Opportunities from Open Source Search

Wray Buntine
Helsinki Institute for Information Technology

September 21, 2005

Acknowledgements

- ALVIS project partners
- Ivana Podnar and P2P group at EPFL
- Ville Tuulos and Jukka Pertiö at CoSCo, HIIT
- attendees of Open Source Web IR Workshop

Disclaimer: we are not anti-Google, anti-Yahoo, etc.

Overview

- Why do we want open source search?
- What are the problems faced?
- Where is open source search working now?
- What do we have working so far?
- What strategies should we use?
- What opportunities are there?

Why do we want open source search?

- Building search engines is some of the best fun a computer scientist can have.
  - Its our moon walk!
  - General students and researchers cannot join the fun unless they work for the “big 4”.
- Search is our window to the world and we want options.
  - town-cryer → newspaper/magazine → internet directory → search
  - “Content bias” means we’re all becoming Americans!

Why do we want open source search? cont.

- Current ranking of results is either secret or paid for.
  - Small businesses can fail if their rankings drop.
- Niches and specialisations are under-served.
  - As for Linux versus Microsoft Windows.
  - Morphologically rich languages (Finnish, Turkish).
  - Special interest groups.
  - Information access to public services.

What are the problems faced?

- Its a changing world.
  - blogs, resource pages, mobile devices, hardware, move to IP
  - People's lives being tracked, "life is the killer app." (Henry Tirri, Nokia)
- Overcoming the *Google barrier*.
  - Small engines lack the global resources to analyse relationships for reputation/authority.
- *Information extraction* is not yet building the semantic web.
  - Automated tagging of content still in its infancy.

What are the problems faced? cont.

- The *Academic-OpenSourceProgrammer gap*.
  - Academics just want some results, programmers want good code

- Getting the right *user interface* (UI) for the next generation.

- *Economics* of operating a large scale search engine.
  - Scaling to both large collections and many queries requires big bucks.
  - “The game of giants” (Henry Tirri, Nokia)

- What *infrastructure* do we need to make open source search work?
  - We know what makes an existing search engine work, but what about mobile search?

Where is open source search working now?

- Want access to tagged meta-data.
  `http://creativecommons.org`
- Want *all* their web-pages made available.
  `http://oregonstate.edu/`
- Want specific services, *e.g.* geographical proximity.

- Cannot/Wont support a development effort or high-end intranet software.
  `http://www.archive.org/`
  `http://www.technorati.com/`

Doug Cutting’s

Nutch experience

What do we have working so far?

- Open source information retrieval
  Lucene, Terrier, Lemur, ...
- Information extraction and natural language programming
  starting to provide better open source
- P2P, social networks, swarm intelligence
  tools for harnessing people power
- P2P, Grid, Google File System
  tools for harnessing computing power
- Linguistic, topical and genre resources
  Wikipedia, DMOZ, ..., *e.g.,* topic discovery
- Crawlers
  WIRE, Heritrix, ...
- Digital Libraries and Internet Archive
  archiving the web

ALVIS: Peer to Peer Semantic-based Search Engine

- Use development path based on open source.
- Target specific user categories, don't compete with the majors.
- Engage information extraction and data mining community.
- Enable different user experiences with simple semantic capability.
- Empower area/language/subject-centric search initiatives with tools readily used.
- Apply P2P technology for query routing and results processing.

What strategies should we use?

- Semantic web for dummies. *e.g.*, semantic web ala `MusicBrainz.ORG` versus semi-automatically tagging content.
- Build more resources using people power. *e.g.*, the genre directory, social networks.
- Borrow computing resources (P2P, Grid).
- Build a search service economy, *i.e.*, a platform for agents.
- Build alternative Authority/Trust/Reputation mechanisms.
- Build other infrastructure.

What strategies should we use: Example

Incoming link and link/anchor text server.

NB. smaller search engines cannot currently get the broader internet/linking context for their web pages, thus results quality is inherently poorer.

- Provide tools to build different ranks: ComputingRank, MusicRank, ...
- Anchor text provision.
- Tools for link spam detection and reporting.
- Crawl seeding (using anchor text for topical detection).

What opportunities are there?

- Topical modeling integrated with search, *e.g.,* Wikipedia search, Exalead's Search News

- Genre based selection of content: FAQs, informative, etc., *e.g.,* Yahoo MindSet

- Efficient and effective topic specific search engines with custom support for their domain.

- A free market of authority and reputation schemes.