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?

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Document Count</th>
<th>Posting Id</th>
</tr>
</thead>
<tbody>
<tr>
<td>ancient</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>artifacts</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>assyrian</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>constable</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>drurer</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>egyptian</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>gallery</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>museum</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>rembrandt</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>turner</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Posting Id</th>
<th>DocId</th>
<th>Term Freq</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
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
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