Contextual Information Retrieval Using Ontology-Based User Profiles

Vishnu Kanth Reddy Challam

Master's Thesis Defense

Date: Jan 22nd, 2004.

Committee

Dr. Susan Gauch(Chair)

Dr.David Andrews

Dr. Jerzy W.Grzymala-Busse

Information and Telecommunication Technology Center

University of Kansas

Presentation Outline

- Search Engines Today
- Search Engine Personalization
- Contributions
- Our Approach for Contextual IR
- Experiments and Evaluation
- Conclusions and Future Work

Search Engines Today

- Return results based on simple key-word matches. No regard for conceptual information.

  For E.g. : If the query is “SALSA”

  Is it.....

Search Engines Today Contd..

- What is the user looking for?
- No personalization mechanism to understand the information needs of the user.

Search Engine Personalization…How?

- Collect and represent information about the user.
- Use this information to either filter or re-rank the results returned from the initial retrieval process or directly use this information in the search process.

Search Engine Personalization ..Challenges

- How can accurate information about the user's interests be collected and represented?
- How can we use this information to deliver personalized search results?

Contributions....

- We present a novel-approach to personalizing search engines using ontology-based contextual user profiles.
- Studied the effect of conceptual ranking versus original keyword based ranking.
- Studied the usage of multiple sources of information to build the user's contextual profile.

Related Work

- Semantic Web
  - Explicitly state meaning of content using Knowledge Representation Languages
  - Domain specific efforts
  - Web is democratic!

Design Criteria

- Monitor and store user information on the client machine or the server.
- Short term vs. Long term
- With server side profiling, privacy is an issue.
- Instantaneous information needs are hard to satisfy.

Contextual Search

- No long term user profiles
- Build contextual profiles that capture the information needs of the user at the time they conduct search...TASK ORIENTED
- Upload the contextual profile to the server.
- Privacy

How to Build Contextual Profiles?

- Monitor the activity of the user on his/her Windows machine. Capture content from Word documents,Web pages, Chat transcripts etc..
- Classify the captured content to build a contextual profile

Monitoring the User Activity

- A Windows application that runs in the background.
- Captured text from open Word, IE, MSN Chat windows.
- Stored the captured content in a special folder on the clients machine.Content is assigned a time-stamp.

Text Classification

- Classifier works in 2 phases: training and classification.
- Training Phase:
  - Classifier is given a series of documents classified manually.
  - Learns about the features (vocabulary) of the various categories into which the text might be classified.

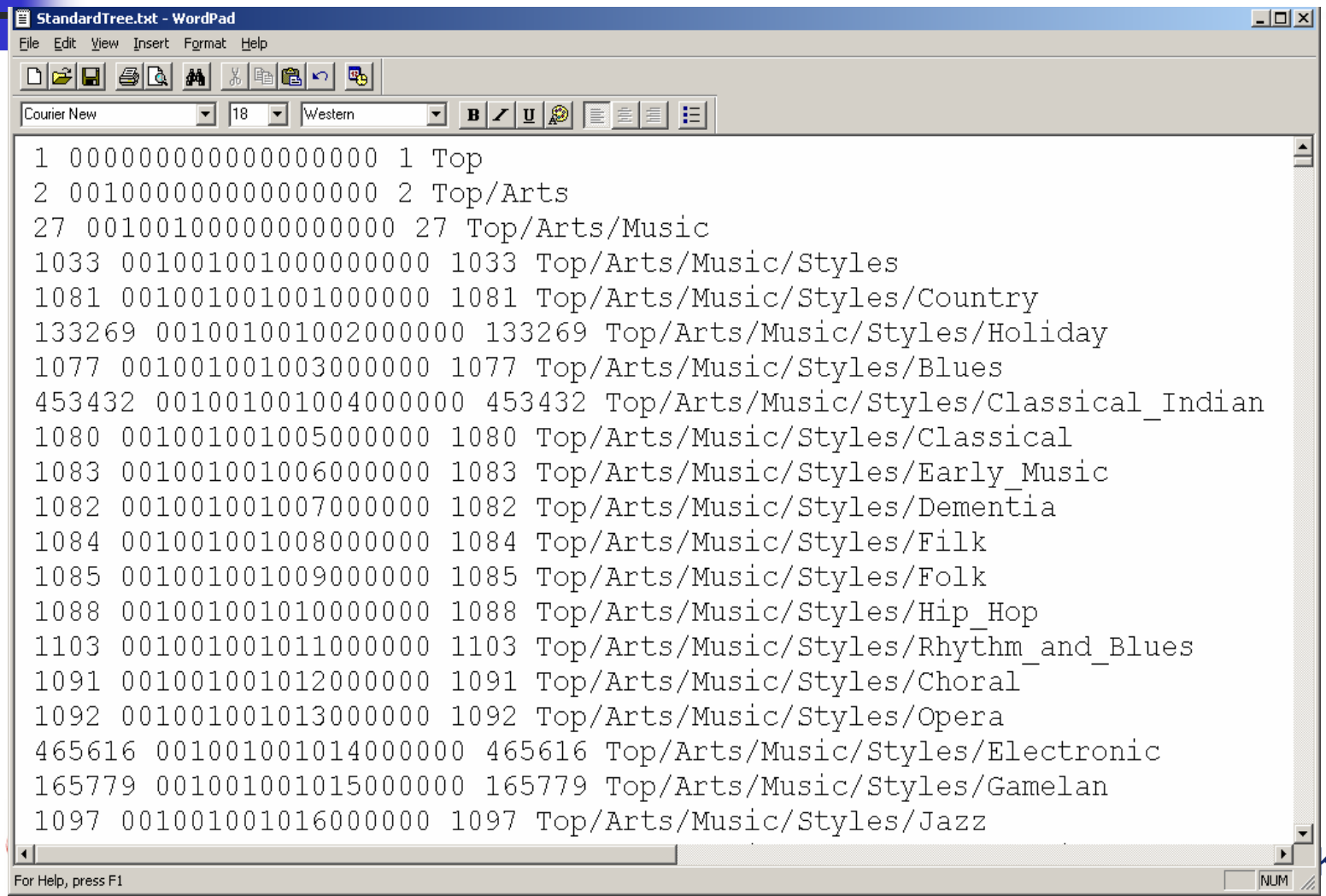
Text Classification Contd...

