International Journal of Environmental Engineering– IJEE
Volume 1 : Issue 4 *[ISSN 2374-1724]*
Publication Date : 27 December,2014

Tracking information on future interests/events

Using web mining, predictive analysis and semantic web

[Ankit Sharma]

***Abstract*— This paper explains the implementation of a mobile application solution on iOS to track future events and interests relevant to brands, weather, entertainment, leisure, persons followed, disaster, etc. and throw fragmented information threads to users using predictive analysis, push notifications, concepts of web mining and semantic web. Public APIs and XML, Json feeds have been used to gather data and push technology transfers processed information threads matched with objects of interests based on the user's predictive analysis. The utility of such a system is to rush the tracked information to the individuals in real time for better activity scheduling and informed decision making. The user's social graph, online activity, mobile data, time of day, location, historical search patterns are selectively accessed for insights into user's behavioral patterns to estimate user's future interests hierarchy. The solution can also be used to generate useful analytics for enterprises.**

Keywords*— *push notification, predictive analysis, mobile iOS app development, semantic web, web mining

I. Introduction

The data on the web is dramatically manifesting and it is increasingly becoming hard to obtain the right information at the right time. Many events and updates that need our attention get missed by, due to the inadequate technologies involved in web mining. Over the last decade, mining the web for easy retrieval of information from thousands of unstructured sources have been a topic of curiosity. Many applications, search engines, RSS(Rich Site Summary) and atom feeder, information aggregation and curation agents have come into existence to feed the user with consumable information.

Many web as well as mobile applications have been deployed that work on solving the problem of connecting users with information such as locating the nearby restaurants, available offers, coupons and shops, outlets from where he/she can buy merchandise.

However, examples of the following set of problems and circumstances involved have not found any comprehensive, quick and easy online solution such as:

1.) when is the weather going to get rough? So that one can schedule their outdoor activity optimally

2.) When is the celebrity or a person you followed died? So that you can offer your condolences on his Facebook page.

3.) When is there a public holiday in the place where your friend is living in? So that you can make your travel plans to visit him.

4.) When the cricket team you are following has its next match?So that you can buy its tickets.

5.) When the football team you follow scored its first goal in a tournament? So that you get real time updates

6.) When the metro transit construction in Mangalore going to get completed? So that you remain informative of one of your interests.

7.) When is the film whose trailer you watched, on YouTube getting released? So that you do not miss its premiere.

8.) When is the release of the next game in a specific game series? So that you buy your copy from an online store.

9.) When the business acquaintance you know is in town? Or When the SENSEX goes down 20000?

A user has to consistently follow up a large set of websites to consume regular updates and information threads. In this paper, information threads are connected pieces of information such as “a. Earthquake in Singapore. b. The earthquake was 7.0 on the Richter Scale. c. Last time, the earthquake came two years ago” is an informative thread.” The most successful route to reach such information is websites on a web browser directly.To reach new updates or most refreshed viral content, the ideal route is a search engine

Again, there are alternate methods involved, such as maintaining an RSS feed or subscribing to the Facebook pages or Twitter accounts of objects of interests such as celebrity, brand, event etc.

The topics and subtopics on which user wishes to be notified are objects of Interests(OI). One facet of this problem is untimely delivery of information from the web.

Suppose, the user consumed Information I at time t_1 and then Information II was generated on the web at time t_2. Now the user happens to consume the Information II at time t_3. The information I and information II were connected and this tells us that there is a time lag in delivery of information II (which, when delivered makes a connected information thread) to the recipient which is t_3-t_2. The user simply missed significant release of information. For ex- a user got to know about a movie and loved the trailer. After one month, the movie got

Ankit Sharma *(Author)*
Maharaja Agrasen Institute Of Technology /GGSIPU University
India

released and after another one month and a random web search revealed to the user that the movie had already been released and he had missed its premiere. This is a time lag of one month and is happening because of traditional internet search systems.

