Databases and Web 2.0 Panel at VLDB 2007

Sihem Amer-Yahia
(moderator)
Yahoo! Research, USA

Alon Halevy
(moderator)
Google Inc., USA

Gustavo Alonso and Donald Kossmann
ETH Zurich
Switzerland

Volker Markl
IBM Almaden

AnHai Doan
Univ. of Wisconsin, USA

Gerhard Weikum
Max Planck, Germany

1. INTRODUCTION

Web 2.0 refers to a set of technologies that enables individuals to create and share content on the Web. The types of content that are shared on Web 2.0 are quite varied and include photos and videos (e.g., Flickr, YouTube), encyclopedic knowledge (e.g., Wikipedia), the blogosphere, social book-marking and even structured data (e.g., Swivel, Many-eyes). One of the important distinguishing features of Web 2.0 is the creation of *communities* of users. Online communities such as LinkedIn, Friendster, Facebook, MySpace and Orkut attract millions of users who build networks of their contacts and utilize them for social and professional purposes. In a nutshell, Web 2.0 offers *an architecture of participation and democracy* that encourages users to add value to the application as they use it.

We held a panel at VLDB 2007 that examined the relationship between Web 2.0 and data management, and explored the opportunities this new medium presents to us. Some of the questions we considered were:

- What are the new research challenges that Web 2.0 presents to the data management community? For example, how should the fact that users are so interrelated in communities change our approach to querying data?
- What existing research problems are emphasized by the challenges faced by Web 2.0? For example, how can we deal with Web-scale data heterogeneity and issues of data quality when content is created by so many people?
- What principles developed by our community can be leveraged to enhance Web 2.0 tools? For example, can the principles of declarative specifications be put to use?
- Given the difficulties of performing academic research on anything related to Web search, what should be our research methodology in addressing Web 2.0?

In what follows, we summarize the position of each panelist and give an overview of the points raised by the lively discussion that followed.

2. ACCESSIBILITY

Gustavo Alonso: *Web 2.0 is about providing devices to access and produce data!*

The generation of new content in the Internet is driven by three different but complementary trends: (1) the Internet reaching a critical mass of users (e.g., blogs or information exchange sites); (2) the proliferation of services that allow users to combine different sources of information (e.g., maps and news feeds); (3) new devices and home appliances that have now become data sources (e.g., digital cameras, video cameras, sensors or automatic data feeds, which produce mostly multimedia data rich in contextual and meta-data information). In trying to predict what will happen next, however, the software is not the defining factor. The question of how this proliferation of data and users can be best supported through services, tools, programming languages, and search technologies, can be answered only by looking at how the people and the devices driving Web 2.0 are evolving.

As the amount of information available increases, there will be more users driven to the Web 2.0 and those already using it will intensify its relation to it as it becomes an indispensable reference in every day life (for information, for entertainment, for communication, etc.). Such a development will demand easier user interfaces, combined multimedia access channels (audio, video, text, images), and the ability to personalize contents.

As technology improves, devices will produce more complex data thereby making data access pervasive (cameras with GPS, cameras that allow the user to record a short audio clip describing the picture, sensors that produce data and its lineage, mobile phones with a flat Internet access rate). The open question is how to better utilize the increasing amount of data to better organize the information and support more precise search.

These are the trends to watch and the ones that will define the appropriate software technology for Web 2.0.

3. COLLABORATIVE EFFORTS

Alon Halevy: *Web 2.0 is about helping the masses manage heterogeneous datasets collaboratively!*

Web 2.0 is all about user-created content. While the prevailing types of content on Web 2.0 continue to be text,

photos and videos, there is a huge potential for creating and sharing more structured data. Structured data can be shared for business, educational and social purposes. Imagine what will hapen to political debates when people can look at *real* data and discuss it!

