Social Information Retrieval

Sebastian Marius Kirsch
kirschs@informatik.uni-bonn.de

25th November 2005

Format of this talk

- about my diploma thesis
- advised by Prof. Dr. Armin B. Cremers
- inspired by research by Melanie Gnasa
- this talk: evolutional rather than technical
- describe the development of my thesis

Outline

Motivation

Social networks

An Algorithm for social IR

Evaluation

Second approach: Associative networks

A model for social IR

Additional work

Conclusion

Outline

Motivation

Social networks

An Algorithm for social IR

Evaluation

Second approach: Associative networks

A model for social IR

Additional work

Conclusion

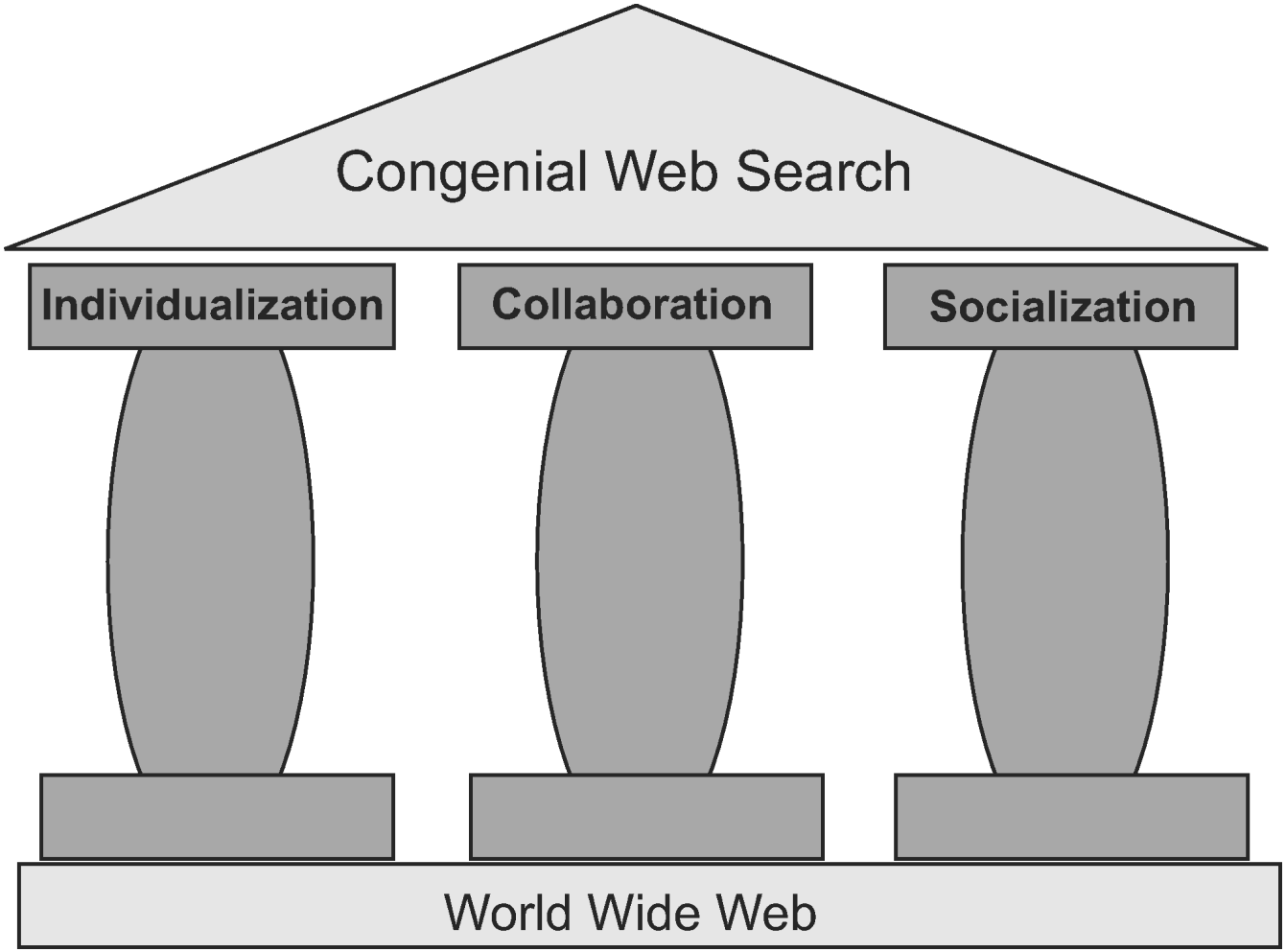
What is information retrieval?

- Popular perception:

  information retrieval = to google for something

  (verb 'to google' is included in the Oxford American Dictionary!)
- The goal of **information retrieval** (IR) is facilitating a user's access to information that is relevant to his information needs.
- [BYRN99]: An information retrieval system 'should provide the user with easy access to the information in which he is interested.'

Three pillars for web search



(source: [GWC04])

Three pillars make a solid edifice?

Individualized (personalized) and collaborative IR:

- prior art exists
  (eg. SearchPad, OutRide, I-SPY)
- slowly becoming mainstream
  (eg. Google Personalized Search, a9.com)

Social IR:

- No prior art exists?
- What is social IR anyway?

Questions:

- What is ‘social’?
- How can we use it for IR?

What is 'social' anyway?

Main Entry: ^{1}**so · cial**
Pronunciation: `'sO-sh&l`
Function: adjective
Etymology: Middle English, from Latin socialis, *from* socius *companion, ally, associate; akin to Old English* secg *man, companion, Latin* sequi *to follow*

source: Merriam-Webster Online Dictionary

What is 'social' anyway?

Main Entry: ^{1}**so · cial**
Pronunciation: `'sO-sh&l`
Function: adjective
Etymology: Middle English, from Latin socialis, *from* socius *companion, ally, associate; akin to Old English* secg *man, companion, Latin* sequi *to follow*

source: Merriam-Webster Online Dictionary

- Every interaction with a fellow human is a **social act**.
- Social interactions form **social ties** between people.
- The entirety of social ties forms a **social network**.

