Mash-Up system, Mash-up system content, metadata triplets, similarity factor

Irina KUSHNIRETSKA[*]

FORMING OF THE SEMISTRUCTURED DATA DYNAMIC INTEGRATION MASH-UP SYSTEM CONTENT

Abstract

This paper describes the method of forming a united dynamic data set that has the general structure and only content. The procedure of forming the triplets with the structure "subject-predicate-object" with the received input information resources descriptions has been proposed. The formula of calculating the similarity factor of user query with information resource semantic metadata has been presented. The structure of the semistructured data dynamic integration system that used "Mash-Up" technology has been designed.

1. INTRODUCTION

Participation of end-users is an essential driving force of web technologies developing. Although the application of Web 3.0 allows the spread of Internet using to more and more areas, is still remain the controversial issue: how not specialized ordinary users can interact with them and be more than just a receiver of data. The current state of Web 2.0 already provides opportunities for end users to evolve from simple consumers of information in information developers. This allows the new data integration systems working using Mash-Up technology. Currently, such systems are actively developing and allow users to collect, transmit and use Web resources. The purpose of these systems is to create new and useful applications from available web resources.

However, existing methodologies and tools for software systems building are focused on the well-structured problem with sufficiently formalized subject areas and permanent local knowledge sources. So many problems associated with providing semantic of information during the semistructured data Mash-Up-system dynamic integration remain unresolved.

[*] Lviv Polytechnic National University, Ukraine, 79013, Lviv, Bandera str., 28a, presty@i.ua

2. THE PROBLEM FORMULATION

The importance of various semistructured data integrating, today, hard to overestimate. The ability to quickly and finding the necessary quality information to make informed decisions adequate and necessary information required to ensure the process of storage, storage, analysis and interpretation of all required data. Using web-search tools you can find a huge number of diverse

Information specific topics, but not always achieved with the necessary coherence and consistency of the data. Integrity, consistency, Data consistency can be effectively achieved only if the use of special methods of information processing single centralized system regardless of the architecture and implementation of information infrastructure organization. Building such infrastructure is lengthy and time consuming process, the complexity of which depends not only on the volume of accumulated historical data, as the number and diversity of sources and different applications. Therefore, crucial in this process certainly plays the correct choice of methods and means of data integration.

Hence, the aim of this work is to research of the application of new approaches for solving the problem of increase the quality of semistructured information received from various web-systems storage and presentation by forming a united dynamic data set that has the general structure and unique content.

3. THE BASIC MATERIAL PRESENTATION

3.1. Dynamic data integration systems based Mash-Up technology

Recently Mash-Up technology became a trend which allows non-professional users to create Web applications, combining functionality with more than one with the important task during Mashup system creating is data getting Web-services to solution the situation and specific tasks.

Mashup-systems often focus on one or more of the following objectives:

- Extract data from existing data sources such as Web-pages, Web-services and channels.
- Data combining from different sources into a single data set. The data from each source must have at least one general area, for example, name of the restaurant, so that one set of records could be matched with corresponding entries in the other sets.
- Data visualization in a way that allows the user to understand the aggregated data. Many collages include geographical data, such as addresses (visualized using Google Maps).

Unfortunately, the general classification of Mashup applications is absent. Because one of the most important tasks when Mashup system creating is data getting, it is wise to make a classification of the data type with which Mashup system operates. There are four main categories: maps; media content, video and photos; news; search and purchase.

The main objectives for the Mashup tools searching [4]:

1. Achieving more efficient Mashup construction;
2. Achieving of integration of Mashup systems developing as part of the software development process;
3. Conduct research and design Mashup existing tools.

