A Mini-project Report

on

Search Engine Vector Voting with OTFDC

carried out as part of the course Semantic Web Technologies (IT412)

Submitted by

Aparna PL (14IT132)
Neha B (14IT224)
Prerana K R (14IT231)
S S Karan (14IT252)

V Sem B.Tech (IT)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

Department of Information Technology

National Institute of Technology Karnataka, Surathkal

Jul - Dec 2016

CERTIFICATE

This is to certify that the project entitled **"Search Engine Vector Voting with OTFDC"** is a bonafide work carried out as part of the course **Semantic Web Technologies (IT412)**, under my guidance by

1. Aparna PL 14IT132
2. Neha B 14IT224
3. Prerana K R 14IT231
4. S S Karan 14IT2522

students of V Sem B.Tech (IT) at the Department of Information Technology, National Institute of Technology Karnataka, Surathkal, during the academic semester ____________Jul - Dec 2016________ in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology, at NITK Surathkal.

Place:

Date: Signature of the Instructor

DECLARATION

We hereby declare that the project entitled "Search Engine Vector Voting with OTFDC" submitted as part of the partial course requirements for the course Semantic Web Technologies (IT412) for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal during the ________Jul-Dec 2016________ semester has been carried out by us. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Further, we declare that we will not share, re-submit or publish the code, idea, framework and/or any publication that may arise out of this work for academic or profit purposes without obtaining the prior written consent of the Course Instructor.

Signature of the Students:

Place:

Date:

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Document clustering is an important technology that can automatically discover groups of similar web documents in a set of the web documents returned by a search engine (Berkhin, 2002). Organizing web documents by clustering can help the user get a sense of the major subject areas covered in the web document set and help the user find the relevant web documents more quickly. To get the set of relevant documents, Search Engine Vector Voting (SVV) algorithm was used and the documents were ranked according to their vote distribution. Search engine Vector Voting (SVV), a MetaSearch engine, can simultaneously retrieve the search results from several search engines.

1.1 Motivation

Search engines generally use information retrieval (IR) techniques to find web pages or documents relevant to a user query in a database of web pages. Traditional IR methods measure the similarity between user query and document contents by models including vector space models, probability models, and fuzzy logic models. These methods are primarily based on the keyword matches and their frequency in documents, and therefore easily result in “keyword spamming”. That is, a document’s rank can be manipulated by duplicating the same set of keywords several times in a document. Collecting search results from several search engines is an effective way to collect and rate relevant web. This method is called metasearch. The metasearch approach saves a lot of time by searching only in one place and eliminating the need to use and learn several separate search engines. The quality of metasearch results depends on the search engines used and how they organize the results. In this project, a metasearch method, called Search Engine Vector Voting (SVV), was developed to rate the search results obtained from several well-known search engines. The relevance of a web page to a given query is determined by its frequency and rankings in the search results returned by each search engine. This result is the base for the document clustering presented in this project.

2 LITERATURE REVIEW

This section reviews some important literature related to the paper. First, we provide a literature review on different concepts of the document clustering. Second, we provide a short description of Search Engine Vector Voting (SVV).

2.1 Different Concepts of Document Clustering

Document clustering is an important data exploration technique that groups similar documents into a cluster. However, document clustering is a difficult problem since the documents contain large amounts of unstructured. Many researchers use different concepts and methods to solve the clustering problem. Some researchers have used the fuzzy concept to solve the clustering problem. Some others have combined the possibilistic and fuzzy formulations of co-clustering for automatic categorization of large document collections. The combined approach allows the word memberships to represent fuzzy word partitions, however it captures the natural word clusters structure. Saraçog ˘lu, Tütüncü, and Allahverdi (2007) designed a two-layer structure and let the documents pass through it to find the similarities. In two-layer structures, they used predefined fuzzy clusters to extract feature vectors of related documents. The similarity measure is estimated based on these feature vectors. Other researchers have used the domain concept to solve the clustering problem. Kerne, Koh, Sundaram, and Mistrot (2005) presented an iterative method for generative semantic clustering of related terms in the documents. Their method for generating semantic clustering is based on a quantitative model that represents mutual information between each new domain and the domains already in the document. Further researchers have used the query concept to solve the clustering problem. Chang, Kim, and Raghavan (2006) constructed the query concepts to express the user’s information needs, rather than trying to reformulate the initial queries. To construct the query concepts, they extracted all document features from each document and then clustered these document features into primitive concepts that are used to form query concepts. Li, Chung, and Holt (2008) used the query sequence to rearrange clustering results. They claimed a query sequence is frequent if it occurs in more than a certain percentage of the documents in all document sets.

2.2 Metasearch

Unlike general search engines (Dreilinger and Howe, 1996; Selberg and Etzioni, 1997;Meng et al., 2002; Zacharis and Panayiotopoulos, 2002; Hai et al., 2004), metasearch engines transmit the keywords submitted in its search box simultaneously to several search engines and present the search results in an integrated format as it operates on the premise that the Web is too large for any one search engine to index it all and that more comprehensive search results can be obtained by combining the results from several search engines. The format lets the users see at a glance which particular search engine returned the most relevant web pages retrieved for a query without having to search each engine individually. Such finding could be used to adjust the rank order of results. Thus, metasearch can significantly save searching time and eliminate the need to use and learn several separate search engines (Hu et al., 2001). So, the users can now just enter their query into a metasearch engine and can get a set of relevant documents that have been gathered from several search engines and ranked according to some logic. The quality of metasearch results depends on which search engines they search i.e. get results from for a given query and how they integrate the results i.e. what factors are taken into considering while giving each web document an aggregate rank and whether the document must be considered at all or not.