- Classification phase:
  - Classifier, classifies the input text and assigns it to a particular category based on similarity between the features of input text and those extracted from training data.

Text Classification : Our Approach

- Vector-space model (tf-idf model).
- Training data are the documents manually assigned into categories of the Standard Tree which is our reference ontology.
- Classifier creates a vector of vocabulary terms and weights associated with the category in an inverted file.

Standard Tree



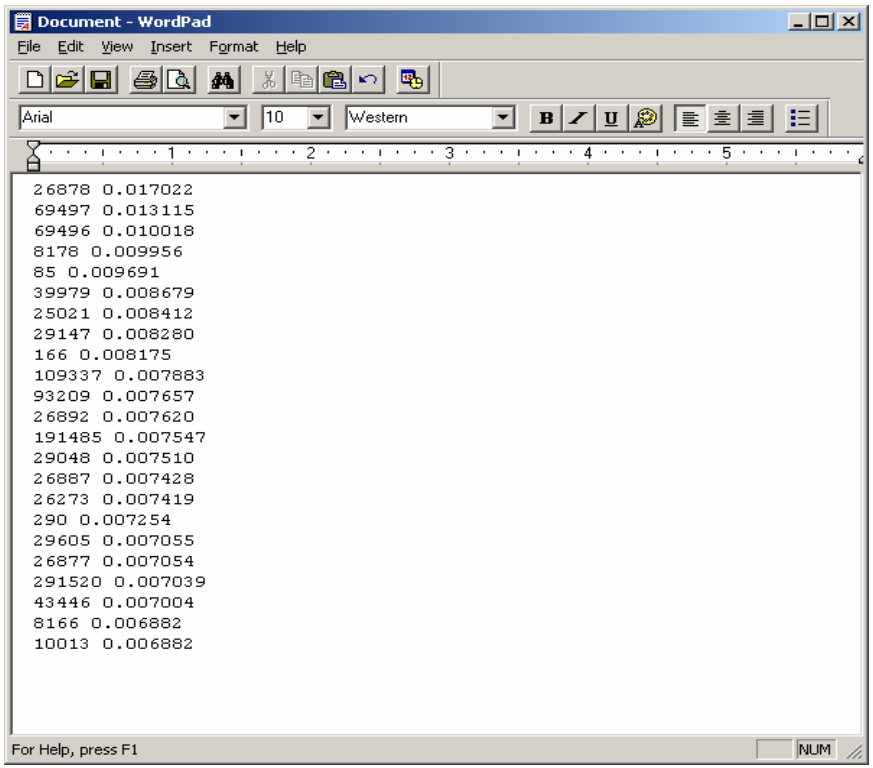
Text Classification: Our Approach Contd..

