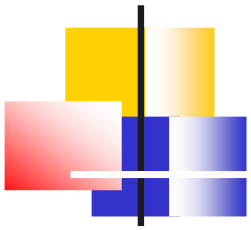


Personalized Search Based on User Search Histories

Mirco Speretta

Master's Thesis Defense

June 17th, 2005.

Committee

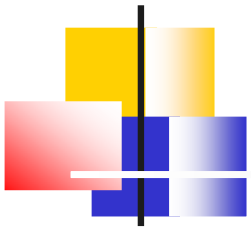
Dr. Susan Gauch (Chair)

Dr. Arvin Agah

Dr. Perry Alexander

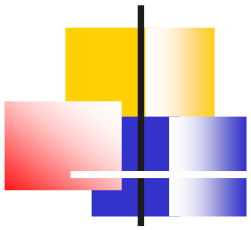
Information and Telecommunication Technology Center

University of Kansas



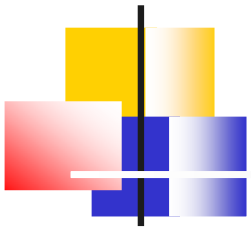
Presentation Outline

- Search engines today.
- Personalization and related work.
- Sources of user information.
- Personalizing search results (Conceptual vs Final Rank).
- System architecture.
- Experimental setup, experiments and validation.
- Summary and conclusions.
- Application demonstration.



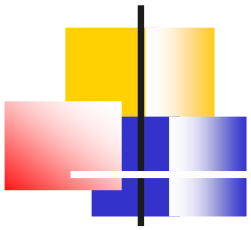
Search Engines Today

- Search engines are used more and more as referrals to web sites rather than direct navigation via hyperlinks [StatMarket].
- According to an analysis conducted by OneStat the most common query length submitted to a search engine is (32.6%) was two words long. 77.2% of all queries were three words long or less. This shows how small information is supplied for a ordinary search.



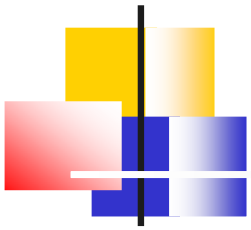
Search Engines Today contd

- Common problems of search engines:
  - ambiguity (e.g., "rock", "canon book");
  - retrieved results are based on web popularity rather than user's interests;
- Goals:
  - Improve search accuracy by retrieving by concept (e.g., "music", "classical studies");
  - Improve search accuracy by matching user interests;



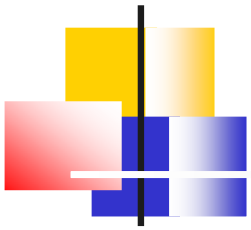
Personalization

- Personalization is the process of presenting the right information to a specific user at the right moment.
- Implicit vs explicit personalization:
  - explicit methods have been used by commercial systems for years (e.g. Yahoo!, Seruku Toolbar, Furl);
  - implicit methods are more complex to explore. Some examples are time spent on a page, page scrolling, click through;



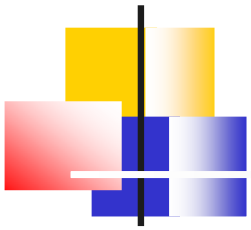
Personalization contd

- Personalization can be applied to search in two different ways:
  - Providing tools that help users organizing their past searches, preferences and visited URLs;
  - Creating and maintaining sets of user's interests (stored in profiles) that can be used by retrieval process of a search engine to provide better results;



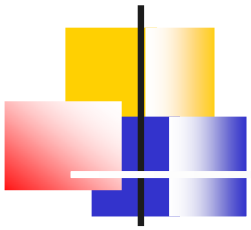
Related Work

- User profiles can be used to
  - to re-rank the results returned from the initial retrieval process;
  - to filter results that better fit user's interests;
- Kuflik and Shoval defines 3 classes of user profiles:
  - Content-based (e.g. documents are represented as vectors of terms);
  - Collaborative (i.e. information gathered from explicit ratings and preferences);
  - Rule-based (i.e. rules are specified from information explicitly provided by users or from session histories);



