



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



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.....



© GMPdigital.com 2002



Information and
Telecommunication
Technology Center

University of Kansas



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.



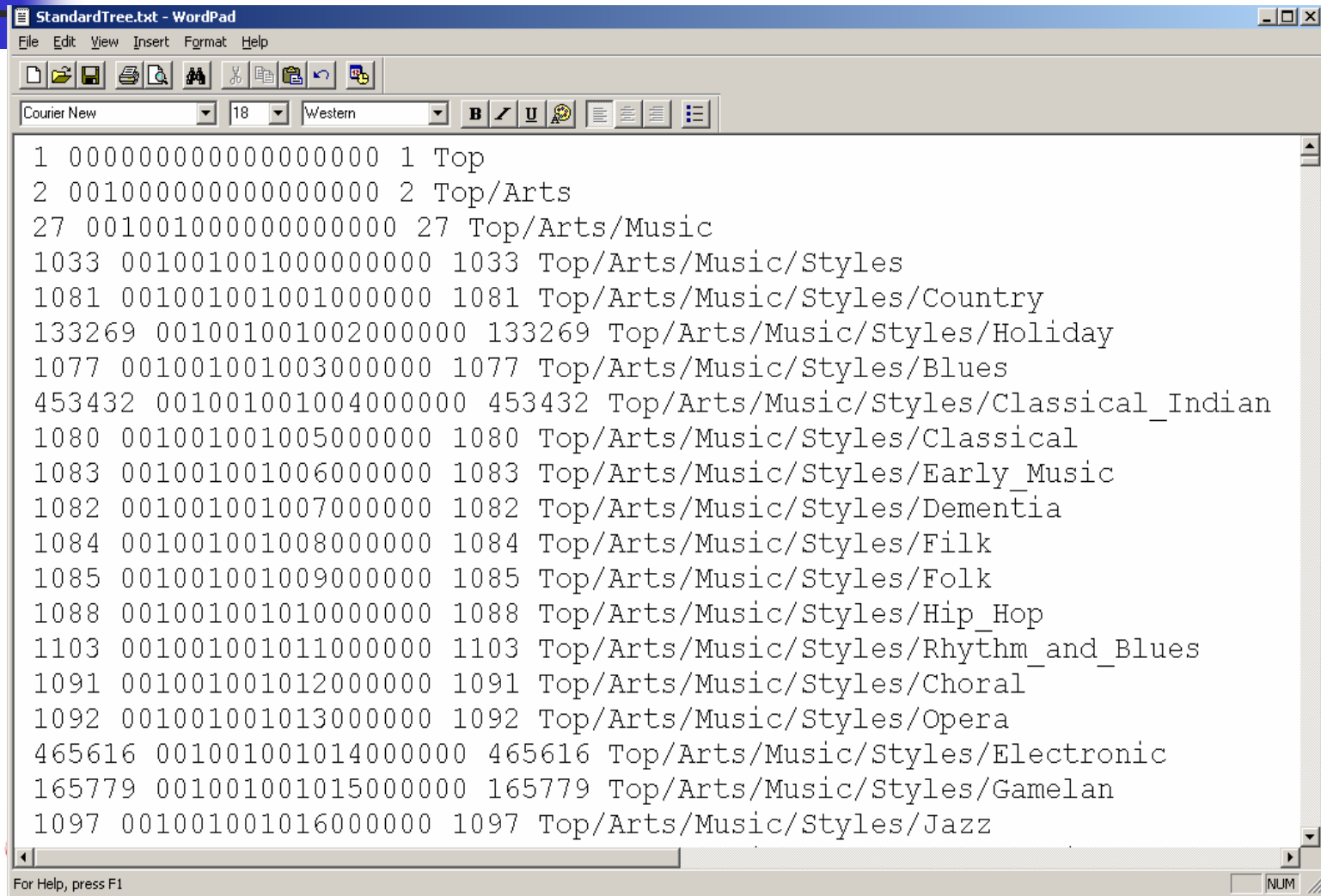
Text Classification Contd...