- During classification phase, vector of input document is created.
- Degree of similarity between training vectors and input document vector calculated using dot product of the vectors.
- Best matches are the concepts into which the input document is assigned.

Building Contextual User Profile

- Content created/viewed within a specific time window is classified.
- The classifier represents the user's contextual profile for the time window as a weighted ontology.
- Weight of a concept in the ontology represents the amount of information recently viewed/created that was classified into that concept.

Sample Contextual User Profile



- Category-id, Weight
- Category-id used to identify the concept in Standard Tree.
- 26878 is Top/Science/Environment/Water_Resources

Personalizing Search Results Using Contextual User Profiles

- Results are re-ranked using a combination of the **original rank** and their **conceptual rank**
- Similarity of the documents to the contextual profile is used to calculate the conceptual rank

Conceptual Rank

- Document's title and summary are classified to create the document profile.
- Document profile is compared to the contextual profile to calculate the conceptual similarity between document and user's context.

$$sim(context_i, doc_j) = \sum_{k=1}^{N} wt_{ik} * wt_{jk}$$

where

wt_{ik} = Weight of $Concept_k$ in $Context_i$

wt_{jk} = Weight of $Concept_k$ in $document_j$

Final Rank

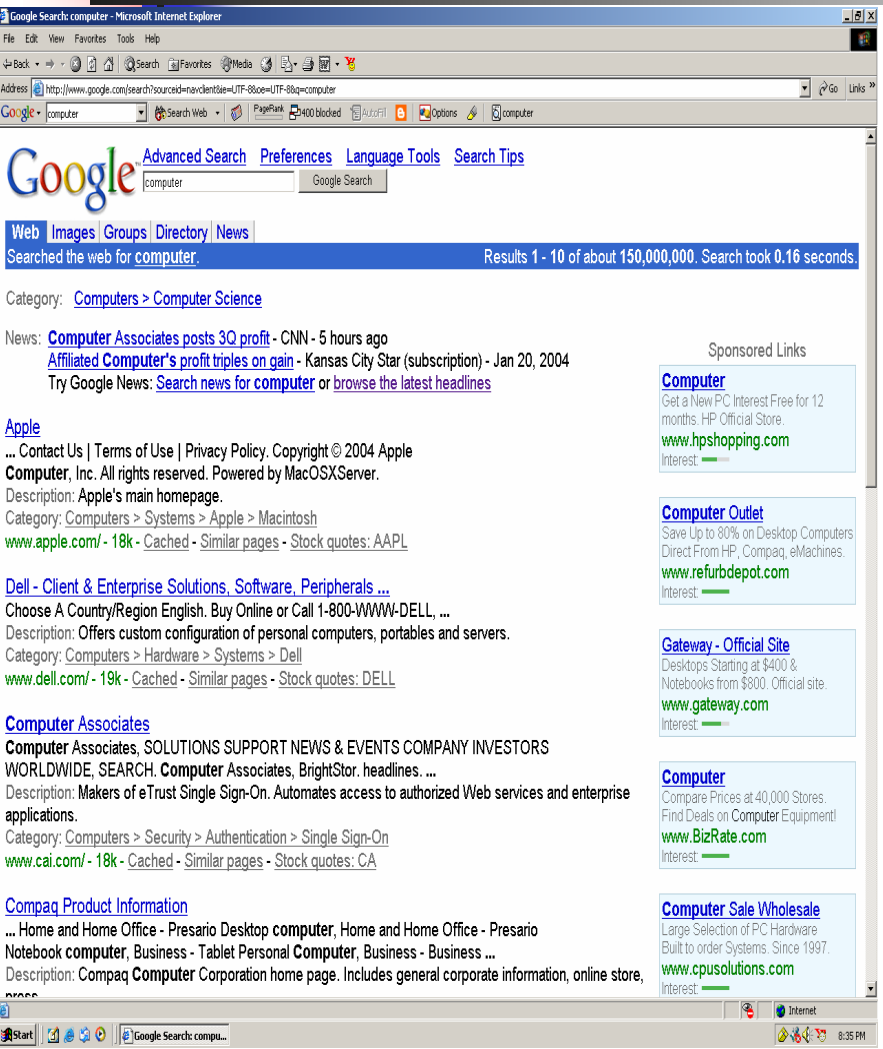
$$\text{Final Rank} = \alpha * \text{Conceptual Rank} + (1 - \alpha) * \text{Keyword Rank}$$

- α has a value between 0 and 1
- Varying the values of α between 0 and 1 conceptual and keyword ranks can be weighted differently.