Several related studies have been conducted on metasearch engines. Lawrence and Giles (1998) proposed a NECI metasearch engine, which can analyze each downloaded document and then display results with the query term shown in a specific context. This format helps users readily determine whether the document is relevant without having to download each document. Glover et al. (1999a) described a metasearch engine architecture which customizes the searching and results ranking strategies based on the user's information need. Compared with a regular metasearch engine, which sends a query to a pre-defined list of search engines and ranks the results in a pre-defined order, this method allows much greater personalization by providing customized ranking of search results from a tailored list of search engines. Zacharis and Panayiotopoulos (2002) proposed a Webnaut metasearch system, which can learn the user's interests and adapt them appropriately as these interests change over time. Braslavski et al. (2004) presented ProThes, which consists of a metasearch engine, a graphical user interface for query specification, and a thesaurus-based query customization system. ProThes also provides simple heuristics for result merging and partial re-ranking.

2.3 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. These candidate clusters are obtained from a metasearch engine that implements the Search Engine Vector Voting algorithm.

2.4 Research Objectives

The proposed model can be used to improve the search performance of any search engine. Organizing web documents by clustering can help the user get a sense of the major subject areas covered in the web document set and help the user find the relevant web documents more quickly. The metasearch approach used for collecting web documents saves a lot of time for the user as the user has to search by only in one place which eliminates the need to use and learn several separate search engines.

3 METHODOLOGY AND FRAMEWORK

3.1 System Architecture

The algorithm to improve clustering performance of search engines requires as input, a list of web documents from various search engines for the given query sorted based on their weights (obtained from their ranking among the various the search engines considered), for which a simple metasearch engine is required.

Implementation of SVV was done in Python and Python's BeautifulSoup4 and Urllib libraries were used. Beautiful Soup is a Python library for pulling data out of HTML and XML files. Urllib module provides a high-level interface for fetching data across the World Wide Web. In particular, the urlopen() function is similar to the built-in function open(), but accepts Universal Resource Locators (URLs) instead of filenames. Queries to the search engine was parallelized to achieve quick results using Python's Multithreading library. Serial querying took approximately 20s and parallelizing it reduced the time to 10s.

All results were stored into database using MySQL. Tables for results from each of the search engines were created and a table for storing combined results along with their weights and vote distribution was created. For combining Python and SQL, MySQLdb was used. MySQLdb is an interface to the popular MySQL database server that provides the Python database API. Front end was designed using HTML, CSS, JavaScript and PHP.

3.2 Algorithms and Techniques

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. Document clustering is an important technology that can automatically discover groups of similar web documents in a set of the web documents returned by a search engine. Organizing web documents by clustering can help the user get a sense of the major subject areas covered in the web document set and help the user find the relevant web documents more quickly.

3.2.1 Search Engine Vector Voting

A search engine generally returns a list of URLs to a user query in decreasing order of relevance; that is, the most relevant answers are on the top. Consequently, users generally prefer top-ranking web pages over others. This phenomenon is called the primacy effect in psychology. Based on the primacy effect, the user behaviour function (UBF) for the ith item l_i 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. When β is small, the UBF value decreases slowly.

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, AOL and Ask. Advertisements and paid placements, which generally show up on the top, are removed before the search results are collected and processed by SVV. 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 four search engines mentioned above and the user behaviour function (UBF). The flow chart of SVV is shown in Figure 1.

Formally, the weight of a web page p for a user query q is defined as follows:

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

where α denotes the user's preference on the first match (which is generally the most relevant to a user query recommended by the search engine) to a given query q returned from search engine i; $x_{i,p,q}$ denotes the ranking of web page p in search engine i's results for query q, and β denotes the user preference of web pages (where $\beta < 0$). 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 six search engines on a particular web page, a web page's vote distribution is also provided with the following formula:

$$P_{p,q} = w_{p,q} / \sum_{i=1}^{6} \alpha_{i,q}$$

where $P_{p,q}$ denotes the vote distribution of web page p for query q in all four search engines, which is represented as a filled Bubble as in Figure 2. If a web page is listed at the top of all the four search engines for a given query, then the bubble to the left of its URL is full with colour.

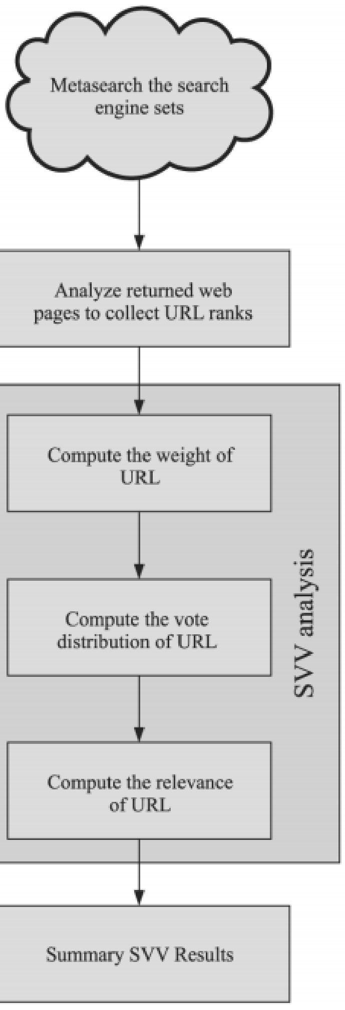

Metasearch the search engine sets
Analyze returned web pages to collect URL ranks
Compute the weight of URL
Compute the vote distribution of URL
Compute the relevance of URL
SVV analysis
SVV analysis
Summary SVV Results

Figure 1: Flowchart of SVV Algorithm

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:

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

where $R_{p,q}$ denotes the relevance of a web page p to a given query q; $\hat{w}$ denotes the average weight of all the web pages returned in this query, and σ_w denotes the standard deviation of the weights of all the returned web pages.

