COVER FEATURE

Toward a PeopleWeb

Raghu Ramakrishnan and Andrew Tomkins
Yahoo! Research

Important properties of users and objects will move from being tied to individual Web sites to being globally available. The conjunction of a global object model with portable user context will lead to a richer content structure and introduce significant shifts in online communities and information discovery.

The Web is evolving into a dynamic repository of information on virtually every topic, including people and their connections to one another as well as to content. Two emerging capabilities will significantly impact online activity. The first involves data and will let users create, reference, annotate, and interact with important objects in a site-independent manner to produce semantically rich content. The second new capability involves people and will let users create portable social environments that follow them as they interact online.

CONTENT AND PEOPLE

On the *content* side of the equation, users are increasingly consuming structured data as more of daily life migrates online. Important types of structured data include information about restaurants, products, songs, videos, finance, user profiles, social networks, and so on. As of late 2006, for example, Google Base and Yahoo!'s vertical properties each contained about 150 million structured commercial listings—for example, homes, jobs, products, and vehicles. Noncommercial listings such as recipes and reviews exist online at a similar scale; eBay reports that it hosted almost 2.4 billion listings during 2006.

Companies are creating search products that rely on extracting structured metadata, such as category tags (Kosmix), product types (Google Base and Yahoo! Shortcuts), and personal attributes (ZoomInfo). Data feeds of real-world and online events are becoming ubiquitous on social networking sites such as Facebook, Upcoming, and Yahoo! Answers and typically include automatically generated structured metadata for user targeting and subscriptions.

On the *people* side, a broad base of users rather than a small number of professional publishers is now producing content at a greater rate than all other forms of textual content both online and offline. User-generated metadata, in which community members employ tools to place cues such as ratings, tags, or reviews on content, is likewise being generated faster than professionally produced anchor text, the traditional workhorse that search engines employ to judge document quality.

Further, *attentional* metadata, which details the pieces of content users are consuming, significantly outweighs all other metadata used for information discovery. Attentional metadata is increasingly sought after and is beginning to accumulate in significant volume, suggesting a paradigm shift—and simultaneously raising serious questions about user privacy.

Finally, social networks are increasingly prevalent as channels of content consumption: Approximately two orders of magnitude more digital information flows daily within these networks than in the public eye.[1]

Users are creating and consuming content at a rapid pace, often within a particular social structure, and this content is increasingly more structured than simple bags of words. At the same time, users must negotiate significant gaps in the Web infrastructure.

0018-9162/07/$25.00 © 2007 IEEE

Published by the IEEE Computer Society

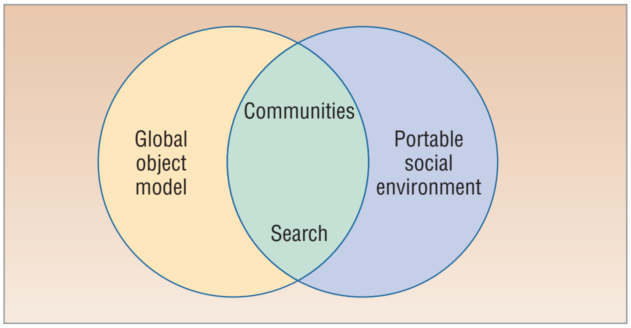

Figure 1. PeopleWeb components. The global object model and portable social environment represent two key emerging capabilities. At the intersection of these capabilities lie two existing aspects of online behavior that will shift significantly in response: communities and search.

Content gaps

All significant structured content repositories are siloed—Amazon for product information, MySpace for profile information, Yelp for local listings, and so on. Repositories for the same type of object do not interlock, and repositories for different types of objects do not communicate. A user wishing to reference a particular digital camera does not have access to an identifier to the camera; she must reference a manufacturer's URL, or more likely a URL on a well-known distributor's site, or a review site, or she must simply describe the camera textually. Another user wishing to add metadata to that camera, such as a review or a rating, must do so in isolation on one site, knowing that most potential consumers of the information will probably never find it. And even if the identifiers are mapped, the attributes might be named or even defined differently.

People gaps

Users must create entirely new personas at each site they visit and re-create from scratch their social networks. Even when they accomplish this, capabilities from different sites do not interact. There is no clean way for a user to share with a set of friends his global activities across the Web—thus, a user cannot put together a travel journal using photos from one site within journal software from another, even if his friends have access to both sites. Automated techniques to identify when one user's content might be of interest to another are also siloed and, consequently, impoverished.