$\Rightarrow$ social network analysis as tool for social IR?

Outline

Motivation

Social networks

An Algorithm for social IR

Evaluation

Second approach: Associative networks

A model for social IR

Additional work

Conclusion

Where do we find social networks?

- traditional sociology/social psychology: fieldwork, conduct interviews, etc.
- electronic media: extract social networks from electronic records
- examples for social media:
  - mailing lists
  - blogs
  - wikis
- much larger and more complex networks than previously available!
- largest well-researched social networks are currently scientific collaboration networks (with more than 1.5 mio. individuals)

Special properties of social networks?

- 'small-world network' [Mil67], 'six degrees of separation': low average shortest path length
- power-law degree distribution: probability of a person having k contacts is proportional to $k^{-\gamma}$ ($\gamma \approx 0.9 \ldots 2.5$)
- giant connected component: 70%–90% of all individuals are part of one connected component.
- high degree of clustering: high probability that two of your friends are friends with each other

$\Rightarrow$ similarities with the web graph!
Use techniques from web retrieval for social IR?

Web retrieval

- the web: a huge collection of semi-structured hypertext
- search engines index up to 20 billion web pages
- 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 [PBMW99]

Outline

Motivation

Social networks

An Algorithm for social IR

Evaluation

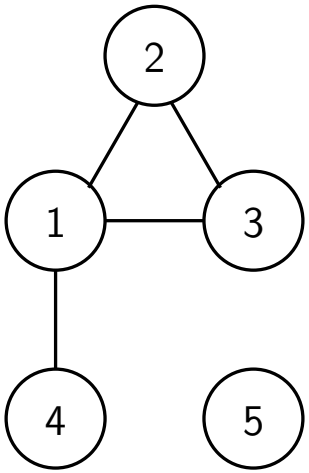
Second approach: Associative networks

A model for social IR

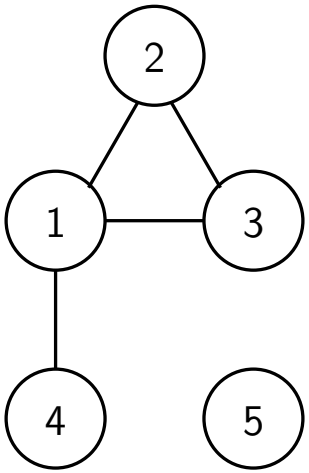
Additional work

Conclusion

PageRank: An authority measure for graphs



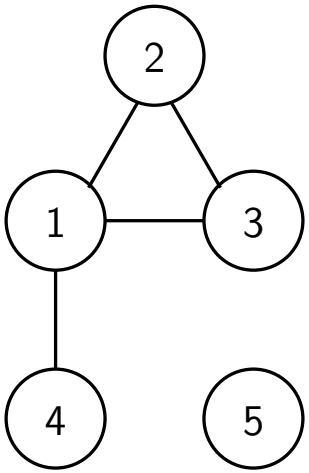
PageRank: An authority measure for graphs



$$\Rightarrow \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

adjacency matrix

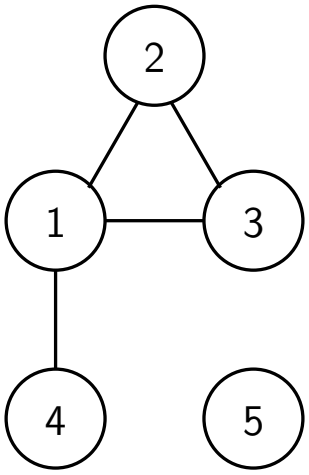
PageRank: An authority measure for graphs



$$\Rightarrow \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

adjacency matrix — row-normalized

PageRank: An authority measure for graphs



$$\Rightarrow \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

adjacency matrix

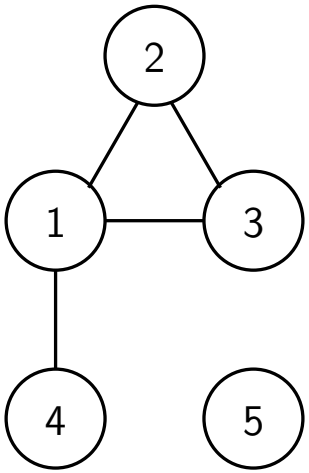
$$\Rightarrow \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix} \Rightarrow$$

row-normalized

$$\begin{pmatrix} \frac{1}{15} & \frac{13}{45} & \frac{13}{45} & \frac{13}{45} & \frac{1}{15} \\ \frac{2}{5} & \frac{1}{15} & \frac{2}{5} & \frac{1}{15} & \frac{1}{15} \\ \frac{2}{5} & \frac{2}{5} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\ \frac{11}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

with teleport ($\epsilon = \frac{1}{3}$)

PageRank: An authority measure for graphs



$$\Rightarrow \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

adjacency matrix

$$\Rightarrow \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix} \Rightarrow$$

row-normalized

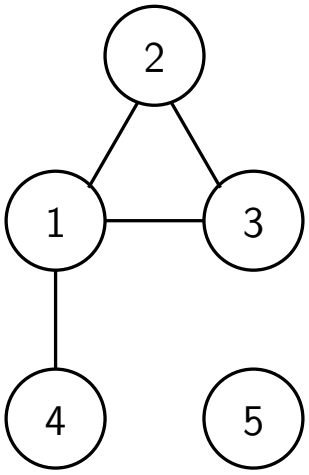
$$\begin{pmatrix} \frac{1}{15} & \frac{13}{45} & \frac{13}{45} & \frac{13}{45} & \frac{1}{15} \\ \frac{2}{5} & \frac{1}{15} & \frac{2}{5} & \frac{1}{15} & \frac{1}{15} \\ \frac{2}{5} & \frac{2}{5} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\ \frac{11}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

with teleport ($\epsilon = \frac{1}{3}$)

