

Participatory Mashups: Using Users to make Data Mashable

Rob Ennals
Intel Research
2150 Shattuck Avenue
Penthouse Suite
Berkeley, CA 94704, USA
robert.ennals@intel.com

Beth Trushkowsky
Computer Science Division
University of California at Berkeley
trush@berkeley.edu

ABSTRACT

Most mashups concern themselves with data that is fairly easy for a machine to understand: addresses, phone numbers, information available through APIs, or information that can be scraped from a small set of popular web sites.

However much of the most interesting information on the web is not in a form that a computer can easily understand. It may be in natural language text, or it may be in the long tail of web sites that developers have not found it worthwhile to write scrapers for.

In this paper, we promote the approach of “participatory mashups” in which the users of the mashup also teach the mashup ways that it can understand new data. As examples, we present our work on Think Link, which connects factual statements on the web to related statements elsewhere, and Mash Maker, which attempts to understand the meaning of structured data on arbitrary web pages.

INTRODUCTION

If we understand what the information on a page means then we can do useful things with it. We can connect the information to information elsewhere that might be relevant, we can compute new information from the information we have, and we can present users with new interfaces that are more useful than that provided by any individual site.

Some kinds of information are easier to understand than others, and mashups have so far understandably focused on the low-hanging fruit. Addresses follow an easily recognized pattern, and have led to a large number of useful mashups which plot various things on maps or work out what things are near to other things. Similarly, many mashups have made use of data that is available from APIs, or that can be easily scraped from popular web pages that the mashup has been hard-coded to understand.

But what about everything else?

Much of the most interesting information on the web is available as unstructured natural language text: arguments, ideas, opinions, factual claims, jokes. Much of this is information that could benefit a lot from being mashed up. For example “which of these claims is likely to be wrong?”, “what other sites disagree with this opinion?”, “is there a more reliable source that backs up this claim?”, or “what ideas on this site would I be likely to find most interesting?”.

Similarly, there is a lot of structured information stored in the “long tail” of web sites that are too small for it to be worth a mashup author writing a custom scraper or API interface for, but which may nonetheless be very interesting. For example, tables of data that have been entered by hand, or one of the many tiny web services that display data from a database backend using their own custom template that nobody has ever written a scraper for.

In this approach we argue for “participatory mashups” as a way to open up access to this data. In a participatory mashup we blur the line between the users and creators of a mashup by allowing the users of a mashup to teach the mashup how to understand new information, and connecting information in new ways. By teaching a participatory mashup the meaning of some data, the user helps themselves, since they can now include this data in their mashup, but they also help the wider community of users, since all other users can now benefit from the system understanding this data.

We motivate this approach using two mashup tools that we have built. Think Link allows users to identify factual claims that they read on web pages and connect them to related factual claims on other sites, including claims that may oppose the original claim. Mash Maker allows users to teach it the meaning of structured data on arbitrary web sites, and then enhances these web sites to make them more useful, guided by example improvements provided by users.

More generally, we consider a tool to be a “participatory mashup” if it relies on users to teach it what information on the web means, and how that information is connected together. We argue that this is an approach that other applications could also use to help them understand the web.

contentious claim: Global warming is causing more hurricanes (click snippet for more info)

All those the question: are storms getting stronger, and if so, what's causing it? According to a new paper in Nature, the answer is yes — and global warming seems to be the culprit. Researchers led by James Elsner, a meteorologist at Florida State University, analyzed

Figure 1. Hovering over a highlighted snippet shows a summary

The screenshot shows a web browser window with a highlighted snippet. A tooltip titled "Investigate Claim" is open, displaying the following information:

- about:** Hurricanes, Global Warming, Effects of Global Warming
- supports:** Global warming is a serious problem; Increasing global temperature is expected to cause changes in weather, agriculture, glaciers, species extinctions, and disease
- opposes:** Global warming has not increased the number of hurricanes
- Global warming is causing more hurricanes** (highlighted)
- supported by:** (empty)
- opposed by:** (empty)
- snippets:**
 - All these hurricanes in such a short period of time begs the question: are storms getting stronger, and if so, what's causing it? According to a new paper in Nature, the answer is yes — and global w...
 - Is Global Warming Worsening Hurricanes? - www.time.com
 - In the North Atlantic, for which we have the best records, there has been a clear increase in the number and intensity of tropical storms and major hurricanes. From 1850-1990, the overall average numb...
 - Hurricanes and Global Warming FAQs: The ... - www.pewclimate.org

Figure 2. Click on a claim to investigate evidence for and against it

THINK LINK

Much concern has been expressed recently about biased or inaccurate information on web sites [5, 11], and the media echo chamber problem in which people read web sites that feed their own views back to them, rather than exposing them to the opposing opinions held by other groups [6, 13].

