Web Search Personalization via Social Bookmarking and Tagging

by Michael G. Noll
Christoph Meinel

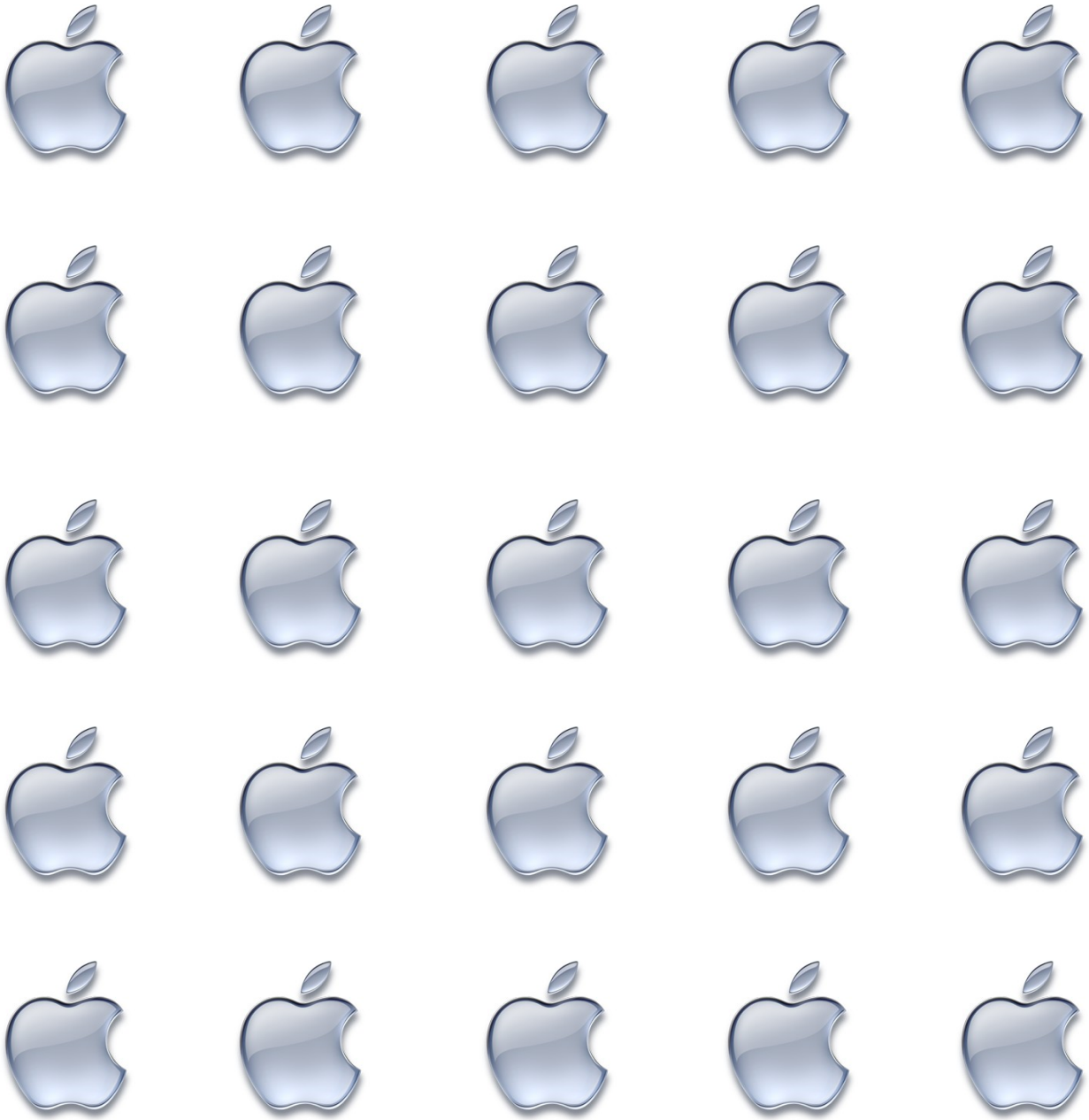
6th International Semantic Web Conference, Busan, South Korea, Nov 2007

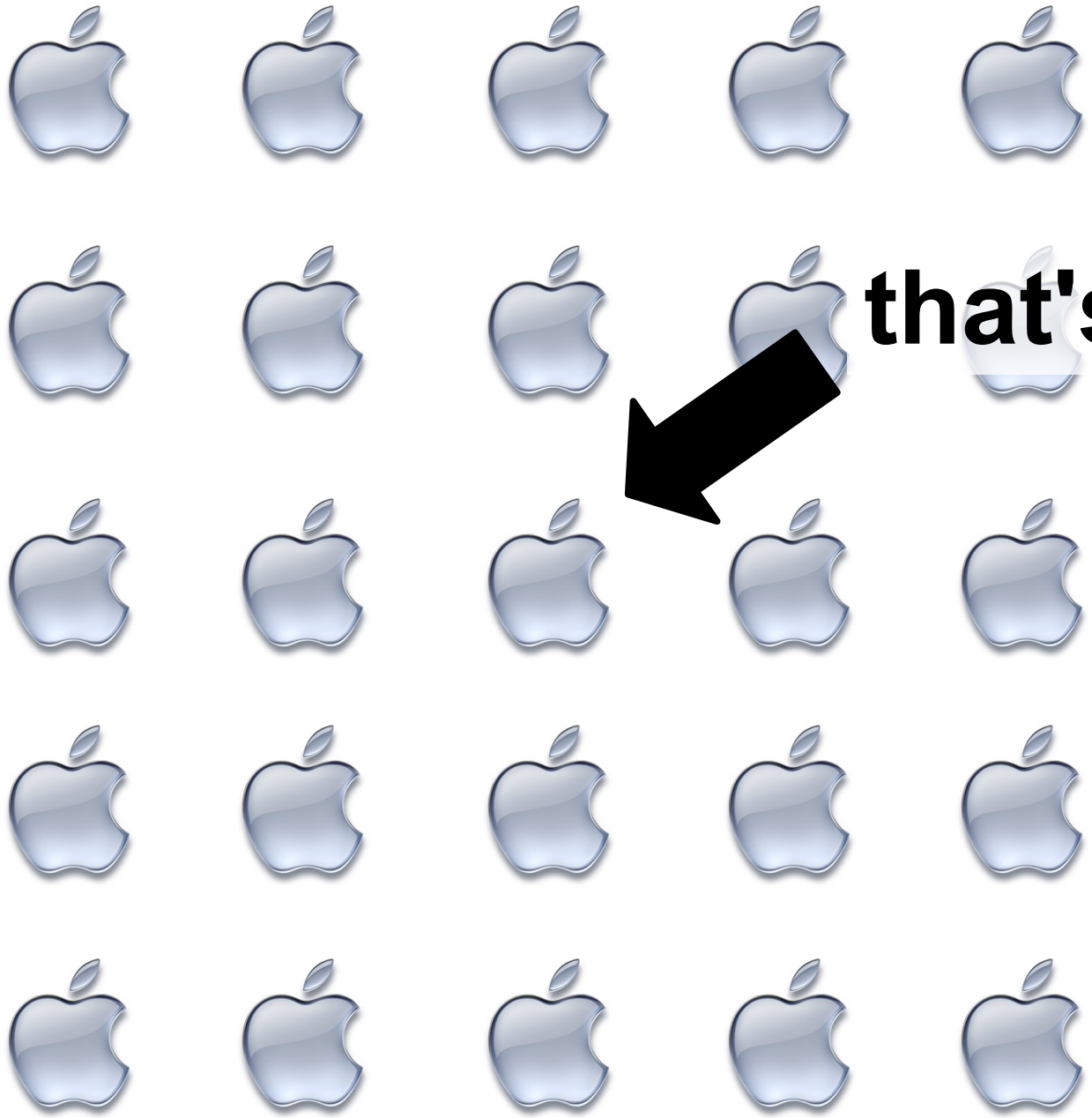
Overview

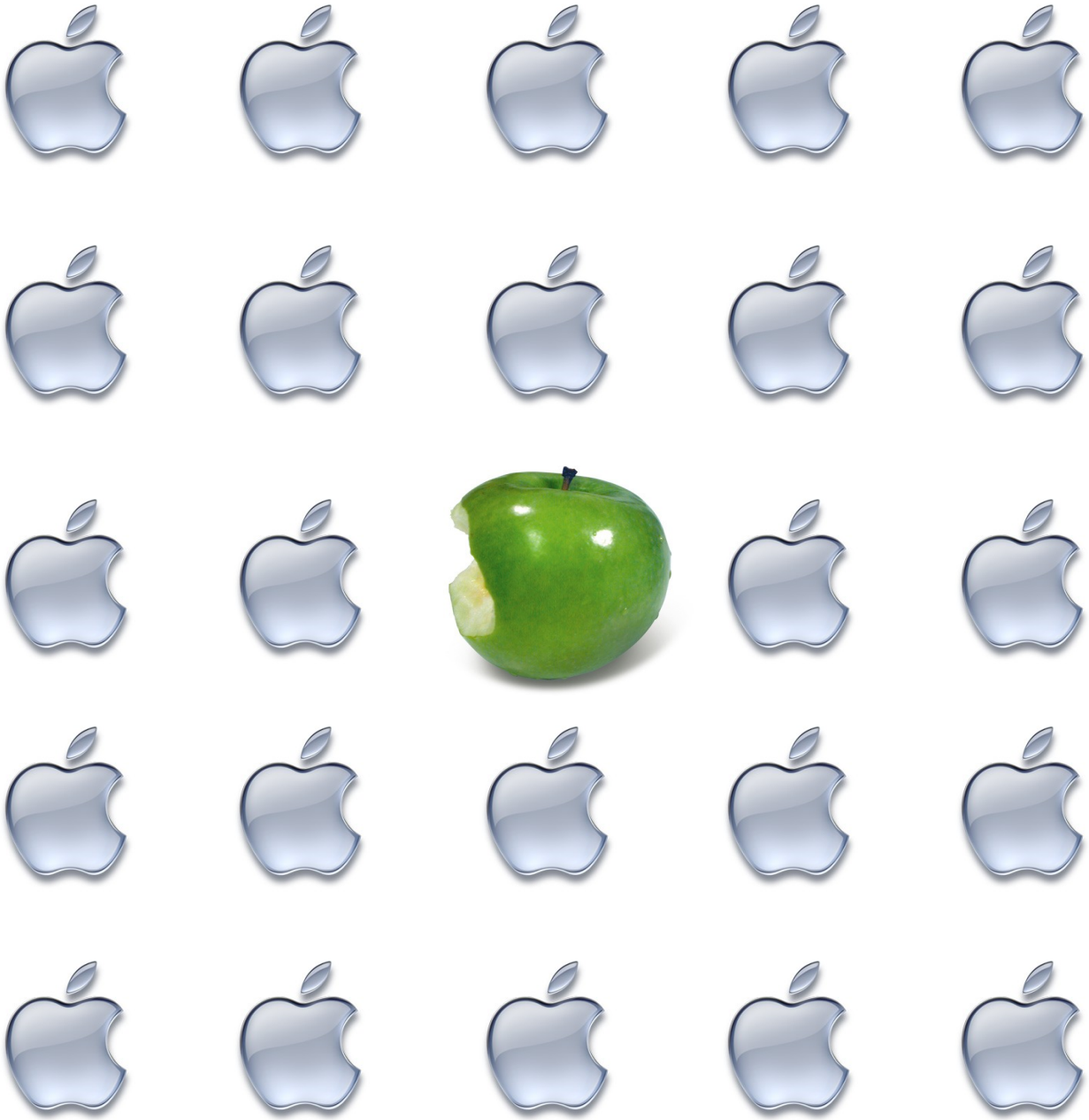
- Introduction
- Personalization “2.0”
- Experiments and evaluation
- Conclusion