The web pages in this category are defined to have high relevance to this query. By contrast, the web pages whose weights fall below average are considered to have low relevance to the query. Other web pages between these two categories are considered to have medium relevance to the query. The three relevance categories are indicated by different color balls, namely green, yellow, and red balls for high, medium, and low relevance, respectively.

The pseudo code of SVV is listed as follows:

```
Algorithm search (query)
{
     Concurrently collect the search results to the query from
     six search engines;
     Does a page analysis to collect a URL rank from the above
     collection;
     Do a SVV analysis on the results of the page analysis;
     Set the SVV results to be the analysis results;
     Break;
     Do a Summary analysis on the analysis results
     {
          Sort the analysis results according to their weight;
          Layout the results;
          Combine the results to be final results;
          Show final results;
     }
     end of Do;
} end of Algorithm;
```

3.2.2 On The Fly Document Clustering

Now, we present 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 not only generate a connective relation between the clusters, but also the relation is a semantic one.

Definition 1 (SR). ($Cluster_i$,$Cluster_t$) have a semantic relation (SR) feature when $Cluster_i$ and $Cluster_t$ satisfy at least one of the following three relations: (a) equivalence, (b) hierarchy, and (c) association. If there is an equivalent relation between $Cluster_i$ and $Cluster_t$, these two clusters express the same concept. For example, (‘‘photocopies”, ‘‘Xeroxes”) and (‘‘freedom”, ‘‘liberty”) exist in an equivalence relation, respectively. A hierarchical relation is based on degree of super-ordination and subordination, where the superordinate cluster may represent a class or a whole, and the subordinate cluster refer to its members or parts. For example, (‘‘mountain regions”, ‘‘Alps”) and (‘‘nervous system”, ‘‘brain”) exist in a hierarchical relation, respectively. An associational relation includes associations between clusters that are neither equivalent nor hierarchical, though the clusters are semantically associated to such an extent that the relation between them should be made explicit. For example, (‘‘flour”, ‘‘wheat”) and (‘‘Greek civilization”, ‘‘Socratic method”) exist in an association relation, respectively.

Definition 2 (HSR). ($Cluster_i$,$Cluster_t$) exist in a high-level SR (HSR) feature if and only if ($Cluster_i$,$Cluster_t$)exist in a SR feature and T 6 R($Cluster_i$,$Cluster_t$) 6 1, where R($Cluster_i$,$Cluster_t$) is the correlation coefficient between two clusters (as defined later in this section) and T is the threshold value (as defined later in Section 4). HSR implies that $Cluster_i$ and $Cluster_t$ not only exist in a SR feature but also R($Cluster_i$,$Cluster_t$) is greater than or equal to the lower bound of T

Definition 3 (CHSR). (($Cluster_i$,$Cluster_t$) exist in a connective HSR (CHSR) feature if and only if ($Cluster_i$,$Cluster_t$) exist in a HSR feature and ($Cluster_i$,$Cluster_t$) 2 HSR . That is, we can find an intermediate cluster $Cluster_d$ so ($Cluster_i$,$Cluster_d$) ϵ HSR* and ($Cluster_d$,$Cluster_t$) ϵ HSR* , where ($Cluster_i$,$Cluster_d$) and ($Cluster_d$,$Cluster_t$) exist in a HSR feature, respectively.

OTFDC uses the reverse links of an important web document to find all the important clusters (Line 6). 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 T (Lines 11 and 12). To realize the similarity degree of any two clusters, OTFDC uses the following equation to calculate the correlation coefficient between any two clusters:

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

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 ($R(Cluster_i,CandidateCluster_t) \geqslant T$ and $CandidateCluster_t \notin FinalClusterSet$)
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;
```

4 WORK DONE

4.1 Detailed workflow

All steps followed, along with description of tools used and challenges faced during the project are described below.

Steps followed:

1. Four popular search engines were chosen: Bing, Yahoo, Ask, AOL.

2. Weights were assigned to each of them based on their popularity.

3. HTML pages were collected using Urllib library.

4. Required fields (namely title of the search result, link and description) from the obtained HTML pages were obtained using BeautifulSoup library using the enclosing HTML tag of the field.

5. Results obtained were stored into 4 different tables, one for each search engine result in MySQL database using MySQLdb.

6. Since scraping results from all search engines serially was time consuming (~20seconds), Python's multithreading library was used to retrieve results parallel and store them to tables.

7. Results were combined in a new table from all the 4 tables and stored in a new table and ranks of each link in the 4 search engines were stored to calculate weight and vote distribution.

8. User interface was designed using HTML, CSS, Javascript and PHP.

Description of tools used:

Urllib library: urllib is a Python package that collects several modules for working with URLs.

BeautifulSoup library: Beautiful Soup is a Python library for pulling data out of HTML and XML files.

Multithreading library: The multiprocessing module includes a relatively simple API for dividing work up between multiple processes. It is based on the API for threading.

MySQLdb: MySQLdb is an thread-compatible interface to the popular MySQL database server that provides the Python database API.

HTML/CSS/Javascript/PHP: Front-end and back-end tools

Challenges faced:

1. Tried to scrape results from Google using Urllib and BeautifulSoup. However, Google's policy did not allow its results to be scraped.

2. Google has deprecated its web search API and provided custom search. However, the ranking algorithm used for Custom Search is different from the actual ranking algorithm that Google uses. Also, it is not allowed to get results from the entire web. The user has to specify the websites from where results are to be scraped.

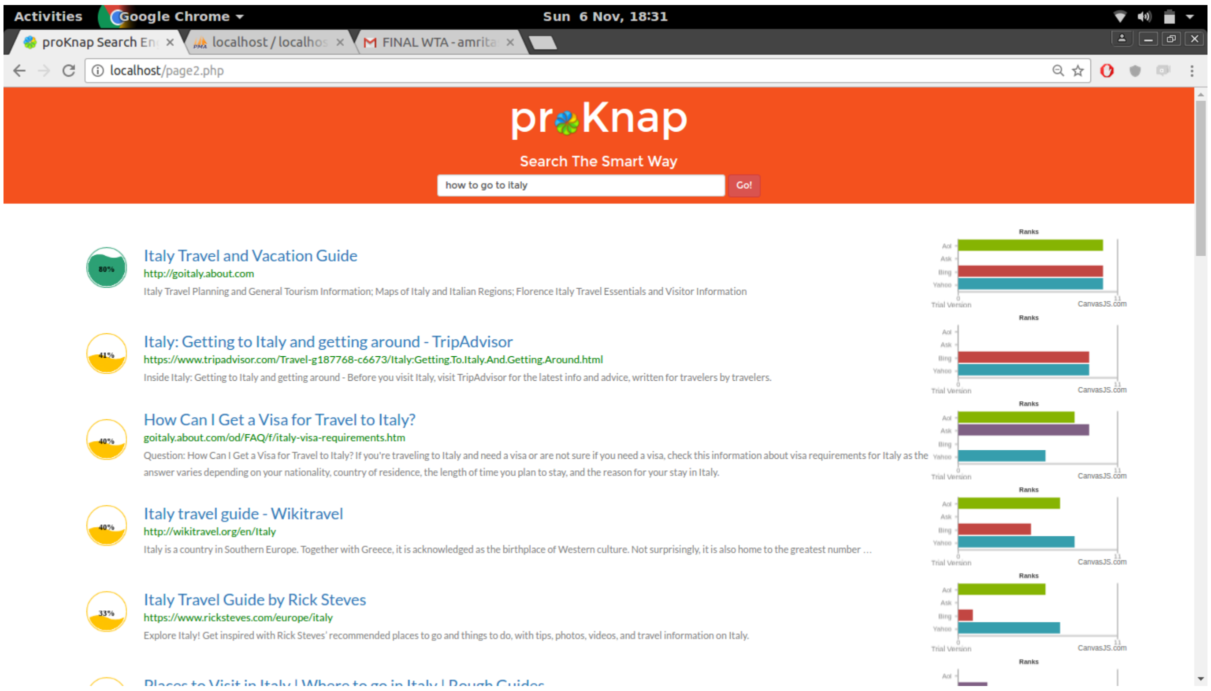
3. Video and image search results were to be removed.

4. All links obtained were to be made uniform.

5. Since it is time consuming to scrape 4 search engines, results were scraped in parallel.

4.2 Results and Discussion

The SVV algorithm yielded successful outputs on queries of the following seven types: "Who", "Which", "When", "Why", "What", "How", and named entities. The results are displayed in descending order of their weight. The first document was found to have the highest weight, and is followed by the document with the next highest weight and so on. The results obtained are shown in Figure 2 and Figure 3 which shows the first few results obtained for the entered query "How to go to Italy" and the remaining set of results obtained for the same query respectively.


