Search Engine Vector Voting

With On The Fly Document Clustering

Aparna PL (14IT132)

Neha B (14IT224)

Prerana K R (14IT231)

S S Karan (14IT252)

Introduction

- Common Techniques used by search engines:

Sl No.	Technique	Drawback
1	IR (Information Retrieval)	Keyword spamming
2	HVV (Hyperlink Vector Voting)	Does not distinguish between high and low quality pages.
3	Henzinger's PageRank algorithm	High computing power

Purpose

- To present a new web page recommendation system to help users reduce navigational time on the Internet, based on the *primacy effect* of browsing behaviour.
- Collect and rate relevant pages, by collecting search results from different search engines, *metasearch* .

Problem Statement

- To improvise the current result set produced by various search engines by presenting a novel clustering algorithm, called *On-The-Fly Document Clustering* (OTFDC), which produces a number of candidate clusters from other web search results. The candidate clusters so produced, not only generate a connective relation between the clusters, but also the relation is a semantic one.

Literature Review

- *NECI metasearch engine*–Lawrence and Giles(1998): analyses each document and displays local context around the query terms.
- Glover's search engine(1999): Metasearch engine architecture, which customises the searching and result ranking strategies based on user's information need.
- Svidzinska's *two-tier metasearch engine(2001)*: *1*-Collect information from topic specific search engines. 2, expand and route user queries to a subset of it.
- *HuddleSearch*-Osdin(2002): New clustering algorithm, which dynamically organises relevant documents into a traversable hierarchy from general to specific.
- Hai(2004): Extracting information from web search interfaces of e-commerce search engines, used to construct e-commerce metasearch engines.

Algorithms and Crux in the Papers

This section describes the proposed search methods in detail. First, the underlying **user behaviour function** is presented. Then, the formulation and **implementation of SVV** is described.

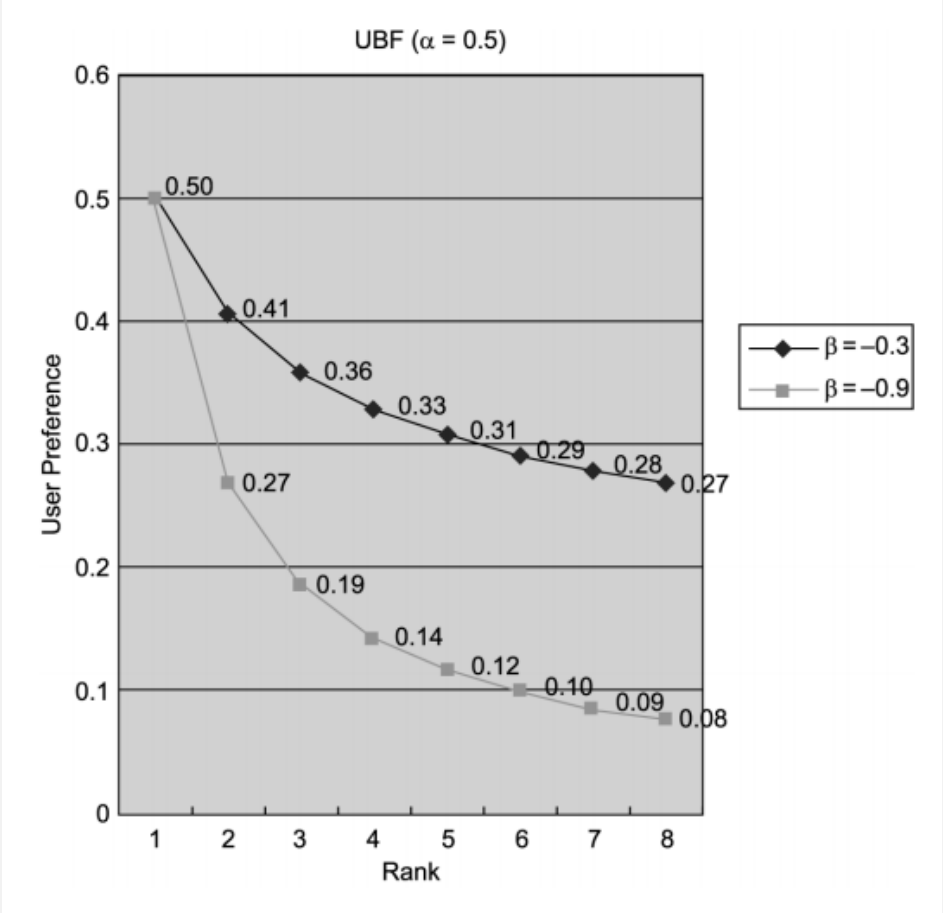
User Behaviour Function

- Primacy effect in psychology
- User behaviour function (UBF) for the i-th item li within an ordered item list l is defined as follows:

$$UBF(l, l_i) = \alpha i^{\beta} (\text{where } \beta < 0)$$

- where α denotes the user's preference of the first item, which represents the user's first impressions on the item list, and β denotes the user preference decay factor.

User Behaviour Function (Contd.)