A. Existing Solutions

It is worth mentioning that services such as Facebook, Twitter and other social media offer notification services to disseminate updates and changes in the subject matter or simply OI.

E.g a Facebook page and Twitter account of 'apple iPhone' will broadcast updates such as the launch of new iPhone.[2] However, in order to better up dying user experiences and increasing competition in social media, these web applications have got changed over a period of time. Also, on the top of it, they do not offer personalized notifications. The news feeds on Facebook wall of a user gets clogged with several posts of objects of interests, brands, celebrities that the user is following and most of this information is either too much or too less.

The organic reach and organic reach percentage of a brand to its follower on Facebook had declined from 12% to 6% recently.[3] The organic reach is defined as the number of people who have seen the post in their news feed.[4]

Another existing solution is “Google Alerts” which are email updates of latest relevant Google search results based on user's query. However, this fails because it tracks search engine results which is not refreshed content (imperfect algorithm) i.e old links get notified and reflect poor information quality plus it is not real time.

Google Now offers an acutely similar solution as been proposed in this paper, but it bases its prediction of user's identity only on historical search queries which is just not enough and the aggregated content is simply the Google search index.[14]

The average internet user is looking for highly specific and selective information with which he can shape his future decisions.

Also, the information consumption of an average internet user through mobile applications is increasing at a very fast pace[1][4]

B. *Suggested Solution*

The user's subscriptions to Facebook pages and Twitter feeds of the OI are captured using predictive analysis of user's Facebook and Twitter account.

Other means of predictive analysis are the assessment of the historical search queries in the web browser and the Google search app for which APIs(Application Programming Interfaces) are available. Location, time of day and mobile data are also tracked to give better justification to the predictions of user's identity.[13]

Once the user's identity is predicted as a probablitity distribution and hierarchy of objects of interests are identified, he user starts receiving notifications in the form of short, summarized and highly specific processed information. The user interest hierarchy has to be contextually tagged with contexts in which these predicted interests lie. The pre-conceptualization of a certain object of interest by the user is also taken in note. This information has been gathered from the web by accessing databases of many sources using APIs and generation of feeds in different file interchange formats. This gathered information is then processed and garbage,noise in the gathering is also cleaned to obtain information threads.Garbage and noise are the useless portion of the information.

The way the information is put in transit is push technology and careful correlation of the notifications is needed to avoid an abundance of unnecessary notifications.

II. Methodology

A major question in implementing a mobile solution is on what operating system such a solution be developed.

The mobile OS available are Android, iOS,Windows, Blackberry etc.

Because of its open source nature, Android can be used so that ready community support can be taken during the research. The penetration of Android is more than any other OS. However, in this paper, the solution is developed on iOS.

The major challenges to the solution are to :

1. Locate the client/device/mobile app user
2. Identify the OIs by tracking time of day, historical user behavior and accessing mobile storage such as music library other than social media analysis
3. Gather data from the web and process(classifying, categorizing semantically and summarizing) the data
4. deliver location wise and query based information threads on the device or to the client

The first challenge is addressed by the use of the Apple’s core location framework while developing the application. This platform helps in gaining access to geographical and mapping data using wifi, cell phone tower triangulation techniques. It is off course high power consuming operation.[8].Location awareness can be added to the implementation by automated location tracking, geofencing and activity recognition. Determining the location and travel routes of the user will explain the places which the user visits more often.

A user registers himself with the application by feeding Facebook, Twitter and other social media accounts to the application's configurational settings.

A user can also add certain keywords which are the keywords he is interested in. The user is not allowed to add ambiguous keywords, but only which are highly specific. For eg- keyword 'cricket' is broad and not allowed to be added by user while 'IPL' and 'EPL' are readily added by human users. This shall mean following example:

User will get updates and notification on matches - live match scores for EPL and IPL matches respectively.

Predictive Analysis takes place at the user end when the app is integrated with Facebook, Twitter, Whatsapp. The user's Twitter streams, Facebook page likes and Whatsapp statuses give latent insights into the perceptible OI.

