RESEARCH ON AGRICULTURE DOMAIN META-SEARCH ENGINE SYSTEM

Nengfu Xie [*] , Wensheng Wang

Agricultural Information Institute, The Chinese Academy of Agricultural Sciences, Beijing, China, 100081

[*] *Corresponding author, Address: Agricultural Information Institute, No.12 Zhongguancun South St., Haidian District Beijing 100081, P. R. China, Tel:+86-10-82109819, Fax:+86-10-82109819, Email:nf.xieg@caas.net.cn*

Abstract: The rapid growth of agriculture web information brings a fact that search engine can not return a satisfied result for users' queries. In this paper, we propose an agriculture domain search engine system, called ADSE, that can obtains results by an advance interface to several searches and aggregates them. We also discuss two key technologies: agriculture information determination and engine.

Keywords: meta search engine, agriculture information, web search engine.

1. INTRODUCTION

With the Web quick development, the information resources on the Web is gradually abundant. In the meantime, the agriculture web sites spread over China like mushrooms after rain. Recently China has about more than 14000 large agriculture web sites. The available enormous amount of Web information requires the use search engine tools such as google, baidu, whose aim is to match general users' requirements, but they suffer from many problems such as, a) low retrieval rate, b) freshness problem, c) poor retrieval rate, d) long list of result which consumes time and efforts, e) huge amount of rapidly expanded information which causes a storage problem, and finally, f) large number of daily hits which makes most search engines not able to provide enough computational power to satisfy each users information need (Eldesouky & Saleh, 2008).

That leads a new generation of search engines called meta-search engines, which provides a united interface that send a user's query to multiple search engines, thus providing the means for a user to search a broader set of documents and potentially get a better set of results (Lawrence & Giles, 1999). A meta-search engine generally does not maintain its own index of web information instead of aggregates the certain search results into a unified result set that is re-ranked based on the relevance to the query. Metasearch engines increase the search coverage, solve the extendibility issues in searching eclectic information sources, facilitate the invocation of multiple search engines and improve information retrieval effectiveness (Meng et al., 2002).

In the paper, we propose an agriculture domain meta-search engine Architecture to build ADSE. ADSE uses different databases for searching information and the results are obtained at a unified interface.

The paper is organized as follows. Section 2 describes the related work. Architecture of the agriculture meta-search engine is proposed in the section 3. Section 4 discusses the domain Information determination algorithm in ADSE. Finally, Section 5 will conclude the paper.

2. RELATED WORK

A meta-search engine transmits user's search simultaneously to several individual search engines' interface, and gets results from all the search engines queried, and then re-ranked the results by relevance to the query. There has been a great deal of work done in making guided searches a reality. One of the best examples is GuideBeam (Guidebeam), which is the result of research work carried out by the DSTC (Distributed Systems Technology Centre) at the University of Queensland in Brisbane, Australia. GuideBeam works based on a principle called "rational monotonicity" that emerged from artificial intelligence research in the early nineties. In the context of GuideBeam, rational monotonicity prescribes how the user's current query can be expanded in a way which is consistent with the user's preferences for information. In other words it is a guiding principle of preferential reasoning (Peter & Bernd, 1999).

There are many meta-search engines in service. Dogpile and Vivisimo are commercial clustering engines. The Meta search engine Dogpile is a meta-search engine that searches multiple search engines at once (Dogpile, 1999). This site was used as it allows users to obtain results from some of the most popular search engines in one attempt, rather than having to perform the search at each of the sites individually. When a search was performed using the syntax e commerce security issues, results were found for most of the major search engines. The Dogpile meta-search engine is good, in that it shows the top search results for some of the top WWW search engines, web directories and Pay-for-Placements engines at one site. Vivisimo provides an innovative document clustering technology that acts as a metasearch engine. It transforms search engine outputs from long, tedious lists into crisply organized categories

(Vivisimo, 2006).One of the oldest meta search services, MetaCrawler began in July 1995 at the University of Washington. MetaCrawler was purchased by InfoSpace, an online content provider, in Feb. 97. Queries other search engines, organizes results into a uniform format, ranks them by relevance, and returns them to the user (Infospace, 1999).Delivers highly relevant results from a variety of different search engines and directories including Inktomi, GoTo, and the Open Directory.

3. NEURAL NETWORKS IDENTIFICATION ALGORITHM

In the web, agriculture information needs of users are stored in the database of multiple search engines. The rate of agriculture information explosion on the internet is much higher than the rate at which web search engines index the web. It is highly inefficient and inconvenient for an ordinary user to manually invoke multiple search engines and identify useful documents from the returned result sets. To support unified access to multiple search engines for getting agriculture information, an agriculture meta search engine can be constructed.