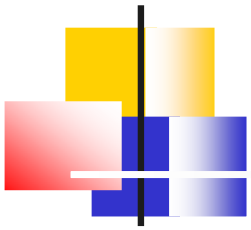
Related Work contd

- Ongoing research to investigate ways
  - to implicitly collect information about the user;
  - to represent information about the user;
- Kelly and Teevan
  - Review of papers about collecting relevance feedback using implicit approaches;
  - Observable feedback behaviours can be classified with respect to *behaviour category* (e.g. examine, annotate, reference) and *minimum scope* (the smallest possible scope of the item being acted upon)



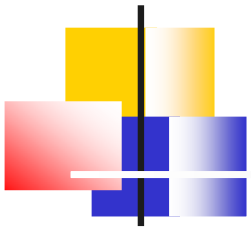
Sources of User Information

- User explicit information
  - users too lazy;
  - information becomes out of date/inaccurate;
- User browsing histories
  - must collect information via desktop robot or have user connect to Internet via a proxy
- User desktops
  - contextual retrieval
- User search histories
  - information available to search engine itself



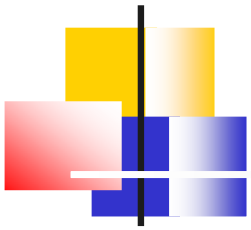
User Profile Creation

- Collect information about the user's interests
  - search history