Introduction

1. “Web Search Personalization”
2. “via Social Bookmarking and Tagging”




that's you!



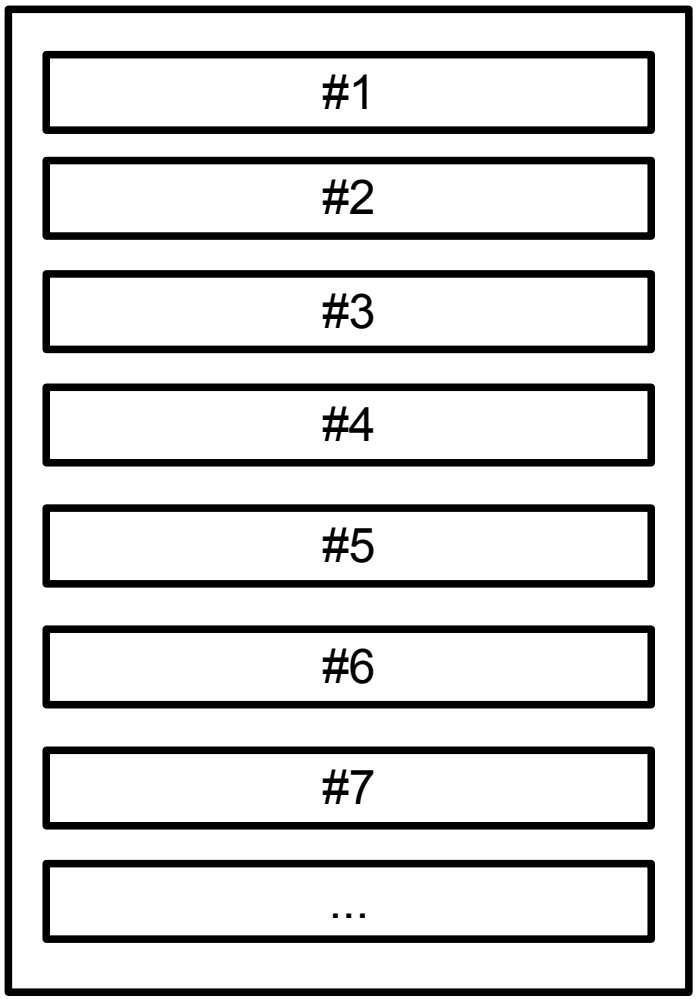
Web search personalization

- integration of user-specific data to improve results
  (and advertising, revenue...)
- two main approaches:
  1. modification of user's query: “nyt” > “new york times”
  2. re-rank search results based on user profile

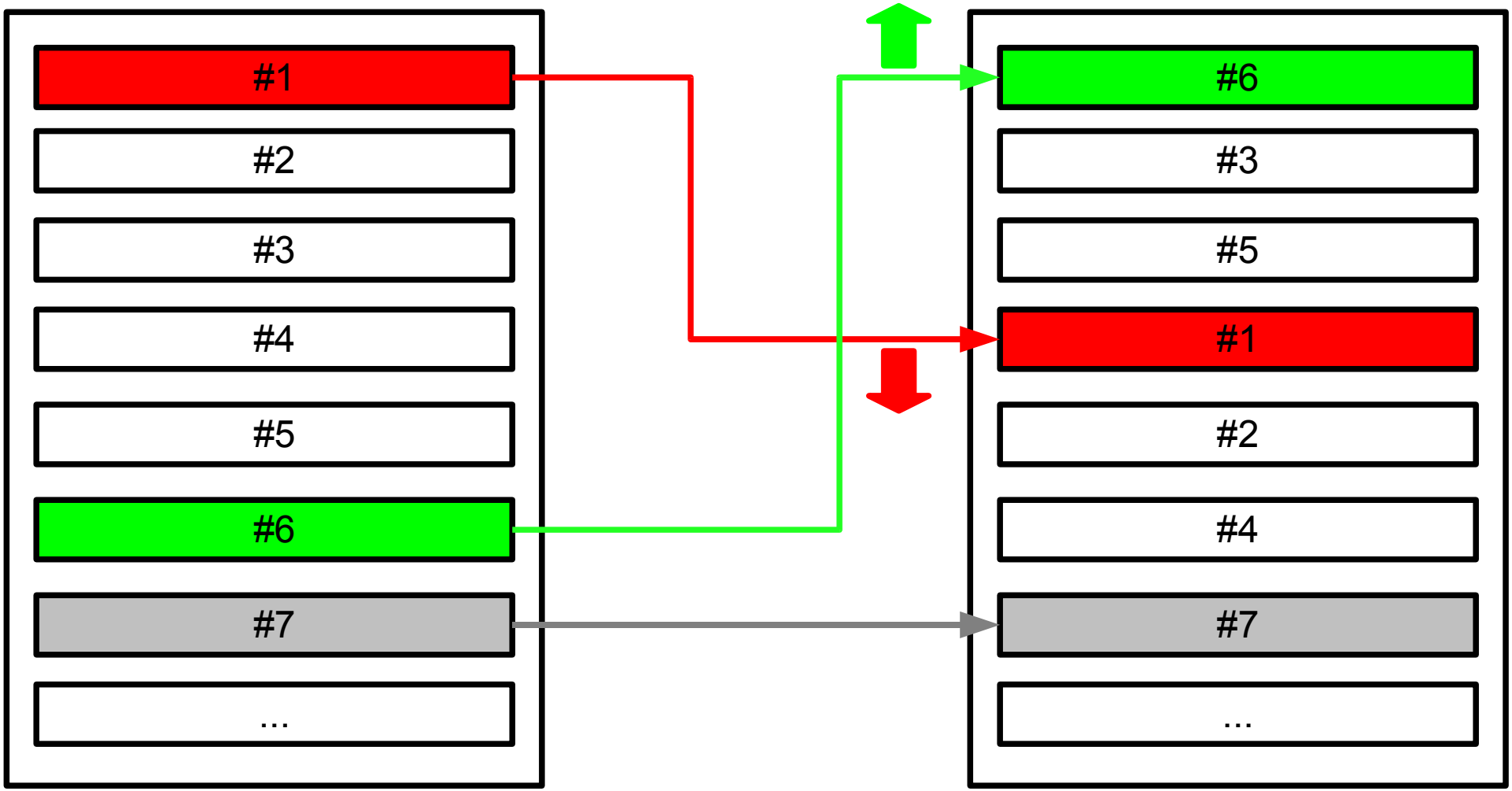
Web search personalization

- integration of user-specific data to improve results
  (and advertising, revenue...)
- two main approaches:
  1. modification of user's query: “nyt” > “new york times”
  2. re-rank search results based on user profile

Web search personalization



Web search personalization



Social bookmarking and tagging

- social bookmarking:
  publicly sharing your bookmarks with others
  (note: social component increases incentive to add metadata)
- tagging / folksonomies:
  Users annotate Documents with with a flat,
  unstructured list of keywords called *Tags*

