http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (19): 5216-5220, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5216.5220

Novel Method of Web Database Redundancy Computing for Web Data Sources Selection

[1,2]Yan Zhang, [1]Qingzhong Li, [3]Rui Zhang and [2]Peiguang Lin
[1]School of Computer Science and Technology, Shandong University, Jinan, 250014, China
[2]School of Com. Sci. and Tech., Shandong Univ. of Finance and Economics, Jinan, 250014, China
[3]Shandong Provincial Institute of Electronic Products Supervision and Inspection, Jinan, 250014, China

Abstract: With the fast increasing number of Web databases (WDBs), it is core issue in the study of Web data integration that we should select the most appropriate composition of databases to query and obtain more targeted data at a smaller cost. In this study, in order to reduce redundant data from different sources, we propose a novel method of Web databases redundancy computing to select proper Web data sources for given keywords. To solve the problem, we propose a web database feature representation model, and based on sample data from the sources, we put forward the deep web redundancy computing method considering three different data types: text attribute, numeric attribute and categorical attribute. Experiments show that this method can achieve the desired objectives and can meet the demand to the integrated system very well.

Key words: Deep web, redundancy computing, data sources selection

INTRODUCTION

As the wide use of web databases, web is speeding up to "deepen" (Ghanem and Aref, 2004). There are mass of pages on the Internet which are generated automatically by the back-end database, and current search engines cannot make indexes to these pages that make the information transparent to users and are called Deep Web. Compared with the Surface Web, deep web contains much richer and "professional" (fix to a specific domain) information. In 2004, the UIUC made a relatively accurate evaluation to the whole deep web, and they suppose that there are 307,000 web sites which offer web database (WDB) and 450,000 web databases. So, it has been a hot study-spot in database domain that to implement the retrieval and use of the information in web database (Lin and Zhao, 2010).

To make good use of these data, there were many works that had been done, such as methods for finding QI (query interface) of WDB (Web Database) (He *et al*., 2003; Lin *et al*., 2008; Lin and Zhao, 2010), classification of QI (Peng *et al*., 2004) and query result extraction (Zhao *et al*., 2005), etc. But facing so massive data, it would be more inefficient if we send the query to QI directly.

Because of the large scale of Deep Web, Deep Web data integration system (DWDI) includes thousands of WDBs and the amount exceeds the traditional data integration system. Furthermore, because we can only visit the WDB by its query interface, DWDI's efficiency would be very low and it need to process massive data if DWDI dealt with each query interface. So, in order to improve the efficiency of DWDI, we need to handle the user's query request like the following style:

- If one WDB had no results to meet the query, DWDI did not need to route the query to the WDB
- If some WDBs had more redundant data among them, DWDI should select some of the WDBs that could possibly cover all the results to query

In order to improve the efficiency of DWDI, we should take more attention to select proper WDBs which were more satisfied with the user's query and they should agree with the following conditions:

- Related to the user's query. To computer the related degree of user's query request and WDB, we should digitize user's query request and the characteristics of WDB
- Returning more data that meet the user's query request
- Having minimum redundant data

To satisfy with the above conditions, the best method was that we had the copy of WDB but that was

Corresponding Author: Yan Zhang, School of Computer Science and Technology, Shandong University, Jinan, 250014, Shandong, China

impossible. Then if we can the independent samples of WDB, we can choose the better WDBs by applying the query request on the samples. In this study in order to reduce the redundancy of returning data from difference sources, we concentrated on computing the redundancy of difference sources with given keywords. To solve the problem, we propose a web database feature representation model based on the samples of different WDBs. Based on that, we gave a web redundancy computing method considering three main data types: Text attributes, numeric attributes and categorical attributes for source selection.

The rest of the study was organized as follows: First, related work was discussed in Section II. In section III, the web database feature representation model was illustrated. In section IV, the redundancy computing based on sample data was discussed and in Section V, the experimental setup and results were shown. Finally, the conclusion is presented.

