RESEARCH ARTICLE | APRIL 18 2018

# Research on the optimization strategy of web search engine based on data mining FREE

Ronghua Chen



*AIP Conf. Proc.* 1955, 040027 (2018)
https://doi.org/10.1063/1.5033691

# Research on the Optimization Strategy of Web Search Engine Based on Data Mining

Ronghua Chen

*Jiangxi Vocational College of Finance and Economics; Jiujiang Jiangxi 332000 China*

**Abstract.** With the wide application of search engines, web site information has become an important way for people to obtain information. People have found that they are growing in an increasingly explosive manner. Web site information is very difficult to find the information they need, and now the search engine can not meet the need, so there is an urgent need for the network to provide website personalized information service, data mining technology for this new challenge is to find a breakthrough. In order to improve people's accuracy of finding information from websites, a website search engine optimization strategy based on data mining is proposed, and verified by website search engine optimization experiment. The results show that the proposed strategy improves the accuracy of the people to find information, and reduces the time for people to find information. It has an important practical value..

**Key words:** search engine; information acquisition; search accuracy; data mining; support vector machine.



## INTRODUCTION

The search engine is a computer system that provides information retrieval service with the demand of the Internet users to quickly query information after the Internet is produced [1-3]. It is an information processing system, with a certain strategy in the Internet, find information, understanding of the information, extraction, organization and processing, and retrieval services for users, which serve the purpose of navigation information, including information search, information collection and user query of three parts[4-6]. From the point of view of the user, it is a tool to help people to retrieve information, and the search engine has become one of the industries in the information field. It mainly used in information retrieval, artificial intelligence, database, data mining, natural language understanding and other fields of theory and technology, which is comprehensive and challenging[7-9].

With the rapid development of the Internet, the search engine has become an essential part of the people. The search engine, with its powerful grasping ability, has covered more than twenty percent of the web pages at the present stage. Google's Web coverage is about 6 billion, and Baidu also reaches 1 billion, and they form their own way of living. Google is known as the greatest company in the Internet. Baidu has become the second largest Internet Co in China, which has declared the great development of search era. Search engine has brought great convenience to the Internet users. On the other hand, the search of website portals is becoming more and more popular. When a website is bigger and stronger, its content is bound to be very rich. Users need to search for the content they want, or want to see the relevant content in this area, all of them need support from search engine.

The search engine is not in a single point, in the case of Baidu, is not only a natural result, show the results of various structured appeared. For example, Liao Fan found, will immediately give encyclopedia, movies, pictures, know, micro-blog, news and other information, which is based on the user logs on, the best search results are given, each of the results which are better than the simple nature, which is one of the many products and program personnel who work hard the results of. Search the movie two words, a very large space of search results appear in front of the user, this is the quality of display technology. The latest, hottest, good users, all kinds of movies you choose. The search team World Cup race, live pregame entrance, and statistics game aspects are covered in this little show in quality. These are all the basic Rank strategies can't reach.

*Advances in Materials, Machinery, Electronics II*
AIP Conf. Proc. 1955, 040027-1–040027-7; https://doi.org/10.1063/1.5033691
Published by AIP Publishing. 978-0-7354-1654-3/$30.00

When users may be interested in only one aspect of knowledge and are not interested in other knowledge, they need special search. For example, users are only interested in news and thematic contents, and news topic search is obviously better able to satisfy users. On the efficiency of search engines due to the special type of content, ratio of web search is more accurate and have received much attention, such as Baidu news, covering the Internet, domestic and international, military, financial and other aspects of the news, users only need to input their interest in the news topic or content, list of major news sites the news of the news immediately. If users are just searching for no purpose, just click on the topic search website, and all kinds of news hotspots are listed. These results are carefully selected by professionals combined with people's interest points, making it difficult for users to leave the page. Baidu knows that it solves a wide variety of problems. As we all know, the Internet contains a variety of knowledge, but how to extract it is a problem that has been difficult to solve. If you pass the basic Rank strategy alone, you may get a lot of related content, but the answer is absolutely not guaranteed.

