
Web Applications: Web 2.0 and Beyond

Thomas Yan
Topix LLC
thomaskyan@gmail.com

COEN 162 Guest Lecture
Santa Clara University
May 24, 2007

Speaker Bio – Thomas Yan

■ Past:

- BS Computer Science (San Jose State University, 2004)
- MS Computer Engineering (Santa Clara University, 2006)
- Carnegie Institution of Washington (Stanford, CA)

■ Present:

- Engineer at Topix (www.topix.com)

Talk Outline

- Web 2.0
- Semantic Web
- Ongoing Discussions
- Outlook

Web 2.0 by Example [Orl05]

Web 1.0	Web 2.0
DoubleClick	Google AdSense
Ofoto	Flickr
Akamai	BitTorrent
mp3.com	Napster
Britannica Online	Wikipedia
Personal websites	Blogging
Evite	Upcoming.org and EVDB
Domain name speculation	Search engine optimization
Page views	Cost per click
Screen scraping	Web services
Publishing	Participation
Content management systems	Wikis
Directories (taxonomy)	Tagging (folksonomy)
Stickiness	Syndication

Example: Content Delivery

- Akamai: replicate content over a network of distributed servers. Brings content closer to users.
- BitTorrent and other P2P: each client is also a server. Files are broken into fragments and served by multiple locations.
- BitTorrent demonstrates a key Web 2.0 principle: service improves automatically as more people use it [Ore05].

What is Web 2.0?

- Tim Berners-Lee (inventor of the Web) referred to “Web 2.0” as useless jargon that nobody can explain and a set of technology that tries to achieve the same thing as Web 1.0 [Cla06].
- Web 2.0 relies on technologies that have been around for years (HTML, HTTP, JavaScript, etc.)
- Marketing buzzword by companies trying to stand out in overpopulated and immature markets.

What is Web 2.0?

- It has acquired a meaning according to Paul Graham [Gra05].
- Ingredients of Web 2.0 according to Graham:
 - Ajax
 - Democracy
 - Don't maltreat users
- Thus, Web 2.0 means using the Web the way it was meant to be used.

What is Web 2.0?

- Term coined by Tim O'Reilly to denote a turning point in Web applications after the dot-com crash [Ore05].
 - The Web as a platform (Web 2.0 design patterns).
 - Harnessing collective intelligence.
 - Importance of owning certain core data.
 - End of the software release cycle.
 - Lightweight programming models.
 - Software above the level of a single device.
 - Rich user experiences.

Web 2.0 and Groupware

- Definition of Groupware: “Computer-based systems that support groups of people engaged in a common task (or goal) and that provide an interface to a shared environment.” [EGR91].
- Keep the experiences of Groupware and CSCW research in mind when evaluating the success of Web 2.0 applications as there is much overlap between CSCW and Web 2.0.

Ajax

- “JavaScript now works.” [Gra05]
- A combination of:
 - XHTML and CSS for markup and styling.
 - A Document Object Model (DOM) for dynamic display and interaction with information.
 - Use of XML as a format for exchanging data.
 - XMLHttpRequest used for asynchronous data exchange.
 - JavaScript to bind everything above together.

Document Object Model (DOM)

- API for representing a document (such as HTML or XML document).
- Allows accessing and manipulation various elements (such as HTML/XML tags) that make up the document.
- Tree structure.

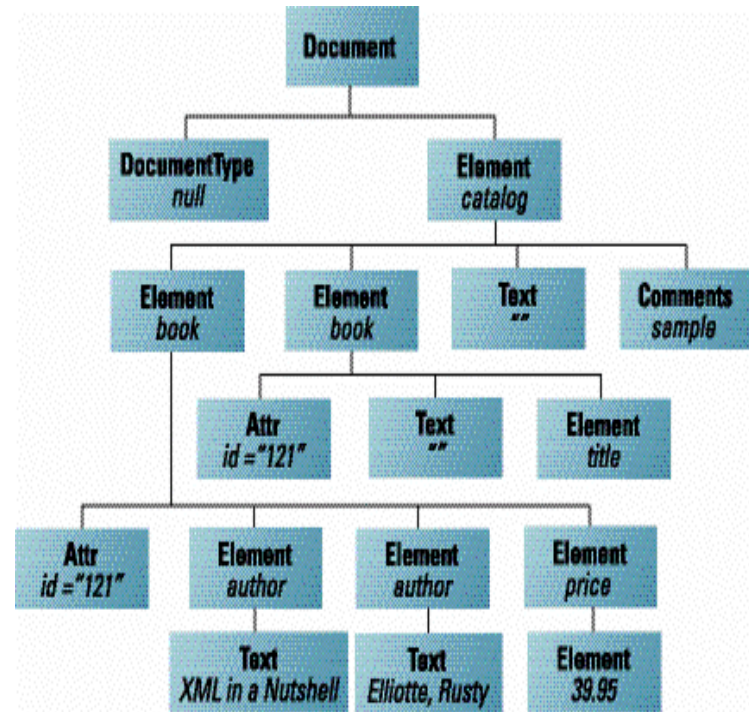


Image Source: www.oracle.com

XMLHttpRequest

- API used to transfer XML or other text data to and from a Web server.
- Enables JavaScript to make HTTP requests without having to reload the page.
- Uses HTTP for communication.
- Applications that use XMLHttpRequest: Google Gmail, Meebo, and Google Maps.

