Views on Static Ranking

*Open Source Web Information Retrieval*

Wray Buntine

Helsinki Institute for Information Technology

September 19, 2005
Background

• Primary ranking factor is “relevance”, based on query. Secondary ranking fact is Authority/Reputation/Trust, often independent of query.

• Launched Google, PageRank$^\text{TM}$, which works well for home page finding.

• Specialised search engines need this kind of infrastructure.

• Theories: random surfer and seeker models, citation analysis, social networks,
Random Surfers and Seekers

- The *random surfer* wanders around the web indefinitely, and we record long term frequency at pages.

- The *random seeker* wanders around the web until they find something, then stop and we record probability of stopping at a page.

- Both can be tuned to specific topics, and start pages, and can use a back button.

Ask me for the full paper “Random Surfers versus Random Seekers” by Wray Buntine
Authority models

• A *hub* links to other pages, with a mixture of topics.

• A *authority* is a good page to link to for a particular topic.

• Build a probabilistic hub-authority model using a variant of Discrete-PCA/ICA/LDA/NMF/PLSI/...
Authority/Reputation services for an open web search

• We need to develop/encourage independent ranking services so that smaller search engines can flourish, and to give users choice.

• *e.g.*, a Yahoo business products and services ranking.

• *e.g.*, an ACM computer science ranking.
OpenSourceSearch.ORG

- Wiki-based.

- Prototype news/blogs search with “search engine” content.

- Started as part of ALVIS project.

- Wants an *Open Source Search Foundation* to manage and focus.