In order to improve people's accuracy of finding information from websites, a website search engine optimization strategy based on data mining is proposed, and verified by website search engine optimization experiment. The results show that this strategy improves the accuracy of the people to find information, and reduces the time for people to find information. It has an important practical value.

## OVERVIEW OF SEARCH ENGINE OPTIMIZATION

Search engine optimization is based on the research of search engine, search ranking rules, and use the search engine search and ranking rules to improve web page ranking method in the search engine is through a special way to optimize and update the site, make the site keywords in the search engine ranking on the relative, to strive for search users click to enter, so as to improve the user traffic website. Its main work is divided into internal optimization and external optimization. Internal optimization refers to the adjustment of basic elements of web pages. External optimization refers to how to increase external links, and the ultimate purpose of optimization is to increase website visits and enhance website.

For a website, there are many factors that can affect its ranking in search engine. There are many important factors that can be used to optimize and influence ranking, such as website domain name, keyword, website structure content and link.



### Web site domain name

The domain name itself also has the weight, generally the nonprofit organization section type website such as: edu and so on suffix domain name is higher than com and so on suffix the domain name weight is high, but the domain name com is higher than the domain name CN weight. The influence of website name selection on search results is mainly reflected in two points: the correlation between domain name and website content and domain name are identified by search engine. The degree of correlation between the name of the domain name and the content of the website mainly shows whether the content of the site can be reflected in the name of the site. Again, don't buy the domain names that have been punished by the search engine.

### Key words

Because of the potential customers or target site users are the specific words or sentences in a search engine to find and access to the site, and the search engine ranking in search of these words or sentences and plays a key role, so named keywords. It can also be seen that in the whole link of search engine optimization, the location and analysis of key words is the most important and at the core. If you choose too popular keywords, may be a waste of manpower, time and promotion expenses, is not easy to get good rankings; if you choose popular keywords, even get good rankings, although can give site bring certain flow, but it is difficult to improve the conversion of potential customers; if you choose keywords are not accurate, it may bring to the site some garbage flows, and even increase the burden of the web server, affect the browsing speed, and will seriously affect the subsequent optimization promotion of all search engines. Therefore, the study of keywords has become the top priority of search engine optimization. The proper selection of keywords and search engine optimization will be much easier.

## Website structure

The structure of a web site refers to the hierarchical relationship between the pages of a web site, which can be divided into logical and physical structures according to their nature. Website structure plays a key role in deciding the weight of a page, which directly affects search engine's collection of pages. Reasonable website structure can effectively guide search engines to grab more and more valuable pages.

The optimization of website structure is a reasonable adjustment of the physical structure and internal link relationship of web pages, in order to reduce the depth of page directory and link depth between important pages. At the same time, add the link entrance of the important page to improve the record and weight of the pages that are included by the search engine.

# DATA MINING

## Data mining technology

Data mining generally refers to the process of searching the information from a large number of data through algorithm search. The concrete steps of data mining are: determine and define the problem. Before data mining, it is necessary to determine what the data mining is done to the search, what the expected effect is, and what the place looks like. Determining and defining problems is a key part of data mining, which will reduce the modification of later development. For example, we want to determine which users are interested in the news and what are interested in the game. So our requirement is to analyze the content of each user for a long time, based on the user's behavior log. Set up a data mining library, that is, to build a data mining library that can be used. It is generally divided into the following steps:

(1) The collection of a large number of credible data;
(2) The description of the demand data;
(3) Select the part that may be used in large data.
(4) To clean up the data quality assessment and messy data.
(5) The merger of the same data;
(6) The construction of authentic and credible initial data;
(7) Data loading into the database.

Determine the data requirements. Analyze the data in the database, find fields that may have a larger impact on the output, and define the fields that need to be exported.

