Data Integration for the Relational Web

Report for the course hy562

Michalis Katsarakis
katsarakis@csd.uoc.gr

26 June 2012
Contents

1. Contribution Summary (Περιλήψη Συνεισφοράς)
2. Range of Result Application (Εύρος Εφαρμογής Αποτελεσμάτων)
3. Powerful Points (Ισχυρά Σημεία)
4. Weak Points (Αδύναμα Σημεία)
5. Open Topics for Future Research and Possible Extensions (Ανοιχτά θέματα για μελλοντική ερεύνα και πιθανές επεκτάσεις)

References
1. Contribution Summary

(Pερίληψη Συνεισφοράς)

The web contains a vast amount of structured information such as HTML tables, HTML lists and deep-web databases. There is enormous potential in combining and re-purposing this data in creative ways. However, integrating data from the relational web raises several challenges that are not addressed by current data integration systems and mash-up tools.

The paper [1], which was published in the conference VLDB2009, presents “Octopus”, a novel system that enables users to integrate relational data extracted from web pages. Except from the traditional relational database operators (e.g.: `project`, `split`, `union`), Octopus provides its users with three database integration operators, namely `Search`, `Context` and `Extend`. The `Search` operator performs 2 tasks: it searches the web for relations relevant to the query string that the user provided and then clusters relations with similar schemes together. The `Context` operator modifies a relation to contain additional columns, using data derived from its source Web page. Finally, the `Extend` operator enriches a relation with more columns by performing a join with additional relations found on the web. Each Octopus operator is implemented by two or three alternative algorithms, which are benchmarked regarding the quality of their results.

Octopus has contributed in the research progress for Data Integration in several ways. It has proved that it is possible to integrate relational data extracted from dirty and incomplete web sources, achieving relatively good quality of results. It has successfully organized the procedure for creating an useful dataset in three operators. For each operator, Octopus has proposed several algorithms, covering a spectrum from naive to very elaborated implementations. For each algorithm, both its computational complexity and the quality of results are examined, giving the reader an overall impression.

Finally, in section 4, the authors discuss the capability of Octopus to be implemented at scale. The difficulties stated regard the runtime complexity of the implemented algorithms and their impact on the potential to provide low latency to a mass audience. The authors also describe the techniques they applied in order to overcome these challenges.
2. **Range of Result Application**  
**Εύρος Εφαρμογής Αποτελεσμάτων**

The Octopus system described in [1] deals with a hot research topic and further progress is needed to be done, in order for new end-user products to be launched. Despite that, the developed algorithms and the results of the performed experiments, as much as the innovative ideas, can be useful for a variety of already existing products (e.g.: search engines, search modules of various applications and information retrieval products). Also, they can be used to support further research on the fields of data integration, information retrieval and knowledge harvesting from the web.
3. **Powerful Points**  
(Ισχυρά Σημεία)

The basic idea of the paper seems to be very interesting and highly novel. The developed algorithms and experiment results also contain some interesting parts.

The writing style used in this paper is very understandable. Always are used the exact words, the appropriate vocabulary and the right terminology.

The implementation of the algorithms is included in the paper in the form of figures, which are supported by detailed captions. In addition to that, all algorithms are explained using the appropriate amount of detail, allowing the reader to fully understand them, while avoiding to become verbose.

The goal and the followed procedure of every experiment is defined clearly. The results are presented using clear plots or tables and are described well, explaining every interesting or “strange” result.

The above merits of this paper successfully make a positive impression to the reader, encouraging it to read more.
4. **Weak Points**

(Αδύναμα Σημεία)

Section 4 discusses the capability of Octopus to be implemented at scale. Despite the mentioned complexity limitations of several algorithms and the used solutions, it would be interesting to include some performance tests and include measurements like the mean delay of each algorithm for given equipment versus the size of the web corpus or the cardinality of the query words. Such a performance analysis would let the reader immediately understand which algorithms could be immediately adopted, which are their scalability limitations and which algorithms need further improvements.

Except from the above, there are not any other weak points detected. The overall rating of [1] should be high, mentioning this with confidence.
5. **Open Topics for Future Research and Possible Extensions**

(Ανοιχτά θέματα για μελλοντική ερεύνα και πιθανές επεκτάσεις)

As mentioned in the previous sections, [1] discusses an open topic of research and thus there are many possibilities for further work. Of course, there are open topics in the optimization of the already implemented algorithms and procedures, but these topics are mentioned in the cited paper. This section focuses in further extensions of the Octopus system.

The first task of Octopus is to extract relational tables from web pages. In its current state, it can retrieve such relations from HTML tables and HTML lists. These two capabilities can provide us a large amount of relations. Despite that, there is room to implement more tools, capable of extracting data from more kinds of elements. Crawling data which is hidden in back-end databases and accessed through valid form submissions, socially created data sets or office-style documents would provide us with new kinds of information, currently unavailable.

Another interesting topic would be to keep anonymous statistics about Octopus users’ activity and request from them feedback about the quality of the data they finally managed to achieve. By recording such feedback, it would be possible to detect which data sources (or combinations of data sources) result high-quality relations. This knowledge could be used to promote qualitative sources in the ranking phase of the Search operator, or even to cache the qualitative results and be able to re-serve them quickly in case of a similar query.

One more idea would be to develop a cloud based service, that would facilitate the creation and publication of structured data on the Web, therefore complementing the Octopus system.

Despite from extracting data from HTML tables, lists or other structures, it would be useful to recognize column names, table schemes and any kind of metadata that could accompany the extracted data. This would open the way to implement, in addition to the graphical user interface, a semantic web service, enabling the automatic use of Octopus services by other applications.
References