

## Attribute Extraction System for Agricultural SEM

Wang Ying<sup>\*1</sup>, Liu Guangli<sup>\*2</sup>, Bai Shengli<sup>\*3</sup>, Yang Zhimin<sup>\*4</sup>

<sup>\*1</sup> *University of Science and Technology Beijing, Beijing, 100083, China*

<sup>\*2, Corresponding author</sup> *China Agricultural University, Beijing, 100083, China*

<sup>\*3</sup> *Cable TV Group Shangqiu in Henan Branch, Shangqiu, 476000, China*

<sup>\*4</sup> *Zhejiang University of Technology, Hangzhou, 310024, China*

*alyce@126.com, liugl@cau.edu.cn, bsl2008@126.com, yzm9966@126.com*

doi: 10.4156/jcit.vol5.issue3.3

### Abstract

*As a rapidly booming industry in recent years, the key problem for Search Engine Marketing (SEM) is to extract web page attributes contributed to the ranking in search engine such as Google and Baidu. The quantitative analysis of these agricultural page attributes is aimed in this paper and the attribute extractor in Java platform is built. Meantime, in order to gain a sample set, a batch gaining method named AAA independent of Search Engine API is introduced and a downloader in Java platform is also designed. The experiment of batch gaining and attributes extracting from samples about certain keywords indicate that our method is effective and has a good performance.*

**Keywords:** *Search engine marketing, Attribute extraction, Agricultural marketing.*

## 1. Introduction

Based on the law of search engines pages crawl and index, search engine optimization (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. And SEM is aimed to promote web page visitors to buy the products in future<sup>[1]</sup>. Currently SEM methods include two types: pay per click (PPC) and SEO.

Agricultural SEM can be defined to promote ranking in search engine and marketing of agricultural web site for certain agriculture-related keywords. China government has worked out several policies for three agricultural problems to promote new rural construction. SEM is to promote agricultural efficiency in agriculture, rural development and an important means for farmers to increase income<sup>[2]</sup>. Therefore, how to improve structure of agricultural site in order to achieve the promotion of agricultural products and e-commerce is an important issue.

## 2. Web page attributes extraction method

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<sup>[3]</sup>.

Keywords-relevant pages attributes are shown in Table 1.

Keywords-independent pages attributes are shown as following: Code of File, Length of Title, Length of Keywords, Length of Description, Length of Body, Internal Link Count, External Link Count, Invalid Internal Link Count, Invalid External Link Count, HTML Validation of Document (to W3C Standards), and Size of Page. And keywords-independent outside pages attributes include: Frequency of Updates to Page, Length of URL, Number of Trailing Slashes (/) in URL, Size of Site, TLD Extension of Site (edu, gov, us, ca, com, etc), Age of Site, Age of Page, Page Rank, Rate of New Pages Added to Site, Rate of New Inbound Links to Site, the number of Into links, site map, and Web traffic. Based on above attributes, we had built the database by SQL Server 2005.

Simplified Chinese web pages are mainly three kinds of encoding formats<sup>[4]</sup>: UTF-8, GB2312, and GB18030 in our system. We use the Mozilla's Java transplant known as Jcharset. Then we build the Html parser in which the path for local files, file coding and keywords search need to be initialized. To improve the efficiency of bulk access to attributes and save the system output, Html parser used one

time scanning algorithm to get a good input stream of web page samples downloaded, one by one character scanning, each sub-label by the corresponding function of processing, thereby extracting the necessary page attributes. In this way, handling with entire page text turned to each sub-label treatment. Dealing sub-tab is mainly divided into Title tags, META tags and Body tags. The main content to deal with Title label is to obtain its contents. In view of the length of Title is usually not very long, we use a String object to save its contents. And the idea of getting attributes about Keywords and Description is same to Title. For Body, the processing is more complex as there are many sub-labels needs to be deal with in <Body>...</Body>, such as <Hx>, <Strong>/<B> and <img> <a> and so on. Therefore, we use marking method in which the appropriate tag is marked by +1 meeting a certain label, and ending at -1. Scanning text displayed we detect the tags, if one is not 0 then adding it to the appropriate buffer. In fact our system does not depend on Google SOAP API or Ajax API to get PR value but by Google Toolbar. At last, in-link number is available through Google's link: prefix achieved. After submitting the Google in the form of link: [URL], Google will return the page containing the URL in-link number.

**Table 1.** Keywords-relevant pages attributes

<i>No.</i>	<i>Attribute</i>	<i>Description</i>
1	Keyword Use in Page URL	Whether the URL contains keywords
2	Keyword Use in Title Tag	The number of specified search keywords contained in Title of Head
3	Keyword Use in Description Tag	Keyword number in description section
4	Keyword Use in Keywords Tag	Keyword number in keywords section
5	Keyword Use in Body Text	Keyword number in body text
6	Keyword Density in Body Tag	Keyword number in body tag
8	Keyword Use in H1 Tag	Keyword number in H1 tag
9	Keyword Use in H2 Tag	Keyword number in H2 tag
10	Keyword Use in H3 Tag	Keyword number in H3 tag
11	Keyword Use in H4 Tag	Keyword number in H4 tag
12	Keyword Use in H5 Tag	Keyword number in H5 tag
13	Keyword Use in Bold/Strong Tags	The number of keywords displayed in bold
14	Keyword Use in Alt Tags and Image Titles	Keywords number in alt tags and image titles
15	Keyword Use in Internal Anchor Text	Keywords number in internal anchor text
16	Keyword Font	The largest font size in the body

### 3. Web page bulk access algorithm based on keywords

For agriculture SEM, we firstly prepared 160 agriculture-related keywords. Then nine keywords and three categories are selected from 160 keywords. We have adopted the idea of Baidu: the higher user attention and the more times keywords search, the higher concern degree<sup>[5]</sup>.