Search engine vector voting(SVV)

- SVV is based on the voting concept that a web page's ranking is dependent on how several selected search engines rank it, rather than its own contents. A web page wins a vote from a particular search engine if it is listed in the search engine's results to a given query.
- SVV currently gathers top 10 items of search results from each of the following four well-known search engines, Yahoo, Bing, Ask and AOL.
- The SVV search method rearranges the returned web pages based on their weights. The weight of a particular web page considers both the voting tendencies of the 4 search engines mentioned above and the user behaviour function (UBF)

Weight and Vote Distribution of Webpage

- The weight of a web page p for a user query q is defined as follows:

$$w_{p,q} = \sum_{i=1}^{4} \alpha_{i,q} x_{i,p,q}^{\beta}$$

- According to this definition, a web page clearly has larger weight if it either wins votes from more search engines or is ranked high in the results of at least one search engine.
- To understand the voting tendency of the four search engines on a particular web page, a web page's vote distribution is also provided with the following formula:

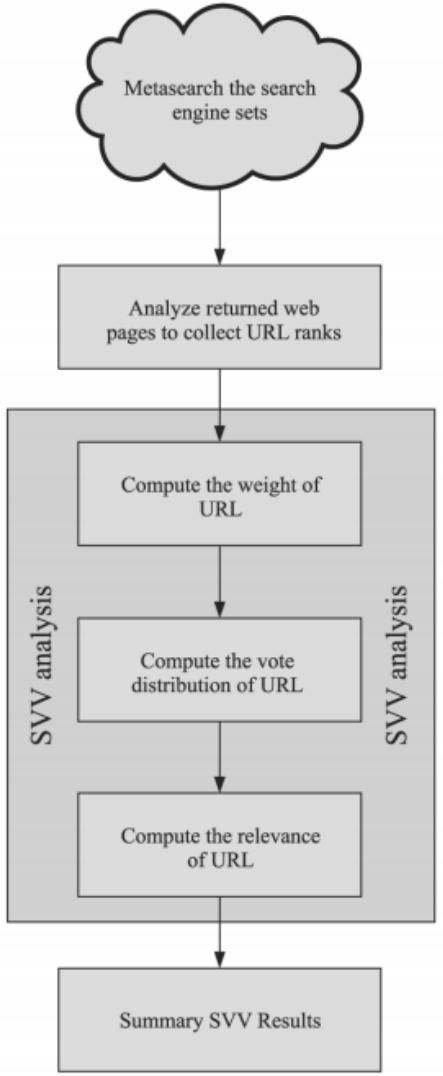
$$P_{p,q} = w_{p,q} / \sum_{i=1}^{4} \alpha_{i,q}$$

Relevance of Webpage

- The vote distribution alone is insufficient for measuring the relevance of returned web pages to a given query. For instance, occasionally the votes on a web page are sparsely distributed among the six search engines, and the resulting almost empty rectangle may lead the user to misinterpret its relevance. Thus the relevance of a web page to a given query is also defined as follows:

$$R_{p,q} = \begin{cases} High, & w_{p,q} > \bar{w} + n\sigma_w \\ Middle, & \bar{w} < w_{p,q} \leq \bar{w} + n\sigma_w \\ Low, & otherwise. \end{cases}$$

- n denotes any value greater than 1, we have taken n=2.



On The Fly Document Clustering (OTFDC)

- This clustering algorithm, called On-The-Fly Document Clustering (OTFDC), produces a number of candidate clusters from other web search results. The candidate clusters not only generate a connective relation between the clusters, but also the relation is a semantic one.
- In OTFDC analysis, we also assumed the candidate clusters derived from an important web document were more likely to be related to the source cluster. Thus, OTFDC only considers the candidate cluster t when the weight of a web document $w_{pt,m}$ is greater than a Threshold. To realize the similarity degree of any two clusters, OTFDC uses the following equation to calculate the correlation coefficient between any two clusters:

$$R(Cluster_i, Cluster_t) = \min(Norm_i, Norm_t)/Norm_i$$

Reverse Links and Building Ontologies

- Three features:
  - SR
    a) Equivalence
    b) Hierarchy, and
    c) Association.
  - HSR
  - CHSR