RELATE WORK

There are mass WDBs and their data are redundant very much. To delete the redundant data is the key technology for DWDI. But if we can select the WDBs that have minimum redundant data and cover the entire query results, DWDI may has low complexity.

In the past ten years, there were many algorithms about data sources selection including GLOSS, gGLOSS/vGLOSS, CORI etc. (Gravano *et al*., 1999). At the same time, Ipeirotis presented the data sources selection method based on topic-classification. Aiming at the redundant information in the same domain, (Miao *et al*., 2009) presented the method to compute the similarity among the databases, which estimated the scope of redundant degree by regression analysis and making use of the sample data. Wang (2009) gave the selection method for Web databases based on TOP queries, which introduced the computational model of un-related degree of number type and text type based on un-related degree of meta-data.

Above works supported our work and supplied the preliminary. But in recent years, the personalization has been becoming the hot pot in the domain of IR and then Web databases should also satisfy user's personalized request. In this study, a new data sources selection method based on redundancy computing was presented, which was based on user's query request and the sample data. Furthermore, the expression method of web database's characteristics was also presented in this study.

EXPRESSION OF WDB'S CHARACTERISTICS

Expression of WDB's text characteristics: In a variety of query interfaces, most had text property as input field, such as title, publisher name, author in Book Search, and job name, company name, job descriptions in the Job Search. These properties mostly describe the contents of the various entities, so the text property in the database, compared with the general document, had its own characteristics:

- The text in web database had a very strong correlation, and mostly for the names and attributes of the various entities, and compared with the general nature and universal nature of the common documents, it had its own unique characteristics
- Texts in Web database often represented entities in real world. Most did not belong to the common words. For example, the "software engineering", although the single word "software" and "engineering" were two common Chinese words, but their frequency counted in the corpus was far lower than in the type of computer database

Therefore, we learned from the representation method about the document feature in Chinese text categorization, and gave the representation about the text feature in web database, such as Eq. 1 showed:

$$Attr_{text} = <tf_1, tf_2,..., tf_n> \quad (1)$$

In the Eq. 1, $Attr_{text}$ represented a text attribute, $tf_i (1 \leq i \leq n)$ represents the frequency, n which referred to all the text segmentation of the text attributes represents the number of the key words.

Expression of WDB's data characteristics: In query interface of Web database, there were a certain number of numerical properties, such as the price of book search, the number of employees in recruitment website. Because the numerical properties had the characteristics of continuity, and the normal distribution had strong universality, we used the expectations and biases to represent the characteristics of numerical properties.

If a property in query interface was a numeric attribute, we calculated μ and variance σ through the sample data, and the characteristics of the numerical property are as follows:

$$Attr_{data} = N(\mu, \sigma^2) \quad (1)$$

Expression of WDB's categorical characteristics: For Categories property, we used the ratio between the number of similar attributes in the Statistical samples and the total numbers to represent it, as Eq. 3 shows:

$$\text{Attr}_{\text{class}} = \frac{\text{No. of similar samples}}{\text{No. of total sample records}} \quad (2)$$

In Eq. 3, $\text{Attr}_{\text{class}}$ was equal to the number of similar samples/the number of total sample records.

REDUNDANCY COMPUTING BASED ON DATA SAMPLES

Firstly, we applied user's query on the sample data and got the result. Secondly, the percentage was calculated by the size of query result and the size of the sample. Thirdly, we estimated the size of web database. And at last, we estimated the amount of result which would meet to the user's query.

Definition: redundancy of WDBs: the percentage of repeated data in all WDBs, which were returned by the WDBs with a given query. For example, after a user submitted query Q1, the amount of returned records in WDB1 was n1 and n2 in WDB2. Suppose n12 was the amount of repeated records in n1 and n2, then the redundancy between WDB1 and WDB2 was:

$$\begin{aligned} \text{redundant}_{\text{WDB}_1}(\%) &= \frac{n_{12}}{n_1} \times 100 \\ \text{redundant}_{\text{WDB}_2}(\%) &= \frac{n_{12}}{n_2} \times 100 \end{aligned} \quad (5)$$

It was obvious that the larger the value of $\text{redundant}_{\text{WDB}}$ was, the more redundant the WDBs were and the lower efficient either one WDB was. To compute the redundancy between the WDBs, we obtained the records which satisfied user's query and then got the number of repeated records. We adopted the samples to simulate the redundancy by computing the similarity.

