

# Open Source, Distributed and Peer-to-Peer IR

Wray Buntine  
National ICT Australia (NICTA)  
Helsinki Institute for Information Technology (HIIT)



## Part I

# Motivation and Open Source IR

# Outline

We cover some case studies and background context for open source IR and for P2P IR.

## 1 Motivation

- Walk-through of Some Systems
- Walk-through of Some Issues
- Summary of Issues

## 2 Open Source IR

- Basic Issues
- Some Standards
- Summary of Open Source IR

# Outline

- 1 Motivation
  - Walk-through of Some Systems
  - Walk-through of Some Issues
  - Summary of Issues
- 2 Open Source IR
  - Basic Issues
  - Some Standards
  - Summary of Open Source IR

# Technorati

- Built on a Lucene backend for search (well, as of 2006).
- Updates indexes every minute, using Lucene's logarithmic indexing.
- Search not always successful, try "[open source search](#)".
- Has added support for tags and "authority".
- *Question:* how does it compare with TREC Blogs track for performance?

## Wikia Search

- Vaporware announcements by Wikipedia umbrella company.
- Developing an [“open” search](#)<sup>1</sup> based on distributed, social, and semantic concepts.
- Have [acquired Grub](#), the P2P crawler, from Looksmart.
- Current Wikipedia search is Lucene.
- For fact style search, where the queries are not obscure, a limited subset of the web would suffice.

**Opinion:** something to watch out for!

---

<sup>1</sup>[http://search.wikia.com/wiki/Search\\_Wikia](http://search.wikia.com/wiki/Search_Wikia)

# PatentLens

- Started out as a [\*patent search engine\*](#)<sup>2</sup> for Bioinformatics to support patent packaging.
- Software is open source, but largely developed in-house at [\*Cambia\*](#).
- Many specific facilities to support patents (organisation/company matching, cross-nation support, gene name search ...).
- The patent landscape is changing, see [\*Open Invention Network\*](#)<sup>3</sup>.

---

<sup>2</sup><http://www.patentlens.net/>

<sup>3</sup><http://www.openinventionnetwork.com/>

## Creative Commons Search

- [Meta search engine](#)<sup>4</sup> from Creative Commons (and multimedia search partners).
- Finds material with Creative Commons licenses (there are many varieties).
- No search engine, but does search for “open content”.

---

<sup>4</sup><http://search.creativecommons.org/>

# Creative Commons, cont.

The screenshot shows the Creative Commons Search interface. At the top, the search bar contains the text 'pyramids' and a 'go' button. To the right of the search bar are two checkboxes: one checked for 'Search for works I can use for commercial purposes.' and one unchecked for 'Search for works I can modify, adapt, or build upon.'

Below the search bar, there are navigation links: 'What is this?', 'Content Directories', and 'Remove Frame'. A row of partner logos includes Google, YAHOO!, flickr, blip.tv, OWL music search, and spinXpress. A 'SUPPORT CC 2007' logo is also present.

The search results are displayed in a grid. The first result is for 'Egypt: Sphinx & pyramids from Matson Collection, ca. 1934-39 (LOC)'. It features a photograph of the Sphinx and pyramids. The text for this result includes: 'Egypt: Sphinx & pyramids from Matson Collection, ca. 1934-39 (LOC)', 'Uploaded on 20 April 2007', 'By pingnews.com', 'See more photos, or visit pingnews.com's profile.', and a list of tags: 'history, archaeology, public, sphinx ...'. Below this result is another image of a pyramid with the caption 'Sphinx and pyramid, Giza, Egypt' and 'Uploaded on 30 November 2006'.

On the left side of the search results, there is a sidebar with a 'Y! SEARCH' box containing 'pyramids egypt' and a 'Search the Web' button. Below the search box, it suggests 'Also try: [cairo](#), [giza](#), [sphinx](#), [camel](#) or [desert](#)'.

At the bottom of the page, there is a 'Creative Commons | Contact' link.

## Social Bookmarks: Del.icio.us

- Uses tagging to provide higher-weighted keywords.
- Uses social bookmarks to get popularity/“authority” for pages.
- Purchased by Yahoo in 2005.

**Opinion:** their search returns best pages on fairly general topic areas (*i.e.*, but not “home page” or “lost page” search).

# Del.icio.us, cont.

The screenshot shows a Del.icio.us search results page. At the top left is the Del.icio.us logo and the text 'del.icio.us / search'. On the right, there are links for 'popular | recent', 'login | register | help'. Below the header is a search bar containing 'information retrieval' and a dropdown menu set to 'del.icio.us'. The main content area lists search results for 'Information retrieval', showing items 1 to 10 of 2227. Each result includes a title, a brief description, and a 'saved by' count. On the right side, there is a 'sponsored results' section with three advertisements: 'Intriever - Surf Faster', 'Information Retrieval', and 'Information Storage Retrieval System'. At the bottom of the page, there are two footer elements: 'Buntine' on the left and 'OS, D\* & P2P IR' on the right.

del.icio.us / search popular | recent  
login | register | help

Search results for **Information retrieval**

**Related tags:** [search](#) [ir](#) [research](#) [information retrieval](#) [reference web software](#) [informationretrieval](#) [google](#)  
showing 1 - 10 of 2227  
« previous | next »

**Introduction to Information Retrieval** [save this](#)  
to [ir book](#) [retrieval](#) [information search](#) ... [saved by 380 people](#)

**The Lemur Toolkit for Language Modeling and Information Retrieval** [save this](#)  
to [ir nlp](#) [linguistics](#) [information search](#) ... [saved by 126 people](#)

<http://www.dcs.gla.ac.uk/Keith/Preface.html> [save this](#)  
to [ir book](#) [informationretrieval](#) [information\\_retrieval](#) [search](#) ... [saved by 202 people](#)

**Information Research: an international electronic journal. Information science, Information management, Information systems, Information retrieval, Digital libraries, Information seeking behaviour, Information seeking behavior, World Wide Web, WWW** [save this](#)  
to [research](#) [information journal](#) [library journals](#) ... [saved by 140 people](#)

**Information Retrieval Resources** [save this](#)  
to [ir](#) [information](#) [information-retrieval](#) [research](#) [retrieval](#) ... [saved by 75 people](#)

**Suchmaschinen-Buch: Inhaltsverzeichnis aus: Dr. Dirk Lewandowski: Web Information Retrieval** [save this](#)  
to [web](#) [informationswissenschaft](#) [suchmaschine](#) [searchengine](#) [recherche](#) ... [saved by 86 people](#)

**Terrier Information Retrieval Platform** [save this](#)  
to [java](#) [search](#) [ir](#) [software](#) [opensource](#) ... [saved by 71 people](#)