OTFDC Algorithm

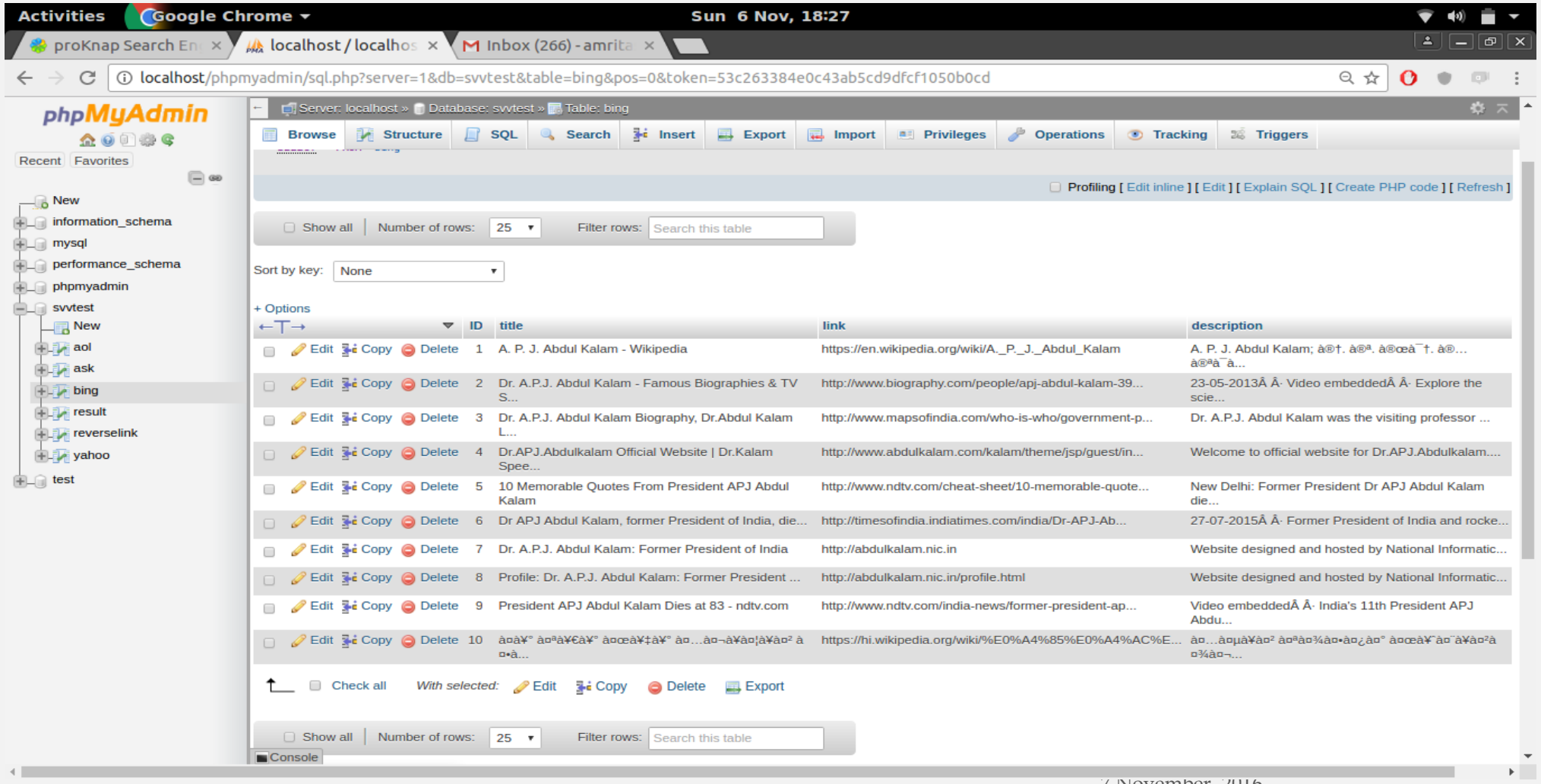
The pseudo code of OTFDC_Core is listed as follows:

```
1   Algorithm OTFDC_Core (Cluster_i as String)
2   {
3   (wp_{i,m}, w_{wp_{i,m},i}) = Call SVV (Cluster_i);
4   (wp_{i,m}, w_{wp_{i,m},i}) = Sort wp_{i,m} according to its weight w_{wp_{i,m},i};
5   Normalize the weight w_{wp_{i,m},i};
6   (CandidateCluster_t) = Use the reverse links of an important web document wp_{i,m} to find out all the candidate
    clusters.
7   Foreach (CandidateCluster_t)
8   {
9   (wp_{t,m}, w_{wp_{t,m},t}) = Call SVV (CandidateCluster_t);
10  (wp_{t,m}, w_{wp_{t,m},t}) = Sort wp_{t,m} according to its weight w_{wp_{t,m},t};
11  If (w_{wp_{t,m},t} < T)
12     Continue;
13  Normalize the weight w_{wp_{t,m},t};
14  Compute the correlation coefficient;
15  If (R(Cluster_i,CandidateCluster_t) ⩾ T and CandidateCluster_t ∉ FinalClusterSet)
16  {
17     Append CandidateCluster_t into FinalClusterSet;
18     Append R(Cluster_i,CandidateCluster_t) into FinalCorCoefSet;
19  } End of If;
20  } End of Foreach;
21  (FinalClusterSet, FinalCorCoefSet) = Sort FinalClusterSet according to its correlation coefficient FinalCorCoefSet;
22  Return (FinalClusterSet, FinalCorCoefSet);
23  } End of Algorithm;
```

Methodology

- 4 popular search engines were chosen –Bing, Yahoo, Ask, AOL
- Assigned weights to each of them based on popularity
- HTML pages of web results were collected using BeautifulSoup and Urllib libraries in Python
- Title, link, description were collected using information of their enclosing HTML tags
- Results from each search engine were stored into separate tables in a database using MySQL

Methodology (continued)