Think Link [14] is a Firefox extension that helps users evaluate the truth of statements they read, helps users be more easily exposed to a range of ideas and opinions that they might not otherwise encounter.

As a user browses the web, Think Link highlights snippets of text that other users have identified as making claims that

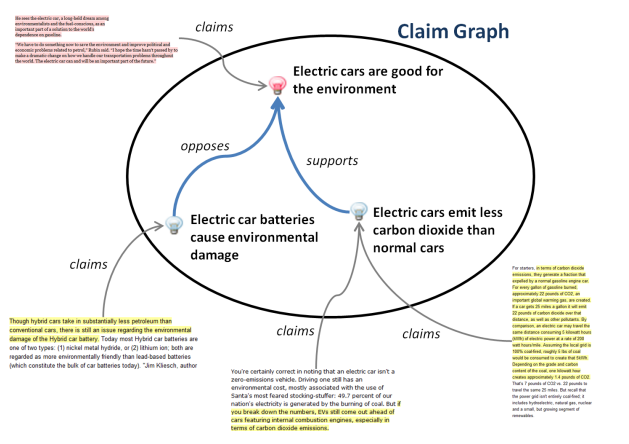


Figure 3. Think link connects claims to each other and to web snippets

The screenshot shows the "Investigate Claim" interface for the claim "Global warming is a serious problem". It displays a network of related claims and snippets:

- Topics it is about:** Global Warming, Effects of Global Warming
- Claims it opposes:** The negative effects of global warming are exaggerated
- Selected Claim:** Global warming is a serious problem
- Claims that support it:**
 - Canada's ice shelves are disappearing due to global warming
 - Increasing global temperature is expected to cause changes in weather, agriculture, glaciers, species extinctions, and disease
 - Global warming is causing more hurricanes
 - Many Americans are concerned about climate change
- Claims that oppose it:**
 - US Government measurements have shown that global temperature has been decreasing slightly
 - Global weather changes can be attributed to El Nino
 - Frozen evidence shows that the arctic ocean was much warmer thousands of years ago, and the arctic cap disappeared in most summers
 - The negative effects of global warming are exaggerated
 - US annual rainfall has increased three inches in the last 100 years, benefitting agriculture
 - Not affecting ecosystem
- Web snippets that state it:**
 - "Global warming is the biggest and most serious problem faced by us in this century. Climate change is happening and its effects are real. If we do not take seriously, it will have serious consequences ..."
 - "Global Warming The Biggest Problem" Br... - www.countercurrents.org
 - "I have said consistently that global warming is a serious problem. There's a debate over whether it's manmade or naturally caused," Bush told reporters. ...
 - Bush: Climate change is 'serious problem' - www.breitbart.com

Figure 4. The claim browser visualizes the claim graph

are interesting or controversial (Figure 1). Interesting claims are highlighted in yellow, and controversial claims are highlighted in red. If a user clicks on a highlighted snippet, Think Link will display an interface that allows a user to easily find snippets on other web pages that make related claims, including snippets that argue against the opinion put forward by the highlighted snippet (Figures 2 and 4). Think Link allows a user to easily access the best arguments for and against a claim, including arguments they might not otherwise have come across because they would not have been voiced by the web sites that the reader normally reads.

Think Link maintains a graph of known factual claims, their relationship to each other, and the snippets that hold these claims (Figure 3). This graph is built entirely by users of the tool. Users create claims, users connect claims, and users identify snippets making claims. Similarly, users vote for snippets, claims, or connections that they believe are particularly important.

This graph allows one to ask interesting questions such as "Which snippets on this page make claims that are opposed by snippets elsewhere on them web?", "What is the most trustworthy news source that makes this claim?", "What do my friends think about this idea?", "What are the most interesting claims about this topic that I haven't read before", or "What is the best argument against this claim, and what would be the best counter-argument to that?".

Identifying factual claims and the connections between them automatically would be very difficult using the current generation of natural language processing and automated reasoning tools; however this is quite an easy task for humans. To create a new snippet, a user selects the text that is making an interesting factual claim, and then uses a claim browser interface to identify the claim being made (Figure 5). Users can connect claims together using a simple drag and drop interface.

