XML-based Data Integration for Semantic Information Portals

Patrick Lehti, Peter Fankhauser, Silvia von Stackelberg, Nitesh Shrestha
Fraunhofer IPSI, Darmstadt, Germany
lehti,fankhaus,sstackel@ipsi.fraunhofer.de

Abstract:
This paper addresses some classical problems to integrate data sources that are heterogeneous in structure with possibly redundant data along a real world example integrating three bibliographic data sources. We show how XML-technology can be applied for the data integration process in a straight-forward manner to populate a data warehouse, how an ontology can be used as common schema, and how a generic, declarative approach can increase flexibility and scalability. Our procedure enables more advanced query functionality for integrated data sources.

1 Introduction

Searching for scientific publication data in the Web is rather time consuming, as most of the available web pages for publication data (e.g., DBLP [Le02], CiteSeer [LGB99], LeaBib, ACM Digital Library, among others) are limited to specific research domains, and some lack in completeness of their collections (e.g., conferences over several years). From the providers' perspective, it is costly to keep these collections consistent and up to date. To overcome these drawbacks, cooperations among the providers have been established. One of these cooperations run within the FIS-I project, aiming to built the so-called IO-Port [HLL+03] to enable uniform access to scientific publication data from several providers (DBLP, CompuScience, LeaBib, among others).

The SemIPort project[1] cooperates with the FIS-I project and aims at the development of semantic methods and tools for portals [AFGO+03]. In this paper, we focus on methods to populate a semantically enriched portal. We address the following issues:

Harvesting of publication data. Publication data that are not directly accessible (e.g., located at external sources) will be collected by crawling corresponding web pages.

Schema Integration/Syntactic Enrichment. Data sources are typically heterogeneous in syntax, structure, and semantics. Therefore these sources are at first transformed to an XML syntax and homogenized in their structure.

Instance Integration/Data Cleansing. As far as data sources provide overlapping data, it is

[1]The project SemIPort (Semantic Methods and Tools for Information Portals) is funded by the Federal Ministry of Education and Research (bmb+f)

needed to detect duplicates. When there exists no globally unique key for this identification, one challenge is to find reasonable content-based keys to specify identical instances. Furthermore it is necessary to detect and reduce or (ideally) remove inconsistent data.

As a result, we arrive at a warehouse comprising three sources which can be queried in an integrated manner, enabling more advanced query functionality.

2 Data Integration Using XML Technology

We currently integrate three bibliographic data sources: CompuScience, DBLP, and CiteSeer. For CompuScience and DBLP, we use local copies, to access CiteSeer data we use a harvesting approach.

2.1 Schema Integration

CompuScience and DBLP both collect bibliographic references. DBLP (around 370 000 references) uses essentially the schema underlying the bibtex format represented in XML, whereas CompuScience (around 220 000 references) uses a proprietary, flat format represented as key value pairs. In order to query the two sources in an integrated manner, both sources have been transformed to XML, and their underlying schemas have been integrated, such that equivalent schema constituents (including title and authors) are represented uniformly, whereas complementary schema constituents (differentiated source information in DBLP, keywords and abstract in CompuScience) are represented in their original (XMLized) form. DBLP sources have been transformed to the integrated schema with XSLT, the CompuScience data have been parsed and transformed to the XML schema with JavaCC.

Altogether, the flexibility of XML in representing deeply nested, heterogeneous structures with partially missing information has largely facilitated the design of an integrated schema that could be instantiated with reasonable effort.

2.2 Instance Integration

Even though CompuScience and DBLP have a different focus, they are overlapping. DBLP specializes on literature in the special fields of logic programming and databases, whereas CompuScience tries to cover the entire field of computer science. The goal of instance integration is to identify bibliographic references occuring in both sources.

With the schemas at hand the most obvious candidate properties for integrating instances are the title, the authors, and the publication year of a bibliographic reference. Other candidate properties such as issn/isbn or volume and number of the source where not available in both sources. The available properties do not constitute a perfect key in the individual

nr.	DBLP	CS	DBLP=CS
all	363445	215562	-
key0	361068	209439	55152
key1	360975	209299	57129

Figure 1: Number of different instances

databases, therefore instances can not be put into a $1:1$ correspondence. On the other hand, the properties author and title are a bit dirty, i.e., they can contain typos, inconsistent usage of special characters, delimiters, and abbreviations. Therefore, these properties can also not detect all possibly corresponding instances. Taking this into account we have experimented with the following two property combinations:

One combination (key0) normalizes authors' last names and titles by normalizing special characters, discarding delimiters, and neglecting case. The other combination (key1) further abstracts titles by taking into account their vowels.

Figure 1 shows the number of overall instances in DBLP and CompuScience (first line), and the number of different key-values for the property combinations key0 and key1 together with the number of instances identified by them (second and third line). We see that about 25 % of CompuScience records correspond to about 15 % of DBLP records. Moreover, the further abstraction performed on titles in key1, only insignificantly decreases its degree of uniqueness ($100 - 150$ fewer values), but increases the number of matching instances by almost 2000.

2.3 Harvesting CiteSeer Data

The CiteSeer site is one external bibliographic resource providing additional features compared to other available bibliographic resources. For example, one can get the link to pdf-files, the abstract, *cited by* and *cited to* information apart from the main bibtex entry.

