Aggregation of Categorized Internet Web Content in the Social Circle of a User

Bhanu Luthra
M.Tech. student
Lingaya's University
Faridabad, Haryana, India

Mandeep Kaur
Assistant Professor, School of Computer Sc.
Lingaya's University
Faridabad, Haryana, India

ABSTRACT

In today's World, the consumption of internet content has increased manifold. There are various applications to track this consumption at individual levels, but none to do it for a closed group like a family or a class or a project team. Our work proposes an application which not only tracks the internet consumption of the individuals but also collaborates it for a social group and displays the summary in a categorical format. The categorization done is based on type and level of the internet content .The result of this application can be customized on the basis of inputs given by a user, providing a very crisp summary which has a plethora of practical advantages to various sections of society.

Keywords

Network, Internet, domain, content based,

1. INTRODUCTION

The usage of internet has grown exponentially in the past few years. Not only the amount of data that is being consumed but also the span of audience has widened. With the growing age, there is a need for tracking and reporting the internet content consumed by an individual, by a group and by the complete society in general. There are many applications that work on similar lines and track internet consumption at individual levels. Also the global tracking is being done by big shots in the technology but nothing is being done at a group level. We have come up with an application that not only tracks the usage for a group of people connected together by some purpose, but also categorizes it into various categories. Such an application can have a lot of practical usage for various subsections of society. It can be used as a vigilance tool, as a suggestion mechanism for social searching, for a group of students studying together in a class or anywhere else where a collection of internet content can be useful.

The basic methodology used for developing this application is divided into five steps. In first step we have developed a chrome extension (any other plug in) which tracks the internet content consumed by every user who installs it. The recorded data is stored at the server side where it is processed in further steps. In the next step we have categorized the internet content into various types and levels using already present categorization algorithms. Once the tracked data is categorized, we are retrieving the social circle of a user and clubbing them into groups on the instructions of the user. The categorized data that is consumed by a certain minimum number of users in a group is collaborated and output is presented in the last step. We have ensure to honor the privacy of individual users and hence kept some criteria for aggregating the information for a group. To maintain the utility, we have kept the searching of the consumed content customizable so that a user gets only the crisp and relevant result as per his demand.

Social search is another field which is based on a similar concept. The difference is that in social searching, the search history of various search engines is tracked for a social circle, using which suggestions are made by the search engine for the relevance of the text. This work tracks the complete consumption of an individual and the aggregates it for a social circle. The paper is organized into following sections: Section II discusses the related work.

2. LITERATURE SURVEY

In the similar domain a lot of work has been done for tracking and categorizing the internet content consumption at various levels for several of reasons.

There is a concept of RSS (Rich Site Summary), in which the activities of frequently updated websites are highlighted in an organized and standardized manner. In a RSS feed there is a full or summarized text, plus meta-data of the web sites getting updated. In RSS feeds the contents are categorized in various formats , like on the basis of type of content like video, audio; or on the basis of type of text in the content like sports , education etc ; or on the basis of cities. RSS basically records the updated content or a new published content and pushes it to the users through various inbuilt plug-ins of various browsers. The internet content tracked here is done at generic levels and not in group levels. There is also a concept of tracking of internet services in many organizations and various other places through which there is a check kept on employees in various organizations. There are many software in the market for the same. Mainly the tracking is done by various techniques.

Tracking Methodologies and Perspectives: Generally, tracking is observing the movements depicting the motion of the objects or persons and supplying a timely ordered sequence of their respective locations. [1]Tracking of internet content is keeping a complete track of the information of internet websites being visited by the users including the URL, time of visits, description of the page visited and the emails etc. There are various ways to track the user online by looking at the IP address of the user's connection, the browsing history that helps to gain information about the user, the cookies that are stored on the system, web bugs in the web pages etc. Since the content and services are immense there is a huge progression from cookies to "super cookies" [2]. The other methods are browser finger printing and device identifiers. The "data deluge" or the massive amount of available data makes tracking even more powerful. As per the previous study [3] a novel technique called "principal-based tainting" is developed which tracks java-script information flow originated from different sources, including that from NPAPI-based plug-ins and browser extensions. A lot of work has been done on extracting the web user behavior. A novel way has been developed [4] to collect web browsing histories, that is by collecting data from randomly selected users called panels and hence the panel logs are generated. They include the panel ID, access time of web pages, URL of accessed web pages etc.