Experiments and Evaluation

- Wrapper around Google built using Google API.
- **Google Wrapper** builds a log of:
    1. Queries given by user
    2. Results & ranks returned by Google
    3. Result clicked by the user
    4. Title & Summaries
- Randomizes the results returned by Google before displaying them to the user.

Google Wrapper



Experiments

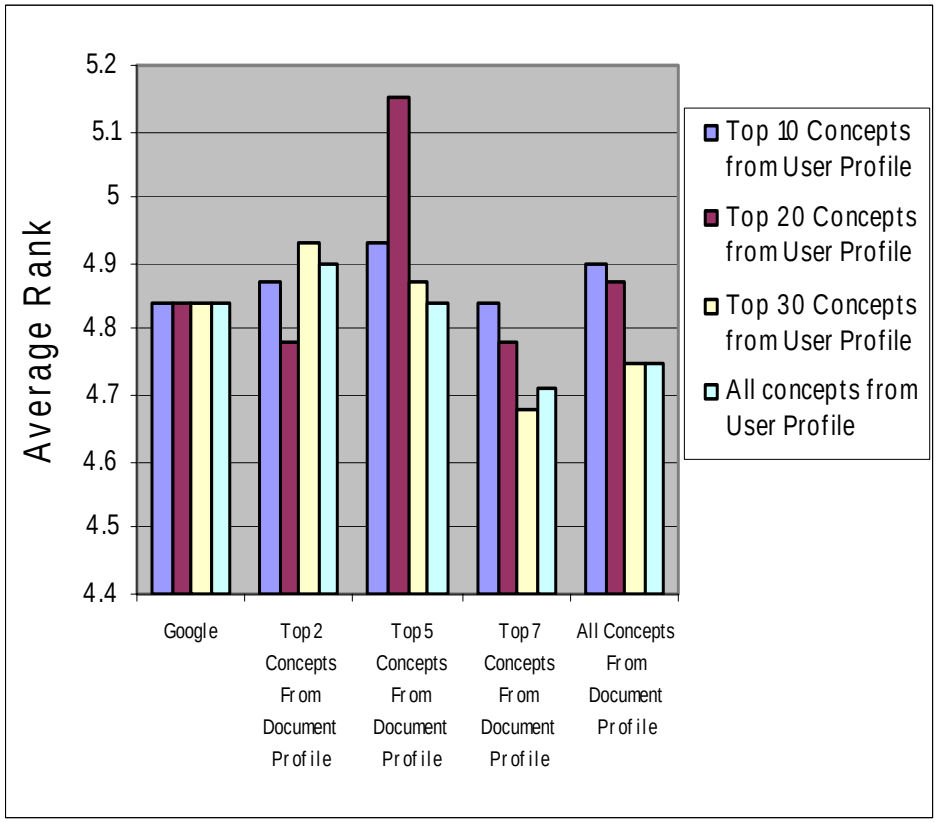
- 5 users asked to write essays on topics ranging from car buying, labs at ITTC to jewelry.
- Windows application monitored their activity
- Queries issued to Google Wrapper
- Result clicked by the user was used as a form of **implicit user relevance** for analysis.

Experiments Contd..

- Log of 50 queries.
- 6 had to be filtered out. 44 queries analyzed
- Evaluate number of concepts for the user's contextual profile, the document profile and the value of α for blending original and conceptual ranks.
- Analysis based on average rank of the result clicked by the user in our conceptual search engine and baseline system Google.

Evaluation

- **Profile built from content of Word documents alone**
- 32 queries analyzed
- Varied the number of concepts for the user profile and the document profile.
- Average Google Rank is 4.84
- Best average conceptual rank is 4.68