$$R \subseteq D \times U \times T$$

Personalization via social annotations

Overview

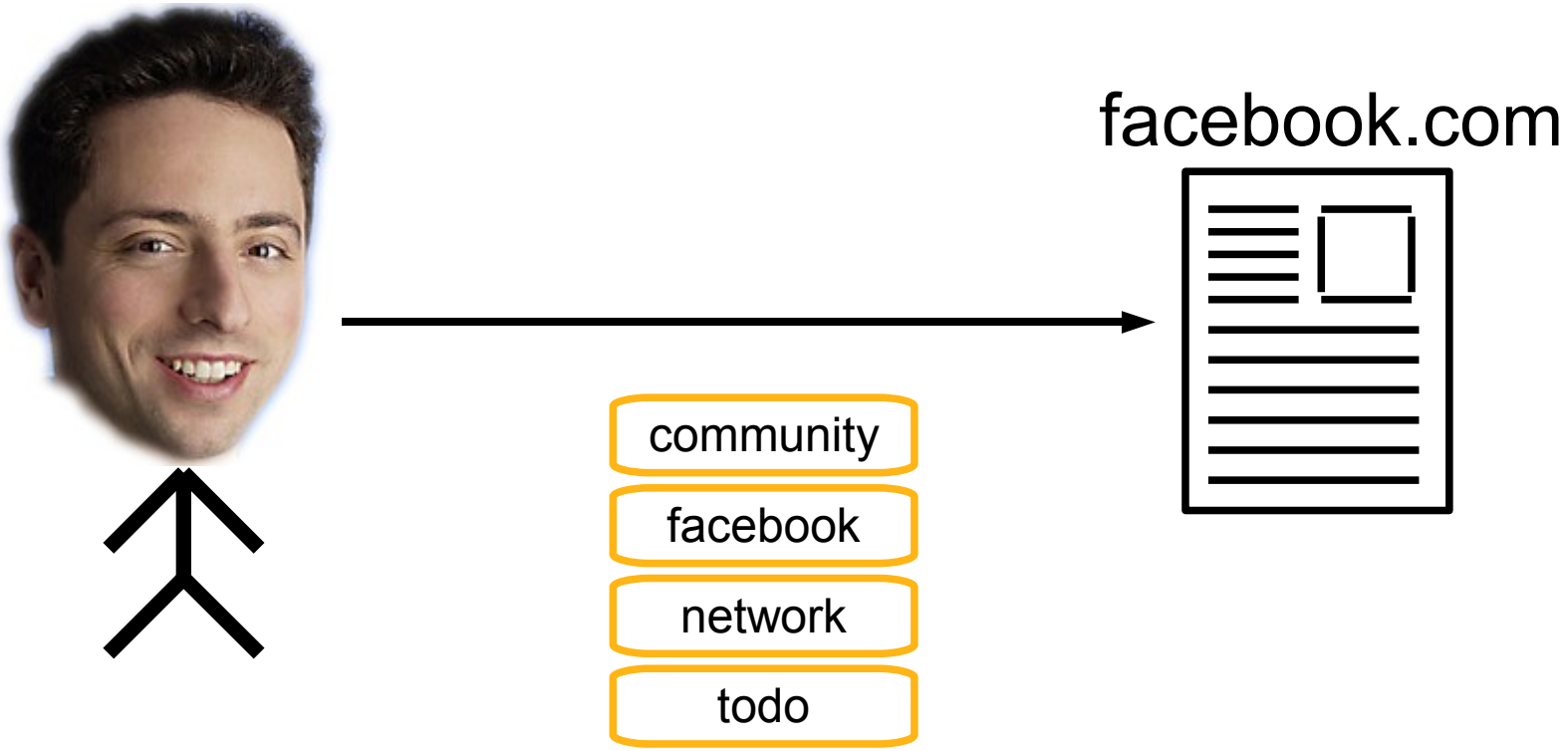
- exploit conceptual links between **web search**, **social bookmarking** and **tagging**
- personalization driven by human users
- separate data collection from personalized information systems – here: search engines
  - no need to give your personal data to Yahoo & Co. (sorry!)
- approach is independent of search engines
  - “semantic overlay on Internet search”, “sitting on (top of) Google”

How it works

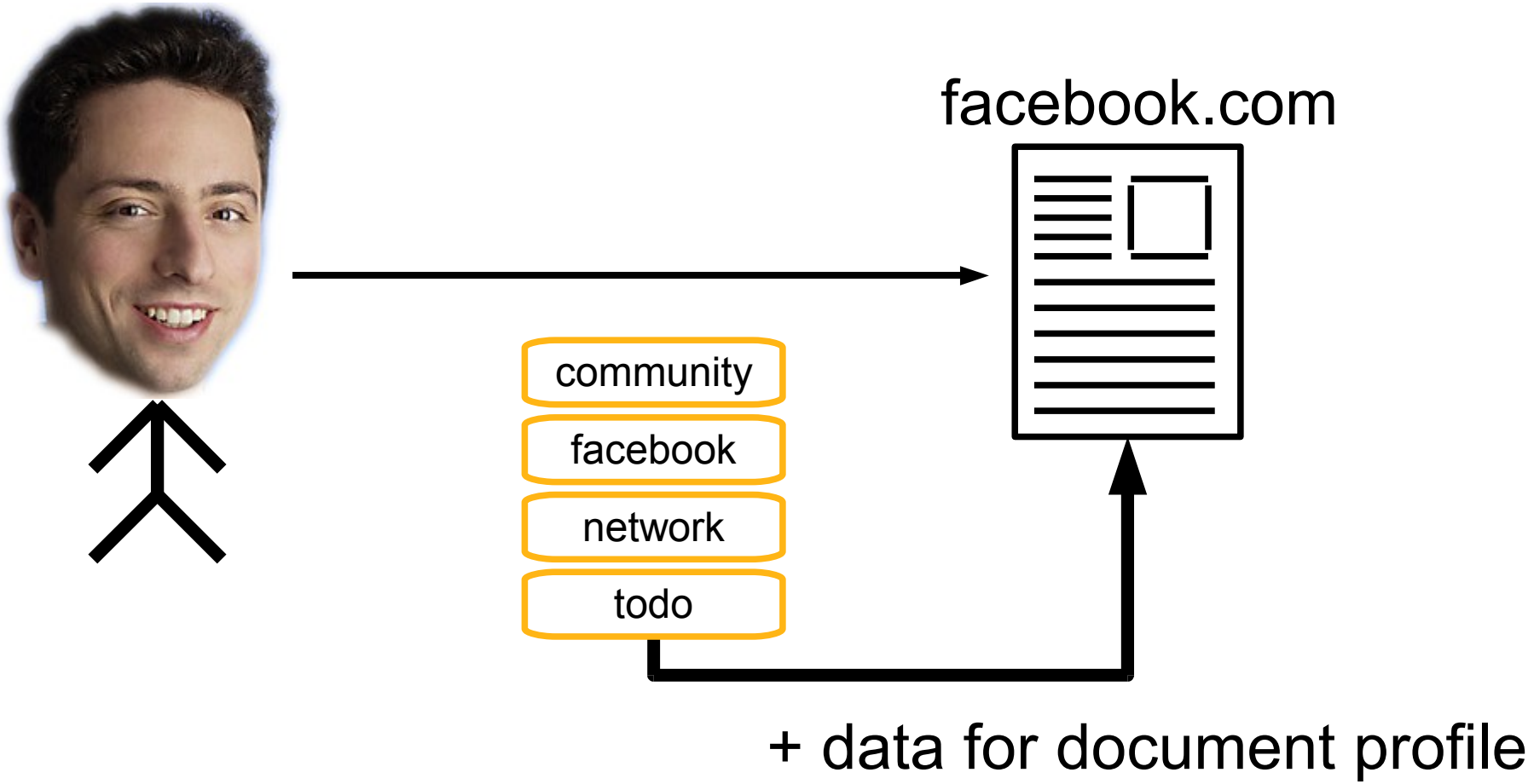
1. collect metadata about users and documents from social bookmarking and tagging
2. build user profiles and document profiles
3. calculate user-document similarity
4. re-rank search results
5. cross fingers!

1. Data Collection

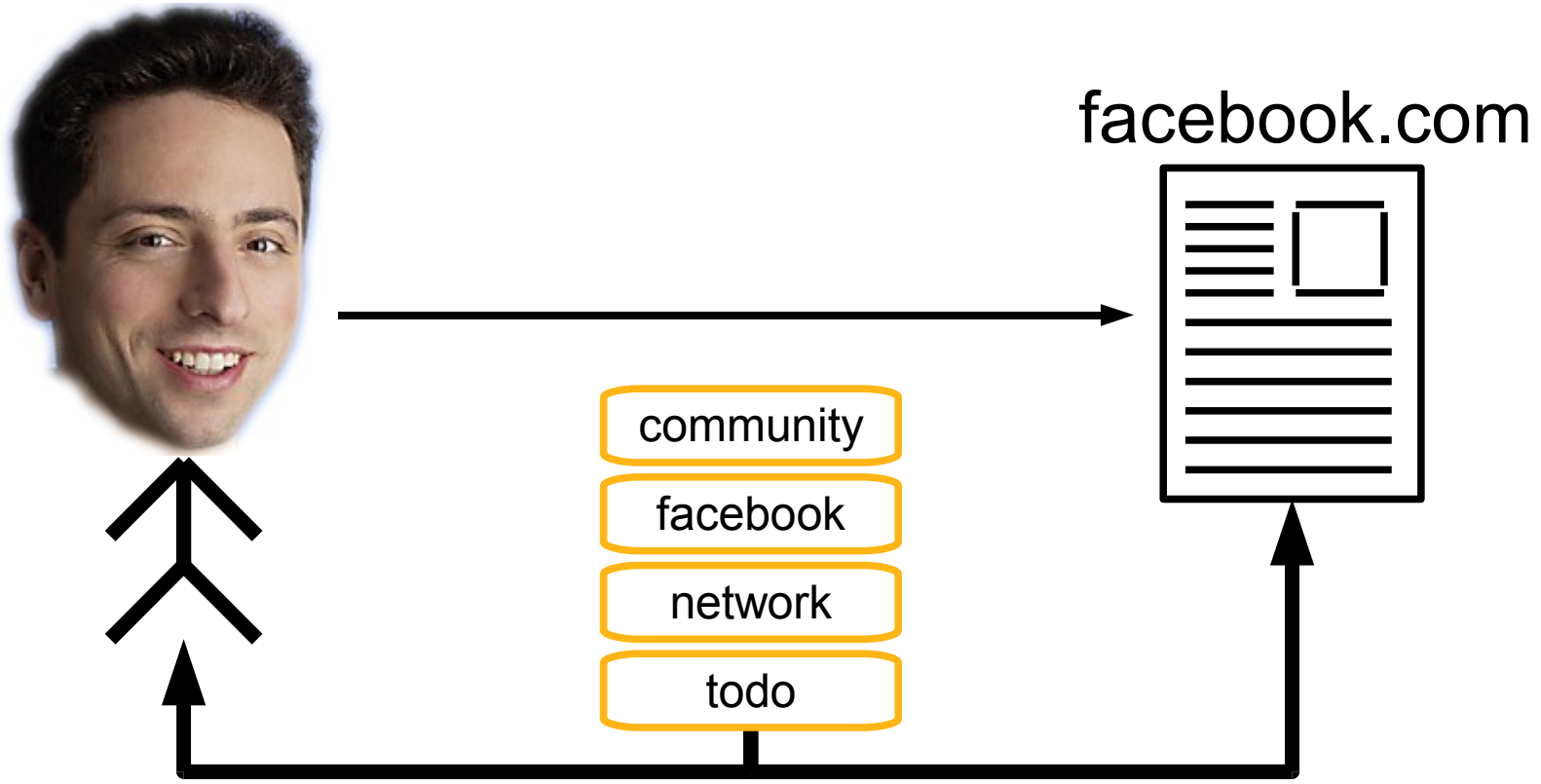
Data Collection



Data Collection



Data Collection



+ data for user profile

+ data for document profile

2. Data Aggregation

User profile

- user's bookmark collection:
  tag-document matrix with *m* tags and *n* docs

$$M_d = \begin{bmatrix} c_{11} & \dots & c_{1n} \\ \dots & \dots & \dots \\ c_{m1} & \dots & c_{mn} \end{bmatrix}, c_{ij} \in \{0,1\}$$

