In Proceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA 2009):595-600, April 21-23, 2009, Brisbane, Australia

Selectivity Estimation for Exclusive Query Translation in Deep Web Data Integration

Fangjiao Jiang[1,2], Weiyi Meng[3], and Xiaofeng Meng[1]

[1]School of Information, Renmin University of China
{jiangfj, xfmeng2006}@gmail.com
[2]College of Physics and Electronic Engineering, Xuzhou Normal University
[3]Computer Science Dept, SUNY at Binghamton, meng@cs.binghamton.edu

Abstract. In Deep Web data integration, some Web database interfaces express exclusive predicates of the form $Q_e = P_i(P_i \in P_1, P_2, \ldots, P_m)$, which permits only one predicate to be selected at a time. Accurately and efficiently estimating the selectivity of each Q_e is of critical importance to optimal query translation. In this paper, we mainly focus on the selectivity estimation on infinite-value attribute which is more difficult than that on key attribute and categorical attribute. Firstly, we compute the attribute correlation and retrieve approximate random attribute-level samples through submitting queries on the least correlative attribute to the actual Web database. Then we estimate Zipf equation based on the word rank of the sample and the actual selectivity of several words from the actual Web database. Finally, the selectivity of any word on the infinite-value attribute can be derived by the Zipf equation. An experimental evaluation of the proposed selectivity estimation method is provided and experimental results are highly accurate.

1 Introduction

The Deep Web continues to grow rapidly [1], which makes exploiting useful information a remarkable challenge. Metaquerier, which provides a uniform integrated interface to the users and can query multiple databases simultaneously, is becoming the main trend for Deep Web data integration.

Query translation plays an important role in a metaquerier. However, due to the large-scale, heterogeneity and autonomy of the Web databases, automatic query translation is challenging. One of the important aspects is that Web database interfaces may express different predicate logics. The integrated query interface and many Web database interfaces express conjunctive predicates of the form $Q_c = P_1 \wedge P_2 \wedge \ldots \wedge P_m$, where P_i is a simple predicate on single attribute. While some Web database interfaces express exclusive predicates of the form $Q_e = P_i(P_i \in P_1, P_2, \ldots, P_m)$, which means any given query can only include one of these predicates. Exclusive attributes are often represented on a Web database interface as a selection list of attribute names or a group of radio buttons each of which is an attribute. A very interesting problem is, among all the Q_es on an interface, which one has the lowest selectivity? It is of critical importance to optimal query translation. In this paper, we mainly focus on the selectivity estimation of infinite-value attribute for exclusive query translation.

Before we carry out our study, we have two important observations: 1) there exist different correlations between different attribute pairs, and 2) the word frequency of the values on an infinite-value attribute usually has a Zipf-like distribution. Based on these observations, we propose a correlation-based sampling approach to obtain the approximate random attribute-level sample and a Zipf-based approach that can estimate the selectivity of any word by Zipf equation.

The rest of paper is organized as follows. Section 2 gives the overview of query selectivity estimation. Section 3 proposes the correlation-based sampling approach. Section 4 proposes a Zipf-based selectivity estimation approach. Section 5 reports the results of experiments. Section 6 introduces the related work. Section 7 concludes the paper.

2 An Overview of Query Selectivity Estimation

The overall flow chart of our approach is given in Fig.1.

Attribute correlation calculation for a domain. For any given domain (e.g., Books), we first calculate attribute correlation for each pair of attributes (***Attribute Correlation calculation)*** and identify the least correlative attribute $Attr_i$ for each specific attribute $Attr_u$. Because attribute correlation of each attribute pair in a domain is usually independent of the Web databases, the attribute correlation can be used for all the Web databases in the same domain.

Selectivity estimation for a Web database. Given an infinite-value attribute $Attr_u$ and a specific Web database, we use a series of query probes on $Attr_i$ in the Web database interface to obtain an approximate random attribute-level sample on $Attr_u$ (***Correlation-based sampling)***. The word rank on $Attr_u$ can be calculated from the sample, which is viewed as the actual word rank on $Attr_u$ of the Web database due to the randomness of the sample. Then several words on $Attr_u$ are used to probe the actual Web database and the frequencies of these words are returned (***Word frequency probing)***. Zipf equation can be estimated using the word ranks and the actual frequencies of several words (***Zipf equation calculation)***. Finally, for any word on $Attr_u$, we can estimate its frequency by the Zipf equation and its rank (***Selectivity estimation)***.