- 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



```
1 000000000000000000 1 Top
2 001000000000000000 2 Top/Arts
27 001001000000000000 27 Top/Arts/Music
1033 001001001000000000 1033 Top/Arts/Music/Styles
1081 001001001001000000 1081 Top/Arts/Music/Styles/Country
133269 001001001002000000 133269 Top/Arts/Music/Styles/Holiday
1077 001001001003000000 1077 Top/Arts/Music/Styles/Blues
453432 001001001004000000 453432 Top/Arts/Music/Styles/Classical_Indian
1080 001001001005000000 1080 Top/Arts/Music/Styles/Classical
1083 001001001006000000 1083 Top/Arts/Music/Styles/Early_Music
1082 001001001007000000 1082 Top/Arts/Music/Styles/Dementia
1084 001001001008000000 1084 Top/Arts/Music/Styles/Filk
1085 001001001009000000 1085 Top/Arts/Music/Styles/Folk
1088 001001001010000000 1088 Top/Arts/Music/Styles/Hip_Hop
1103 001001001011000000 1103 Top/Arts/Music/Styles/Rhythm_and_Blues
1091 001001001012000000 1091 Top/Arts/Music/Styles/Choral
1092 001001001013000000 1092 Top/Arts/Music/Styles/Opera
465616 001001001014000000 465616 Top/Arts/Music/Styles/Electronic
165779 001001001015000000 165779 Top/Arts/Music/Styles/Gamelan
1097 001001001016000000 1097 Top/Arts/Music/Styles/Jazz
```

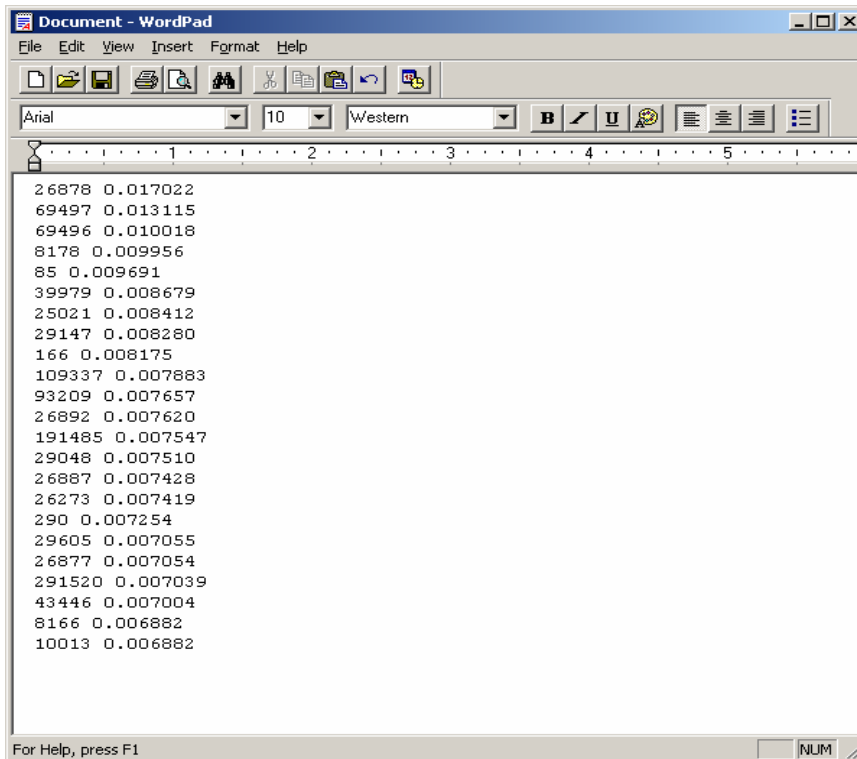

Text Classification: Our Approach Contd..

