ISSN 2319-2720
Volume 2, No.4, October - December 2013
International Journal of Computing, Communications and Networking
Available Online at http://warse.org/pdfs/2013/ijccn01242013.pdf

Efficient Search Based On Relativity Context Using Web Mining Techniques

Varun Mishra
Dept. of CSE, Amity University, Gwalior (MP),India, varund5@yahoo.co.in

ABSTRACT

In this paper, I present an efficient method for keyword based searching using the relativity context of keyword on the basis of information available on social networking sites. To do that I designed an standard application and by using social networking sites by which I can get information about a user's activities and interests. I use these information and put it in our own database to categorize in an efficient form ,so that the searching time can be reduced comparatively. I also can find the location for the search term by using mobile technology.

There has been much work already done in trying to personalize search results based on user profiling. The top search engines attempt to build a user profile based on usage history and then use that to customize search. However, there have been fewer attempts to personalize search based on a deep understanding of a user's activities and preferences. In this work, I attempt to do this.

1.INTRODUCTION

Since I know that very well lot of research is going on in the field of Data mining and Searching is essential for mining a data and resulting data should be in well presentable form. Search has been one of the most widely studied problems of computer science.
Towards this direction, I have used social networking sites like Facebook to get detailed information about the activities and preferences of users. This information has been combined with the ODP directory project to create a detailed user profile. This user profile is used to disambiguate between different contexts for a search query and then execute the search query biasing it towards the contexts best suited for the user. The results are then rendered for both online and mobile consumption [3, 5]
Taking a separate initial direction, I have built a location based search engine. Here a user SMSes the search term, his/her location is retrieved using triangulation via the telecom operator. A database of Location tagged points is built and it is searched with the keyword and the user's location to return him the nearest relevant points of information. Finally, I have combined both the approaches to build a location sensitive personalized search engine which can be accessed over SMS.

2. RELATED WORK

2.1 Existing Search Engines

The Internet started with a directory listing of all the web pages. But as the size of the network and the content hosted on it grew, information retrieval became a challenge. Archie was the first search engine for finding and retrieving computer files. Others being Gopher and Wais. All had the following common characteristics.
• They had a spider which traversed the network and retrieved documents from different servers.
• Built up databases of directories or web pages.
• Ran ranking algorithms

There were two main directions in which industry preceded - directories and search engines. Directories are lists of sites input by human beings. Some of the most famous ones were WWW virtual library and Yahoo. Slowly directories lost out because of the exponential growth rate in web content. This rendered it impossible to browse web content using directories.
Social networks provide rich data on a person, his activities and interests. Sites like Facebook have been integrated now into the overall web experience with Facebook plugin being present on several sites. Often users end up 'liking' or commenting about a link or web page on Facebook, Twitter or other similar social media. But often this information is shared only with the 'friends' of the person. Hence, though a person's Facebook account provides a great deal of information about the user, this information is not available publically. Social networks like Facebook have included functionality to search the web but their utilization of the user profile information in personalizing search is very limited [2, 16].

There has been some research on using the ODP database itself in using personalization though that it is only limited to information sources like bookmarks.

2.2. Existing Location Based Services

Location based services [23] delivered through mobile devices have been a subject of interest in both the academic and application development community for a decade. The central idea is to know the location of the user via GPS/ cell triangulation or explicitly being told by the user and then

provide services to the user based on his/her location. The following application areas have been explored amongst others like Navigation, Information, Tracking People/Vehicle tracking. Games, Advertising. Technology giants like Google and Facebook have native mobile applications for maps, navigation, status updates about places, photo tagging which rely heavily on location data to be pulled from the cellphone. A number of startups have emerged and attracted a lot of investor attention in this domain.
However, most if not all of the existing LBS applications rely on rich computing devices like smartphones to be available with the user. In a developing country like India this is not the case and I propose an alternate method of providing LBS information to people with basic mobile phones. On this front, recently, both the application and the business case for LBS has been understood by telecom operators in India who are now willing to provide the location of the user through the cell tower infrastructure itself.