- Categorize representative texts into concept hierarchy
  - use Open Directory Project for concepts
  - train classifier on training pages
  - compare representative texts to training texts to identify the concepts discussed
- Concept weights represent user interests



Conceptual Rank

- Submit query to Internet search engine
- Categorize each result into same concept hierarchy (e.g., ODP) to create result profiles
- Conceptual match is calculated based on similarity between each result profile and user profile
- Rerank results based on conceptual match
  - rank order produced called "conceptual rank"

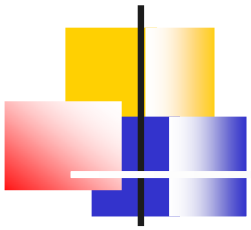


Final Rank

- The final rank of the document is calculated by combining the conceptual rank with Google's original rank using the following weighting scheme

$$\text{Final Rank} = \alpha * \text{ConceptualRank} + (1 - \alpha) * \text{GoogleRank}$$

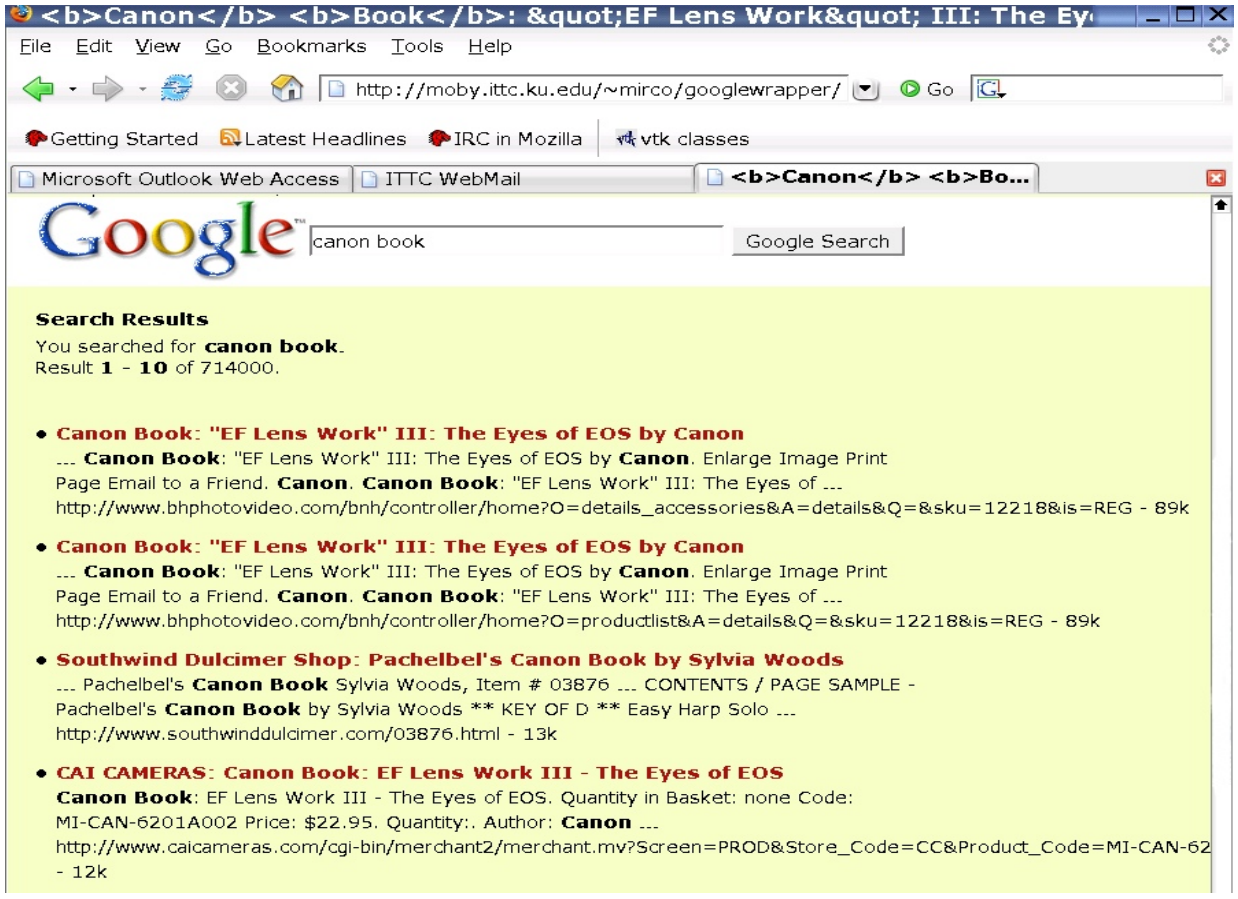
- α has a value between 0 and 1
- The conceptual and search engine based rankings can be blended in different proportions by varying the value of α