$$\Rightarrow \begin{pmatrix} \frac{1}{15} & \frac{2}{5} & \frac{2}{5} & \frac{11}{15} & \frac{1}{5} \\ \frac{13}{45} & \frac{1}{15} & \frac{2}{5} & \frac{1}{15} & \frac{1}{5} \\ \frac{13}{45} & \frac{2}{5} & \frac{1}{15} & \frac{1}{15} & \frac{1}{5} \\ \frac{13}{45} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{5} \\ \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{5} \end{pmatrix}$$

transposed

PageRank: An authority measure for graphs



$$\Rightarrow \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

adjacency matrix

$$\Rightarrow \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix} \Rightarrow$$

row-normalized

$$\begin{pmatrix} \frac{1}{15} & \frac{13}{45} & \frac{13}{45} & \frac{13}{45} & \frac{1}{15} \\ \frac{2}{5} & \frac{1}{15} & \frac{2}{5} & \frac{1}{15} & \frac{1}{15} \\ \frac{2}{5} & \frac{2}{5} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\ \frac{11}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{1}{15} & \frac{2}{5} & \frac{2}{5} & \frac{11}{15} & \frac{1}{5} \\ \frac{13}{45} & \frac{1}{15} & \frac{2}{5} & \frac{1}{15} & \frac{1}{5} \\ \frac{13}{45} & \frac{2}{5} & \frac{1}{15} & \frac{1}{15} & \frac{1}{5} \\ \frac{13}{45} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{5} \\ \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{5} \end{pmatrix} \Rightarrow \begin{pmatrix} 1.63 \\ 1.12 \\ 1.12 \\ 0.75 \\ 0.38 \end{pmatrix}$$

with teleport ($\epsilon = \frac{1}{3}$)

transposed

dom. eigenvector

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.

Outline

Motivation

Social networks

An Algorithm for social IR

Evaluation

Second approach: Associative networks

A model for social IR

Additional work

Conclusion

Evaluation of IR systems

- not a clear-cut problem
- different methodologies, settings and metrics exists
  eg. evaluation of interactive performance vs. evaluation in a batch setting
- comparability of results not always ensured between different IR systems or even between different experiments with the same system
- for our experiments: use batch setting
  - determine query terms and relevant documents beforehand
  - evaluate whether the system finds the relevant documents
  - take position in result list into account
  - compare performance with performance of a baseline method
  - task: known-item retrieval
    find a single document
  - metrics: average rank and inverse average inverse rank

Corpus and queries

- mailing-list archive
- messages from years 2000–2005
- 44108 messages
- 1834 different email addresses
- used two subsets for evaluation:
  1. messages from 2004
  2. messages from 2000–2005
- choosing query terms and the ‘known item’:
  1. consider only messages from 2004
  2. extract frequent bi- and trigrams from subject lines
  3. choose 10 bi- and trigrams which are frequent, but not correlated with author of message
  4. consider messages with chosen bi- or trigram in subject
  5. have two human experts choose one of the messages as ‘known item’

Results (expert searcher)

method:	VS	PR×VS
searcher:	expert	expert
	on messages from 2004:	
$\overline{\text{rank}}$:	14.75 ± 0.25	17.95 ± 0.05
$\overline{\text{rank}}$ change [%]:		$+21.7 \pm 2.4$
IAIR:	7.548 ± 0.032	7.082 ± 0.010
IAIR change [%]:		-6.2 ± 0.5
	on messages from 2000–2005:	
$\overline{\text{rank}}$:	24.4 ± 0.3	41.45 ± 0.05
$\overline{\text{rank}}$ change [%]:		$+69.9 \pm 2.3$
IAIR:	8.787 ± 0.040	6.697 ± 0.012
IAIR change [%]:		-24.6 ± 0.5

Results (novice searcher)

method:	VS	PR×VS
searcher:	novice	novice
	on messages from 2004:	
$\overline{\text{rank}}$:	17.5 ± 0.3	15.2 ± 0
$\overline{\text{rank}}$ change [%]:		-13.1 ± 1.5
IAIR:	4.670 ± 0.013	4.599 ± 0
IAIR change [%]:		-1.5 ± 0.3
	on messages from 2000–2005:	
$\overline{\text{rank}}$:	39.35 ± 0.35	39.6 ± 0
$\overline{\text{rank}}$ change [%]:		$+0.6 \pm 0.9$
IAIR:	4.962 ± 0.013	7.86 ± 0
IAIR change [%]:		$+58.4 \pm 0.4$

Outline

Motivation

Social networks

An Algorithm for social IR

Evaluation

Second approach: Associative networks

A model for social IR

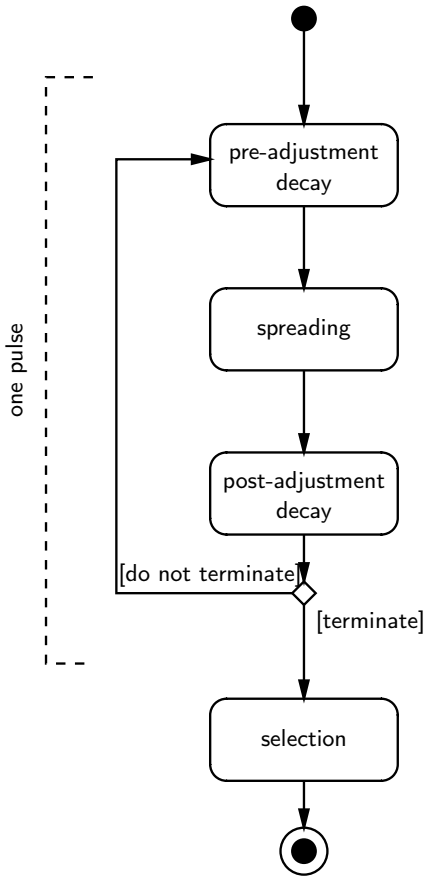
Additional work

Conclusion

A second approach: Associative networks

- first approach was motivated by web retrieval
- also explored a second approach motivated by associative retrieval
- treat problem domain as associative network containing documents, authors and queries
- use spreading activation search:
  - search algorithm motivated by neural networks
  - based on concept of 'activation energy'
  - energy spreads through network via links
  - constraints and adjustments limit and direct spread of activation