XMLHttpRequest

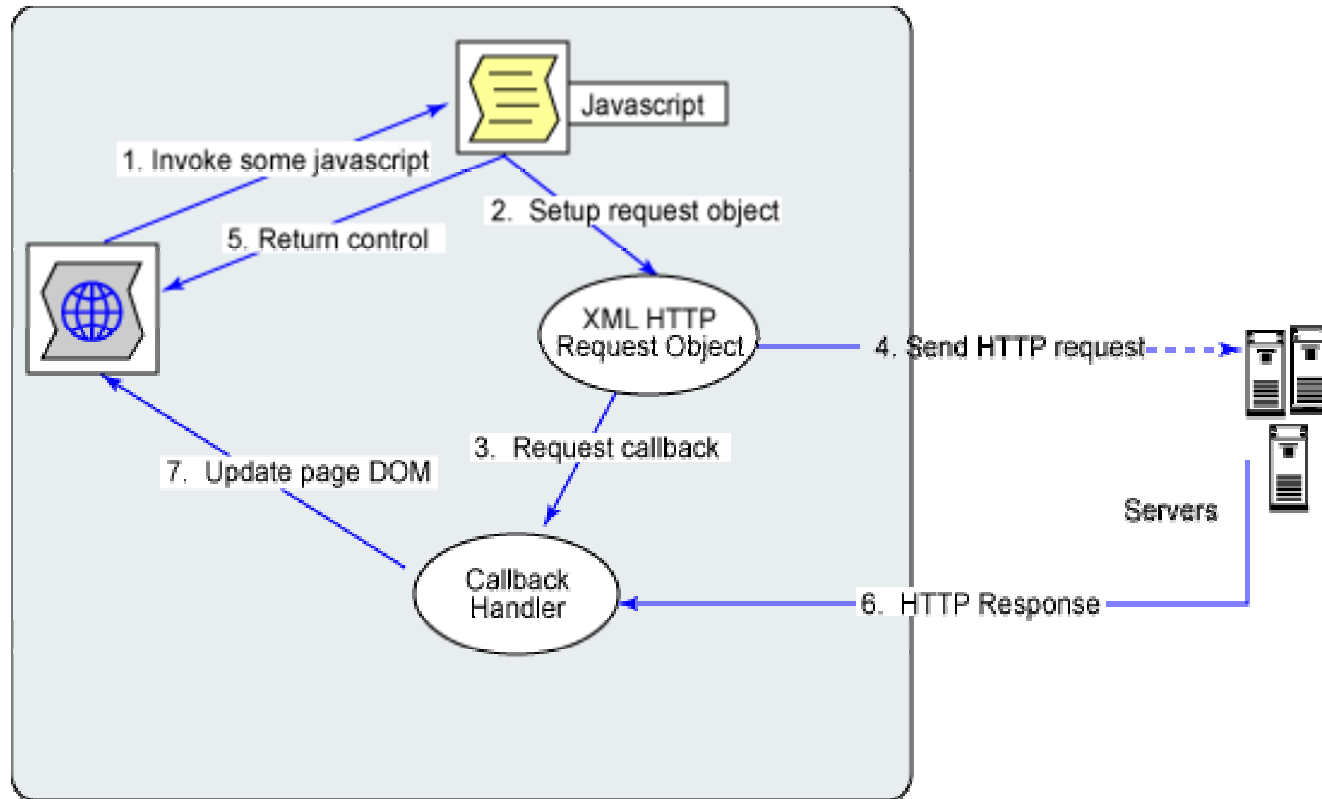


Image Source: www.ibm.com

Ajax Enables Rich User Experiences

- Ajax allows for much more interactive user interfaces on the Web.
- Expect to see Web reimplementations of PC applications [Ore05].
- Examples: Yahoo's new e-mail interface (Outlook-like functionality), Google Documents

Syndication with RSS and Atom

- **Atom:** XML language for Web feeds.
- **RSS:** family of XML-based Web feed formats.
- Both formats are used to publish frequently updated Web content.
- Users may subscribe to a variety of feeds in order to get updated headlines and posts:

Mashups

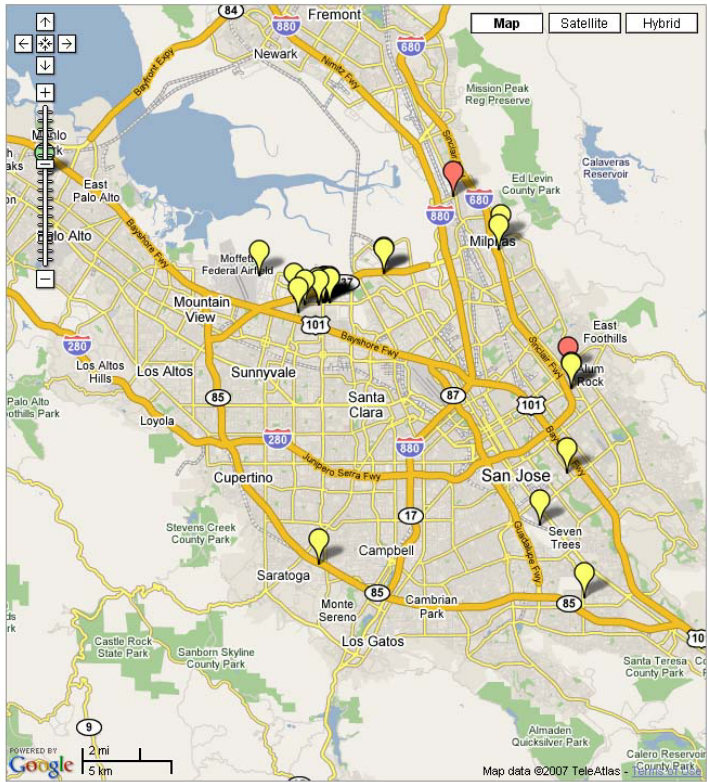
- A website that combines the contents of more than one source to provide an integrated experience.
- Content source: public APIs, Web feeds (RSS or Atom), Web services, screen scraping

Mashups

For Rent For Sale Rooms Sublets

City: SF - South Bay Price: \$150K - \$300K Show Filters^{New} Refresh Link

Powered by craigslist and Google Maps
(this site is in no way affiliated with craigslist or Google)
[About / Feedback](#)



pics	price	description	city	date
●	\$300K	Charming Condo in Fantastic Complex!	San Jose	5/19
●	\$179K	Beautifully Maintained Home	Sunnyvale	5/19
●	\$210K	Mobile Home - You Own The Land	Campbell	5/19
●	\$165K	4271 North First Street #144 - San Jose 95134	San Jose	5/18
●	\$170K	Homestead High School Area, Affordable 3 Bdrm 2 Bath 1560 sq ft Home	Sunnyvale	5/18
●	\$300K	Fabulous 1/1 Condo in Heart of San Jose	San Jose	5/18
●	\$289K	Near Ocean, 55+ Resident Owned Park - Open House Sat/Sun 1-5	Santa Cruz	5/18
●	\$195K	Large beautiful manufactured home - Just like new inside!	Sunnyvale	5/15
●	\$169K	Stunning Manufactured Home in Pristine Park with fenced in yard!	Sunnyvale	5/15
●	\$185K	Beautiful manufactured home with cathedral ceiling and large rooms!	Sunnyvale	5/15
●	\$300K	Price Reduced! Lease Option - Milpitas Condo	Milpitas	5/14
●	\$295K	Comfortable Condo, Pool & Spa Near Shops	Watsonville	5/14
●	\$205K	10 minutes to Beach Boardwalk, 30,000 first time buyer credit program	Scotts Valley	5/13
●	\$165K	4271 North First St #144 - San Jose 95134	San Jose	5/12
●	\$293K	Crossroads Incentives Continue to Excite	Milpitas	5/12
●	\$190K	Land Priced for Profit - Yours	Watsonville	5/11
●	\$219K	Open House -- Saturday 12-4pm -- New Home (Cda703)	Sunnyvale	5/10
●	\$189K	New Home -- 3 bd / 2 bth -- Sat Open House (Aw518)	Sunnyvale	5/10
●	\$249K	3 bd / 2 bth -- Open House Sat 12-4pm (Cl507)	San Jose	5/10
●	\$300K	Fabulous 1/1 Condo in Heart of San Jose	San Jose	5/10
●	\$159K	Beautiful two car garage manufactured home with fenced in yard!	Sunnyvale	5/10
●	\$300K	Updated Condo, Gated Community	San Jose	5/10
●	\$159K	Feel Right at Home in This Charming Unit	Sunnyvale	5/10

- housingmaps.com combines housing listings from craigslist with map information from Google Maps.

Representational State Transfer (REST) [BT02]

- A collection of architectural principles that attempts to minimize latency and network communication while maximizing independence and scalability of component implementations.
- Addresses the observation that performance in network-based applications are dominated by network communication.
- REST served as a design guideline of much of modern Web architecture (ex. HTTP/1.0, HTTP/1.1).

REST Principles

- Application state and functionality are defined as resources.
- Resources are addressable using a universal syntax.
- Resources share a uniform interface for transfer of state which is made up of well-defined operations and content types.
- Protocols should be: client/server, stateless, cacheable, layered.

Benefits of REST

- Improved response time due to support for caching.
- Improved server scalability by eliminating the need to maintain communication state.
- Requires less client-side software (everything done through a browser).
- No separate resource discovery mechanism needed due to use of hyperlinks in content.
- Provides long-term compatibility.

REST vs. RPC

- REST is a lightweight programming model compared to RPC-based communication.
- REST allows for less tightly coupled systems than RPC.
- REST-based protocols are focused on syndicating data outwards and not on what happens when the data reaches the other end of the connection [Orl05].

The Importance of Data [Orl05]

- Companies are racing to own a certain core class of data.
 - Companies that are able to do so may become the single source for data if there is a significant cost to create the data (ex. maps).
 - Companies also enhance data (ex. Amazon enhances ISBN data with publisher-supplied data such as tables of contents and sample material, in addition to user annotations in the form of reviews).
 - Thus, Amazon, not the ISBN registry provider, is the primary source for bibliographic data on books.
-

End of Software Release Cycle [Orl05]

- Dynamic programming languages (Perl, Python, PHP, Ruby) are the tool of choice for developers building systems that require constant change.
- Perpetual Beta: Web 2.0 products and features released on a monthly, weekly, or daily basis.
- Real-time monitoring of the system to see which features are used and how they are used.
- Radical departure from previous software development cycles. Flickr deploys new builds up to every half hour!

