A Novel Method on Incremental Information Acquisition for Deep Web

Fang Wei, Pan Wu-bin, Cui Zhi-Ming

Journal of Convergence Information Technology, Volume 6, Number 6, June 2011

Abstract

Deep Web is autonomous, independently updating, and its data are always in a state of frequent update. However, the user always hopes to obtain the newest content in the current Web database. Different from previous research, this paper wants to emphasize the importance of updating frequency in the study of Deep Web information acquisition. And, an approach on incremental information acquisition based on logical reinforcement learning has been proposed. Then, we find in our research that under the same condition of constraint resources, the novel approach can improve the freshness of data, discovery efficiency of new data and the service quality of Deep Web information integration.

Keywords: Deep Web, Logical Reinforcement Learning, Information Acquisition, Update Frequency

1. Introduction

Traditional search engines can only search the indexed information on the surface of Internet at present; however, a large amount of information cannot reach by traditional search engines, which is implied in the depth of the Internet. This information is called Deep Web (Deep Web is also known as Invisible Web or Hidden Web)[1]. Bright Planet’s [2] research has shown that the amount of Deep Web information, which is 500 times more than the indexed information, is extremely huge, and 95% of the Deep Web contents are publicly accessible through the Internet without registration and paying. Deep Web's information is generally stored in the Web database of server, which is usually greater in the amount of information, more specific in theme, better in information quality and structure when compared with the static page. In order to facilitate the user to use the Deep Web information fast and efficiently, some scholars have implemented a wide range of research on Deep Web data integration [3] [4] [5] [6]. As we all know, Deep Web information integration mainly has two implementation approaches currently: one is a method based on meta search [7], unified query interfaces are provided for a domain, the user queries are transmitted to each Deep Web data source by semantic mapping, and the results returned by extraction, semantic annotation and overlap removal are presented to the users. Then, the case doesn’t need to maintain the local database, but there are following shortcomings: query response time is determined by the service quality of remote data source, and response time is uncontrollable; at the same time, establishing and maintaining semantic mapping between a unified query interface modes and each interface mode of data source is costly. The other is the same as an establishment of traditional search engine, contents in Deep Web database are crawled, stored in a local copy library of dynamic web pages and indexed. Users’ query requests can be responded in the shortest time [8] [9]. Under the case, Jayant Madhavan and others have proposed a new data integration framework based on DataSpace-PayGo[10], an integration framework is characterized by Pay-As-You-Go, evolution, low establishment cost, domain independence and so on. The second case has been paying increasing attention from scholars at home and abroad currently, which will be the mainstream of the study on Deep Web data integration, and a key issue in the case is how to synchronize local data with the data in remote data source. This paper will focus on...
the research of the key issue, under the same condition of updating resource. The maximization synchronization between local data and remote data is maintained.

Due to Deep Web is autonomous, updating independently, the data is always in the state of frequent update, and user always hopes to obtain the newest contents in current Web database. Therefore, a new method to update the local Deep Web data automatically and incrementally is more desiderata to be proposed, so that the freshness of the local data and the discovery efficiency of new data are improved under the same condition of resource constraints. The new method proposed can effectively improve the service quality of Deep Web information integration, so that Deep Web information can serve the scientific research, production and policy making better.

There are five parts in this paper. We start with a brief introduction the related work on information acquisition. Then we propose a novel approach to incrementally obtain the Deep Web information. An experiment of this novel approach is also provided. We end the study with the conclusion.