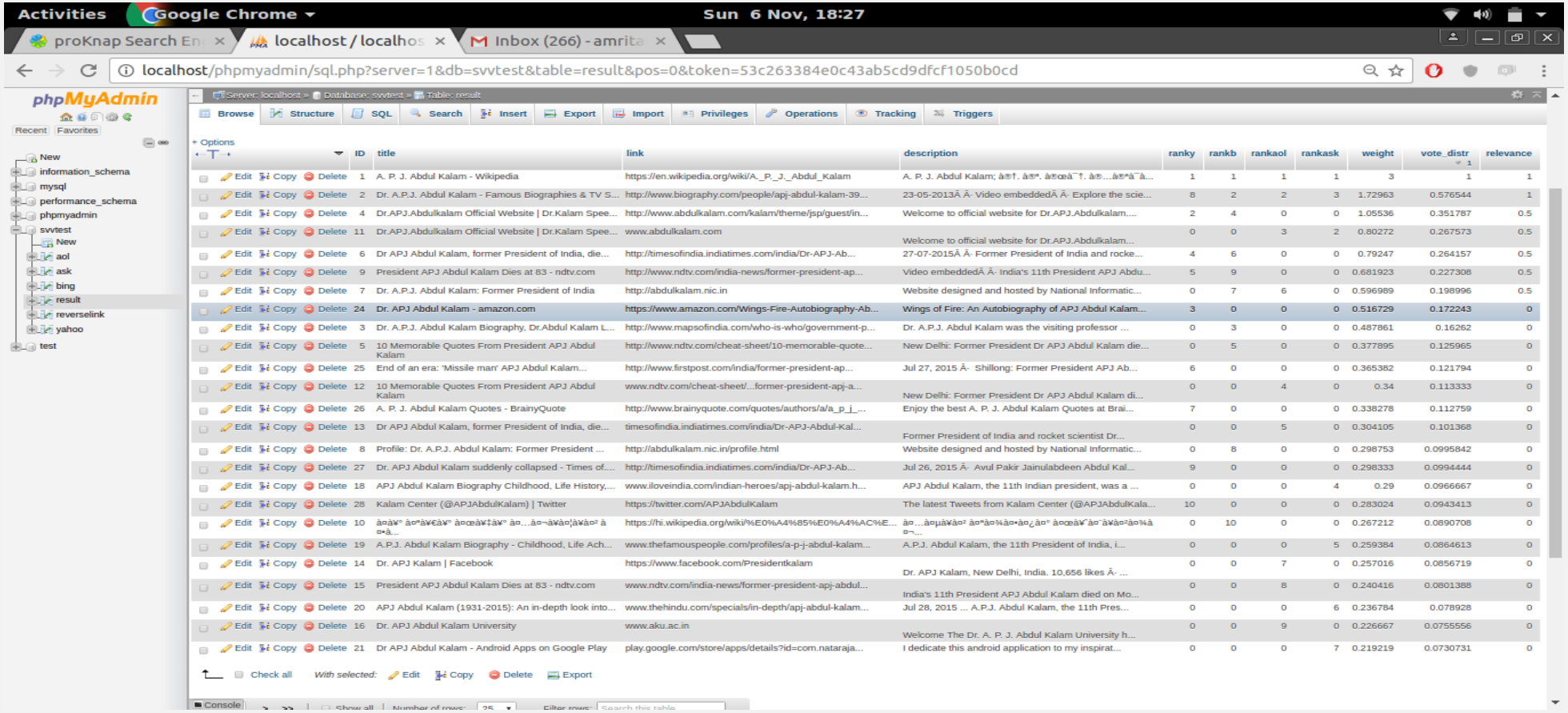


Methodology (continued)

- Serial Scraping of all 4 search engines : 20 seconds
- Parallel Scraping: 10 seconds
- Parallelization achieved through Python's Multithreading library

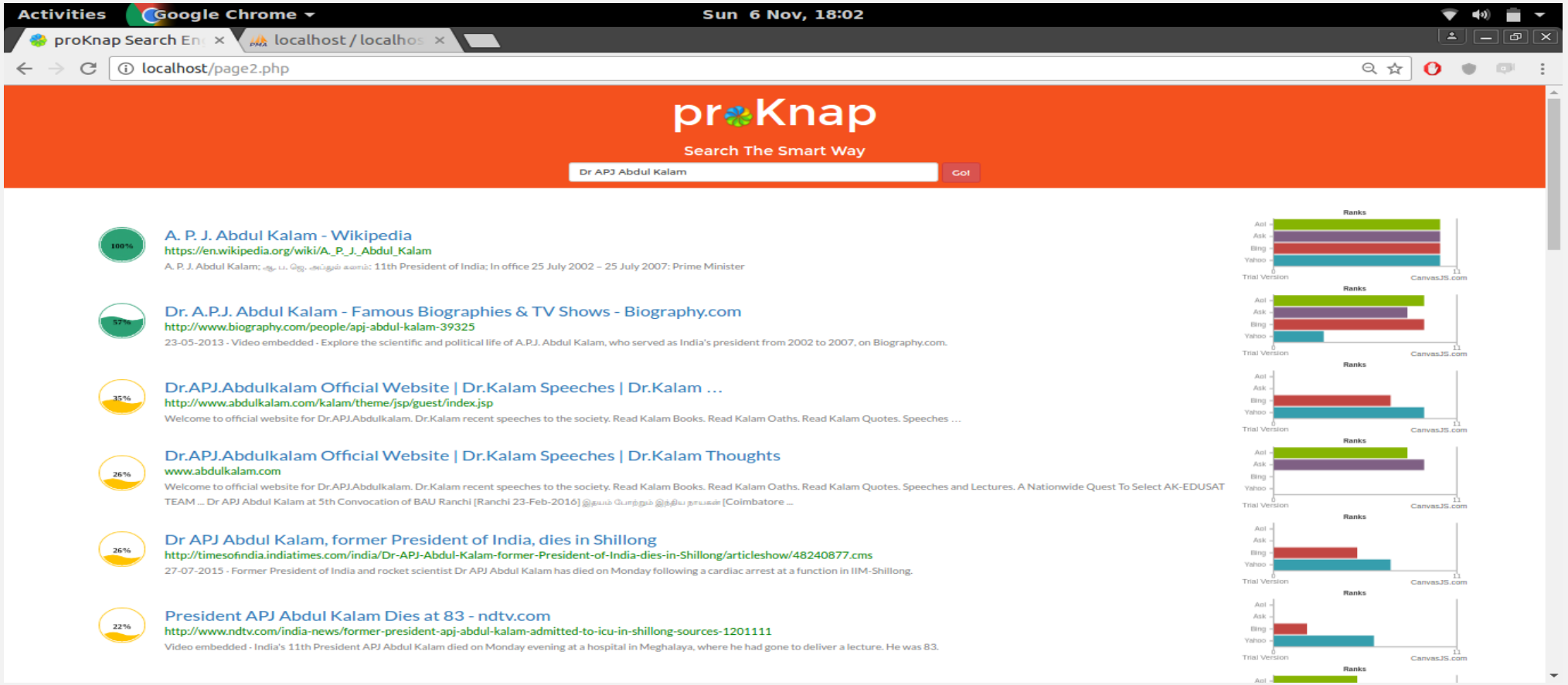
Methodology (continued)