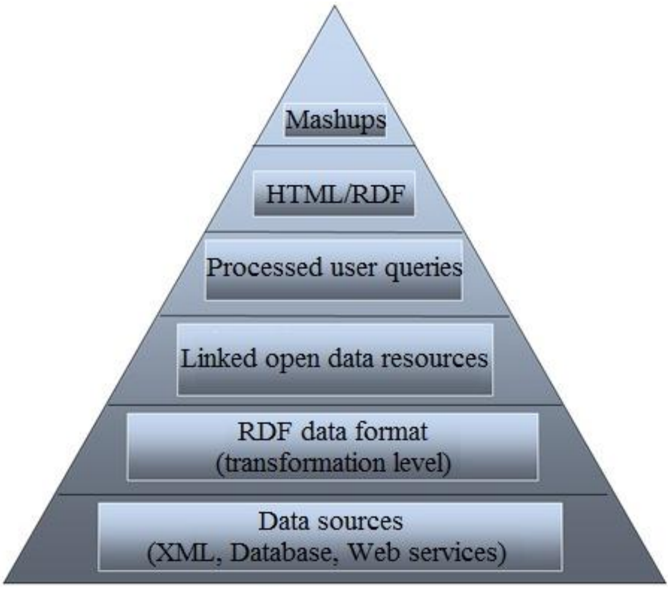

Fig. 1. "Pyramid values" of the semistructured data dynamic integration system based on "Mash-Up" technology

Today we know a number Mashup systems that help the user to create the data mashups. We chosen the following systems for three main reasons:

1. These systems are most popular when this analysis was performed (based on discussion forums, blogs, etc.).
2. Information about some other systems could not be complete because there is unavailability of systems data at certain stages of research, not allowing us to experiment and report results according to our analysis.
3. The reason for this is motivated by the fact that our aim is not to analyze all systems, but provide an indication of the current state of these systems and understand their overall approach to data integration.

IFTTT [5] ("if this, then that") – Mashup service that allows users to connect to various web applications using simple conditional statements, known as "recipes" and create simple automated sequence that runs when performing an action. IFTTT was developed by the American programmer Linden Tibetson and launched in 2010. IFTTT allows users to create and share "recipes" that correspond to the judgment: "if this, then that" "it" part of the recipe, called trigger. Pretty simple to use and consists of only three tabs:

- Tasks – a list of your active tasks.
- Recipes – a list of the most popular tasks that you can use as their own, that is something like pieces tasks.
- Channels – a list of supported services at the time of the study 54 (e.g., Facebook, Twitter, LinkedIn, RSS, Google RSS Reader, Evernote, Gmail, Google Calendar, WordPress, etc.).

Example working with IFTTT: When we update status on Facebook, want it to be immediately twitt on Twitter. In addition to its Web application, released a mobile version for iPhone service in July 2013. Android-version of the application was released on April 24, 2014.

One with the worthy analog IFTTT can be called a service called Zapier [6], which was created in 2012 by the American development team: Brian Helmihom, Wade Foster and Michael Knop. It fully inherited the trigger-action circuit and allows you to build relationships between different web applications. Unlike IFTTT Zapier supports 147 channels, 2 times more than ifttt, and also its interface for many people seem more intuitive and simple. But unlike the completely free IFTTT, Zapier has 4 packages - one free and three paid with a limited number of services, tasks and trigger's actions for each packet.

Yahoo pipes [7] – web-tool provided by Yahoo. Users can create collages by aggregating and processing of web-channel, web-pages and other services. Pipe consisting of one or more modules, each of which performs a certain task, for example, such as: receiving data from web-sources, filtering, sorting or merging channels. The resulting system data can be accessed by the client with a unique URL as RSS or JSON, or through visualization on Yahoo map.

Google Alerts [8] – service from the search giant. It is based on the idea of query results monitoring according to the time changing. In fact, you can set up "alerts" to new results for. The system is able to filter out "all rubbish" and select the most relevant data.

So, in order to get the Mashup system work quality results should provide the necessary for this integration, coherence and consistency of the data. So critical for this problem solving is certainly in the right choice of methods and means of data receiving, storing and presenting.

3.2. Materials and research methods of the combined dynamic data set content forming that has the general structure and united content

For received input information storing in the system in a structured way useful will be using such semantic-oriented technologies such as ontology and description logic [9].