2. Related Work

With the rapid growth of the Internet information, incremental information acquisition technology becomes an effective way to obtain network information, time and resource waste bought by the repeated collection of unaltered web pages can be avoided. For surface web, page change and incremental information acquisition technology have already attracted vast attention and research [11] [12] [13]. However, as far as we know, the research on Deep Web incremental information acquisition is very little. Some research related to this issue includes the following aspects:

As to incremental information acquisition of Surface Web, Brandman et al [12] have proposed a server implementation method, that is, a URL and its list file of updating rate are reserved on a Web server. Then, the crawler downloads the URL list before crawling the site, identifies the modified URL set since last accessing and crawls the modified pages only. This method is extremely efficient, resource waste of server and crawler is avoided, but its biggest defect is in need of modifying the Web Server implementation. While the client technology is not dependent on the function of server, Cho and Molina [13] have made a profound research on how to improve the freshness of the local data copy if the data source is autonomy and independently updating, an incremental Web crawler has been proposed, several data update strategies were implemented, and their efficiency was analyzed and compared.

In terms of Deep Web information acquisition, Ntoulas Alexandros [14] has proposed a theoretical framework about Deep Web query generation problem for unstructured Deep Web, an automatic and efficient strategy for query generation was formulated based on this framework. Wu Ping et al [15] have proposed a method crawling Deep Web data efficiently by query selection for structured Deep Web, structured Web database is considered as a property - value graph model, the Web database crawling problem was transformed into a graph traversal problem. Much Deep Web content was obtained with a minimum of query submission. Recently, Google has proposed a multi-field and cross-lingual approach on Deep Web information acquisition [16]. The crawled Deep Web content was surfaced, then crawled content was put into the Google index, so that the user could search part of the Deep Web content by Google. Currently, Google could crawl hundreds of Deep Web dynamic pages per second.

In Deep Web incremental information acquisition, Capi et al[5] have applied the reinforcement learning to the assessment calculation of theme correlativity; the correct direction of search was controlled. Taiyo Maeda, et al[17] constructs a PSE System for Developing Reinforcement Learning Algorithms. Yu Chen, et al[18] present a method that automatically obtain new patterns of term. These methods are represented by pos tags and indicated term. Hierarchical reinforcement learning method has made significant progress in solving the "dimension disaster" problem, but for the Deep Web incremental information acquisition problem, it was difficult to resolve by traditional reinforcement learning method in virtue of state explosion and hierarchical disposal difficulty. Therefore, the logical reinforcement learning technique with the status and activity of logical expression has been proposed, which could solve such problems effectively.

3. A novel method on Incremental Information acquisition

Because the Poisson process can be used to describe the variation regularity of the Web object, which has been verified in the previous section. Then, for each object in the local database, we will
follow the object change of original database in a period of time. And the average change frequency of the object will be calculated approximately by the formula $\lambda = \frac{X}{T}$.

After the average change frequency of each object was counted, the object's synchronization frequency could be determined, so that the object information in local database obtained the optimal synchronization effect. Mathematical description of the problem is as follows: the known average change frequency of the object is $\lambda_i (i = 1, 2, ..., n)$, the goal is to determine the corresponding synchronization frequency $f_i (i = 1, 2, ..., n)$ of each object, so that the average novelty $\bar{F}(S)$ of local database is maximal in the condition of satisfying simultaneous resource constraints.

$$\bar{F}(S) = \frac{1}{n}\sum_{i=1}^{n} F(e_i) = \frac{1}{n}\sum_{i=1}^{n} \bar{F}(\lambda_i, f_i)$$ (1)

### 3.1. Deep Web data update

According to the characteristics of Deep Web, Deep Web data update strategy could be formulated based on two different granularities. One is that the update frequency was determined based on the importance weight and change frequency of the data source. The other is that the update frequency was determined based on the historical change frequency of the data record in data source.

Data’s freshness was regarded as an assessment indicator of Deep Web data updated strategy, for the Deep Web information acquisition system, if the local data record is the same as actual content of data record in remote Deep Web at the same moment, the data record called fresh. The freshness of a data record $r_i$ maintained by data crawling system could be defined as follows:

The freshness of element $e_i$ and database $S$ at a moment $t$ is defined in literature [19], which was calculated according to the following formula:

$$F(r_i, t) = \begin{cases} 1, & \text{up-to-date} \\ 0, & \text{otherwise} \end{cases}$$ (2)

According to the definition of the above formula, the average freshness of set $S$ constituted by $N$ data records was further defined as follows:

$$F(S, t) = \frac{1}{N}\sum_{i=1}^{N} F(r_i, t).$$ (3)

Theory synchronization frequency $f_i$, $i = 1, 2, ..., n$ of each object could be calculated by Lagrange multiplier, then the object data was synchronized according to $f_i$ so that the average freshness of local database could be maximal.