Results were combined, and weights and vote distribution was calculated and stored



Methodology (continued)

- GUI designed with HTML, CSS, JavaScript and PHP
- All results sorted according to their weights and displayed



Approaches / Challenges faced in SVV

- Scraping Google
- Tried to use search engine API
- Removing video and image results
- Making URLs uniform
- Time consuming to scrape 4 search engines

Methodology (Continued)

- Implemented OTFDC in Python

Challenges:

- Get reverse links
- Time Consuming
- Most important link not present in successive iterations, or relevance does not exceed threshold

Solution:

- Used semantically related queries based on HSR values
- Use OTFDC on closed domain

Results

- The SVV algorithm was successfully implemented
- Some of the queries that received successful output are
  - Who
  - When
  - What
  - Why
  - Which
  - How
  - Named Entities

Activities Google Chrome Sun 6 Nov, 18:31

proKnap Search En | localhost / localhos | FINAL WTA - amrita

localhost/page2.php

proKnap

Search The Smart Way

how to go to italy Go!

80%

Italy Travel and Vacation Guide

http://goitaly.about.com

Italy Travel Planning and General Tourism Information; Maps of Italy and Italian Regions; Florence Italy Travel Essentials and Visitor Information

41%

Italy: Getting to Italy and getting around - TripAdvisor

https://www.tripadvisor.com/Travel-g187768-c6673/Italy:Getting.To.Italy.And.Getting.Around.html

Inside Italy: Getting to Italy and getting around - Before you visit Italy, visit TripAdvisor for the latest info and advice, written for travelers by travelers.

40%

How Can I Get a Visa for Travel to Italy?

goitaly.about.com/od/FAQ/f/italy-visa-requirements.htm

Question: How Can I Get a Visa for Travel to Italy? If you're traveling to Italy and need a visa or are not sure if you need a visa, check this information about visa requirements for Italy as the answer varies depending on your nationality, country of residence, the length of time you plan to stay, and the reason for your stay in Italy.

40%

Italy travel guide - Wikitravel

http://wikitravel.org/en/Italy

Italy is a country in Southern Europe. Together with Greece, it is acknowledged as the birthplace of Western culture. Not surprisingly, it is also home to the greatest number …

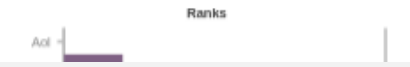
33%

Italy Travel Guide by Rick Steves

https://www.ricksteves.com/europe/italy

Explore Italy! Get inspired with Rick Steves' recommended places to go and things to do, with tips, photos, videos, and travel information on Italy.

Places to Visit in Italy | Where to go in Italy | Rough Guides



33%

Italy Travel Guide by Rick Steves

https://www.ricksteves.com/europe/italy

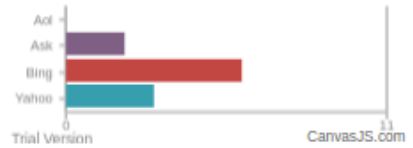
Explore Italy! Get inspired with Rick Steves' recommended places to go and things to do, with tips, photos, videos, and travel information on Italy.

29%

Places to Visit in Italy | Where to go in Italy | Rough Guides

https://www.roughguides.com/destinations/europe/italy

Explore Italy with Rough Guides: find out the best places to visit, where to go, view itineraries and read about Rome, Venice, pizza and Pompeii.



19%

Planning your first trip to Italy - Lonely Planet

www.lonelyplanet.com/italy/travel-tips-and-articles/planning-your-first-trip-to-italy

Oct 30, 2015 ... Where to go? How to travel? Here's everything you need to know to get the absolute utmost out of your first-time Italy trip. One-of-a-kind Venice ...

16%

The Best Time to Go to Italy • The Blonde Abroad

http://theblondeabroad.com/2016/06/02/best-time-go-italy

Ciao, bella! Ready for the ultimate Mediterranean getaway? From Milan to Sicily, Italy is one of the best destinations on the planet. With its fascinating history, breathtaking …

16%

Italy: Getting to Italy and getting around - TripAdvisor

https://www.tripadvisor.com/Travel-g187768-c6673

Inside Italy: Getting to Italy and getting around - Before you visit Italy, visit TripAdvisor for the latest info and advice, written for travelers by travelers.