As per the study, In the West, Monitoring/tracking appears in all types of firms to control and rationalize work (Rule & Brantley 1992).In various organizations it is being done so that they do not face difficulties by getting virus due to downloaded material and excessive chatting on the internet. [5]The monitoring can be done by network surveillance or by e-mail monitoring. The former is done by tracking down the sites visited by the employees by tracking the keystrokes and latter can be done by reviewing the e-mails and storing the mails sent. As per a recent study, [6], semi-automated and automated methods are developed for capturing, classifying, organizing and analyzing the usage and the content of extremists and hate groups' web sites. It is found that the inter-organizational structure and cluster affinities can be identified by analyzing the hyperlink structures and content of users and constructing their social network maps. Also, as the content on the internet is expanding, researchers are finding it challenging and time-consuming [7, 8] to track new existing new sites. Manual, automatic, semi-automatic processes are applied for gathering the Web sites, classifying them and visualizing the patterns. Web harvesting approaches are applied, firstly to gather web sites back to a local repository for further analysis. [9]After websites are harvested, analysis of the web is being done by web link analysis and web content analysis. Web link analysis is based on hyperlink structure and is used to find links between the communities. [10, 11].

The user expects that many of their online activities are anonymous. But in reality every search, query, click, page view and links are analyzed and used by third parties [12].

As per the study [13] Third-party trackers bring tremendous value to the web: but as are unlimited in scope, they give rise to privacy concerns [14].

Categorization of Internet Content: Categorization of contents is to define under which topic the website or webpage belongs. By Organizing search results into categories users do not need to browse through all the results. It allows users to focus on items on categories. For that we need to have some idea about the document. To retrieve the general idea about the document web page we need to extract the extraneous data in the html document and evaluate that relevant data by counting the words frequency or it can be done by using the metadata tags which are present in the html document and describes the content of a particular website and after evaluation, categorization of the document takes place. [15]

Traditionally, categorization was done manually, by trained human classifiers[16],but as new documents are published every day and new categories emerge, old ones are either removed or updated and take up new meanings, it is very difficult for human categorizers to keep pace with it. Hence, automatic classification is performed by comparing documents representations with representation of categorization and computing the similarities between them. It is being performed in various areas like text categorization [17, 18, 19, 20, 21, 22].It combines information retrieval and machine learning techniques. Various Web search engines use automatic method for classification like Northern Light uses the technique "Custom Search Folders"where it organizes/categorizes document on the basis of similar subject, type, source, language. To implement it on the basis of (a) subject, it divides into 20,000 categories hand –crafted by specialists.(b)type, it uses 150 different document types. Similarly, InfoSeek uses a tool called CCE (Content Classification Engine), which organizes information automatically into categories. It uses two techniques for categorization:

1. It imports and analyzes information from site map and categorize the document.

2. It examines the directory structure and makes guesses about the categorizing the document.

Categorization can also be done by context analysis[23] and for that document must contain hints about it. It extracts contextual information by analyzing the structure of the Web. Today, web directories like Yahoo and Look Smart are used to classify Web pages. An example for a hierarchal structure is "look smart's Web directory" which takes the web pages returned by a search engine classify them into categories. To categorize/organize document into different contexts various techniques are used. Some of them are "structural information"(meta data),table of contents[24].These meta data are used by various systems like Dynacat System by Pratt[25],Allen[26] etc. Link Structure of Web pages are also used [27].