Search Engine Marketing

- Goal is to increase the visibility of a website in search engine results pages.
 - Search Engine Optimization
 - Pay-per-click (advertisers bid on keywords that they believe their target market would enter into search engines)
 - Paid inclusion (feed listings into search engines)
 - Social media optimization (placing ideas in online communities in hopes that it spreads virally)
 - Video search marketing (strategically placing short video clips on sites such as YouTube)

Search Engine Optimization (SEO)

- Consider how search algorithms work and what people search for.
- Fix problems on site that may prevent search engines from fully crawling it.
- Adding unique content to a site and making sure it is easily indexed by search engines.

White Hat SEO

- Intended to produce results that last for a long time.
- Use techniques that search engines recommend as part of good design.
- Includes no deception.
- Create content for users, then making that content accessible to Web crawlers.

Black Hat SEO

- Short term: sites anticipate that they will be banned once the search engines find out what they're up to.
- Uses techniques that are disapproved by search engines in order to improve rankings.

Term Spamming

- Repetition
- Dumping unrelated terms
- Weaving spam terms into content copied from other sites (such as news)
- Stitching together sentences and phrases from other sites.

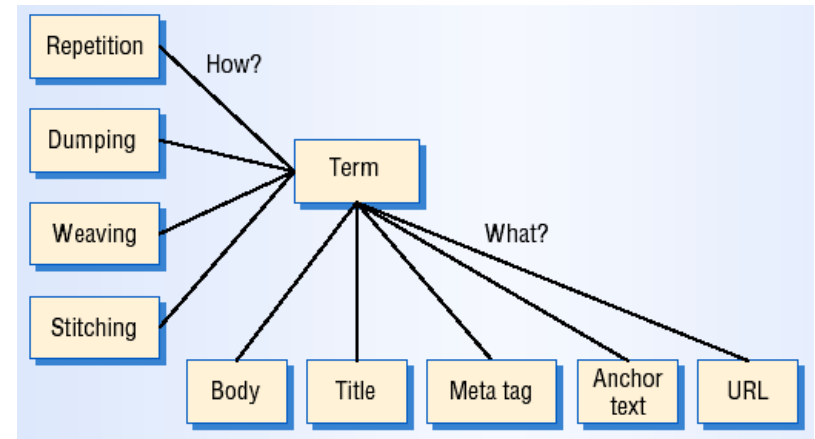


Image Source: [GG05]

Link Spamming

- Creating link structures in hopes of boosting importance of pages.
- **Directory cloning:** replicating the content of well-known directories (like DMOZ).
- **Honey pot:** hide spam links on page with useful information.
- **Spam farm:** group of sites with a link structure that boosts ranking.

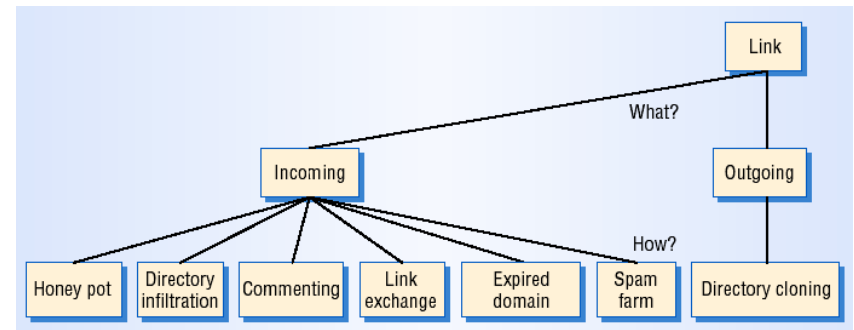


Image Source: [GG05]

Anti-Spam Techniques

Proposed solution	Spam targeted	Function	Automatic?	Useful for
Statistical language model	Term spamming, in particular blog infiltration	Identifies unnatural word distribution	Yes	Spam detection
Analysis of link-count distribution	Link spam farms	Looks at in-degree and out-degree distribution outliers	Yes	Spam detection
Analysis of PageRank distribution	Link spam farms	Looks for unnatural PageRank score distributions of in-neighbor pages	Yes	Spam detection
Collusion detection	Link spam farms	Identifies groups of strongly interconnected pages	Yes	Rendering specific spamming technique ineffective
TrustRank	All types	Separates reputable pages from spam on the basis of connectedness to a set of known reputable pages (seed)	Semi (requires manually compiled seed set)	Spam demotion

Source: [GG05]

Google AdSense

- Websites enroll in Google's AdSense program to enable advertisements on their site.
- Google uses its search technology to serve ads based on Web content.
- Website owners get paid when ads are clicked (revenue shared with Google).
- Websites with more traffic get paid more when ads are clicked.

[Cupertino homes for sale](#)

See current home listings & MLS in Cupertino, CA. Free!
homegain.com

[Cupertino People Search](#)

Find anyone in Cupertino, CA. Get current address, phone, & more!
www.usa-people-search.com

[Cupertino Homes](#)

Instantly View Hundreds of CA Homes Search the Cupertino CA MLS
www.ZipRealty.com

[Delivery in One Day](#)

Same Day Delivery Courier 24/7 Call 408 846-8158
www.streetwisesd.com

Ads by Google

Click Fraud

- Websites display lists of ads (from Google and Yahoo) and little else.
- These sites are part of click-fraud rings, where hundreds of thousands of participants are paid to click on ads.
- Automatic software “clickbots” are also utilized.
- Advertisers pay each time the ad is clicked.
- 10% - 15% of ad clicks are fake, representing \$1 billion in annual billings [BE06].

Collective Intelligence

- Taking advantage of collective intelligence is a key Web 2.0 concept.
- Successful “Web 1.0” companies such as Amazon and eBay did this [Orl05].
- Web 2.0 applications take this even further by allowing users to control the data.

Wikis

- Websites that allow users to add, remove, and edit content.
- Tool for massive collaborative editing.
- Vandalism can be a problem (I once looked up “fascism” in Wikipedia and was redirected to the page for George W. Bush).
- Possible to determine reputation of editors based on how well their edits are preserved and flag changes made by low-reputation editors [AD07].

User-Defined Content (Wikipedia)

- Perhaps most well-known example of user-defined content.
- Online encyclopedia that can be edited by any Web user.
- Low barriers to entry.
- Radical experiment in trust [Ore05].

User Participation on Topix

topix Edit took 0.453s on rcx85 welcome, **tyan** : your page | sign out

Santa Clara News

Local news for Santa Clara, CA continually updated from thousands of sources on the web.

Santa Clara, CA News Forum Wire Classifieds

Most Residents Overweight In Santa Clara Co.

The survey conducted in 2005-2006 focused primarily on high-risk behaviors, called the single greatest factor in a person's health. [via ABC7 - KGO-TV](#)

Posted by [tyan](#) 16 hrs ago | [Permalink](#) | [Comment?](#) | [Kill Story](#) | [Edit](#)

Santa Clara Group Wants 49ers Stadium Put To Vote

SANTA CLARA A group of Santa Clara residents who call themselves "notwithmymoney" wants voters to be able to decide whether a costly new San Francisco 49ers stadium is built in the city. [via Cbs5.com](#)

Posted by [tyan](#) on Thursday | [Permalink](#) | [Comment?](#) | [Kill Story](#) | [Edit](#)

Santa Clara Forum

Transient sentenced for sex with girls	mad	2 hr
Google revamps its Internet search	ko ko	4 hr
Airplanes Spraying San jose CA	san jose report	4 hr

Search

This Topic [Find A Topic](#) [Change City](#) [Search All](#)

Search within Santa Clara, CA

Who's Editing This Page?