The key motivation behind the construction of our agriculture meta search mechanism is to increase the coverage and scope of search, handle the search of agriculture information efficiently, provide a single united access for several search interfaces. When a user issues a search query, the agriculture meta search engine fetches the results from various search engines, and the results relevant to the query. The results are unified into a individual result set with a single ranked list, and order them according to relevance.

Apart from the underlying search engines, a meta search hierarchy has four primary software components: a database selector (we select google, baidu, msn as our main databases), a document downloader, query dispatcher and a result merger. But ADSE, domain information determination is also a primary component. Reference ADSE architecture of an agriculture meta search engine is illustrated in Figure 1.

S1, S2, … Sn is the given search engine database from which the agriculture information will be fetched. The database selector is to identify one or more databases that are likely to contain useful documents for the query. The objective of performing database selection is to improve efficiency as by sending each query to only potentially useful search engines, network traffic and the cost of searching useless databases can be reduced.

The downloader component is responsible for establishing a connection with each selected search engine and passing the query to it, and downloads all URL obtained as result set.

The query dispatcher component focuses query management which parser the query, and specify options for each search provider, and finally

send the result to the corresponding downloaders. Query dispatcher may also try to adjust the relative weights of query terms in the original query to fetch optimal results.

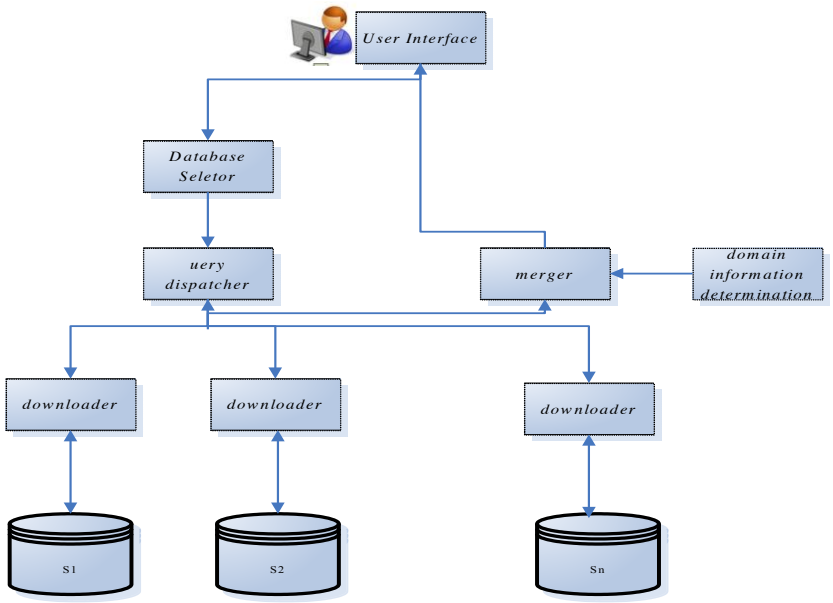

Fig.1. Agriculture Meta search Component Architecture

The merger component combines the results into a single ranked list by similarity between the query and documents in the result. The merger uses a ranking algorithm to re-rank the results.

The domain information determination algorithm component is to determine whether a document belongs to the agriculture domain based on a website url, which quickly collects the minimum closure of the domain-based web information, compared with other web page clustering algorithms in many domain search engines.

4. DOMAIN INFORMATION DETERMINATION ALGORITHM

For agriculture domain, one different point from other general met-search engines is to provide an interface for agriculture users to get agriculture information from the web. The general key technologies in meta search engines have been discussed in many papers. Here we proposed domain information determination algorithm based on website homepage content.

In a certain domain-specific website, whose homepage has a domain characteristic, and describes the abstract of the website's content, which is

used to determine the website's domain. But some websites' homepage description is very brief or only contains login information. The following page' content is the main domain-determined information description after clicking the homepage.

A domain website homepage contains many links, whose descriptions can be extracted to form a agriculture word set according to the domain lexicons (Xie et al., 2008). The set contains enough semantic information to measure the website domain type. In our method, a website homepage can be defined by the frame description as following:

```
def category webHome
{
attribute: title
        :type string
   attribute: anchorInfo
        :type ArrrayAnchor
attribute: copyright
      : type string
}
```

The above website homepage description can be stored in XML format. The attribute "title" represents the title of the homepage, and the attribute "ArrrayAnchor" describes the information about an anchor, represented as 2-tuple <url, urlInfo>, the first coordinate url represents a link in a anchor, and the second coordinate urlInfo represents the responding link anchor text in the anchor, for example:

<a href="http://[URL goes here]">Link Anchor Text

In which, the value of url is "http://[URL goes here]", and the value urlInfo is "Link Anchor Text".