- 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



A screenshot of a WordPad window titled "Document - WordPad". The window displays a list of 20 rows, each containing a category ID and a weight. The text is as follows:

26878	0.017022
69497	0.013115
69496	0.010018
8178	0.009956
85	0.009691
39979	0.008679
25021	0.008412
29147	0.008280
166	0.008175
109337	0.007883
93209	0.007657
26892	0.007620
191485	0.007547
29048	0.007510
26887	0.007428
26273	0.007419
290	0.007254
29605	0.007055
26877	0.007054
291520	0.007039
43446	0.007004
8166	0.006882
10013	0.006882

- 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.

$$\text{sim}(\text{context}_i, \text{doc}_j) = \sum_{k=1}^N \text{wt}_{ik} * \text{wt}_{jk}$$

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

Google Search: computer - Microsoft Internet Explorer

Address: <http://www.google.com/search?source=ev&client=ie=UTF-8&oe=UTF-8&q=computer>

Google Google Search

Web | Images | Groups | Directory | News

Searched the web for computer. Results 1 - 10 of about 150,000,000. Search took 0.16 seconds.

Category: [Computers](#) > [Computer Science](#)

News: [Computer Associates posts 3Q profit](#) - CNN - 5 hours ago
[Affiliated Computer's profit triples on gain](#) - Kansas City Star (subscription) - Jan 20, 2004
Try Google News: [Search news for computer](#) or [browse the latest headlines](#)

Apple
... Contact Us | Terms of Use | Privacy Policy. Copyright © 2004 Apple
Computer, Inc. All rights reserved. Powered by MacOSXServer.
Description: Apple's main homepage.
Category: Computers > Systems > Apple > Macintosh
www.apple.com/ - 18k - Cached - Similar pages - Stock quotes: AAPL

Dell - Client & Enterprise Solutions, Software, Peripherals ...
Choose A Country/Region English. Buy Online or Call 1-800-WWW-DELL, ...
Description: Offers custom configuration of personal computers, portables and servers.
Category: Computers > Hardware > Systems > Dell
www.dell.com/ - 19k - Cached - Similar pages - Stock quotes: DELL

Computer Associates
Computer Associates, SOLUTIONS SUPPORT NEWS & EVENTS COMPANY INVESTORS
WORLDWIDE, SEARCH. Computer Associates, BrightStor. headlines. ...
Description: Makers of eTrust Single Sign-On. Automates access to authorized Web services and enterprise applications.
Category: Computers > Security > Authentication > Single Sign-On
www.cai.com/ - 18k - Cached - Similar pages - Stock quotes: CA

Compaq Product Information
... Home and Home Office - Presario Desktop computer, Home and Home Office - Presario Notebook computer, Business - Tablet Personal Computer, Business - Business ...
Description: Compaq Computer Corporation home page. Includes general corporate information, online store, ...

Sponsored Links

- Computer**
Get a New PC Interest Free for 12 months. HP Official Store.
www.hpshopping.com
Interest:
- Computer Outlet**
Save Up to 80% on Desktop Computers Direct From HP, Compaq, eMachines.
www.refurbdepot.com
Interest:
- Gateway - Official Site**
Desktops Starting at \$400 & Notebooks from \$800. Official site.
www.gateway.com
Interest:
- Computer**
Compare Prices at 40,000 Stores. Find Deals on Computer Equipment
www.BizRate.com
Interest:
- Computer Sale Wholesale**
Large Selection of PC Hardware Built to order Systems. Since 1997.
www.cpusolutions.com
Interest:

Google Wrapper - Microsoft Internet Explorer

Address: <http://www.ktc.ku.edu/~vichallan/GoogleWrapper/submitquery.php?q=computer&btnG=Google+Search>

Google Google Search

Search Results

You searched for **computer**.
Result 1 - 10 of 63600000.

