Open Source and Search

Paul Pham

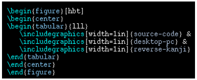
University of Washington
Computer Science and Engineering

2 May 2006

Overview

- Open Source Software (OSS)
- The Social Web
- Internet Search
- How are they related
- Wild speculations about the future of search

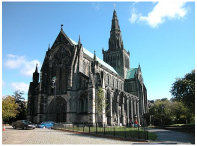
The Desktop Software Model



```
\begin{figure}[hbt]
\begin{center}
\begin{tabular}{lll}
   \includegraphics[width=1in]{source-code} &
   \includegraphics[width=1in]{desktop-pc} &
   \includegraphics[width=1in]{reverse-kanji}
\end{tabular}
\end{center}
\end{figure}
```

- The release cycle: Lather, rinse, repeat
  1. Developers write program in source code.
  2. Users buy/download and use compiled binaries.
  3. Users send feedback to developers, who incorporate into next release.
- Users and developers are separate.
- Everyone interacts with their own machine.

Closed Source Software - The Cathedral



- Source code is kept secret, often duplicated.
- Development is centralized and done by experts/pros.
- Problems are concealed, no guarantee of fixes.
- Software is a product, connected to the business.
- Success is determined by profit.

Open Source Software - The Bazaar

- Source code is freely available, effort is shared.
- Development is decentralized and done by anyone (experts and amateurs).
- Problems are discovered through independent review.
- Support is a service, freedom to fork.
- Success is determined by users and actual practice.

Social Concerns of Open Source

- Privacy - does it revealing personal information about me?
- Truthfulness - does it work the way it claims?
- Personalization - can I make it work the way I want? ("scratching an itch")
- Independence - am I free to switch software and still access my data?
- Cost - am I paying for what I want?

OSS can answer positively on all counts.

Open Source Success Stories

- BSD - Solaris, Mac OS X, Windows
- Apache - runs 54% of the world's websites (Netcraft)
- Mozilla - Firefox and Thunderbird.
- Linux - RedHat, SuSE, Novell, Debian, Ubuntu.
- Companies that invest in OSS: IBM, Oracle, Novell
- Companies that use OSS: Amazon, eBay, Yahoo!, Google

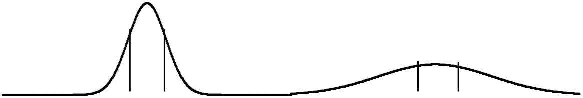
Disadvantages of Open Source

- Notoriously hard to use (UI is ugly).
- Not available for high-end, specialized applications (video editing, photo manipulation).
- Bad for entertainment content (movies and music).
- Games are different (plugins, mods, open source engines, etc.)
- So how can we become beautiful, rich, and famous by using OSS?

The Social Web (Web 2.0)

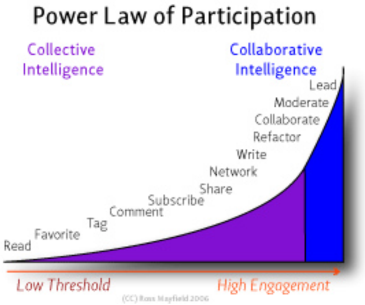
- Low-key user content - Craigslist
- Reputation mechanisms - eBay ratings
- Blogs - Blogger, LiveJournal
- Tagging (folksonomy) - Delicious, Flickr
- Wikis - Wikipedia, FAQs
- Podcasts
- Syndication (RSS)
- Social networking - MySpace, Friendster, Orkut
- Web Platform (GMail/Calendar vs. Outlook)

Chris Anderson's Long Tail



- Traditional markets are mediated geographically.
- Business target the mainstream (average under a normal curve).
- Dissipated communities have no voice.
- Internet user participation and automation make niche markets profitable.
- Success stories: Amazon reader lists and recommendations, Netflix

Ross Mayfield's Power Law of Participation



- Spectrum of "group" intelligence for Social Web.
- Also a long tail here to exploit.

Report Card of the Social Web

- Privacy - no (privacy policies)
- Truthfulness - no (privacy policies)
- Personalization - yes
- Independence - yes
- Cost - yes

Questions?

- Do we need more ideas from OSS?
- Is the Social Web useful for search?

Current Weaknesses of Search

- Dynamically generated websites (online databases).
- Multimedia files (video, sound, images)
- "Islands" of content with no external links.
- Personalized search and training.
- Search engine optimization (SEO) and spam.

Recent Search Trends

- User submission: Google SiteMap
- User content and tagging: Google Base (Craigslist clone)
- Provides web services through open APIs (maps, search, etc.)
- Personalized search (privacy concerns)
- Censorship issues (truthfulness concerns)

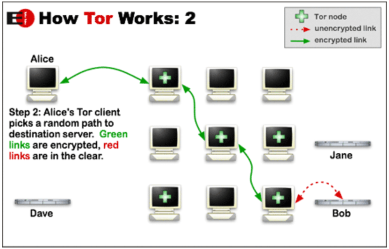
Open Source and Google

- Google benefits from open source
  - Runs on modified RedHat Enterprise Linux.
  - Google Web Server (GWS) is modified Apache.
  - Uses scripting languages like Perl and Python.
- Open Source benefits from Google.
  - Member of foundations for Apache/Java/Python/Mozilla.
  - Funds student “externships” on OSS projects (Summer of Code).
  - Releases tools on SourceForge.

Lessons Learned By Google

- The Web is a platform (open standards, commodity browser).
- Use and extend OSS, don't recreate it, and give back.
- Release early and often (and keep things in Beta forever).
- Let your users tinker, and listen to their feedback.
- Advertising is a better revenue model than support.
- Influence programmers when they're young.
- Leadership and a common vision keep a project/company together.

Anonymous Communication: Onion Routing and Tor

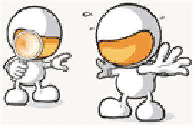


Step 2: Alice's Tor client picks a random path to destination server. Green links are encrypted, red links are in the clear.

- Founded by Roger Dingledine and Nick Mathewson (MIT)
- Like anonymized U.S. Postal Service.
- Protects against privacy attacks.
- Most famously motivated by Google search tracking.
- Solves privacy.

Open Source Search: Nutch

nutch



Lucene

- Founded by (our own) Mike Cafarella and Doug Cutting.
- Used publicly by Oregon State University and MozDex (dmoz).
- Makes it easy to conduct search research.
- Solves privacy and truthfulness.

Speculation on the Future of Search

- User crawling instead of link-crawling.
- Collaborative searching and tagging.
- Peer-to-peer, distributed, social crawling / searching.
- Privacy and truthfulness as defaults on new PCs.