Information and Telecommunication Technology Center

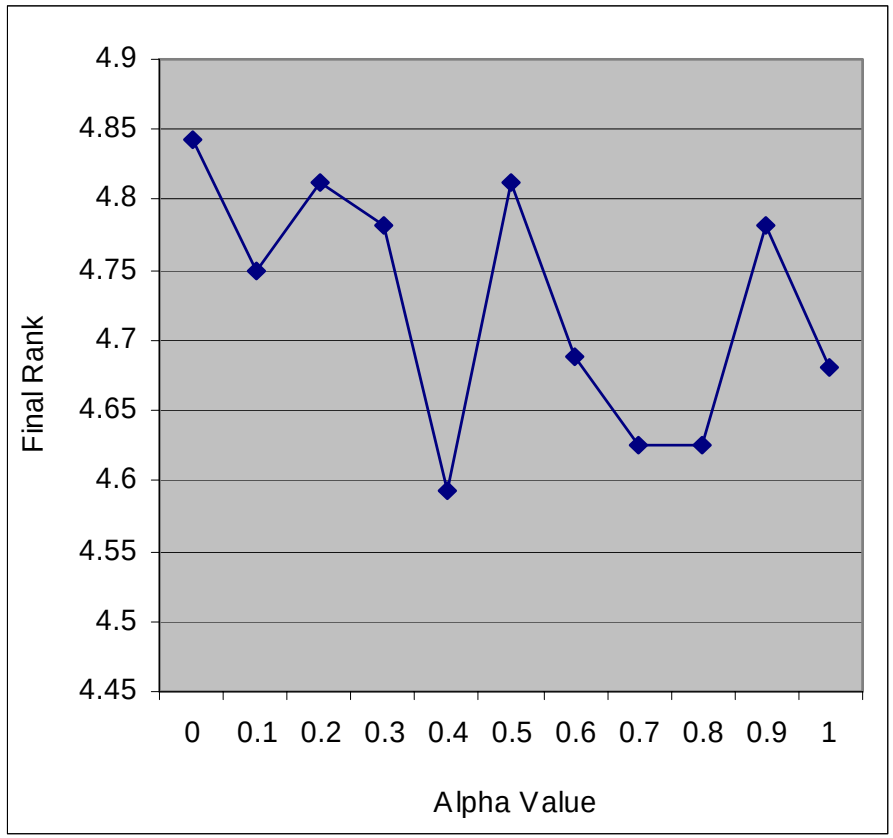
University of Kansas

Evaluation Contd...

- Final Rank calculated using the formula

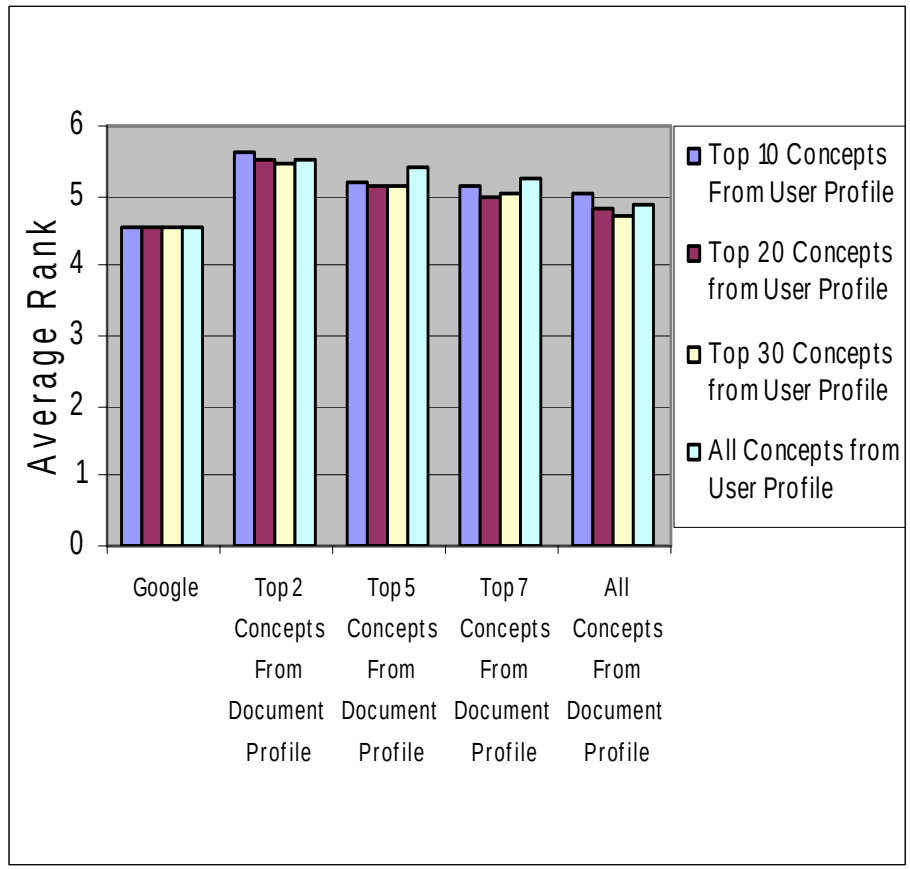
  $FR = \alpha * CR + (1-\alpha) * KR$

- Best final rank of 4.59 when $\alpha = 0.4$
- 5.16 percent improvement over Google's rank of 4.84
- Contextual information from Word documents can be used to improve web queries.