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 …

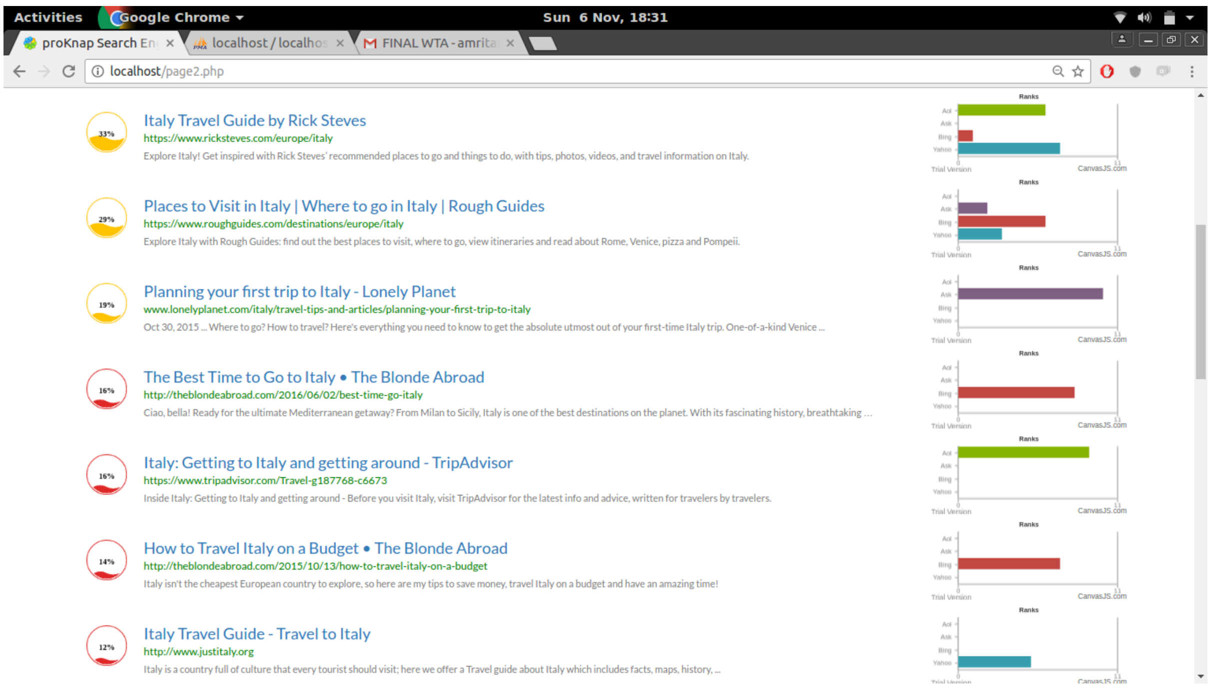
33% Italy Travel Guide by Rick Steves
https://www.ricksteves.com/europe/italy
Explore Italy! Get inspired with Rick Steves' recommended places to go and things to do, with tips, photos, videos, and travel information on Italy.

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

Figure 2: Results obtained for query "How to go to Italy"

The circular water bubble on the left hand side of each document shows the relevance of the document to the user entered query among the retrieved set of documents. The numeric label on the bubble indicates the vote distribution percentage, which has also seen depicted graphically as the level of water in the water bubble.

A document has also been assigned a relevance value to indicate how relevant it is among the retrieved set of documents. A green coloured bubble indicates that the document has a high relevance to the query, a yellow bubble indicates intermediate relevance and a red bubble indicated low relevance. The green and yellow bubbles can be observed in Figure 2 whereas the yellow and red bubbles can be viewed in Figure 3. It is seen that documents with high relevance are displayed first, followed by those with intermediate relevance and lastly the ones with low relevance.



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.

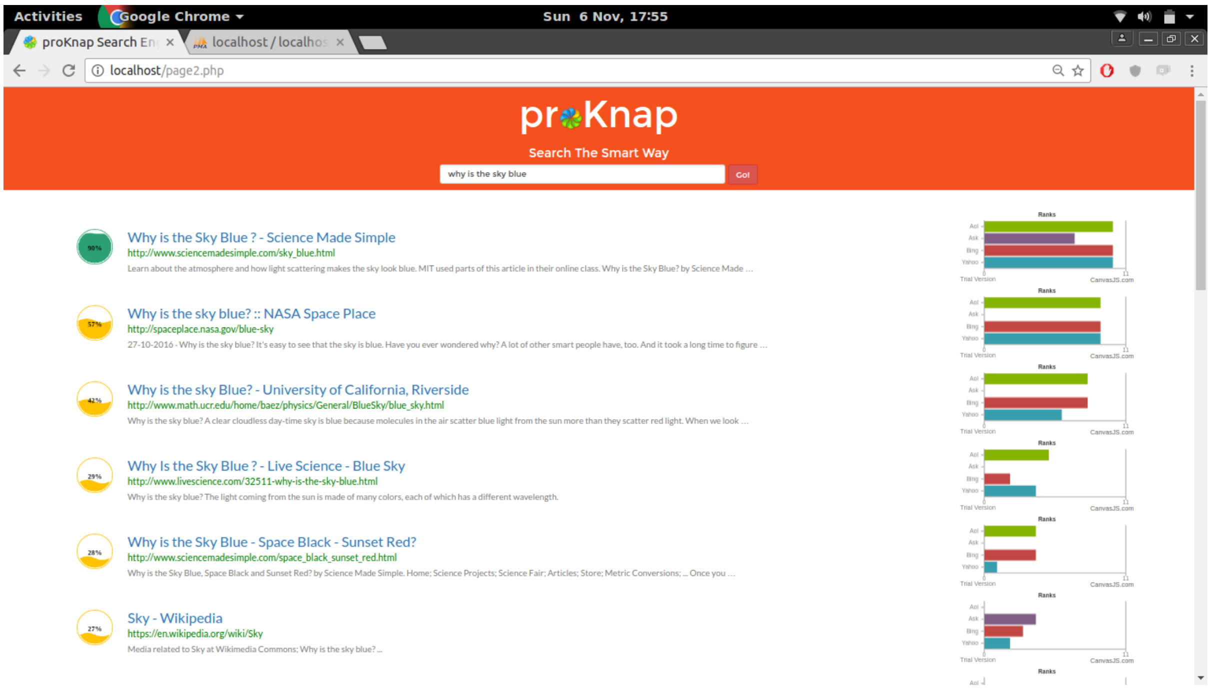
14% How to Travel Italy on a Budget • The Blonde Abroad
http://theblondeabroad.com/2015/10/13/how-to-travel-italy-on-a-budget
Italy isn't the cheapest European country to explore, so here are my tips to save money, travel Italy on a budget and have an amazing time!

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, ...

Figure 3: Results obtained for query “How to go to Italy”

The bar chart shown on the right hand side of each document displays the vote distribution of the corresponding document among the four search engine selected, namely, AOL, Ask, Bing and Yahoo!. A longer bar indicates in the bar chart indicates that the web document was among one of the top results for the search query entered in a particular search engine. The bar height could have values from 1 to 10 with 10 indicating that it was the first document in the search engines result page for the query entered by the user and 1 indicates that it was the last document in the first page of the search engines result page. These can be seen in Figures 2 and 3.

It has been observed that the documents for two different queries having around similar vote distribution percentages need not necessarily have the same relevance. Also a document with a high vote distribution can have an intermediate or even a low relevance and a document with a lower vote distribution can have high relevance. This is because vote distribution only takes into account whether the given document is present or not in the respective search engines result, whereas relevance also takes into account the rank of the document in the results. This can be seen in Figures 4 and 5.