**Information retrieval - Wikipedia, the free encyclopedia** [save this](#)  
to [search](#) [ir](#) [information-retrieval](#) [reference](#) [wikipedia](#) ... [saved by 75 people](#)

sponsored results

**Intriever - Surf Faster**  
Retrieve **Information** from Web sites faster. Get a free trial now.  
[www.intriever.com](http://www.intriever.com)

**Information Retrieval**  
Retrieve files and folders accidentally lost on MS Windows. Try FREE.  
[www.handyrecovery.com](http://www.handyrecovery.com)

**Information Storage Retrieval System**  
Buy Anything at SHOP.COM. Shop OneCart(TM) Trusted Merchants.  
[www.SHOP.com](http://www.SHOP.com)

**Searchblox Content Search Software**  
Add a search engine to your Web site, Intranet or Portal in minutes. Ajax based Admin console. Download free edition now.  
[www.searchblox.com](http://www.searchblox.com)

Buntine OS, D\* & P2P IR

# Internet Archive

- Founded in 1996 as an internet library. Some countries have related national efforts.
- Open source shop as a non-profit with institutional and rich individual donations.
- Crawl data comes from [Alexa](#).
- Uses Heritrix an open source crawler and Nutch/Lucene for search.
- Indexes past web as well as digital media such as internet radio.

## World Wide WebLibrary

- [WorldCat](#)<sup>5</sup> is the largest union catalogue of library holdings, also available online.
- Maintained by [Online Computer Library Center](#)<sup>6</sup>, a nonprofit organization founded **1967**.
- Catalogue holdings exported to major search engines.  
*i.e.* the catalogue content is open, not the system.
- Component of a larger system [FirstSearch](#) which has full digital rights management (DRM).
- Runs on Oracle database with a [Z39.50](#) interface, with semi-structured and typed content.
- For a better digital library, see [the European Library](#).

---

<sup>5</sup><http://www.worldcat.org/>

<sup>6</sup><http://www.oclc.org/>

# WorldCat, cont.

The screenshot shows the WorldCat search interface. At the top, there is a search bar with the text 'information retrieval' and buttons for 'Search' and 'Advanced Search'. Below the search bar, the results are sorted by 'Relevance'. The search results list five items, each with a checkbox, a title, author, language, type, and publisher information.

**WorldCat** (Beta)

Home Search You are not signed in ([Sign In to WorldCat](#) or [Register](#))

Search for items:  [Search](#) [Advanced Search](#)

Search results for 'Information retrieval' Sort by:

Results 1-10 of about 72,151 (.18 seconds) << First < Prev 1 2 3 Next >>

[Select All](#) [Clear All](#) Save to:  [Save](#)

1. [Advances in information retrieval recent research from the Center for Intelligent Information Retrieval](#)  
by W Bruce Croft; NetLibrary, Inc.; Center for Intelligent Information Retrieval.  
Language: English Type: Internet Resource Computer File  
Publisher: New York : Kluwer Academic, ©2002.

2. [Find it fast : how to uncover expert information on any subject](#)  
by Robert I Berkman  
Language: English Type: Book  
Publisher: New York : HarperPerennial, ©1997.

3. [Student guide to research in the digital age : how to locate and evaluate information sources](#)  
by Leslie F Stebbins  
Language: English Type: Book Internet Resource  
Publisher: Westport, Conn. : Libraries Unlimited, 2006.

4. [Bioinformatics a practical guide to the analysis of genes and proteins](#)  
by Andreas D Baxeavanis; B F Francis Ouellette; NetLibrary, Inc.  
Language: English Type: Internet Resource Computer File  
Publisher: New York : John Wiley, ©1998.

5. [Information architecture for the World Wide Web](#)  
bv Louis Rosenfeld; Peter Morville.

**Refine Your Search**

**Author**  
[United States](#) (2192)  
[West Publishing Comp...](#) (528)  
[International Busine...](#) (135)  
[American Chemical So...](#) (104)  
[Inc Mead Data Centra...](#) (93)  
[Show more...](#)

**Content**  
[Library Science, Gen...](#) (12771)  
[Computer Science](#) (4900)  
[Law](#) (2924)  
[Business & Economics](#) (2769)  
[Engineering & Techno...](#) (1916)  
[Show more...](#)

**Format**  
[Book](#) (41929)

# Outline

- 1 Motivation
  - Walk-through of Some Systems
  - **Walk-through of Some Issues**
  - Summary of Issues
  
- 2 Open Source IR
  - Basic Issues
  - Some Standards
  - Summary of Open Source IR

# The Internet Society and Search

*(need I say this!)*

- Primary school students are given internet search tasks as assignments.
- Internet news and blogs overtaking paper, but the business models are unclear. Search and topic tracking common.
- E-government and business and consumer e-services booming, and search a necessary complement.
- Search and multimedia now a significant form of entertainment.  
*e.g.* 8 year-old boy and keywords “dinosaur”, “meteor”.

# Advertising and Search

*(need I say this!)*

- Advertising on specialist websites, on particular keyword searches, or on your email based on its content, is well focussed.
- Targeted advertising through the web, for instance Google AdSense, is considered the best value for money for advertising.

# Information Warfare

Definition: "the use and management of information in pursuit of a competitive advantage over an opponent."

*e.g.* where the opponent is the consumer, voter, etc.

- [WorldPublicOpinion.ORG survey](#) (Oct. 2003):
  - 80% of the watchers of FOX news had one or more major misconceptions over Iraq war,
  - compared with only 23% for PBS/NPR.
- [Pew Research Center on "news"](#) (Aug. 2007):
  - "more than half of Americans say US news organizations are politically biased, inaccurate, and don't care ...,"
  - "poll respondents who use the Internet as their main source of news – roughly one quarter of all Americans – were even harsher with their criticism."

# Wikiality

## Information Warfare, cont.

- Wikiality: the reality that exists if you make something up and enough people agree with you (Stephen Colbert, 2006).
- “Interesting” contributors to Wikipedia, and some IP addresses from government sites even banned from Wikipedia for vandalising content.
- On any controversial issue, impartiality is relative. Who makes the editorial decisions?

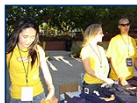
## Information Warfare, cont.

- These issues are reported w.r.t. current affairs and politics, but what about consumer products, health, business and economics?
  - The Internet-savvy populace particularly is more distrusting.
- ⇒ Search is now one tool for use in untangling information.  
*e.g.* Using search to research side-effects of a new drug you have been given.
- ⇒ Complementary growth industry is “opinion tracking” for corporations.  
*e.g.* Major forestry industry player searches/tracks “green” sites for potential environmental scandals.

# The Google Dance

## GOOGLE DANCE 2005

Action, Antics, & Atmosphere - [Demos](#) - [The Garden](#) - [Club G](#)



# The (Other) Google Dance

Gaming the big search engines, and trusting them?