[Edits History](#)

tyan (Guest editor)
is editing the Santa Clara News page.

Become an editor today
The **Santa Clara News** readers need your help!

Got a local story for Santa Clara News?
[Submit your news here](#)
or email it to the editors at zipcode@topix.com.
(example: 94043@topix.com)


- Editors post news stories on a Topix page for a city or a subject of interest.
- Any user may apply to edit a Topix page.

User Participation on Topix

topix

Sunnyvale, CA

Sunnyvale, CA [News](#) [Forum](#) [Wire](#) [Classifieds](#)

 Posted by [Groucho](#) on Sunday Apr 29 | [Permalink](#)


Demolition of Sunnyvale Town Center Mall Approved

One Macy's, one Target and one very unused shopping mall on 25 acres of prime downtown real estate.

■ Full story: <http://www.1siliconvalley.com/demolit...>, published Sunday Apr 29

COMMENTS

Showing posts 1 - 4 of 4

<p>Groucho</p>  <p>"Man Bites Dog" JOINED: Dec 12, 2005 COMMENTS: 85 Sunnyvale ISP Location: Hayward, CA</p>	<p>Monday Apr 30 #1 Flag Reply »</p> <p>Wow, it would be so cool if this finally actually truly really ever happens... How many years has it been now?</p>
<p>MrJLH Hayward, CA</p>	<p>Monday Apr 30 #2 Flag Reply »</p> <p>First, I'm not one of those "Tree" saving nuts. But I just want to point out that the "Five" trees that the old mall surrounded were planted by old man Murphy in the 1800's. They represent the five native species of trees in California. A great deal of effort and money was spent back in the 70's to save them at taxpayers expense. I know because I had to water them every day during construction. The original developer did not care and nearly cut them all down. If they are stressed and dying and needs removing thats one thing, but is there any plans to save them? Or are we just going to destroy them?</p>
<p>Kaz Santa Cruz, CA</p>	<p>Tuesday May 8 #3 Flag Reply »</p> <p>Rumor has it that there was (is?) a communications tunnel that ran underneath, the mall. Someone actually suggested that it belonged to Lockheed? Can anyone confirm or deny that rumor?</p>

- Users are able to post comments in response to an article that was posted in a Topix page.
- Comments may be posted by both registered and unregistered users to encourage participation.

Participation in Wikipedia

- Notion of activity of users transforms as their participation increases [BFB05].
 - Editing what they know -> Building the Wikipedia
 - Using search to find articles and edit articles that have something missing -> Using tools to maintain the integrity of a set of articles.
 - Wikipedia is a collection of articles with random people adding information -> Participation makes one a member of the Wikipedia community.

Folksonomy

- A user-generated taxonomy used for the categorization and retrieval of Web content.
- Users use their own terms “tags” to categorize content.
- Applications: del.icio.us (Web pages) and Flickr (photos).

Advantages of Folksonomies

- Non-hierarchical
- Low barriers to entry

...when compared with formal taxonomies.

Disadvantages of Folksonomies

[GH06]

- **Polysemy:** tagging with words that have many related senses (window: hole in wall or the glass?)
- **Synonymy:** multiple words with close meanings
- **Basic level variation:** tag depends with familiarity and expertise (tag object as “programming” or “perl”?)

Collaborative Tagging

- Tag: a text label
- Resource: a Web object (ex. Web pages, images, media files).
- Resources are tagged with 0 or more tags by users.

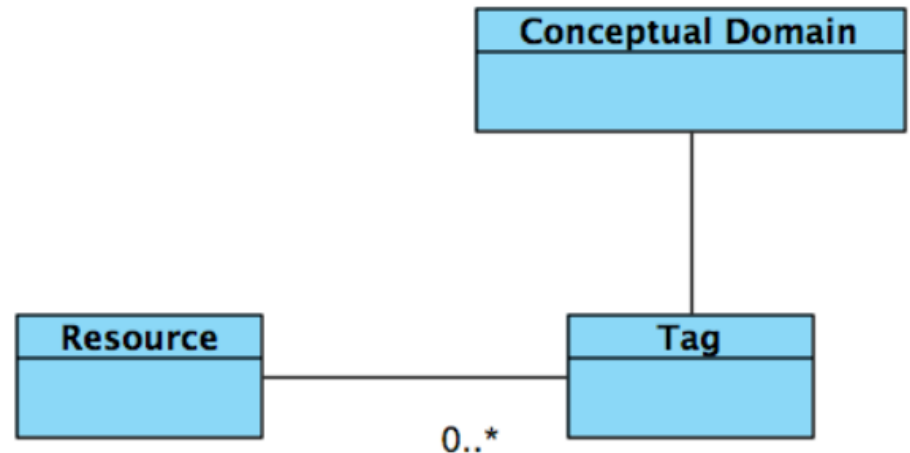


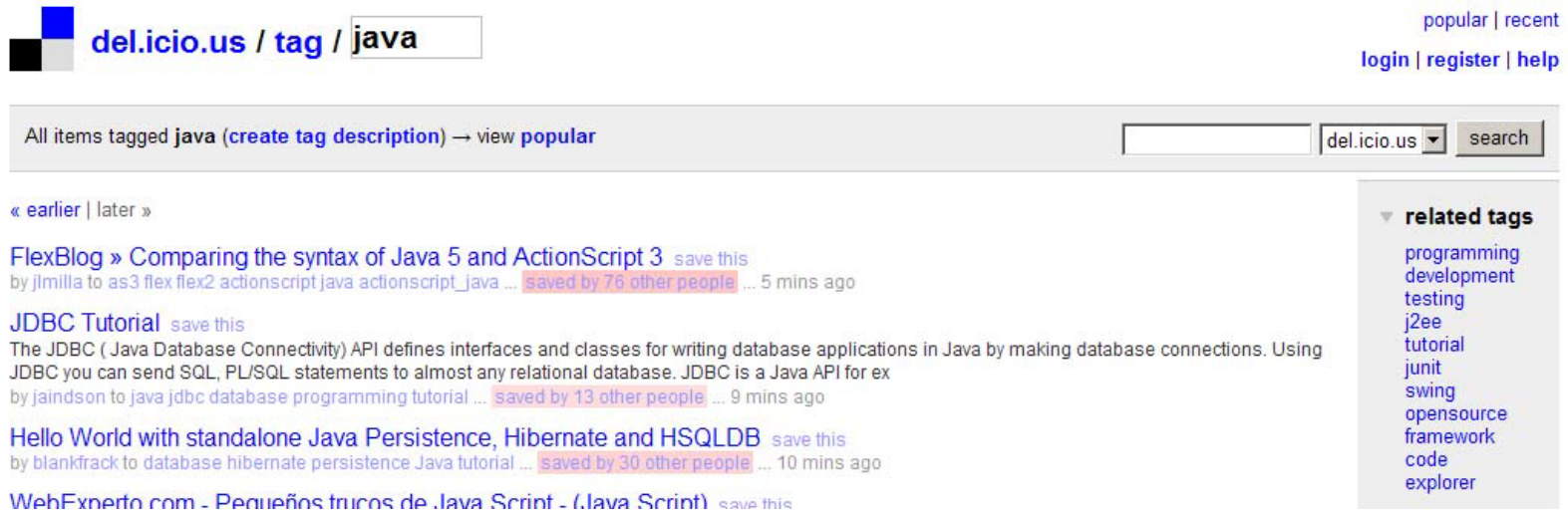
Image Source: [Nic07]

Characterizing Tagging Systems

[MNBD06]

- Tagging rights (Who can tag a resource? Creator? Anyone?)
- Tagging support (Blind tagging or suggested tags?)
- Aggregation (Duplicate tags from different users allowed?)
- Type of object (What's being tagged? Web pages, images, videos?)
- Source of object (Users? System? Anything?)
- Resource connectivity (Resources linked together independent of tags?)
- Social connectivity (Users in the system linked together?)