Our user study participants identified several reasons why

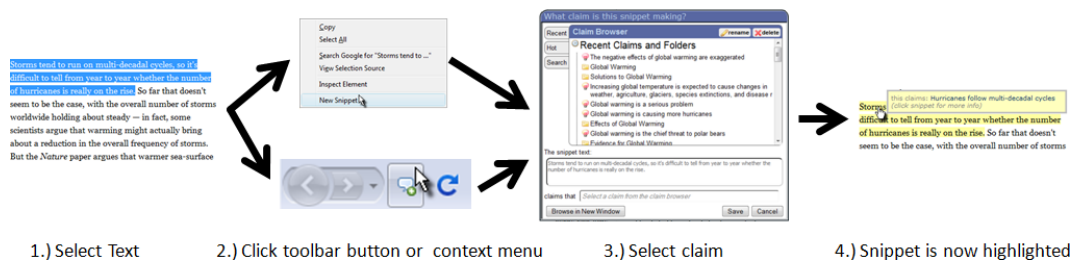


Figure 5. Process for creating a new snippet

they would want to mark up factual claims in documents they found. The most common reason people gave for identifying and organizing factual claims was if the person was writing an article or researching a topic and wanted to keep track of the information they had found. One user (who is an active blogger) also expressed an interest in finding and marking up instances of claims that he disagreed with, so that readers would see the arguments highlighted in red and be directed to the counter-arguments. The same user also expressed an interest in marking up claims in documents he had made in his own articles so that readers could quickly see the evidence he had found in support of his claims.

Like the work of Luis von Ahn [15, 10] or the “invisible hand” or Adam Smith the intention is that people should be motivated by self interest to do something that benefits the wider community.

MASH MAKER

Mash Maker [3] is a browser extension that understands the meaning of the pages that users browse and suggests ways that it can improve the current page so as to be more useful to the user. Mash Maker relies on users to teach it both how to understand web pages, and also how particular kinds of web page might be improved in interesting ways.

As views a web page, Mash Maker will suggest ways that it can make the page more useful, and suggest these improvements on its tool bar. If the user clicks on the button for such an improvement then Mash Maker will apply it to the current page, potentially using other web sites and remote APIs, and potentially applying widgets that produce new visualizations or compute new data (Figure 6). Mash Maker suggests improvements based on the meaning of the current page, the meaning of pages that the user has recently browsed, and the behavior of other users.

Our intention is that users should be able to use Mash Maker to enhance arbitrary web sites in arbitrary ways. Unlike “mashup creation tools” like Yahoo Pipes [8] or Microsoft Popfly [9], we see Mash Maker not a tool for creating special-case mashup sites, but as a generalized tool that understands what you are doing and infers new mashups that can enhance your current experience, based on what others have tried in the past.

Users can contribute to Mash Maker in several different ways. They can teach Mash Maker how to understand new

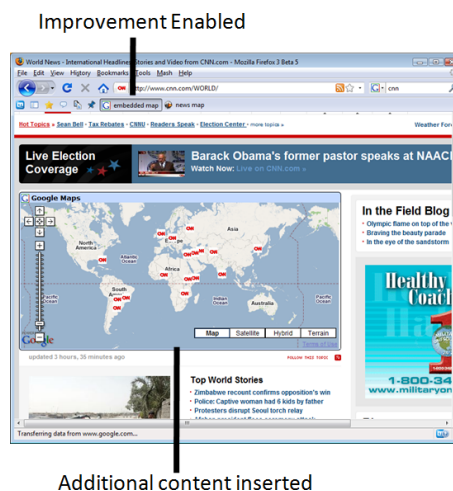


Figure 6. Using Mash Maker to add a map to a page

kinds of web site, they can write new widgets, and they can show Mash Maker new ways that it can improve web pages that contain particular kinds of data. A user teaches Mash Maker how to extract information from a particular kind of web page by opening the Mash Maker extractor editor (Figure 7) and picking out examples of things on the page that the user thinks is interesting. As more users give Mash Maker more examples, Mash Maker gets better at understanding what pages mean. Unlike tools like Dapper [2] extractors are not owned by particular users. Instead an extractor is built up from examples given by all users on the web.

Given the huge number of differently formatted web sites on the web, it would not be practical for Mash Maker to be able to understand such a large range of web sites without users participating to teach in what different sites mean. For the users, the benefit of teaching Mash Maker how to understand a web site is that they get to use an enhanced version of the web site that Mash Maker has improved for them.