- bookmarks are column vectors
- c_{ij} = 1 if tag t_i is assigned to document d_j

User profile

- user profile: vector with *m* tags

$$p_u \stackrel{\text{def}}{=} M_d \cdot \omega_d = \begin{bmatrix} c_1^* \\ .. \\ c_m^* \end{bmatrix}, c_i^* \in N_0$$

- in our implementation, weight vector

$$\omega_d{}^T \stackrel{\text{def}}{=} 1^T = [1 \ldots 1]$$

= equal importance to all *n* documents

Document profile

- analogue to user profile - cool!

$$p_d \stackrel{\text{def}}{=} M_u \cdot \omega_u = \begin{bmatrix} c_1^* \\ .. \\ c_m^* \end{bmatrix}, c_i^* \in N_0$$

- weight gives equal importance to all users

Profile examples

User jsmith	
„open source“	13
„programming“	19
„proprietary“	2
„research“	10
„security“	21
„semantic web“	34

http://iswc.semanticweb.org/	
„iswc“	156
„computing“	48
„programming“	66
„conference“	90
„research“	111
„semantic web“	140

3. Similarity

Similarity

- user-document similarity is:
  - dimension-less score
  - used for relative weighting and re-ranking of documents within a given search result list

$$similarity(user, document) \stackrel{\text{def}}{=} p_u^T \cdot \| p_d \|$$

Similarity

- user-document similarity is:
  - dimension-less score
  - used for relative weighting and re-ranking of documents within a given search result list

$$similarity(user, document) \stackrel{\text{def}}{=} p_u^T \cdot \| p_d \|$$

- naïve “normalization” of document profile simply sets all non-zero components down to 1:
  => user profile as key factor for personalization

Similarity example

User jsmith	
„open source“	13
„programming“	19
„proprietary“	2
„research“	10
„security“	21
„semantic web“	34

http://iswc.semanticweb.org/	
„iswc“	156
„computing“	48
„programming“	66
„conference“	90
„research“	111
„semantic web“	140

Similarity example

User jsmith	
„open source“	13
„programming“	19
„proprietary“	2
„research“	10
„security“	21
„semantic web“	34

http://iswc.semanticweb.org/	
„iswc“	156
„computing“	48
„programming“	66
„conference“	90
„research“	111
„semantic web“	140

similarity (“jsmith”, “http://iswc...”) = 63

Similarity

- similarity score properties:
  - favors documents with tags that are applied frequently by the user himself
  - promotes known*, similar documents and demotes non-similar or unknown documents
  - score of 0 (zero) for unknown documents (!)
  - most critical factor in practice:
    “do we have sufficient data to make all this work?”

*“known” = bookmarked and tagged by users

4. Personalization

Personalization

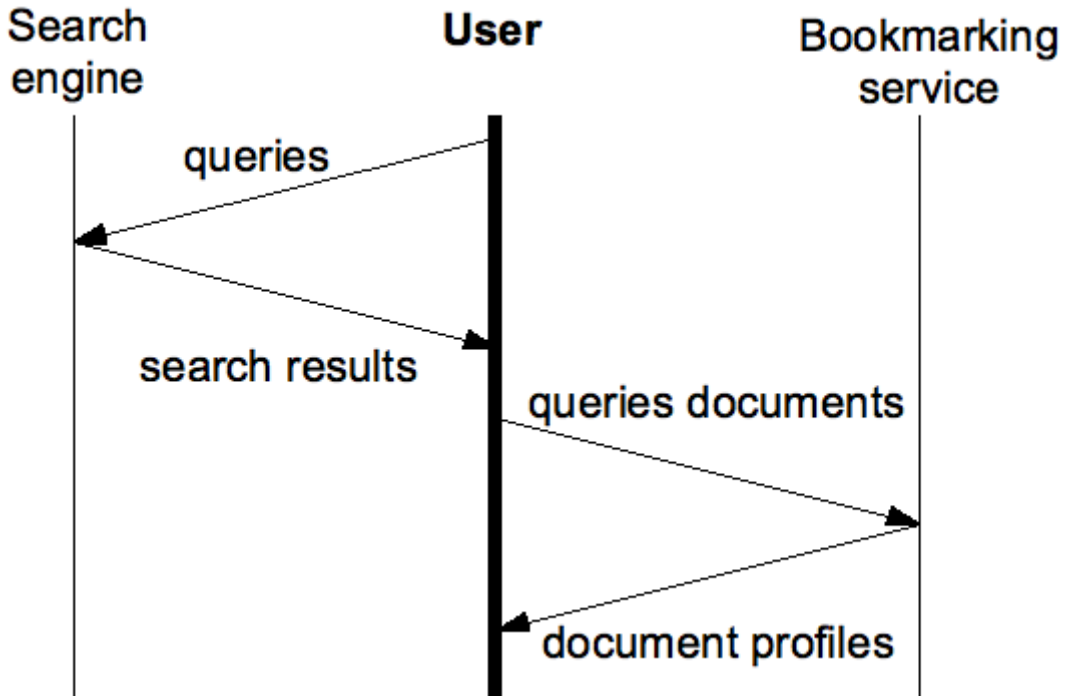
- input:
  user profile + ordered list of *n* document profiles
- algorithm:
  - calculate *similarity(user, document)* for all docs
  - sort documents by similarity from highest to lowest
- output:
  re-ranked search result list

Personalization

- system setup
  - server: social bookmarking service
  - client: browser add-on
- implements all the previously described stuff
- modification of search engine UI by updating the DOM tree of the search result pages in real-time