2.3. SMS Based Search Engine

Considering the dominance of basic feature phones in developing countries like India, there have been some attempt to provide search functionality using short messaging service technology. Amongst bigger technology companies Google and Yahoo both started SMS based search. However, these services have met with mixed response. Google has closed down it's service since. Some other local search engines have also forayed into SMS search including Asklaila, Guruji etc. However, none of them have been successes possibly partly due to the lack of personalization on user profiling and location.

3. PERSONALIZED SEARCH

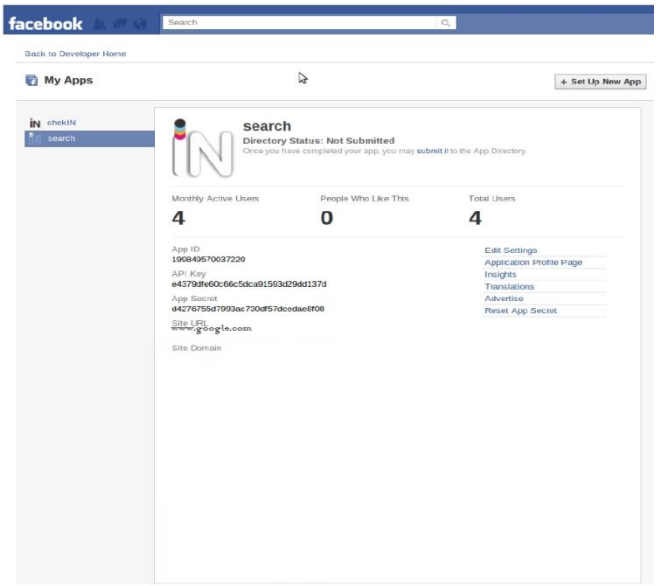
Personalizing search requires one to understand the preferences and activities of the user. A search term typically has several contexts associated with it. For eg. take the search term ‘AAMIR’. Now, if one runs a Google search on it the top results one obtains are:

- AAMIR cricketer.
- AAMIR actor.
- AAMIR movie name.

Now, different people could have different intentions with the same keyword. One of the ways to disambiguate between the different queries is on the basis of broad interest. One person might have an interest in educational content; another person might have an interest in entertainment etc. Several such niche search engines exist which allow querying a specific domain. However, using them requires the additional effort of remembering each of them. Further, most people have interests spawning multiple categories but have specific preferences in those categories. For e.g. most people would conduct searches for movies, sports, and educational content but in those specific categories, people would only have watched or want to search about certain movies, or certain sports [2].

3.1. Sources of information on user

There are several online sources to get information about the activties and preferences of a user. Users maintain profiles on professional sites such as Linkedin, some maintain their personal home pages. However, Iwere amazed at the amount of information people are willing to share on social networking sites like Orkut, Facebook and Twitter[2,9].
However, most of the information people share on sites like Facebook is available only to their friends and explicit permission is required by an outside party to get this information. For this purpose a Facebook 'app' (a software running on Facebook) was created to authenticate users and obtain permission to access their profile information.



3.2. Using the ODP database for characterizing keywords

Now, the keywords themselves are of limited utility. Take for instance the keyword 'Nirvana' above. I desire that whenever the user searches for anything which has a direct link to Nirvana, given the fact that Nirvana is amongst his liked bands, our system figures out that as a possible context. For example, if the user searches for 'Kurt' the lead guitarist of the band Nirvana, then the context should be understood to be the Nirvana band amongst others.
To accomplish this, I need to gather more information about every keyword. Here the ODP database comes in handy. The ODP database is a large directory of weblinks organized in a hierarchical manner. The ODP directory allows us to find weblinks which can help us obtain detailed information about a keyword. The actual implementation involved dumping the ODP directory to a MYSQL database. This has to be done in a category specific manner since dumping the complete ODP directory created such a large database that searching weblinks for every keyword in it took several minutes.