- Computerworld**
Computerworld ...
<http://www.computerworld.com/> - 89k
- Dell - Client & Enterprise Solutions, Software, Peripherals ...**
Choose A Country/Region English. Buy Online or Call 1-800-WWW-DELL, ...
<http://www.dell.com/> - 19k
- IEEE Computer Society**
... The user-friendly **computer's** \$12,000 price tag kept it out of the consumer market, but many of its graphical features were passed down to the Mac in 1984. ...
<http://www.computer.org/> - 69k
- ASUSTeK Computer**
<http://www.asus.com/> - 11k
- Apple**
... Contact Us | Terms of Use | Privacy Policy. Copyright © 2004 Apple
Computer, Inc. All rights reserved. Powered by MacOSXServer.
<http://www.apple.com/> - 18k
- Compaq Product Information**
... Home and Home Office - Presario Desktop **computer**, Home and Home Office - Presario Notebook **computer**, Business - Tablet Personal **Computer**, Business - Business ...
<http://www.compaq.com/> - 32k
- Computer Associates**



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.

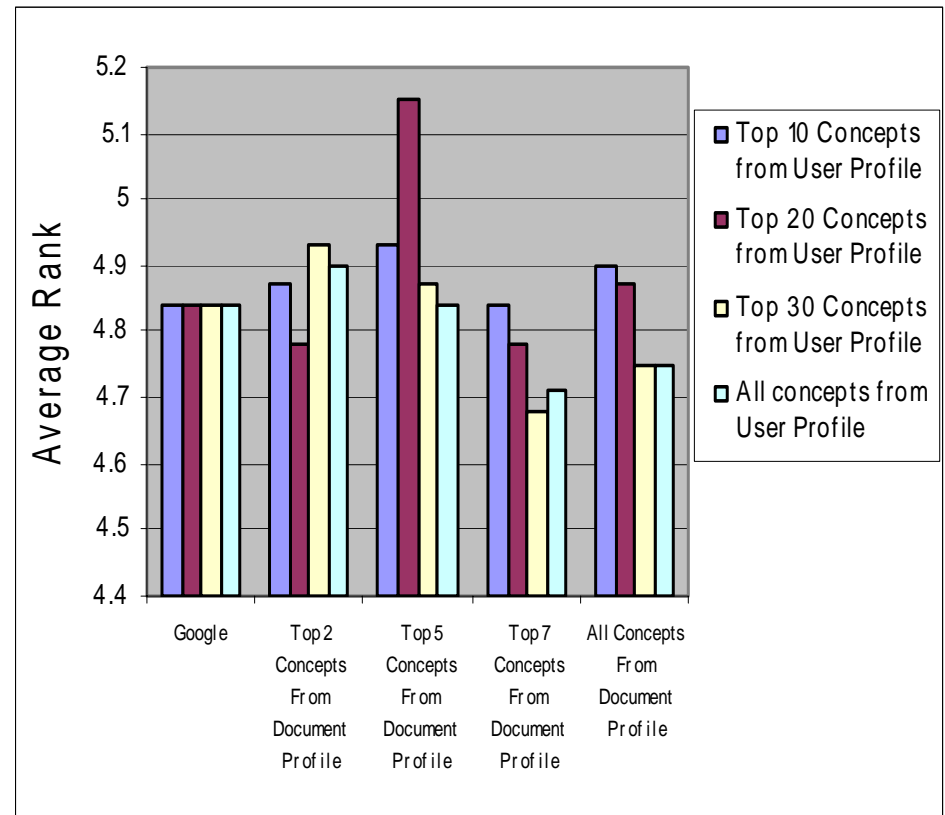


Experiments Contd..