A mission of the database community should be to build the tools that enable people to create, share and analyze such data. This effort will require collaborations with human-computer interaction, and information visualization researchers at the very least and possibly other communities. Opportunties for ground-breaking research are huge. Below I mention three important directions.

The first research challenge is to design systems where heterogeneity is the rule, not the exception. There are millions of heterogeneous sources of structured data on the Web. In addition, text-mining algorithms are producing structured and very heterogeneous collections from text documents. Research on heterogeneous databases has been a healthy sub-field of ours for almost three decades, but the focus has always been on reconciling heterogeneity and the scale has been limited. The challenge here is to deal with millions of data sources in arbitrary domains with no hope of enforcing common schemas or terminologies. Furthermore, since the Web is not static, evolution is a key component. To make things more concrete, we need to build systems that can perform gracefully with high degrees of heterogeneity and create mechanisms for that incent users to reconcile heterogeneity when they see fit and without central authority.

Second, we need to build data management and integration tools that can be used by the masses. Discussions on un-usability of database systems are a part of our community's favorite pastimes. It is time to make a leap and build usable systems. This will mean a lot of *compromises* in functionality from a traditional database point of view. The challenge is to capture the most important use cases and design interfaces that will make those as easy as working with a spreadsheet. There is a budding industry set of tools in this area already (e.g., Many-eyes, Swivel, and several tools for easily creating mashups).

Finally, a more open-ended (and somewhat more vague) challenge is to imagine the kinds of databases that can be created when millions of people spread across the planet are collaborating. For example, suppose you want to build a database that stores where in the world people have access to clean water, or where certain diseases are currently prevalent. Such data is incredibly hard to collect right now and varies considerably when you drive for one hour from location to another. But with Web 2.0, you can imagine people in villages entering data through mobile devices and obtaining a live picture of access to clean water or prevalence of disease. Of course, this data will often be dirty (no pun intended), inaccurate and possibly maliciously doctored. Hence, we need methods that enable us to explore such data and leverage techniques for modeling uncertain data, data lineage and inconsistency.

4. DATA QUALITY

Gerhard Weikum: *Web 2.0 is about content-production democracy and a data-quality crisis!*

The proliferation of user-provided content opens up unprecedented opportunities for harvesting the "wisdom of crowds" [10]. In principle, these mega-trends are turning Web 2.0 into the world's most comprehensive knowledge base, with a wealth of intellectual wisdom and freedom of opinions. Let Web democracy find out about the best MP3 players and drugs against HIV!

Social wisdom of this kind is, of course, not new at all. Wikipedia is a wonderful success story of user-provided content at large scale with relatively little explicit control [4]. And already 250 years ago, about 140 people collectively wrote l'Encyclopédie with 70,000 articles in 28 volumes [1]. But these kinds of high-quality endeavors do not scale up with the increasing Web 2.0 population. All kinds of would-be experts offer their opinions in unmoderated or leisurely moderated forums, and blogs are a rich source of rambling babbles. There is certainly also a growing amount of valuable content, but it tends to be hidden in noise. Web 2.0 is about to create a major *data-quality crisis*. Understanding and analyzing trust, authority, authenticity, and other quality measures in social networks will pose major research challenges.

Tags assigned to photos, videos, and Web sites vary from highly informative to meaningless meta-tags (e.g., toRead) typos and misspellings (e.g., Brittnye, AngelinaJolly), trivialities (e.g., myVideo), and intentionally misleading annotations (e.g., bestPresident). With currently millions of low-profile users and potentially further growth to billions, we are witnessing rapid degradation of the noise/content ratio. This will make it increasingly difficult to find valuable information. Thus, notwithstanding the brave hopes for more database-style structure on the Web and the grand vision of a Semantic Web, the huge variance in information quality will make *search* on Web 2.0 a lot harder than on today's Web.