According to [10], ontology based on description logic is a signs system:

$$O_{DL} = \langle C, CD, R, A, I, V, R_I, A_I, L, P_C, P_R, P_A, P_{IC}, P_{LC}, P_{LR}, P_{LA}, P_{LI} \rangle \quad (1)$$

where: $C = \{c_1, ..., c_n\}$ – concepts finite set in ontology,

$CD = \{cd_1, ..., cd_k\}$ – standard data types set, including two types {string, integer},

$R = \{r_1, ..., r_m\}$ – binary relations final set between concepts,

$A = \{a_1, ..., a_w\}$ – finite set of attributes (binary relations between concepts and standard data types),

$I = \{i_1, ..., i_z\}$ – finite set of instances in the ontology,

$V = \{v_1, ..., v_q\}$ – standard type specific values finite set,

$R_I = \{ri_1, ..., ri_m\}$ – concretized relations finite set (binary relations between instances) $ri_I(i_x, i_y)$,

$A_I = \{ai_1, ..., ai_w\}$ – concretized attributes finite set (binary relations between instances and i specific values $ai_I(i_x, v_y)$,

$L = \{l_1, ..., l_f\}$ – lexical label final set (ontology dictionary),

$P_C \subseteq C \times C, P_C \in R$ – anti symmetric, transitive, anti reflexive binary relation, which is the relation of the partial order on the concepts set C,

$P_A \subseteq A \times A$ – anti symmetric, transitive, anti reflexive binary relation, which is the relation of the partial order on the attributes set A,

$A = \{a_1, ..., a_w\}$ – finite set of attributes (binary relations between concepts and standard data types),

$P_R \subseteq R \times R$ – anti symmetric, transitive, anti reflexive binary relation, which is the relation of the partial order on the relations set R,

$P_{IC} \subseteq I \times C$ – incidence binary relation between sets I and C,

$P_{LC} \subseteq L \times C$ – incidence binary relation between sets L and C,

$P_{LR} \subseteq L \times R$ – incidence binary relation between sets L and R,

$P_{LA} \subseteq L \times A$ – incidence binary relation between sets L and A,

$P_{LI} \subseteq L \times I$ – incidence binary relation between sets L and I.

The ontology definition in formula (1) is given with regard to the description logic properties.

When the user's query analyzing and answer to it forming in the form of information resources integrated set there arises the problem of comparison on the similarity of the query and integrated data. To solve this problem is proposed to use the similarity strings metric. Currently there are a number of commonly used universally metrics of determining the similarities of two text strings [11]: Lowenstein distance or edit distance, Zhakkar-Winkler coefficient, Tanimoto coefficient and Severensen-Dicey coefficient.

Let X and Y - two strings of length m and n. For Lowenstein distance getting are calculated the distances matrix D, in which each element D [i, j] contains the distance between the first character I of string X and the first character j in the string Y. Lowenstein distance is determined by the following formula:

$$D(i,j)=\left\{\begin{array}{l}\max(i,j), if\ \min(i,j)=0\\ \min\left\{\begin{array}{l}D(i,j-1)+1\\ D(i-1,j)+1\\ D(i-1,j-1)+m(X[1],Y[j])\end{array}\right\}\end{array}\right\} \qquad (2)$$

where: $m(a,b)=0$ if $a=b$,
$m(a,b)=1$ if $a\neq b$.

Tanimoto coefficient or Zhakkar-Winkler also often are used to determine the similarity of one string to another. In this case, the string is seen as a set of characters, and the similarity metric determines the amount of the same characters as follows:

$$p=\frac{c}{a+b-c} \qquad (3)$$

where: p – strings similarity factor: $0\leq p\leq 1$,
c – number of joint characters in strings,
a and b – number of characters in strings A and B respectively.

By the same principle operates Severensen-Dicey coefficient, which is a binary measure of strings similarity and has in general the following entry:

$$p=\frac{2\times n(X\cap Y)}{n(X)+n(Y)} \qquad (4)$$

where: p – strings similarity factor: $0 \le p \le 1$,
X and Y – compared strings,
n(a) – function that determines the number of characters in the string a.