Text attribute redundancy computing: For this kind of attribute, we adopted the ratio of the sum of the frequency of the same keywords and the one of all keywords, shown in Eq. 6:

$$\text{redt}_{\text{text}}(\text{WDB}_1, \text{WDB}_2) = \frac{\sum_{i \in \{\text{sameKeys}\}} \text{tf}_i}{\sum_{j \in \{\text{WDB}_1.\text{text}\}} \text{tf}_j} \quad (6)$$

In the above equation, $\text{redundant}_{\text{text}}(\text{WDB}_1, \text{WDB}_2)$ was the redundancy of WDB_1 related to WDB_2. sameKeys was the same keywords in WDB_1 and WDB_2; WDB_1text was of all keywords of current attribute in WDB_1. tf_i was the frequency of the i^{th} keyword.

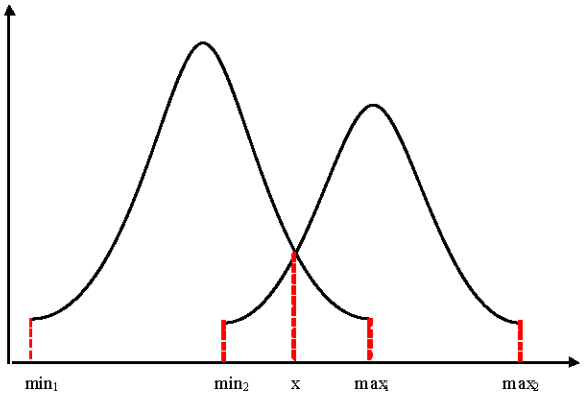

Fig. 1: Two normal distributions of the two number attributes

Numeric attribute redundancy computing: By the expression of the characteristics of WDB, the Normal distribution was used to express the Number data. Suppose $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ were the Normal distribution of the attributes in WDB1 and WDB2, and the miniature value and the maximum value in the related attributes of WDB1 and WDB2 were min1, max1, min2 and max2.,and x was the cross part, as showing in Fig. 1. Then, we gave the redundancy of Number attribute as the ratio of the area of overlap and the all. And the computation was given in Eq. 7:

$$\begin{aligned} \text{redt}_{\text{data}}(\text{WDB}_1, \text{WDB}_2) = {} & F_1\left(\frac{\max_1 - \mu_1}{\sigma_1}\right) - F_1\left(\frac{x - \mu_1}{\sigma_1}\right) \\ & + F_2\left(\frac{x - \mu_2}{\sigma_2}\right) - F_2\left(\frac{\min_2 - \mu_2}{\sigma_2}\right) \end{aligned} \quad (7)$$

Categorical attribute redundancy computing: We gave the redundancy of two classification attributes as the ratio of the number of records in the current classification value to all records, as was:

$$\text{redt}_{\text{class}}(\text{WDB}_1, \text{WDB}_2) = \frac{\text{No. of records in sameClass}}{\text{No. of all records in allClassValues}} \quad (8)$$

Considering all the three types of attribute, we got the overall equation of measuring the redundancy of WDB_1 and WDB_2:

$$\text{redt}(\text{WDB}_1, \text{WDB}_2) = \sum_i \beta_i \times \text{redt}_{\text{att}_i} \quad (\sum_i \beta_i = 1) \quad (9)$$

EXPERIMENT

Experiment dataset: In order to verify the validity of the characteristics expression method and data sources selection algorithm proposed in this study, sample data are crawled from Internet and analyzed. The data sources are four websites which include national recruitment website ChinaHR (www.chinahr.com), local recruitment website QingDaoHR(www.qindaohr.com) and general website with job information Ganji (www.ganji.com), Qilu talent net (www.qlrc.com), they were crawled base industry attributes (classification attributes) with the Watir tools. The data include about 6000 samples which contains position information (position title, number of recruit, work area). For statement simplicity, the four sources were represented by CHR, QDHR, GJ and QL, respectively.