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 …

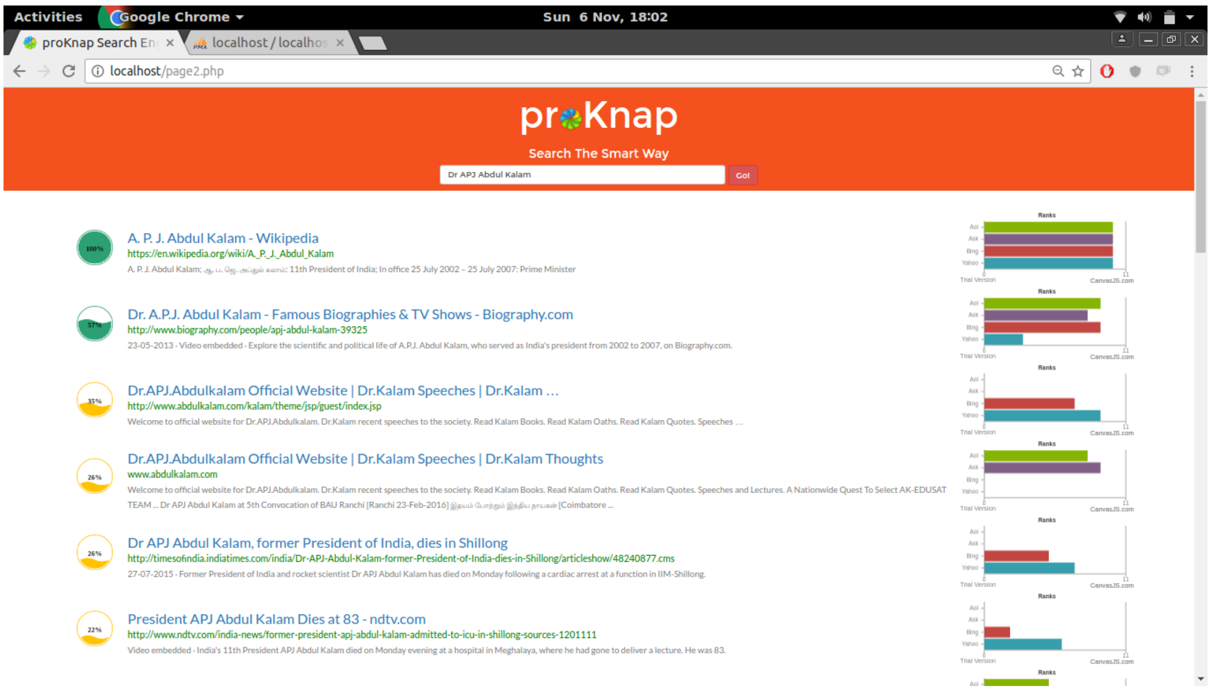
42% Why is the sky Blue? - University of California, Riverside
http://www.math.ucr.edu/home/baez/physics/General/BlueSky/blue_sky.html
Why is the sky blue? A clear cloudless day-time sky is blue because molecules in the air scatter blue light from the sun more than they scatter red light. When we look …

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? ...

Figure 4: Results obtained for query “Why is the sky blue”


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. ப. 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 ...

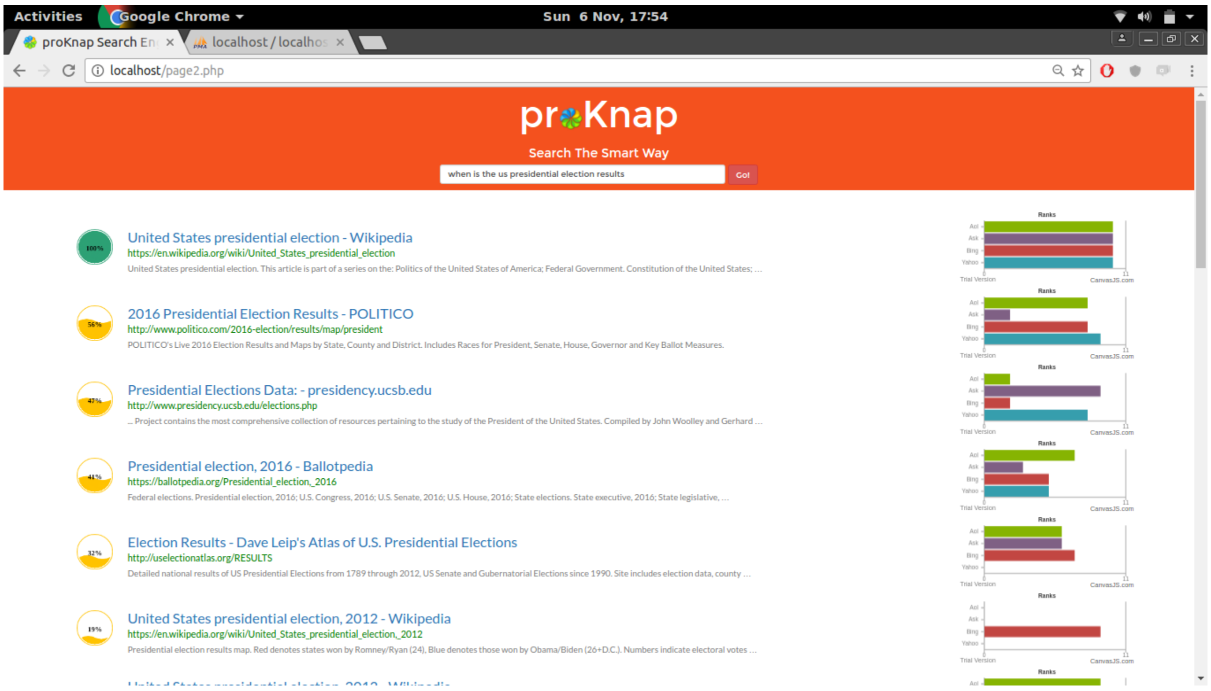
26% Dr APJ Abdul Kalam, former President of India, dies in Shillong
http://timesofindia.indiatimes.com/india/Dr-APJ-Abdul-Kalam-former-President-of-India-dies-in-Shillong/articleshow/48240877.cms
27-07-2015 · Former President of India and rocket scientist Dr APJ Abdul Kalam has died on Monday following 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.

Figure 5: Results obtained for query “Dr APJ Abdul Kalam”

As is clear, the second document in Figure 4 has a vote distribution percentage of 57 and has an intermediate relevance whereas the second document in Figure 5 also has a vote distribution percentage of 57 but has a high relevance.

It has also been observed that queries of the type "When" (shown in Figure 6), "What" (shown in Figure 7), "Who" (shown in Figure 8), and named entities (shown in Figure 5) had better results than the others in terms of vote distribution percentage as well as relevance.



when is the us presidential election results

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.

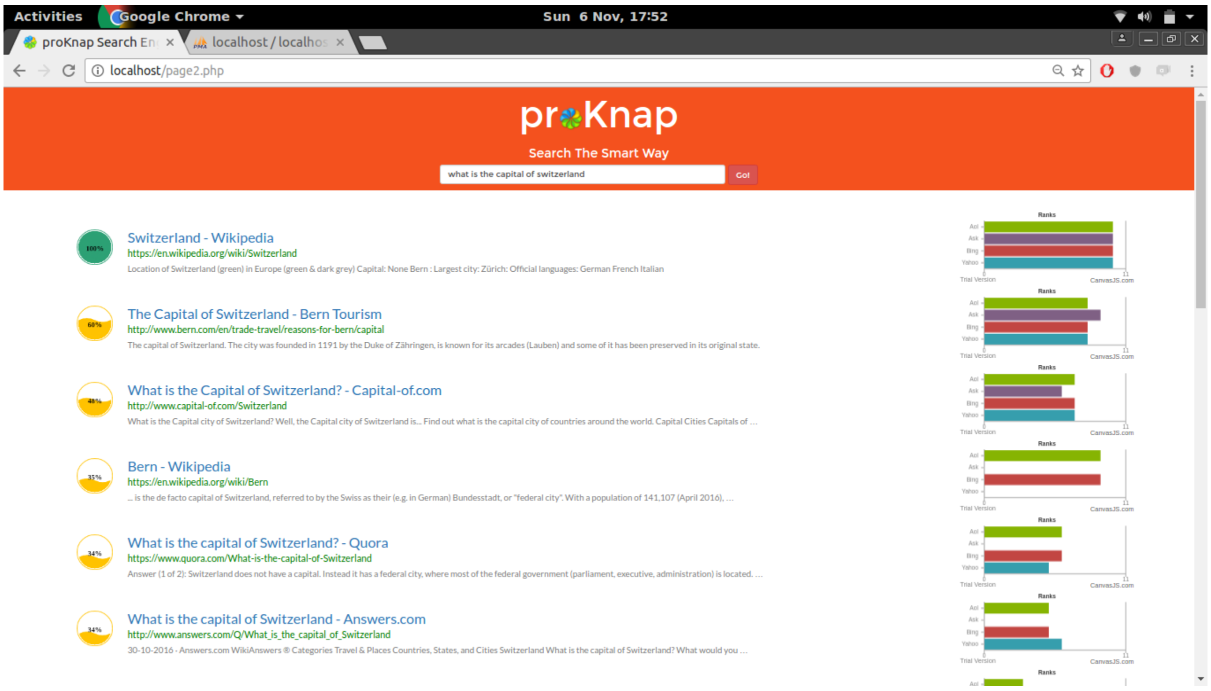
47% Presidential Elections Data: - presidency.ucsb.edu
http://www.presidency.ucsb.edu/elections.php
... Project contains the most comprehensive collection of resources pertaining to the study of the President of the United States. Compiled by John Woolley and Gerhard ...

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 ...

Figure 6: Results obtained for query "When is the US presidential election result"



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.

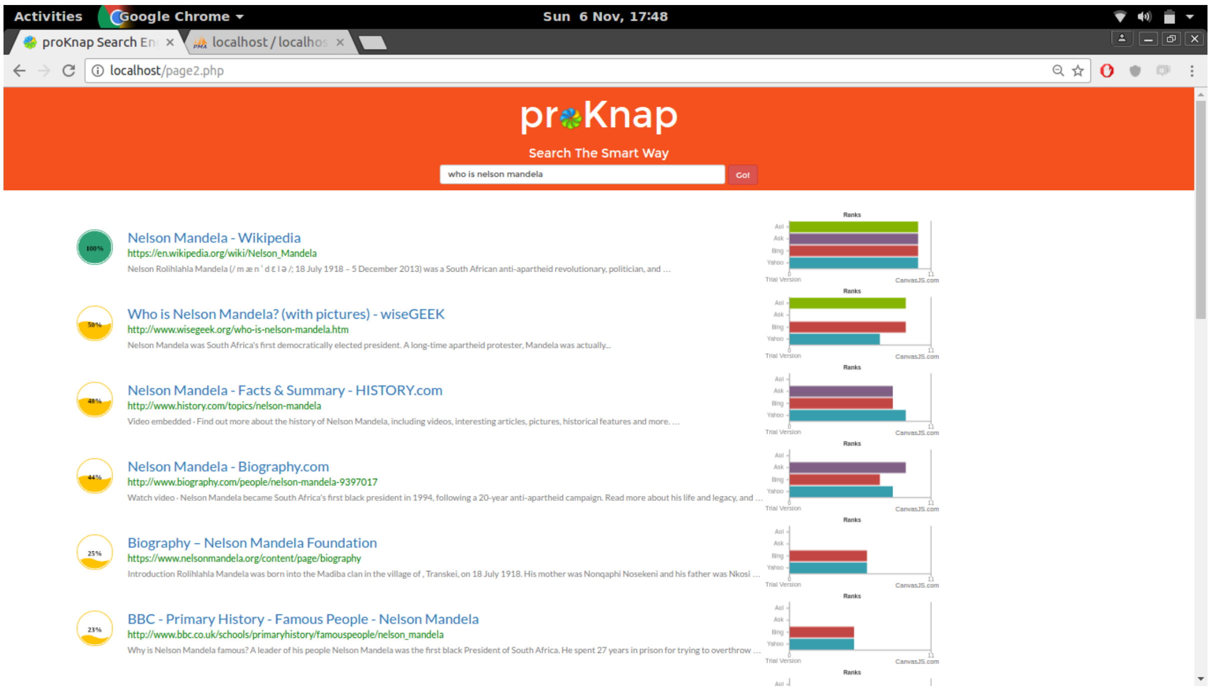
48% **What is the Capital of Switzerland? - Capital-of.com**
http://www.capital-of.com/Switzerland
What is the Capital city of Switzerland? Well, the Capital city of Switzerland is... Find out what is the capital city of countries around the world. Capital Cities Capitals of …

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 …

Figure 7: Results obtained for query "What is the capital of Switzerland"



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 …

Figure 8: Results obtained for query "Who is Nelson Mandela"