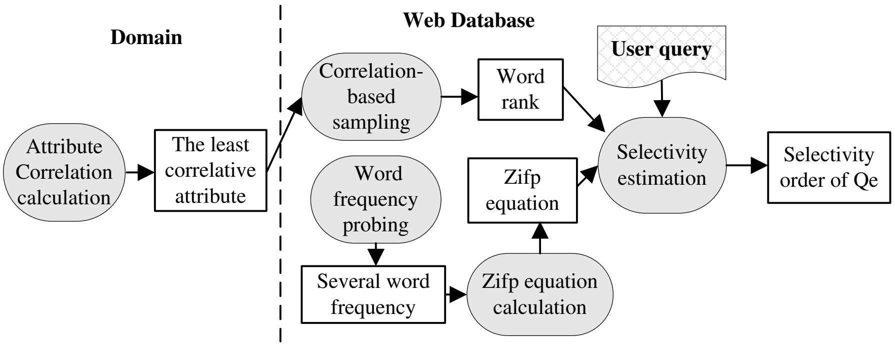

Fig. 1. The processing flow of our approach.

3 Correlation-based Sampling for Word Rank

In this paper, we use *Attribute Word Distribution* of different attributes to define the concept of attribute correlation.

Definition 1 Attribute Word Distribution (AWD). *Given all the words w_1, w_2, ..., w_m of the values of attribute A in a database D, the Attribute Word Distribution for A is a vector $\overrightarrow{v}(v_1, v_2, \ldots, v_m)$, each component of which v_i is the frequency of the word w_i. Under the assumption that no word appears more than once in an attribute value, the frequency of the word w_i is the number of tuples returned by the query $\sigma_{A=w_i}D$.*

Definition 2 Attribute Correlation. *Attribute Correlation is the dependence between any attribute pair($Attr_u$, $Attr_v$) and is measured by the difference of the Attribute Word Distributions of the returned results on an attribute ($Attr_u$).*

A measure of the distribution difference is Kullback-Leibler(KL) divergence. If we submit different queries Q_1, Q_2, ..., Q_s on $Attr_v$, we will gain the corresponding result sets S_1, S_2, ..., S_s on $Attr_u$. Suppose that S is the union of S_1, S_2, ..., S_s and S consists of a set of words w_1, w_2, ..., w_k. Then the KL-divergence of $Attr_u$ from S to S_j is:

$$D_{KL}(S||S_j) = \sum_{l=1}^{k} prob(Attr_u = w_l|S) log \frac{prob(Attr_u = w_l|S)}{prob(Attr_u = w_l|S_j)}$$

where prob($Attr_u$ =w_l| S) refers to the probability that $Attr_u$=w_l in S and prob($Attr_u$=w_l| S_j) refers to the probability that $Attr_u$=w_l in S_j.

Attribute correlation is the average of the KL divergence of $Attr_u$ from S to S_j:

$$Correlation(Attr_u, Attr_v) = \frac{1}{s} \sum_{j=1}^{s} D_{KL}(S||S_j)$$

After discovering the least correlative attribute $Attr_i$, we submit some query probes on $Attr_i$ to the Web databases and collect the returned results on attribute $Attr_u$ as the attribute-level sample of $Attr_u$, which is the approximate random sample. Then we order the words of the sample by their frequencies and the word rank can be viewed as the actual one due to the randomness of the sample.

4 Zipf-based Selectivity Estimation

It is well known that English words of a general corpus satisfy the Zipf distribution. However, it is not clear if the words of text attributes in different domains also follow this distribution. Our experiments indicate that they do.

Zipf distribution can be represented by $N = P(r + p)^{-E}$ [9], where N represents the frequency of the word, r represents the rank of the word and P, p and E are the positive parameters. As Fig.2 shows, we submit word i, word j to the Web database and obtain their frequencies F_{wi}(i.e., N_i) and F_{wj}(i.e., N_j), respectively. And we know the ranks of these two words (i.e., r_i and r_j) from the sample obtained in section 3. Then, we can estimate the parameters P, p and E as follows.

- Equation Transformation: After the logarithm transformation, the Zipf equation is changed to $ln(N) = lnP - Eln(r + p)$. Because the parameter p $(0 < p < 1)$ is usually much smaller than word rank r(i.e., some applications even assume $p = 0$), the parameter E is approximately viewed as the slope of the line $ln(N) = lnP - Eln(r)$ as shown in Fig.3.
- Parameter E: E can be calculate by the equation $E \approx \frac{ln(N_i)-ln(N_j)}{ln(r_j)-ln(r_i)}$.
- Parameter p: When E is estimated, parameter p can be derived from the equation $\frac{N_i}{N_j} = \frac{P*(r_i+p)^{-E}}{P*(r_j+p)^{-E}}$. So we have $p \approx \frac{r_j - r_i * e^m}{e^m - 1}$ $(m = \frac{1}{E} * \ln \frac{N_i}{N_j})$.
- Parameter P: Finally, the parameter P is derived. $P \approx N_j * (r_j + p)^E$.