We do extensive crawling to the CiteSeer site to extract the bibliographic data. One can pose a query to the CiteSeer portal by giving the author name, or any keyword as string. Such a query results in publication links. Our aim is not to limit the crawling to these matched links, but to follow the links to parse their addressed publications and to extract links to further publications. For this reason, the developed CiteSeer crawler is recursive. The given query extracts the links to publications which are stored in a so-called links reservoir. All links are then crawled to harvest bibliographic information, which is extracted by a bibliographic parser developed in JavaCC. The bibliographic parser generates XMLized data as well as accumulates the available *cited by* and *cited to* links. The crawler stops, when it reaches the defined depth or number of links (provided by the crawler software), or when the links reservoir gets empty. Finally, the harvested data (publication data and links) is locally stored in XML.

3 Beyond XML

3.1 Lessons Learned

The described, straight-forward approach provides reasonable results that can be achieved with low implementation effort. But this procedure leads to a monolithic architecture. Schema mapping and instance integration are expressed in a procedural way which makes the validation of the transformation as well as the addition of new sources difficult. Therefore an approach for declarative and incremental integration of data sources is developed in parallel, that will now be described in more detail.

3.2 Declarative and Incremental Integration of Data Sources

This approach relies on an explicitly defined global schema [LF04]. This schema is formulated as an OWL ontology, as a very generic and powerful schema language. The global schema is either developed based on the application requirements or more typically, based on the schemas of an initial set of data sources. The querying/presentation step is only based on this global schema and therefore independent of the actual sources.

For posing queries against an OWL schema, we use SWQL (Semantic Web Query Language). SWQL is a declarative, strongly typed and functional query language, based on XQuery syntax. SWQL uses OWL as its type system, and a corresponding graph data model instead of a tree based XQuery data model. This language enables to query ontologies and their instances [LSH03].

In a first step every data source is translated into valid instances of the global schema. This translation is done based on a declarative mapping, between the local schema and the global schema. These mappings are again specified using SWQL, which enables checking these mappings for consistency, due to the declarative and strongly typed nature of SWQL.

The instance integration step (re-)analyses the quality of single or combined attributes as ID for all integrated data sources. The quality of an ID can be measured e.g., by its uniqueness. This results in weighted attributes, which can now be used for automatic instance alignment.

4 Related Work

As opposed to the FIS-I portal, where manually collected bibliographic data are integrated in one database, the CiteSeer [LGB99] portal uses search engines and crawler to efficiently locate and collect bibliographic data from the Web. Consequently, CiteSeer has a high coverage of bibliographic data, but these data have a less explicite structure as opposed to data sources such as from DBLP or CompuScience. CiteSeer employs full-text of publications, which results in *cited by*-relationships, we consider only bibliographic metadata. Niedereé

et. al [NSM04] focus on the use of rules to support the identification of implicit, domain specific relationships. To this end, an ontology is used to describe the domain and the type of relations that can be plausibly inferred. This work might be used to provide more advanced queries on integrated data sources.

5 Conclusion and Future Work

We are currently investigating in the evaluation of our tools within the FIS-I portal. In the future, we want to investigate in several improvements of our tools. One of them is to increase recall by using similarity search based on edit distances for the publication alignment, as spelling errors are likely. Another aspect is to further semantically enrich the data. For example, for aligned publications, knowledge on authors' first name can be employed to enrich data sources in which authors' name is abbreviated. The same procedure can be applied for publication data (e.g., given volume and number of a journal, or the given ISBN).

References

[AFGO+03] Agarwal, S., Fankhauser, P., Gonzales-Ollala, J., Hartmann, J., Hollfelder, S., Jameson, A., Klink, S., Lehti, P., Ley, M., Rabbidge, E., Schwarzkopf, E., Shrestha, N., Stojanovic, N., Studer, R., Stumme, G., Walter, B., und Weber, A.: Semantic methods and tools for information portals. In: *Proc. of Informatik 2003*. volume 1. pp. 116–131. September/October 2003.

[HLL+03] Horn, L., Ley, M., Liksch, P., Maas, J., Mayr, E. W., Oberweis, A., Ortyl, P., Pfingstl, S., Rossi, E., Rüssel, F., Rusnak, U., Sommer, D., Stucky, W., Vollmar, R., und von Mevius, M.: Konzeption und Betrieb eines Kompetenz- und Dienstleistungsnetzes für die Informatik. In: *Proc. of Informatik 2003*. volume 1. pp. 132–147. September/October 2003. in German.

[Le02] Ley, M.: The DBLP computer science bibliography: Evolution, research issues, perspectives (invited paper). In: Laender, A. H. F. und Oliveira, A. L. (Eds.), *SPIRE*. volume 2476 of *LNCS*. pp. 1–10. Springer. September 2002.

[LF04] Lehti, P. und Fankhauser, P.: XML data integration with OWL: Experiences and challenges. In: *Symposium on Applications and the Internet (SAINT 2004)*. pp. 160–170. IEEE Computer Society. January 2004.

[LGB99] Lawrence, S., Giles, C. L., und Bollacker, K.: Digital libraries and autonomous citation indexing. *IEEE Computer*. 32(6):67–71. 1999.

[LSH03] Lehti, P., Shreshta, N., und Hollfelder, S. The semantic web query language SWQL. September 2003. available at $http$: $//www.ipsi.fraunhofer.de/oasys/projects/semiport$.

[NSM04] Niedereé, C., Stewart, A., und Mehta, B.: A multi-dimensional, unified user model for cross-system personalization. In: Ardissono, L. und Semeraro, G. (Eds.), *Proc. of the Workshop on Environments for Personalized Information Access (AVI)*. pp. 34–54. May 2004.