Web 2.0 - the people's Web - is not a Web of facts; it is a Web of opinions [7, 3]. Blogs, for example, seem to be a blatant invitation for spamming; and some impertinent users even post contributions or make up entire blogs under someone else's name (celebrities being the preferred object, but even database researchers from Wisconsin have been targeted). But also tagging, rating, and recommendations are game to opinionated and manipulative minds. Moreover, masses of short-minded or easily influentiable people may follow, creating a flood of "truthiness": statements that one believes to be true regardless of how much they disagree with bare facts. Web 2.0 is bound to violate one of the axioms of the "wisdom of crowds", namely, independence of opinions. It is not uncommon that users in social-tagging communities blindly copy someone else's tags, thus reinforcing initial falsehoods. A similar situation may arise with mashups, as they critically depend on the data quality of their underlying sources and on the correctness of the corresponding mappings and matchings (between schemas as well as entities and attribute values). Thus, mashups over mashups over mashups may serve as amplifiers for inaccuracy and distortion. For all these reasons, it is paramount to identify not only the best information authorities but also to analyze and track the authenticity and lineage of annotations and recommendations.

5. SOCIAL RELEVANCE

Sihem Amer-Yahia: *Web 2.0 is about leveraging social ties to find the right content to serve to the right user!*

The recent advent of "Web 2.0", that is, the evolution of

the Web from a technology platform to a social milieu, has been accompanied by an explosion in the number and reach of *social content sites* such as *collaborative tagging sites* and *collaborative reviewing sites*. The unprecedented popularity of these sites is the source of a wealth of user-generated content. Some statistics: 24M people added on FaceBook since 12/06; 60M users on Yahoo! Answers and 120M answers; 100M views/day in YouTube/65K new videos/day; 7M groups/190M users/12M emails daily; 2.7 tags/user/resource in del.icio.us. The ability to sift through large amounts of content is a challenging problem that has a big impact on the *survival* of these sites [8]. For example, in del.icio.us, a social book-marking and tagging site, users can subscribe to their friends' feeds in order to learn about their latest bookmarked URLs. They can also view hotlists [1] [3], as well as browse tags to find related content.

The quality of a hotlist can be measured by estimating its *scope* (set of people for whom it is intended) and its *coverage* (average overlap of the hotlist with the user's interests.) Consequently, the ability to model users and their interests is a key challenge [9]. While Databases and Information Retrieval rely on the assumption that content is static and user interests are dynamic and expressed using keyword search, Information Filtering techniques have been developed to address dynamic content and static user interests [5]. In social content sites, *both content and user interest* are dynamic: people review and tag new content every day. This presents a unique opportunity for re-thinking search, query processing and content recommendation in the context of collaborative sites.

Collaborative Filtering (CF) is a popular method that uses machine learning to determine interest overlap between users based on their behavior such as common ratings of items, or common purchasing and browsing patterns. In social tagging sites, a user's interest can be modeled in terms of the tags he uses to annotate content, and in terms of his explicitly stated and derived social ties. We advocate the need to build *common interest networks* that link two users if the sets of items they tagged overlap significantly. We argue for exploring different kinds of networks which model different users behaviors, and using them to generate higher quality hotlists.

One factor that limits the effectiveness of deriving interest overlap between users in CF is *sparsity*: there are often many more items in the system than any one user is able to rate. This issue is further aggravated in the context of a collaborative tagging site such as del.icio.us, where the set of items corresponds to a potentially infinite set of Internet sites. Another important reason is that people rarely agree on everything: you may agree with your mother on cooking, and with your adviser on research, but your adviser's opinion on food is hardly relevant. This argues for combining tags and item overlap to construct *per-tag common interest networks*. Such networks have wider applicability than *item-only interest networks*, and can be used to construct hotlists of higher quality.

In summary, databases need to be enhanced by adding the social dimension (tags, reviews, explicit and implicit social ties) and incorporate recommendation mechanisms.

[1] a list of most popular items among a set of users in a given period of time.