Spreading activation search

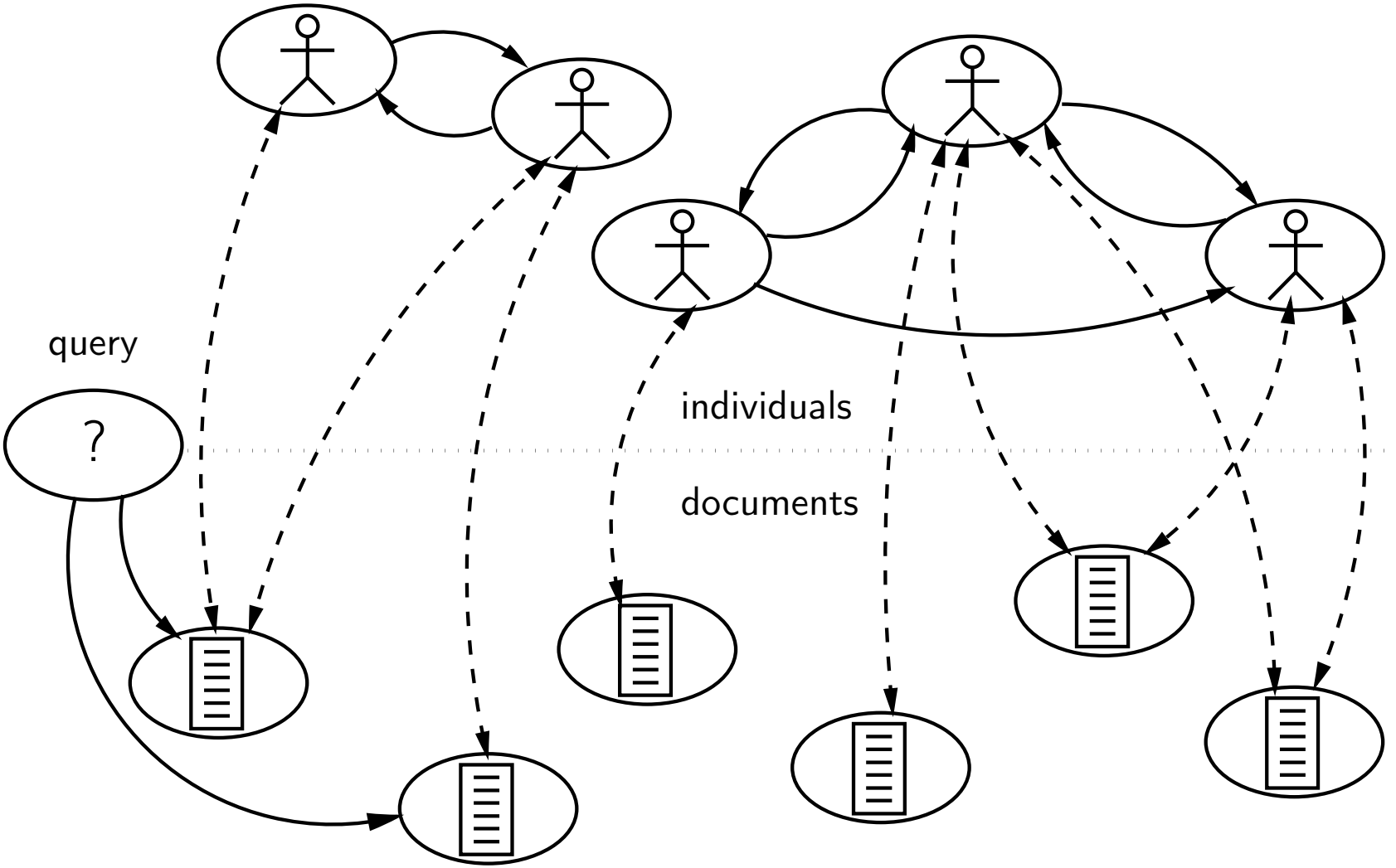


- iterative process
- four steps in every iteration
  1. pre-adjustment, decay:
     output energy of a node is computed from activation level in the previous iteration
  2. spreading:
     input energy is accumulated for each node on the network
  3. post-adjustment, decay:
     new activation level is computed from input energy and activation level in previous interation
  4. termination check:
     after a fixed number of iterations, or other when other conditions are met, iteration stops.

Spreading activation search

- not a search algorithm per se
- method for formalising different search algorithms
- often employed in an interactive fashion: user reviews newly activated nodes after each iteration and decides direction of search
- constraints and adjustments must be carefully chosen
- common problems: whole network gets activated *or* activation decays to fast.
- large number of possible adjustments and constraints makes systematic choice difficult
- parameters often mimic an inference process