TOWARD A PEOPLEWEB

As people and objects acquire metadata while moving across Web sites, a new kind of interwoven community fabric will emerge. Data objects will become richer, with interactions occurring in the context of the people involved. Reputation-weighted authorship and both explicit and implicit user-generated metadata will inform object quality measures, the social environment will inform access control and information dissemination, and broader community interactions will yield more, and higher-quality, content creation. An individual's profile will grow to reflect activities across a range of topics and sites in a unified manner; information about an object will similarly grow to reflect perspectives accumulated across several communities.

The result will be a rich new PeopleWeb formed by users and their interactions with increasingly rich content. Consider the following scenario. On the PhotoManiacs site, Andrew can tag a given attribute of an object—for example, "num-pixels" of the Nikon D80—with the value "10.2M," add a review, and provide access to this metadata to the group AndrewPhotoBuddies, which he is managing on a different site. Any of those buddies who visit PhotoManiacs and join this group will be able to see the review. Further, this object can be viewed by a different user, say Raghu, at another site—for example, on Mike's vitality page on Facebook—and all the metadata that it accumulated at PhotoManiacs will be accessible, subject to the access rules.

Components

Figure 1 shows the PeopleWeb's main components, which include two emerging capabilities:

- *Global object model*. Users will be able to reference a broad range of objects from anywhere on the Web, and they will do so based on a common identity for both objects and individuals, and in many cases even commonly accepted attributes (such as num-pixels for a digital camera).
- *Portable social environment*. As users move from one site to another, their personas and the social networks they belong to continue to be maintained, unless they choose to assume a different persona. All of a user's activity in a given persona might be aggregated, at the user's behest, leading to more robust models of user reputation and social structure.

Two existing aspects of online activity will change significantly in response to these new capabilities:

- *Communities*. These will expand to exploit people and objects that move seamlessly from site to site, leading to novel kinds of interwoven communities and increasingly richer content structure.
- *Search*. Targeted information discovery through search will leverage rich reputation-weighted metadata of user creation, modification, and consumption of content. Similarly, serendipitous information discovery through push channels will interpret these signals in the context of the portable social environment.

There are sufficient economic incentives for all current Web participants to contribute to the PeopleWeb. Likewise, there are deployment models that provide incremental return on investment for individual components rather than requiring the entire network to come into being before any value can be derived from it.

As we consider how these components might evolve, the theme of a centralized versus a distributed infrastructure will arise frequently. While the former approach has the appeal of technical simplicity, the Web has repeatedly shown itself to be anarchistic, and distributed solutions are viable for many of the problems we consider. The choices made with respect to centralized or distributed management of identity will profoundly impact the future shape of online communities and information discovery.

Privacy concerns

The PeopleWeb raises potential privacy concerns, in particular the notion that an individual's identity is portable and that activity is tracked across sites. It is important—and entirely feasible—to ensure that users retain explicit control over the management of their identity.

There are sufficient economic incentives for all current Web participants to contribute to the PeopleWeb.

A PeopleWeb user can have multiple personas, just as in today's Web; the key difference is that a persona is not synonymous with the user's activity on a single site. The user must continue to control the choice of which personas to assume in a given context. Thus, a user might choose to consistently use one persona in all sites that deal with his professional interests and a distinct second persona in all online fantasy sports. Both personas are part of the user's portable social environment and are available in a site-independent manner—conflating these identities is not permissible without the user's explicit opt-in. Responsible third-party sites will be diligent in ensuring that events tied to an identity are carefully controlled according to the policy that identity establishes. Irresponsible sites will suffer as users vote with their feet.

The ongoing tension between users and publishers regarding privacy is also likely to continue. We expect users to continue sharing much of their created content with private groups, and we expect the PeopleWeb to actually facilitate limited sharing in many situations where the Web currently forces consideration of more public alternatives because social context is not readily accessible. That said, there are serious considerations around the sharing of rich and high-volume metadata such as attentional metadata, and major vendors such as Yahoo! are exploring ways to offer more capabilities without stretching the social contract with users.

CURRENT TRENDS

Four key trends are emerging with respect to textual content. First, user-generated public content has surpassed "traditional" content creation in volume. Second, novel forms of explicit social media metadata such as tagging and reviewing still lag behind anchor text (itself a form of explicit social media metadata), even using a conservative estimate of anchor-text generation rates. Third, attentional metadata has considerably more volume than anchor text, and thus potentially represents the most valuable untapped source of information about resource quality. Fourth, structured listings are actually arriving on the Web at a daily rate greater than that of Web pages themselves.