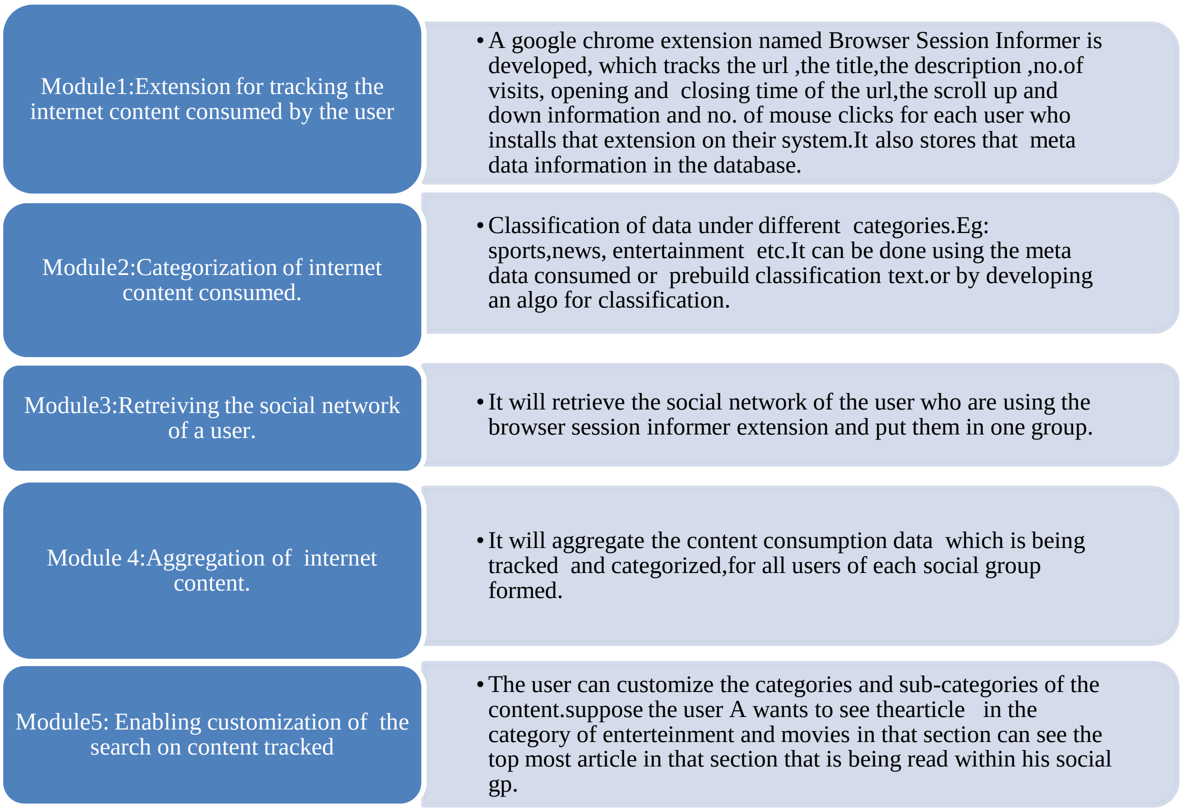
3. PROPOSED METHOLOGY

The desired functionality will be attained with help of five modules as shown in Figure1. Module 1: Tracking of internet Content : This can be done by building a browser plug-in or an extension (depending upon the internet browser). Every user will be given an extension ID and will have a common username as that for his face book / social network. Through extension we can track the basic information about the internet page like its URL and metadata along with other information linked with a user like opening time , closing time, scroll up time , no. of clicks , session id etc. The entire data picked up by the tracking extension can be stored on the server side for the given extension ID and username, where further processing of the data can be done for the next modules.

Module 2: Categorization of internet content consumed by a user into various types and levels: This will be done by developing an algorithm first, on the basis of which categorization will be done. The categorization on the basis of type of content is quite easy, but the challenging part is the level categorization. The algorithm will first categorize the content into various types. For example, all the urls tracked will be put into various categories like entertainment, sports, news, knowledge, medicare, social network etc and further subtypes like in entertainment movies, plays, online games etc. Once the type based categorization is done, then the algorithm will categorize the content into levels like amateur, professional, naive, research or on the basis of age. The second part of algorithm might involve artificial intelligence.

Once the algorithm is developed, the tracking record and data generated in module 1 will be categorized into various types and levels.

Module 3: Fetching the social network of a user, who also use the tracking extension, and making a social group of all such linked members: Using the extension username, the facebook / social network of a person will be retrieved.