Collaborative Tagging (del.icio.us)



del.icio.us / tag / java

popular | recent
login | register | help

All items tagged java (create tag description) → view popular

« earlier | later »

[FlexBlog » Comparing the syntax of Java 5 and ActionScript 3](#) save this
by jlmilla to as3 flex flex2 actionscript java actionscript_java ... saved by 76 other people ... 5 mins ago

[JDBC Tutorial](#) save this
The JDBC (Java Database Connectivity) API defines interfaces and classes for writing database applications in Java by making database connections. Using JDBC you can send SQL, PL/SQL statements to almost any relational database. JDBC is a Java API for ex
by jainson to java jdbc database programming tutorial ... saved by 13 other people ... 9 mins ago

[Hello World with standalone Java Persistence, Hibernate and HSQLDB](#) save this
by blankfrack to database hibernate persistence Java tutorial ... saved by 30 other people ... 10 mins ago

[WebFxnerto.com - Pequeños trucos de .Java Script - \(.Java Script\)](#) save this

related tags
programming
development
testing
j2ee
tutorial
junit
swing
opensource
framework
code
explorer

- Social bookmarking.
- Tag resources to bookmark them for yourself.
- Bookmarks shared with other del.icio.us users.

Social Bookmarking in del.icio.us

[GH06]

- It is expected that the combination of the varying tag collections of users, individual preferences, and a large number of users would yield a chaotic pattern of tags.
- In fact, there are stable patterns: a nascent consensus forms after a relatively small number of bookmarks of a resource (fewer than 100 tags).
- Constitutes a “social proof.”
- Commonly used tags and more personal tags coexist.

Types of Tags in del.icio.us [GH06]

- Identifying topic of item (what it is about).
- Identifying what an item is.
- Identifying the owner of the item.
- Refining categories (tags that do not stand alone, like numbers).
- Characteristics (tags such as *funny*).
- Self reference (such as *mystuff*).
- Task organizing (such as *toread* or *jobsearch*).

Motivations for Tagging

- Social dimensions of tagging: Self and Social
- Functional dimensions of tagging: Organization and Communication

		<i>Function</i>	
		Organization	Communication
<i>Sociality</i>	Self	* Retrieval, Directory * Search	* Context for self * Memory
	Social	* Contribution, attention * Ad hoc photo pooling	* Content descriptors * Social Signaling

Image Source: [AN07]

Motivations for Tagging in del.icio.us

- Self: organize bookmarks using tags created by the user which can be retrieved from any computer.
- Social:
 - Attract attention with common tags or express opinions about a particular Web object [MNBD06].
 - Even without a social motivation, other users benefit when many people use del.icio.us to organize their own bookmarks.
- Succeeds in overcoming the “work vs. benefit disparity” [Gru94]: del.icio.us succeeds because it doesn’t require additional work from people who do not directly benefit from using the system.

Tag Spam

- Misleading tags generate to increase visibility of certain resources.
- Example: homemade videos on YouTube tagged with names of popular movies or TV shows.
- Develop a models of good users and targeted attacks, and develop a method of ranking documents matching a tag based on tagger's reliability [KEGH+07].

Feedback Systems

- Made popular by eBay user feedback and Amazon product reviews.
- University of Michigan study shows that feedback scores generated a collection of unrelated people matters [New07].
- **Digg:** users submit articles that make it to Digg's front page if they get enough votes.

Crowdhacking [New07]

- **Buddy System:** users organize into groups to vote up stories on Digg.
- **Geek Baiting:** companies publish geek-friendly articles that have nothing to do with their business to drive traffic to an ad-filled Website thru Digg.
- **Network for Hire:** recruit networks of Digg users willing to sell their vote.
- **Pump-and-Chump:** build up a high reputation on eBay selling inexpensive items and defraud customers on expensive items.

Online Communities



- Social Networks
- Blogs

Image Source:
<http://xkcd.com/c256.html>

Social Networks

- First studied by social scientists to understand patterns of relationships between people.
- Two types of online social networks [YD06]:
 - **Egocentric:** each user has personal social network perspective.
 - **Aggregate:** a single social network capturing relationships of an entire group.

Characteristics of Typical Social Networks

- Users create a profile of themselves.
- Add other users who are registered with the site as “friends.”
- Users try to collect as many “friends” as possible, so it’s common to receive “friend requests” from people they don’t know.
- Some mechanism to leave messages to friends or even other users.
- Ability to create groups based on interest.

Motivations for Participation in Social Networks

- Keep in touch with people when you don't have anything substantial to say to them.
- Meet new people (friends, dating, etc.).
- Express interests and personal information through a public (or semi-public) profile.

Privacy Issues with Social Networks

- People don't always realize the information they put online is public.
- Potential employers sometimes search social networks for information.
- How people want to represent themselves to their friends isn't always how they want to represent themselves in general.
- Possibility of stalking and harassment on social networks.
- Controversy over newsfeed on Facebook made many people aware of the privacy issues.

Communication Issues with Social Networks

- How will the way people build and maintain social relationships be affected?
- New technologies change the way communicate and also affect older methods of communication.
- Are forms of communication designed for commercial appeal as good as traditional communication?

Communication Issues with Social Networks

- Profiles designed to attract as many people as possible.
- Are people replaceable?
- Can we demonstrate commitment through social networks and not depend on friends to constantly amuse us?

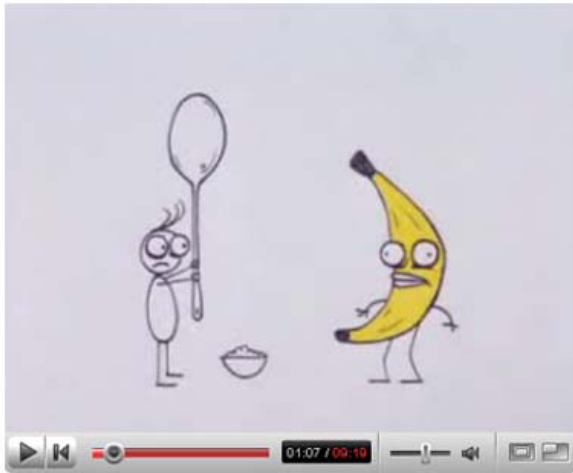


Web Logs (Blogs)

May 3, 2007

"My spoon is too big!"

I must be wrong in the head to like this so much. I couldn't stop laughing.



In particular the end apocalypse sequence (starting at 7:00 min) is amazing.

Update: *I'm not wrong in the head!* (not about this, anyway :)

A little digging for this post turned up the fact that the animator, [Don Hertzfeldt](#), was nominated for an academy award for this short (which is titled "[Rejected](#)"). Apparently it's received over 27 awards, and is the #3 most popular short of all time according to IMDB.

Posted on May 3, 2007 2:18 AM | [Permalink](#) | [Comments \(3\)](#) | [TrackBacks \(0\)](#)

Blog: www.skrenta.com

- Entries in chronological order.
- Commentary on news, politics, random topics, or online diary.
- Can be integrated with social networks.
- Allows readers to comment on a blog entry.

Blog Terms

- **Permalink:** URL pointing to a particular blog entry. Makes it easy for others to cite a specific entry on a blog.
- **Trackback:** Enables authors to keep track of who is linking to their entries.