- *Search engine optimization*: SEO, getting clients website's higher up in results.
- There are “acceptable” and questionable methods:  
*e.g.* link farms, fake websites, flogs.
- PageRank™ is only 1 facet of 200 used by Google, not as important now. In 1998 it was a good proxy for “popularity.”
- As Google changes ranking, so do SEOs, and everybody dances.

# Everything is Relative



If everyone thought the same,  
nothing would ever change.

[yourpointofview.com](http://yourpointofview.com)

HSBC   
The world's local bank

# Apriori/Static Ranking

## Making it Relative

Static Ranking: the query independent part of a (usually additive) ranking score.

- One-size-fits-all proprietary ranking not ideal. Others:
  - authority and trust (what are these?)
  - popularity (a proxy is click-throughs from search results),
  - genre and topic,
  - country (now addressed by the major engines).
- Static ranking allows bounding when forming results, huge savings in computation, and thus essential on very large engines.
- Personalisation feasible via modifiable static ranks.
- Social bookmarking an alternative to assess “popularity.”
  - Like the ODP, *an ideal candidate for open systems.*

## Other Issues

- Fear of a duopoly in the market, or at least dominance by a few players.
- Personalisation,  
*e.g.* user-centric data stores.
- Digital rights management,  
*e.g.* allowing players like BBC and libraries to contribute more content.
- Next generation search and Web 2.0,  
*e.g.* tagging, Flickr, AJAX, [mashups](#), *etc.*
- Open APIs.
- Desktop search, *e.g.*, [Xesam](#) (eXtensible Search And Metadata specification).

# Outline

- 1 Motivation
  - Walk-through of Some Systems
  - Walk-through of Some Issues
  - Summary of Issues
- 2 Open Source IR
  - Basic Issues
  - Some Standards
  - Summary of Open Source IR

## Summary of Issues:

- Different strokes for different folks:
  - genres,
  - ranking metrics, and
  - search modes/interfaces.
- Search is a vital part of the information landscape, warfare and all.
- Transparency and authority in ranking desirable for some.
- Open source IR the basis of many commercial ventures, and a key enabling factor in them (like other OS systems).
- “Open content” a related theme.
- Alternative search a hot topic for Silicon Valley.

# Open Source IR: Outline

Some general aspects of open source IR.

- 1 Motivation
  - Walk-through of Some Systems
  - Walk-through of Some Issues
  - Summary of Issues
- 2 Open Source IR
  - Basic Issues
  - Some Standards
  - Summary of Open Source IR

# Caveats

- Not going into specifics of particular systems.
- Not covering crawling, document analysis, or digital libraries, but these have very healthy open source and distributed communities.

# Outline

- 1 Motivation
  - Walk-through of Some Systems
  - Walk-through of Some Issues
  - Summary of Issues
- 2 Open Source IR
  - Basic Issues
  - Some Standards
  - Summary of Open Source IR

# Academia vs. Industry

## Academia:

- Developed in universities as part of a group project.
- Can migrate into industry as open/closed source.
- High on advanced features.
- Used in TREC competitions.  
*e.g.* systems at *TREC 2006 Terabyte Track*.

## Industry:

- High on stability and good coverage of features.
- Developmental sustainability.  
*e.g.* easily customised, well documented, good web support, active news groups, ...
- Typified by the Apache projects ([Lucene](#), Nutch, Solr), [Xapian](#), but many others!

## Licensing Issues

**Copyleft:** (or viral) whether modifications by others automatically become part of the main code, e.g. GPL.

**Patents:** when software is contributed, whether or not the contributor's relevant patents can be claimed against subsequent users, e.g. Apache says patent rights must be waived with contribution.

**Reuse:** whether the code can be reused commercially and a fee charged for use. "Free for use by nonprofit or research" is common, but not "open source".

**Proliferation:** many licenses exist and some organisations resent the legal barriers implicit here. Managed by [Open Source Initiative](#).

## Industry Acceptance

- More and more business development and commercial departments and organisations accepting open source software distribution and coding participation.

*e.g.* took a fight, but [Terrier](#) could be released under MPL.

- Non-copyleft licenses acceptable to industry. Some companies ban internal development on copyleft-licensed software.
- Heavy users of open source software also tend to be key developers.
- Like other OS software, open source IR and search is an important *infrastructure* component for industry.

⇒ Open source IR and search creates new jobs and industries.

# Academic Systems

Selected from *TREC 2006 Terabyte Track*. See Beigbeder *et al.*<sup>7</sup>

[Indri](#): U.Mass+CMU C++ system with language models on BSD license, part of Lemur.

[Lucene](#): Java-based industrial system sometimes used in academia due to its popularity.

[MG4J](#): “Managing Gigabytes for Java” system from U.Milano under LGPL.

[Terrier](#): feature-laden Java-based system from U.Glasgow on MPL.

[Wumpus](#): scalable desktop-oriented system from U.Waterloo on GPL.

[Zettair](#): simple, fast C-based system from RMIT University on BSD style license.

---

<sup>7</sup> “Open Source Search and Research”, Beigbeder, Buntine and Yee, IWRIDL, 2006.

## Popular Features for Industry

- Generally, systems need to offer something over and above use of Google, Yahoo, Scirus, *etc.*
- Fielded search (*i.e.* primitive support for XML-like structure) and types (*e.g.* dates, numbers, RegExp searchable) important.
- Incremental database support (*e.g.* for news).
- Scalable, *e.g.*, use of [Hadoop](http://hadoop.apache.org/)<sup>8</sup>.
- Completely customisable and easily scriptable ranking.

---

<sup>8</sup><http://lucene.apache.org/hadoop/>

## Workshops

OSWIR at WI-IAT 2005: organisers Michel Biegbeder and Wai Gen Yee; well-run and enthusiastic workshop.

OSIR at SIGIR 2006: organisers Biegbeder, Buntine and Yee; good cross-section of OS IR community and great discussions; workshop report in *SIGIR Forum*.

The future:

- TREC and INEX support many of the operational needs of the community.
- Apachecon *etc.* supports the industry community on Apache IR tools.
- Cooperation needed amongst academic groups to support interoperability?

# Outline

- 1 Motivation
  - Walk-through of Some Systems
  - Walk-through of Some Issues
  - Summary of Issues
- 2 Open Source IR
  - Basic Issues
  - **Some Standards**
  - Summary of Open Source IR

## SRU and CQL

*Search/Retrieve via URL*<sup>9</sup> (SRU).

*Common Query Language*<sup>10</sup> (CQL).

**Target:** search and information retrieval over the web, intended as a user-friendly replacement for the older Z39.50 protocol.