PATTERNS VS INSTANCES

Users teach Mash Maker “patterns” that Mash Maker can use to understand a large number of similarly formatted web pages (e.g. how to understand a flight listing on Expedia). This differs from Think Link, in which users pick out individual instances of factual claims which are not generalized to automatically find more snippets.

Extractor Editor

Use the extractor panel to teach Mash Maker how to understand this kind of web page.

Selection

The current selection in the extractor editor is highlighted on the page

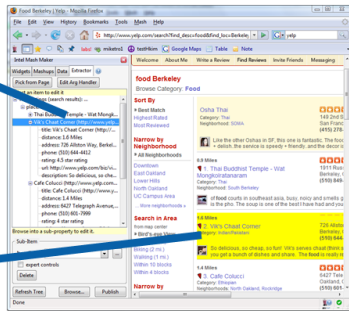


Figure 7. Any user can teach Mash Maker how to understand new pages

In the future, we plan to extend Think Link with natural language processing technology so that Think Link can automatically find new snippets, based on the examples it has been given. For example, if many users have shown Think Link example snippets that make the claim “global warming is man made”, then Think Link could use these examples to find other snippets that may also be making this claim. At the simplest level, this could be done using simple textual similarity, similar to that done by Google Book Quotations [7].

RELATED WORK

The notion of a “participatory mashup” is somewhat fuzzy, and indeed some previous work can be considered to fit this description to some extent. To be a “true” participatory mashup, we believe a tool should not just allow users to contribute content, but allow users to teach it how to understand the meaning of content that already exists on the web, and use this knowledge to help other users of the tool.

reCAPTCHA [10] gives users the task of correctly transcribing scanned books as a test to see if they are human. Users thus unintentionally help machines better understand this information. Luis von Ahn has also shown how users can be persuaded to teach a system what data means by making it fun [15]. Our approach to persuading users to teach us what data means is to make this data immediately useful to them.

Online mashup creation tools such as Yahoo Pipes [8] or Microsoft Popfly [9] allow users to contribute new mashups. These tools treat each mashup as a separate entity and do not enable users to contribute knowledge about the web that can be used by other mashups.

Dapper [2] allows users to create new web scrapers, effectively creating APIs from web sites. This is similar to the extractor editor in Think Link. The key difference is that each scraper users create is treated as a separate entity, owned by a particular user, rather than as a pool of knowledge about the web that can be shared and edited by all users.

Like Think Link, SpinSpotter [1] also uses a browser plugin to help users identify and share instances of media bias and misrepresentation. SpinSpotter allows users to annotate web pages with comments about the spin they think is present and suggest textual edits. This differs from Think Link, which

attempts to understand what claims are being made and relate them to claims made elsewhere on the web.

FreeBase [4] is an open database of knowledge that anyone can contribute to. While FreeBase does not understand the meaning of existing web sites, it does allow users to contribute knowledge that mashups can use.

ACKNOWLEDGMENTS

We would like to thank Allison Woodruff, Tye Rattenbury, Prashant Gandhi, Eric Brewer, and all our user study participants for all their help during the design of Think Link and Mash Maker. Think Link and Mash Maker both use icons from the free FamFamFam Silk [12] collection.

REFERENCES

1. B. X. Chen. Spinspotter combats unethical, biased journalism. *Wired Magazine*, Sept. 2008.
2. Dapper: The data mapper. <http://dapper.net>.
3. R. Ennals, E. Brewer, M. Garofalakis, M. Shadle, and P. Gandhi. Intel mash maker: Join the web. *ACM SIGMOD Record*, 36(4), Dec. 2007.
4. Freebase: an open, shared database of the world’s knowledge. <http://freebase.com>.
5. P. Ghosh. Warning sounded on web’s future. *BBC News*, Sept. 2008.
6. K. H. Jamieson and J. N. Cappella. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press, June 2008.
7. O. Kolak and B. N. Schilit. Generating links by mining quotations. In *Proceedings of the nineteenth ACM conference on Hypertext and Hypermedia (HYPERTEXT '08)*. ACM, 2008.
8. Yahoo Pipes. <http://pipes.yahoo.com>.
9. Microsoft popfly. <http://popfly.com>.
10. reCAPTCHA. recaptcha.net.
11. J. Seigenthaler. A false wikipedia biography. *USA Today*, Nov. 2005.
12. Silk icons. <http://famfamfam.com>.
13. SourceWatch. Sourcewatch entry on “echo chamber”. http://www.sourcewatch.org/index.php?title=Echo_chamber.
14. B. Trushkowsky and R. Ennals. Browsing the web of factual claims. Submitted to CHI 2009.
15. L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8), 2008.