- 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

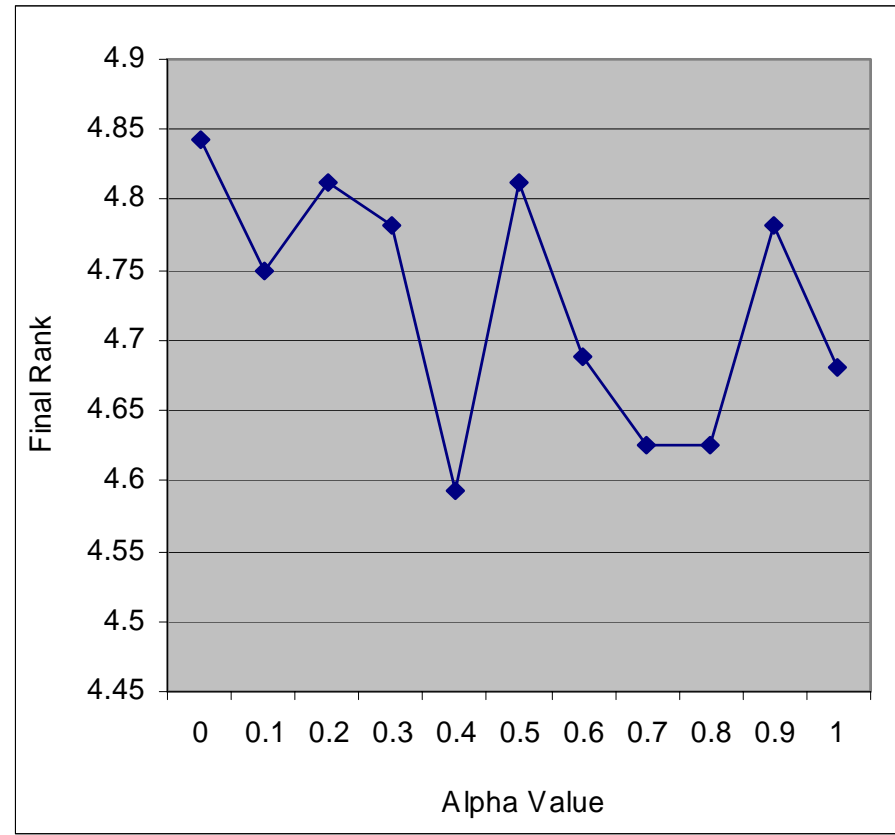


Evaluation Contd...

- Final Rank calculated using the formula

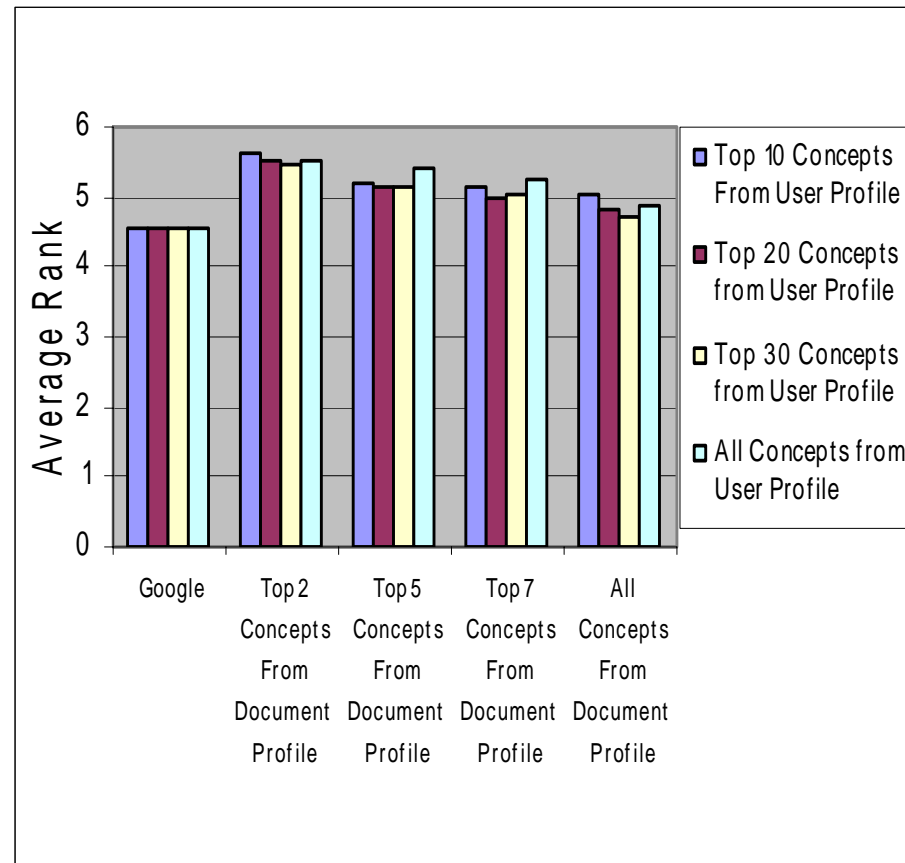
$$FR = \alpha * CR + (1 - \alpha) * KR$$