**Advantages:** Builds on the experience of the Z39.50 community from digital libraries.

**Barriers:** Information retrieval and digital libraries are not strongly overlapping communities, where, for instance, evaluation of search results and the nature of content is quite different.

---

<sup>9</sup><http://www.loc.gov/sru>

<sup>10</sup><http://www.loc.gov/cql>

# OAI-PMH

## *Open Archives Initiative Protocol for Metadata Harvesting*<sup>11</sup> (OAI-PMH)

**Target:** Publication of meta-data about digital (and non-digital) resources. Intended as a means to support distributed crawl.

**Advantages:** Bulk access to sets of metadata. Is used by search engines as a means of accessing some digital libraries.

**Barriers:** Not considered successful within digital libraries community itself.

---

<sup>11</sup><http://www.openarchives.org/OAI/openarchivesprotocol.html>

# OpenSearch

## OpenSearch<sup>TM</sup><sup>12</sup>

**Target:** Smaller search engines, to allow syndication on A9.com.

**Advantages:** Allows syndication of content to meta search engines. Uses RSS to return results.

**Barriers:** Lacks a more general strategy for results aggregation, and other aspects are too simple for more general search use.

---

<sup>12</sup><http://www.opensearch.org>

# Outline

- 1 Motivation
  - Walk-through of Some Systems
  - Walk-through of Some Issues
  - Summary of Issues
- 2 Open Source IR
  - Basic Issues
  - Some Standards
  - Summary of Open Source IR

## Summary

- Open source IR tools are the ideal tool for academic research, and some of these we expect to migrate to industry in the future.
- Commercial open source IR tools are widely used and are vital infrastructure tools.
- New handling of genres, query modes, personalisation, *etc.*, are needed.
- A common meme: How do we federate lots of smaller, topic specific search engines, for instance with P2P?  
*e.g.* For this, common query protocols and web interfaces are needed at least, as in [ALVIS](http://www.alvis.info)<sup>13</sup>.
- Another, related meme is “give the Web back to the people.”

---

<sup>13</sup><http://www.alvis.info>

## Part II

# Distributed and Peer-to-peer IR

A grand challenge problem for Computing Science.

# Distributed IR: Outline

General overview of distributed IR in its many forms. A necessary prerequisite for P2P.

## 3 Distributed IR

- Distributed Computing Basics
- Distributed Search Basics
- Literature
- Summary of Distributed IR

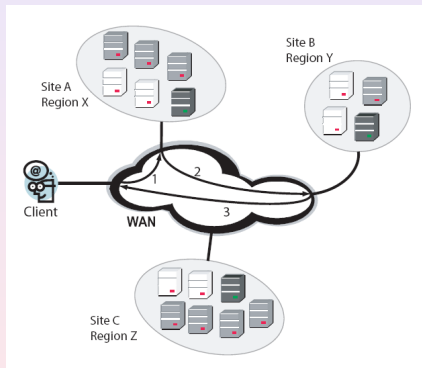
## 4 Peer-to-peer IR

- Basics of P2PIR
- Useful Papers
- Summary and Assessment

# Outline

- 3 Distributed IR
  - Distributed Computing Basics
  - Distributed Search Basics
  - Literature
  - Summary of Distributed IR
- 4 Peer-to-peer IR
  - Basics of P2PIR
  - Useful Papers
  - Summary and Assessment

## Distributed Computing, Example



**Query processor** : matches documents to the received queries



**Coordinator** : receives queries and routes them to appropriate sites



**Cache** : stores results from previous queries

A large, distributed search engine, from Baeza-Yates *et al.* 2007.

# Distributed Computing

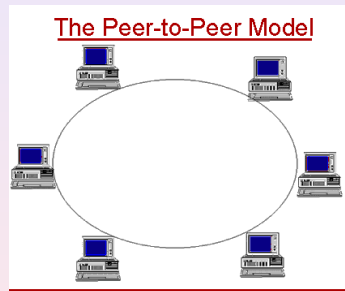
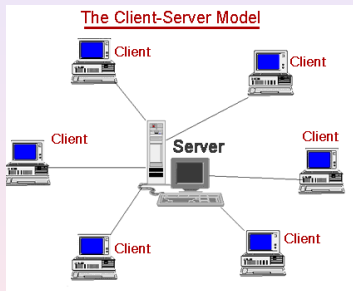
Distributed computing has multiple computers connected by a network.

**Clustered:** Distributed, but with group of tightly coupled computers that work together closely (almost like a single computer) and (oftentimes) connected by high-speed network.

**Peer-to-peer:** Distributed, but with *ad hoc* connections and nodes, usually does not have the notion of clients or servers but instead “peers”.

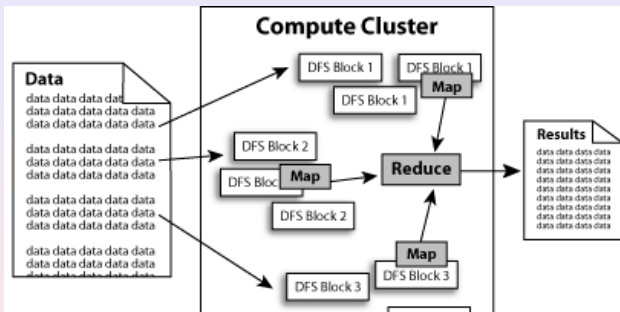
**Geographically distributed:** Sets of nodes (e.g. clusters) isolated by lower-bandwidth connections.

# Client Server vs. Peer to Peer



Cartoon schematic from <http://www.ibiblio.org>.

# The Map-Reduce Paradigm for Clustered Computing



From [Hadoop website](#).

- Framework for simplified applications on large clusters of commodity hardware.
- Application is divided into many small fragments of work, each assigned to any node.

## The Map-Reduce Paradigm for Clustered Computing, cont.

- Map operates on fragments of data to produce (key,value) pairs, Reduce sorts and summarises these to produce results.
- Built-in failure handling and recovery, as when 1,000's nodes, some may fail during a run.
- First published in “MapReduce: Simplified Data Processing on Large Clusters” by Dean and Ghemawat, 2004.
- [Hadoop](http://lucene.apache.org/hadoop/about.html)<sup>14</sup> is an open source implementation in Java started by Doug Cutting, founder of Lucene.

---

<sup>14</sup><http://lucene.apache.org/hadoop/about.html>

## Impetus for Distribution

Each form has characteristic cost-benefit issues contributing to its adoption.

**Clustered:** cost effective high-performance when a single machine is too small, and super computing too expensive.

**Peer-to-peer:** where much of the resources are contributed by the users themselves, and no single points of failure exist.

