Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 9 (4): 716-723, 2009
ISSN 1812-5654

Web Information Extraction for Question and Answering System about Prices of Chinese Agricultural Products

[1]Wen-Sheng Wang, [2]Li Liu, [3]Qing-Tian Zeng, [1]Xiao-Rong Yang and [1]Neng-Fu Xie
[1]Agricultural Information Institute, China Academy of Agricultural Sciences,
Beijing, People's Republic of China
[2]School of Computer Sciences and Technology,
Beijing Institution of Technology, Beijing, People's Republic of China
[3]College of Information Science and Engineering, Shandong University of Science and Technology,
Qingdao, Shandong, People's Republic of China

Abstract: This study presents an approach to Web information extraction for the question and answering system about prices of Chinese agricultural products. This approach first uses the training corpus to product keyword dictionary and then matches the sample pages to find key information path to format automatic information extraction. The advantage of this approach is effectively to avoid the nonstandard Hyper Text Marked Language (HTML) pages and to obtain high extraction accuracy in specific domain.

Key words: Web information extraction, wrapper, automatic rule learning, format recognition, question and answering

INTRODUCTION

Question and answering (QA) is one kind of typical application of information retrieval. Given a collection of documents (such as the World Wide Web or a local collection), the system should be able to retrieve answers to questions posed in natural language. QA is regarded as requiring more complex Natural Language Processing (NLP) techniques than other types of information retrieval such as document retrieval and it is sometimes regarded as the next step beyond search engines (Srihari and Li, 2000). Today, QA technologies attract the interest of many research centers and companies. Since the Text Retrieval Conference (TREC) added a question answering track, QA is becoming one of the most popular tracks of TREC.

There are many well-developed QA systems such as START, AnswerBus and IBM's Statistical question answering system and so on. START is a natural language processing system that analyzes English text and produces a knowledge base which incorporates information found in the text. Users can retrieve the information stored in the knowledge base by querying it in English. The system will then produce an English response. In addition, by annotating free-form text with English phrases and sentences, then matching these annotations with incoming queries, the power of sentence-level natural language processing can be effectively put to use in the service of multimedia information access. AnswerBus is an open-domain question answering system based on sentence-level information retrieval. It accepts users' natural-language questions in English, German, French, Spanish, Italian and Portuguese and extracts possible answers from the Web. It can respond to users' questions within several seconds. Five search engines and directories, Yahoo, WiseNut, AltaVista and Yahoo news are used to retrieve Web pages that potentially contain answers. From the Web pages, AnswerBus extracts sentences that are determined to contain answers. In China, many universities and research centers are having the research on QA, such as Fudan University, ICT and HIT University. Compared with English QA, Chinese QA is less developed because of the complexity of Chinese syntax.

Information Extraction (IE) (Freitag, 1998) is a new research area of information processing that can be used as the data source of QA, so it is an important component of QA. There are five kinds of popular technologies for Web information extraction that are based on Natural Language Processing (NLP) (Laender *et al*., 2002), Ontology (Srihari *et al*., 2008), HTML structure (Arocena and Mendelzon, 1998), Web search (Zhao *et al*., 2005), semantic patterns (Kim and MoNovan, 1993) and wrapper (Liu *et al*., 1999), respectively. These technologies have already been widely used, however none of these is suitable for all the web data.

Corresponding Author: Qingtian Zeng, College of Information Science and Engineering,
Shandong University of Science and Technology, 579 Qianwangang Street,
Huangdao, Qingdao 266510, China

We have developed a Web-based question and answering system about agricultural price (APQA). Web extraction for the price information about Chinese agricultural products is the core component of our system. This study presents an approach for Web information extraction within APQA about the price of Chinese agricultural products. This approach first uses the training corpus to product keyword dictionary and then matches the sample pages to find key information path to format automatic information extraction. We will address the process of extracting data from corpus and generation of extraction rules in the study. Experimental results show that our approach is effectively able to avoid the nonstandard HTML pages and to obtain high extraction accuracy in specific domain.