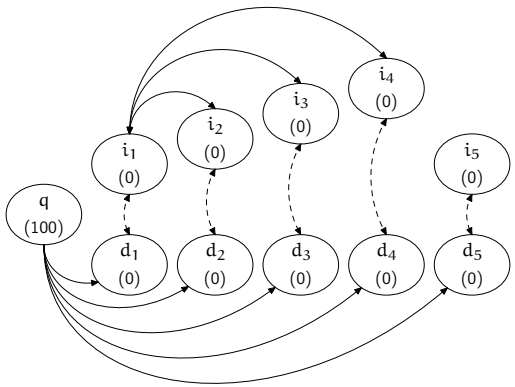
Associative network for the domain



Spreading activation for social IR

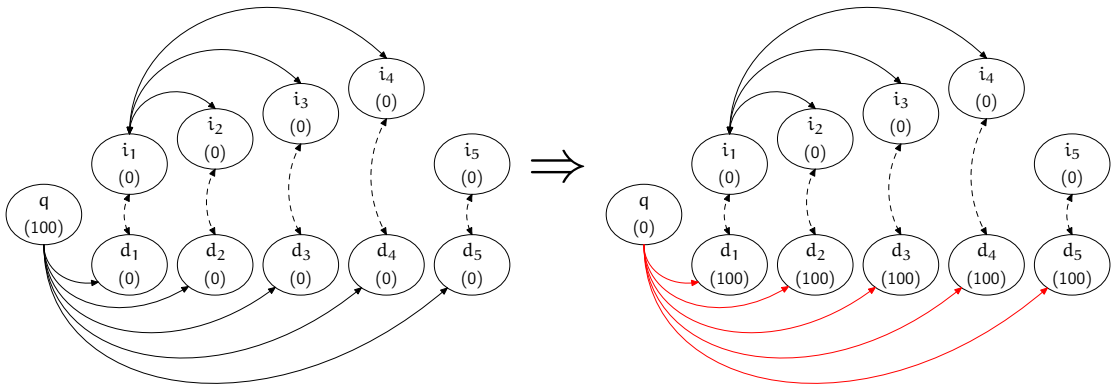
- mimic an inference process we would use to infer the relevance of a document:
  - initial relevance is determined by keyword retrieval (conventional IR)
  - authors of relevant documents are presumed experts
  - an author is authoritative if he has social ties with many experts in the topic, and if he has written many documents about the topic
  - relevance of a document depends on its initial relevance and the authority of its author
- implement these rules as a set of five constraints and adjustments; terminate after four iterations.