4. LOCATION BASED SEARCH

Given a set of points positioned in 2-D space representing places, events, people, etc, the goal of the problem is to find out the nearest points around the user matching the user query. Some of the points are static, for e.g. places such as ATMs, petrol pumps, shops etc. Others are constantly changing, for e.g. the offers in different shops, events like concerts, people etc. Example query - 'ATM' returns to the user the nearest two or three ATMs to his location. 'Petrol Pump' returns to the user the nearest Petrol Pumps to his location. 'Shoes' returns to the user the nearest shoe stores to his location.

4.1. Getting user location

Retrieving the user location can be done in a variety of ways with each having a different accuracy associated with it. Existing web search engines often get the user location explicitly specified by the user to the granularity of a city. Web based systems also do this by IP tracking where the IP address of the user is used to geocode his location.

In mobile devices, the methods are much more diverse. A coarse accuracy can be obtained via Cell Tower mapping. The cellphone typically knows the cell tower ID of the tower it is connected to, the NAC (network area code) etc. This can be matched against a database of cell towers to find the location of the cell tower the user's mobile phone is connected to. The location of the cell tower itself can serve as a good proxy for the location of the user. The accuracy can range from a few hundred meters in dense urban areas to several kilometers in rural areas.

The second approach to obtain the location of a mobile user is by using cell phone triangulation. This is similar to the first approach of using cell tower location but here instead of the location of one cell tower, the locations of several cell towers are used.
The highest accuracy is typically obtained by using a GPS device (Global positioning system). These devices are increasingly being integrated into smartphones and are the basis for a number of applications using location data. The GPS accuracy can reach upto a few meters. However, GPS devices require a clear sky to work properly and are dysfunctional inside buildings for instance.

5. LOCATION SENSITIVE PERSONALIZED SEARCH ENGINE

I have seen the implementation of two different search engines. One of them performs user personalization based on the data available about the user on social networking sites like Facebook and makes available a personalized search engine which can be accessed via SMS and Web. The other builds a location aware SMS based search engine which allows a user to send a query for nearby locations or points of interest in SMS form and after locating the user via the telecom operator, answers the query using both an internal location database and external location databases. I review the location aware personalized search engine as a combination of two funnels.

Consider a general search query done on a location tagged database. This will return all points of interest matching the search query. Our first funnel is location which limits the results to the local domain of the user. So, after the first funnel I get location tagged results which match the user query which are limited to the spatial locality of the user. The second funnel is user personalization; this filters the results to return us those results which have the most matching content with the user profile.

The building of the user profile is as detailed earlier –
- The user agrees to provide information to the application.
- The information is pulled by using the FB graph API.
- The user profile is built by using the ODP directory and the Nutch-Lucene crawler/indexer.

The search workflow is as follows -
- The user SMSes the search query to the gateway.
- The gateway redirects the query to our HTTP server.
- process the message of obtain the search query and the number of the user.
- The number is used to obtain the user's location from the cell phone provider and to obtain his profile.
- The search term is matched with the user profile to determine possible contexts.
- A search query is run on our internal and external location databases for the following -
  - The actual search term
  - The category in which the context falls
  - Context obtained
- The results are rendered back to the user via SMS