FRAMEWORK OF APQA

The technologies of knowledge storage, knowledge representation, information extraction and NLP are integrated in QA system. Generally speaking, a QA system usually contains three components including question analysis, information retrieval and answer extraction. The framework of APQA is shown in Fig. 1. And then this study will introduce the details about the automatic information extraction technology in present APQA.

AUTOMATIC INFORMATION EXTRACTION ABOUT PRICES OF CHINESE AGRICULTURAL PRODUCTS

The framework of the Web based information extraction about the price of Chinese agricultural products is shown in Fig. 2. The main process of the Web based information extraction includes:

- Since the main technology is based on wrapper, the wrapper is customized manually first
- Data are extracted from corpus and filtered into generate keywords
- Travel all the websites ready to be processed, filter and analyze them
- Match keywords and generate extraction rules of the website so as to customize wrapper automatically

Next, we will focus on the process of extracting data from corpus and the generation of extraction rules.

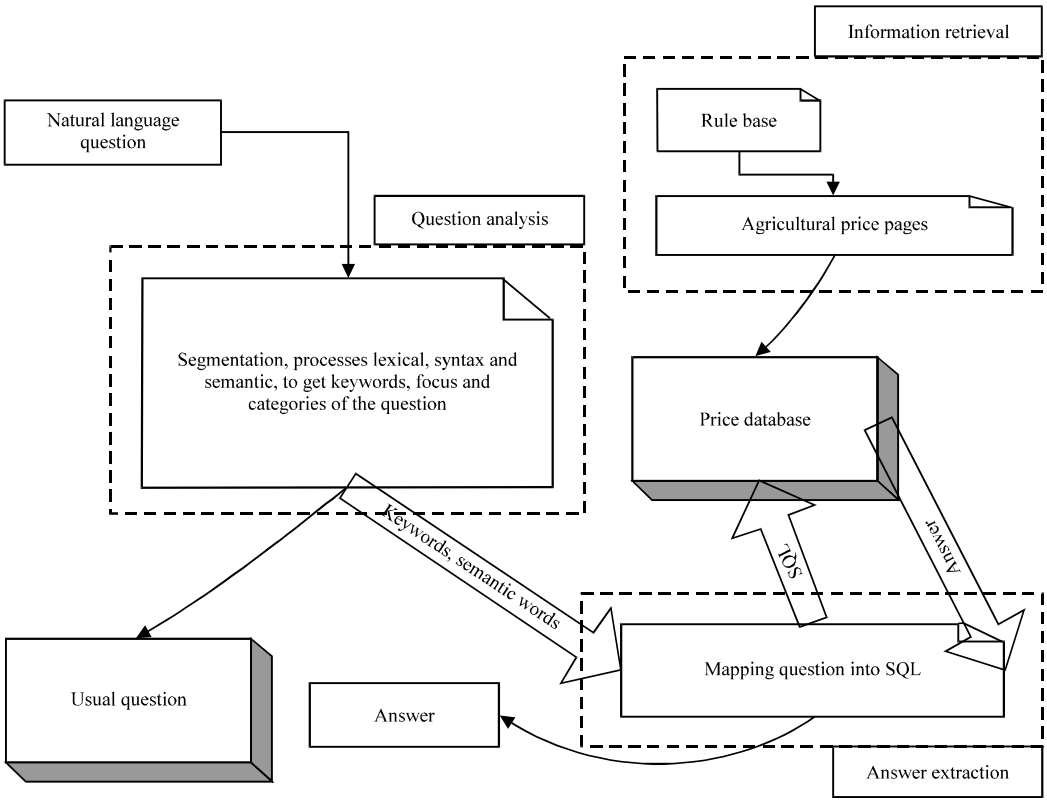

Fig. 1: Framework of APQA

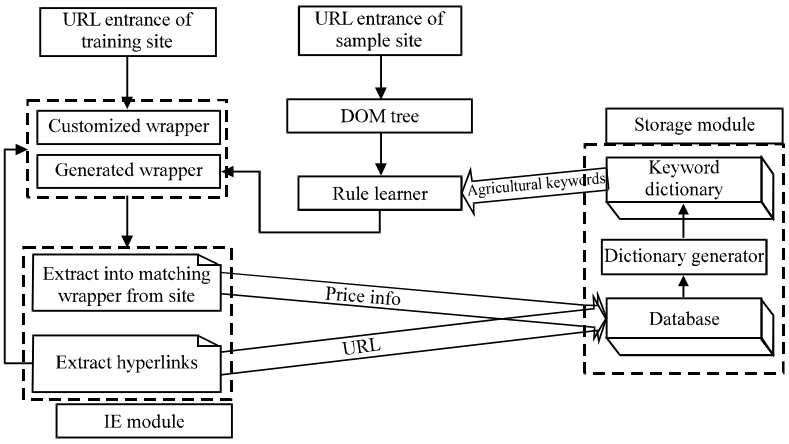

Fig. 2: Framework of automatic information extraction about the price of Chinese agricultural products

Table 1: Some of the websites we analyzed

Website	URL
China price information network	http://www.chinaprice.gov.cn/fgw/chinaprice/free/index.htm
Fujian price information network	http://www.se-price.gov.cn/ncpxx/
Henan agricultural information network	http://www.haagri.gov.cn/price/
China vegetable network	http://www.vegnet.com.cn/Price/

TRAINING CORPUS PROCESSING

The keywords are obtained by training corpus. In order to cover most of the keywords about agricultural products, we choose agricultural websites which have huge amount of information and simple structure. Some of the websites are shown in Table 1.

Features of agricultural websites: By analyzing a mount of Chinese agricultural price websites, we find most of them have the following features:

- The prices about Chinese agricultural products are published in unified format. What is more, there is information about the prices published in table. For example, the price is shown in Table 2
- Information Webpages about Chinese agricultural products are updated in a fixed time interval