Module1:Extension for tracking the internet content consumed by the user

- A google chrome extension named Browser Session Informer is developed, which tracks the url ,the title,the description ,no.of visits, opening and closing time of the url,the scroll up and down information and no. of mouse clicks for each user who installs that extension on their system.It also stores that meta data information in the database.

Module2:Categorization of internet content consumed.

- Classification of data under different categories.Eg: sports,news, entertainment etc.It can be done using the meta data consumed or prebuild classification text.or by developing an algo for classification.

Module3:Retreiving the social network of a user.

- It will retrieve the social network of the user who are using the browser session informer extension and put them in one group.

Module 4:Aggregation of internet content.

- It will aggregate the content consumption data which is being tracked and categorized,for all users of each social group formed.

Module5: Enabling customization of the search on content tracked

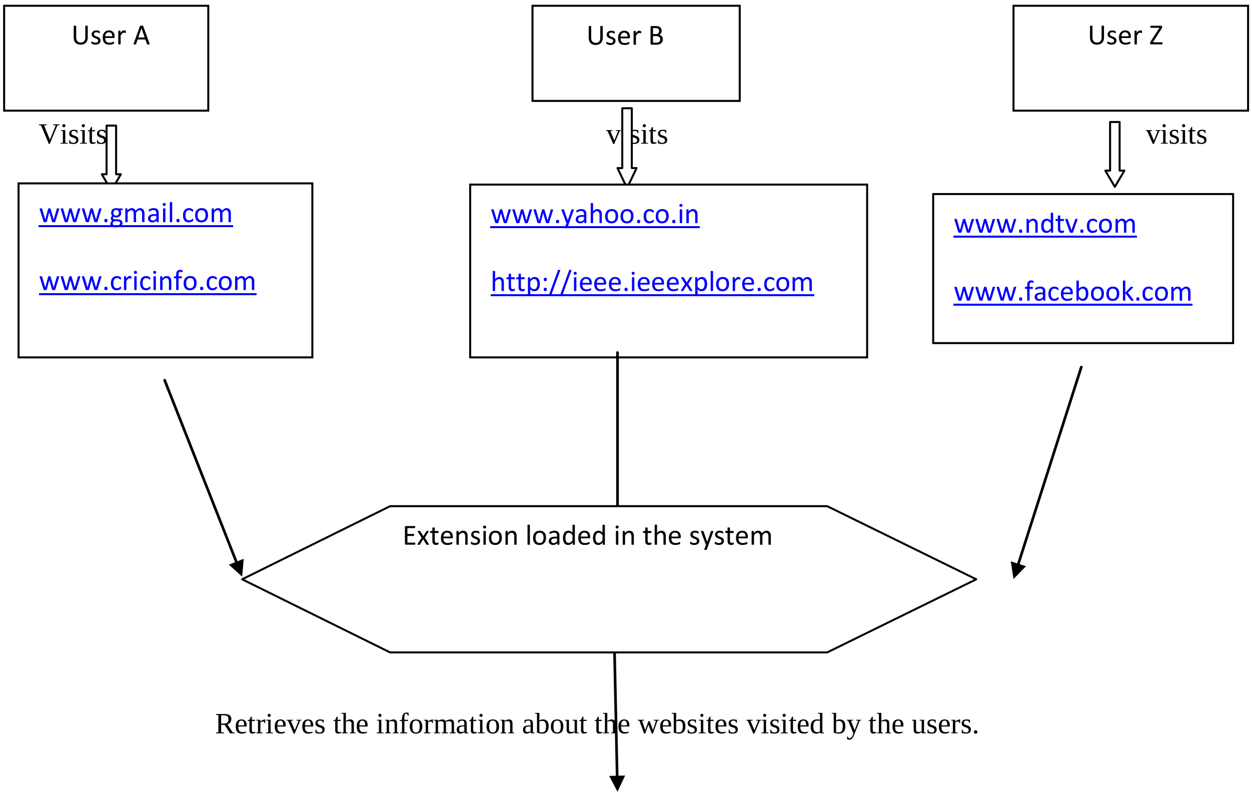
- The user can customize the categories and sub-categories of the content.suppose the user A wants to see thearticle in the category of enterteinment and movies in that section can see the top most article in that section that is being read within his social gp.