Web database feature representation: We first extract features for each websites' text data (position title), numerical data (company scale) and categorical data (work area), the results shown in Table 1 (we only list top 5 in each types).

The data of each website are estimated by classification attribute based estimate method for t that the domain of human resource, a position rarely belongs to two different industries.

In addition, based on above characteristics, we calculated the redundancy matrix between each two website as shown in Table 2. In this table, the data in the cell of row i, column j intersecting expresses relatively redundancy vector of i-th database and j-th database, each component represents text, numerical and classification attributes, respectively.

It can be seen from Table 2, the position has large repeat between each database, it can reflect difference of position distribution; On the scale of company, it easy to find by data combined that: CHR has a lot of large companies, QDHR and QL has more medium companies, GJ has a wide distribution of companies, it's a medium repeatability between each library; On the regional distribution, CHR and QDHR has large redundancy, GJ and QL has large redundancy.

Web data source selection using redundancy: We design four queries to test the efficiency of the use of redundancy, which are shown in Table 3. The four queries covered three types of data attribute. Query3 and query4 focus on national recruitment and local (Shandong) recruitment, respectively. Meanwhile, the column of "parameter settings" verifies different situations with different values, especially for query3 and query4, two sets of values are fetched for test; Due to database redundancy calculation does not consider query requests, so just one set of values are fetched to verify the value of parameter.

According to the above queries, we calculated the similarity, amount of returned data and redundancy between query condition and each website, respectively. The results of calculation and execution are shown in Table 4 and 5.

Table 1: Feature extraction results for each website

	Position title	Company scale	Work area
CHR	(Limited company,5883), (Sell,2067), (manager,1974), (Beijing, 1922), (Technology, 1212)	(671,253)	(Shenzhen,247), (Jinan,269), (Guangzhou, 333), (Shanghai, 941), (Beijing, 2146)
QDHR	(Software,7240), (Engineer,6554), (Limited company,6306), (Technology,2697), (Shanghai,1709)	(155,174)	(Shanghai,1863), (Beijing,1300), (Shenzhen, 799), (Guangzhou, 493), (Hangzhou, 355)
GJ	(Limited company,5162), (Jinan,1929), (Shandong, 1632), (Manager,1121), (Technology,1016)	(229,327)	(Jinan, 4032), (Binzhou, 308), (Qingdao, 270), (Taian, 211), (Dongying, 147)
QL	(Limited company,4923), (Jinan,1469), (Technology, 1233), (Shandong,968), (Manager,803)	(140,232)	(Jinan,2124), (Qingdao,647), (Shandong, 551), (Yantai, 450), (Weifang, 375)

Table 2: Redundancy component matrix between each website

	CHR	QDHR	GJ	QL
CHR	-	(0.8611, 0.1248, 0.94)	(0.86, 0.3485, 0, 78)	(0.8078, 0.084, 0.64)
QDHR	(0.9302, 0.0948, 0.9966)	-	(0.892, 0.8462, 0.87)	(0.905, 0.945, 0.74)
GJ	(0.8666, 0.3042, 0.88)	(0.8369, 0.5632, 0.86)	-	(0.9032, 0.4528, 0.89)
QL	(0.8673, 0.1052, 0.90)	(0.8482, 0.942, 0.96)	(0.9037, 0.7542, 0.98)	-