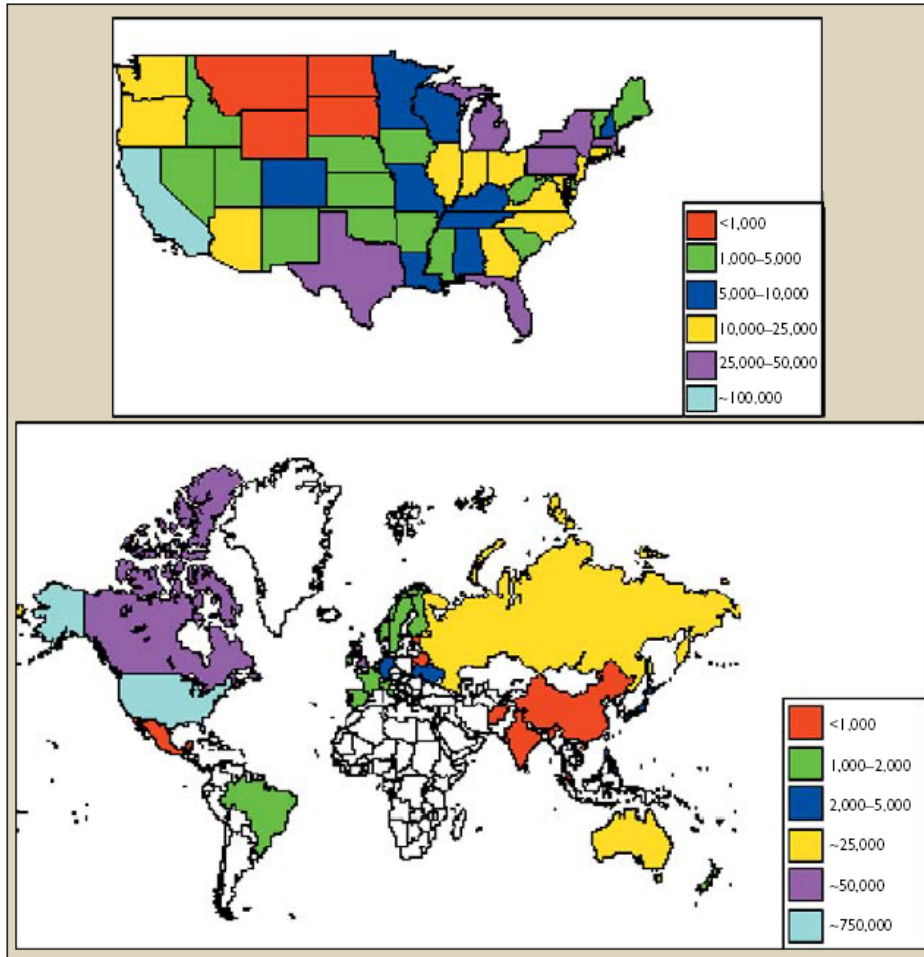
How Trackback Works

- User A posts a blog entry that links User B's blog entry.
- Server hosting User A's entry extracts link to User B's post and send notification to server hosting User B's entry using HTTP POST.
 - Information sent:
 - Linking site name
 - Linking post title
 - Linking post excerpt
 - Linking post URL
 - Linking post ID number

Motivations for Blogging [NSGS04]

- “Document my life.”
- Commentary (expression of opinions).
- Outlet for thoughts and feelings (ranting to strangers).
- Muse (thinking by writing, testing ideas by writing them for an audience).
- Community forum.

Geographic Distribution of Bloggers



- Distribution of bloggers on Livejournal in Feb. 2004.

Image source: [KNRT04]

Age Groups and Interests of Bloggers

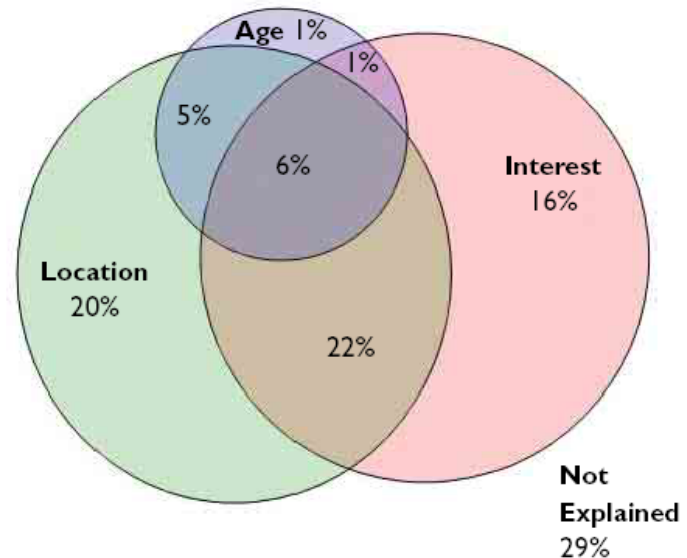
Age	%	Representative Interests
1-3	0.5	treats, catnip, daddy, mommy, purring, mice, playing, napping, scratching, milk
13-15	3.5	Web designing, Jeremy Sumpter, Chris Wilson, Emma Watson, TV, Tom Felton, FUSE, Adam Carson, Guyz, Pac Sun, mall, going online
16-18	25.2	198(6, 7, 8), class of 200(4, 5), Dream Street, drama club, band trips, 16, Brave New Girl, drum major, talking on the phone, high school, Junior Reserve Officers' Training Corps
19-21	32.8	198(3, 5), class of 2003, dorm life, frat parties, college life, my tattoo, pre-med
22-24	18.7	198(1, 2), Dumbledore's army, Midori sours, Long Island iced tea, Liquid Television, bar hopping, disco house, Sam Adams, fraternity, He-Man, She-Ra
25-27	8.4	1979, Catherine Wheel, dive bars, grad school, preacher, Garth Ennis, good beer, public radio
28-30	4.4	Hal Hartley, geocaching, Camarilla, Amtgard, Tivo, Concrete Blonde, motherhood, SQL, TRON
31-33	2.4	my kids, parenting, my daughter, my wife, Bloom County, Doctor Who, geocaching, the prisoner, good eats, herbalism
34-36	1.5	Cross Stitch, Thelema, Tivo, parenting, cubs, role-playing games, bicycling, shamanism, Burning Man
37-45	1.6	SCA, Babylon 5, pagan, gardening, Star Trek, Hogwarts, Macintosh, Kate Bush, Zen, tarot
46-57	0.5	science fiction, wine, walking, travel, cooking, politics, history, poetry, jazz, writing, reading, hiking
>57	0.2	death, cheese, photography, cats, poetry

Source: [KNRT04]

- Percentage of bloggers in different age groups on Livejournal.
- Representative interests of each age group.
- People create blogs for their children (1-3 age group).

Social Network of Bloggers on Livejournal

- Friends are clustered tightly.
- Clustering coefficient of 0.2 (20% of the time, two friends of the same blogger are friends themselves) [KNRT04].
- Friends are clustered together based on location, age, and sharing the same interests.

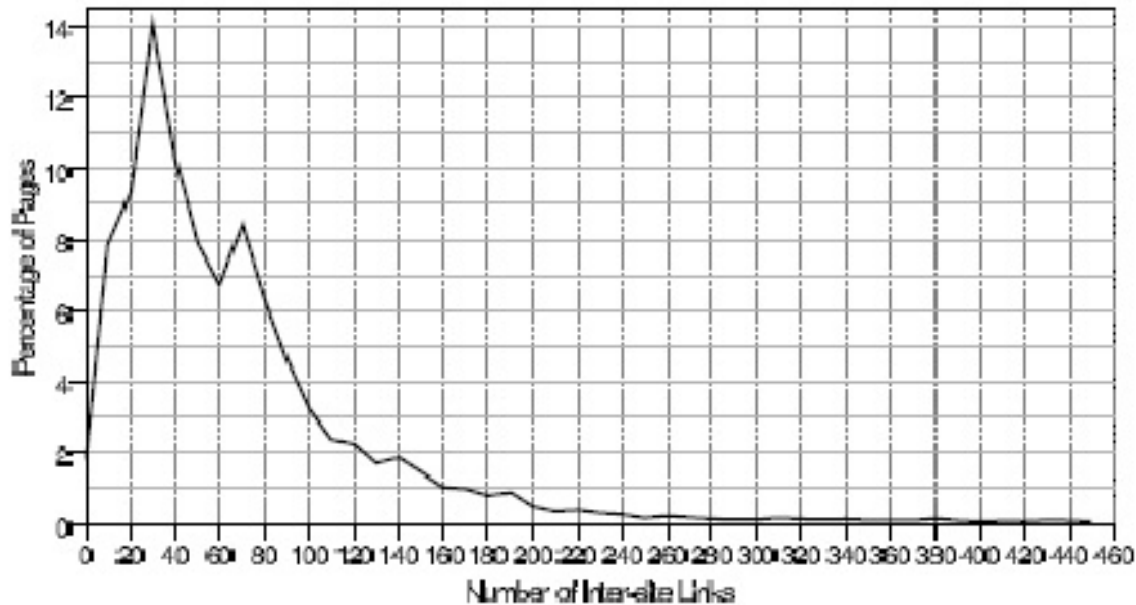


Source: [KNRT04]