- The same kind of information publishing pages are located in the same directory level of websites. And, there are hyperlinks from one page to other information publishing pages
- Titles of the webpage show information about publishing time, agricultural product types and other important information

Format recognition: The information contents provided in HTML are always formatted by tags like <Td> and </Td>, which is one of the important indications for format recognition in the process of table extraction. However, the existence of tag <Td> and </Td> doesn't mean the existence of the real effective data since sometimes it just represents some partition lines, pictures, hyperlinks and so on. This kind of tables without any contents is named as a non-data table or false form, otherwise table containing real practicing significant data will be called as true form. By statistic, the number of true form is less than 30 percent in a fixed field of HTML form (Myllymaki, 2001). The purpose of form recognition is to delete all the false forms in the webpage so as to get the true form.

Table 2: One of the published formats of prices about Chinese agricultural products

Market	Product	Price (RMB kg^{-1})
Nanqiao Suzhou	Lettuce	0.60
Yuegezhuang Beijing	Water spinach	3.80
Huilongguan Beijing	Dongguashan	3.20
Mawangdui Changsha	Taro	1.20

Extraction of attribute value: By recognition of page formats, we get the formats of the webpage about Chinese agricultural products. Next, we customize a data frame (Califf and Mooney, 1999) to match the attribute value of these webpage. The advantage of this method is suit for different page formats.

We use several large Chinese agriculture websites as training corpus. These websites cover most of the Chinese supply and demand information and agricultural product keywords.

Question and answering system is different from a general website, it should have the ability of publishing former information, so we need to store collected information into a database. By statistic, the number of supply and demand information of a large agriculture website is generally about 100,000-1,000,000. Because the information sources of our APQA are from several large agriculture websites, the capacity of our database should be at least 10,000,000. In consideration of the retrieval speed of database, we choose MySQL as database. Indexes for each attribute of records are built up to raise retrieval speed. Experiments shown in Table 3 prove that the retrieval speed rises greatly after building up indexes.