Content creation

Imagine that each of the world's six billion people spends four hours per day typing aggressively at 100 words per minute. The total annual output of this process would be about 52 petabytes per year, assuming words are represented using a code that takes, on average, one byte per word (the entropy of English text is roughly five bits per word). At current storage costs of $500 per terabyte, sufficient capacity for all this text could be purchased for $25 million per year. By 2010, buying storage to hold all global textual output would be financially equivalent to maintaining 10 people on payroll. Thus, any company that could afford to hire 10 more workers for a business-critical purpose could choose instead to store the planet's entire textual output going forward to eternity.

This somewhat facile analysis does not consider the cost of managing and using the storage, but it seems reasonable to conclude that within a small number of years, any company that sees business value in preserving all produced text could realistically do so. Access rights will be a more serious impediment than scale.

Content can be divided into five distinct classes:

- *published*—professionally printed content such as books, magazines, and newspapers;
- *professional Web*—produced by somebody being paid to do so, such as a corporate site's Web master;
- *user-generated*—produced by individuals and posted publicly, such as a MySpace profile, book review, blog, comment, or personal Web page;
- *private text*—produced by an individual but visible only to a limited set of other individuals, such as instant messages or e-mails; and
- *upper bound*—in our hypothetical scenario, content produced by six billion people typing four hours per day at 100 words per minute.

Table 1. Daily content creation.

Content type	Amount of content produced per day
Published	3-4 Gbytes
Professional Web	~2 Gbytes
User-generated	8-10 Gbytes
Private text	~3 Tbytes (300× more)
Upper bound	~700 Tbytes (200× more)

According to a study by Peter Lyman and Hal R. Varian, total textual published content, including duplicates, currently equals about 14 Gbytes per day, mostly from newspapers (www.sims.berkeley.edu/how-much-info-2003). Due to syndication and other causes of duplication, we apply a 4× correction factor to estimate unique published content at around 3 to 4 Gbytes per day.

Each week, a high-quality Web site will publish new content amounting to about 5 percent of its total content.[2] We estimate high-quality sites contribute about 1 billion pages to the Web, and that 90 percent of these pages come from catalogs and similar database-backed mechanisms, spam, and other automated sources. Of the remaining 5 million pages per day, we assume 2 million pages per day are paid professional content and the rest are user-generated content or other forms of unpaid content. Assuming 1,000 Kbytes per page of unique professional Web content, this results in 2 Gbytes per day.

For user-generated content, we note that Yahoo! Groups contains about 7 billion posts; assuming 5.4 million posts per day, at roughly 200 bytes per post of unique text, this amounts to 1 Gbyte per day. BoardReader also indicates roughly 5 million posts per day. We assume these contributions overall represent perhaps 40 percent of the total posts—including both group-hosting organizations such as MSN Groups and privately hosted forums not completely indexed by BoardReader—to yield about 5 Gbytes per day of board postings. Technorati quotes 1.6 million indexed posts per day and BlogPulse 1.1 million per day. Assuming 1,000 Kbytes per post of unique content results in 1.6 Gbytes per day. Estimates of Netnews production of textual content range from about 500 Mbytes per day to 2 Gbytes per day, of which some nontrivial fraction are duplicates.[3]

MySpace hosts more than 140 million users, each of whom can introduce and update a profile page. Assuming that a quarter of users contribute 200 bytes of text to profiles and message boards during a given week yields 1 Gbyte per day of new textual content. Without performing a detailed analysis of Web page hosting organizations, related profile and social network sites such as LinkedIn, Facebook, and Friendster, and other forms of user-generated content like comments, we conservatively lower-bound such content at 10 Gbytes per day. Note that Wikipedia is not a significant contributor to this volume.

For private communication, e-mail remains the dominant form. Lyman and Varian estimated 60 billion e-mails per day in 2006. At a conservative 50 bytes per e-mail message of novel text, this yields a total of 3 Tbytes per day of private text.

Thus, at a high level, it is possible to characterize publicly visible text creation as being around 10 Gbytes per day, while private text creation is about two orders of magnitude higher at 3 Tbytes per day, and the upper bound on text creation is roughly another two orders of magnitude higher at 700 Tbytes per day.

Table 1 summarizes daily content creation rates for all categories.

Metadata creation

By its nature, metadata has fuzzy boundaries—it is possible to argue that, for example, an insightful book review published in *The New Yorker* is content rather than metadata about a book. However, online metadata typically has four key forms:

- *anchor text*—the underlined text in a hyperlink that can be clicked to take the user to another page;
- *tags*—single words or short phrases placed on a resource such as a picture or URL to aid in retrieving or sharing the resource;
- *page views*—the act of viewing a page; and
- *reviews/comments*—free-form text associated with a resource, such as a book, movie, URL, or product.