Mobile storage is also accessed to analyze the storage patterns of the user. e.g- the music library on a hand held device, which consists of songs and artists and mentions of artist albums, give an accurate insight to what kind of music,musician the user is interested in. The predictive modelling[18] over this data can estimate the hierarchy of objects of interests such as the artist, genre of music,albums etc.

The user has a social graph made up of his social network and likes on Facebook brand pages . The Twitter account of the user contains his followings and the subjects to which he is following. The Twitter stream also contains user's tweets.The search patterns and consumption of search results by the user is also taken in account to estimate the future interests. All these sources list a lot of data over which predictive, descriptive and decision modelling can result in estimation of certain objects of interests like the genre of movies and novels the user is interested in.

A web crawler crawls websites which are pertaining to objects of interests around the world, music industry websites, imdb i.e internet movie database among others.

The crawling process has to be automated, distributed, highly scalable, extensible and performance efficient[6]. A web crawler gathers dynamic web pages to support indexing process,in our case, priming activated indexing[7].

Priming Activated Indexing is the method that can extract keywords representing the author's main point in an information thread, regardless of 'Term Frequency'.The term frequency is the total no. of times a keyword is detected in a piece of information.

Once the feed is obtained, it is processed by semantically categorizing and summarizing large articles into smaller pieces of information. The basis of this activity is a set of algorithms.There are many kinds of algorithms which can mine the data. The textual categorization can be done using an algorithm which uses the concept of weighted mean.[15]

This means that a weight is associated with a certain keyword. The term frequency is used to calculate the total weight of a keyword in a large article. Based on these weights, contexts can also be defined and synchronous rules tie contextual tagging to the articles or pieces of content. Then, a summarization algorithm can summarize the article into least bit of information. e.g- in processing of news articles, once they are categorized semantically, the article can be reduced to only the headline.

In many cases of processing such as processing of weather or sports data, synchronous rules are followed. All kinds of parsed and processed information are then connected and information threads as small as 256 bits are generated.

When this solution is implemented, the user harnesses the potential of a push notification instead of pull notification. He does not need to access a website and its huge amount of data only to check whether new content has been furnished or not. This saves time. Push notification means publishing of information by the server to the client without any request unlike pull notification where the client requests for release of information.[5] A cloud infrastructure is used for hosting the mobile application and push technology is the direct information transfer from this cloud to the device without request.

In this paper, important future events worth tracking(OI)

1.) Sports events- sports wise, player wise, tournament wise-scores notification

2.) Holidays – notification while travelling with summary

3.) Bollywood movie/music events – notification with movie summary, release venue and tickets information

4.) Hollywood movie/music events - notification with movie summary, release venue and tickets information

5.) Persons followed and abrupt weather changes

7.) Natural disaster, manmade disaster

Once configured , the user receives notifications for objects of interest(OI). These notifications are highly customizable, i.e the user can decide to withdraw from following a particular OI so that the notifications are stopped. Factors such as post query navigation and general browsing behavior have been used to evaluate recommender system's performance

Public REST APIs are used which offer many services and the ease of development. Every information thread/notification is given an etag which is a unique way the notification is remembered by the system. The etag contains the time stamp, information about OI, source,links etc. This etag is the similarly structured like what the HTTP protocol gives to every document on the web from where the nformation thread has been derived.