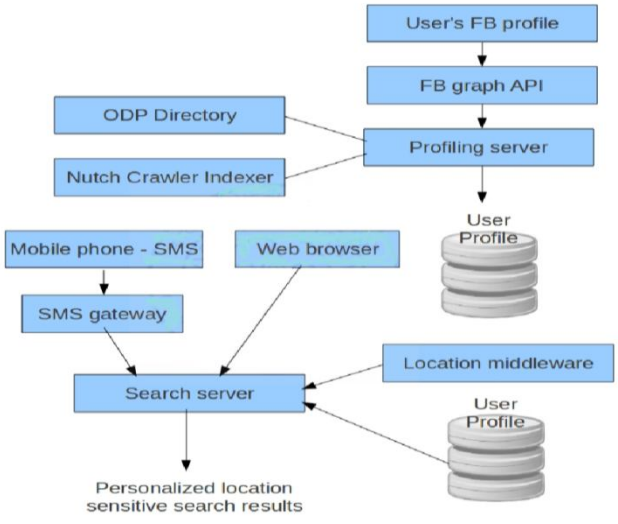


Figure A: System architecture

6. RESULT AND ANALYSIS

The system was tested in two broad aspects -
• Functionality testing. Here I demonstrate using an authors profile whether the intended functionality is being achieved or not.
• Performance testing. Here I choose 9 subjects and ask them to rate the performance of the system.

For the first part, the system was analyzed in detail using the author's profile. The profile obtained from Facebook.
On referencing the 150 keywords with the ODP database a list of links describing the activities of the author was obtained. This contained about 34,000 links. Written below is a compilation of some of the search queries and the corresponding contexts identified by the personalized search engine.

Query	Context found	Type of relationship demonstrated
Winner	Tennis	Winner is a type of stroke in tennis
Watson	Harry Potter	The actress Emma Watson played a role in the movie Harry Potter.
Kirk	Metallica	Kirk Hammett is the lead guitarist of Metallica.
Blackjack	21	21 was a movie based on the card game blackjack.
Dictator	Fidel Castro, V for Vendetta	Fidel Castro was considered a Cuban dictator and V for Vendetta is a political movie
Pumba	Lion King	Pumba is a character in the movie Lion King
MIT	A beautiful mind 21, Goodwill hunting	MIT was pictured in the movies 21, A beautiful mind. MIT was mentioned in the authors personal site hence that was also identified as a context

The actual weblinks have not been reported since they are being fetched from an external search engine. The above searches considered only the personalization part and not the location context. The following are simulations of queries run from the mobile phone when combined with the user profiling and location context. It should be noted that sufficient location tagged content was assumed/ added to the internal database. The amount of content in the internal location database or the coverage of external location sources is not within the purview of this work and hence this was assumed to be good.

Query	Context found	Relationship	Actual result returned
James Hetfield	Metallica Music	Performer band and category	**Metallica concert, Music Court,** Bhopal (belonging to Music category)
Pasta	Italic Food	Cuisine and broad food category	**Food Plaza**, New Market,bhopal
Ritesh Deshmukh	Indian Film actor		**PVR Theatre**, Lal chok,bhopal **Sangeeta** Theatre,Gandhi Road,Bhopal
Jeans	Levis jeans Clothing		**Levis jeans store,** Birhana Road,Vidisha (Preference of the user for Levis jeans)
Enfield	Royal Enfield Automobile		**Royal Enfield showroom, VIP Road** (Royal Enfield bike) VIP Road , Bike and car rental, (Broader automobile category.)

Here I can see the benefits of using a Hierarchical system like the ODP through which I am able to identify multiple layers of context and use each of those to return results applicable to the user.

The second part involved performance testing using test subjects. The system was tested by 9 other evaluators. They were each asked to submit their Facebook profile which was used to make a profile for the user. In the interest of time, only 1000 randomly selected links were crawled to build the user profile from the set of links returned by the ODP system. For heavy users, the total number of links was seen to be as high as 80,000. They were then asked to run 20 queries related to the content they have uploaded in their facebook profile and answer whether the context which the system returned was their intended context. They were also asked to run 10 queries which should not return any context from the system according to their search intention and information on the facebook profile page and report their result.