We estimate the rate of metadata generation as follows. Yahoo! generates about 8 percent of worldwide clicks, representing some 110 billion clicks monthly. This equals about 46 billion clicks per day worldwide. Assuming four bytes of data indicate a click's location, this yields 184 Gbytes per day of click data, without including information about the user, time of click, and so on.

For anchor text, we conservatively estimate that clearly valuable anchor text arises from the top 10 links on each of the top billion pages, resulting in 10 billion total links. We assume that 5 percent of this amount is created each week, yielding 71 million new links per day.[2] Of these, we estimate 10 to 20 percent are links to other sites and thus represent social media metadata.[4] This results in approximately 10 million new pieces of valuable anchor text per day, at about 10 bytes per anchor, resulting in 100 Mbytes of new daily anchor text.

Tag information comes from estimates of tagging growth rates on the Yahoo! network. For reviews, we consider as surrogates some popular review Web sites such as Epinions and Amazon. The latter has around 2.1 million reviewers,[5] and the number of reviews falls off quite rapidly—reviewers in the top 1,000 may have only 70 reviews in their life. Thus, we estimate 2 million reviewers, with on average three reviews each, for a total of

6 million reviews. Even assuming a few paragraphs each, this comes to 2 Mbytes per day of Amazon review content, scaled up 5× to 10 Mbytes per day of total review and rating content.

Table 2 summarizes daily metadata creation rates for all categories.

Finally, for structured listings, eBay reports an average of 7 million new structured listings per day for 2006, and Yahoo! and Google are within an order of magnitude of this number across their various structured properties.

Table 2. Daily metadata creation.

Metadata type	Amount of metadata produced per day
Anchor text	100 Mbytes
Tags	40 Mbytes
Page views	180 Gbytes
Reviews/comments	~10 Mbytes

GLOBAL OBJECT MODEL

For certain types of commonly referenced objects, a global name scheme greatly enhances what social interactions can achieve with respect to creating meaningful descriptions of these objects. Referring to the URL of an object such as a digital camera is straightforward, but referring to the object in a way that is common across Web sites is more difficult; a camera, like many other objects we interact with online, is not a first-class object in today's Web.

Even if a human can correctly interpret the reference—which is by no means clear as, for example, identical products may have different names in different geographies—a Web search almost surely will not generate all references to the underlying object. A cursory attempt to gather all discussion, reviews, and pricing information for such an object will illustrate this problem. The same is true for many other types of objects, such as movies, restaurants, and even people themselves.

New capabilities

A canonical reference scheme for certain key types of objects enables several new capabilities. First, simple objects can aggregate metadata and consumption patterns from across the Web. Objects can expose Web service calls providing information necessary to display the object on a page in a remote Web site so that remote applications can easily benefit from the presence of a clean, high-quality repository. If the display information allows viewers to potentially contribute ancillary metadata (reviews, ratings, and so on) to the repository, an ecosystem might emerge around the universe of objects.

Once the simple objects are in place, creating richer compound objects with embedded references to other objects becomes easier. This could be as simple as an event object that contains references to a venue, or as sophisticated as a guidebook capturing the HDTV market, with embedded references to all relevant models, manufacturers, and distributors.

While individual users might benefit from accessing information about an object and might in some cases be willing to contribute some data to the object repository, there are other use cases in which an entire community forms around a set of objects—for example, types of cars, tech gizmos, or geographically proximate restaurants. Work performed in these communities will benefit the global object universe.

Representing structure

A schema is, informally, the set of attributes used to describe a collection of similar objects. If cameras are described in terms of num-pixels, manufacturer, and price, this set of attributes constitutes a schema for cameras. There are a number of complications in maintaining a structured view of objects. For example, we might not know num-pixels for a given camera, and the price for another is only an estimate; we must develop graceful ways of dealing with such missing and uncertain values. Another challenge is that several schemas will likely emerge—especially across sites—for describing the same class of objects, leading to inconsistencies. For example, a second schema for describing cameras might refer to price as cost and represent it in yen rather than dollars, or it might not distinguish digital cameras from traditional cameras, and therefore not even recognize the attribute num-pixels. These integration issues have been widely studied, and their difficulty is well-recognized.[6]

Metadata

With a global object model in place, it becomes possible to place certain types of metadata on all objects. On today's Web, four metadata types apply broadly to all object types:

- *stars*—three stars, thumbs up, "I digg this," and related forms of low-information-content positive or negative feedback;
- *tags*—short textual words or phrases associated with an object to support retrieval or sharing;
- *attention*—a user viewed this object, clicked on it, or interacted with it, implying some level of satisfaction or interest; and
- *text*—a review, comment, or other piece of textual information associated with the object.