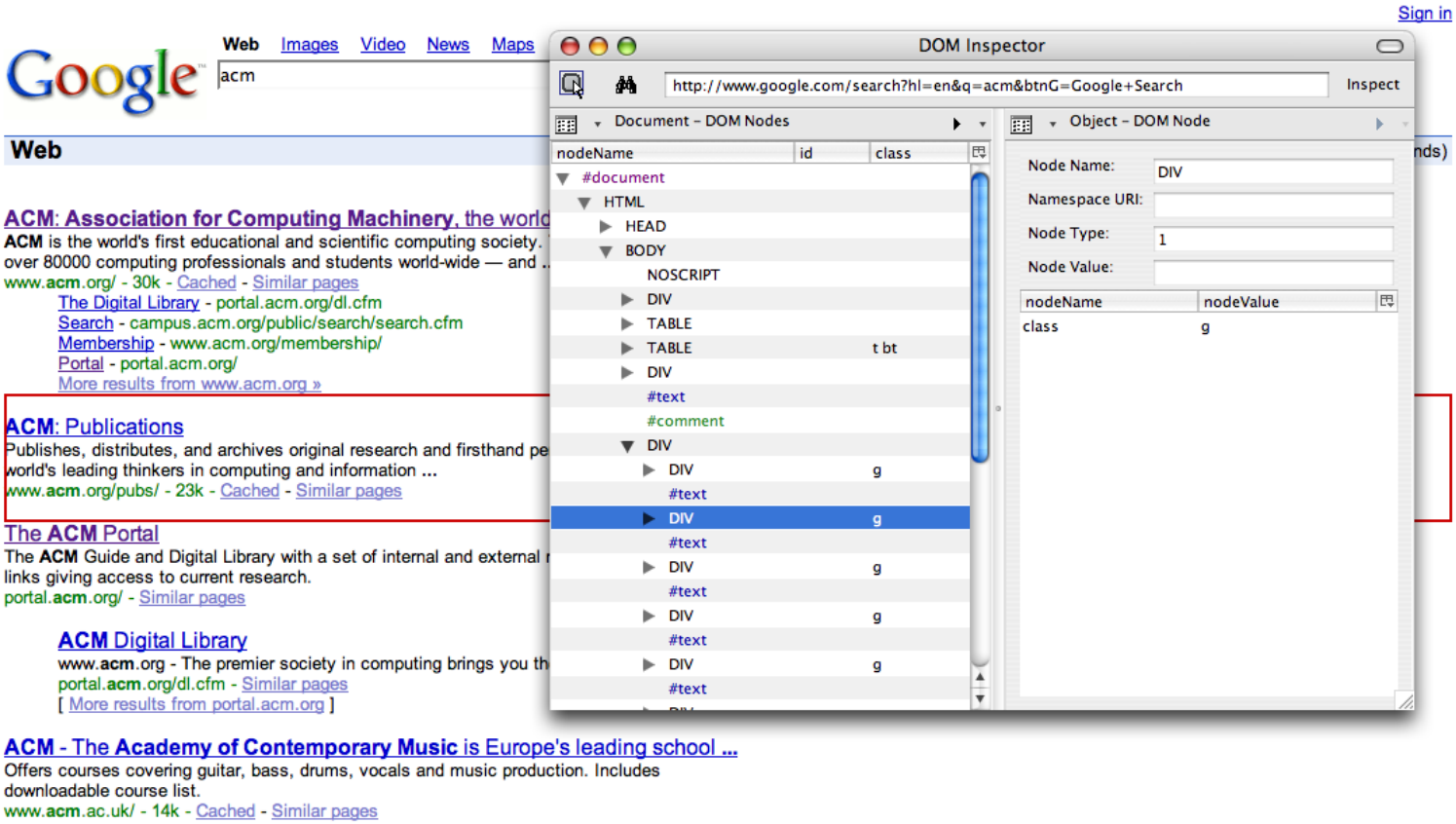
Personalization

- communication flow



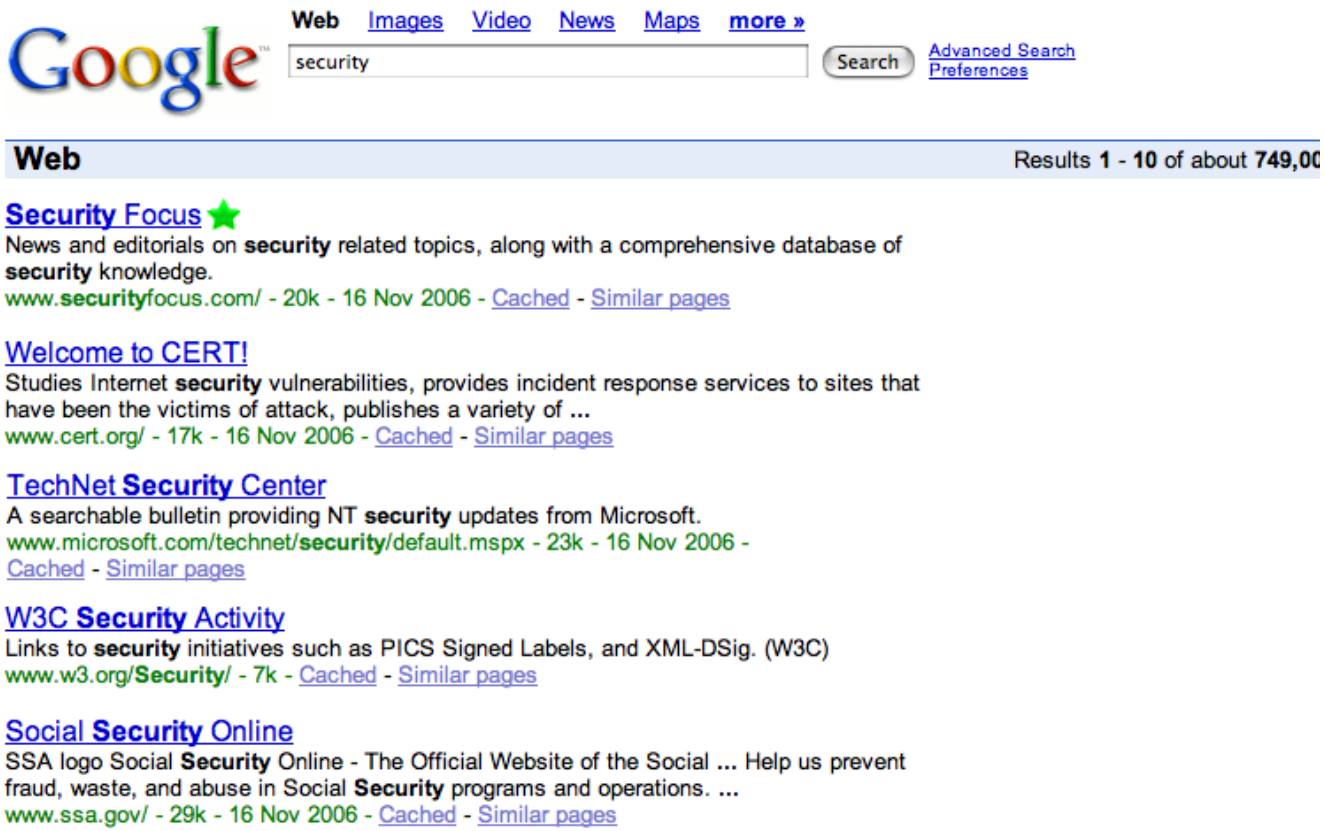
everything done in real-time:
in practice, users do not notice the personalization overhead

Personalization



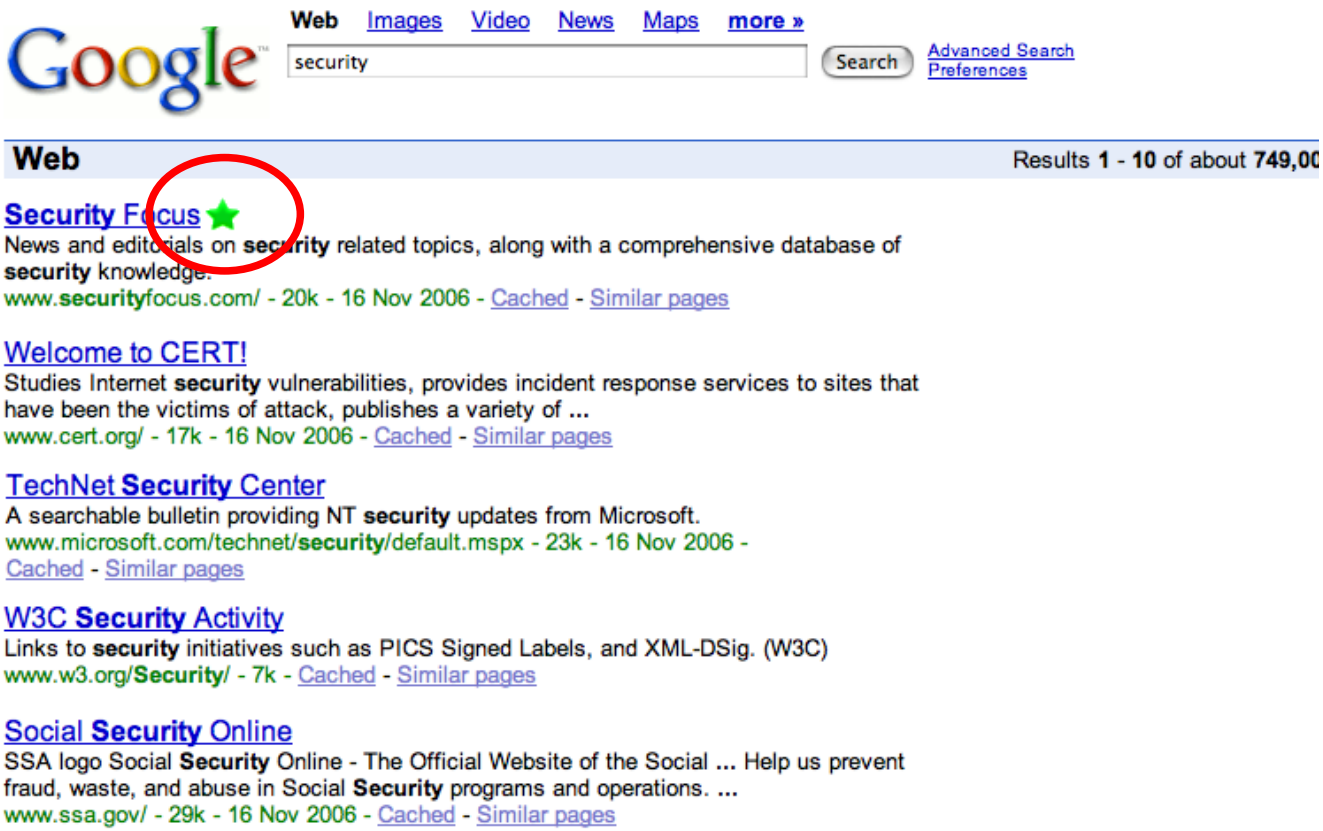
DOM tree of Google search result page

Personalization



personalization is transparent to the user

Personalization



personalization is transparent to the user

Experiments and Evaluation

Evaluation

- quantitative analysis:
  “critical mass of social annotations in practice?”

- qualitative analysis:
  “if so, how good is the personalization?”

Evaluation

key question!

- quantitative analysis:
  “critical mass of social annotations in practice?”

- qualitative analysis:
  “if so, how good is the personalization?”

Quantitative analysis

- DMOZ100k06 [Noll and Meinel, 2007]:
  - random sample of 100,000 web documents with social bookmarking and tagging data
    + Google PageRank for document popularity
- AOL500k [AOL research, 2006]:
  - subset of full corpus, giving us:
    1,750,000 web searches by AOL users with
    1,000,000 clicked search results

