ and $d_j$. When $sim(d_i, d_j) = 1$, $d_i$ and $d_j$ are parallel or coinciding. At this time they are most similar. If $sim(d_i, d_j)$ is near 0, it means vectors $d_i$ and $d_j$ are vertical .They have the smallest value of similarity.

We can compute the seniorities of classes in the k neighbors from the new version in turn. The formula of seniorities is:

**Corresponding Author:** Hongdan Fan, College of Information, zhejiang Sci-Tech University, Hangzhou, 310018, China

$$P(\bar{x}, C_j) = \sum_{\bar{d}_i \in KNN} sim(\bar{x}, \bar{d}_i) y(\bar{d}_i, C_j)$$

$\bar{x}$ is the basic characteristic quantity of new pages. $sim(\bar{x}, \bar{d}_i)$ is the formula of similarity. $y(\bar{d}_i, C_j)$ is type property function it means if class $C_j$ includes $\bar{d}_i$, then $y(\bar{d}_i, C_j) = 1$, else $y(\bar{d}_i, C_j) = 0$.

**Analysis of common classification algorithm:** The advantage (Lifang and Yang, 2009) of the traditional algorithm is that it can make use of the correlation directly between the two given samples, thus reducing the influence caused by inadequate choices from classification feature and also the error term during the process of classification. But this algorithm compares with every sample vector in sample space in order to find k neighbors of sample classification, so it causes computing times to increase and system performance degrades.

## IMPROVED ALGORITHM

**The basic idea of improved categories algorithm:** When searching k neighbors of one sample, only find those which have overlapping words with those unsorted pages, thus reducing the search scope and accelerating the speed of search. The structure of the improved algorithm includes term arrays and their lists. Term array is the ID of feature entry that stored in arrays undergone the feature extraction after dividing the training texts into words. Every entr (ti) in term arrays has its own pointer ,which points to the list formed by all of the texts. The text list includes two parts, such as ID and the seniorities ti in text. After list of the text including ti is finished, sorting decreasingly according to the value of seniorities, then make a further optimization to narrow the search range of classification algorithm.

**Description of improved classification algorithm:** Page d ready to be sorted is expressed as text vector V($w_1$, $w_2$, … , $w_n$), search each document list li ($1 \le i \le n$) of term ti($1 \le i \le n$) in vector V, then merge list li and remove the ID of same texts in lists, so we can obtain the similarity between the set of texts ID and texts in the set.

**Analysis of improved classification algorithm:** Similarity is only between the improved algorithm and the documents vectors of intersectional training texts ready to category. So it can reduce the time complexity (Weimin and Wu, 2008) and accelerate the speed of classification on a certain extend. But the improved classification algorithm is more similar with part of training samples, so there are many certain overlaps in sample vectors. The improved algorithm is a compromised algorithm compared with common algorithm and has a better practicability.

**Application of improved algorithm in vertical search engines:**

**Theoretical description:** Secondary Data Processing and Participle Module is a process of structured analysis that the mode separation of the text content on web pages stored in the database, adjustment of data (Stephen *et al*., 1999) and analysis of related links (Claudio et al., 2008). It will deal the information further by re-crawling and categories based on the basic of structured data. This improved classification algorithm stresses the responsiveness and the accuracy of vertical search when it's used in vertical search engines model.

**Experiments:** This experiment designs two evaluation indicators, precision ratio and recall ratio, for the improved algorithm. Precision ratio (Ning *et al*., 2005) is defined as the percents of texts which meet with the result of artificial classification in all given documents (Fig. 1). Its mathematical formula is Precision ratio = (the number of correct classification texts/all classification texts). Recall ratio is defined as the rate of texts which meet with classification system in all the deserved results of artificial classification. Its mathematical formula is Recall rate = (a mount of correct classification texts/ deserved texts). Precision rate and recall rate reflect the quality of classification in two different aspects. We should balance the two aspects. Thus there is a new evaluation indicator, the value of F1, and its mathematical formula is F1 (Zhili *et al*., 2007) = (2*precision rate*recall rate)/deserved texts.

This experiment conducts a test of 300 web pages of the given vertical search engine model. In the algorithm experiment, we set the threshold (Fengjing and Yu, 2003) k = 100 and the vectors dimension threshold V are given different values.

## EXPERIMENTAL RESULTS AND ANALYSIS

The precision and recall of improved Algorithm are showed as Fig. 2 and 3.