We expect to see significant work combining this STAT metadata with user reputation measures to produce overall scores of object quality in various contexts.

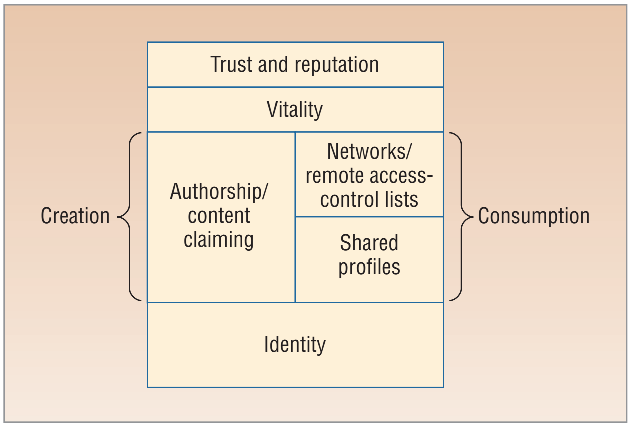

Figure 2. Portable social environment. Users could navigate the Web or employ a new site to benefit from a wide range of social connections in the current context without reconfiguring the environment each step of the way.

Such scores will of course be augmented by additional type-specific metadata, but they represent a nontrivial starting point for understanding where, and to what extent, objects are relevant.

De facto standardization

The PeopleWeb is fundamentally about people and how they reference content in the context of their social neighborhood. If the Semantic Web were to reach a tipping point and gain significant traction, this would greatly contribute to the global object model.[7] Failing this, however, we expect the model to arise from other mechanisms.

Rob McCool proposes a Named Entity Web[8] as a highly simplified form of the Semantic Web in which pages can easily declare objects using a few new HTML attributes and can assign a type from an initially small set of choices. Such a simple system, if successful, could also bootstrap the PeopleWeb.

A third and more likely possibility is that large organizations with substantial content repositories will be the key players in introducing the global object model. Consider, for example, a company that has significant market share in selling digital cameras. Suppose that it introduces a naming scheme for the cameras in its catalog, develops and publishes a simple schema, then opens a Web service allowing queries to attributes for the schema. The company pursues this path because the Web service will also provide prebuilt, customizable, embeddable HTML snippets that can be used on enthusiast and other sites, potentially driving traffic back to the organization.

Such a schema and naming system could potentially become a de facto standard, and as adoption grows, ancillary support structures would emerge to, say, map references from this company to a competitor pursuing the same strategy. In fact, in the PeopleWeb global identities might emerge in a decentralized fashion; several organizations could create naming schemes for the same object—for example, Nikon and PhotoManiacs might both introduce naming schemes for a Nikon camera—and over time, as versions of the object using these schemes move across the Web and grow, they are likely to be identified in a community-driven manner.

Thus, we expect that a handful of organizations will seed the global object model with a small number of high-value object types, such as products and local listings, and the system will begin to grow as usage of this seed set expands. Increasing Web developer familiarity with Web services and user consumption of structured content make the opportunity to develop a de facto standard object model for a particular domain attractive to competing organizations.

PORTABLE SOCIAL ENVIRONMENT

Users create online content in numerous formats and access-control environments. Much of this content is to be shared with others, but not publicly. Historically, such sharing was point-to-point, as in e-mail or instant messages. Today, however, content is also shared with groups. Many of these groups are relatively small and managed by the user creating the content—as in, for example, Flickr's friends and family features. However, also common are online forums such as those provided by Yahoo! or MSN Groups, or by various niche sites such as www.fredmiranda.com for digital cameras and PriusChat for the Toyota Prius.

The number of users with privileged access to particular content is often quite large. For example, more than 18,000 people have access to the cell phone number of one of this article's authors via Facebook's MIT network. The PeopleWeb will clearly increase the volume of broadly disseminated but nonpublic content. Simple access-control approaches that assume information is available to all, or to a limited number of close friends, are simply inadequate in this setting. New technical approaches will be required.

Moving beyond access control to the broader issue of social networks, these range from close friendships to family ties to interactions that might not even require acquaintance. Given the complexity of interpersonal relationships, users cannot be expected to reproduce their place in the social universe at every new site they visit, especially as the number of sites a user visits is likely to increase. As the venue of consumption continues to fragment, cross-site identity and credentials will become increasingly important.

A portable social environment, shown in Figure 2, would let users navigating the Web or employing a new site benefit from a wide range of rich social connections in the current context, without needing to reconfigure this environment each step of the way. However, each layer in the stack also presents unique challenges.