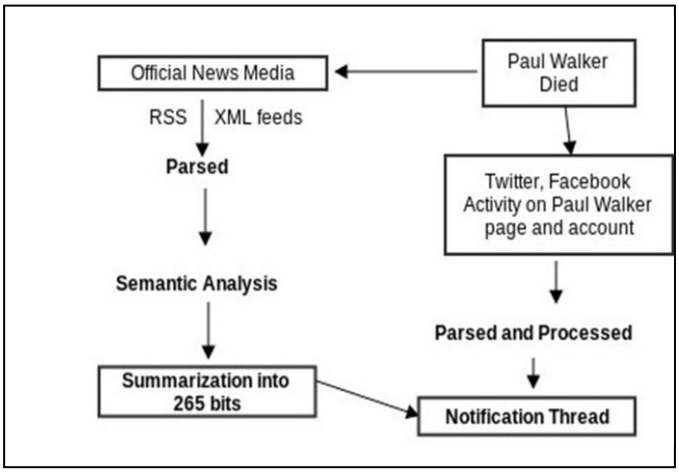


Fig i. example for generation of notification thread

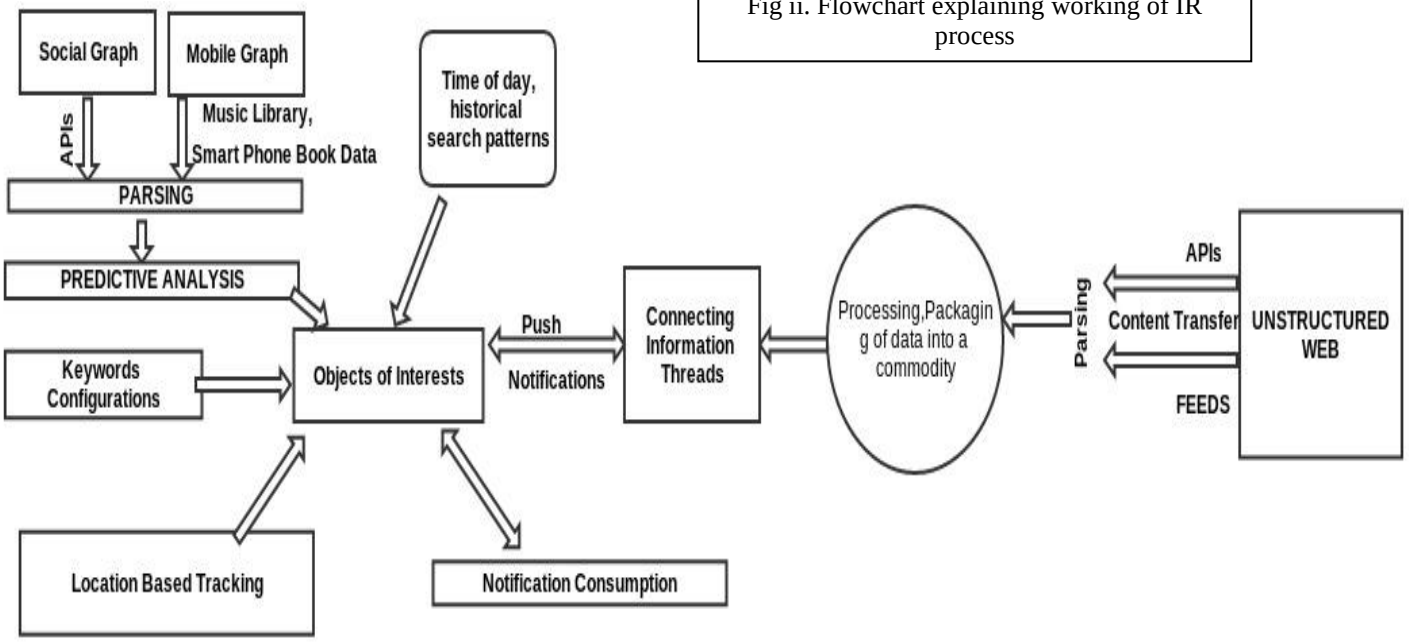


Fig ii. Flowchart explaining working of IR process

When the document updates on the web so does its etag. As a result, information threads coming from the dynamic web pages get uniquely identified by different etags and thus are structurally different.

III. Development

The development is divided in following modules:

1.) Web Crawler i.e to gather the data from the unstructured web by making HTTP requests to public APIs. The crawler will gather the data in XML, KML or JSON formats.