**Geographically distributed:** provide global/local coverage and/or ability to handle a full cluster failure.

## Network Practicalities

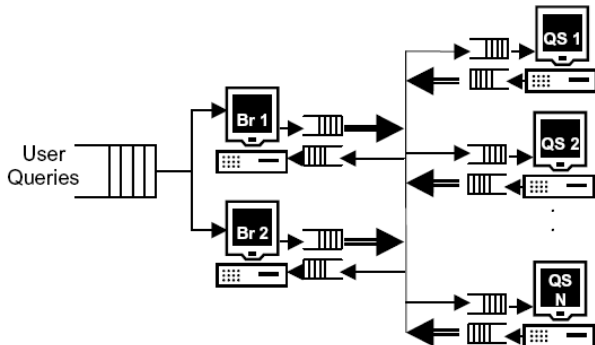


Fig. 1. Switched LAN model for a distributed IR system (Br: broker, QS: query server).

Characteristics of your network and systems is critical, from Cacheda *et al.* 2007.

# Outline

- 3 Distributed IR
  - Distributed Computing Basics
  - Distributed Search Basics
  - Literature
  - Summary of Distributed IR
  
- 4 Peer-to-peer IR
  - Basics of P2PIR
  - Useful Papers
  - Summary and Assessment

## Factors in Distributed IR

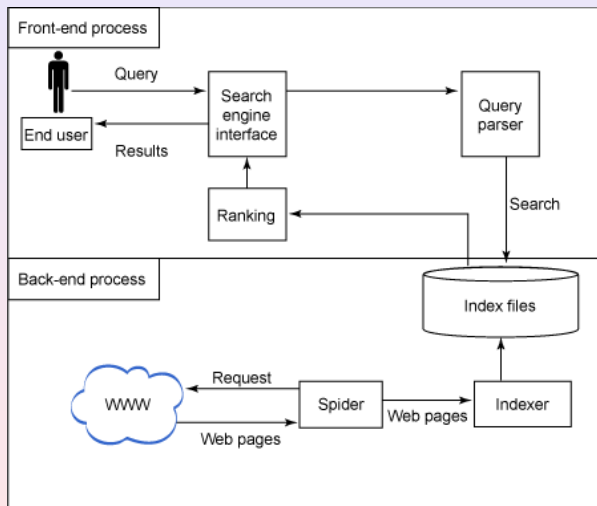
- Mode of distributed computing: clustered, P2P, geographically distributed, ...
- *Co-operating* versus *competing* nodes.
  - e.g.* cannot trust the scores nodes provide.
  - e.g.* rogue crawler injects spam into the collection.
- *Homogeneous* versus *heterogeneous* nodes.
  - e.g.* share software and standardised scoring.
- *Centralised* versus *decentralised* versus *no* control.
  - e.g.* can/cannot control key partitioning of computing elements.
- Nature of content: independent collections, topically factored collections, ...

## Historical Factions

- Federation of search in disparate *digital libraries*.
- Success of *P2P* in video and audio, and the idea of grass-roots computing “suggested” adoption in IR.
- *Very large scale search engines* needed clusters to serve large indexes, and geographical distributed to support international/regional access and fail-safing.
- *Meta-search engines*, a feasible start-up model for novel interfaces.
- The *Grid* as a computational platform.

**NB.** lots of interaction too.

# Review: A Search Engine



From  
[www.ibm.com](http://www.ibm.com)'s  
Lucene pages.

## Stages of Search Processing

**Crawling:** or gathering documents, lots of dirty details and source formats.

**Pre-processing:** any information extraction or HTML cleaning, *etc.*, sometimes a no-op.

**Indexing:** forming the indexes from processed documents, can be batch or incremental; dependent on query method.

**Querying:** query handling and building result set.

**Result serving:** showing snippets or fields of results, requires accessing document store.

## Methods of Distribution in Search

**Crawling:** by URL, or domain.

**Pre-processing:** by document.

**Indexing:** a very large distributed sort on (term,doc) elements.

**Querying:** by term or by document, or some combination.

**Result serving:** by document.

**NB.** the hard stuff to distribute is indexing and search, the IR core!

## Review: Query scores with TF-IDF

Consider a basic scoring function like  $tf.idf$ , whose component for a term  $t$  in document  $d$  is (something like)

$$tf.idf_{t,d} = \frac{tf_{t,d}}{dl_d} \log \frac{N}{df_t} \qquad tf.idf_d = \sum_{t \in \text{query}} tf.idf_{t,d}$$

$tf_{t,d} ::$  the count of term  $t$  in the document,

$dl_d ::$  the total terms in the document,

$df_t ::$  the total number of documents with term  $t$ .

- To compute  $tf.idf_d$  locally, need vector of global  $df_d$ 's.
  - Only need to compute *top K results* of all document scores.
  - For multi-term query, problem if terms are on different nodes.
- ⇒ Need to distribute document-term entries (parts of inverted lists) to make this efficient.
- ⇒ Computational cost is give by computation of score averaged over the queries (e.g., estimate via query logs).

## Basic Federated Querying

“Federating” disparate, independent IR systems.

- Assumptions:
  - No control over collections.
  - May be geographically distributed and heterogeneous.
- Similar to meta-search, but with digital library focus.
- Tasks:

**Resource Selection:** Choosing which nodes to route queries to. Good algorithm is CORI (Callan, 2000).

**Results Merging:** Assembling results from separate result sets. Confounded because they are scored differently.

**Discovery:** Sampling to estimate collection statistics for nodes.

## Basic Query Logs

**N.B.** use to assess cost of distributed IR,

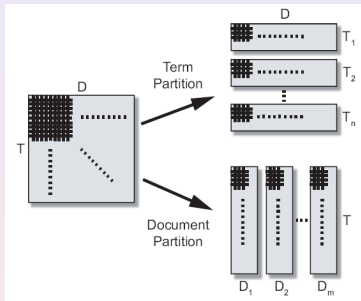
- Major search engines collect massive query logs.
- Through search toolbars, also collect click-through data.
- Query frequencies very skewed with a long tail.
- Average query length between 2 and 3.

⇒ Term correlations w.r.t. queries are critical!

- 70% of queries get over one million results on major search engines.
- Thus perhaps 30% of queries could be handled by a core 2% of the web (an educated guess).

⇒ Caching of queries is critical!

# Basic Distributed Querying



Term-document matrix from Baeza-Yates *et al.* 2007.