Thus we refer to Baidu's index (<http://index.baidu.com>) and determine nine keywords in higher concern degree which are divided into three categories, as shown in table 2.

On April 11, 2002, Google published a Web API's beta test version followed SOAP protocol and WSDL rules. After users applied for a serial number, he and she can use the Google Web API application development in .NET or Java platform and enjoy a certain keyword search results, even access to Google cache copy. But Google SOAP API did not support unfriendly to the Chinese keywords, we do not consider this algorithm. Meanwhile, Google Ajax Search API instead of SOAP API is not considered in our system as serial number with limited times is needed too. And Baidu has not open API<sup>[6]</sup>.

Thus we proposed an algorithm noted AAA which principle is to summit http request to search engines (Google or Baidu) by simulation browser, and obtain the pages containing search results, and then by parsing, matching obtain the corresponding search structure. For Chinese Google, its search URL format is in 'http://www.google.cn/search?q=keywords ( UTF-8 ) ] &hl= zh-CN&lr =&as\_qdr=all&num= [per page displayed number] &newwindow=1&start=[start index] &sa=N'. Pay attention to converting Chinese keywords to UTF format, such as Fertilizer should be in

‘%E5%A4%8D%E5%90%88%E8%82%A5’. Per page displayed number has four selections 10, 20, 50 and 100 in Google. And start index means to start from 0 and end in N.

**Table2.** Concern degree for Agricultural SEM

<i>Categories</i>	<i>Keywords</i>	<i>Baidu concern degree</i>
Agricultural Materials	keywords	Baidu concern degree
	Fertilizer	230
	Phosphate	111
	Potash	182
Agricultural Machinery	Oilpress	535
	Milking Machine	96
	Bander	93
Flowers	Rose	1049
	Lily	841
	Carnation	1034

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<sup>[7]</sup>.

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. These results need to filter. See table 3.

**Table 3.** Search results needing to filter

<i>Filter Item</i>	<i>URL head</i>
Picture	http://images.google.cn/images?
News	http://news.google.cn/news?
Map	http://ditu.google.cn/maps?
Video	http://video.google.cn/videosearch?
Music	/url?q=http://www.google.cn/music/
Blog	http://blogsearch.google.cn/blogsearch?

Downloader. 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.

Feature Extractor. 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.

Database. 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 100records.

## 4. Conclusions

Agricultural SEM has certain practical significance regarding the farmer additionally receiving agriculture efficiency. A new attribute extraction method and system for agricultural SEM is introduced in this paper. Data test shows the algorism AAA is feasible.

Next research would aim to propose a real-time quantitative evaluation method for ranking in search engine, Kernel Principal Component Analysis (KPCA), and realize the software system. Based on the KPCA algorithm, the idea is to obtain nonlinear combination of features, and by adjusting kernel function and the parameter to guarantee the biggest factor on higher contribution. In fact, we would

build a sub-copy to Google or Baidu in agricultural keywords though the algorithm of ranking can not be known. But our goal is to agricultural marking not others<sup>[8]</sup>.

## 5. Acknowledgement

This paper is supported by “Eleventh Five-Year” National Science and Technology Support Program funded projects (2008BADA8B01-2, 2006BAJ07B09) and supported by Chinese Universities Scientific Fund (2009-1-99) and National Natural Science Foundation (10926198).

## 6. References

- [1] J. Wang and J. Peng, Research on the Structural Design of Web Crawler. Beijing: Technology Information, 2007.
- [2] W. Zhang. Simple Analysis on Application of Search Engine Optimization Technology. Chengdu: Sichuan University, 2005.
- [3] Y.Y. Lin, "Simple Analysis on Application of Search Engine Optimization Technology", Software Guide, Vol. 8, No.11, 2009, pp.147.
- [4] A.Y. Gao, "Talking About Writing Search Engine Optimization Techniques", Information Science, Vol. 11, 2008, pp.49.
- [5] B. Sergey, P. Lawrence, "The Anatomy of a Large-scale Hyper Textual Web Search Engine", Seventh international World-Wide Web Conference, 1998.
- [6] K. Li and F.L. Hao, "PageRank-Pro: An Improved Page Rank Algorithm", Journal of Jilin University, Vol. 21, 2003, pp.12.
- [7] B. Albert, C. Carlos, "An Analysis of Factors Used in Search Engine Ranking", First International Workshop on Adversarial Information Retrieval on the Web, 2005.
- [8] Hyung Tae Kim, Sang Bong Kim, Jong Sik Go, Yang Dam Eo, Byoung Kil Lee, "Building 3D Geospatial Information using Airborne Multi-Looking Digital Camera System", JCIT: Journal of Convergence Information Technology, Vol. 5, No. 1, pp. 15 ~ 22, 2010.