Generation of keyword dictionary: The purpose of training corpus is to generate keyword dictionary of agricultural products. In the process of extraction rules generating, the keyword dictionary is the basis of judging a word whether is an attribute value. The reason why the keyword dictionary is extracted manually but not by an official agricultural dictionary is that there is too much non-norm information existing on the Web, which is out of accordance with official information. This problem can be avoided by collecting information directly from the network.

There are many kinds of attribute values getting by information extraction: time, specie, publisher name, product name, market, location and so on. It is not all the attribute values are suitable to be added into keyword dictionary. For example, if we add keyword date into dictionary, it needs too many spaces. It is more suitable to be matched with regular expression. For the keywords without unified format, such as product name, market name and so on, rules generating has to be carried on with keywords.

The keywords are obtained by statistics of attribute values stored in the database and the keyword dictionary should be updated together with the update of database. Keyword dictionary includes the word frequency, which is taken as the threshold of rules generating. Word frequency and keywords are stored in the hash-table, which will be preloaded into memory before carrying out to match.

The target of training corpus is to generate agricultural keyword dictionary. According to Chinese agricultural market information, the training process generates the keyword dictionary DictAgri and DictMark, which represent agricultural product name and market location, respectively. Combining with the official classification standard of Chinese agricultural products and market location information collected by statistics, the process of the keyword dictionary are summarized and shown in Fig. 3.

Table 3: Extraction efficiency with or without index

Question	No. of answers	Time without index (sec)	Time with index (sec)
What is the price of cabbage in Beijing today?	189	13.557	0.003
How much is the taro yesterday?	93	13.164	0.002
What is the lowest price of watermelon in Shandong province?	1	9.192	0.001

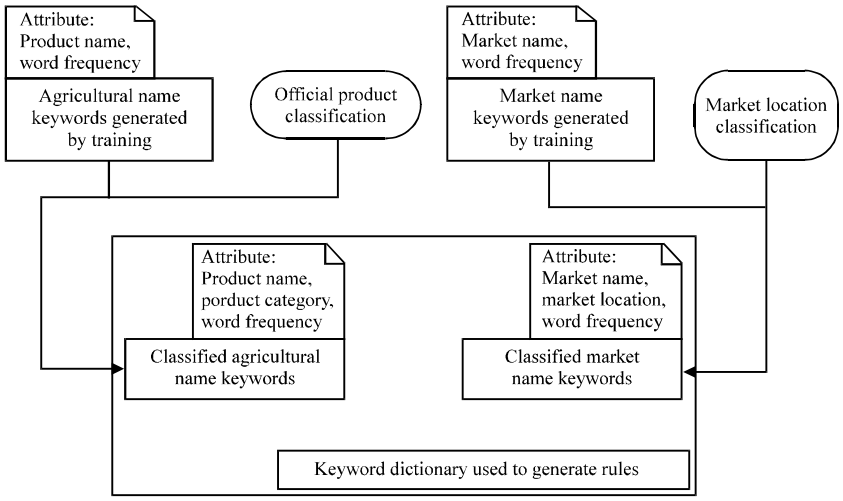

Fig. 3: Process of generating keyword dictionary

INDUCTION LEARNING OF EXTRACTION RULES

Selection of sample pages: The way of automatic rules generation is to analyze sample pages. The first step is to select suitable sample pages. This process is completed manually and it follows the following principles:

- The features of sample pages should represent most conditions of webpage about Chinese agricultural products
- The format of the key information area is obvious and easy to be extracted
- The sample pages are located at the same level of the website

Page filtered into and resolution: Generally speaking, most webpage are formatted by HTML and all of labels should appear in couple in the W3C HTML technical specification. Today many new web browsers' fault tolerance are becoming stronger and they accept some webpage which do not follow HTML standard (Yi *et al*., 2003). However, normal desolators can only resolute the standard HTML pages, otherwise there will be mistakes occurred. In present approach, the webpage are filtered before resolution and then JDOM will complete the resolution. The process of the whole resolution is shown in Fig. 4. Figure 5 shows an example of Document Object Model (DOM) tree.