Figure 1: Proposed Methology

This can be done by various methods like using the facebook apis. Once the list of the entire social network of a user is there , groups will be formed of the some users as per the instructions given by the users. These users may be connected to each other through a common cause, a common relation or something else that connects them. The purpose of forming them into one group is to have a collective usage summary for this group rather than for an individual.

Module 4: Aggregating all the internet content tracked by module 1 and categorized by module 2 for a social group formed by module 3: Once the social group is formed, the entire tracking data recorded and categorized in the previous two modules will be clubbed together so as to get a complete group record. This record will give the complete consumption analyses for a social group. Privacy of an individual will be maintained and therefore only those records will be picked up this aggregate engine which specifies certain criteria set to ensure the privacy.

Module 5: Developing an ability to customize the searching from the internet content consumed by a group: Since the data recorded and stored can be very large and of various types, there is a need for customization in the presentation of the


User A
User B
User Z
Visits
visits
visits
www.gmail.com
www.cricinfo.com
www.yahoo.co.in
http://ieee.ieeexplore.com
www.ndtv.com
www.facebook.com
Extension loaded in the system
Retrieves the information about the websites visited by the users.

Login ID	Url	Description		Opening Time	Closing Time
hariluthra@gmail.com	www.cricinfo.com	Keeps you updated with all information about the cricket.		02:30 06-04-2013	14:00 06-04-2013