Table 3: Query condition in experiment

	Query condition			Parameter settings	
Condition id	Position title	Company scale	Work area	α	β
1	Engineer		Beijing	$\alpha_1 = 1$	$\beta_1 = 0.4$, $\beta_2 = 0.3$, $\beta_3 = 0.3$
2	Engineer	Over 100	Qingdao	$\alpha_1 = 0.6$, $\alpha_2 = 0.4$	
3	Engineer			$\alpha_1 = 0.7$, $\alpha_3 = 0.3$	
4	Engineer			$\alpha_1 = 0.3$, $\alpha_3 = 0.7$	

Table 4 Similarity of WDB and user query

Compute Content Data lib	Text	Numerical	Categorical (Beijing)	Categorical (Jinan)
CHR	0.00637	0.3478	0.28720	0.03600
QDHR	0.10800	0.2846	0.18900	0.01803
GJ	0.01290	0.3810	0.00784	0.56480
QL	0.01430	0.3347	0.00275	0.36570

Table 5: Experiment results

Selection result ID	Auto selection	Artificial selection	α
1	QDHR, GJ	QDHR, GJ	$\alpha_1 = 1$
2	QDHR, QL	QDHR, QL	$\alpha_1 = 0.6, \alpha_2 = 0.4$
3	QDHR, GJ	QDHR, GJ	$\alpha_1 = 0.7, \alpha_3 = 0.3$
	CHR, QDHR	CHR, GJ or QDHR, GJ	$\alpha_1 = 0.3, \alpha_3 = 0.7$
4	QDHR, GJ	QDHR, GJ	$\alpha_1 = 0.7, \alpha_3 = 0.3$
	QDHR, QL	GJ, QL	$\alpha_1 = 0.3, \alpha_3 = 0.7$

As the experiment results showed, the data sources selection method basically in line with actual requirement. Although, an individual data in Table 5 (row4) is inconsistencies. Also it can be seen from Table 5, different value of parameter α has influence to the result of data sources selection, for example, the selection result direct at query3 and query4. Their reason is obviously shown in Eq. 9, different α_i value take influence to the calculation result of query similarity; similarly, the effect of parameter β for data sources selection can be further verified.

CONCLUSION

With the rapid development of Internet technology, the web data integration has becoming the hot research. In order to improve the efficiency of data sources selection, this study proposed a redundancy computing method to route the user query to proper sources based on data samples. With the method in this study, we could calculate the redundant degree between user's query request and WDB, and could further estimate the amount of WDB. Experiment shows that this method can help to auto-select proper data sources.

REFERENCES

Ghanem, T.M. and W.G. Aref, 2004. Databases deepen the Web. Computer, 37: 116-117.

Gravano, L., H. Garcia-Molina and A. Tomasic, 1999. GlOSS: Text-source discovery over the internet. ACM Trans. Database Syst., 24: 229-264.

He, H., W. Meng, C. Yu and Z. Wu, 2003. Wise-integrator: An automatic integrator of web search interfaces for e-commerce. Proceedings of the 29th International Conference on Very Large Data Bases, Volume 29, September 9-12, 2003, pp: 357 368.

Lin, P., R. Xu, Z. Hong and Y. Zhang, 2008. Finding the WDB's query interface in deep web automatically. Proceedings of the International Conference on Internet Computing in Science and Engineering, January 28-29, 2008, Harbin, China, pp: 195-200.

Lin, P.G. and L. Zhao, 2010. Research on the expression and extraction of WDB's query interface based on ontology. J. Convergence Inform. Technol., 5: 103-113.

Miao, Z.Y., P.P. Zhao, P.Y. Hu and Z.M. Cui, 2009. Estimation for overlapping rate of deep web databases based on attribute high-frequency words. Comput. Eng., 35: 28-30.

Peng, Q., W.Y. Meng, H. He and C. Yu, 2004. WISE-cluster: Clustering E-commerce search engines automatically. Proceedings of the 6th ACM International Workshop on Web Information and Data Management, November 12-13, 2004, Washington, DC., USA., pp: 104-111.

Wang, B., 2009. Research on database choice and query conversion for deep web. Master's Thesis, Dalian Institute of Technology, Dalian, China.

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