Identity

Most Web users employ many different logins online to authenticate with multiple sites. While OpenID (http://openid.net) offers a partial solution to this problem by connecting users to a particular online identity, only a small fraction of logins currently use the protocol. In fact, the current Web has no persistent and omnipresent notion of identity. This will emerge through a single provider such as OpenID gaining critical mass, or through widespread adoption of browser-side tools that manage a given user's multiple identities as the user moves across sites.

Shared profiles and networks

Even if a protocol to let users sign in to any Web site using the same user ID and password were adopted globally, no widely adopted solutions exist to making profile information available to all Web sites. As sites commonly offer capabilities that increase in value as users provide a social network, users find themselves manually re-creating similar or identical social networks on every new Web 2.0 site they visit.

It is natural, however, to imagine sharing profile information by simply asking users to provide an optional URL referencing some profile data, such as a list of contacts expressed using global identities. Such schemes are clean if the new site need only consume the URL's content, perhaps refreshing it on new logins. However, once users begin to update their social network or other profile data, synchronizing the copies requires a more sophisticated protocol that may be undesirable.

> **The nature of online communities will inevitably change to exploit richer data.**

Authorship

Users can author one or more blogs under various identifiers—some kept separate by design, some for technical reasons. They can post comments to numerous other blogs as well as post content to various forums. They can also enter reviews of books, movies, or other products online and install tools through companies like del.icio.us or StumbleUpon to place lightweight metadata on URLs throughout the Web. *Content claiming* is the capability to aggregate all content authored by a user into a single stream so that friends, family, or other interested parties can see the user's public activities wherever they occur.

There are two basic primitives for content claiming. The first is to claim a source of information, such as a blog or personal homepage. Typically, the verifier creates a random string and asks the user to make the string appear somewhere in the content source that only the owner controls.

The second basic primitive concerns authorship on non-user-controlled sources such as comments, reviews, or forum posts. The simplest approach is for the user to place a unique signature such as a URL or e-mail address in all such posts and then ask a central repository to ascribe all such content segments to the user. However, anybody can pretend to be the given user or plagiarize the user's text. Also, much of the content might not be crawled in a timely manner, and even if crawled efficiently, the user's contribution to the page might not be segmented properly. Alternately, a user can introduce browser-based agents to send notifications when posting new content to address coverage and segmentation concerns and can adopt simple cryptographic protocols to combat spoofing and plagiarism.

Vitality

Once a content-claiming scheme exists, users can subscribe to a particular person and consume all activities that person performs. However, more nuanced forms of subscription would be valuable, including the following:

- Show me all of Bill's activities in groups to which I also belong.
- Show me all activities on this site by CMU graduates.
- Show me important updates of these 500 people I knew at my last job.

Vitality platforms currently provide these capabilities on a single site, but to date no clean and scalable solution has been implemented across the entire Web. With identity and authorship in place, however, this becomes possible.

Reputation

Reputation will be necessary to interpret the many interactions between users and data in the PeopleWeb. eBay's reputation-management system has been shown to provide an 8.1 percent average boost in price to high-reputation sellers over newcomers,[9] but this and other one-dimensional reputations common on today's Web are probably not high-fidelity representations of reality. More research is required on this topic.

COMMUNITIES

The nature of online communities will inevitably change to exploit richer data. As a community provider in a particular vertical domain reaches out to incorporate relevant structured content and offer a better user experience, other community providers must either conform or perish. Similarly, as identity and social environment become portable, communities will evolve to become more user-centric.

Communities around data

The changes we anticipate are likely to most impact communities of shared interest, such as academic communities in a given discipline or photo enthusiasts, and communities of purpose, such as a technical support group for a company's products. Content in these communities is often richly structured, and the global object model and portable social environment will facilitate focused interactions among community members to capture and share such content.

Core sites. A given community's content-creation activity typically centers on a few core sites organized in a way that reflects the community's interests. For example, a photography site might be organized by camera products or photo types, while a university's alumni site might be organized by discipline, year of graduation, and location. The site content reflects the community's interests as well: The photography site might contain collections of photo objects and digital camera objects (described in terms of make, model, pixels, and focal range); the alumni site might contain lists of alumni with name, address, year of graduation, major discipline, and current employer.

The evolution of a community often leads to the creation of other related communities.

Attributes. This begs the question: What are the sources of structured content? The community application typically captures some common attributes automatically as a by-product of user activity—for example, for each question or answer, the author ID, time of posting, and so on; for each object, such as a camera, all reviews and products named in associated user-generated tags. This information can be aggregated by author to create personal profiles that reflect the user's cumulative activity and to construct social networks based on explicit links such as buddy lists and implicit links such as a user responding frequently to another's postings.