In some cases, instead of APIs, RSS i.e Rich Site Summary and atom feeds are available. Or wherever they are not available, tools such as Draper can create RSS feeds for the implementation.

KML is keyhole markup language while Json stand for javascript object notation emerging as the most relevant format for REST services. Json is a lightweight interchange format.

REST is representational state transfer, a software architectural style consisting of a co-ordinated set of constraints applied to data elements, connectors and components, within a distributed hypermedia system. [9]

API	Source
Weather	http://www.worldweatheronline.com/
Sports	http://www.espncricinfo.com/
Movie	http://www.rottentomatoes.com/
News feed	Several sources
Disaster	http://comcat.cr.usgs.gov

2.) The crawled XML feed is then parsed using the NSXML parser available with the iOS SDK.

The NSXML is a forward only reader, i.e event driven parser. This means that when the parser comes across the start of an element, an event is raised and the delegate of NSXML parser implements the events to capture the XML data.

3.) Analysis and processing of the XML feed data such as text based categorization and semantic analysis is performed using lexer.

Lexers are used to recognize words that make up the language element because structure of such words is simple. Regular expressions are extreme in content sensitive and syntactic parsers.

4.) The data is stored in a relational database management system, in this paper SQLite.

It is an embedded RDBMS. It is provided in the form of a library linked into the application. No database server is running in the background.

SQL i.e structured query language is used to access data stored in SQLite.

5.) So now, the database contains the information threads of objects of interests i.e processed data as commodity is stored in this database.

Determination of OI

6.) A predictive analysis consiting of predictive, descriptive and decision modelling from the data from social media APIs is performed. Facebook Platform's Graph API has been used in the implementation so that the application can connect to the social graph of the user and analyze the usage pattern and news feed consumption to iterate the push notifications efficiently.

Also in later functionality, the notification threads can be shared with the user's social graph.The API used is Graph API.

The ideal example of graph API's usage in the current implementation is to understand whether the user has liked a specific object of interest's Facebook page or not.

7.) Similarly, the Twitter's Streaming API has been used to get a real time stream of the user's Twitter feed and predictive modelling has been applied on the user's Twitter data to understand the user's likes and dislikes.