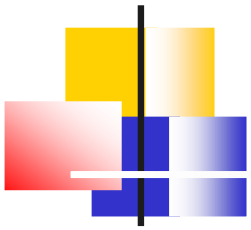


System Architecture contd

- GoogleWrapper: wrapper around Google built using Google's APIs.
- GoogleWrapper monitors users maintaining a log of:
  - submitted queries for which at least one result was visited;
  - user-selected snippets from retrieved results;
  - top 10 results (title and summary) retrieved;

System Architecture contd



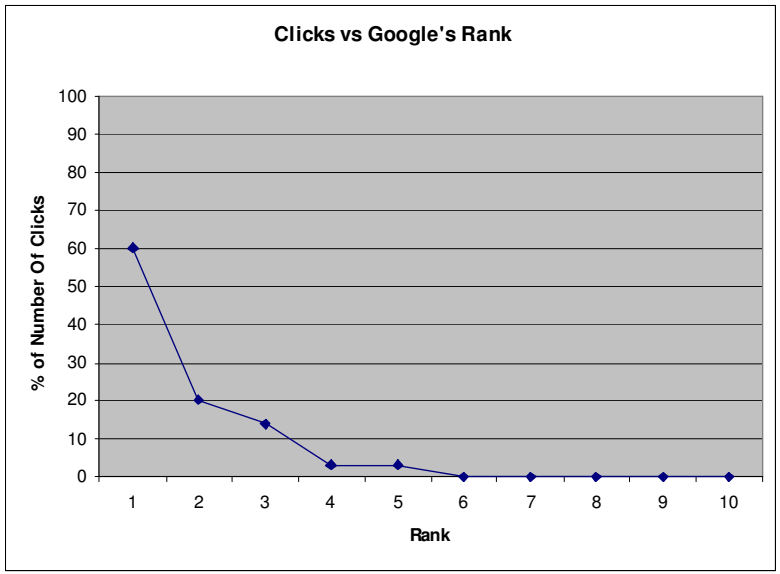


Experimental Setup

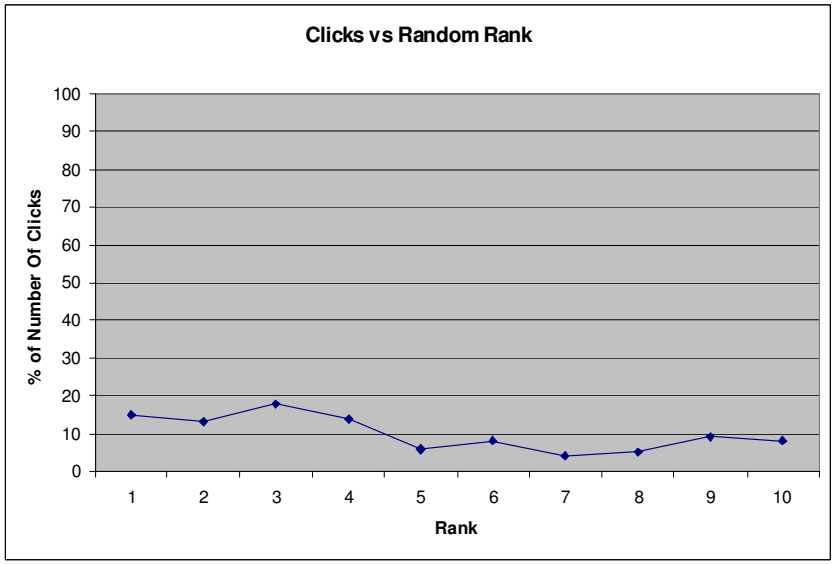
- Preliminary study
  - randomly selected 100 queries out of 576 collected
  - the top-ranked result was the most frequently selected
  - (60%), the second (20%), and the third (14%).
- Conclusions
  - *users rarely look at results beyond the first page;*
  - *user judgments are affected by presentation order;*
- Reimplement GoogleWrapper so that only top 10 results are displayed in random order;

Experimental Setup contd

Percentage of user selections versus rank for the top 10 results from Google



Percentage of user selections versus rank for the top 10 results from Google displayed in random order