6. DECLARATIVE MASHUPS

Volker Markl and Donald Kossmann: *Web 2.0 should leverage database expertise to define mashups declaratively!*

Web 2.0 is all about people providing content. The logical next step is that users will try to combine the content in interesting ways in order to provide new content and more importantly, provide new *services*. Consequently, users will try to combine services to provide more specialized services. This process is typically described by another buzz word: *mashups*. A Web of mashup services is the logical next step after the Web of documents.

In order to facilitate the Web of mashup services, it must be just as easy to create a mashup as it is to put a photo on Flickr or ask a question on Yahoo! Answers today. Just as the digital camera has created several billion "Steven Spielbergs", the Web of mashups will create several billion hackers. Not only must it be easy to create mashups, it must also be cheap to run and operate them.

There is a need for a declarative language to build scalable and reusable mashups. Unfortunately, it is still difficult to write code. One big problem of today's situational applications is that they are not created in a declarative fashion. Instead, programming languages like JavaScript, Java, PHP, or Ruby are used to program mashups. These models are clearly not appropriate for Joe Doe's grandma. The situation becomes even more confusing as some of these languages are intended for client-side mashups (e.g., JavaScript only runs in the browser) whereas others are intended to run on servers. (Grandma does not care about clients and servers.) Furthermore, these models prevent mashups from being properly indexed and found in search engines. In addition, it limits the re-use and combination of existing mashups in new applications.

The database community has been strong in making declarative programming a mass market. Clearly, SQL is not going to be the winner on the Web, but the SQL success has shown: (a) logical and physical data independence so that applications can evolve over time and survive technological shifts; (b) increased productivity using a declarative programming language; and (c) reduced cost of operation and increased scalability because of automatic optimization. Yahoo! Pipes [2] or IBM DAMIA [3] are examples which attempt to enable such mashup specifications. However, they fall short of several aspects. A comprehensive infrastructure for the specification of mashups must facilitate data management and presentation logic in addition to data and control flow specification. Any patchwork of different technology will make it difficult to index mashups and to migrate mashups in response to new hardware and architectural developments; e.g., moving more computing to mobile clients.

Well, if Joe Doe's grandma can build situational applications, so can Joe Doe's boss. There will be a new separation of work between large software vendors (i.e., vendors of so-called "standard" software such as IBM, Microsoft, Oracle, and SAP), independent software vendors (ISVs), and customers. Technologies to facilitate going to go from the ISVs to the customers have been called *software mass customization* [6, 2], adopting a term from manufactural engineering [4].

[2] *http://pipes.yahoo.com*
[3] http://services.alphaworks.ibm.com/damia/
[4] *http://en.wikipedia.org/wiki/Mass_customization*

7. METHODOLOGIES

AnHai Doan: *Web 2.0 opens up many compelling opportunities for database research. But how should we proceed?*

I completely second the Web 2.0 challenges raised by my fellow panelists. Creating more structures, adding social dimensions, finding high quality data, developing declarative mashups – these constibute many compelling opportunities for database research on Web 2.0.

But how should we proceed? Doing research on the Web scale requires getting access to real data of social content sites which can be cumbersome. How do we find "fundamental" Web 2.0 problems to work on? And if we find a solution, how do we know that it has not been employed at a Web company, and how do we evaluate the solution anyway? To successfully maximize our impact on Web 2.0, we need multiple "attack plans" with a low "barrier of entry".

As a possible "attack plan", I propose to explore managing *unstructured data* at the *community* scale. To manage such data (e.g., Web pages, newsgroup postings, memos, articles), extraction to generate more structure is fundamental, because otherwise the data cannot be fully utilized and there is little for us to "play with". Integrating the extracted structures will then become important. Further, since extraction and integration often are imperfect, we should engage users to assist with the process, in a mass collaboration fashion. In general, we should make it very easy for users to help extract, integrate, contribute, combine, query, visualize data and services, and to network with one another within the community.