7. CONCLUSION

The location retrieval is currently done in collaboration with a telecom partner, but this could be done based on a separate method. A front end smartphone application can be developed which gets the location via GPS or GPS and sends the location along with the text query. The location information sources currently are our internal database and some external sources. Multiple other sources can be added to the system with ease without any significant changes to the overall implementation. A simple approach has been used for rendering the information for the user to view, one by sending him an SMS about the basic details of every search result. This can also be modified without any changes to other parts of the application. A different delivery mechanism could be used. One can have a category specific focus on different details. Hence, overall the modular design of the system leaves sufficient room for the customization of its different components and including more parallel components to do the same task.

REFERENCES

1. U. Manber, A. Patel, and J. Robison. **Experience with personalization of Yahoo!** *in Communications of the ACM*, 43(8):35-39, 2000. ISSN 0001-0782.
2. Robin Singh Bhadoria, Premnarayan Arya, **Blogger: Enable Searching in Blog Using Cluster Algorithm** in *World Applied Programming: International Journal*, Volume 1, No.4, 2011, pp. 237-245.
3. W. Chu and S.T. Park. **Personalized recommendation on dynamic content using predictive bilinear models**. *in Proceedings of the 18th international conference on World wide web*, pages 691-700. ACM, 2009.
4. A. Sieg, B. Mobasher, and R. Burke. **Web search personalization with ontological user profiles**. *in Proc. of the sixteenth ACM conference on Conference on information and knowledge management*, pages 525-534. ACM, 2007.
5. F. Liu, C. Yu, and W. Meng. **Personalized web search by mapping user queries to categories**. *in Proc. of the eleventh international conference on Information and knowledge management ACM, 2002*. pages 558-565.
6. F. Liu, C. Yu, and W. Meng. **Personalized web search by mapping user queries to categories.** *in Proc. of the eleventh international conference on Information and knowledge management, ACM, 2002.*, pages 558-565.
7. F. Yang and Z.M. Wang. **A mobile location-based information recommendation system based on GPS and WEB2. 0 services**. *WSEAS Transactions on Computers*, 8(4):725-734, 2009. ISSN 1109-2750.
8. O. Masutani and H. Iwasaki. **BEIRA: An Area-based User Interface for Map Services**. *World Wide Web*, 12(1):51-68, 2009. ISSN 1386-145X.
9. Robin Singh Bhadoria, Ram Kumar, Manish Dixit **Analysis on Probabilistic and Binary Datasets through Frequent Itemset Mining** in Proc. of *IEEE World Congress on Information and Communication Technologies (WICT 2011),* Mumbai, 11-4 Dec, 2011, pp. 263-267.
10. H. Xu, H.H. Teo, B.C.Y. Tan, and R. Agarwal. **The role of push-pull technology in privacy calculus: the case of location-based services.** *in Journal of Management Information Systems*, 26(3):135-174, 2009. ISSN 0742-1222.
11. M. Gruteser and D. Grunwald. **Anonymous usage of location-based services through spatial and temporal cloaking**. *in Proc. of the 1st international conference on Mobile systems, applications and services*, ACM, 2003, pages 31-42.
12. L. Barkhuus and A. Dey. **Location-based services for mobile telephony: a study of users' privacy concerns**. *in Proc. Interact, volume 2003,. Citeseer* pages 709-712.
13. M. Dalal. **Personalized social & real-time collaborative search.** *in Proc. of the 16th international conference on World Wide Web, ACM, 2007*, pages 1285-1286.
14. P. Bonhard and MA Sasse. **Knowing me, knowing youUsing profiles and social networking to improve recommender systems.** *BT Technology Journal*, 24(3):84- 98, 2006. ISSN 1358-3948.
15. E. Diemert and G. Vandelle. **Unsupervised query categorization using automatically-built concept graphs**. *in Proc. of the 18th international conference on World wide web*, pages 461-470. ACM, 2009.
16. Robin Singh Bhadoria, Deepak Sain, Rahul Moriwal **Data Mining Algorithms for Personalizing user's profiles on Web**, in *International Journal of Computer Technology & Electronics Engineering (IJCTEE),*Vol.1 , Issue 2, 2011, pp 120-125