In our method, the copyright information is not considered. So a domain website homepage can be represented as hInfo = <title,{<url, anchorInfo >}>. In order to compute the confidence of a website belonging to a certain domain, the content of url and anchorInfo is segmented to a set of words in the domain lexicons using our a fast algorithm for Chinese Word Segmentation. So hInfo can be also described as hInfo = <{ w_i },{<url, { w_j } >}>. For each word w_i, we define a word weight, called word rank (wr), so the vector of hInfo is divided two parts: 1) the vector of title as [wr_{t1}, wr_{t2},…, wr_{tn}] in which n is the vector dimension; 2) the anchor vector space (AVS), defined as:

$$AVS_{k\times n} = \left\{ \begin{bmatrix} wr_{l11} \\ wr_{l22} \\ . \\ . \\ . \\ wr_{lk1} \end{bmatrix} \begin{bmatrix} wr_{l12} \\ wr_{l22} \\ . \\ . \\ . \\ wr_{lk2} \end{bmatrix} \cdots \begin{bmatrix} wr_{l1n} \\ wr_{l2n} \\ . \\ . \\ . \\ wr_{lkn} \end{bmatrix} \right\}$$

In which, n is the number of anchors in the homepage and k is the number of components for each anchor data. So the steps of the domain website determination algorithm are described as following:

1) Determining the weight of the title and anchor Information belonging to a certain domain, called W_t and W_a;

2) Given a website web address, a crawler was used to download the website homepage, and then store the homepage content (HC);

3) Applying the webpage parser to produce a DOM tree equivalent to HC, and then get the description of webpage (hInfo). For Chinese website, we use word Segmentation to split the content of the title and anchor Information into words;

4) Assigning the weight to each word;

$$\text{Confidenc} \quad \Pr = W_t \frac{\sum_i wr_i}{N_t} + W_a \, {}^{\text{Conf}}a^{(\text{AVS}_{k\times n})}$$

5) Using the following formula to compute the confidence of a website belonging to a certain domain.

In which, ${}^{\text{Conf}}\boldsymbol{a}^{(\text{AVS}_{k\times n})} = \sum_j^n \sum_l^k \boldsymbol{w}_{jl}$, and Nt is the number of components in a title vector, n is the number of links in the homepage. For a given threshold confidence value (TCV) of a website belonging to a certain domain, if the confidencePr of a website is equal to or above TCV, we can conclude that the website belonging to the domain.

5. CONCLUSION

This paper proposes an agriculture domain search engine system, called ADSE. We have studied the agriculture domain search engine system hierarchy, in which we propose a domain information determination algorithm to solve agriculture information selection. The new approach significantly improved the domain-specific information closure, reducing the non agriculture information returned. The papers have also studied the domain Information determination algorithm in ADSE. Future work involves, building the agriculture domain meta search engine, incorporating more number of search engines in the study, studying the running performance of ADSE.

ACKNOWLEDGEMENTS

This work is supported by Special Fund Project for Basic Science Research Business Fee, AIIS, the Chinese Academy of Agricultural Sciences (Grant No 2008211) and The National Science & Technology Program (Grant No.2006BAD10A06).

REFERENCES

B. Peter and V.L.Bernd. Preferential Models of Query by Navigation. Chapter 4 in Information Retrieval: Uncertainty & Logics, The Kluwer International Series on Information Retrieval. Kluwer Academic Publishers, 1999. http://www.guidebeam.com/ preflogic.pdf.

Dogpile Home Page. Internet. 16 Jun. 1999. Available: http://www.dogpile.com/

E. Eldesouky, A. Saleh, N. El Gendy.An Efficient Strategy for Mobile Focused Crawling (MFC) Based on Mobile Agent Technology. The Sixth International Conference on Informatics and Systems (INFOS2008), 2008.

Infospace, Inc. Infospace. http://www.infospace.com/, January 1999.

N.F. Xie, et al.. .Journal of Jiangxi Normal University (Natural Sciences Edition), 2008, Vol. 32:192-196.

S. Lawrence and C. L. Giles. the NECI meta search engine. Computer Networks and ISDN Systems, 1998,(30):95～105..

S.Lawrence, & C. L.Giles (1999b). Searching the Web: General and scientific information access. IEEE Communications, 37(1):116–122.

Vivisimo clustering engine, 2006. Available from: http://vivisimo.com/demos/ PubMed@ NIH.html.

W. Meng, C. Yu, and K. Liu. Building Efficient and Effective Metasearch Engines. ACM Computing Surveys, 34(1), March 2002, pp.48-84.

Zheng Ye Lu.et all..The Strategies for Building Agriculture Web Sites. FITA2002, 2002. Guidebeam - http://www.guidebeam.com/aboutus.html.