

Beyond the Web: Retrieval in Social Information Spaces

Sebastian Marius Kirsch

`kirschs@informatik.uni-bonn.de`

Institut für Informatik III

Rheinische Friedrich-Wilhelms-Universität Bonn

10th April 2006

Outline

Social Information Spaces

Retrieval with Social Networks

An Algorithm for Social Retrieval

Evaluation

Conclusion

Social Information Spaces

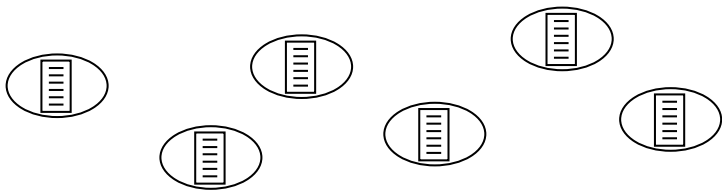
- ▶ 'We live, work, play in social spaces – both online and offline.'
[Lueg and Fisher, 2003]
- ▶ 'Man is a social animal.'
- ▶ online group interaction predates the internet (email mailing lists, Usenet)
- ▶ today: surge in web-based social software
 - ▶ wikis (Wikipedia, ...)
 - ▶ blogs (LiveJournal, Blogspot, MySpace, ...)
 - ▶ social networking platforms (Friendster, orkut, openBC, ...)
 - ▶ 'social' bookmarking (del.icio.us, simpy, ...)
 - ▶ more added every day
- ▶ realize vision of the 'read-write web' [Lawson, 2005]

Beyond the web?

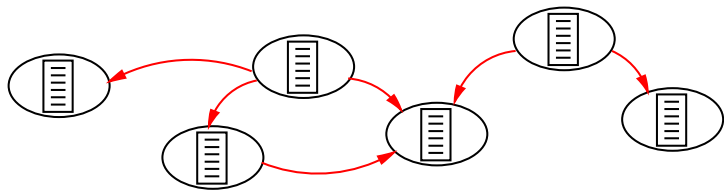
- ▶ web is a document-centric system
- ▶ documents authored individually, joined by hyperlinks
- ▶ web is just a user interface for social information spaces
- ▶ underlying information space lives in a database
- ▶ social information spaces: users, their documents, and relations between them.