A data record set $S$ is maintained by incremental information crawling system, the average freshness and age of $S$ at a moment were concerned. At the moment, it could be measured by the average in time:

$$\bar{F}(S) = \lim_{t \to \infty} \frac{1}{t}\int_{0}^{t} F(S, t)dt.$$ (4)

Based on hierarchical logical reinforcement learning algorithm, an important indicator of evaluating information incremental acquisition system, Deep Web data incremental update objective was abstracted as an optimal problem. For all $(s, a)$ initialization table entry $Q_{0}(s, a) = 0$, returned by the range of data source representing its activities in each plot:

$$r_i = \frac{1}{N}\sum_{i=1}^{N} F(r_i, t)$$ (5)
The logical reinforcement learning $Q$ was updated at a period of time:

$$q_j = r_j + \lim_{t \to \infty} \frac{1}{t} \int_0^t F(S, t) dt. \quad (6)$$

In the premise of resource constraints according to the algorithm, that is, the maximum number of interactions with the server was $M$, making $\sum_{i=1}^{N} \omega_i f_i \lambda_i$ maximal, $f_i$ and $\lambda_i$ were collection frequency and change frequency of data record $i$ respectively, $\omega_i$ was the importance weight, defined as the importance of data source or the entity.

### 3.2. Discovery of New Deep Web data

Based on logical reinforcement learning algorithm, online learning was carried out in the process of the Deep Web data acquisition. The corresponding reward value was set according to the new records in the result returned by keyword or a combination of keywords. According to the learning results, more resources were allocated to the keyword or a combination of keywords which may emerge new data. The discovery efficiency of new data could be improved effectively in the premise of same resource constraints.

In order to avoid the expansion of the search tree in the process of data acquisition, the reinforcement learning techniques were applied to the controllable web crawler method of data acquisition. Some control "experience information" was obtained by this method through reinforcement learning techniques, the posterior return was predicted according to the information, searched according to a subject, so that the return value returned from accumulation was maximal. Concrete steps were shown in Figure 1.

Figure 1 showed the process of training and crawling based on logical reinforcement learning crawler (LQ-Spider), including the following steps:

1. The theme of remaining query data was provided, the site initial training queue URL was built respectively, then the front queue URL was extracted. The link address URL in pages was extracted by analyzing page structure, and the immediate return was calculated according to key information of page. The value of future return was drawn from combining the experience, and then the comprehensive $Q$ value of this link address was calculated by combining the future return in a value word library.

2. The trust of the immediate return value and the future return value were weighed. That is, whether the present is the phase of disposal and use or exploratory, the trust was

![Figure 1. Training Module Flow Chart](image-url)
controlled. According to whether the depth factor of URL address is greater than 5, if the depth factor is greater than 5, then abandoned, did not put into the URL queue remaining to extract. According to the survey, 91.6% of the depth of the page of the deep Web query interface were within the 5-layer, so when the depth of the URL link is greater than 5, did not deal with the link, disposal capacity was reduced effectively in the premise of ensuring the accuracy.

3. When the depth factor of the URL link is less than 5, then judged whether the comprehensive Q value is greater than a theme value, if yes, then the original property value of Value word library was updated, and the future return was calculated with the new value word library, and then put into the URL queue remaining to extract on priority, and trained so forth until obtained the final URL queue remaining to extract, then the Deep Web incremental information was crawled by crawler program purposefully. If the comprehensive Q value was less than a theme value, then the URL was elided. It returned to step (1) and continued to the next round of training.

