Design and Realization of SVM Topic Crawler Based on Incremental Learning

Zhou Ping

College of Computer Science and Technology. NanJing University of Aeronautics and Astronautics, NanJing, JiangSu ,China

Keywords: topic crawler, support vector machine, incremental learning.

Abstract. Using information technology such as search engine for collecting and monitoring of network public opinion is a practical and effective method. This paper puts forward an improved algorithm of SVM classifier based on incremental learning online, and then implements a topic crawler system for network public opinion and to grab the public opinion.

Concept of SVM Topic Crawler

SVM (Support Vector Machine) is first proposed by Vapnik in 1995. It exhibits many unique advantages in addressing the small sample, nonlinear and high dimensional pattern recognition. SVM pattern recognition is based on statistical learning theory method has important applications in computer pattern recognition field. The SVM research is not perfect, cannot efficiently solve pattern recognition problems. But with the study of the application of statistical learning theory and neural networks than the new machine learning methods encounter some significant difficulties, such as how to determine the issue of network structure, through learning and learning problems due to local minima problems, leading to many researchers added to the study of SVM classification algorithm to improve. This effectively promotes the rapid development of SVM classification algorithm and continuous improvement, and SVM classifier was quickly extended to other machine learning problems fitting in function today SVM classification algorithm on text categorization has been successful applications. Since the 1990s, the rapid development of Internet technology, automatic text classification research has entered a new stage, based on machine learning text classification technology gradually replaced the method based on knowledge engineering automatic text classification has become the main form. Carried out in comparison with the classification Bayesian k-nearest neighbor and decision tree, the support vector machine method achieved the best classification accuracy since more and more researchers began to pay attention to them, and for the support of two standard corpus SVM and text classification were studied and put forward a number of new methods. In recent years, the introduction of SVM classifier topic crawlers, used to guide and supervise the theme crawler got the attention of many scholars. Johnson first SVM classification algorithm supervision focused crawler conducted theoretical research, proposed a SVM classifier model to guide the crawling reptile theme, and a lot of related experiments. More and more scholars began to use support vector machine guidance and supervision topic reptiles, Michelangelo and other support vector machine to guide the theme reptiles, and proposes a support vector machine classification algorithm page. Topic relevance determination method of obtaining a direct impact on the rate of theme crawler, traditional themes reptiles crawling in the website, acquisition rate has been low. Based on SVM classifier, this paper further studies SVM classification algorithm in the prediction of the page subject classification and proposes an incremental learning of SVM classification algorithm, and finally applies it to the network crawl on public opinion.

Improved Algorithm of SVM Topic Crawler

Classification Algorithm of SVM Topic Crawler

The problem of Web text representation is a good method to construct a web page classifier model. In the selection process of the text feature words in the web pages, we must first deal with some of the words that have no substantive meaning. Feature selection must have certain principles. They should consider the number and frequency of feature words. This section will define the lexical features of the number and frequency of closed values, the frequency and number of the selected feature must meet the qualifications. In the process of solving the support vector machine, the weight value is a single web page; in the process of establishing a small class model, the entry value is a category of web pages. The detailed steps of classification algorithm are as follows: (1) first, we should be trained to carry out the training of web pages, and then use the technology of web page purification and Chinese word segmentation in the form of web pages, which will be processed in a candidate list. (2) Set feature item weight of each feature list word list, mainly by TF-IDF weight calculation method, and from the candidate list selected m most gifted special eigenvalue. (3) The subject vector model of the training document is generated and stored in a training list. (4) Select the appropriate kernel function and parameters, and use it to train the SVM classifier. (5) The classification vector model after the training. By using the classified vector model of the above process, it calculates the similarity of the web pages, and then to determine the correlation degree of the web pages.

Improved Classification Algorithm of SVM Topic Crawler

Because of the rapid and large scale of Web pages in the Internet, it is not practical to obtain a complete training set at the initial stage. Over time, some of the spatial vector related to the theme of the web page may be more and more features, in the process of a large number of related topics, so we need to carry out effective incremental learning on the Web page. For the Internet in a burst of an event, in human to collect the training set samples may have some positive and negative unclear boundary samples, which also led to the impact on the training results. If the process of incremental learning is just the union of history and the incremental set, then the training result may not just as one wishes, so we propose a set of new history for learning process combined with incremental learning, if a theme of the history of the training set, increase the amount of learning the training set is N_j, then SVM subject classification algorithm are as follows: (1) The SVM trained classification vector model and historical training set text vector of the training samples in Ni to calculate the similarity, and the removal of large page effect to some positive and negative examples according to the set threshold, a new training set W; (2) The new the training set Wi and incremental learning union training set N_j as the new training set. (3) A new training set is trained by SVM classifier. (4) After the training, the new classification vector model is saved and a new grasping task is carried out. In the process of the actual crawl, the correct classification of the training set and the increment set is improved.

Realization of SVM Topic Crawler Based on Incremental Learning

The framework is divided into the following modules: crawler module, web pretreatment module, correlation calculation module, SVM classifier module and the results display module. It is shown as follows:

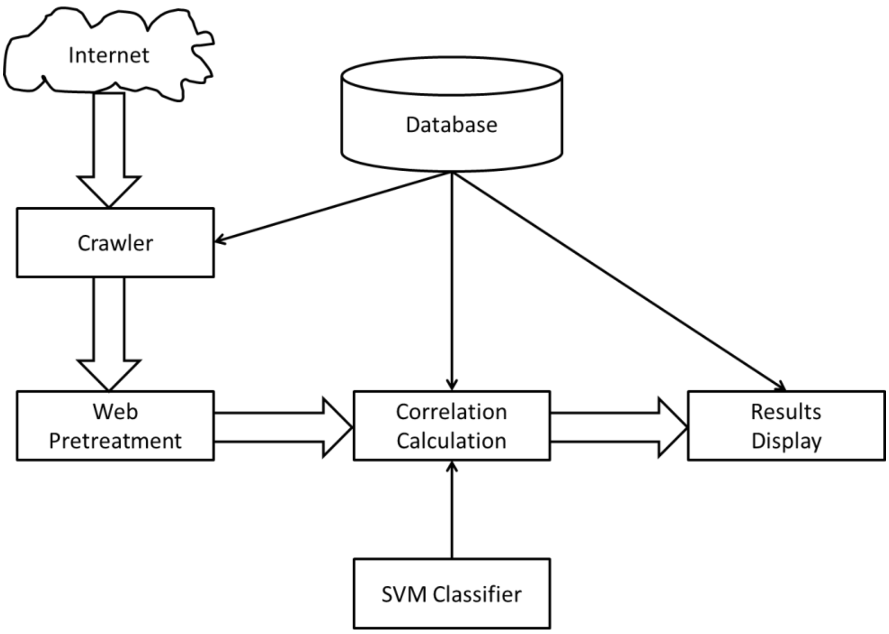

Figure 1. Framework of SVM Topic Crawler System

(1) Crawler is used to search and download web information. According to the requirement of user information, the collection of information is collected and stored in the web database to achieve more excellent results. (2) Web pretreatment module uses third chapters to introduce the technology of web page purification and the Chinese word segmentation technology, mainly to achieve the following three functions: first, the content of the text is extracted from the content, such as script information, pictures, advertising and some useless links. Second: the Chinese word segmentation processing through the word segmentation dictionary SDIC.txt load, using the dictionary based word segmentation method in the forward maximum matching algorithm, the text word processing. Feature extraction through the segmentation of text feature extraction, using TF-IDF text is represented as an n-dimensional vector. (3) The module of SVM classifier is trained to be the subject vector of the web pages, and the vector representation of the subject pages is obtained. (4) Incremental learning can improve the accuracy of the web pages, and then add some high related web pages to the training set, and constantly improve the training set of the related subjects. (5) The correlation degree calculation will be carried out by the analysis and processing of the web pages. The obtained vector model and the SVM classifier are based on the topic vector model. According to the preset threshold, the high degree of correlation is taken out, and the URL is taken out. (6) After the display module function has been associated with a high degree of relevance of the web page, the module will be the main theme of the crawler to get to the theme of the web page displayed to the user.

The operating example of the system is as follows: firstly, the subject vector model is added to the program. System interface is simple and clear, there are two main tabs, the first tab is the main parameters of the setting, threads is the number of threads, depth is the theme of the crawl depth. Site is the site of the crawl. The second tab is the main page of the URL and the URL of the related topics. Public opinion crawl, select the different characteristics of M and incremental learning on the Internet for the public opinion on the Internet, grab the theme for the "Sino Japanese dispute", thread selection 100, crawl depth is 5, respectively from Phoenix, global and Net ease network for public opinion to set the time fixed in 30 minutes, the results can be seen from the three table, the results can be seen, the threshold value of a theme crawler. Different threshold setting results are different. In the real process of grasping the opinion, we set the appropriate threshold in order to improve the number of topic web pages.

Table 1. System Program Results (Threshold=0.7)

incremental learning number	m=10			m=20			m=30		
	Download Page Number	Related Page Number	Acquisition Rate	Download Page Number	Related Page Number	Acquisition Rate	Download Page Number	Related Page Number	Acquisition Rate(%)
0	29563	8396	28.4	28892	7685	26.6	27201	7371	27.1
1	29332	9122	31.1	28401	8179	28.8	27449	7663	27.7
2	29701	9564	32.2	28523	8528	29.9	27368	8293	30.3

Table 2. System Program Results (Threshold=0.75)

incremental learning number	m=10			m=20			m=30		
	Download Page Number	Related Page Number	Acquisition Rate	Download Page Number	Related Page Number	Acquisition Rate	Download Page Number	Related Page Number	Acquisition Rate(%)
0	28491	7521	26.4	28692	6943	24.2	27006	6778	25.1
1	29147	8394	28.8	28223	7083	25.1	27422	7239	26.4
2	28803	9564	30.7	28149	7684	27.3	27138	7788	28.7

Table 3. System Program Results (Threshold=0.8)

incremental learning number	m=10			m=20			m=30		
	Download Page Number	Related Page Number	Acquisition Rate	Download Page Number	Related Page Number	Acquisition Rate	Download Page Number	Related Page Number	Acquisition Rate(%)
0	29563	7449	25.2	28874	6439	22.3	27201	6283	23.1
1	29212	7801	26.7	28401	6027	23.6	27449	6258	22.8
2	29811	8108	27.2	28523	6874	24.1	27368	6568	24

References

[1] Zhu Ting, Teng Guifa, Lu Hao, Zhang Changli, Zeng Dajun, On Adaptive Focused Crawler Based on Online Incremental Learning[J]. Computer Applications and Software, 2009(5)25-29.

[2] Gao Zhaoqiong, The Focused Web Crayviing Strategy Based on Incremental Learning [D]. Xihua University, 2010.

[3] Fan Jin, The Design and Analysis of Topic Crawler for Electronic Public Opinion [D]. Tianjin University of Science & Technology, 2014.

[4] Ye Y, Ma F, Lu Y, et al. iSurfer: A Focused Web Crawler Based on Incremental Learning from Positive Samples[J]. Lecture Notes in Computer Science, 2004.

[5] Zhang X. The Design of Hyper-sphere SVM Based on Incremental Learning[J]. Computer Engineering & Applications, 2006, 42(13):66-69.