3.3. The method of the combined dynamic data set, which has the general structure and united content forming

In [12] based on the Mashup systems activity analysis are the following activity states:

1. Registration. When there is the successful registration – move to the second step, the unsuccessful registration – returning back to the registration beginning.
2. Authorization, if all goes well – move on, if not – back to the authorization beginning.
3. Task formulation. If the task is formed according to the rules of the system, we move on, if not – return to the beginning of the third step.
4. Mashup system sources forming.
5. In selected sources of the required information finding. If the results are satisfactory – we go further, if not – return back to the searching.
6. Relevant information extracting and to the next state transition.
7. Received information storing in the service form.
8. The finished Mashup system result visual presentation.

The most important states in Mashup system work, according to [12], are: required information finding (fifth state), found information extracting (sixth condition) and storage as a service (seventh state). To improve of the result receiving for seventh state work in [12] the procedure of determining the structure and content of incoming information resources has been proposed. This procedure can be interpreted as a method of determining the structure and content of the received input information. The usage of this method makes it possible to increase the quality indicators of the fifth state result. And, as every subsequent state depends on the previous one – it increases of the two next important activity states productivity.

User request analysis and answer forming to it – some of the tasks that need to be solved at the seventh state of Mashup system work. To improve of seventh state (storing as a service) work result quality indicators we propose a method of forming a united dynamic data set that has the general structure and unique content.

The method of forming the united dynamic data set that has the general structure and unique content consists of the following steps-tasks:

Step 1: Analysis of the received metadata information resources.

If semantic metadata is generated based on a textual description of the object and no clear selection of concepts and instances, then the data will process according to the second step tasks. And if all concept and instances meta description are selected, then the data will process according to the third step tasks.

Step 2: Searching of the concepts and instances in the information resources textual descriptions.

The textual description is analyzed for the presenting of concepts and instances that can serve as subjects in the semantic metadata elements. To solve the problem at this step was used the instances and concepts in the text search function. The objective of this function is to search of lexical labels of concepts and instances with ontology in the object textual description to form a set of possible subjects in the semantic metadata elements. Of course, the result of this function can’t completely replace human work and after using search function of concepts and instances you have to use of additional technologies for:

- to remove of all objects that do not reflect the essence of the object description;
- to eliminate multiple meanings if the set contains items with the same lexical labels;
- to complete the set of concepts and instances that were not found as a result of the function.

Then the elements of set can be used to form triplets according to certain rules of predicates and objects selection that we describe a little later.

Step 3: Forming of triplets with the structure "subject-predicate-object" with received semantic information resource metadata.

We need that the received our semantic metadata items be in form of the triplets with the structure "subject-predicate-object" or separate concepts or instances of ontology, which will be called the "subject". In order to structure your content semantically is necessarily need to be given "subject".

If the subject is given for the essence of the subject description presentation, then on the selection of predicate and object are imposed additional restrictions that arising from the rules of formation of the description logic statements.

The set of possible predicates in triplet is limited by chosen triplet subject. That is, as a predicate can be selected those relations or attributes, that in ontology are defined for the subject - concept or instance.

After predicate selecting is necessarily indicate the triplets object. The set of possible objects depends on the selected predicate. The rules of forming the set M_O of possible objects in triplets based on the definition of ontology in formula (1) are as follows:

1. If the predicate value – is the relation r_x, then

$$M_O = \{o_i \in C \cup I \mid r_x(c_x, o_i) \vee (r_x(c_x, c_y) \wedge P_{IC}(o_i, c_y))\} \tag{5}$$

2. If the predicate value - is the attribute a_x, then

$$M_O = \{o_i \in cd_j \mid a_x(c_x, cd_j)\} \tag{6}$$

That is, the predicate possible values are defined or the attribute values concrete area, or the relation values area.

When described rules are followed, are formed in form "subject-predicate-object" of the content semantic metadata elements. Restrictions on the items number in semantic metadata are not imposed.

Step 4: Similarity determining of user request and information resources semantic metadata triplets.