Filter of key information nodes: Before presenting the approach, we first define a set of related variables:

- The set of DOM tree nods is $N = \{n_{body}, n_{href}, n_{tr}, YY, n_{td}\}$, where, the subscript represents the category of a node
- Threshold M_1 represents the quantity limit between the probable keyword nodes and regular nodes. The nodes will be recognized as probable keyword nodes if their quantity is not less than M_1
- If the quantity of matching times between a kind of probable keyword nodes and keyword dictionary is more than threshold M_2, this kind of nodes will be recognized as keyword nodes
- The path from root to keyword nodes is denoted as iPath

There are two steps to filter keyword nodes. First, we choose nodes whose quantity $n>M_1$ from N. The nodes obtained by the first step may include non-key information, such as table head, headline and so on. The second step is to match all the nodes obtained by the first step using keyword dictionary. The number of nodes matched successfully by the second step is denoted as n*. If $n^*>=M_2$, it means these nodes are keyword nodes and the paths from root node to them are iPath.

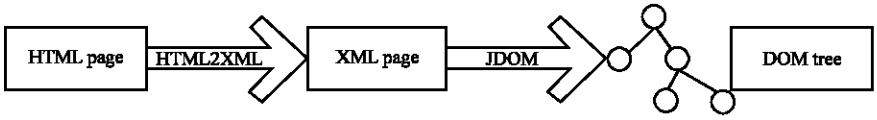

Fig. 4: Process of webpage resolution

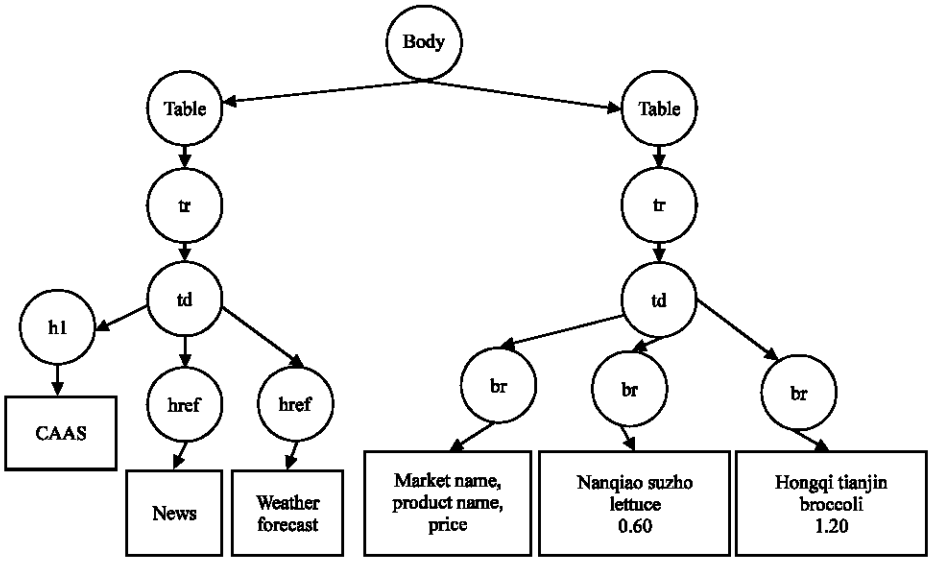

Fig. 5: Example of DOM tree

Automatic rules learning: According to the analysis to the features of Chinese agricultural webpage, keywords about agricultural products usually exist in the same position in the webpage. We have already found this particular district in the DOM tree that can be located by the information that users interested in. Usually, information extraction faces complicated conditions and different districts may be contained in a single webpage. However, there is only one kind of district appearing on the webpage about the price information of Chinese agricultural products.

