sMash: Semantic-based Mashup Navigation for Data API Network

Bin Lu, Zhaohui Wu
College of Computer Science,
Zhejiang University,
Hangzhou, 310027, China
{lb, wzh}@zju.edu.cn

Yuan Ni, Guotong Xie
IBM China Research Lab,
Zhongguancun Software Park,
Beijing, 100193, China
{niyuan, xieguot}@cn.ibm.com

Chunying Zhou, Huajun Chen
College of Computer Science,
Zhejiang University,
Hangzhou, 310027, China
{cyzhou, huajunsir}@zju.edu.cn

ABSTRACT
With the proliferation of data APIs, it is not uncommon that users who have no clear ideas about data APIs will encounter difficulties to build Mashups to satisfy their requirements. In this paper, we present a semantic-based mashup navigation system, sMash that makes mashup building easy by constructing and visualizing a real-life data API network. We build a sample network by gathering more than 300 popular APIs and find that the relationships between them are so complex that our system will play an important role in navigating users and give them inspiration to build interesting mashups easily. The system is accessible at: http://www.dart.zju.edu.cn/mashup.

Categories and Subject Descriptors
H.4.m [Information Systems Applications]: Miscellaneous

General Terms
: Algorithms, Design, Experimentation

Keywords
Mashup Navigation, Data API network, Social, Semantic

1. INTRODUCTION
An increasing number of information sources have published their data in the form of open data APIs which facilitate users to fetch public data as well as their personal data. If all the data APIs[1] get together according to a certain kind of relationship to form a real-life data API network, it will be a novel idea to solve difficulties many researches currently meet, such as mashups and linked data exploitation.

As an essential transformation of the Web [1, 2], mashups, which are typically drawn upon content retrieved from external data sources by means of data API calling, bring increasing interest to users. Even though current mashup tools [3-5] are sometimes efficient and convenient for mashup building, it makes users, especially non-developers, feel confused when they have little knowledge about APIs.

Besides, our statistic result shows that because of the difficulty of being discovered and mastered by users, more than 4/5 data APIs are rarely used to build mashups even if they may supply more abundant information to satisfy users' requirements. If we can visualize these APIs and their relationships, users may build more amazing mashups.

[1] We regard data APIs as any information source that could offer their data in a RESTful way.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

To make mashup building easier and more interesting, we propose to construct a data API network that enables users to build mashups by navigation. In this network, each API is represented as a node; a link between two APIs means they have the mashupable relationship; each mashup can be regarded as a path. Base on this vision, we present a semantic-based mashup system, sMash which integrates conventional techniques: social community, semantic and collective intelligence. Our system has three main advantages:

- An automatic mashup navigation system: What users need to do is a little bit "fuzzy-match-keyword-search", and then the network is constructed and visualized around the matched APIs. The navigation is provided for users by an automatic link of mashupable APIs and a detailed mashup candidate recommendation. Besides, a global view of all the related data APIs and their relationships is presented to give users inspiration in deciding which APIs to use and how the path should be like.
- A precise way to describe the metadata of API: a RDF model is proposed as the "schema" model of API to incorporate rich semantics of metadata.
- An extendable and flexible real-life data API network: sMash provides a user friendly "schema" editor to facilitate users to contribute the "schemas". sMash keeps the definition of link between APIs configurable, which makes it easy for us to focus on semantic data search by reconstructing the network to exploit more data and links in the future.

2. MASHUP ON DATA API NETWORK
There are three main steps for providing mashup navigation for data API network: (a) Data collection. (b) Data API network construction and visualization. (c) Mashup candidate recommendation.

Data collection: Until now, we have analyzed and described more than 300 APIs using the "schema" editor in order to construct a sample network.

To describe the data content, we bring the idea of microformats to predefine the frequently-used semantic data types, e.g., "geo", "photo" and "event", and provide a data type editor to enable users to add new data types if they cannot find a proper one from the predefined data types. "Schema" editor, data type editor and the "schema" model are illustrated in figure 1.

Data API network construction and visualization: The data API network is constructed and visualized in the following three steps:

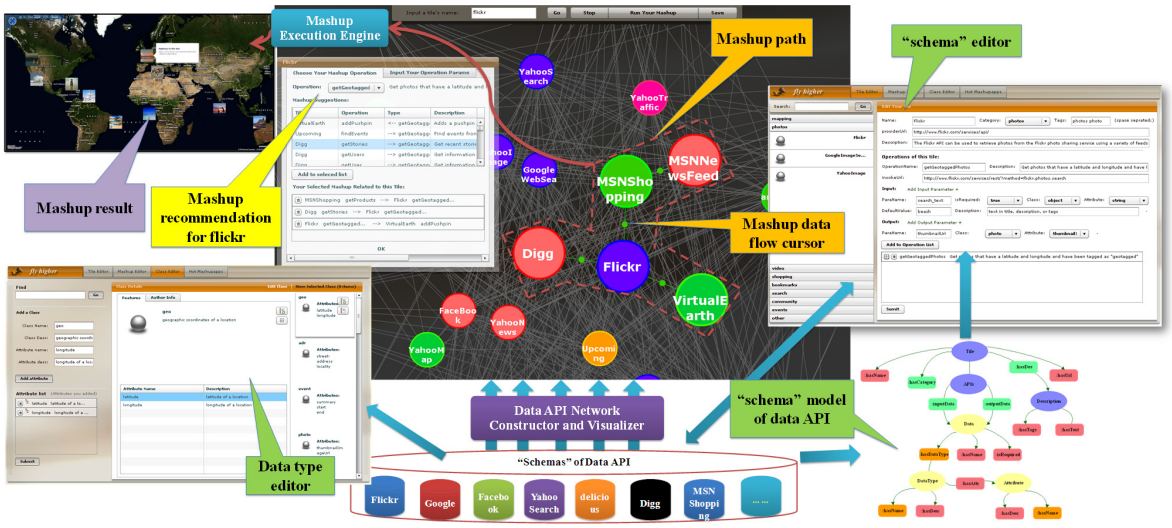

Figure 1. Overview of sMash Implementation

Step 1: Discover loosely links. Let $Node_1$, $Node_2$ be two nodes; API_1 and API_2 be one API of $Node_1$ and $Node_2$ respectively. $Node_1$ can loosely link to $Node_2$ if data types of output data of API_1 match data types of required input data of API_2. This means two nodes are mashupable on the basis of semantic data types and we find that approximately 51% of the links is meaningful.

Step 2: Remove the meaningless loosely links. We make use of tags and categories of APIs to refine the link's definition and extracted the meta-information, e.g., tags and data APIs each mashup used, of roughly 4100 existing mashups from ProgrammableWeb and Yahoo Pipes[2]. By virtue of collective intelligence, we clustered the tags that may be mashupable and built a mashupable tag model. If two nodes can be loosely linked, we check whether their tags are mashupable, if true, a link is added. Whenever user contributes a data API, sMash will search the dataset and record all the linkable nodes using above two steps.

Step 3: Visualize the network in the form of graph. sMash does a breadth-first-search starting from user specified nodes and ending in three depth-level to collect the linkable nodes. Considering each mashup as a path, we found that about 80% of 4100 existing mashup depth was no more than 3, so we decided to make the depth level of the breadth-first-search be 3. In this graph, vetexes and edges represent nodes and links respectively. The main part of figure 1 shows a visualized network around flickr.

Mashup candidate recommendation: sMash shows users a sorted list of mashupable APIs according to their selection in the form of {Data API, type [mashup from or mashup to], mashupable input and output data}, as illustrated in recommendation part of figure 1. To recommend the most related APIs first, we make a mashupable evaluation model based on loosely links; mashupable tags model and user behaviors.

User scenario: Suppose a user has no clear ideas about APIs, he knows a mashup source called flickr. So he starts from flickr, and then sMash will show him a global view of all the mashupable paths around flickr, navigate him which way to go next and remind him what his mashup currently looks like by means of data flow preview. All he needs to do is to select his interesting nodes to form a mashup path and set some key parameters, e.g., his flickr id and search term of the photos.

A brief analysis of data API network: In this network composed of more than 300 APIs, the total number of links is about 6500, the average number for each API is 21 and more than 55% of APIs have over 30 links. With the growth of the network, the links will be more complex.

3. CONCLUSION AND ONGOING WORK

We bring the idea of data API network into mashups and implement a semantic-based mashup navigation system which aims at facilitating all the users to build interesting mashups easily. Currently, we are improving mashup recommendation to give users a more intelligent navigation. In the future, to exploit more interesting data and links, we intend to focus on semantic data search based on this real-life data API network. We believe this work may bring semantic data search into a wider field.

4. ACKNOWLEDGMENTS

This work is funded in part by Zhejiang University and IBM CRL joint project called social enabled semantic data mashup, NO. 2006AA01A123, NO.51306030101, No.NSFC60525202 and No.9140A06060307JW0403.

5. REFERENCES

[1] D. Butler. Mashups mix data into global service. Nature, 439(5):6–7, 2006.

[2] A. Jhingran. Enterprise information mashups: Integrating information simply. In VLDB, pages 3–4, 2006.

[3] Microsoft Popfly: http://www.popfly.com/

[4] QEDWiki: http://www.alphaworks.ibm.com/tech/qedwiki

[5] Yahoo! Pipes: http://pipes.yahoo.com

[2] We developed an extraction engine to extract 3358 existing mashups from ProgrammableWeb, and 800 from Yahoo! Pipes.