Overview of Open Source and Information Retrieval

Andy MacFarlane

Centre for Interactive Systems Research

Department of Information Science

andym@soi.city.ac.uk

Talk Structure

- What is Information Retrieval (IR)?
- Motivation for using OS Development in IR
- A brief survey of OS IR Systems
- Problems and obstacles
- Summary

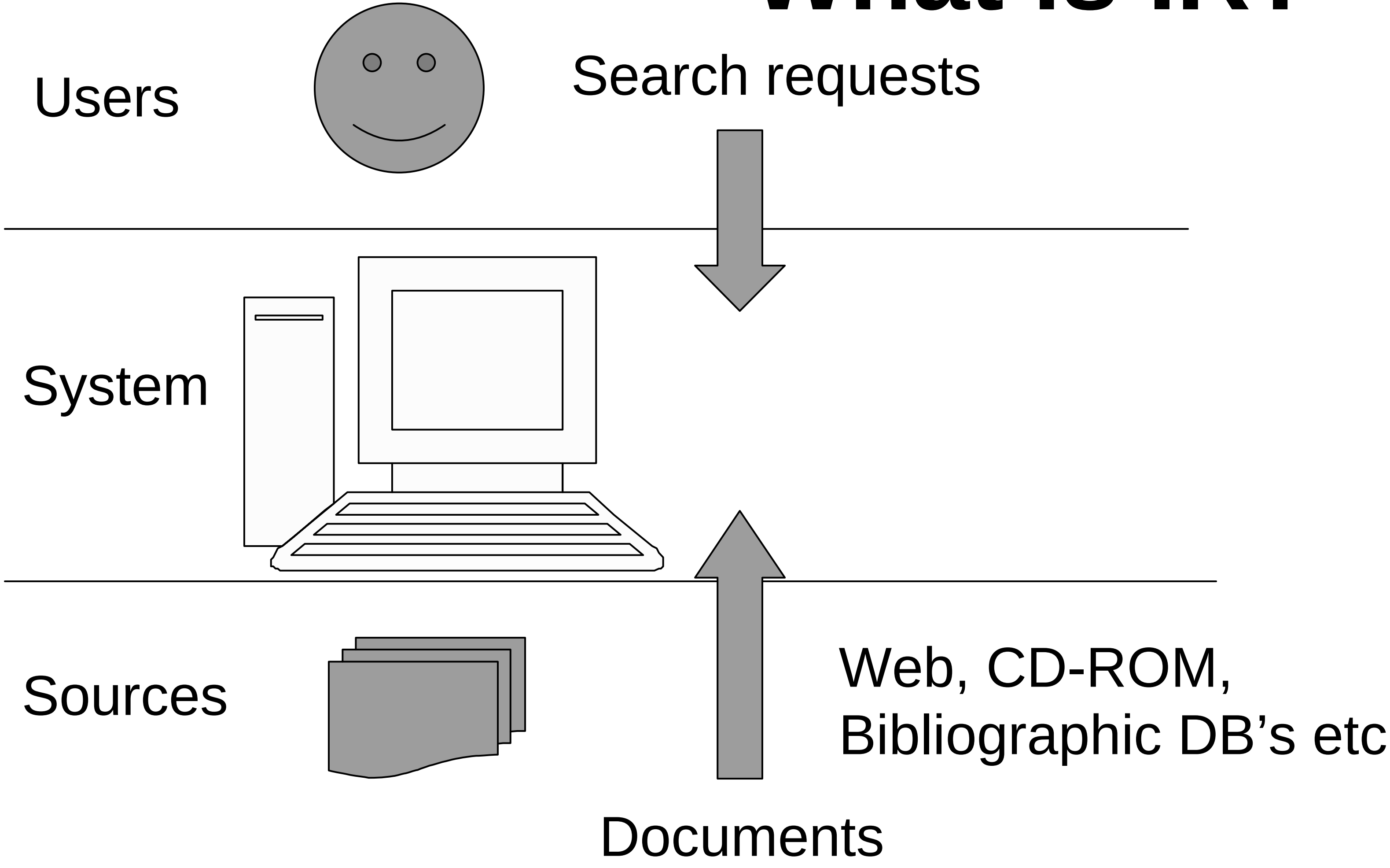
Talk Structure

- **What is Information Retrieval (IR)?**
- Motivation for using OS Development in IR
- A brief survey of OS IR Systems
- Problems and obstacles
- Summary

What is IR?

- Google! ?
- Just one example of an IR system in a particular area (Web Search)
- For the most part it means - keyword access to text
- Is some interest in other media e.g. images
- I will concentrate on text here
- Issues: Definitions,Technical, evaluation, tasks

What is IR?