In general, however, there are attributes whose values must be explicitly provided or inferred in some way. For example, a user can provide the price of a camera, or perhaps information extraction techniques can infer it from a product description Web page, but it is not obtainable by simply recording normal user activity. Users can provide such structured data by means of a catalog or a feed in which attribute-value pairs for several objects are input in some agreed-upon format, through APIs such as Google Base and Google Co-op, or by using a mechanism for per-object attribute tagging.

Integration. Community sites that capture or use such structure must be aware of an underlying model of certain classes of entities and relationships among them.[10] For example, an academic community like DBLife (http://dblife.cs.wisc.edu) or Rexa (http://rexa.info) is aware of entities such as authors, conferences, and publications, and relationships such as program committee membership and coauthorship.

Integrating structured data from multiple sources is problematic. In the PeopleWeb, incremental conflict-resolution techniques such as dataspaces[6] will likely prove more viable. Further, attributes ultimately derived from user input—directly via a feed or attributed tagging, or indirectly via information extraction from a Web page—also have questionable fidelity to an underlying reality. Thus, a user-provided camera price can be inaccurate, extraction error can occur when inferring an individual's phone number and e-mail address from his home page, and aggregating bibliographic entries from multiple Web pages, in which several distinct authors might share the same name, can lead to inaccurate coauthor listings.

Community interactions

Once identity and social environment become portable across sites, members can participate in a given logical community and interact around objects of common interest regardless of which site they are on. This could be disruptive, as it contravenes the current model in which a given community exists on a single site.

Tools. We anticipate that tools for in-situ consumption and creation of structured content—for example, for attributed tagging, rating, and reviewing—will become commonplace. These tools could be packaged and distributed as toolbar extensions or as callable APIs supported by community sites focused on a particular type of content. For example, a site such as LinkedIn might offer APIs that let someone view and comment on user profiles while on a Facebook page. Such a capability exists today on Facebook but is not globally available.

Site features. Coming changes will greatly impact community features at the site level as well as on the Web level. Sites must involve users in creating and maintaining structured content. While site content might come from editorial activity or standard catalogs, successful organic growth of communities around objects and data imposes some requirements.

First, community members must be able to shape relevant data—for example, through wiki-style editing and annotations. Second, community managers must be able to help identify relevant data sources and channel the community's interactions to produce high-quality feedback from members. Third, a site that seeks to expose and maintain structured descriptions ideally will let users correct errors and explain how values were obtained—for example, by identifying the user who provided the value or the page from which it was extracted, perhaps with some indication of the user's reputation or the confidence in the extraction procedure and the source page's credibility.

Federations. The evolution of a community often leads to the creation of other related communities. For example, a Nikon enthusiasts club is likely to foster similar clubs for Canon, Leica, and other camera makers. Users are likely to belong to several such clubs and to want to search across all camera clubs.

To support this, community sites must have mechanisms to form federations. Some federations are loose, for example, cooperating to share search capabilities. Others have closer ties. For example, Freecycle (www.freecycle.org) is a community recycling organization that allows people to announce items available for free, and others to claim these items. The ability to organize clubs at the city, district, and state levels online, with shared hierarchical search, common moderation policies, and default moderator privileges, would be immensely useful.

Creation platforms. More flexible community-creation platforms are also likely to emerge. This trend is already reflected to some degree in the emergence of sites such as Ning (www.ning.com), but future communities are likely to exhibit even more customizability with respect to content structure.

The PeopleWeb presents numerous challenges and opportunities from both a technical and a commercial perspective.

SEARCH

Information discovery through search is and will remain the driving force behind rich patterns of access to Web content. However, we envision that search will change significantly. Much content of interest to users, both at the level of serendipitous consumption through a network or recommendation, and at the level of targeted information discovery via a search engine, will be created by other users rather than professionally produced. The objects to be retrieved will have structure, ownership, nontrivial access-control restrictions, and a broad range of heterogeneous metadata gathered from many sources.

Current state

Blogs and bulletin boards represent a good case study of the current state of Web search. The data model for both content types is multilevel, including individual time-stamped posts and higher-level structures. The HTML delivering the content consists of posts wrapped in a rich, templated envelope containing blog rolls, thread or forum information, and the like.

Authors may differ from post to post or comment to comment, and readers will probably understand that authors have various degrees of expertise. Forums often provide formal cues to the level of experience and engagement of the author of a particular piece of content—member, senior member, and so on. Each author can also be tracked historically through the blog or forum, and across the Web, to get a sense of common topics, quality level, and user response. Most of these natural inputs to ranking are not used today because they are difficult to extract, and carefully tuning a state-of-the-art ranking function for this type of data requires a high level of sophistication.

