Application of MashUp Technology for Dynamic Integration of Semi-Structured Data

Irina Kushniretska[1], Andriy Berko[2]

[1]Information Systems and Networks Department,
Lviv Polytechnic National University, UKRAINE, Lviv,
S. Bandery street 12, E-mail: presty@i.ua

[2]General Ecology and Ecoinformation Systems Department,
Lviv Polytechnic National University, UKRAINE,
Lviv, Generala Chuprynky street 130,
E-mail: berkoandriy@yandex.ua

Abstract – The usage mashup-technology for the dynamic integration of semi-structured data has been considered in this paper. Architecture of web-mashup, which consists of the three parts has been described. The process of the creation the mashup-service on data level, on process level and on presentation level has been presented. Four specific search strategies that use different kinds of information for query ranking and selection have been proposed.

Key words – dynamic integration, mashup-technology, mashup-service.

I. Formulation of the problem

Today, the quite large amounts of semi-structured data of various nature is accumulated in web-systems. Due to the continuous development of information technology this data continue to grow rapidly. Hence, there is a need for operational, dynamic integration of these data for ease of presentation and further use.

The problem of dynamic data integration is extremely multifaceted and diverse. The complexity and nature of the methods used to solve it essentially depends on the level of integration that is necessary to ensure, of the properties of individual data sources and the totality of sources in general and of the methods of integration, that is necessary. The actual and unsolved problem is to overcome the heterogeneity of sources of information resources and the development of mechanisms of the dynamic and the semantic integration of the semi-structured data in web-systems. Another critical task is measures and means of the semantic data coherence as on the local (source repository) and on the general (source registers) levels.

II. Analysis of recent research and publications

Data integration in information systems is understood, as the providing of the single unified interface to access a certain set, generally speaking, independent heterogeneous data sources [2, 3]. Thus, for user the information resources of the totality of the integrated sources are presented as a single new service. A system that provides the user such opportunities called the data integration system.

For many years, the scientific and engineering community of the world are interested of the problem of dynamic integration of information resources, but only in recent years the emergence and the active development of modern infrastructure (WEB and GRID environment) and open service-oriented architectures in the field of information technologies (SOA, OGSA), and also significant progress in the development of relevant international basic standards for information exchange (XML, RDF, TM, OWL) allow you to create a fundamentally new models of information systems. Such models allow to build the globally distributed applications, that implement the technological chains, in which can be used not only own IP, but also those that may be offered by other organizational structures. Take into account the fact, that when working with this application you can replacing of one information service to others, delete the ones that have lost value or relevance, and add new ones.

Today, if speaking about the dynamic integration of data in web, the acquire increasing importance the research and development the systems of the integration that are using technology MashUp.

MashUp is the technology that integrates data from multiple sources into a single integrated tool [2]. MashUp is a form of data integration technology, that adapts them to integrate many technologies and languages of. Some mashups are the combination of JavaScript-code with XML and they create new innovative web service. Other, larger mashups that are the basis of relevant web sites use the technology such services as Gooogle Maps and the address database, linking them together and displaying information about the project on the map [2].

Unfortunately, there is no generally accepted classification mashup-applications. Since the one of the most important tasks, when creating mashup, is the receiving of data, it is reasonable to make the classification of the data types with which operates mashup. There are four basic categories:

- maps;
- media content, video and photo;
- news;
- search and buy.

Application should not belong to any specific category. For example, an application that receives of the news feed from the multiple sources, extracts from which the information about the scene of the event and makes a mark on a map, obviously, will belong to the first and the last type.

Currently there are several mashup platforms that help the user to create mashups. This, for example: GoogleApps, IBMLotusMashups, Intel Mash Maker, Yahoo! Pipes or others.

Architecture web mashup always consists of three parts [1]:

1. Content provider (data level) is a data source. Data is available through the API and various web protocols such as RSS, REST and web-services.

Content provider is one of the most important parts of the system, because it determines how and in what way will the data come from third-party services. It is believed that mashup should use only those services that provide the special API.

2. Mashup site (process level) is a web-based application that provides a new service that uses a data source that does not belong to him.

3. Browser of client (presentation level) - the actual user interface mashup. In web applications, content can be "mashuped" by the client browser using the client programming language, for example JavaScript.

