African Journal of Agricultural Research Vol. 5(13), pp. 1536-1538, 4 July, 2010
Available online at http://www.academicjournals.org/AJAR
ISSN 1991-637X ©2010 Academic Journals

Mini Review

Research on agricultural search engine optimization

Wang Daoping[1], Wang Ying[1], Liu Guangli[2]*, Shen Cuihua[2] and Liu Tong[2]

[1]University of Science and Technology Beijing, Beijing, 100083 China.
[2]China Agricultural University, Beijing, 100083 China.

Accepted 7 June, 2010

Search engine optimization (SEO) has practical significance for promoting farmers income and agricultural efficiency in China. Firstly, how to extract web page attributes contributed to the ranking in search engine is considered. And the attribute extractor in Java platform is built. Then, a batch gaining method noted AAA is proposed independent of Search Engine API by which a downloader is also designed. Third, a new kernel principal component analysis (KPCA) method is proposed to rank these agricultural web pages on keywords, in which the non-linear combinations of search engine ranking factors can be obtained. By adjusting the kernel function and its parameters in order to ensure maximum contribution rate of variance. Fourth, the software system is developed for agriculture to provide decision support for search engine marketing. Data experimental results show that our method has a good performance.

Key words: Search engine optimization (SEO), kernel principal component analysis (KPCA), attributes extraction.

INTRODUCTION

Search engine optimization technology is a new network marketing and in-depth researched by scholars (Wang and Peng, 2007; Zhang, 2005). Agricultural SEO plays an important role for improving agricultural products and brands marketing by increasing search engine ranking of agriculture-related information.

Based on the law of search engines pages crawl and index, SEO is the technology to promote the web search engine ranking and website traffic and ultimately marketing for products, by reasonable adjusting to the site structure and optimizing to web page content and elements including the title and description etc (Lin, 2009; Gao, 2008).

China government has worked out several policies for agricultural problems to promote new rural construction. SEO is to promote agricultural efficiency in agriculture, rural development and an important means for farmers to increase income. Therefore, how to improve structure of agricultural site in order to achieve the promotion of agricultural products and e-commerce is an important issue.

AGRICULTURAL WEB PAGE ATTRIBUTES EXTRACTION

The factors affecting the ranking in search results page can be classified keywords- relevant and keyword-independent, but also can be divided into for pages within or outside (Sergey and Lawrence, 1998; Li and Hao, 2003). For agriculture SEM, the authors firstly prepared 160 agriculture-related keywords. Then nine keywords and three categories are selected from 160 keywords. The authors have adopted the idea of Baidu: the higher user attention and the more times keywords search, the higher concern degree. Thus the authors refer to Baidu index and determine nine keywords in higher concern degree which are divided into three categories, as shown in Table 1.

The authors proposed an algorithm noted AAA which principle is to summit http request to search engines by simulation browser, and obtain the pages containing search results, and then the corresponding search structure by parsing, matching is obtained. Http request can be divided into Get and Post. Get format is simpler than Post for URL including most of its information. We adopt Get request by the class HttpURLConnection in Java and

*Corresponding author. E-mail: liugl@cau.edu.cn.

Table 1. Concern degree of keywords.

Categories	Keywords	Baidu concern degree
Agricultural materials	Fertilizer	230
	Phosphate	111
	Potash	182
Agricultural machinery	Oilpress	535
	Milking Machine	96
	Bander	93
Flowers	Rose	1049
	Lily	841
	Carnation	1034

set the property user-agent to achieve the effect of analog Browser. The idea of matching search results is to find label before search result which can be separate from other text. In fact, it can be realized by Element Review in Google Chrome. Search engine returns a page that contains the required in addition to the search results, it also contains a number of other search engines add pages, as Google would add focus, music and pictures.

The aim was to obtain a sample set of web pages and extract the required attributes and conduct further ranking analysis. The system has three parts: Downloader, Feature Extractor and Database. Downloader's input is the pages on the keywords inputted by user to engine submission. Its work flow is as follows:

1) Encode the keywords format acceptable for search engines (Google for tUTF-8 and Baidu is GB2312), generate the appropriate URL format and submit requests to http. Extract the titles and URL of search results.
2) According to URL, Downloader saves these pages locally.

As the core of system, these pages are input of Feature Extractor. Based on attribute extract method above, we can get the attributes by analyzing the source code of pages. Microsoft SQL Server 2008 is adopted as data management tool in our system to save and manage the attributes. For agricultural SEM, ten tables are built and each table has 100 records.