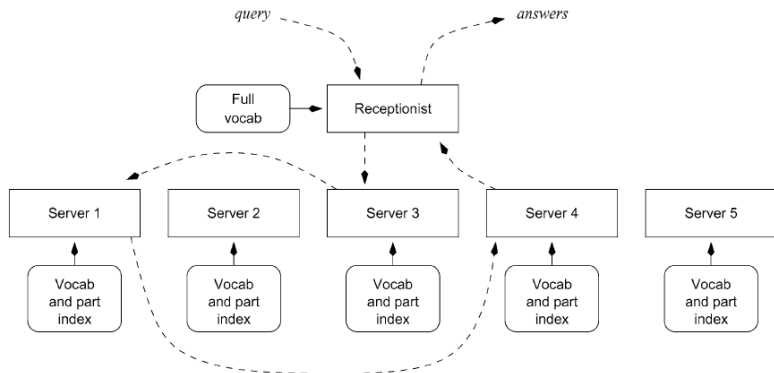
## By document partitioning:

- a parallel “find top K” algorithm,
- statistically, is well load-balanced when docs randomly distributed.

## By term partitioning:

- handling of multi-term queries:
  - *centralised broker*: creates a bottleneck at some terms and difficult to load-balance,
  - *pipelining*: more difficult to schedule but better load-balanced;
- term partitioning an optimisation task determined by query distribution.

# Pipelining in Term Partitioning



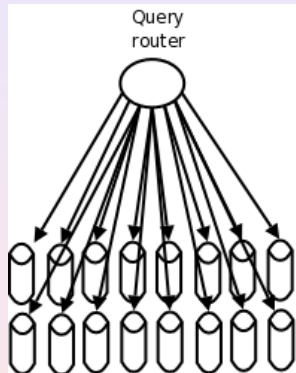
**Fig. 2** Pipelined query processing with a term-partitioned index. In this example, the query contains terms that necessitate routing the query bundle through processors 3, 1, and 4 in a system containing five servers and five index partitions

From Moffat *et al.* 2007.

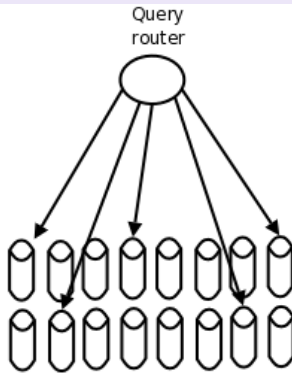
## Making Term Partitioning Efficient

- Naive partitioning leads to poor *load balancing*.
- Need to minimise transfer of document lists across nodes. Multi-term queries cause the problem.
- More correlated terms (w.r.t. queries) should co-reside at nodes.
- Duplicate some terms across nodes if they cannot be partitioned effectively.  
*e.g.* higher frequency terms (with long inverted lists).
- Can index multi-word terms (e.g., in 2005, "Britney Spears").
- Use query logs to estimate correlations and thus cost.
- Large vocabularies (tens of millions) is a problem of scale.

## Document Partitioning Styles



*Flooding of Queries*

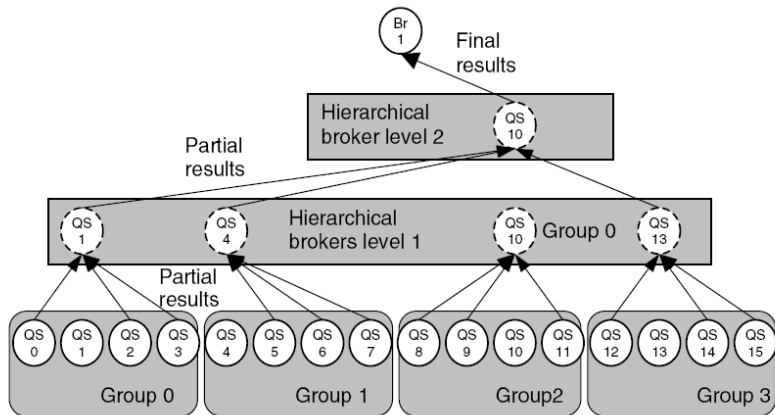


*Selective Routing*

# Document Partitioning

- Three tasks: setting up data, **partitioning**, and during runtime, **query routing** and **obtaining results**.
- Randomly distributing documents is a good strategy when all nodes are to respond to a query (flooding).
- Other partitioning option is to selectively route queries to a subset of nodes. Thus:
  - 1 *during routing predict which nodes (i.e., document sets) will respond well to a query, and, thus,*
  - 2 *partition the documents across nodes so that subsequent routing will be efficient.*
- The query routing task is called *resource selection* in the digital libraries community, but they have no control over document partitions.

# Assembling Results in Document Partitioning



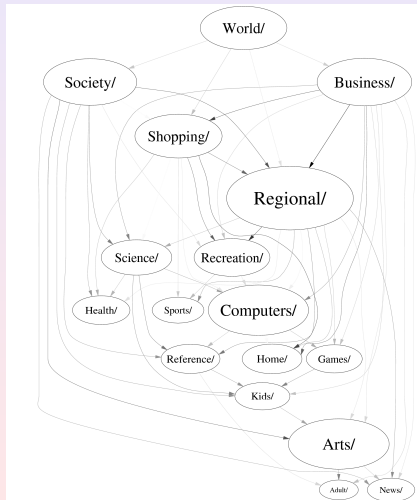
Distributed hierarchical model for assembling results using top-K processing, from Cacheda *et al.* 2007. BR=final broker, QS=query server.

## Making a Document Partition for Query Routing

Partition documents across nodes so that subsequent querying will be more efficient.

- **Interpretation:** We need to *partition the documents across nodes* so that fewer nodes respond to each query, and the load is more balanced during querying.
- Documents can be partitioned by factors like *topic, language, and major genre*.
- Documents partitions can overlap where it helps subsequent query routing.  
*e.g.* why not duplicate up to 5 times?
- *Explicit optimization* of a document partition could be made based on query logs.

# Document Partitioning by Topic



- 75,000 internet queries (no sex) used to build query topics.
- Graph is Bayesian network, nodes represent top-level DMOZ/ODP topics for a query.
- Size of node is topic frequency.
- Shade of arc gives correlation coefficient with highest 0.18. Very low!

Strong document partitioning almost impossible! So need (1) 3-5 topics/nodes to respond to queries or (2) user interaction to disam-

# Outline

- 3 Distributed IR
  - Distributed Computing Basics
  - Distributed Search Basics
  - **Literature**
  - Summary of Distributed IR
  
- 4 Peer-to-peer IR
  - Basics of P2PIR
  - Useful Papers
  - Summary and Assessment

## Workshops Literature

Most have good workshop reports online giving historical context.

- *Distributed IR* at SIGIR 2004
- *P2PIR* at SIGIR 2004
- *Heterogeneous and Distributed IR* at SIGIR 2005
- *P2PIR* 2005 and 2006 at CIKM
- *Large-Scale Distributed Systems for IR* at SIGIR 2007
- *Adversarial IR on the Web*, at WWW 2007
- **NB.** deals with different kinds of spam and competing nodes.
- Other: IPTPS, international workshop series on P2P.