Each data source needs to be first analyzed and modeled in order to perform the required actions of retrieval and preprocessing [2]. The data level is mainly concerned with accessing and integrating heterogeneous web data sources. These sources can provide structured, semi-structured or unstructured data. Existing data mashup tools cannot deal with structural and semantic diversities of heterogeneous data sources. However, to the best of our knowledge, there exists no data mashup tool which allows the user to formulate queries over web data sources using their respective query languages and at the same time deals with the heterogeneity of the data sources.

III. Presentation of the basic material

The process of creating web-mashup consists of a tree-level building: at the data level and at the process level and at the presentation level (Fig. 1).

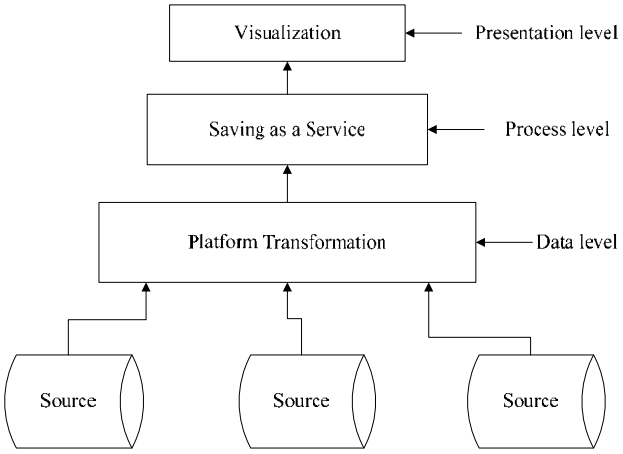

Fig. 1. Tree-level building web-mashup

The process of the creating of mashup-service:

At the data level:

1. Convert iCal in RDF (if necessary);

2. Filtering the data (eg. name, event name, start date, etc.) SPARQL queries;

3. Mashup! (Identify the same events, the involved people) SPARQL design;

4. Saving and publishing.

At the process level:

1. Extraction of information from the data level;

2. Definition of the choreography between the involved applications;

3. The Describing of the composition using either programming languages such as Java, or dedicated workflow languages such as WS-BPEL;

4. Preparing data to visualization and mashup again!

At the presentation level:

Data visualization (HTML page, or a more complex web page developed with Ajax, Java Script, etc.).

Basic of the work of mashup service is a combination of information from different data sources, as a result of user's query. To this combination was the most successful you need to apply the right of search strategy. After analyzing the pros and cons of mashup service we select four specific search strategy:

1. Parallel - this search strategy fulfills all the demands of all available generators request.

2. Serial - this strategy executes queries according a fixed order of query generators.

3. Optimistic - this search strategy executes queries according to the number of covered entities and thereby prefers queries with a large coverage over other queries.

4. Preliminary assessment - This strategy makes the most advanced searches based on performance (ie, efficiency and effectiveness) previously executed queries of the same query generator. The approach is based on a preliminary evaluation of the search results for all queries and generators on a common training set of data input.

All search strategies use the function of the evaluation of query and the function of the selection of query. The evaluation function assigns a score to each request. All requests are ranked by their estimates (in ascending order) and then the selection function filters the requests according to their ranking.

Conclusion

This paper describes the MashUp technology for dynamic integration of semi-structured data. MashUp is a relatively new technology that is gaining in popularity among web-developers. Its advantages include the possibility of convenient simultaneous visual analysis of data from multiple sources, it has relatively simple implementation and modification in accordance with the needs of the user and also it enhances the visualization of data from existing systems. Technology "mashup" allows you to quickly create new useful web-application, but at the same time and it makes higher demands on the security, integrity and availability of technical data to be provided with data integration systems. Availability of data in the open access is the main engine of growth in popularity "Mashup applications".

References

[1] A. Vancea, M. Grossniklaus, and M. C. Norrie, “Database-driven web mashups,” in Proc. ICWE, pp. 162–174, IEEE, 2008.

[2] G. D. Lorenzo, H. Hacid, H.-Y. Paik, and B. Benatallah, “Data integration in mashups,” SIGMOD Record, vol. 38, no. 1, pp. 59–66, 2009.

[3] Levy A.Y. Logic-Based Techniques in Data Integration. Logic-based Techniques in Data Integration. In: Logic Based Artificial Intelligence. Edited by J. Minker. Kluwer Publishers, 2000.