- 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)

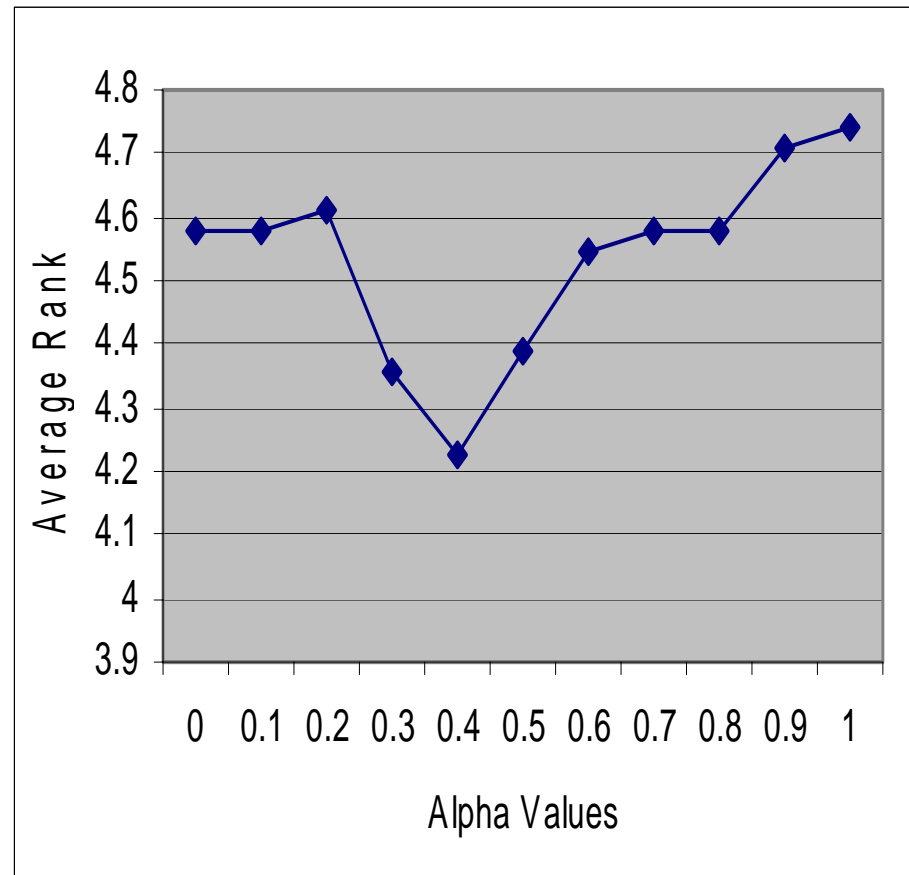


Evaluation Contd...

- 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.