KERNEL PCA FOR AGRICULTURAL SEARCH RANKING

The natural ranking rule of search engine is a multi-factor and multi-target optimum selection and scheduling problem. As the core of search engine, ranking algorithm is commercial secret. Principal component analysis (PCA) is a selection method to resolve search engine ranking which may have a problem of low contribution rate for a single index. Kernel Principal Component analysis using nonlinear combination can overcome the problem (Berhard et al., 1998; Albert and Castillo, 2005). Suppose the number of webs on one keyword is l. And p indexes are gotten by downloader above for each page, construct a vector $x=(x_1,\ldots, x_p)^T$. New vector is $y=(y_1,\ldots,y_p)^T$ by linear transformation noted by $Y=AX$. And $A=(a_{ij})$ $(i,j=1,\ldots,p)$ is orthogonal matrix. Let variance of y_i beλ_i $(i=1,\ldots,p)$. We have $\lambda_1>\lambda_2>\ldots>\lambda_p>=0$ due to independent for each other. Then the index y_i is i-th principal component. Let sample data vector be x_k□R.p $(i=1,\ldots,p, \sum x_k=0)$ and C= $(\sum x_j\, x_j^T)/l$. And linear PCA is to solve feature value and vector of C.

Now we extend PCA to nonlinear situation: kernel PCA. In fact, nonlinear PCA may be considered as doing PCA in high dimensional space by nonlinear transformation (Muller et al., 2001; Twining and Taylor, 2003). Let $D=(\sum\varphi(x_j)\varphi(x_j)^T)/l$, then we have $\lambda V=DV$ where λ is feature value and V is feature vector of D. Define matrix $K=(k_{ij})$ and $k_{ij}=(\varphi(x_i)\varphi(x_j)^T)$. We have $l\lambda K\alpha=K^2\alpha$ where we only solve $l\lambda\alpha = K\alpha$. Principal Component extraction for one test point $\varphi(x)$ can be computed by its projection $(V^k\cdot\varphi(x)) = \sum\alpha_i^k(\varphi(x_j)\,\varphi(x_j))$. Let matrix $K'=K-1_lK-K1_l+1_lK1_l$ instead of K where $(1_l):=1/l$ and $\sum\varphi(x_k)=0$ would be gotten. In fact, different kernel such as RBF and Sigmoid can be adjusted to a suitable contribution rate of variance.

The software system has four components: data loading, data pre-processor, data analyzer and results memory. In our data test, we use Gaussian kernel function, where σ = 999999, contribution rate contributes over 95% for one principal component. The result of test is shown in Table 2. In Table 2, Precision1 means the precision of the top ten Google rankings. For top 100 search simulation ranking results, as long as top ten under simulated results is same to real ranking we would think it accurate. And precision 2 is floating scope precision. That is, in top 100 results, if situation result for one page is within 5%, it is accurate.

Table 2. Test results.

Keywords	Contribution rate	Precision 1 (%)	Precision 2 (%)
Fertilizer	0.97861	10	18
Phosphate	0.99025	40	57
Potash	0.98427	20	15
Oilpress	0.98203	20	24
Milking machine	0.98601	30	18
Rose	0.99173	20	15
Lily	0.99021	10	26
Carnation	0.99184	20	25

Conclusions

Agricultural SEO has certain practical significance regarding the farmer additionally receiving agriculture efficiency. A new attribute extraction method and system for agricultural SEO is introduced in this paper. Data test shows the algorism AAA is feasible. And a new agriculture search ranking method: Kernel PCA is presented. Data results show that the method has certain theoretical and practical significance.

ACKNOWLEDGEMENT

Supported by “Eleventh Five-Year” National Science and Technology Support Program funded projects (2008BADA8B01-2, 2006BAJ07B09) and Special Funds for Basic Scientific Research Project of CAU (15059110).

REFERENCES

Albert B, Carlos C (2005). An Analysis of Factors Used in Search Engine Ranking [J]. In Proc. of the Workshop on Adversarial IR on the Web. April 8.

Berhard S, Alexander S, Klaus- Robert M (1998). Nonlinear Component Analysis as a Kernel Eigen Value Problem [J]. Neural Comput., 10(11): 1299-1319.

Gao A (2008). Talking about writing search engine optimization techniques [J]. Info. Sci., (11): 49-49.

Li K, Hao F (2003). Page Rank-Pro: An Improved Page Rank Algorithm. [J]. J. Jilin Univ. 4.

Lin Y (2009). Simple Analysis on application of search engine optimization technology [J]. Software Guide, 8(11): 147-149.

Muller KR, Mika S, Ratsch S (2001). An introduction to kernel-based learning algorithms [J]. IEEE Trans on Neural Network, 122, 12(2): 181.

Sergey B, Lawrence P (1998). The anatomy of a large-scale hyper textual web search Engine. seventh international World-Wide Web Conference (WWW 1998).4

Twining C, Taylor C (2003). The use of kernel principal component analysis to model data distributions [J]. Pattern Recognition, 361, 36(1): 217-227.

Wang J, Peng J (2007). Research on the structural design of Web crawler [M]. Technol. Info.

Zhang W (2005). Simple Analysis on application of search engine optimization technology [D]. Chengdu: Sichuang University.