Users

Search requests

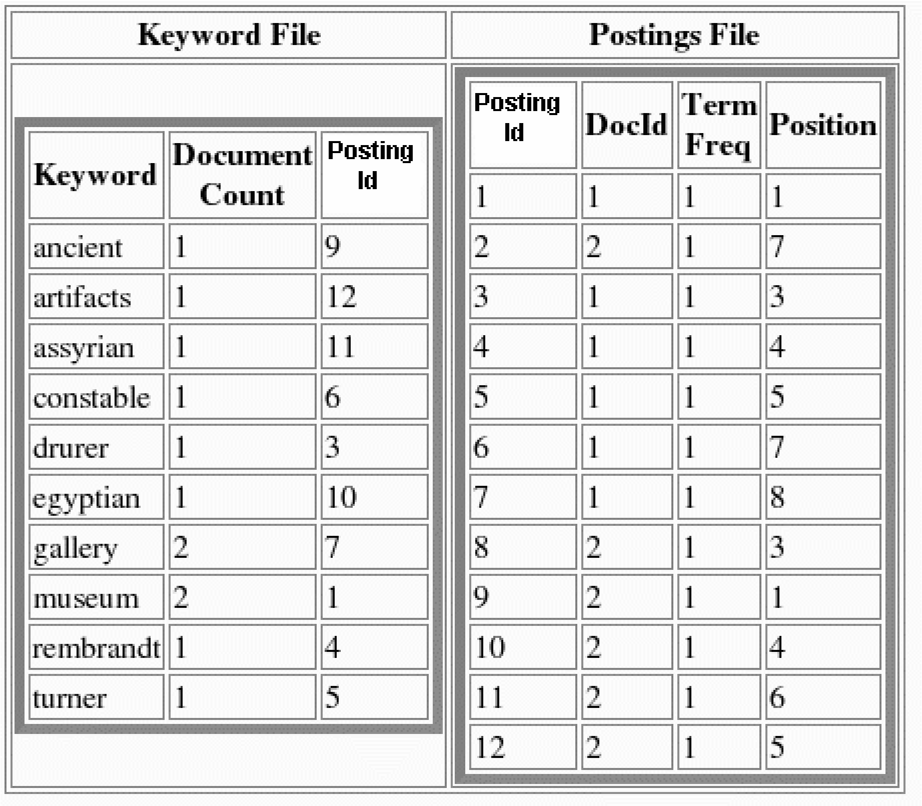
System

Sources

Web, CD-ROM,
Bibliographic DB’s etc

Documents

What is IR?



Keyword File

Keyword	Document Count	Posting Id
ancient	1	9
artifacts	1	12
assyrian	1	11
constable	1	6
drurer	1	3
egyptian	1	10
gallery	2	7
museum	2	1
rembrandt	1	4
turner	1	5

Postings File

Posting Id	DocId	Term Freq	Position
1	1	1	1
2	2	1	7
3	1	1	3
4	1	1	4
5	1	1	5
6	1	1	7
7	1	1	8
8	2	1	3
9	2	1	1
10	2	1	4
11	2	1	6
12	2	1	5

What is IR?

- Evaluation - a big issue in IR
- Types: Operational, laboratory
- Cranfield Model: collections, information needs and relevance judgements
- Measures: Precision/Recall
- TREC: operational and laboratory
- TREC drives evaluation in IR now.

What is IR?

- Tasks - a few examples
- Web Search (Ad-hoc)
- Enterprise Search
- Domain Specific Search (Medicine, Law)
- Cross Language Search
- Question and Answering
- Source code search

Talk Structure

- What is Information Retrieval (IR)?
- **Motivation for using OS Development in IR**
- A brief survey of OS IR Systems
- Problems and obstacles
- Summary

Motivation for using OS in IR

- Technical, economic and political
- Technical primarily IMHO
- Argument for OS, take advantage of OS software development process e.g.
  - Parallel development, prompt feedback etc
- Share knowledge (Porter and Boulton)
  - Developers
  - Academic community

Motivation for using OS in IR

- Economic, political reasons impact as well
- 70/80% of software costs is on maintaining it, makes sense to share these costs (Raymond)
- On personal level improve reputation, personal economic benefit
- Political: some people don’t like working with or on proprietary software!

Talk Structure

- What is Information Retrieval (IR)?
- Motivation for using OS Development in IR
- **A brief survey of OS IR Systems**
- Problems and obstacles
- Summary

Survey of OS IR Systems

- My paper: on Open Source IR (2003)
- Found 36 in 2003 (mostly from Sourceforge, Freshmeat)
- Tried my query again recently - didn’t work :(
- Technical division: inverted files and DB systems
- DB can systems can be used for small collections

Survey of OS IR Systems

- Some examples
- Libraries: Senda, Xapian (more later), Lucene
- Web search: Ht://Dig, swish, swish++ (more in a moment)
- Academic research: MG, Lemur, Terrier
- Licenses: GPL mostly + private, apache, LGPL
- Range of activity varied from very active to moribund.

Swish++

- Web site search program
- Has crawler/indexer
- Only supported English (2003)
- Can index non-ASCII document e.g. word
- No ability to merge intermediate results
- Supports
  - Boolean search
  - Term weighting (model is not specified)

Ht:/Dig

- Web site search program (like Swish++)
- Quite widely used, less so now with Google
- Like Swish++ has crawler/indexer
- Supports Boolean/term weighting search
- Ranking function
  - Higher weights for terms which occur higher up the document than those which occur lower
- Bug ridden and hard to use (2003)

Talk Structure

- What is Information Retrieval (IR)?
- Motivation for using OS Development in IR
- A brief survey of OS IR Systems
- **Problems and obstacles**
- Summary

Problems and obstacles

- Google: can be used for web site search - make sense because of the evidence available
- Forks: Xapian and Open Muscat/Omsee
- Proliferation: number of systems found in my survey (some collaboration however)
- Effectiveness of OSS development in the light of Forks/Proliferation
- Usability: Nichol et al - what about the user?

Talk Structure

- What is Information Retrieval (IR)?
- Motivation for using OS Development in IR
- A brief survey of OS IR Systems
- Problems and obstacles
- **Summary**

Summary

- OS software development is a good way to develop IR systems and share ideas
- There are clear benefits in terms of the development process and issues in IR
- Not without problems: forks, usability
- These can be tackled and are!
- Open source in IR workshops: Two so far, OSWIR 2005, OSIR 2006

References

- A. MacFarlane (2003). On Open Source IR. ASLIB proceedings: New Information Perspectives. Vol. 55, No. 4 2003.
- Porter and Boulton (2000). Open Muscat, an Open Source search engine, SIGIR Forum, 34(1), 16-17.
- Nichols et al (2001) Usability and open source software development: http://www.cs.waikato.ac.nz/~daven/docs/oss.html