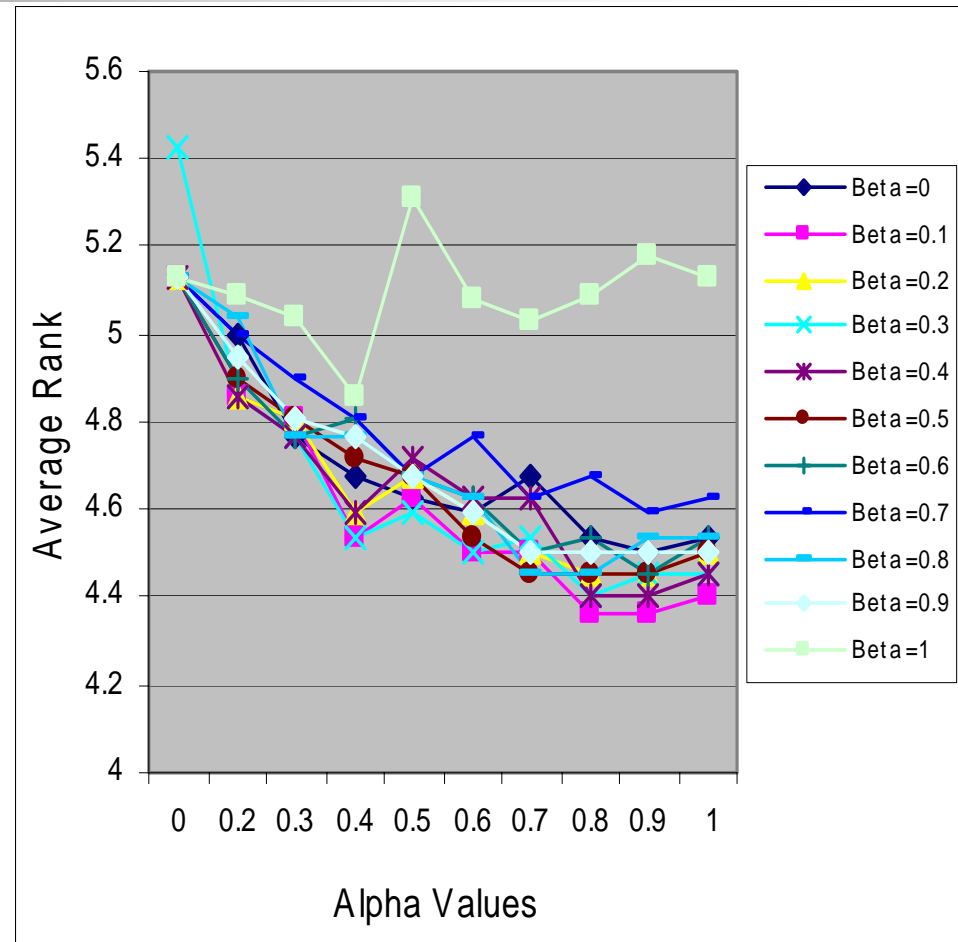


Evaluation Contd...

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

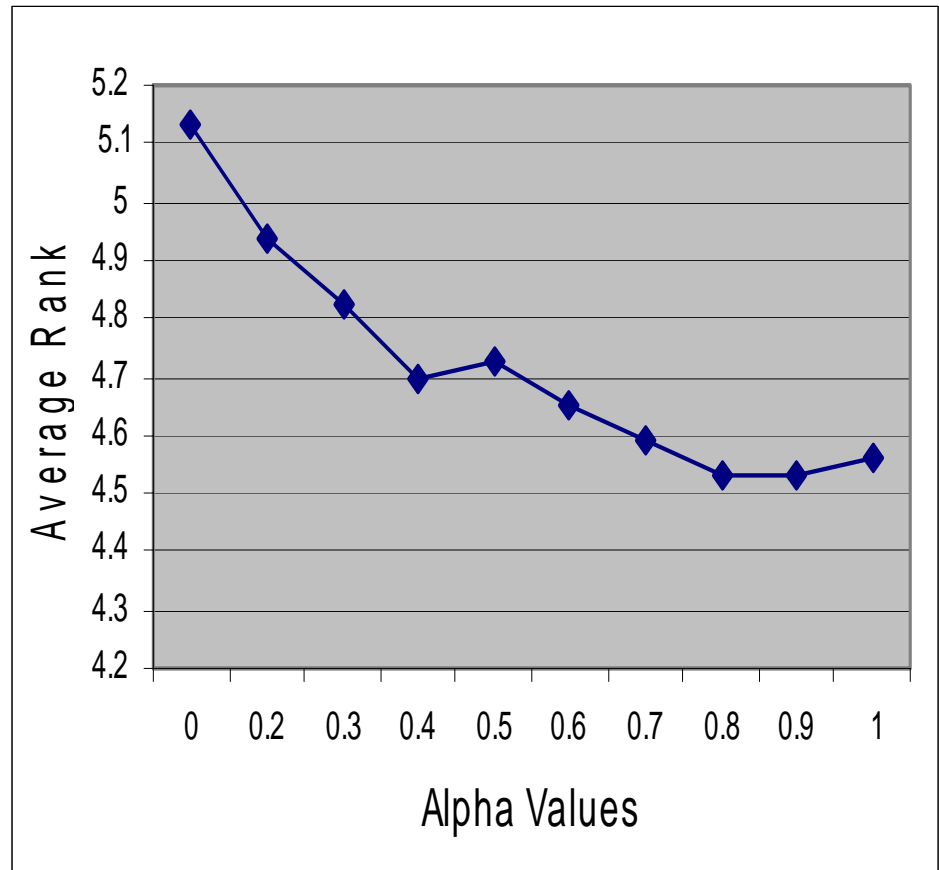
Evaluation Contd....