## Useful Publications: Caveats

- One persons opinions.
- Not necessarily located the earliest or best proponents of ideas.
  - e.g.* Many good ideas have several independent inventors, but history records just one.
- Aims to give a spread of different ideas.

## Useful Publications

- “The anatomy of a large-scale hypertextual Web search engine”, Brin and Page, 1998.  
Describes indexing on a cluster as a distributed sort. Good historical context.
- Chapter 9, “Parallel and Distributed IR” by Brown in *Modern information retrieval* Baeza-Yates and Ribeiro-Neto, 1999.  
Nuts and bolts summary of standard models. Good introduction and framework.
- “Distributed IR”, Callan, 2000.  
Sets up the task of IR for distributed digital libraries with heterogeneous, geographically distributed, oftentimes topical and independent collections, with no overall control. Covers the standard subtasks of *database selection* and *results merging*. Many follow-up papers.

## Useful Publications, cont.

- “Building efficient and effective metasearch engines,” Meng, Yu, and Liu, *ACM Computing Surveys*, 2002.  
Review article. More recent metasearch engines general provide novel interfaces.
- “Web search for a planet: The Google cluster architecture,” Barroso, Dean, and Hoelzle, *Micro IEEE*, 2003.  
More recent review of the Google architecture giving scale and distributed aspects.
- “Three-Level Caching for Efficient Query Processing in Large Web Search Engines,” Long and Suel, *WWW Conference*, 2005.  
Review and discussion of caching in distributed search engines, plus a new method.

## Useful Publications, cont.

- “Query-Driven Document Partitioning and Collection Selection,” Puppin, Silvestri and Laforenza, *Infoscale* 2006. Includes nice review of document partitioning. Connects the document partitioning problem with query logs, and experiments reporting improvements on CORI. Query-document matrix partitioned using co-clustering.
- “Challenges on Distributed Web Retrieval,” Baeza-Yates, Castillo, Junqueira, Plachouras, Silvestri, 2007. General current review of web search for large search engines, requirements, and current research. Essential reading.

## Useful Publications, cont.

- “A pipelined architecture for distributed text query evaluation,” Moffat, Webber, Zobel and Baeza-Yates, 2007.  
A more realistic term-distributed search architecture, and discussion of intricacies.
- “Performance analysis of distributed information retrieval architectures using an improved network simulation model,” Cacheda, Carneiro, Plachouras, Ounis, *Information Processing Management*, 2007.  
Sophisticated simulation model for distributed IR, tested against live systems, used to analyse and improve document partitioning.  
Addresses the open problem given by Baeza-Yates *et al.*

## Useful Publications, cont.

- “Optimized Inverted List Assignment in Distributed Search Engine Architectures,” Zhang and Suel, *IPDPS*, 2007. Connects the term partitioning problem with query logs, and experiments using TREC GOV2. Good review and experimental work

# Outline

- 3 Distributed IR
  - Distributed Computing Basics
  - Distributed Search Basics
  - Literature
  - Summary of Distributed IR
  
- 4 Peer-to-peer IR
  - Basics of P2PIR
  - Useful Papers
  - Summary and Assessment

# Summary

- Distributed IR is a growth area, out of necessity, with active workshops.
- Many different styles and goals exist: federating independent systems, meta-search, clusters, P2P, etc.
- Practical intricacies of distributed computing make good simulation and implementation vital.
- Basic tasks:
  - indexing: a sort,
  - querying: find top K,
  - distributing data: term and/or document partitioning.
- Term/document partitioning:
  - has query logs to estimate “cost”,
  - can allow duplication for redundancy and to co-locate highly correlated terms,
  - should drift in time to reflect changes.

## Peer-to-peer IR: Outline

P2PIR is not technically optimal, but it may be economically feasible.

- 3 Distributed IR
  - Distributed Computing Basics
  - Distributed Search Basics
  - Literature
  - Summary of Distributed IR
- 4 Peer-to-peer IR
  - Basics of P2PIR
  - Useful Papers
  - Summary and Assessment

# Outline

- 3 Distributed IR
  - Distributed Computing Basics
  - Distributed Search Basics
  - Literature
  - Summary of Distributed IR
  
- 4 Peer-to-peer IR
  - Basics of P2PIR
  - Useful Papers
  - Summary and Assessment

# BitTorrent Peer-to-peer



- BitTorrent is a [P2P protocol](#) that allows distribution of multimedia data.
- Many different clients exist.
  - Its the **protocol** that matters!
- Media distribution costs shared by peers.
- Use often banned at university/commercial sites.

## Definition: Peer-to-peer

- Each computing node is a peer of the others, so they all have similar capabilities or responsibilities.
- Usually no centralised control, but sometimes have *super-peers*.  
*i.e.* some nodes are more equal than others.
- *Robustness* important, thus nodes join and drop off and functionality degrades gracefully.
- Not client-server model, but a *community model*.
- Auxiliary goals: *user-contributed computing power* and *resource sharing*.

# Glossary

**Brokered system:** a central router coordinates queries.

**Content-based locality:** nearby nodes are similar in content.

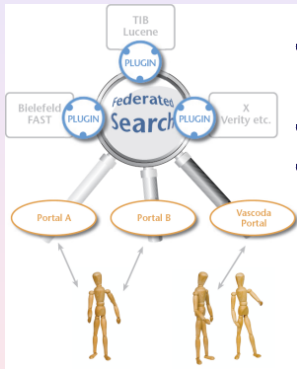
**DHT:** distributed hash tables, software such as Chord, to route queries.

**Hierarchical system:** a P2P style using "super-peers" to coordinate queries through a hierarchy.

**Network overlay:** a virtual network over a base network.

**Small world property:** any two peers likely to have a short path between them (*i.e.* w.r.t. the network overlay).

# Characteristic System: Federated Search



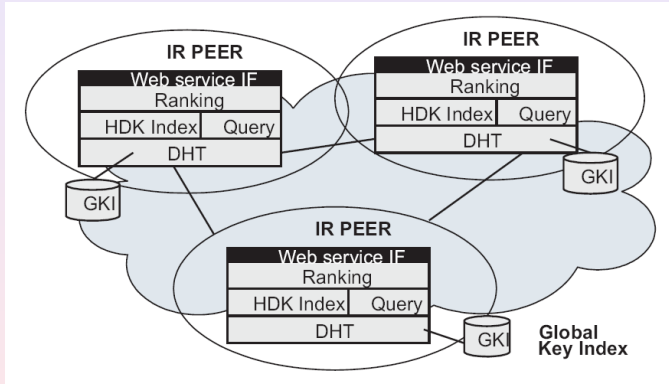
From [www.L3S.DE](http://www.L3S.DE).