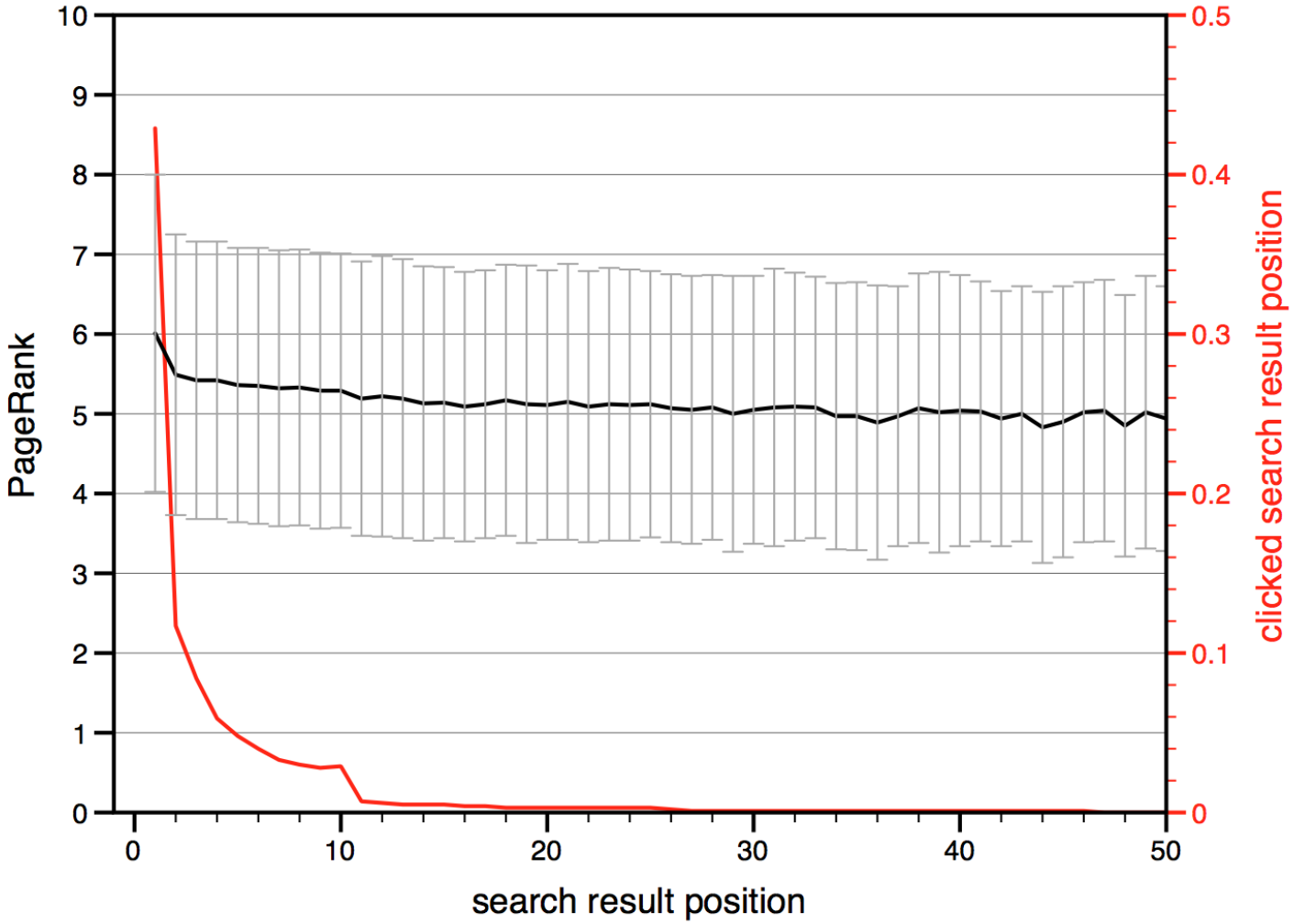
Quantitative analysis

- background
  - previous work: positive correlation of #{bookmarks, tags} and document popularity
  - “the more popular, the more bookmarks and tags”

- documents:
  analyze popularity of web documents and user click frequency for each search result position

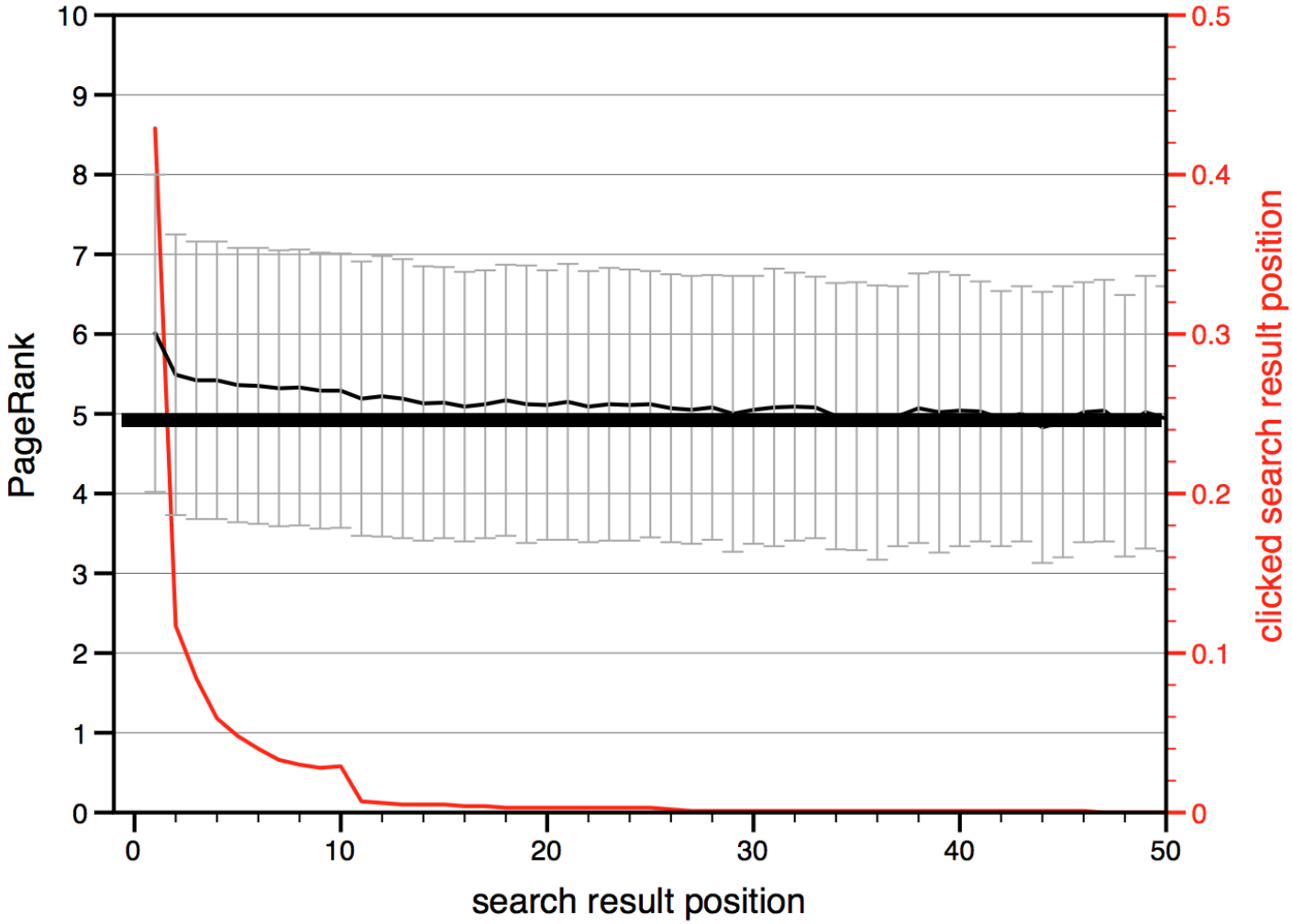
Quantitative analysis

- avg PageRank: 5 – 6 (of 10)
- sufficiently high!
- top 5 docs: 75% of all clicks
- top 10 docs: almost 100%
- first search result page is enough!



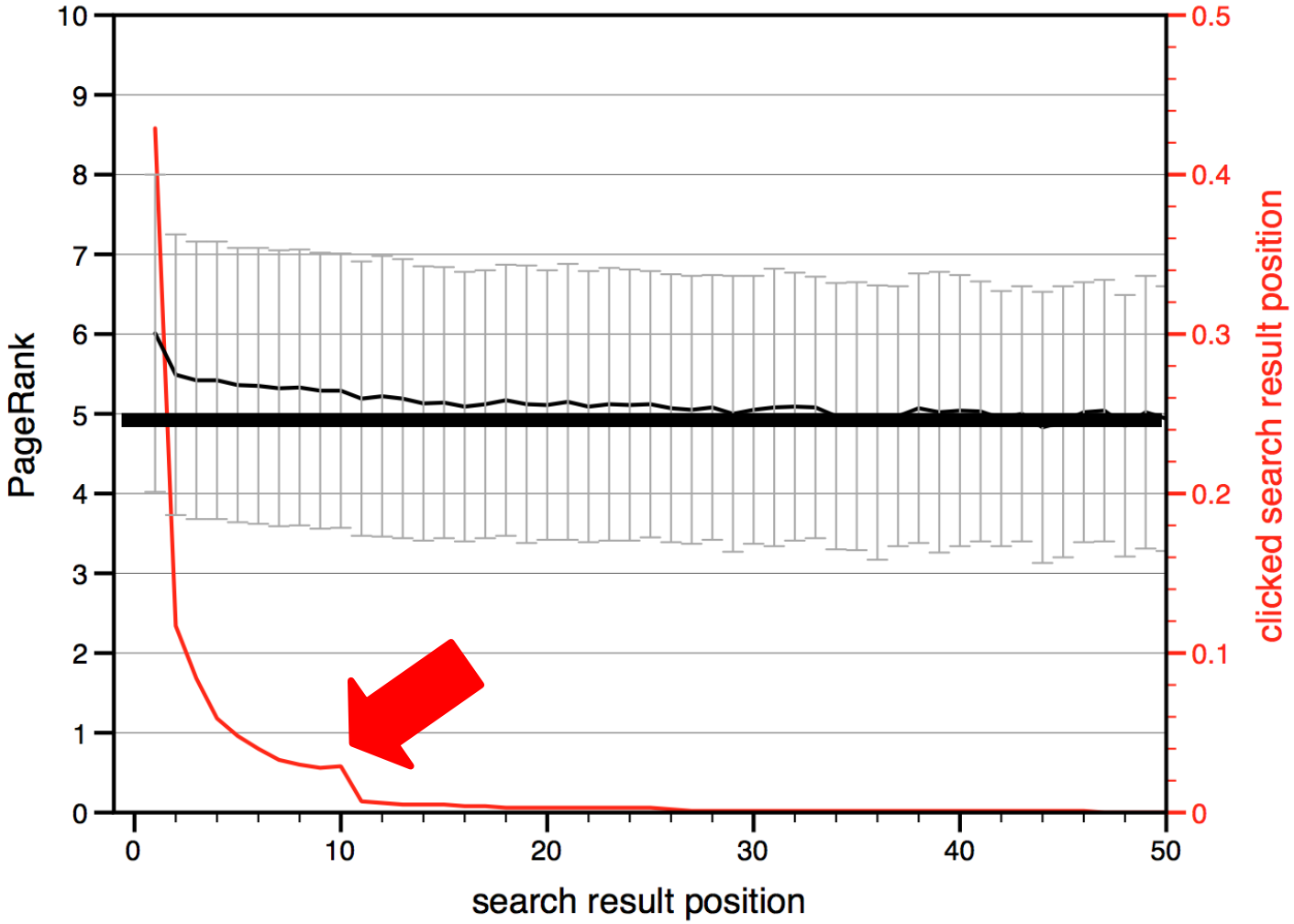
Quantitative analysis

- avg PageRank: 5 – 6 (of 10)
- sufficiently high!
- top 5 docs: 75% of all clicks
- top 10 docs: almost 100%
- first search result page is enough!