Further, the authors of the research paper, had conducted an experiment among 23 volunteers in order to judge and make a comparison of the results for certain hot key words. The popularity of certain search engines like Google and Yahoo had been incorporated by adjusting their α and β values. The testers were asked to perform queries on each search engine with popular query terms obtained from Lycos 50 (2004) including "Dragonball", "Pamela Anderson", "Britney Spears", "Las Vegas", "Harry Potter", "Kazaa", "WWE", "Jennifer Lopez", "Final Fantasy", "Yu-Gi-Oh!", "The Bible" and "NBA". The testers independently rated the search results of each search engine for a given query term on a score ranging from 1 to 5. It was seen that the scores of the SVV was significantly better than most of the other methods. The experiment confirmed that users are more satisfied with the proposed search methods than with general search engines, because of the depth of the information provided.

After the successful implementation of the SVV algorithm, the On the Fly Document Clustering algorithm was implemented to improve clustering performance of search engines. Presently only a set of pre decided queries work well for the mentioned algorithm as it currently functions only a closed domain. It is seen that on making use of ontologies for the pre decided set of queries, in a way simulates finding the reverse links of an important web document. It is observed that on using the OTFDC algorithm, the search results improve greatly and the results include web documents related to not only the search query entered but also to the related concepts of the search query. This can be useful when the user has minimal knowledge of the domain of the query he/she wants to enter.

The authors of the OTFDC paper had conducted three types of performance test using various clustering algorithms. First, 12 students were selected from National Dong Hwa University to judge the quality of clustering results for 39 top queries on the Internet during 2007. Second, a computer simulation was used to determine a suitable Threshold (T) value to use in OTFDC. Third, OTFDC applied to Google, Yahoo, and Vivisimo in order was simulated to verify whether OTFDC can improve the clustering performance of any search engines or not. In all the three tests it was seen that on making use of the proposed algorithm yielded significantly better results when compared to a non-semantic search.

4.3 Individual Contribution of Project Members

Although there is no clear division in the work done, as it mostly a team effort with regular meet ups and group discussions, the following is a vague distribution of the work that was supervised by the respective team members.

Register numbers 14IT132 and 14IT252 were involved with the scraping of results from the four search engines used and also comparing them in order to obtain an integrated and improved ranking of the web documents.

Register numbers 14IT224 and 14IT231 were involved with handling the entry and retrieval of the scraped search results to and from database. The integration of the python code with PHP and the user interface were also handled.

5 CONCLUSION AND FUTURE WORK

5.1 Complete Work Plan of the Project

Major tasks involved in the project are:

1. Analysing and understanding research paper.
2. Obtain results from popular search engines.
3. Combine results obtained and store them to database.
4. Make a user friendly GUI.