14%

How to Travel Italy on a Budget • The Blonde Abroad

http://theblondeabroad.com/2015/10/13/how-to-travel-italy-on-a-budget

Italy isn't the cheapest European country to explore, so here are my tips to save money, travel Italy on a budget and have an amazing time!

12%

Italy Travel Guide - Travel to Italy

http://www.justitaly.org

Italy is a country full of culture that every tourist should visit; here we offer a Travel guide about Italy which includes facts, maps, history, ...

Activities Google Chrome Sun 6 Nov, 17:55

proKnap Search En × localhost / localhos ×

localhost/page2.php

proKnap

Search The Smart Way

why is the sky blue Go!

90%

Why is the Sky Blue ? - Science Made Simple

http://www.sciencemadesimple.com/sky_blue.html

Learn about the atmosphere and how light scattering makes the sky look blue. MIT used parts of this article in their online class. Why is the Sky Blue? by Science Made …

57%

Why is the sky blue? :: NASA Space Place

http://spaceplace.nasa.gov/blue-sky

27-10-2016 - Why is the sky blue? It's easy to see that the sky is blue. Have you ever wondered why? A lot of other smart people have, too. And it took a long time to figure …

42%

Why is the sky Blue? - University of California, Riverside

http://www.math.ucr.edu/home/baez/physics/General/BlueSky/blue_sky.html

Why is the sky blue? A clear cloudless day-time sky is blue because molecules in the air scatter blue light from the sun more than they scatter red light. When we look …

29%

Why Is the Sky Blue ? - Live Science - Blue Sky

http://www.livescience.com/32511-why-is-the-sky-blue.html

Why is the sky blue? The light coming from the sun is made of many colors, each of which has a different wavelength.

28%

Why is the Sky Blue - Space Black - Sunset Red?

http://www.sciencemadesimple.com/space_black_sunset_red.html

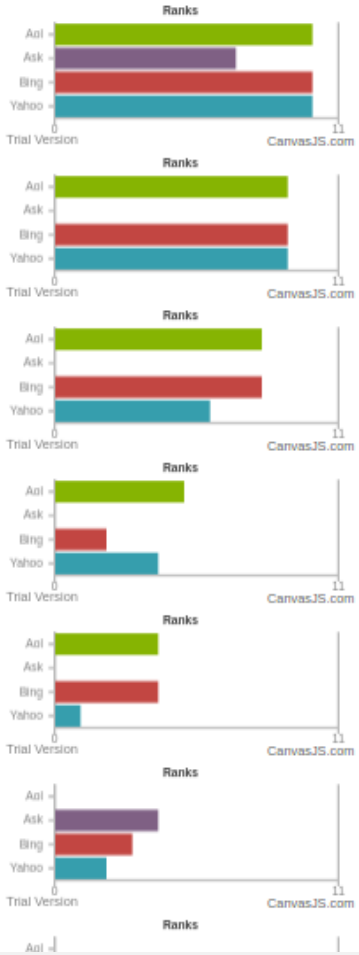
Why is the Sky Blue, Space Black and Sunset Red? by Science Made Simple. Home; Science Projects; Science Fair; Articles; Store; Metric Conversions; ... Once you …

27%

Sky - Wikipedia

https://en.wikipedia.org/wiki/Sky

Media related to Sky at Wikimedia Commons; Why is the sky blue? ...



Activities Google Chrome Sun 6 Nov, 18:00

proKnap Search En × localhost / localhos ×

localhost/page2.php

proKnap

Search The Smart Way

which bird flies backward Go!

94%
What Bird Can Fly Backwards? - It May Suprise You
http://www.ponderweasel.com/what-bird-can-fly-backwards
To fly backwards is quite a challenge it turns out. With only a fraction of 1% of birds being able to do so, which among the birds has the agility to do it?

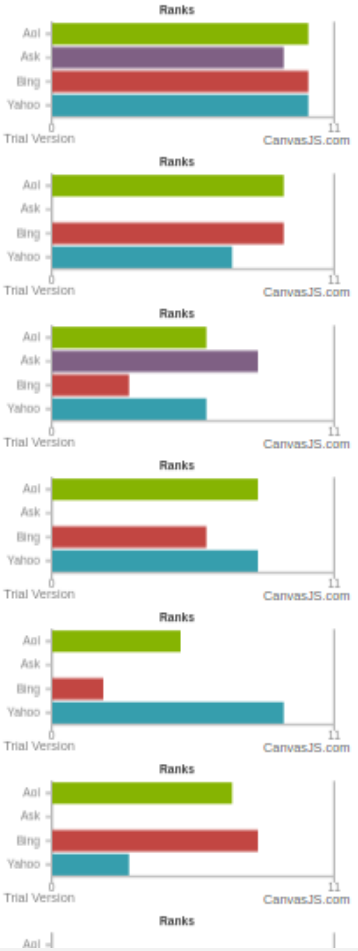
50%
What Is The Only Bird That Can Fly Backwards - Uttertrivia
http://www.uttertrivia.com/birdthatfliesbackwards.php
The Hummingbird is the only bird that can fly backwards. All other birds use the down stroke to power their flight. Hummingbirds can fly up and down, left and right, …

44%
Only two birds can fly backwards. - Sheffield Forum
http://www.sheffieldforum.co.uk/showthread.php?t=758276
Only two birds can fly backwards. General Discussions. Site Contents. Home Page; Sheffield Businesses; Advertise on SF; Forum Team List; Help ... One is the …

42%
This Kind of Bird Flies Backward: Paintings by Joan Brown | …
http://sjmusart.org/joan-brown
This Kind of Bird Flies Backward is presented in honor of Barbara and Ken Oshman.