By working at the community scale – that is, mini-Web, rather than the entire Web scale, this plan should incur a relatively low "barrier of entry", especially for academic research groups. At this scale, we should be able to build community-centric data management systems, then apply them to real-world applications to drive and evaluate the research (just like what we did in the relational world).

We should also be well-positioned to make significant impact on Web 2.0, in two ways. First, the Web is fundamentally the largest database of unstructured data, managed by the largest user community on Earth. Hence, many lessons we learn in managing unstructured data at the community level should also be applicable to Web 2.0.

Second, Web 2.0 includes not just the "Infotainment" Web of *Flickr* and *Youtube*. It also includes the myriad communities of users (that we have rarely heard of) in "Science 2.0", "Government 2.0", "Spy 2.0", etc., who are collectively acquiring and managing their community data. Examples include *ecolicommunity.org*, which is trying to build the largest E. Coli database in the universe, *umasswiki.com*, which collects all information about the University of Massachusetts, Amherst and the surrounding area, and *Intellipedia*, the largest wiki-based spy database. Our community-centric tools can immediately be applicable to these cases.

8. THE AUDIENCE VERDICT

The presentations by the panelists was followed by a lively audience discussion. The issues discussed centered around several main areas in which data management technology is relevant to Web 2.0: building scalable back-ends for Web 2.0 services, building platforms on which others can build services, constructing new Web 2.0 data-oriented services, and studying user behavior to improve services.

In addition, the audience reacted as follows:: (1) *How are we going to evaluate our solutions for these Web 2.0 problems, especially if they involve many users?* and (2) *There is some concern that we are no longer leading data management trends.* The main implication is: should our community change our paper evaluation practices, if we want to promote work where users are the main drivers?

Users in Web 2.0 tend to adopt new technology quickly and easily and before it is even understood. In that regard, usage precedes deep thinking as we, researchers, are used to. We thus find ourselves in an after-the-fact situation which is quite typical of Web technologies and the natural sciences, where we strive to understand the natural world. Web 2.0 encompasses a wide array of ideas and approaches, not all of them directly related to technology and some with deep social implications. For a computer scientist in general and a database researcher in particular, it is difficult to see where a contribution can be made as many of the discussions around Web 2.0 are not technology-oriented (e.g., the political relevance of blogs). This can be viewed as a unique opportunity for computer scientists to see the wider impact of their work and look at users for inspiration on where the next challenges lie. Naturally, this may lead to some change in how we evaluate our work and the work of our peers.

9. REFERENCES

[1] J. B. l. R. d. e. a. Denis Diderot. Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers. pages 1751–1772.

[2] G. A. Donald Kossmann. Software Mass Customization (in German). In *Datenbank Spektrum*, 2006.

[3] N. K. (Editor). Special issue on data management issues in social sciences. In *IEEE Data Engineering Bulletin*, volume 30, 2007.

[4] J. Giles. Internet encyclopaedias go head to head. In *Nature 438*, 2005.

[5] J. A. Konstan. Introduction to recommender systems. In *SIGIR07: Proceedings of the 30th Annual International ACM SIGIR Conference*, 2007.

[6] C. Krueger. Software mass customization. In *White Paper, BigLever Software Inc.*, 2005.

[7] B. A. H. Scott A. Golder. Usage Patterns of Collaborative Tagging Systems. In *Journal of Information Science*, volume 32, 2006.

[8] P. B. Sihem Amer Yahia, Michael Benedikt. Challenges in searching online communities. In *Special Issue on Data Management Issues in Social Sciences, IEEE Data Engineering Bulletin*, volume 30, 2007.

[9] J. Stoyanovich, S. A. Yahia, C. Marlow, and C. Yu. Leveraging Tagging to Model User Interests in del.icio.us. In *AAAI Social Information Processing Workshop*, 2008. To appear.

[10] J. Surowiecki. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies. In *SN*, 2004.