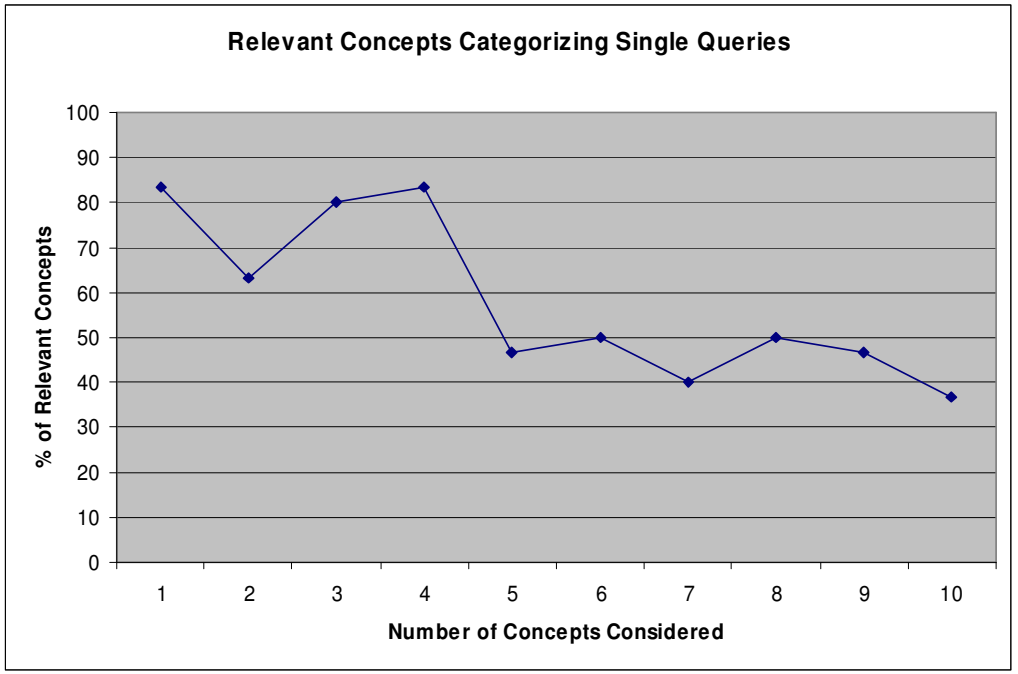


Information and Telecommunication Technology Center

University of Kansas

Experimental Setup contd

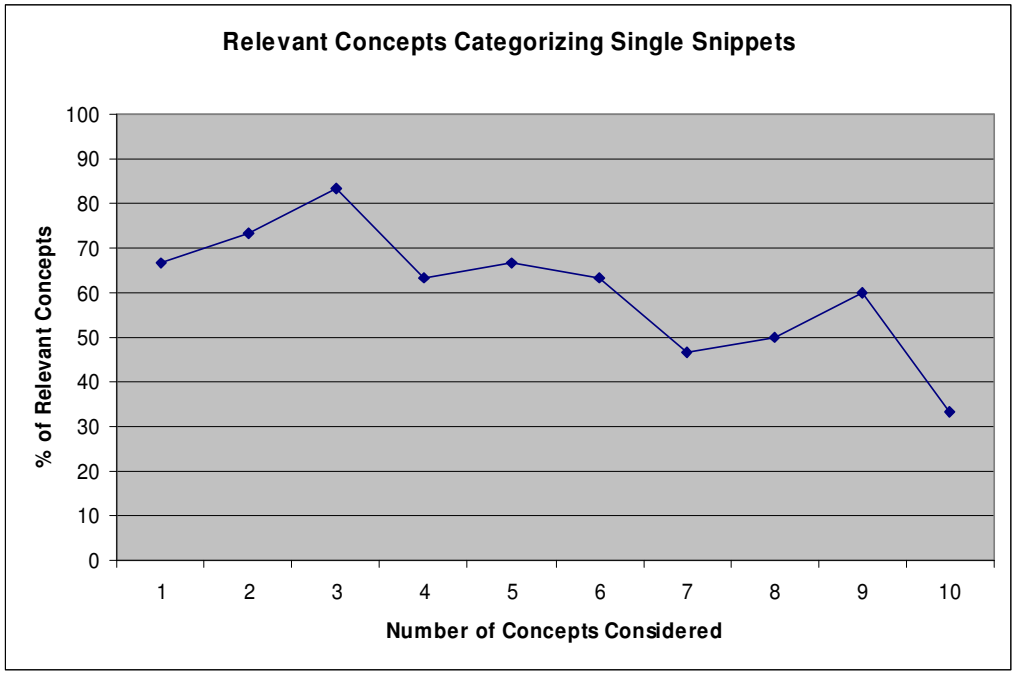
- Percentage of relevant concepts versus number of categories considered per single query



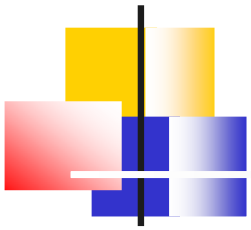
- 30 queries (5 from each user)
- Manually examined the top 10 concepts

Experimental Setup contd

- Percentage of relevant concepts versus number of categories considered per single snippet