39%
What is the only bird that can fly backwards - Answers.com
http://www.answers.com/Q/What_is_the_only_bird_that_can_fly_backwards
The only bird that can fly backwards is the humming bird' 14 people found this useful Edit. Share to: On the Wallaby. Passionate about all things Australian. Which birds …

38%
Bird flies backwards - YouTube
http://www.youtube.com/watch?v=QR_3eemXA_U
19-10-2011 · Video embedded · In my science book they say only humming birds can fly backwards.They are wrong



Activities Google Chrome Sun 6 Nov, 17:54

proKnap Search En | localhost / localhos

localhost/page2.php

proKnap

Search The Smart Way

when is the us presidential election results Go!

100%

United States presidential election - Wikipedia

https://en.wikipedia.org/wiki/United_States_presidential_election

United States presidential election. This article is part of a series on the: Politics of the United States of America; Federal Government. Constitution of the United States; …

56%

2016 Presidential Election Results - POLITICO

http://www.politico.com/2016-election/results/map/president

POLITICO's Live 2016 Election Results and Maps by State, County and District. Includes Races for President, Senate, House, Governor and Key Ballot Measures.

47%

Presidential Elections Data: - presidency.ucsb.edu

http://www.presidency.ucsb.edu/elections.php

… Project contains the most comprehensive collection of resources pertaining to the study of the President of the United States. Compiled by John Woolley and Gerhard …

41%

Presidential election, 2016 - Ballotpedia

https://ballotpedia.org/Presidential_election,_2016

Federal elections. Presidential election, 2016; U.S. Congress, 2016; U.S. Senate, 2016; U.S. House, 2016; State elections. State executive, 2016; State legislative, …

32%

Election Results - Dave Leip's Atlas of U.S. Presidential Elections

http://uselectionatlas.org/RESULTS

Detailed national results of US Presidential Elections from 1789 through 2012, US Senate and Gubernatorial Elections since 1990. Site includes election data, county …

19%

United States presidential election, 2012 - Wikipedia

https://en.wikipedia.org/wiki/United_States_presidential_election,_2012

Presidential election results map. Red denotes states won by Romney/Ryan (24), Blue denotes those won by Obama/Biden (26+D.C.). Numbers indicate electoral votes …

Activities Google Chrome Sun 6 Nov, 17:52

proKnap Search En × localhost / localhos ×

localhost/page2.php

proKnap

Search The Smart Way

what is the capital of switzerland Go!

100%

Switzerland - Wikipedia

https://en.wikipedia.org/wiki/Switzerland

Location of Switzerland (green) in Europe (green & dark grey) Capital: None Bern : Largest city: Zürich: Official languages: German French Italian

60%

The Capital of Switzerland - Bern Tourism

http://www.bern.com/en/trade-travel/reasons-for-bern/capital

The capital of Switzerland. The city was founded in 1191 by the Duke of Zähringen, is known for its arcades (Lauben) and some of it has been preserved in its original state.

48%

What is the Capital of Switzerland? - Capital-of.com

http://www.capital-of.com/Switzerland

What is the Capital city of Switzerland? Well, the Capital city of Switzerland is... Find out what is the capital city of countries around the world. Capital Cities Capitals of …

35%

Bern - Wikipedia

https://en.wikipedia.org/wiki/Bern

... is the de facto capital of Switzerland, referred to by the Swiss as their (e.g. in German) Bundesstadt, or "federal city". With a population of 141,107 (April 2016), …

34%

What is the capital of Switzerland? - Quora

https://www.quora.com/What-is-the-capital-of-Switzerland

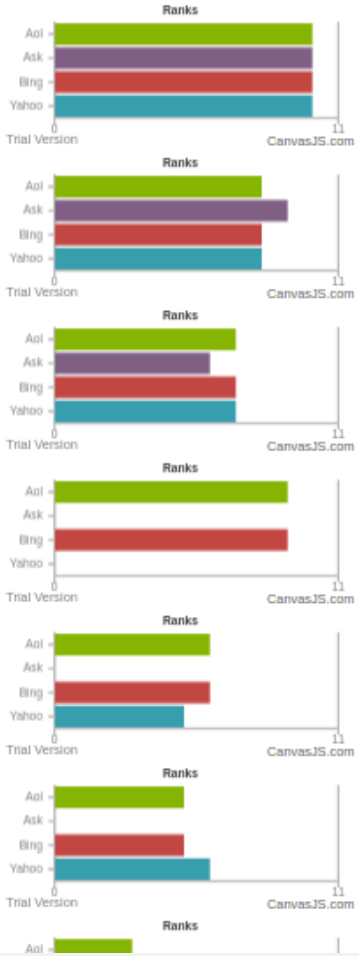
Answer (1 of 2): Switzerland does not have a capital. Instead it has a federal city, where most of the federal government (parliament, executive, administration) is located. …

34%

What is the capital of Switzerland - Answers.com

http://www.answers.com/Q/What_is_the_capital_of_Switzerland

30-10-2016 · Answers.com WikiAnswers ® Categories Travel & Places Countries, States, and Cities Switzerland What is the capital of Switzerland? What would you …



Activities Google Chrome Sun 6 Nov, 17:48

proKnap Search En × localhost / localhos ×

localhost/page2.php

proKnap

Search The Smart Way

who is nelson mandela Go!