Example search



initial activation

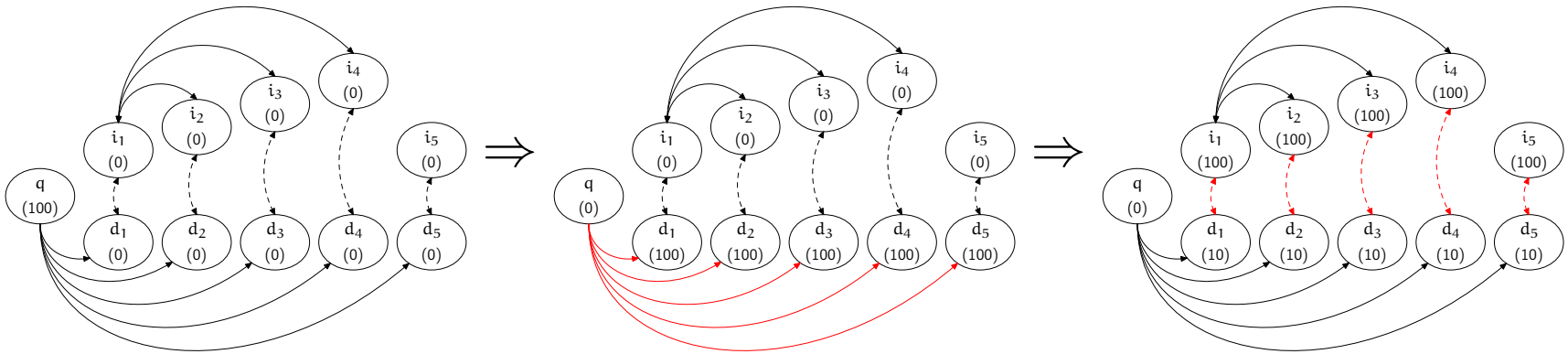
Example search



initial activation

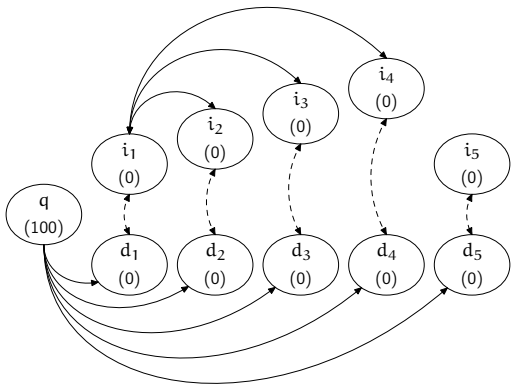
after 1st pulse

Example search

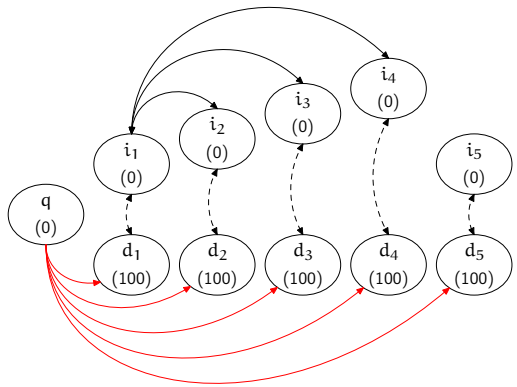


initial activation | after 1st pulse | after 2nd pulse

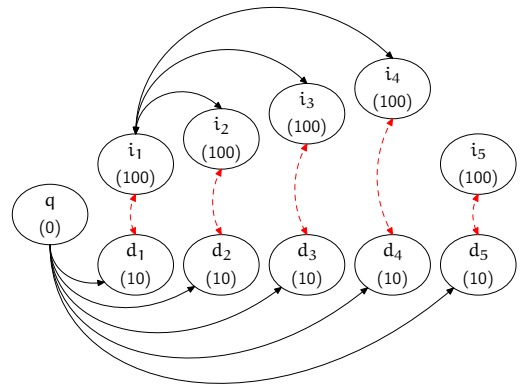
Example search



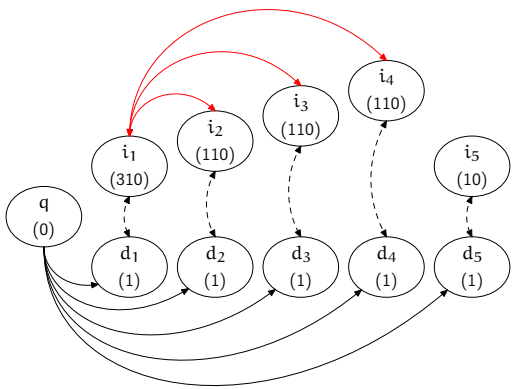
initial activation



after 1st pulse

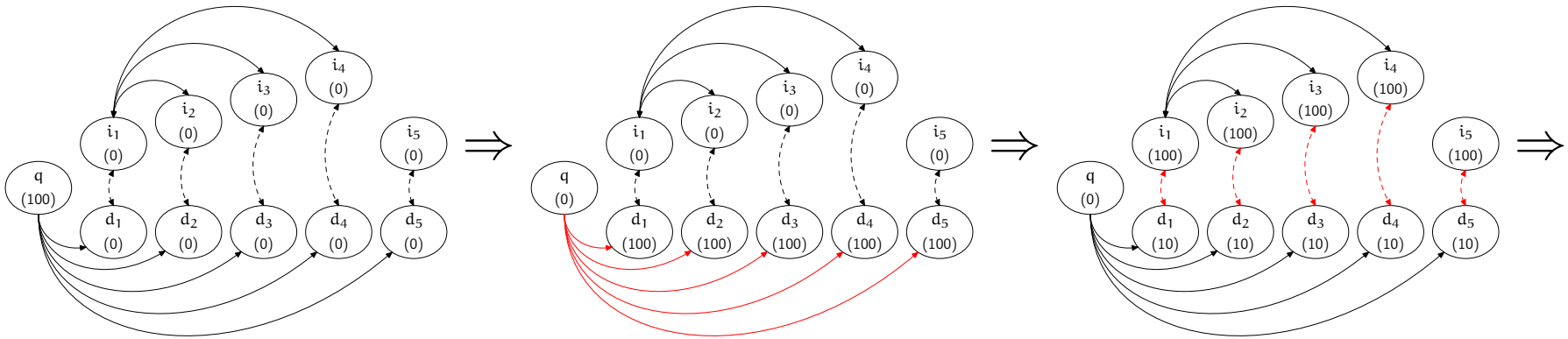


after 2nd pulse

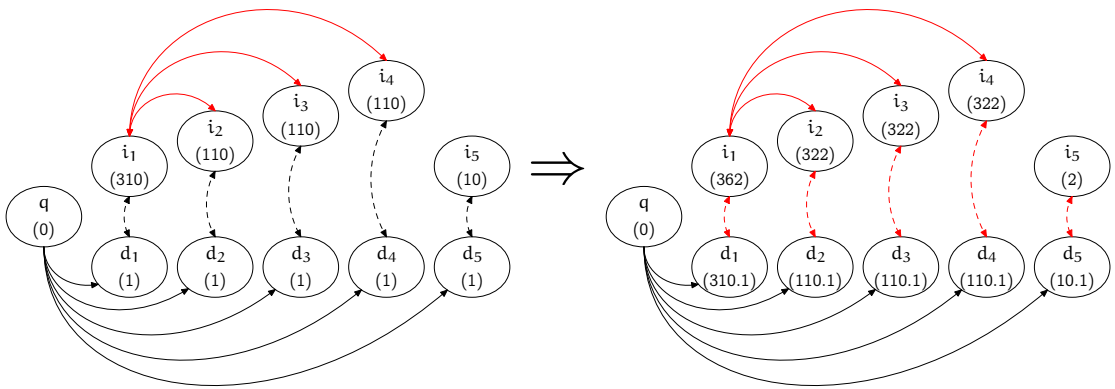


after 3rd pulse

Example search

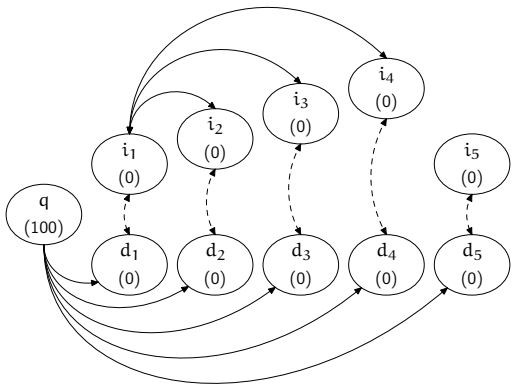


initial activation | after 1st pulse | after 2nd pulse

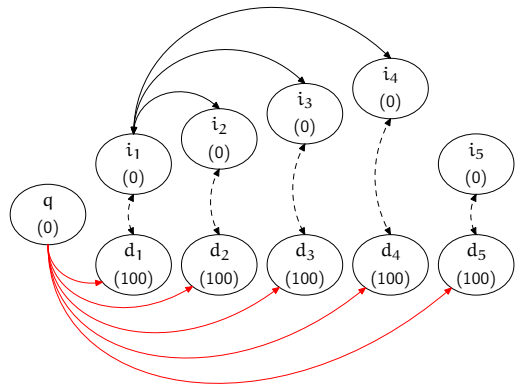


after 3rd pulse | after 4th pulse