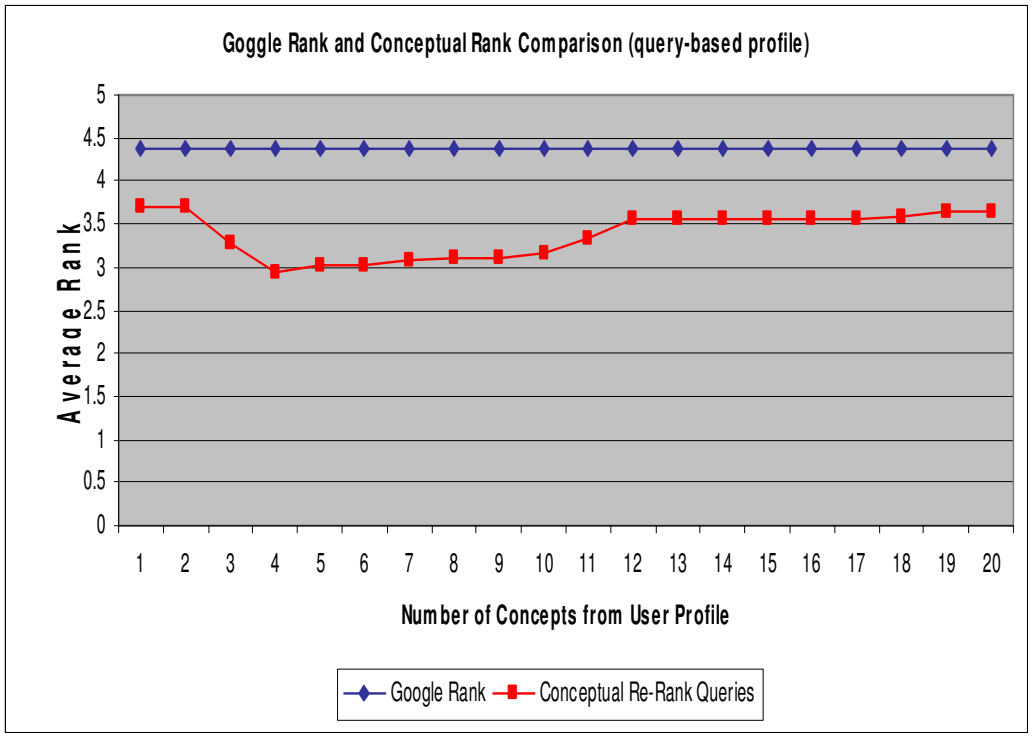
- 30 snippets (5 from each user)
- Manually examined the top 10 concepts



Experiments

- Monitored six volunteers.
- 609 queries collected.
- We removed duplicate queries for each user
- 47 queries per user (282 total) were distributed into the following sets:
  - 240 (40 per user) queries were used for training the 2 user profiles (query-based and snippet-based);
  - 30 (5 per user) queries were used for testing personalized search parameters;
  - 12 (2 per user) queries were used for validating the selected parameters