Evaluation Contd…

- **Profile built from content of Web pages alone**
- 31 queries analyzed
- Varied the number of concepts for the user profile and the document profile.
- Average Google Rank is 4.58
- Best average conceptual rank is 4.74(30 concepts for contextual profile and all concepts for document profile)

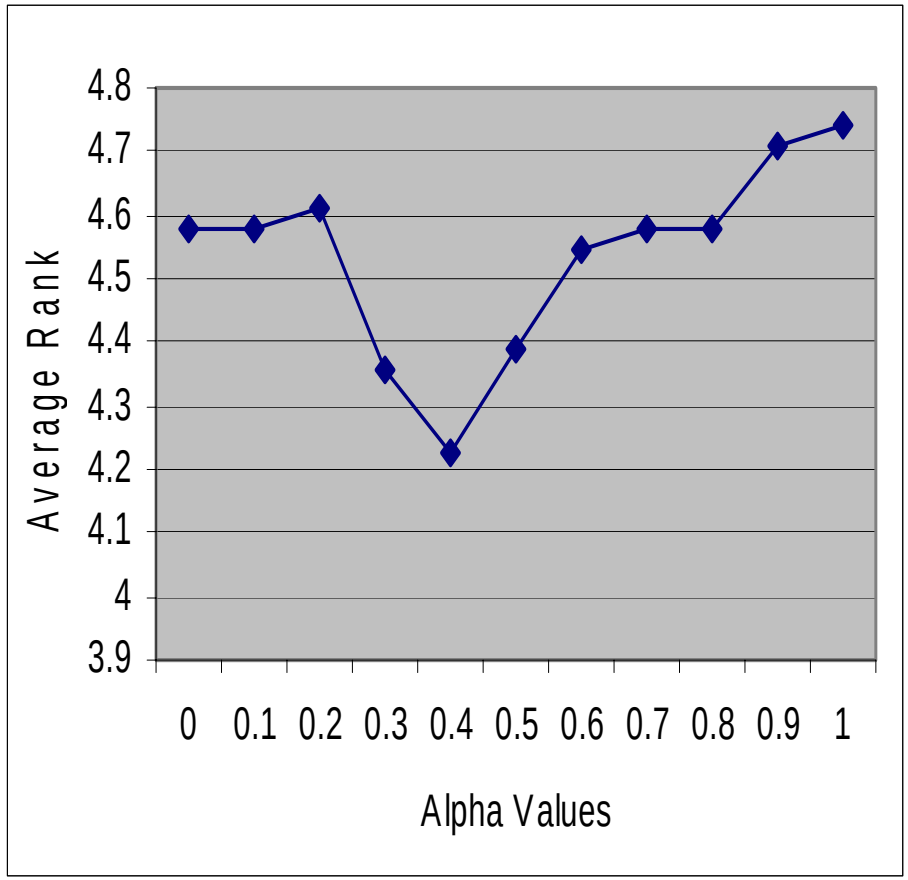


Evaluation Contd...

- Final Rank calculated using the formula

  $FR = \alpha * CR + (1-\alpha) * KR$

- Best final rank of 4.22 when $\alpha = 0.4$
- 7.86 percent improvement over Google's rank of 4.74
- Contextual information from Web Pages can be used to improve web queries.