- 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!**



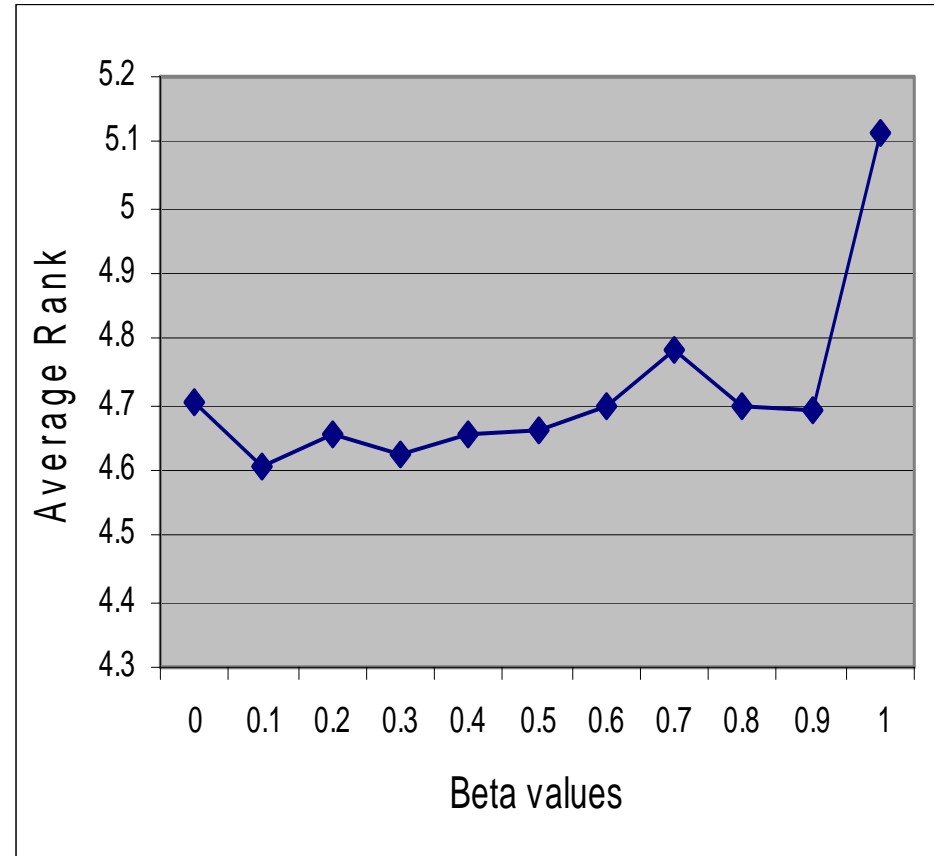
Evaluation Contd...

- 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!