It can be seen from Fig. 2 obviously, the improved algorithm is slightly more common in precision, but when the dimensions of vectors reached a certain value, the accuracy decreased significantly. So in the case of certain threshold, the vector dimension should be appropriate. We

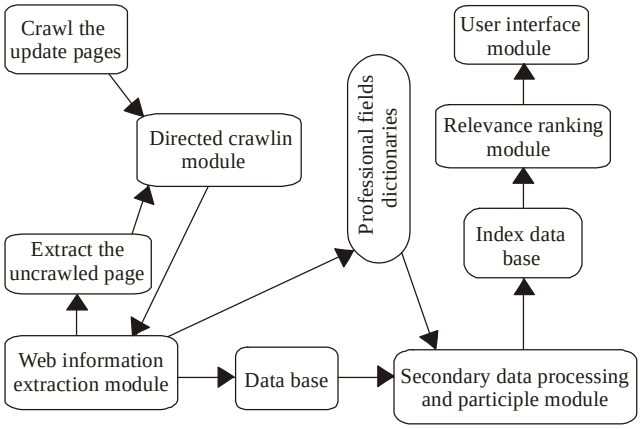


Fig. 1: Design of application model (Yubo *et al*., 2011)

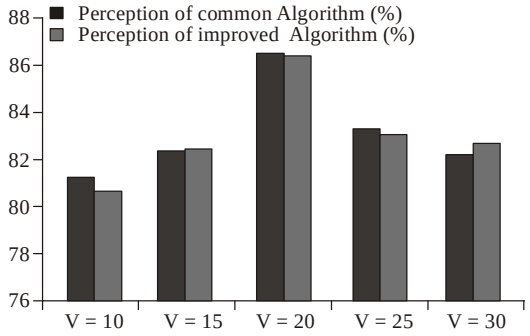


Fig. 2: Compared precision of algorithm

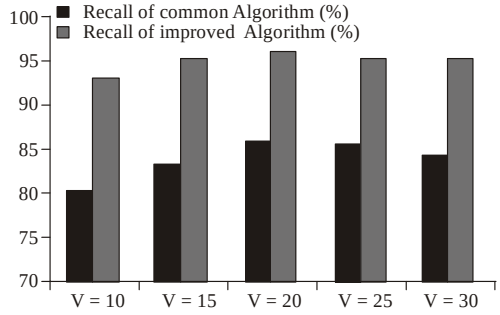


Fig. 3: Compared recall of algorithm

can discover from Fig. 3 that recall is less affected by the dimension of vectors and keeps balance overall, but recall of our improved algorithm increases clearly.

## CONCLUSION

This improved algorithm enhances the recall and efficiency of categories significantly on the basis of slight loss of precision. In the vertical search engines, improved classification algorithm can achieve classification function of the secondary processing and Participle n module. Therefore, the improved algorithm has a better practicability for vertical search engines.

## REFERENCES

Claudio, B., V. Crescenzi and P. Merialdo, 2008. Crawling programs for wrapper-based application (C). IEEE IRI, pp: 160-165.

Fengjing, S. and Z. Yu, 2003. Principle and Algorithm of Data mining (M). Waterpub Press, Beijing, pp: 126-176.

Jincheng, Y. and P. Ling, 2009. Improvement of pagerank algorithm for search engine (J). Comput. Eng., 35(22): 35-37.

Lifang, P. and B. Yang, 2009. Study on KNN arithmetic based on cluster (J). Comput. Eng. Design, 30(18): 4260-4261.

Ning, Z., Z. Jia and Z. Shi, 2005. Text categorization with KNN algorithm (J). Comput. Eng., 31(8): 171-185.

Stephen, S., C. Cardie and R. Mooney, 1999. Learning Information Extraction Rules for Semi-strucrured and Free Text, Machine Learning.

Weimin, Y. and W. Wu, 2008. Data Structures (C Edition) (M). Tsinghua University Press, Beijing, pp: 13-17.

Yoshida, M., O. Satoshi, A. Nskao and K. Kswashima, 2010. Controlling file distribution in the share network through content poisoning. Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications (AINA2010), pp: 1004-1011.

Yubo, J., H. Fan and G. Xia, 2011. Design of an Application Model Based on Vertical Search Engine. Proceeding of the 2011 2nd International Conference on Networking and Distributed Computing, (NDC' 2011), pp: 57-60.

Zhili, P., X. Shi, M. Maurizio and L. Yanchun, 2007. An enhanced text categorization method based on improved text frequency approach and mutual information algorithm (J). Prog. Nat. Sci., 17(12) pp: 1494-1500.