Quantitative analysis

- avg PageRank: 5 – 6 (of 10)
- sufficiently high!
- top 5 docs: 75% of all clicks
- top 10 docs: almost 100%
- first search result page is enough!



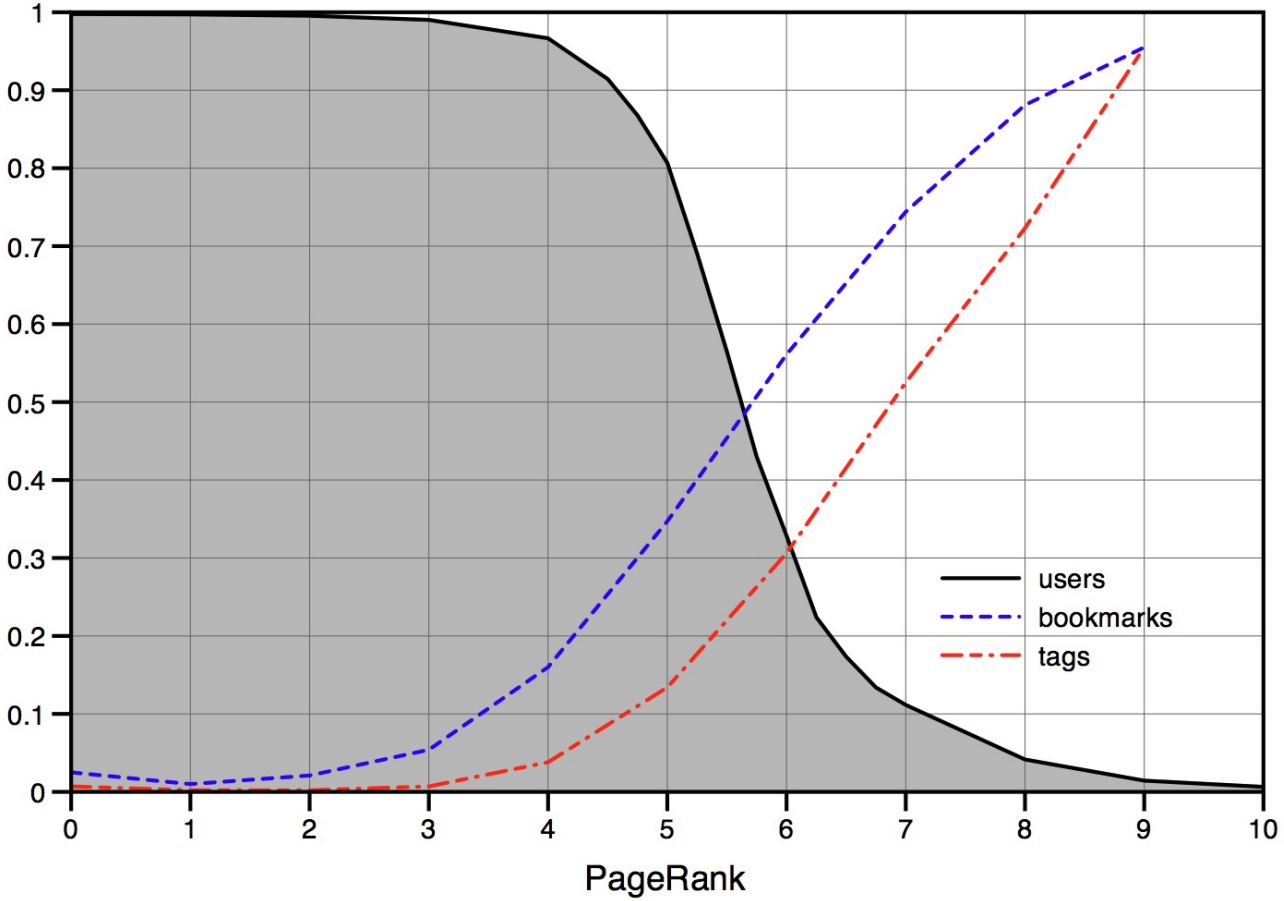
Quantitative analysis

- users:
  analyze popularity of clicked search results
  for each user in the data set

  = individual click preferences *regardless* of
  a document's search result position

Quantitative analysis

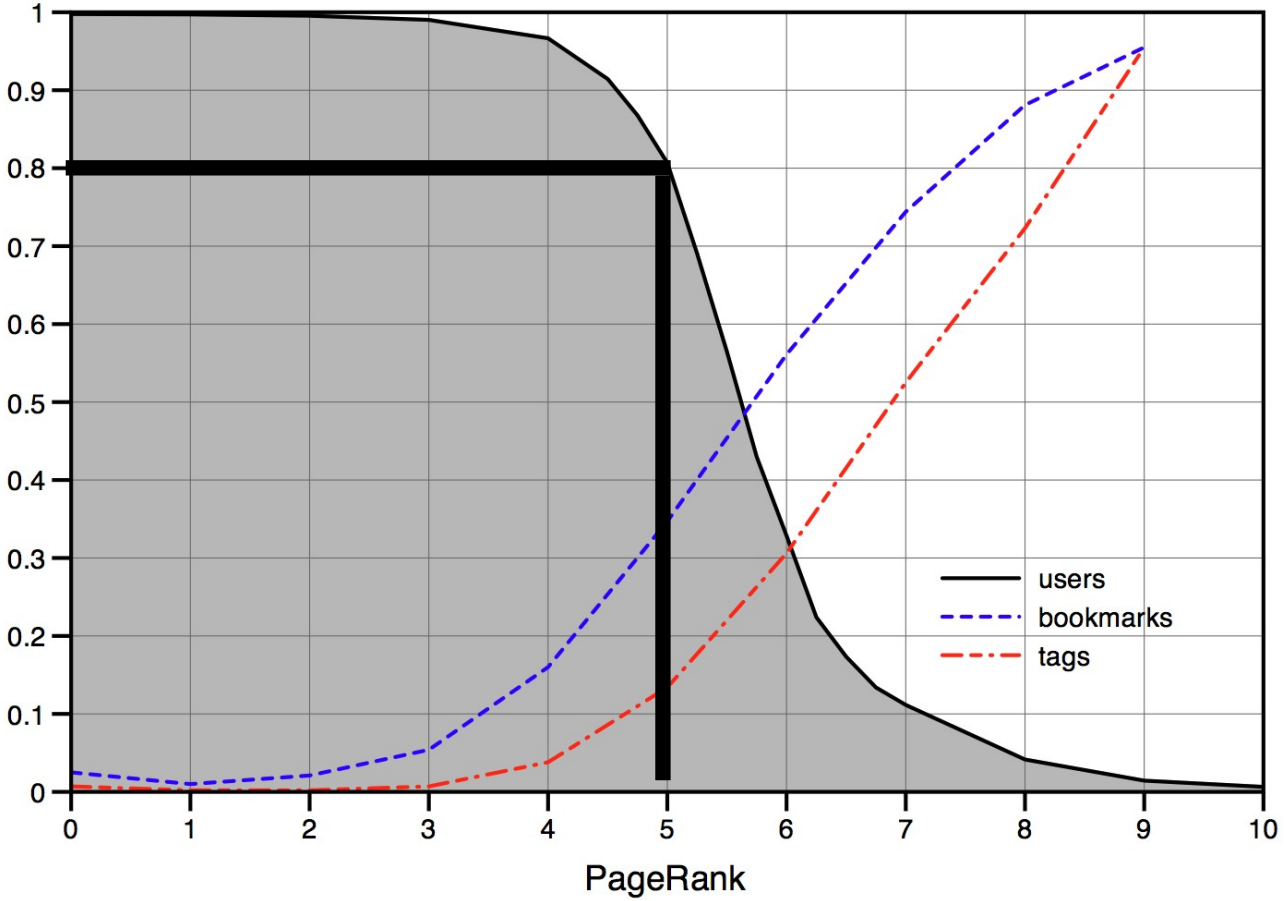
- 80% of users with PageRank >= 5
- 33% of users with PageRank >= 6
- *combined* probability of *n* docs to be bookmarked or tagged is high enough in practice!