Experiment 1

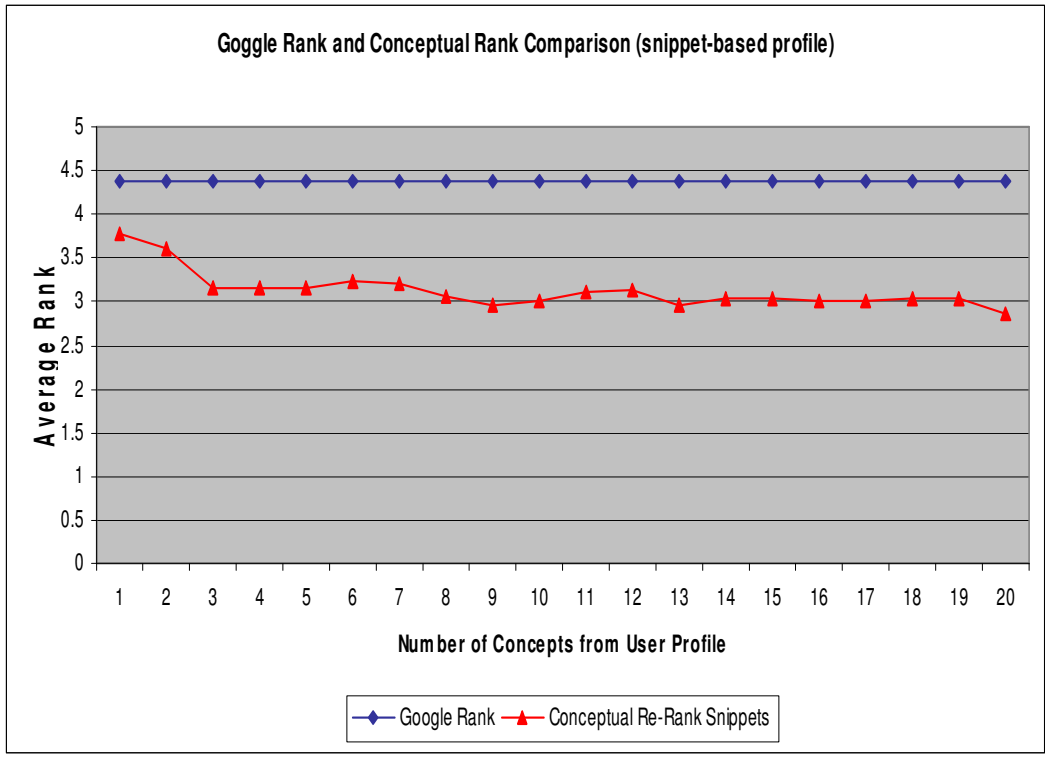


- *Profile built classifying and combining each query (4 concepts from each classified query).*
- Varied
  - the number of queries used to create profile;
  - the number of concepts for the user profile.
- Average Google Rank is 4.4
- Best average conceptual rank is 2.9 (using 4 concepts from user profile and 30 queries to create the profile)

Information and Telecommunication Technology Center

University of Kansas

Experiment 2

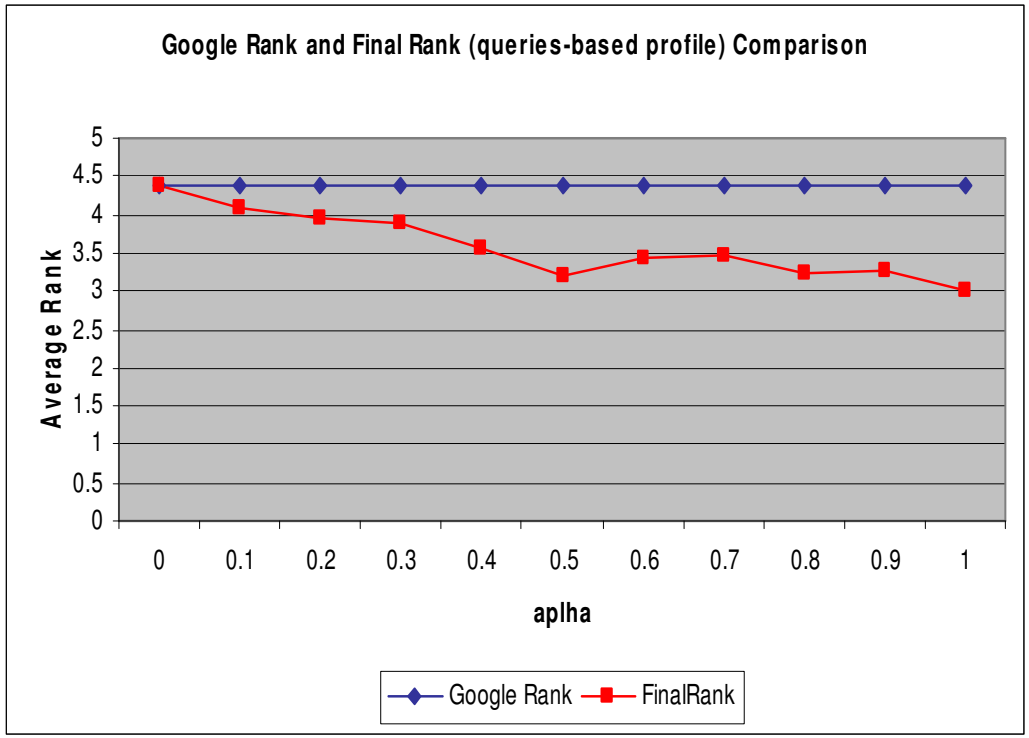


- *Profile built classifying and combining each snippet (5 concepts from each classified snippet).*
- Varied
  - the number of snippets used to create profile;
  - the number of concepts for the user profile.
- Average Google Rank is 4.4
- Best average conceptual rank is 2.9 (using 20 concepts from user profile and 30 snippets to create the profile)