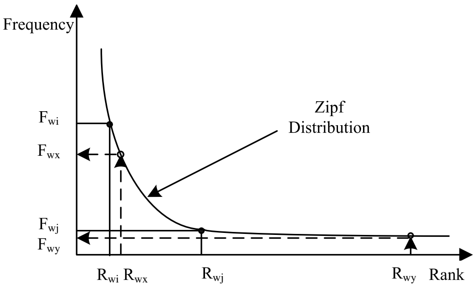

Fig. 2. Zipf-based Selectivity Estimation.

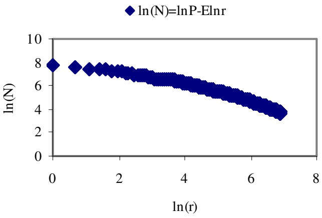

Fig. 3. Distribution transformation

Consequently, we can use the Zipf equation and the word ranks to compute the selectivity of any word on the attribute.

It is worth noticing that the parameters P, p and E are not unique. We study the relationships among the precision, word ranks and rank distances. The results show that the precision will go down when the rank increases and to keep the precision stable, the distance of two word ranks should be increased with the increase of the word ranks.

5 Experiments

We evaluate our approach with the precision measure which is defined as follows.

$$Precision = \frac{1}{N} \sum_{n} \left| \frac{N_r - E_s}{N_r} \right|$$

where N_r is the number of results when submitting the word on the attribute to the actual Web database, E_s is the selectivity of the word on the same attribute estimated by our approach, and n is the number of the words that we test in the experiments.

We select the top 100 words on *Title*, *Conference* attribute of Libra, *Title*, *Director* attribute of IMDb and submit them to actual Web databases. Meanwhile, we estimate the selectivity of these words using our approach. Overall, as we can see from Fig.4, the precision of our approach is generally good.

However, there is still some deviation on estimation values. The reasons are that any two attributes are somehow correlative with each other and the words on some infinite-value attributes do not satisfy Zipf distribution perfectly.

Given that our approach can cope with selectivity estimation of all the infinite-value attributes and it is domain independent, it is generally feasible to be applied in query translation for exclusive query interface.

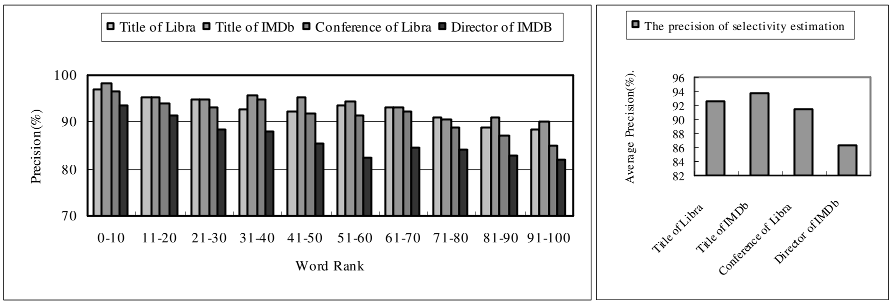

Fig. 4. The precision of selectivity estimation.

6 Related Works

The problem of selectivity estimation through uniform random sampling has received considerable attention [2, 6]. [2] cannot be applied as we do not have full access to the Web databases. [6] proposes a random walk approach to sampling the hidden databases, which is a database-level sampling and relatively complex compared with our attribute-level sampling. [3] focuses on the selectivity estimation of the text type attribute with several constraints (e.g., *any*, *all* or *exactly*, etc.) in Web database interfaces.

7 Conclusions

In this paper, we study the query translation problem of the exclusive query interface and present a novel Zipf-based selectivity estimation approach for infinite-value attribute. Experimental results on several large-scale Web databases indicate that our approach can achieve high precision on selectivity estimation of infinite-value attribute for exclusive query translation.

Acknowledgments. This work is supported in part by China 863 High-Tech Program(No.2007AA01Z155); NSF of China (No.60833005, 60573091); China National Basic Research and Development Program's Semantic Grid Project (No.2003CB317000). US NSF grants IIS-0414981 and CNS-0454298.

References

1. The Deep Web: Surfacing Hidden Value. http://www.completeplanet.com/Tutorials/
2. Oliken. F.: Random Sampling from databases. PhD Thesis, University of California, Berkeley (1993).
3. Zhang. Z., He. B., Chang. K. C. C.: On-the-fly Constraint Mapping across Web Query Interfaces. In: IIWEB (2004).
4. Mandelbrot. B. B.: Fractal Geometry of Nature. W. H. Freeman and Co. (1988).
5. Si. L., Callan. J. P. : Relevant Document Distribution Estimation Method for Resource Selection. In: SIGIR. (2003) 298-305.
6. Dasgupta. A., Das. G., Mannila. H.: A random walk approach to sampling hidden databases. In: SIGMOD. (2007) 629-640.