⇒ analyze the information space directly for information retrieval

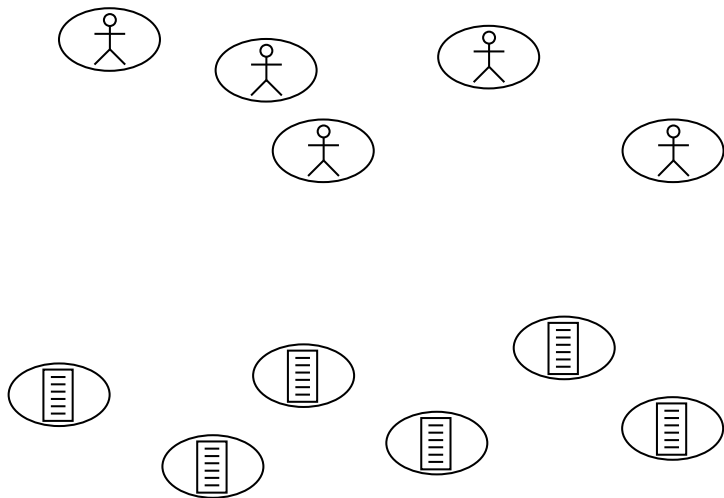
Information Spaces



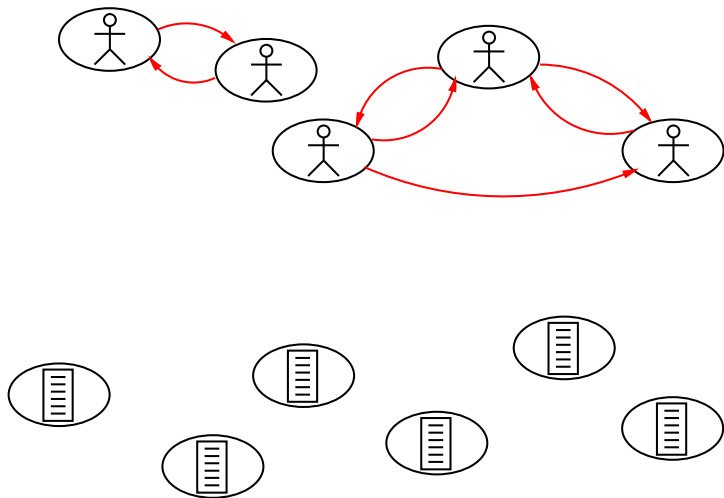
Information Spaces



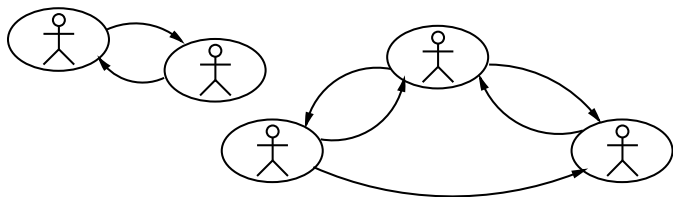
Information Spaces



Information Spaces

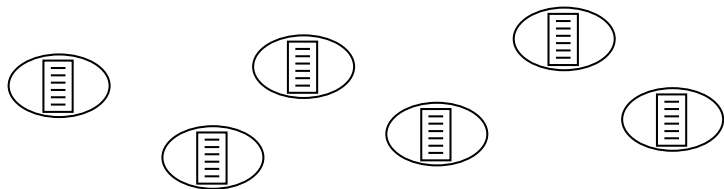


Information Spaces

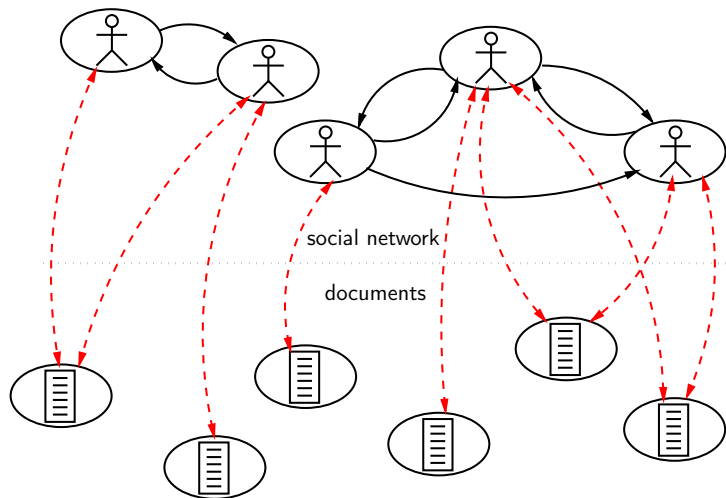


social network

documents



Information Spaces



Web retrieval vs. social retrieval

- ▶ web retrieval
 - ▶ content and keywords not sufficient to determine relevant pages
 - ▶ algorithms analyse hyperlink structure
 - ▶ try to infer authority of a page from the pages linking to it
 - ▶ most prominent example: PageRank [Page et al., 1999]
- ▶ social networks
 - ▶ graph-based retrieval, like web retrieval
 - ▶ social networks share many statistical properties with the web graph (small world, power-law distribution, clustering)

⇒ apply techniques from web retrieval

⇒ use PageRank as authority measure on social network

PageRank as an authority measure for social networks?

PageRank scores extracted from coauthorship network of 25 years of SIGIR proceedings, normalized, with a teleportation probability of $\epsilon = 0.3$:

rank	name	PageRank
1.	Bruce W. Croft	7.929
2.	Clement T. Yu	4.716
3.	James P. Callan	4.092
4.	Norbert Fuhr	3.731
5.	Susan T. Dumais	3.731
6.	Mark Sanderson	3.601
7.	Nicholas J. Belkin	3.518
8.	Vijay V. Raghavan	3.303
9.	James Allan	3.200
10.	Jan O. Pedersen	3.135

PageRank-based algorithm for social IR

1. Extract authors and social network from corpus.
2. Compute PageRank scores r_i for authors in the social network.
3. Assign PageRank scores to documents: $r_d \leftarrow r_i$ if i is author of d .
4. For a query q , determine set of relevant documents D_q and relevance scores $\text{score}(q, d)$ for $d \in D_q$
5. Combine PageRank scores with relevance scores:

$$r_d \cdot \text{score}(q, d)$$

6. Sort D_q by $r_d \cdot \text{score}(q, d)$ and return it.

Evaluation

- ▶ task: known-item retrieval
- ▶ metrics: average rank and inverse average inverse rank
- ▶ compare performance with performance of a baseline method
- ▶ mailing-list archive (44108 messages from 2000–2005, 1834 different email addresses)
- ▶ semi-automatic method for choosing query terms and known items
- ▶ results for expert searcher
 - ▶ average rank increases (up to 70%)
 - ▶ up to 25% decrease in IAIR
 - ▶ better results for larger collections
- ▶ results for novice searcher are inconclusive
 - ▶ increase in both average rank and IAIR for larger collections
 - ▶ no trend as regards collection size

Conclusion

- ▶ social networks are an integral part of information retrieval
- ▶ social network analysis can lead to significant performance improvements
- ▶ further research is necessary
 - ▶ evaluation
 - ▶ application to different domains
 - ▶ perhaps combine with community approaches?
 - ▶ privacy implications?
- ▶ rise of social software will necessitate retrieval algorithms using social networks
- ▶ generate tangible advantages from using social software

Questions? Feedback?

Thank you very much for listening!

slides for this talk are available at

[http://www.sebastian-kirsch.org/moebius/docs/
ecir2006-slides.pdf](http://www.sebastian-kirsch.org/moebius/docs/ecir2006-slides.pdf)

Beyond the Web: Retrieval in Social Information Spaces

Sebastian Marius Kirsch

`kirschs@informatik.uni-bonn.de`

Institut für Informatik III

Rheinische Friedrich-Wilhelms-Universität Bonn

10th April 2006

Mark Lawson. Berners-Lee on the read/write web. broadcast by Newsnight on BBC Two, August 2005. URL <http://news.bbc.co.uk/1/hi/technology/4132752.stm>. Interview with Tim Berners-Lee.

Christopher Lueg and Danyel Fisher, editors. *From Usenet to CoWebs. Interacting with social information spaces*. Springer, 2003. ISBN 1-85233-532-7.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford University, November 1999. URL <http://dbpubs.stanford.edu:8090/pub/1999-66>.