In relation to the traditional web spider model, the immediate return and future return were fully integrated in this training module, the feature of the "calculation of URL comprehensive Q value" was that the immediate return was calculated according to the structure and key information of the page. Future return value was obtained by combining the experience. The link Q value was calculated; and the trust of the immediate return value and the future return value were weighed, that is, whether the present was the phase of disposal and use or exploratory, the trust was controlled, so that the URL was put into the queue remaining to extract on priority, its sequence was maintained. If you wanted LQ-Spider to obtain a more reasonable Q value corresponding to the URL from the module "calculation of the URL comprehensive Q value ", the weighed problem of the immediate return value and future return value needed to solve, that is, two corresponding value of "trust" should be given respectively.

4. Experimental Validations

The efficiency and the feasibility of the proposed method were evaluated through specific experiments in this section. First, the feasibility of the new approach was verified by analyzing their theoretical basis, then the performance advantages and disadvantages of using this method was further validated on this basis.

Currently, performance index of focused crawler was mainly evaluated by calculating the ratio of the relevant pages and related pages crawled, then superior or inferior of crawler were weighed. In this paper, the method of harvest rate was used to evaluate the performance of different crawlers. And, the value harvest-rate∈[0,1]. The formula represented that the ratio of the pages crawled effectively and the total crawled pages [20].

Deep Web focused crawler based on logical reinforcement learning was used by experiment, Deep Web crawler (such as Best-first and Breadth-first) was usually used to crawl and analyze the data source of two domains(Forums, Jobs),the harvest rate that crawler acquired information of different methods was compared, the experimental results were shown in Figure 2.

In each practical crawling, the links of internal and external sites were all needed to be tackled. Human intervention was relatively small, the scalability of experimental scale was better, and crawling was not limited to a few sites. After a period of time of crawling, with the number of links in access queue was increased in geometrical status, memory consumption was very fast, CPU utilization became very low. Therefore, the relative data structure occupying the capacity of memory needed to limit, when its capacity was greater than a certain value, data should be written to disk by persistence technologies (serialization).

In order to test crawling effect of Deep Web focused crawler in different methods, two different areas (such as forums, jobs) were selected in this paper. Then, Best-first, Breadth-first strategy and our crawling strategy (LQ) were applied to crawl respectively. Figure 4 has shown the comparison of the crawling performance of several crawlers, X axis
A Novel Method on Incremental Information Acquisition for Deep Web
Fang Wei, Pan Wu-bin, Cui Zhi-Ming
Journal of Convergence Information Technology, Volume 6, Number 6, June 2011

represents the number of results page crawled by crawlers, y-axis represents harvest rate. In addition, each result page concluded at least ten data records taken from the backend database. Experimental results have shown that the crawler based on logical reinforcement learning could obtain more information, the proposed crawling strategy has better application effect, and the harvest rate of the crawler was increased significantly.

5. Conclusions

Aim to an incremental information acquisition problem of Deep Web, which mainly is a difficulty in the data updates and discovery of new data. And their key is the update frequency and new data discovery. Due to the problem of state explosion and problem of slicing while solving this kind of problems, it is difficult to resolve by using traditional reinforcement learning method. Therefore, a novel approach on Deep Web incremental information acquisition based on logical reinforcement learning was proposed to analyze the Deep Web data update frequency in this paper. The result has shown that the method can improve the freshness of data and also can improve the discovery efficiency of new data, improve service quality of Deep Web information integration effectively. Several areas of future work such as improving the acquisition accuracy by further optimizing the crawler algorithm, the update frequency and other value function criteria will be studied in future.

6. Acknowledgment

The Project of Science Research Foundation of Nanjing University of Science &Technology(NO. 201000331); The 2010 Opening Project of JiangSu Province Support Software Engineering R&D Center for Modern Information Technology Application in Enterprise(NO. SX201003); The 2011 University Science Research Project of Jiangsu Province; A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.
7. References


