Mini Review

Research on agricultural search engine optimization

Wang Daoping¹, Wang Ying¹, Liu Guangli², Shen Cuihua² and Liu Tong²

¹University of Science and Technology Beijing, Beijing, 100083 China.
²China 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
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
thttp. 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.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Keywords</th>
<th>Baidu concern degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural materials</td>
<td>Fertilizer</td>
<td>230</td>
</tr>
<tr>
<td></td>
<td>Phosphate</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>Potash</td>
<td>182</td>
</tr>
<tr>
<td>Agricultural machinery</td>
<td>Oilpress</td>
<td>535</td>
</tr>
<tr>
<td></td>
<td>Milking Machine</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>Bander</td>
<td>93</td>
</tr>
<tr>
<td>Flowers</td>
<td>Rose</td>
<td>1049</td>
</tr>
<tr>
<td></td>
<td>Lily</td>
<td>841</td>
</tr>
<tr>
<td></td>
<td>Carnation</td>
<td>1034</td>
</tr>
</tbody>
</table>

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 \( x=(x_1, \ldots, x_l)^T \). 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 \( y_i \)
be \( \lambda_i \) \ (\( i=1, \ldots, p \)). We have \( \lambda_1>\lambda_2>\ldots>\lambda_p \geq 0 \) due to
independent for each other. Then the index \( y_i \) is i-th
principal component. Let sample data vector be \( x_i \in \mathbb{R}^p \)
\( (i=1,\ldots, p, \sum x_i=0) \) and \( C=(\sum x_i x_i^T)/l \). 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 \( D=(\sum \phi(x_i)\phi(x_j))/l \), then we have \( AX=DV \) where \( \lambda \) is feature
value and \( V \) is feature vector of \( D \). Define matrix \( K=(k_{ij}) \) and \( k_{ij}=(\phi(x_i)\phi(x_j))^T \). We have \( l\lambda K=K\lambda \) a where we only solve \( l\lambda = K\lambda \). Principal Component extraction for one
test point \( \phi(x) \) can be computed by its projection \( (V^\lambda \phi(x)) = \sum a_i \phi(x_i) \phi(x_j) \). Let matrix \( K=K_1+K_1L+K_1 \) instead of \( K \) where \( (1):=1/l \) and \( \sum \phi(x_i) =0 \) 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 \( \sigma = 999999 \), 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 1. Concern degree of keywords.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Keywords</th>
<th>Baidu concern degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural</td>
<td>Fertilizer</td>
<td>230</td>
</tr>
<tr>
<td>machines</td>
<td>Phosphate</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>Potash</td>
<td>182</td>
</tr>
<tr>
<td>Agricultural</td>
<td>Oilpress</td>
<td>535</td>
</tr>
<tr>
<td>machinery</td>
<td>Milking Machine</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>Bander</td>
<td>93</td>
</tr>
<tr>
<td>Flowers</td>
<td>Rose</td>
<td>1049</td>
</tr>
<tr>
<td></td>
<td>Lily</td>
<td>841</td>
</tr>
<tr>
<td></td>
<td>Carnation</td>
<td>1034</td>
</tr>
</tbody>
</table>
Table 2. Test results.

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

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 algorithm 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