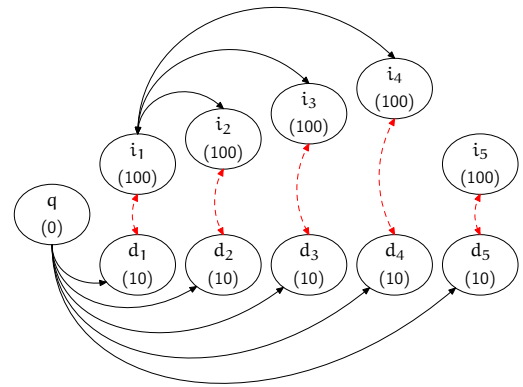
Example search



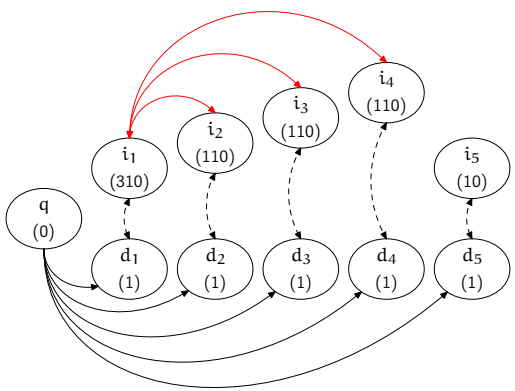
initial activation



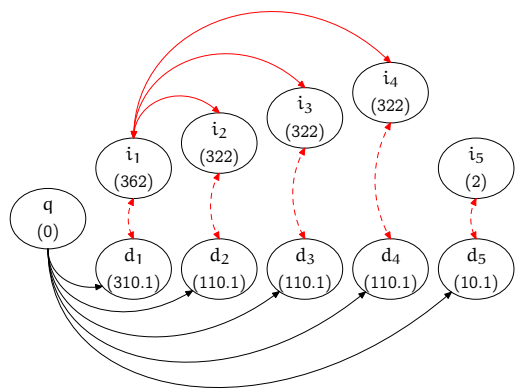
after 1st pulse



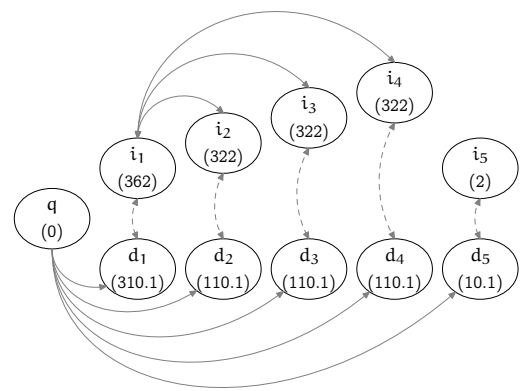
after 2nd pulse



after 3rd pulse



after 4th pulse



result

Evaluation

Note: Evaluation was performed on messages from 2004 only.

method:	VS	SA
	expert searcher:	
$\overline{\text{rank}}$:	14.75 ± 0.25	15.75 ± 0.05
$\overline{\text{rank}}$ change [%]:		$+6.8 \pm 2.1$
IAIR:	7.548 ± 0.032	3.814 ± 0.008
IAIR change [%]:		-49.5 ± 0.3
	novice searcher:	
$\overline{\text{rank}}$:	17.5 ± 0.3	12.4 ± 0
$\overline{\text{rank}}$ change [%]:		-29.1 ± 1.2
IAIR:	4.670 ± 0.013	3.831 ± 0
IAIR change [%]:		-18.0 ± 0.2