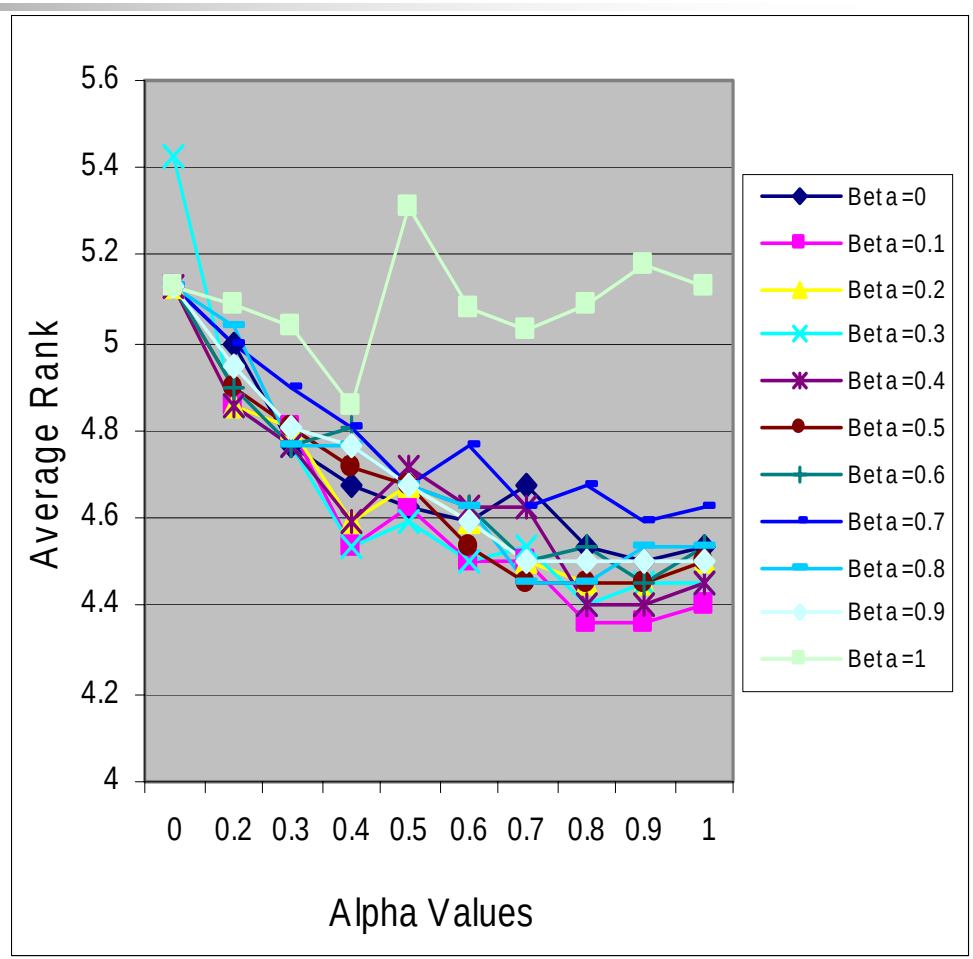


Evaluation Contd...

- **Profile built by combining content of Web pages and Word Documents.**
- Final Profile = β *Word Profile + (1 - β) * Web Profile
- β has values between 0 and 1

Evaluation Contd....

- Effect of α and β
- 22 queries analyzed
- Best Conceptual Rank 4.36 when α is 0.8 and β is 0.1
- **15% improvement over Google's rank**!

