REPORT

Overview:

The report talks about two papers related to web and data integration systems-namely: "*Web-scale Data Integration: You can only afford to Pay As You Go*[1]" and "*Bootstrapping Pay-As-You-Go Data Integration Systems*[2]".

The authors of the first paper, "*Web-scale Data Integration: You can only afford to Pay As You Go"*, have said that the content in the World Wide Web is increasingly becoming structured, resulting in heterogeneity of the structured data. They have concluded that the traditional integration techniques cannot be applied in the face of such heterogeneity and scale. They have proposed a new data integration architecture PAYGO which is inspired by the concept of dataspaces and have emphasized *pay-as-you-go* data management as means for achieving web-scale data integration. In order to illustrate their claim, they have provided examples of Deep Web, Google Base and have cited annotation services like Flickr Service of Yahoo!. After discussions on each of them, the authors have cited the various issues pertaining to them and have contended that traditional system is incompetent to handle them. They then have discussed the PAYGO architecture and also a research prototype built at Google for managing web-scale heterogeneity. The goal of the PAYGO architecture described in this paper is to show that data management techniques can be applied to these collections of data, but to do so, one is required to change some of the assumptions that one typically makes when dealing with heterogeneous data.

While the first paper has talked about using data integration architecture called Pay As You Go in the web, the second paper has dealt with means of bootstrapping this process. It is noted that the most significant and time-consuming task in setting up the data integration system, is the creation of the mediated schema and semantic mappings from the data sources to the mediated schema. It is further observed that many application contexts involving multiple data sources (e.g., the web, personal information management, enterprise intranets) do not require full integration in order to provide useful services, motivating a pay-as-you-go approach to integration. This paper "*Bootstrapping Pay-As-You-Go Data Integration Systems*[2]" has described the first completely self-configuring data integration system. The goal was to investigate how advanced of a starting point one can provide a pay-as-you-go system. The system is based on the new concept of a probabilistic mediated schema that is automatically created from the data sources. The paper also discussed techniques to automatically create probabilistic schema mappings between the sources and the mediated schema. It has further discussed the various algorithms that have been developed to facilitate such a process and also the experimental results along with the performance metric they have considered.

1. Jayant Madhavan, Shawn R. Jeffery, Shirley Cohen, Xin (Luna) Dong, David Ko, Cong Yu, Alon Halevy Google, Inc
2. Anish Das Sarma, Stanford University California, Xin Dong ,AT&T Labs–Research New Jersey, Alon Halevy Google Inc. California,

Comments:

The first paper “Web-scale Data Integration: You can only afford to Pay As You Go” is a “high-level” paper abstracting the implementation details of the PAYGO based data integration system, merely touching the issues faced and illustrating the drawbacks of the traditional data integration systems and why PAYGO architecture is superior. The paper lucidly describes in a structured manner, the issues faced by traditional data integration systems in the wake of high content heterogeneity in the web; however it does not provide any experimental results regarding the performance of PAYGO architecture.

Some of the questions that were asked during my presentation of this paper are:

- *Why is this system required, as there are already existing techniques to handle heterogeneity?*

Yes, there are techniques to handle heterogeneity; however the important point here is the scale of the heterogeneity. The existing techniques do not take into consideration such a large scale of heterogeneity.

- *How the PAYGO system is able to identify which content is value and which is attribute in a query?*

The Structure index, Value index provide us with knowledge of which is a data value and which is an attribute value. Further using schema clusters one can identify which domain has to be selected.

- *The selection of domain is based on the clustering. Can selection of domain go wrong because of clustering?*

The paper does not discuss about the clustering algorithm, however technically it can.

The second paper “Bootstrapping Pay-As-You-Go Data Integration Systems” discusses in great detail the implementation of the first self configuring data integration system.

Some of the questions that were asked during my presentation of this paper are:

- *Have the authors checked for the performance of the system if the number of data sources is increased?*

There is no mention of this in the paper; however in my opinion the time taken should not exceed linear growth rate.

- *Is the consolidated mediated schema prepared at runtime?*

I do not think so. The entire process is created beforehand which allows the users to pose queries on the consolidated schema. However I am not sure about this.