Outline

Motivation

Social networks

An Algorithm for social IR

Evaluation

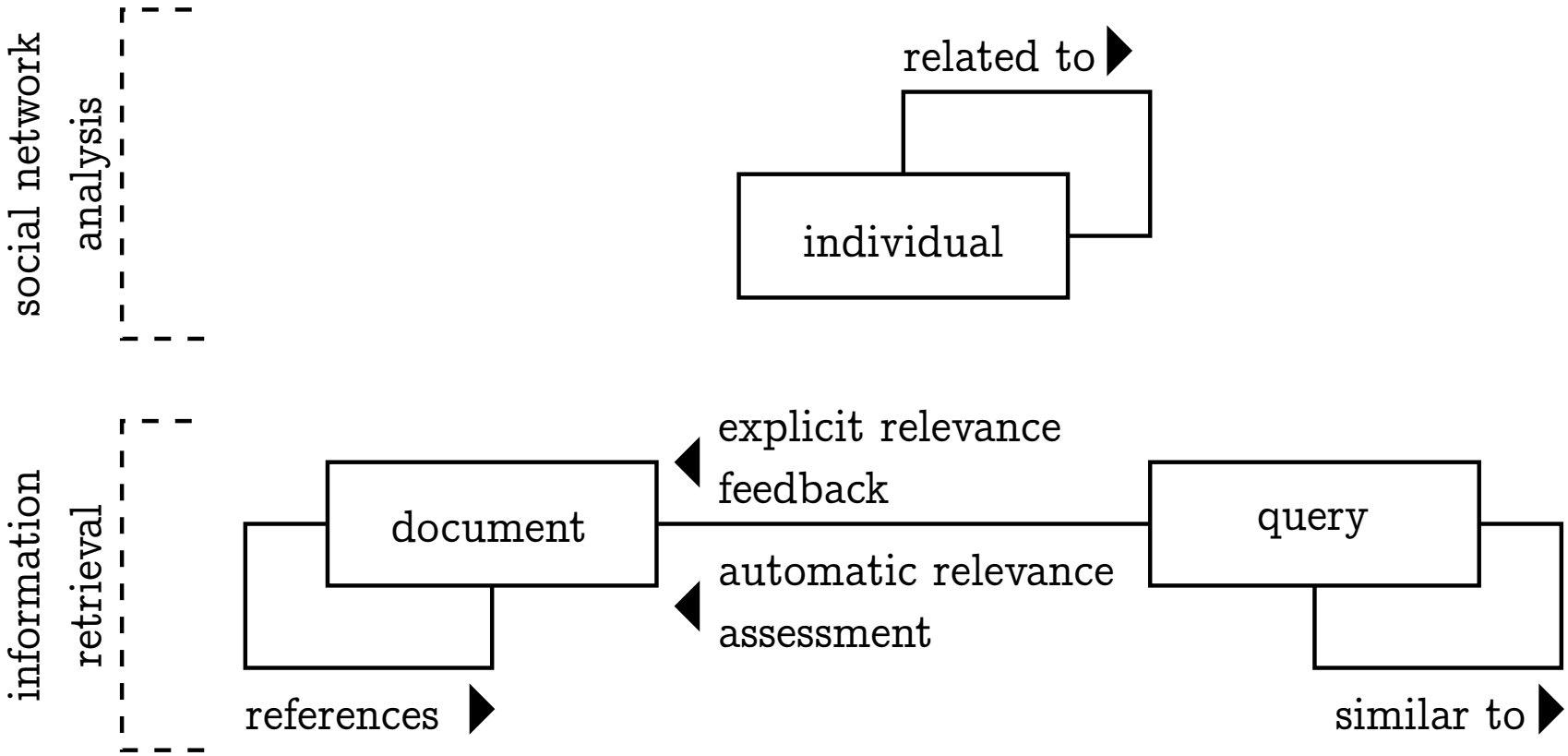
Second approach: Associative networks

A model for social IR

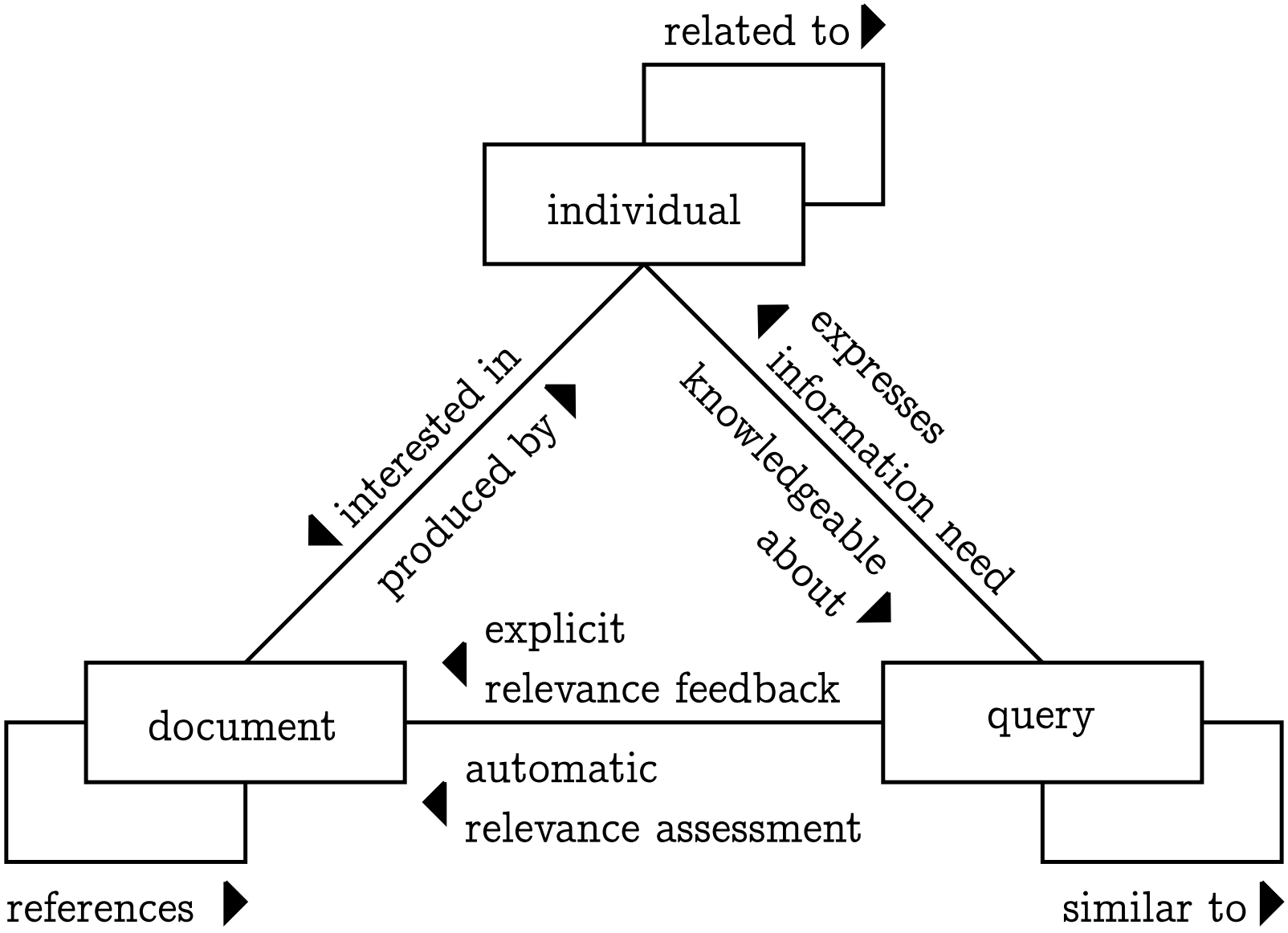
Additional work

Conclusion

Traditional domains



Domain model for social IR



Implications of the domain model

- Individuals appear in two roles: information producers and information consumers
- Queries and documents are essentially interchangeable (notion already present in the vector space model)
- queries and/or documents may be used to model an information need *or* an area of expertise.
- most systems will use only some of the relations in the model
- for a social IR systems, modelling relations between individuals is mandatory
- is it possible to design a unified IR system which makes use of all relations?

Outline

Motivation

Social networks

An Algorithm for social IR

Evaluation

Second approach: Associative networks

A model for social IR

Additional work

Conclusion

Additional work

- describe domain model for social IR and develop criteria for social IR systems
- prepare corpora (no standard corpora available)
- evaluate methods on second corpus (containing 25 years of SIGIR proceedings)
- examine statistical properties of social networks extracted from corpora
- implement prototype (in Java)

Outline

Motivation

Social networks

An Algorithm for social IR

Evaluation

Second approach: Associative networks

A model for social IR

Additional work

Conclusion

Conclusion

- social networks are an integral part of information retrieval
- social network analysis can lead to significant performance improvements
- rise of social software will necessitate retrieval algorithms using social networks
- my thesis contains a description of the problem domain and proposes two algorithms
- further research is necessary (esp. evaluation)

Questions? Feedback?

Thank you very much for listening!

slides for this talk are available at

`http://www.sebastian-kirsch.org/moebius/docs/socialir-slides.pdf`

Social Information Retrieval

Sebastian Marius Kirsch
kirschs@informatik.uni-bonn.de

25th November 2005


Ricardo Baeza-Yates and Berthier Ribeiro-Neto.
Modern Information Retrieval.
Addison-Wesley, 1999.


Melanie Gnasa, Markus Won, and Armin B. Cremers.
Three pillars for congenial web search. Continuous evaluation for enhancing web search effectiveness.
Journal of Web Engineering, 3(3&4):252–280, 2004.


Stanley Milgram.
The small-world problem.
Psychology Today, 2:60–67, 1967.


Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd.
The PageRank citation ranking: Bringing order to the Web.
Technical report, Stanford University, November 1999.