In a world of collaborative authorship, rich content types, and embedding of objects within numerous other objects, each with its own reputation and consumption patterns, available tools will clearly be inadequate. Recent work on search over semistructured content is applicable to this problem,[11] but a wide range of new issues have not received thorough study to date.

Future search

Objects to be discovered in the future must have structured attribute values available from a broad range of sources. Will information discovery involve centralizing this vast body of data or developing a distributed platform in which agents cooperate to respond to a user's search, track an alert, and so on? Although global objects and portable contexts can realistically be hosted in a distributed manner, even in today's relatively simple search ecosystem there is no credible distributed search platform at the scale supported by the major search engines.

Ranking fundamentals will become more complex in the future, suggesting that despite the many compelling advantages of a distributed paradigm, technical feasibility will demand centralization, at least during the first few generations of this evolution. To be successful, such a centralized system must provide many touch points for both receiving and sending information, and it must present open standards that let it return value to the ecosystem.

Further, the Web almost completely fails to address the critical problem of access control. Major search engines index the public Web and leave the rest untouched, allowing individual sites to provide search over private content as they see fit. Information will probably be meaningfully restricted to groups that range in size from two to tens of thousands or even larger. It is not feasible to first load all content matching a query and then restrict access, as today 99 percent of online content is private. Search over all relevant private repositories, which could easily number in the tens of thousands, is likewise impossible.

In short, search in the PeopleWeb will be a very different problem than it is today, with significant shifts in technology and approach.

The emergence of two new capabilities on the Web—a global object model that enables creation of richer structured content, and a portable social environment that facilitates user-centric rather than site-centric communities—will radically transform the way people interact online and discover information. This PeopleWeb presents numerous challenges and opportunities from both a technical and a commercial perspective. ■

Acknowledgments

We gratefully acknowledge the help of many colleagues who have contributed to this work, in particular Phil Bohannon, Andrei Broder, AnHai Doan, Ravi Kumar, Cameron Marlow, Tim Mayer, Jan Pedersen, Prabhakar Raghavan, Neal Sample, Joshua Schachter, and Eckart Walther.

References

1. S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications,* Cambridge Univ. Press, 1994.
2. A. Ntoulas, J. Cho, and C. Olston, "What's New on the Web? The Evolution of the Web from a Search Engine Perspective," *Proc. 13th Int'l Conf. World Wide Web,* ACM Press, 2004, pp. 1-12.
3. E. Sit, F. Dabek, and J. Robertson, "UsenetDHT: A Low Overhead Usenet Server," *Proc. 3rd Int'l Workshop Peer-to-Peer Systems,* LNCS 3279, Springer, 2005, pp. 206-216.
4. K. Bharat et al., "Who Links to Whom: Mining Linkage between Web Sites," *Proc. 2001 IEEE Int'l Conf. Data Mining,* IEEE CS Press, 2001, pp. 51-58.
5. N. Jindal and B. Liu, "Review Spam Detection," *Proc. 16th Int'l Conf. World Wide Web,* ACM Press, 2007, pp. 1189-1190.
6. A. Halevy, A. Rajaraman, and J. Ordille, "Data Integration: The Teenage Years," *Proc. 32nd Int'l Conf. Very Large Databases,* VLDB Endowment, 2006, pp. 9-16.
7. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific Am.,* May 2001, pp. 34-43.
8. R. McCool, "Rethinking the Semantic Web, Part 2," *IEEE Internet Computing,* Jan./Feb. 2006, pp. 93-96.
9. P. Resnick et al., "The Value of Reputation on eBay: A Controlled Experiment," *Experimental Economics,* June 2006, pp. 79-101.
10. A. Doan et al., "Community Information Management," *IEEE Data Eng. Bull.,* Mar. 2006, pp. 64-72.
11. S. Amer-Yahia and M. Lalmas, "XML Search: Languages, INEX and Scoring," *SIGMOD Record,* Dec. 2006, pp. 16-23.

***Raghu Ramakrishnan** is chief scientist, Audience, and a Research Fellow at Yahoo! Research. His interests include data mining, online communities, and Web-scale data management. Ramakrishnan received a PhD in computer science from the University of Texas at Austin. He is a Fellow of the ACM. Contact him at ramakris@yahoo-inc.com.*

***Andrew Tomkins** is vice president of search research at Yahoo! Research. His work focuses on the measurement, modeling, and analysis of content, communities, and users on the World Wide Web. Tomkins received a PhD in computer science from Carnegie Mellon University. He is a member of the IEEE and the ACM. Contact him at atomkins@yahoo-inc.com.*