Effect of Blogs on Web Crawlers

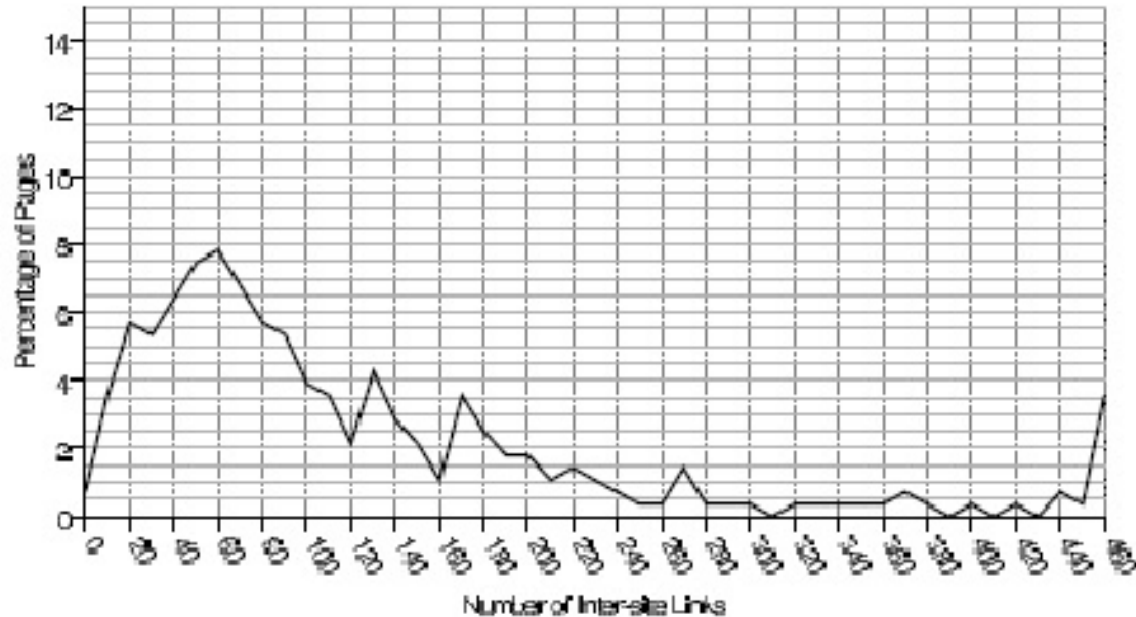
- Studies of the Web graph exhibit “small world” features and properties such as distribution of site sizes and distribution of hyperlinks on pages follow the Power Law [ENMP+04].
- Algorithms such as PageRank utilize these findings to rank Web pages [BP98].
- What effect do blogs have on the Web graph?

Effect of Blogs on Web Crawlers



- Inter-site links for homepages (blog and non-blog). Power Law is less pronounced than that for just non-blog homepages [ENMP+04].

Effect of Blogs on Web Crawlers



- Power Law no longer exists in inter-site link distribution on blogs homepages [ENMP+04].
- Will have an effect on Web crawlers and ranking algorithms if blogs rise as a proportion of all Web pages, as expected.

Looking Ahead...

- What will “Web 3.0” look like?
- **The Web as a Database:** the Web will be made up of structured XML-based data records that can be queried with languages such as SPARQL.
- **Artificial Intelligence:** AI will be able to reason about the Web in a human-like fashion.

- The Semantic Web?

The Semantic Web Vision

- Tim Berners-Lee *et al.* [BHL01].
- World Wide Web was developed as a medium of documents for people.
- The Semantic Web will add metadata targeted at computers.
- Computers will find the meaning of semantic data by following hyperlinks to definitions and rules.
- Computers use definitions and rules to reason about them logically.

The Semantic Web Vision

- Overcome the problems of search engines:
 - Too many irrelevant results returned.
 - Important results may not be returned.
 - Search results are highly sensitive to vocabulary.
 - Results are single Web pages, if the information we need is spread over several pages we must extract the partial information from each page and put it together.
- Better information retrieval achieved through meaning of Web content that is machine-accessible.

The Semantic Web Vision

- Knowledge Management on the Semantic Web:
 - Knowledge organized into conceptual spaces according to its meaning.
 - Keyword-based search will be replaced by query-answering.
 - Query-answering over several documents will be supported.

The Semantic Web Vision

- Scenario from [BHL01]:
 - Lucy and Pete need to schedule physical therapy appointments for their mother.
 - Lucy's Semantic Web agent retrieves information about her mom's prescribed treatment and looks up a list of providers that are within her mom's insurance plan and within a 20 mile radius from her home.
 - Lucy's Semantic Web agent tries to find a match between Pete and Lucy's schedules and available appointment times.
 - Pete doesn't like the first appointment returned, the hospital is across town and he would be driving during rush hour. His Semantic Web agent tries to schedule another time using stricter preferences about location and time.

Semantic Web Components

- **Metadata:** captures the meaning of Web content in a machine-readable format (RDF).
- **Ontologies:** formal model of knowledge defining the relationships between concepts and logical rules for reasoning about them.
- **Agents:** software that runs without constant human-supervision to accomplish goals provided by the user.

Ontology Engineering Issues [Yan07]

- **Construction:** building ontologies is difficult, time-consuming, and expensive, especially if they're designed to support the inferencing in the Semantic Web vision.
- **Mapping and Merging:** Semantic Web will be made up of a great number of small ontologies created in an anarchic manner. Need to map concepts from different ontologies.
- **Evaluation:** few widely used techniques to compare and evaluate ontologies.

Is Formality Harmful?

- Groupware research has revealed some problems with formal representations [SM99]:
 - Cognitive overhead with adding formalized information.
 - Tacit knowledge not represented.
 - Enforcing premature structure due to incomplete understanding.
 - Situated nature of knowledge means formalism appropriate for one task may not be suitable for a similar task.

How Groupware Fails

- Groupware may fail if [Gru94]:
 - Disparity in work and benefit.
 - Disruption of social processes.
 - Lack the “critical mass” of users required to make it useful due to high barriers to entry.
- Similar rules will apply to the formalism required for the Semantic Web.

Ontologies are Overrated?

- Ontologies don't work well when [Shi05]:
 - Large corpus that is ill-defined.
 - Unstable and unrestricted entities.
 - No formal categories.
 - Uncoordinated and amateur users.
 - Naïve catalogers.
 - No authority.
- Ontologies work well in certain domains, as in bio-ontologies [Yan07].

Metacrap?

- Some problems with metadata [Doc01]:
 - People lie (metadata with false information).
 - People are lazy (barriers to adding metadata).
 - People are stupid (lack of care in metadata creation).
 - More than one way to describe something.
- Observational metadata (metadata extracted from Web structure and content) is more reliable?