percentage of users with average PageRank of *x* or higher

Quantitative analysis

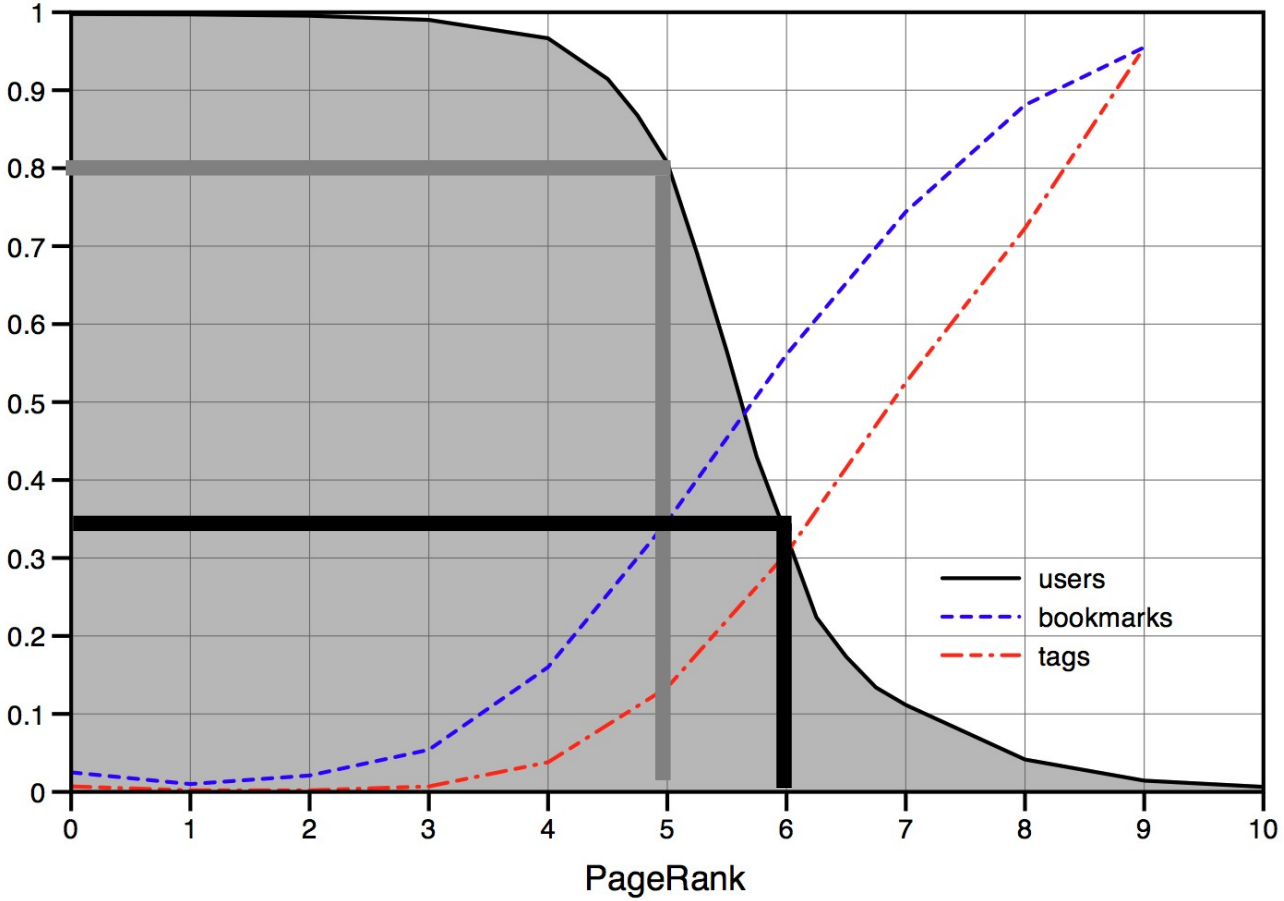
- 80% of users with PageRank >= 5
- 33% of users with PageRank >= 6
- *combined* probability of *n* docs to be bookmarked or tagged is high enough in practice!



percentage of users with average PageRank of *x* or higher

Quantitative analysis

- 80% of users with PageRank >= 5
- 33% of users with PageRank >= 6
- *combined* probability of *n* docs to be bookmarked or tagged is high enough in practice!



percentage of users with average PageRank of *x* or higher

Quantitative analysis

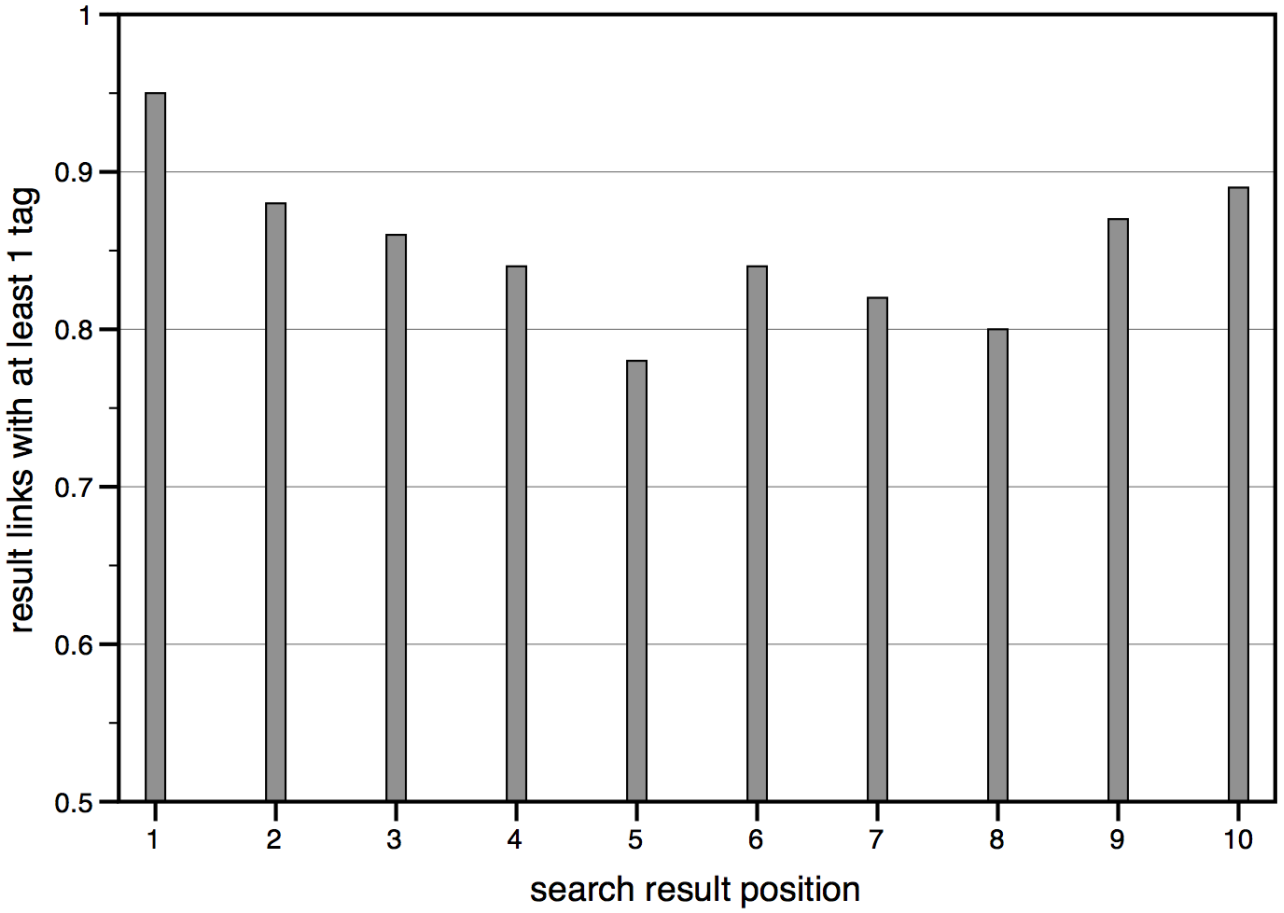
- search queries for “popular tags” of the social bookmarking service del.icio.us (> 1M users)
- idea: upper bound for personalization approach?

Quantitative analysis

- test set
  - 140 “popular tags”
  - 1400 search queries
- totaling
  - 981,989 bookmarks
  - 20,498 tag annotations
  - 2,300 unique tags

Quantitative analysis

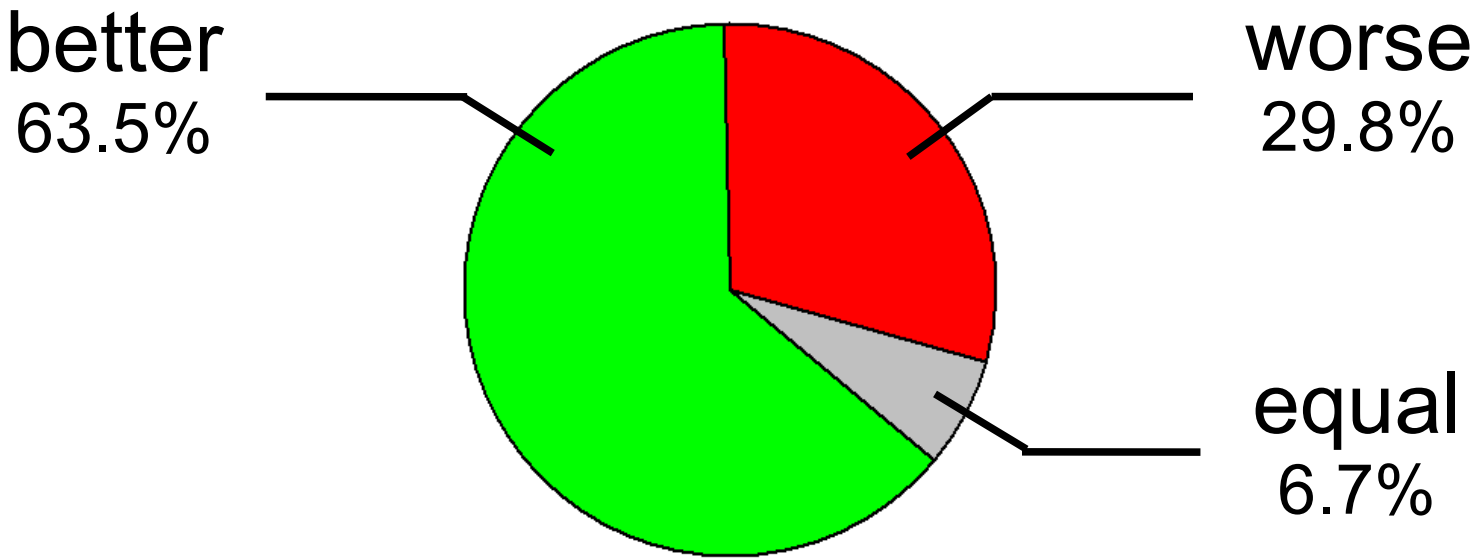
- we can expect to personalize approx. 85% of the search result documents in this scenario



Qualitative analysis

- user study:
  participants evaluate top 10 search results, i.e. first result page, for 13 search queries each
- blind test: direct comparison of unmodified vs. personalized result list => user picks better one
- N = 8
- total queries = 104
- total documents = 1040

Qualitative analysis



- personalization *better or as good* in 70% of queries
- interestingly low percentage of “equal” results

Conclusion

- will not repeat results from previous slides :-)
- proposed personalization approach is feasible and viable in practice:
  - already sufficient user-supplied metadata available
  - initial evaluation of personalization quality shows very promising results
- Open Access on steroids
  - http://www.michael-noll.com/dmoz100k06/ - data set
  - http://www.michael-noll.com/delicious-api/ - scripts

Future work

- “proof of concept” – we're at the start
- synonyms, ambiguity, emergent semantics, <insert your favorite topic of last days here>
- compliment with other personalization techniques – strength & weaknesses?
- more evaluation
- more playing around