mashpoint: Browsing the Web Along Structured Lines

Igor O. Popov, m.c. schraefel, Wendy Hall, Nigel Shadbolt
School of Electronics and Computer Science,
University of Southampton
{ip2g09, mc, wh, nrs}@ecs.soton.ac.uk

ABSTRACT

Large numbers of Web sites support rich data-centric features to explore and interact with data. In this paper we present mashpoint, a framework that allows distributed data-powered Web applications to linked based on similarities of the entities in their data. By linking applications in this way we allow browsing with selections of data from one application to another application. This sort of browsing allows complex queries and exploration of data to be done by average Web users using multiple applications. We additionally use this concept to surface structured information to users in Web pages. In this paper we present this concept and our initial prototype.

Author Keywords
WWW; Data Mashups; User Interfaces.

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: Miscellaneous.

General Terms
Design, Human Factors.

INTRODUCTION

A large segment of applications on the Web are data-oriented, i.e. they help users browse, explore or interact with structured information. Examples range from online shopping sites (e.g. Amazon), media content sites (e.g. Youtube) to social networking sites (e.g. Facebook). The majority of these applications typically deploy rich features that allow users to browse the underlying data. However, finding and integrating information about the same entities on related websites still requires manual effort. For example, if we visit an online travel website and narrow down the results to a few hotels of interest, finding information about exactly these hotels on other web sites still requires substantial effort. A typical scenario would involve finding sites that may hold such data (through using search engines), go through the process of finding the related information for each hotel separately (e.g. through using a search tool provided on the Web site) and checking if returned results corresponds to the entities found in the first application. The purpose of our research is to introduce the notion of entity based browsing to the Web that would allow browsing from one website to another with a set of entities (e.g. People, Places, etc.). In this way we are providing an extension to the traditional way of linking web pages, one that reflects the data-oriented nature of a large segment of applications the Web. We embodied these ideas into mashpoint[1], a framework that allows data-oriented Web sites to be linked based on the similarities in the their respective datasets.

RELATED WORK

Existing work in this area is most closely related to end-user mashup tools. For example, current mashup tools, such as, Intel's Mash Maker [3] allows users to scrape data for Web pages and do various visualisations. In the area of Semantic Web, several data browsers such as provide ways of browsing through raw data sources publicly available on the Web. For example the Tabulator [1] allows data in RDF format to be browsed, aggregated and displayed in various visualisation widgets. Tools such as Parallax [4] allows data from the Freebase[2] databased to be explored and visualised in different ways. Unlike mashup tools that are focused on mashing data from different sources, our framework proposes way of querying for information by combining and browsing through distributed applications that have data about same entities.

MASHPOINT: LINKING DATA-CENTRIC APPLICATIONS ON THE WEB

In the mashpoint framework, a data-oriented application is loosely defined as an application that allows data to browsed, explored and visualised. mashpoint is built on the idea that each data-oriented application at any given time displays information about one or more sets of entities. For example, an application can display data about entities such as countries, cities, people, events etc. At any given time an application shows a subset of these entities. Using an embedded widget, an application can surface up a window that suggests other applications that can show information exactly about those entities or a subset of those entities. For example, one application might show economic data about countries such as GDP, GDP per capita, growth levels etc and allow exploring this data by filtering on those facets. Another application, on the other hand can show demographic data about countries such as birth rate, death rate, age distribution. So if a user wants to query for the age distribution in countries with a low

[1] A demo is publicly available online at `http://mashpoint.net`

[2] `http://www.freebase.com`

UIST'12, October 7–10, 2012, Cambridge, Massachusetts, USA.
ACM 978-1-4503-1582-1/12/10.

GDP per capita a user can select the low GDP per capita in the first application and the *pivot* with those selection of countries to the other application. The user can then chose to continue to pivot to another application and get further data about those countries or a subset of those countries. By browsing between applications in this way, a user can do complicated and expressive queries that cannot be accomplished with any single application.

Implementation

In order to be able to perform such data-oriented operations between applications we devised our framework to require as little as possible effort to integrate an existing application. To integrate an application in the framework requires two things. First, to communicate, applications need to use canonical identifiers for the entities in their datasets. In order to identify entities we require that all applications reconcile their entities agains a common identifier (in our case we are using Semantic Web[2] standards and thus we use URIs as identifiers). In our particular case, we require that all application reconcile their data to Freebase URI identifiers. Thus if an application talks about an the country entity "UK", this is identified in all applications as `http://freebase.com/en/united_kingdom` or the relative URI `/en/united_kingdom`. In order to reconcile their data easily, application provides can use tools that support automatic or semi-automatic reconciliation against a particular set of named entities. For Freebase in particular such support is provided through an API of through tools such as Google Refine[3].

Second, the framework requires that an application must be able to express and establish its state based on the entities on which the data is about. For example an application that can display data about countries must be able to show any selected subset of countries on demand. To accomplish this we mandate that each applications URL contains a *"mashpoint"* parameter in encodes the state of the application in terms of sets of entities. For example an application displaying data about countries and their capitals which currently focuses on data about three countries (UK, US and France) and their would express that state in the URL in the following

Once these criteria are satisfied pivoting between applications is possible. However to make the process of finding applications to which to pivot to easy for users, we've augmented the framework to include a discovery service that could be launched from any mashpoint enabled application. Each application adds a mashpoint button, that when pressed displays a list of applications that can take the current selection of entities as input. In order to enable the discovery service each application registers the sets of entities on the discovery service and embeds a javascript widget in the form of a button that communicates with the discovery service, retrieves the available applications and displays these choices to users

Surfacing structured information directly form a Web page

Some Web pages today embed structured information and metadata about the page. Examples include Facebook Open Graph Protocol[4] and Schema.org[5] For the current subset of Web pages that still do not embed any structured data, however, we've devised a way of surfacing structured content using the mashpoint framework. A bookmarklet can be installed by users on their web browser that when pressed takes the content of the Web page and does named entity recognition on the text of the Web page. Once entities are found and identified to Freebase entities these can be passed. Thus, for example if user browsers through a news article that talks about certain countries, these will be identified to Freebase entities and the same widget will be surfaced to the users allowing them to view the identified countries in any number of applications. In order to to the named entity recognition we currently rely on AlchemyAPI[6] which does match concepts found in text to Freebase identifiers.

CONCLUSION

In this paper we presented an alternative way of mixing and exploring data by using multiple distributed publishing applications. We believe that the proposed framework has several strengths. First, rather than the current breed of mashup tools, which has a constraint on the ways data can be represented, as a mashup tool mashpoint has a potentially infinite ways of viewing data. Second, it follows a very simple, Web like model of browsing while providing powerful data-oriented queries. Third its distributed nature, allows potentially an organic growth beyond the scope of any one single application.

ACKNOWLEDGMENTS

This work was supported by the EnAKTing project, funded by EPSRC project number EI/G008493/1.

REFERENCES

1. Berners-lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A., and Sheets, D. Tabulator: Exploring and analyzing linked data on the semantic web. In *In Procedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06* (2006), 06.

2. Bizer, C., Heath, T., and Berners-Lee, T. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS) 5*, 3 (MarMar 2009), 1–22.

3. Ennals, R., Brewer, E., Garofalakis, M., Shadle, M., and Gandhi, P. Intel mash maker: join the web. *SIGMOD Rec. 36* (December 2007), 27–33.

4. Huynh, D., and Karger, D. Parallax and Companion: Set-based Browsing for the Data Web. 2009.

[3] `http://code.google.com/p/google-refine/`

[4] `http://ogp.org`

[5] `http://schema.org`

[6] `http://www.alchemyapi.com/`