Web 2.0 and the Semantic Web

- Web 2.0 and the Semantic Web are commonly seen as competing visions.
- Can the two visions complement each other?

Web 2.0 and the Semantic Web

- The “two cultures” are compatible [AKTV07]:
 - Semantic Web needs to move away from an overly machine-centered approach and can draw insights from the community-centered approach of Web 2.0 in order to take hold.
 - Semantic Web technologies can enhance Web 2.0 applications by improving data interchange, data distribution, and facilitating creative reuse of data.

Web 2.0 and the Semantic Web

- Social bookmarking systems could benefit from a Semantic Web trust network ontology so that users can search for information based on a social notion of trust [YD06].
- Established ontologies may be enriched with methods for finding emerging ontologies in social bookmarking systems that reflect community dynamics and contain new phrases used by a community [Mik05].

Microformats: On the way to the Semantic Web?

- Encoding semi-structured information into XHTML.
- Using existing XHTML elements and a class-attribute system to make it easier to describe people, places, events, and other common types of semi-structured information [Kha06].
- Bloggers are intended users.
- Expected to be natively supported by Firefox version 3 and Internet Explorer 8.

Why Microformats?

- There is already an RDF-based format (FOAF) for describing people in social networks that can be used by bloggers.
- XHTML Friends Network is weaker than FOAF, but can be used by anyone with knowledge of HTML, while FOAF is still too complex with most blogging tools [Kha06].
- Semantics added by microformats still allows computer programs to extract meaning from Web pages marked up with microformats.

Microformat Example

```
<div class="vcalendar vevent">
  <span class="summary">Microformats: What the Hell Are
  They and Why Should I Care?</span>
  <p class="description">Ryan King will explain why
  microformats are important and how you can mark up specific
  kinds of content in ways that make it easier for the right people
  to find your stuff.</p>
  <abbr class="dtstart" title="20050926T050000-
  0700">September 25th, 2005, 5</abbr>-
  <abbr class="dtend" title="20050926T060000-
  0700">6PM</abbr>
  in the <span class="location">Balder Room</span>
</div>
```

Source: [KC06]

- An event in microformatted XHTML.
- Class names come from the vCalendar standard.

Thank You

- Questions?

References

- [AD07] Adler, B.T. and de Alfaro, L., "A Content-Driven Reputation System for the Wikipedia," in *Proc. WWW 2007*, Banff, Alberta, Canada, May 2007.
- [AKTV07] Ankolekar, A., Krotzsch, M., Tran, T., Vrandecic, D., "The Two Cultures: Mashing up Web 2.0 and the Semantic Web," in *Proc. WWW 2007*, Banff, Alberta, Canada, May 2007.
- [AN07] Ames, M. and Naaman, M. "Why We Tag: Motivations for Annotation in Mobile and Online Media," in *Proc. CHI 2007*, San Jose, CA, May 2007.
- [BE06] Grow, B. and Elgin, B., "Click Fraud: The Dark Side of Online Advertising," *BusinessWeek*, Oct. 2, 2006.
- [BFB05] Bryant, S.L., Forte, A., and Bruckman, A. "Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia," in *Proc. GROUP'05*, Sanibel Island, FL, Nov. 2005.

References

- [BHL01] Berners-Lee, T., Hendler, J., and Lassila, O., "The Semantic Web," *Scientific American*, 284(5), pp. 34-43, May 2001.
- [BP98] Brin, S. and Page, L., "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in *Proc. WWW 1998*, Brisbane, Australia, Apr. 1998.
- [Cla06] Clarke, G., "Berners-Lee calls for Web 2.0 calm," The Register.
http://www.theregister.co.uk/2006/08/30/web_20_berners_lee/
- [Doc01] Doctorow, C., "Metacrap: Putting the Torch to the Seven Straw Men of the Meta-Utopia," <http://www.well.com/~doctorow/metacrap.htm>
- [EGR91] Ellis, C.A., Gibbs, S.J., and Rein, G.L. "Groupware: Some Issues and Experiences," *Communications of the ACM*, 34(1), pp. 38-58, Jan. 1991.

References

- [ENMP+04] Evans, M.P., Newman, R., Millea, T.A., Putnam, T., and Walker, A., "The Effect of Web Logs and the Semantic Web on Autonomous Web Agents," in *Proc. ISCIS 2004*, Antalya, Turkey, Oct. 2004.
- [FT02] Fielding, R.T. and Taylor, R.N., "Principled Design of the Modern Web Architecture," *ACM Transactions on Internet Technology*, 2(2), pp. 115-150, May 2002.
- [GG05] Gyongyi, Z. and Garcia-Molina, H., "Spam: It's not Just for Inboxes Anymore," *IEEE Computer*, 38(10), pp. 28-34, Oct. 2005.
- [GH06] Golder, S.A. and Huberman, B.A., "Usage Patterns of Collaborative Tagging Systems," *Journal of Information Science*, 32(2), pp. 198-208, Apr. 2006.
- [Gra05] Graham, P. "Web 2.0," <http://www.paulgraham.com/web20.html>
-

References

- [Gru94] Grudin, J. "Groupware and Social Dynamics: Eight Challenges for Developers," *Communications of the ACM*, 37(1), pp.92-105, Jan. 1994.
- [KC06] Khare, R. and Celik, T., "Microformats: A Pragmatic Path to the Semantic Web," in *Proc. WWW 2006*, Edinburgh, Scotland, May 2006.
- [KEGH+07] Koutrika, G., Effendi, F.A., Gyongi, Z., Heymann, P., Garcia-Molina, H., "Combating Spam in Tagging Systems," in *Proc. AIRWeb'07*, Banff, Alberta, Canada, May 2007.
- [Kha06] Khare, R., "Microformats: The Next (Small) Thing on the Semantic Web?" *IEEE Internet Computing*, 10(1), pp. 68-75, Jan-Feb. 2006.
- [KNRT04] Kumar, R., Novak, J., Raghavan, P., Tomkins, A., "Structure and Evolution of Blogspace," *Communications of the ACM*, 47(12), pp. 35-39, Dec. 2004.
-

References

- [Mik05] Mika, P., "Ontologies are us: A Unified Model of Social Networks and Semantics," in *Proc. ISWC 2005*, Galway, Ireland, Nov. 2005.
- [MNBD06] Marlow, C., Naaman, M., Boyd, D., and Davis, M., "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read," in *Proc. HT'06*, Odense, Denmark, Aug. 2006.
- [New07] Newitz, A., "Herding the Mob," *Wired*, pp. 110-113, Mar. 2007.
- [Nic07] Nickull, D. "Web 20-20: Architecture Models and Patterns for the new Internet," http://www.web2expo.com/presentations/webex2007/nickull_web.ppt
- [NSGS04] Nardi, B.A., Schiano, D.J., Gumbrecht, M., and Swartz, L., "Why we Blog," *Communications of the ACM*, 47(12), pp. 41-46, Dec. 2004.

References

- [Ore05] O'Reilly, T. "What is Web 2.0? Design Patterns and Business Models for the Next Generation of Software," O'Reilly Network.
<http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>
- [Shi05] Shirky, C., "Ontology is Overrated: Categories, Links, and Tags,"
http://shirky.com/writings/ontology_overrated.html
- [SM99] Shipman III, F.M. and Marshall, C.C., "Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems," *Computer Supported Cooperative Work*, 8(4), pp. 333-352, Dec. 1999.
- [YD06] Yan, T.K. and Dommel, H.-P., "A Social Network Ontology for Semantic Web-Enabled Collaboration," in *Proc. SWWS'06*, Las Vegas, NV, Jun. 2006.
<http://ww1.ucmss.com/books/LFS/CSREA2006/SWW4884.pdf>
- [Yan07] Yan, T.K., "Practical Issues in Ontology Engineering," in *Proc. ICAI'07*, Las Vegas, NV, Jun 2007 (*to appear*).