- Large number of independent search engines, perhaps on different topics.
- Share protocols, e.g. query protocol.
- Same tasks as general federated IR. In addition:
  - P2P system to organise peers into a network overlay.
  - P2P system to do query routing (resource selection) and results merging.

## Characteristic System: Web Search

- An entire search engine based on P2P.
- Separate P2P protocols and components for:
  - crawl,
  - document partitioning and indexing,
  - querying and term partitioning, and
  - results serving.
- Independent, small topical search engines might join through a gateway.

# A Querying Architecture



Indexes built on DHTs, from Luu *et al.* 2006.

## Role of P2P

P2P system may play multiple roles and have different network overlays. The P2P system may:

- Do crawling, e.g., Grub.
- Do indexing of a document. But what about its anchor text?
- Do document partitioning for querying.
- Build a network overlay for querying, e.g., for federated IR.
- Do querying.
- Manage term partitioning during querying.

## Computational Tricks for Querying in P2P

Approximate and exact methods to improve P2P performance of a term/document partitioned system during querying.

- *Bloom filters* and *gap compression* to speed up transfer of document lists.
- *Pruning term lists* for a document *e.g.*, via *tf.idf*.
- *Top-K* processing to reduce size of document lists  
*e.g.* reducing local results sets to be less than required global result set size.
- Duplicate inverted lists or documents across nodes.

# Outline

- 3 Distributed IR
  - Distributed Computing Basics
  - Distributed Search Basics
  - Literature
  - Summary of Distributed IR
  
- 4 Peer-to-peer IR
  - Basics of P2PIR
  - **Useful Papers**
  - Summary and Assessment

## Useful Papers

- “On the Feasibility of Peer-to-peer Web Indexing and Search”, Li, Loo, Hellerstein, Kaashoek, Karger and Morris, 2003. (Out of date) discussion of the order of magnitudes required, and summary of some standard tricks to speed up term-distributed query processing.
- “Efficient Peer-To-Peer Searches Using Result-Caching,” Bhattacharjee, Chawathe, Gopalakrishnan, Keleher and Silaghi, *IPTPS*, 2003.  
In term partitioned P2P system, cache document lists from multi-term queries and access the cache using the tree of conjunctions.
- “Hybrid Global-Local Indexing for Efficient Peer-to-Peer Information Retrieval,” Tang and Dwarkadas, *NSDI'04*, 2004. Combines both term and document partitioning with some approximation schemes to achieve the benefits of both.

## Useful Papers, cont.

- “Efficient Query Evaluation on Large Textual Collections in a Peer-to-Peer Environment”, Zhang and Suel, *IEEE International Conference on Peer-to-Peer Computing*, 2005. Combines bloom filters with top-K processing to improve term partitioned P2P system.
- “Self-organizing distributed collaborative filtering,” Wang, Reinders, Legendijk, Pouwelse, *SIGIR*, 2005. P2P collaborative filtering, may be needed to support information retrieval.

## Useful Papers, cont.

- “P2P Content Search: Give the Web Back to the People,” Bender, Michel, Triantafillou, Weikum, Zimmer, *IPTPS*, 2006. Term partitioning and indexing with DHTs. But nodes also keep correlation statistics for their terms, then highly correlated term pairs keep pre-computed “peerlists” (inverted lists, but for peers not documents). *i.e.*, the inverted lists for highly correlated term pairs are cached with the first term. More experimentation and analysis needed. Related to Bhattacharjee *et al.* 2003.

## Useful Papers, cont.

- “Distributed Cache Table: Efficient Query-Driven Processing of Multi-Term Queries in P2P Networks,” Skobeltsyn, Aberer, *P2PIR*, 2006.

A cache-centric version of Bender *et al.* 2006 and related to Bhattacharjee *et al.* 2003. The cache stores multi-term inverted lists in a DHT. One attempts to answer queries using the cache only, and using the most specific (or largest) conjunctions available, and resorts to a full network request to build a single term inverted lists only when none exist in the cache. Good experimentation using Wikipedia data.

## Useful Papers, cont.

- “ALVIS Peers: A Scalable Full-text Peer-to-Peer Retrieval Engine,” Luu, Klemm, Podnar, Rajman, Aberer, *P2PIR*, 2006. Term distributed system using DHTs. Only indexes full term sets called *highly discriminative keys* known to have smaller inverted lists. Full discussion and experimentation.
- “Size Doesn’t Always Matter: Exploiting PageRank for Query Routing in Distributed IR,” Parreira, Michel, Bender, *P2PIR*, 2006.  
Using PageRank for peers in a document partitioned system.

## Useful Papers, cont.

- “Content-based peer-to-peer network overlay for full-text federated search”, Lu and Callan, *8th RIAO Conference on Large-Scale Semantic Access to Content*, 2007.  
P2P method for document partitioning. Local algorithms to form an overlay network with content-based locality and small world properties. Uses KL distance on unigrams as similarity. Thus CORI and resource selection algorithms will be effective.
- “Homepage Finding in Hybrid Peer-to-Peer Networks”, Bragante and Melucci, *8th RIAO Conference on Large-Scale Semantic Access to Content*, 2007.  
Incorporates anchortext and URL links into the P2P process for a document partitioned system.

## Useful Papers, cont.

- “Design Alternatives for Large-Scale Web Search: Alexander was Great, Aeneas a Pioneer, and Anakin has the Force,” Bender, Michel, Triantafillou, Weikum, *SIGIR Workshop on Large Scale Distributed Systems for IR*, 2007.  
Discusses hardware environments and architectures for document and term partitioning P2PIR systems. No precise simulation is done, but the numbers and discussion is detailed. Suggests term partitioning with newer, cheaper Flash-RAM memory may work well.

# Outline

- 3 Distributed IR
  - Distributed Computing Basics
  - Distributed Search Basics
  - Literature
  - Summary of Distributed IR
  
- 4 Peer-to-peer IR
  - Basics of P2PIR
  - Useful Papers
  - Summary and Assessment

## Summary of P2P IR

- P2P IR builds on technology from distributed IR.
- Really only makes sense in the context of a full search engine (*i.e.*, with a crawler *etc.*)
- An exciting part of distributed IR due to its potential.
- Not because we believe it is the best technical approach, but because it is an economically feasible approach.
- Many aspects of the IR system subject to P2P treatment.

## Assessment of P2P IR

- Should see a convergence of term partitioning, clever caching (e.g., Skobeltsyn *et al.* 2006), and document partitioning methods, plus the use of novel user interaction.
- Successful systems will employ alternative and next generation search methods such as social bookmarking, tagging, “poor man’s” semantic web, multimedia, genre, *etc.*, to add value.
- Functional and practical P2PIR should be achieved in the next 5 years, but will not compete with current major search engines in many respects,  
*i.e.* it may remain an alternative search technology.