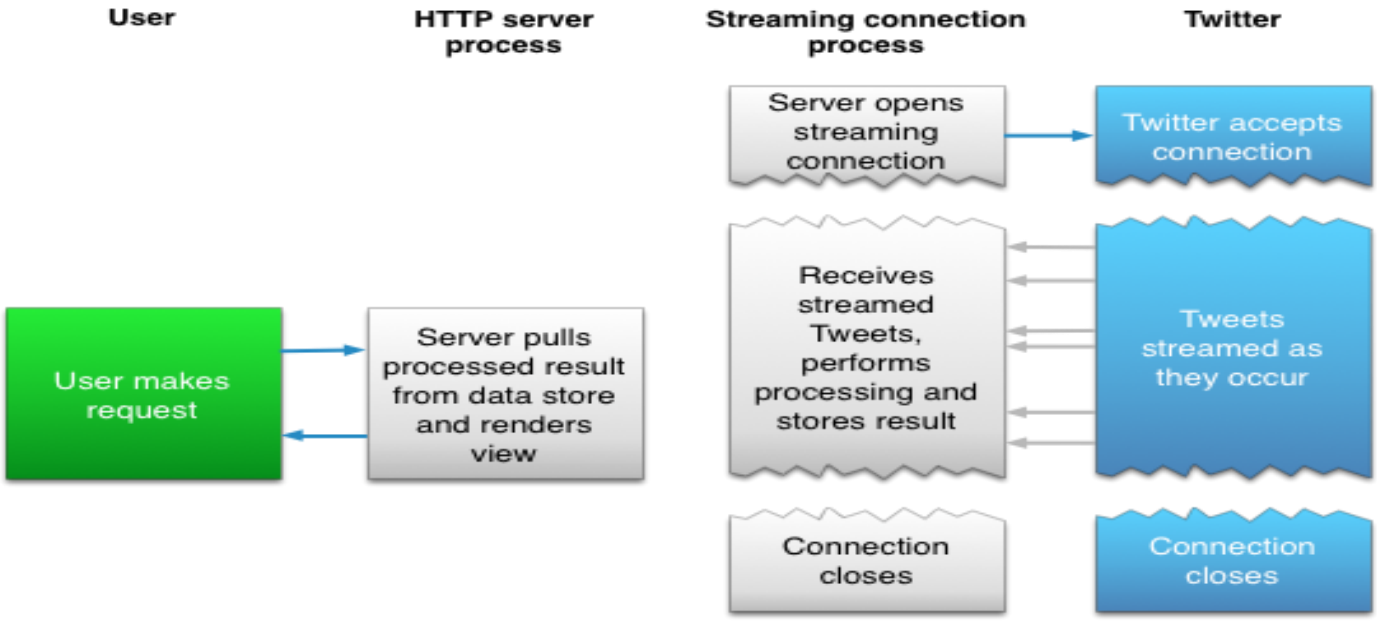


Fig iii.) how Twitter streaming API is harnessed[12]

Even Whatsapp offers Yowsup python library and documentation integration APIs to share mobile implementation notifications on Whatsapp. This create a custom Whatsapp client.

However, in this paper the predictive model has been limited to the usage of Graph API, Streaming API, mobile data, online activity, time of day and location of the hand held device.

The time of day is very significant in obtaining contextual relevancy of the notification being pushed to the client from the server.

The notification detailing abrupt changes in weather would be very important to be delivered just before the user is scheduled to go outdoors, e.g- morning and evening hours. This is the reason, capturing time of day is a vital contribution to the total process of anticipating user's identity hierarchy.

Also the search patterns of the user using search engines and web browser are analyzed. These historical search patterns significantly boost the accuracy of the predictive analysis and estimation of user's identity. A mobile knowledge graph is also created during the implementation when the system accesses the mobile storage of the user and deduces certain user identity based prognostics.

8.) Location tracking has been achieved by using the core location framework that lets us locate the current position of the device.

However, this exercise is power intensive.

Since this solution does not have GPS level accuracy and continuous tracking as critical to its implementation, significant change location service has been used instead of the standard location service.This saves power.

9.) The core of this application has been written in Objective C 2.0 and the IDE used is Xcode . Since the aim was to develop a native iOS app, iOS 7.0 SDK with XCode 5.0.2 suite has been used.

Objective C is similar to C with an addition of small talk style messaging. It is off course object oriented language with objects delegating their tasks on other objects.

Cocoa Touch framework for UI has been used. The Mac OS X is at the heart of this development, as all these resources work on Mac OS X.

10.) There were other available options such as using Appcelerator Titanium and Adobe's Phonegap. In Appcelerator Titanium, HTML, CSS and javascript could have been used to build the native iOS app. This method also takes lesser time and the integration of all APIs as discussed so far at the client end as well as crawling and mining process would remain the same.[10]

But in recent times, Apple created issues against development of iOS apps using third party SDKs and not being written in Objective C.

Therefore, in this paper programming the solution in Objective C in XCode has been done.[16]

The MVC architectural pattern of the mobile application development is also used in the XCode IDE.

This is Model, View, Controller pattern.

11.) Testing and debugging has been done in XCode as it provides a remarkable Simulator simulating iPhone, iPad touch.

APIs have been used in the current implementation.

We can get access to any data gathered by different organizsation using their APIs for the application by using Json/XML API.[9]

e.g in our current implementation, we used APIs for weather, disaster, movie and cricket updates.

Most of the APIs used provide the system access to data via a simple RESTful web interface. Data is available in many languages. Data responses are returned in JSON, JSONP and XML. SSL encryption is also available for secure communication between web browser and the server.[9]SSL stands for Secure Socket layer.

The pull parser or the document object modelling are used for the subsequent parsing of web pages which contained many data types.