(1) Prepare the data. This part not only establishes the data that needs to be prepared, but also needs to determine the final results after the data mining. This ensures a reasonable feedback on the data results and determines whether the data mining model can meet the requirements.

(2) Set up a data mining model. When building a model, you need to think carefully which model can solve the problem we need. Then, after some data to build data mining model, and use the remaining data to verify the model is correct. Training and testing data need to be divided into two parts: one is training for data mining, one is executable model, and the other is completely used for model testing.

(3) Evaluation model. After the model is established, the model data is analyzed by step step step by step, and the accuracy of the model is obtained. The evaluation model needs to consider all kinds of problems that may be encountered in practical application. It can try out mining models in a small range, and then extend it in a wide range after satisfaction.

With the changing needs of people, it is likely that a period of time, the model may need to be rebuilt. At this time, a new test of the model is required to determine whether it is valid. If it fails, it is possible to reestablish the mining model. Data mining brings a new field to the search engine by analyzing the most interesting aspects of the user from the user log. For different users, the search engine can give completely different search results.

## Support vector machine(SVM)

The choice of the key words determines the location of your website. The most intuitive Market a demand survey is the number of times a particular keyword on a search engine is searched. The more the main key words that are related to the product are searched for the market demand the greater the user's attention is, the higher. At present, the search engine usually uses a fixed weight algorithm to determine the importance of the retrieved pages, and the results

are important. The level of sex is given. And the Support vector machine can give variable ranking weights for each user, so that everyone's search is done. The results of the cable will be completely different, and the neural network can give the user more accurate results suitable for themselves, no doubt very much good to satisfy the user.

Let the given training samples be represented by (xi, yi), i=1,2…,n, where xi∈Rd is an input vector, yi∈R is its corresponding desired output, and n is the number of training data. In SVM, the original input space is mapped into the high dimensional space called feature space by nonlinear mapping x→g(x), Let f(x) be the output of SVM corresponding to the input vector x, in the feature space, then a linear function is constructed:

$$f(x) = \omega^t g(x) + b \tag{1}$$

Where, ω is a coefficient vector, b is a threshold.

SVM learning can be obtained by the minimization of the empirical risk on the training data, and the ε-intensive loss function is used for the minimization of empirical risk. The loss function is defined as:

$$L^{\varepsilon}(x, y, f) = \left|y - f(x)\right|_{\varepsilon} = \max(0, \left|y - f(x) - \varepsilon\right| \tag{2}$$

Where, ε is a positive parameter that allows approximation errors smaller than ε. The empirical risk is

$$\mathrm{R}_{emp}(\omega) = \frac{1}{n}\sum_{i=1}^{n} L^{\varepsilon}(y - f(x_i)) \tag{3}$$

Other than the $\varepsilon$-intensive loss, SVM tries to reduce the model complexity by minimizing $|\omega|^2$. This can be described by slack variables $\xi_i$ and $\xi_i^*$, which measure training data $x$i whose deviations exceed the constant $\varepsilon$. Subsequently, the SVM approximation is obtained as the following optimization problem.

$$\begin{aligned} &\min \frac{1}{2}|\omega|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &s.t. \\ &\begin{cases} y_i - f(x_i) \le \varepsilon + \xi_i \\ f(x_i) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \end{aligned} \tag{4}$$

Where, $C$ is a positive constant to be regulated.

By using the Lagrange multiplier method, the minimization of formula (4) causes the problem of maximizing the following dual optimization

$$\begin{aligned} &\max \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) - \varepsilon\sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) \\ &s.t. \\ &\sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0, C \ge \alpha_i^*, \alpha_i \ge 0 \end{aligned} \tag{5}$$

Where, $\alpha_i^*$ and $\alpha_i$ are Lagrange multipliers, and $K(x_i, x_j)$ is a kernel function, which is equivalent to the dot product in the feature space, namely

$$K(x_i, x_j) = g(x_i)g(x_j) \tag{6}$$

Here the Gaussian function is used as kernel

$$K(x,y)=\exp(-\frac{\|x-y\|^2}{2\sigma^2}) \tag{7}$$

Where, $\sigma$ is kernel parameter.