All the team members went through both the research papers and a week was spent on the same. To obtain search results from the search engines, Google Web Search API was tried. However, it was found to be deprecated. Search APIs for most search results was also not found to be useful as the ranking algorithm used was different from that of their actual search results. Figuring this out took approximately 3-4 days. So, BeautifulSoup was found to be a suitable alternative to Search API. However, scraping Google results was not permitted and we chose the next 4 most popular search engines, namely – Bing, Yahoo, Ask and AOL. Implementing scraping algorithms took approximately 2 weeks. The results then had to be combined and stored. MySQLdb was used to integrate Python and with SQL. This job was completed within 15 days. All results were displayed using a GUI which was implemented through various technologies like HTML, CSS, JavaScript, PHP which took roughly 1 week to build.

The SVV algorithm was implemented completely and successfully. However, some challenges were faced while implementation. An attempt to collect the search results from Google, the most widely used search engine was made. However, Google has stricter policies and hence, its search results could not be obtained. Google has deprecated it Search API and provided a custom search API. However, the entire web could not be crawled and the ranking algorithm was not same as that of Google's actual search engine. As a matter of fact, all Search API's do not use the same ranking algorithm as their actual search results. Also, as search results included video, image and news results, they had to be removed while storing into database. The links collected while scraping had to be made uniform as some links have http, some https and so on since we store only distinct links in the database. Querying all 4 search engines took around 20s. So sending queries to the search engines was made parallel using Python’s Multithreading library and it was observed that the time taken after parallelization reduced to 10s.

OTFDC Implementation

During implementation of OTFDC, a few hurdles were faced during the implementation of the OTFDC algorithm. Mainly obtaining reverse links to an important web document posed a challenge. So, related searches were taken from one of the search engines instead of reverse links. The process of running SVV algorithm for each of the related queries is a time consuming process and it could not be parallelized as it involves write operations to the database.

Currently, the algorithm has been implemented by taking a few queries (closed domain) and creating relevant ontologies for them so as to obtain candidate clusters for each cluster. It is proposed that the domain can be improved by making use of relevant ontologies as they become available.

REFERENCES

[1] Chen, L. C., & Luh, C. J. (2005). Web page prediction from metasearch results. Internet Research: Electronic Networking Applications and Policy, 15(4), 421–446.

[2] Chen, L. C. (2011). Using a new relational concept to improve the clustering performance of search engines. Information Processing and Management 47 (2011), 287-299