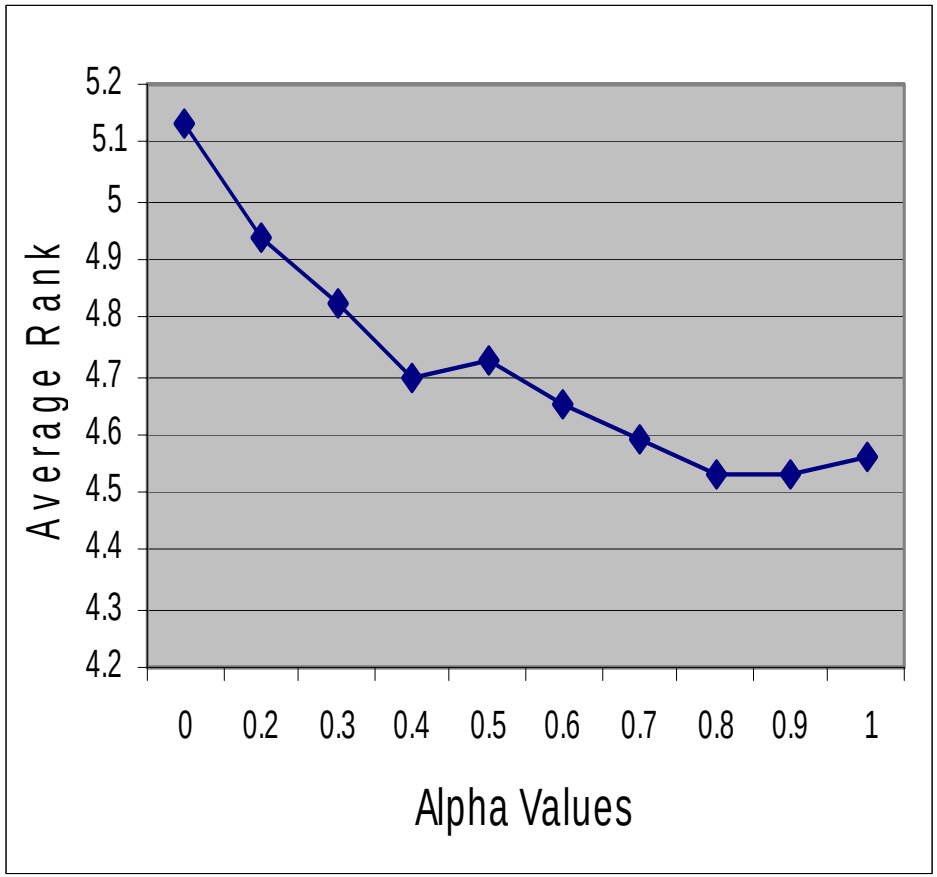


Information and Telecommunication Technology Center

University of Kansas

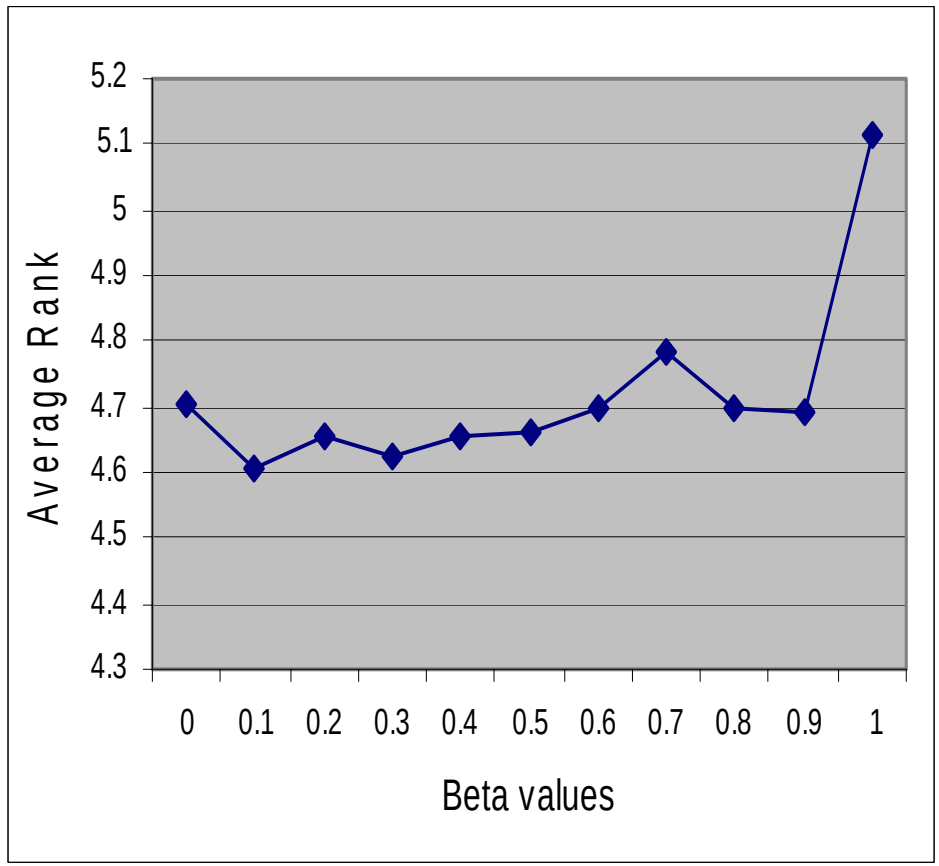
Evaluation Contd...

- Effect of α on final rank
- High value of α indicates that conceptual rank should be given more importance.
- Re-ranking among top 10, all of them match the user's query equally well.
- Primary distinguishing factor is conceptual similarity to contextual profile.



Evaluation Contd…

- Effect of β on final rank
- β values between 0.1 and 0.5 produce roughly comparable results.
- Increased importance of Web content maybe because Word documents were short.
- If more content available in Word documents a higher value of β might have been observed.



Conclusions

- Contextual profiles improve Web searches.
- 15% improvement over Google when profile is built by combining content from Word documents and Web pages
- Within top 10 results of Google, re-ranking should be done giving more weight to conceptual similarity between documents and the contextual profile

Conclusions Contd..

- All users were expert search engine users. Query length was long.
- Longer queries tend to disambiguate themselves.
- System performs better for shorter queries more common on the Web as a whole

Future Work

- Best time window within which documents captured should be included in the contextual profile
- Analyze content from other sources like Chat transcripts, Excel spreadsheets, PowerPoint slides etc..
- Combination of user's current context, long and short term interests.

Questions or Comments

?? or !!

Thank You!