By replacing $\beta_i=\alpha_i^*-\alpha_i$ and relation $\alpha_i^*\alpha_i=0$, the optimization of formulae (5) is rewritten as

$$\begin{aligned}&\max \sum_{i=1}^{n} y_i\beta_i-\varepsilon\sum_{i=1}^{n}|\beta_i|-\frac{1}{2}\sum_{i=1}^{n}\beta_i\beta_j k(x_i,x_j)\\&s.t.\\&\sum_{i=1}^{n}\beta_i=0,-C\le\beta_i\le C\end{aligned} \tag{8}$$

The learning results for training samplesD can be derived from Eq. (8). Note that only some of the coefficients βi are not zeros and the corresponding vectors z are called support vectors (SV). That is, vectors xi whose corresponding coefficients $\alpha_i^*-\alpha_i$ are not zero are SV. Then the approximation function is represented by Lagrange multipliers, namely

$$f(x)=\sum_{i=1}^{P}(\alpha_i^*-\alpha_i)k(x_i,x_j)+b \tag{9}$$

Step1: the data of web search engine optimization are collected and divided into the training samples and validation samples.

Step2: the range of parameters of SVM which are C and σ are selected.

Step3: The initial parameters are produced randomly, which represented parameters of SVM.

Step4: The training samples are input into SVM to train and establish the web search engine optimization prediction mode.

## SIMULATION EXPERIMENTS

### Web search engine optimization data

In order to avoid web search engine optimization data in greater numeric ranges dominating those in smaller numeric ranges and avoid numerical difficulties during the SVM training. Generally, web search engine optimization data are scaled to the range [0, 1] by formula (10).

$$x_i^{'}=\frac{x_i-\min(x_i)}{\max(x_i)-\min(x_i)} \tag{10}$$

Where, xi is original value, xi' is scaled value, max (xi) and min (xi) represent the max and the min value of web search engine optimization data separately.

### Results and analysis

The simulation results as shown in Fig.1 and fig.2, from Fig.1 and fig.2, in all the strategies, the search engine optimization strategy of information search and information search is highest at the correct rate, error rate is the lowest,

which indicates that the search engine optimization strategy can achieve very accurate identification of categories of information, obtain the ideal in the search engine optimization the results also can be seen from Fig.3, the search engine optimization to reduce working time, can satisfy the requirement of real-time search engine optimization, because the Support vector machine can give variable ranking weights of each user in the proposed method, The results of the cable will be completely different, and the neural network can give the user more accurate results suitable for themselves, no doubt very much good to satisfy the user, so search engine optimization strategy the superiority is very obvious.