100%

Nelson Mandela - Wikipedia

https://en.wikipedia.org/wiki/Nelson_Mandela

Nelson Rolihlahla Mandela (/ m æ n ˈ d ɛ l ə /; 18 July 1918 – 5 December 2013) was a South African anti-apartheid revolutionary, politician, and …

50%

Who is Nelson Mandela? (with pictures) - wiseGEEK

http://www.wisegeek.org/who-is-nelson-mandela.htm

Nelson Mandela was South Africa's first democratically elected president. A long-time apartheid protester, Mandela was actually...

48%

Nelson Mandela - Facts & Summary - HISTORY.com

http://www.history.com/topics/nelson-mandela

Video embedded · Find out more about the history of Nelson Mandela, including videos, interesting articles, pictures, historical features and more. …

44%

Nelson Mandela - Biography.com

http://www.biography.com/people/nelson-mandela-9397017

Watch video · Nelson Mandela became South Africa's first black president in 1994, following a 20-year anti-apartheid campaign. Read more about his life and legacy, and …

25%

Biography – Nelson Mandela Foundation

https://www.nelsonmandela.org/content/page/biography

Introduction Rolihlahla Mandela was born into the Madiba clan in the village of , Transkei, on 18 July 1918. His mother was Nonqaphi Nosekeni and his father was Nkosi …

23%

BBC - Primary History - Famous People - Nelson Mandela

http://www.bbc.co.uk/schools/primaryhistory/famouspeople/nelson_mandela

Why is Nelson Mandela famous? A leader of his people Nelson Mandela was the first black President of South Africa. He spent 27 years in prison for trying to overthrow …

Activities Google Chrome Sun 6 Nov, 18:02

proKnap Search En × localhost / localhos ×

localhost/page2.php

proKnap

Search The Smart Way

Dr APJ Abdul Kalam Go!

100%

A. P. J. Abdul Kalam - Wikipedia

https://en.wikipedia.org/wiki/A._P._J._Abdul_Kalam

A. P. J. Abdul Kalam; ஆ. பி. ஜெ. அப்துல் கலாம்: 11th President of India; In office 25 July 2002 – 25 July 2007: Prime Minister

57%

Dr. A.P.J. Abdul Kalam - Famous Biographies & TV Shows - Biography.com

http://www.biography.com/people/apj-abdul-kalam-39325

23-05-2013 · Video embedded · Explore the scientific and political life of A.P.J. Abdul Kalam, who served as India's president from 2002 to 2007, on Biography.com.

35%

Dr.APJ.Abdulkalam Official Website | Dr.Kalam Speeches | Dr.Kalam …

http://www.abdulkalam.com/kalam/theme/jsp/guest/index.jsp

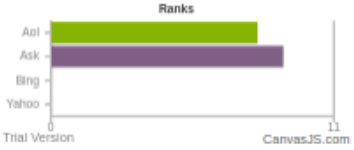
Welcome to official website for Dr.APJ.Abdulkalam. Dr.Kalam recent speeches to the society. Read Kalam Books. Read Kalam Oaths. Read Kalam Quotes. Speeches …

26%

Dr.APJ.Abdulkalam Official Website | Dr.Kalam Speeches | Dr.Kalam Thoughts

www.abdulkalam.com

Welcome to official website for Dr.APJ.Abdulkalam. Dr.Kalam recent speeches to the society. Read Kalam Books. Read Kalam Oaths. Read Kalam Quotes. Speeches and Lectures. A Nationwide Quest To Select AK-EDUSAT TEAM ... Dr APJ Abdul Kalam at 5th Convocation of BAU Ranchi [Ranchi 23-Feb-2016] இதயம் போற்றும் இந்திய நாயகன் [Coimbatore ...



26%

Dr APJ Abdul Kalam, former President of India, dies in Shillong

http://timesofindia.indiatimes.com/india/Dr-APJ-Abdul-Kalam-former-President-of-India-dies-in-Shillong/articleshow/48240877.cms

27-07-2015 · Former President of India and rocket scientist Dr APJ Abdul Kalam has died on Monday following a cardiac arrest at a function in IIM-Shillong.

22%

President APJ Abdul Kalam Dies at 83 - ndtv.com

http://www.ndtv.com/india-news/former-president-apj-abdul-kalam-admitted-to-icu-in-shillong-sources-1201111

Video embedded · India's 11th President APJ Abdul Kalam died on Monday evening at a hospital in Meghalaya, where he had gone to deliver a lecture. He was 83.

Conclusion

- A survey conducted by authors of the paper, showed that users were more satisfied with the proposed search method than general search engines
- Ultimately the ‘goodness’ of the search results depends on the user behavior input given

Future Work

- Results can be improved by using ontologies that would be created when the net becomes fully semantic
- Collect user feedback to get accurate values of α and β
- Reduce execution time

Thank you