The process of automatic rules learning is presented as follows:

Input: DOM tree of sample webpage and keyword dictionary
Output: iPath
{
Set threshold M1 and M2;
iPath = Null;
i=0;
Travel DOM tree;
*Take the nodes whose quantity is not less than M1 as set Ntemp;
while (Ntemp is not null) {
Select a node from Ntemp, matching with keyword dictionary, if success, i++;
if (i>= M2) {Set this kind of nodes as keyword nodes and put the path from root node to key information node into the set iPath. Set i =0 and break ;}
else {Select the other kind of candidate nodes, go to *}
}
return iPath;
}

The set iPath obtained records the keyword path in sample webpage. Next, it is only necessary to extract keywords from the whole site by matching iPath. The advantage of this approach is that it requests less semantic and the information filter (filter out non-important picture and information).

EXPERIMENTS

We have programmed to implement present approach in the following configure environment: (1) Hardware: CPU: Pentium 4 3.06G; Memory: 768M and Network: ADSL 2Mbit/s; (2) Software: System: Windows XP Pro with ServicePack2; Developing tools including Eclipse3.2.2 + Dreamweaver8 + JDK1.6.0 + Tomcat5.5 + Access2003.

Website www.chinaprice.gov.cn is used as training site. Agricultural information is published in uniform, shown in Fig. 6.

市场名称	品种名称	批发价
城北回龙观	莴笋	1.00
北京岳各庄	莴笋	1.40

Fig. 6: Example of information published in website

Every attribute of the information is published in uniform, so we customize wrapper using regular expression, it is able to describe market name and product name published in Chinese and price published in price pattern. The regular expression we make is: \\s*([\u4e00-\u9fa5]+)\\s+([\u4e00-\u9fa5]+)\\s+(\\d*.?\\d+)\\s*.

We extract keywords about agricultural products such as vegetable, fruit, egg, meat and so on to cover general agricultural keywords. Finally, 230,000 records are obtained about the price information of Chinese agricultural products in the first half year of 2007. After filter, 417 keywords about the name of Chinese agricultural products are generated. Then we extract several local agricultural sites, the number of keywords does not increase. It indicates that these 417 keywords have covered most of Chinese agricultural products.

Next, we choose www.agri.gov.cn as object website to do experiment and choose three pages as sample:

http://www.agri.gov.cn/jghq/sc/t20060705_642866.htm
http://www.agri.gov.cn/jghq/sc/t20060705_642859.htm
http://www.agri.gov.cn/jghq/sc/t20060705_642855.htm

We turn them into XML and use JDOM to resolute them by setting the threshold $M_1 = 30$ and $M_2 = 20$. The number of tags is shown as follow:

Tr-5 Td-711 Body-1 Href-16(S.....

It shows that we should choose nodes <Td> to extract. By matching the keyword dictionary, we get the format of nodes <Td> is ><Td>market name, product name, price <\Td>'. Use this rule to extract the whole site, 38 webpage and 207,215 records are obtained. By statistic, there are about 230,000 records in this site. The recall is nearly 90%. Figure 7 shows the results of information extraction. The records in ACCESS database is Chinese agricultural price information, including attributes such as >id, market, product name, price and so on.

We choose some general questions to test APQA:

- Questions containing product name, market name and time. For example, How much is beef in Huilongguan market today?

id	market	name	price	data	url
1480154	北宁窟隆台	大白菜	.04	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1588696	河南濮阳	菠菜	.08	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1550293	山东宁津	菠菜	.08	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1550298	河南濮阳	菠菜	.08	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1555881	山东宁津	菠菜	.08	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1583197	河南濮阳	菠菜	.08	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1587593	山东滕州	水萝卜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1593868	河南濮阳	白萝卜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1588586	河南濮阳	白萝卜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1592903	山东滕州	水萝卜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1550243	冀魏县天仙	菠菜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1534684	冀魏县天仙	菠菜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1539876	山东宁津	菠菜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1567312	岳阳花板桥	洋白菜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1566009	山东滕州	水萝卜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1561453	河南濮阳	菠菜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1555826	冀魏县天仙	菠菜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1561754	岳阳花板桥	洋白菜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1572561	岳阳花板桥	洋白菜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1555885	河南濮阳	菠菜	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=
1437841	山东寿光	西葫芦	.1	2007-6-5	http://www.zynw.com/lanmu2.php?lmid=

Fig. 7: Experiment results of information extraction

Fig. 8: APQA experiment result

- Questions containing product classification or market classification. For example, How much is vegetable in Beijing?
- Questions containing negative words. For example, How much is vegetable except tomato in Beijing?
- Questions querying the maximum or minimum result. For example, What is the cheapest vegetable in Beijing?
- Questions containing comparison. For example, How more much expensive is tomato in Shanghai than in Beijing?