12.) The notification threads thus consumed by the user as push notifications can be accepted or rejected by means of the GUI. Again with this user behaviour, notification usage patterns and data consumption patterns can be determined.

IV. Conclusion

It is therefore concluded that an iOS based advanced mobile implementation is presented which

a.) Predicts the objects of interests of the user by modelling data from user's social graph, past online search activity,mobile graph, location and time of day.

b.) Processes the crawled web content into useful information threads using semantic text categorization and other processing and classifying techniques available. The textual content can be gathered using APIs and directly accessing large ecosystems on the web such as Facebook, Twiter, news websites, specific domain organizations.

c.) Supplies the information threads in form of push notifications keeping location into account.

V. Future Work

In future work, I will be using the concepts of artificial intelligence and natural language processing to improve the efficiency of predictive analysis at client and processing of gathered data. Machine Learning must be introduced at the text categorization involved in the processing of collected data from the web. The profiling should be replaced by powerful semantic analysis to better the quality of information in transit to the client. The algorithms need to be precisely improved to improve contextual relevancy factor of notification results.

Better rule synchronization improves the contextual relevancy and also implementing this solution on other mobile platforms such as Android, Blackberry etc.

References

[1] Nielsen Holding's Cross Platform Report, March 2014- http://www.nielsen.com/us/en/reports/2014/an-era-of-growth-the-cross-platform-report.html

[2] The Complete Social Media Community Manager's Guide by Marty Weintraub and Lauren litwinka ISBN 111860582X

[3] The Time Magazine - http://time.com/34025/the-free-marketing-gravy-train-is-over-on-facebook/

[4] Digital Marketing Analytics: Making Sense Of Consumer Data in Digital World by Chuck Hemann and Ken Burbarry ISBN 0133150925

[5] Event Based Programming: Taking events to the limit by Ted Faison. ISBN 1430201568

[6] Mining the Web:Discovering Knowledge From Hypertext Data By Soumen Chakrabarti. ISBN 1-55860-754-4

[7] Research Paper-PAI: Automatic Indexing for Extracting Asserted Keywords from a Document :Naohiro Matsumura ,Yukio Ohsawa ,Mitsuru Ishizuka ISSN:0288 3635

[8] Geopositioning and Mobility by Ahmed-Nai-Sidi- Moh, Maxime Wack, Jaafar Gaber

[9] Restful Web APIs by Leanord Richardson, Mike Amundsen, Sam Ruby. ISBN 1449359736

[10] Building Iphone Apps with HTML, CSS and Javascript: Making Appstore Apps without Objective C and Cocoa By Jonathan Stark .ISBN 1449382916

[11] www.wikipedia.org

[12] www.dev.twitter.com/docs/api/streaming '

[13] Budzik, J. & Hammond, K. (1999). Watson: anticipating and contextualizing information needs. Proc. ASIS, 727-740.

[14] The Forbes - http://www.forbes.com/sites/kashmirhill/2012/07/03/google-nows-terrifying-spine-tingling-bone-chilling-insights-into-its-users/

[15] A News categorization system Philip J Hayes, LauraE Knecht and Monica J Cellio

[16] Cocoa and Objective C by Jeff Hawkins. ISBN 1849690391

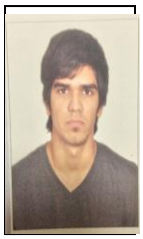
[17] https://developer.apple.com/

[18] Applied Predictive Analytics by Dean Abbott ISBN 1118727932

[19] Multivariate Data Analysis by Hair ISBN 8131715280

[20] Information Retrieval:Searching in the 21st Century by Ayse Goker, John Davies. ISBN 0470033630

About Author (s):



Ankit Sharma is a student of Computer Science engineering and successfully developed a similar web based solution as proposed in this paper and worked with companies like DRDO and Fortis Healthcare etc. His research includes attempts at making personalization of information retrieval processes more efficient.