Information and Telecommunication Technology Center

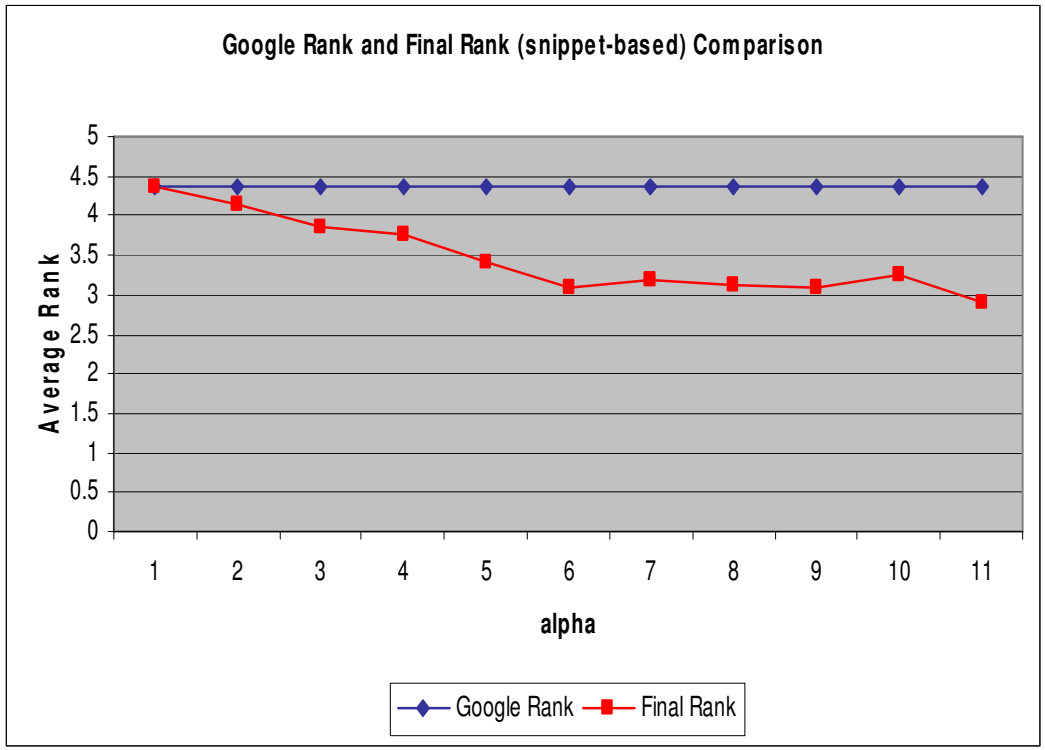
University of Kansas

Experiment 3



- *Final rank: combination of original rank with conceptual rank. Query-based profile.*
- Varied α from 0.0 to 1.0 (0.1 step);
- Parameters that gave best improvements in Experiment 1 were considered
- Best value when α is 1.0 (only conceptual rank is applied);

Experiment 4

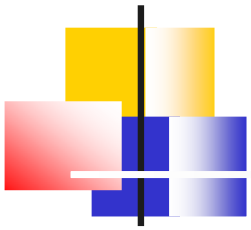


- *Final rank: combination of original rank with conceptual rank. Snippet-based profile.*
- Varied α from 0.0 to 1.0 (0.1 step);
- Parameters that gave best improvements in Experiment 2 were considered
- Best value when α is 1.0 (only conceptual rank is applied);

Validation

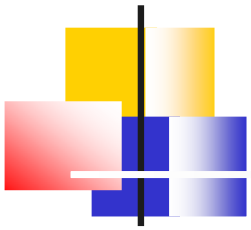
- To verify that the 2 types of profiles (query-based and snippet-based) are able to improve queries never seen before.
- 12 (2 per user) testing queries never seen before;
- Query-based profile (30 queries and 4 concepts per query);
- Snippetd-based profile (30 snippets and 20 concepts per snippet);

Ranking Algorithm	Avg. Rank	Improvement	# Training Queries/Snippets	# Profile Concepts Used
Google (Original)	4.8	--	--	--
Query-based Profile	1.8	37%	30	4
Snippet-based Profile	3.5	27%	30	20



Summary

- Built user profiles based on queries submitted and snippets of user-selected results.
- This information was sufficient to build user profiles that were able to significantly improve personalized rankings.
- A query-based profile produced an improvement of 33%.
- A snippet-based profile produced an equivalent improvement of 34%.



Conclusions

- Search engines can capture information submitted to their site in order to create personalized search.
- Users need not install proxy servers or desktop bots.
- Privacy issues arise with any personalized service.
- Need to look at combination of short-term, long-term user interests with current task focus.

http://www.ittc.ku.edu/~mirco/demo