Figure 2: Module1

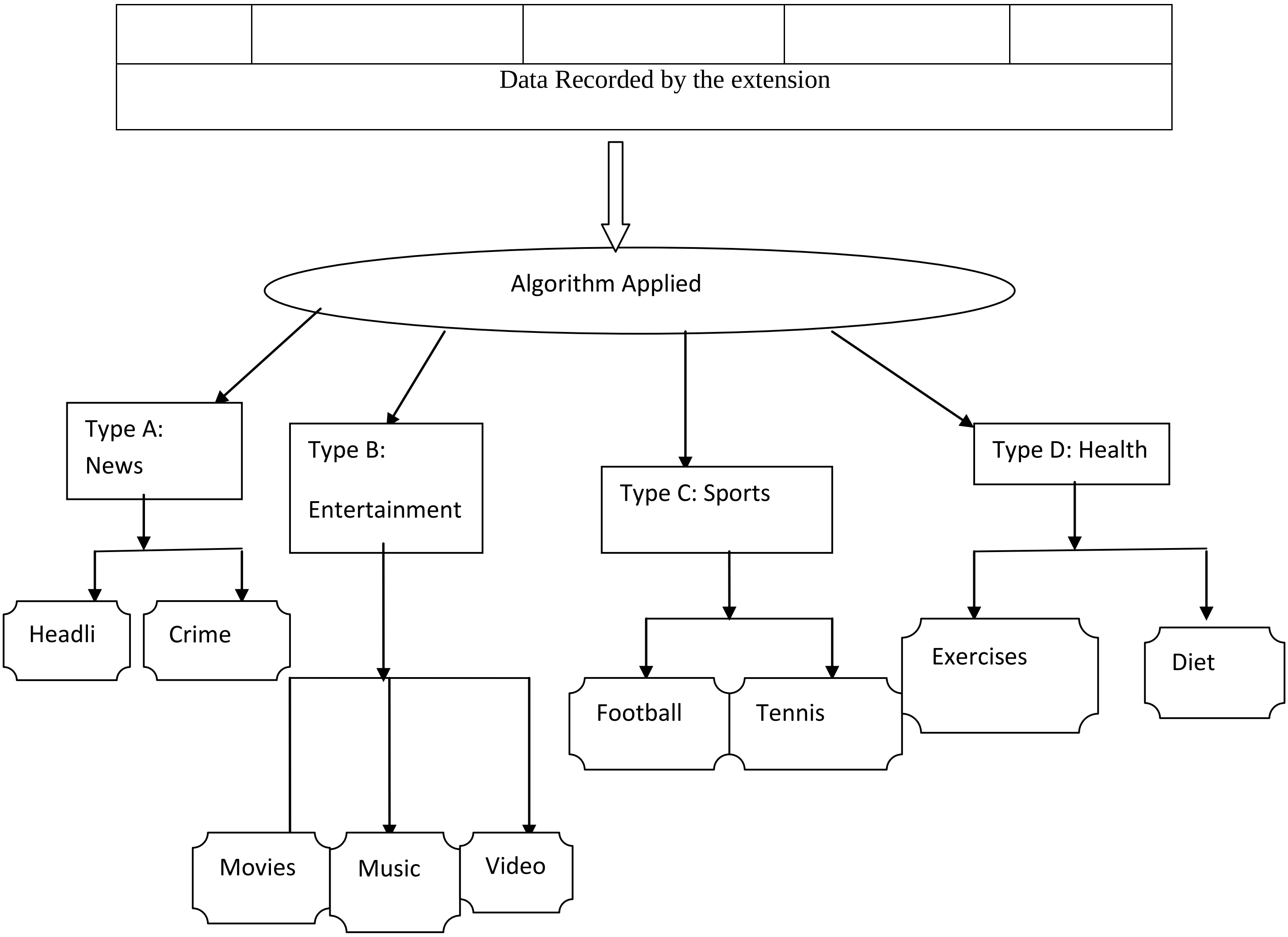


Figure 3: Module 2

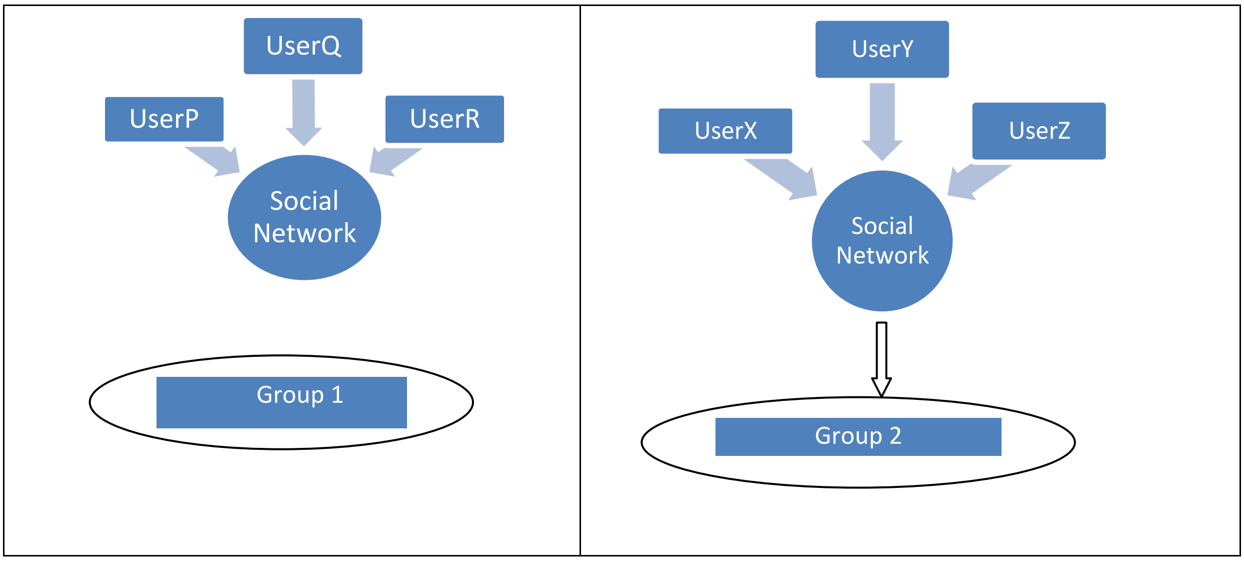
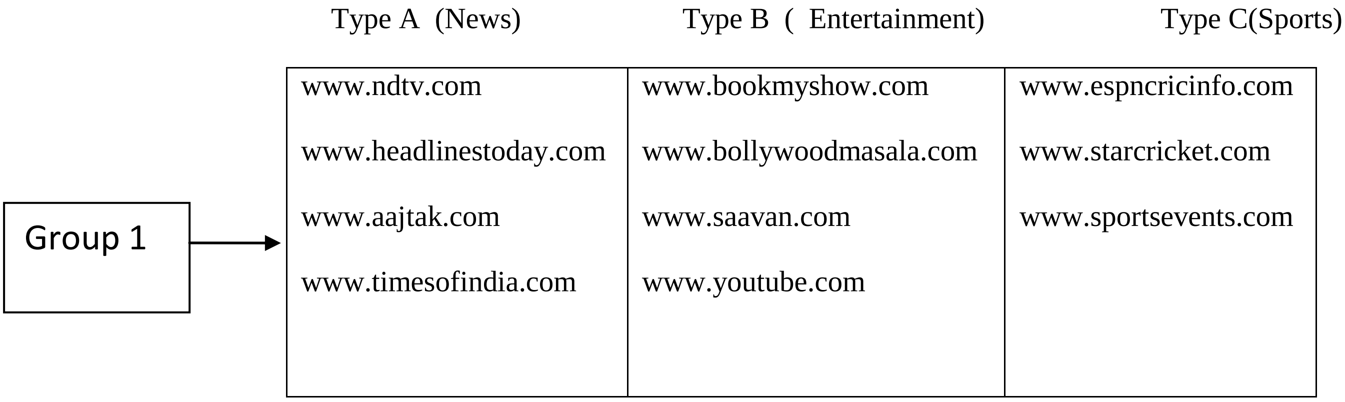