Figure 8 shows the experiment results of APQA. For each question, the APQA can return time, market name, product name, price and reference URL of every fitted record. Compared with other open domain QA, APQA can

get more recall and precision because it is restricted in a very specific domain about Chinese agricultural products.

CONCLUSION

Focus on a specific domain, price information about Chinese agricultural products, an approach of Web information extraction is proposed. This approach can effectively avoid the nonstandard HTML pages and obtain high extraction accuracy. Although the approach is proposed for information extraction about the prices of Chinese agricultural products, it can be widely used in other specific domains where the information is published normally on the Web. The system can also get high recall (about 80%) in e-commercial web site. This APQA system still needs improvement to answer more complex questions, such as 'the price of beef increases, isn't it?'. And it is necessary to learn more extraction rules for the Web information in the future research.

ACKNOWLEDGMENTS

This research is supported by the National Science Foundation of China under Grant (No. 60603090 and 90718011), subproject of the National Key Technology R and D Program (2006BAD10A06-2) and the Taishan Scholar Program of Shandong Province.

REFERENCES

Arocena, G.O. and A.O. Mendelzon, 1998. WebOQL: Restructuring documents, databases and webs. Proceedings of the 14th IEEE International Conference on Data Engineering, Feb. 23-27, ICDE. IEEE Computer Society, Washington, DC., pp: 24-33.

Califf, M.E. and R.J. Mooney, 1999. Relational learning of pattern-match rules for information extraction. Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence, July 18-22, American Association for Artificial Intelligence, Menlo Park, CA., pp: 328-334.

Freitag, D., 1998. Information extraction from HTML: Application of a general machine learning approach. Proceedings of the 15th National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence, July 26-30, AAAI Press/MIT Press, Menlo Park, CA, pp: 517-523.

Kim, J. and D. MoNovan, 1993. Acquisition of semantic patterns for information extraction from corpora. Proceedings of the 9th IEEE Conference on Artificial Intelligence for Applications, Mar. 1-5, Los Alamitos, CA, IEEE Computer Society Press, pp: 171-176.

Laender, A.H., B.A. Ribeiro-Neto, A.S. da Silva and J. Teixeira, 2002. A brief survey of web data extraction tools. ACM Sigmod. Rec., 31: 84-93.

Liu, L., W. Han, D. Buttler, C. Pu and W. Tang, 1999. An XJML-based wrapper generator for Web information extraction. SIGMOD Rec., 28: 540-543.

Myllymaki, J., 2001. Effective web data extraction with standard XML technologies. Proceedings of the 10th International Conference on World Wide Web, May 1-5, WWW '01. ACM, New York, USA., pp: 689-696.

Srihari, R. and W. Li, 2000. A question answering system supported by information extraction. Proceedings of the 6th Conference on Applied Natural Language Processing, April 29-May 04, Morgan Kaufmann Publishers, San Francisco, CA, pp: 166-172.

Srihari, R.K., W. Li, T. Cornell and C. Niu, 2008. Infoxtract: A customizable intermediate level information extraction engine. Nat. Lang. Eng., 14: 33-69.

Yi, L., B. Liu and X. Li, 2003. Eliminating noisy information in web pages for data mining. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, Washington, DC. New York, pp: 296-305.

Zhao, H., W. Meng, Z. Wu, V. Raghavan and C. Yu, 2005. Fully automatic wrapper generation for search engines. The 14th International Conference on World Wide Web, May 10-14, Chiba, Japan, pp: 66-75.