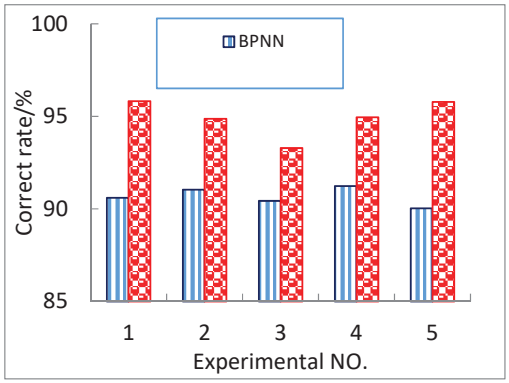


**FIG.1** correct rate of information search

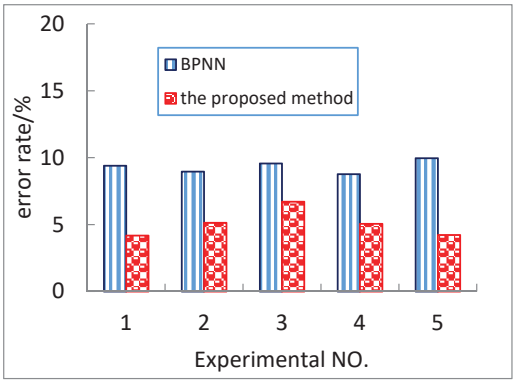


**FIG. 2** error rate of information search

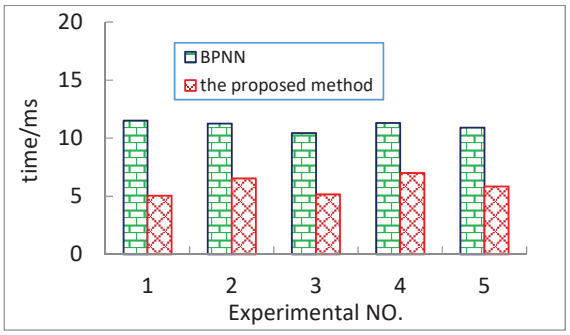


**FIG.3** the average time for information search

## ICONCLUSION

With the advent of the information age, the demand for information is becoming more and more common and widely used in the traditional sense. The sources of information are mainly libraries, newspapers, magazines and books, which are far from being able to meet this growing growth. Since the birth of Internet, people have found a more convenient and higher information dissemination and access. With the arrival of twenty-first century, Web information has become an important way for people to obtain information. When Web information becomes an important way for people to obtain information, it is found that the increasing explosion is increasing. It is difficult to find the information you need in the long Web information, which sets off the network surpassing. The convenience brought by time and space takes a lot of time. Therefore, there is an urgent need for a tool. To solve this problem, the search engine came into being. The search engine greatly improves the Web information to the user efficiency, people no longer worry about looking for information like look for a needle in the ocean. With the development of information processing technology, the development trend of the information on the website is magnanimous, traditional methods cannot adapt to the development requirements, this paper proposes a search engine optimization strategy based on Web data mining, specific experimental results show that this strategy greatly shortens the website search engine optimization time, improve the website search engine optimization efficiency, improve information search accuracy, provides an important research tool for website search engine optimization.

## ACKNOWLEDGMENTS

Project Funded by Jiangxi Provincial Education Department: “Research on the Optimization Strategy of Web Search Engine Based on Data Mining” (No. GJJ161399).

## REFERENCES

1. Robert Cooley, Pang-Ning Tan, Jaideep Srivastava. Discovery of interesting usage patterns from web data [M]. Germany: Springer-Verlag, 2000.
2. Andrea Garratt, Mike Jackson, Peter Burden, Jon Wallis. A Survey of Alternative Designs for a Search Engine Storage Structure [J]. Information and Software Technology, 2001,43(11):661-677
3. Pal S K, Talwar V, Mitra P. Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions[J]. IEEE Transactions on Neural Networks, 2002, 13(5):1163-1177.
4. Huang Yu, Wang Hong, Xu Duanyi, et al. Search engine technology and new development of multi search engine system [J]. Computer Engineering, 2002, 33(1): 4-5.
5. Clement Yu, Weiyi Meng, Wensheng Wu, King-Lup Liu. Efficient and Effective Met search for Text Databases Incorporating Linkages among Documents [J].ACM SIGMOD, 2001, 8(2):187-193.
6. Li Yong Ping, the Kunmei. Lead climbing in integrated search result sorting optimization analysis [J]. Huazhong University of Science and Technology newspaper (Natural Science Edition), 2003, (11): 28-30.
7. Song Aibo, Dong Yisheng, Chen Jing. Based on Weblog Research on pattern discovery and Application [J].Small micro Type computer system, 2002, 23 (11): 1331-1335.
8. Jiang Ping, Cui Zhiming. Analysis and Research on user interest model in intelligent search engine [J]. Microelectronics and planning Computer, 2004, 2 1(11):24-26.
9. Cho Li, road wave. Search engine optimization strategy research [J]. Productivity research, 2010, (7): 118- 119.