Analyzing the most popular of the strings similarity determining indicators, notting: none of them can't guarantee a good result when is changing the order of words in a sentence and using of multiple languages in a text string. In this regard, it is proposed not to use these metrics directly, and use the modified metrics that are worked based on the method of splitting words at the N-grams [13]. Using the so-called Shinhlinh method [14], which would create average N-grams, is modified of the formula for calculating the Severensen-Dicey coefficient which is a binary measure of strings similarity, using N-grams instead of characters:

$$p = \frac{2 \times (f_{NG}(X) \cap f_{NG}(Y))}{f_{NG}(X) + f_{NG}(Y)} \tag{7}$$

where: p – strings similarity factor: $0 \le p \le 1$,
X and Y – compared strings,
$f_{NG}(a)$ – function that determines the length of N-grams in the string a.

The usage of non-individual characters but word combinations allows to reduce the number of errors during of the current text string analysis.

The result that we get after the fourth step - is fully formed and stored in the system for further dynamic visual presentation combined data set that has the general structure and unique content.

4. CONCLUSIONS

The current methods of finding and receiving the data in dynamic data integration systems that are worked using "Mash-Up" technology have been researched and analyzed.

To improve of the quality indicators of Mashup system information storing and presentation result, the method of forming the united dynamic data set that has the general structure and unique content has been designed. The rules of forming the set of possible objects in triplets have been described. The procedure of the similarity determining of user request and information resources semantic metadata triplets has been characterized.

REFERENCES

[1] KUSHNIRETSKA I., BERKO A.: *Application of Mash-Up Technology for Dynamic Integration of Semi-Structured Data*. Proceedings of the Sixth International Conference of Young Scientists CSE-2013, Lviv Polytechnic National University Publisher, Lviv, 2013, pp. 220-221.

[2] ABITEBOUL S., GREENSHPAN O., MILO T.: *Modeling the mashup space*. WIDM, 2008, pp. 87-94.

[3] MAXIMILIEN E., WILKINSON H., DESAI N., TAI S.: *A domain-specific language for web apis and services mashups*. Proceedings of ICSOC'07, Berlin, Heidelberg, 2007, pp. 13-26.

[4] FISHER T., BAKALOV F., NAUERZ A.: *An overview of current approaches to mashup generation*. Proceedings of the Fifth Conference Professional Knowledge Management: Experiences and Visions, Vol. 145, Solothurn, Switzerland, 2009, pp. 254-259.

[5] *About IFTTT*. IFTTT Inc., 2015 [electronic resource] – Available on: https://ifttt.com/wtf

[6] *Zapier*. Zapier Inc., 2015 [electronic resource]. – Available on: https://zapier.com

[7] *About Pipes*. Yahoo! Inc., 2015 [electronic resource]. – Available on: http://pipes.yahoo.com/pipes

[8] *Alerts. Follow for new interesting content on the Internet*. Google Inc., 2015 [electronic resource]. – Available on: https://www.google.com/alerts

[9] BAADER F., CALVANESE D., MCGUINNESS D., NARDI D., PATEL-SCHNEIDER P.: *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003, pp. 132-136.

[10] LEVY A.: *Logic-Based Techniques in Data Integration*. Logic Based Artificial Intelligence. Kluwer Publishers, 2000, pp. 74-76.

[11] RECCHIA G., LOUWERSE M.: *A Comparison of String Similarity Measures for Toponym Matching*. Proceedings of the First ACM SIGSPATIAL International Workshop on Computational Models of Place, 2013, pp. 54-62.

[12] KUSHNIRETSKA I., KUSHNIRETSKA O., BERKO A.: *Determination of the structure and content of input information resources for Mashup system*. Technology Audit and Production Reserves, No. 6/3 (20), 2014, pp. 4-9.

[13] BARASHEW D.: *The Similar Documents Search*. Computer science. Big Data'13, 2013 [electronic resource] – Available on: http://compscicenter.ru/sites/default/files/materials/2013_04_18_BigData_lecture_09.pdf.

[14] GUDKOV V., GUDKOV E.: *N-grams in Linguistics*. Herald of Chelyabinsk State University, No. 24 (239), 2011, pp. 69-71.