Figure 4: Module3



Type A (News)	Type B (Entertainment)	Type C(Sports)
www.ndtv.com	www.bookmyshow.com	www.espncricinfo.com
www.headlinestoday.com	www.bollywoodmasala.com	www.starcricket.com
www.aajtak.com	www.saavan.com	www.sportsevents.com
www.timesofindia.com	www.youtube.com	

Figure 5: Module 4

data. This module will basically present the data on the demands of a user. For example, if one has to see the usage of a group A for a period of two weeks in the technology type, then through this engine the user will be presented the demanded data. Once the desired functionality is attained, the generated information can be very useful

1. The information gathered can be used for suggestions in social searching. Anybody willing to search for a reference text at a search engine can be suggested the articles most read by his /her group or by people in his close nexus.

2. The information gathered can be useful for sales and marketing team for various corporate as this information gives an idea about the internet consumption of a group. So a corporate/businesses can post advertisements on the most visited websites by a particular group that is to be targeted.

3. The entire information gathered can also be used for detective purposes. If one has to keep a track on the content consumed by a particular social circle then it can be done through this method.

4. ISSUES AND CHALLENGES

There are many challenges and hurdles in the developing the entire modules. Some of the expected challenges are

1. In Module 1, it is hard to detect the real time spent on a particular page. A person might open up a page and then go for a stroll. Only the closing time and opening time is definitely not sufficient to actually check the importance of the page.

2. In Module 2, it is very hard to judge the level of page. It is a very subjective thing. Tough level for a particular person may be easier for somebody else. Moreover just by building a piece of code, the exact replacement for the human judgment is hard to be accomplished.

3. In Module 3, list of people in one group can be inadequate. Less numbers will not generate a conclusive consumption pattern. On the contrary large numbers may test the optimization part of the code.

4. In Module 1, the data generated can be very bulky. So there needs to be some optimization on the code part to cater to the challenges of large data entries.

5. In Module 3, the groups formed can be irrelevant, as a social network may have a large set of acquaintances.

5. CONCLUSION AND FUTURE SCOPE

We proposed to collaborate various systems and tools to get the desired tool. There were tracking engines, there were categorization algorithms, there were social network retrieval APIs but there was nothing to collaborate them and provide the desired result. Our entire focus is in seem less integration of the present sub tools with the required tweaking and modifications. The future work will focus on following issues:

1. The group formation in the third module is not customizable in my project. Third module can be tweaked in order to form groups on demand of users willing to have a common content consumption analysis.

2. Presence of large data always opens the scope of optimization of the codes and algorithms.

3. Suggestions to the search engines can be extended so as to optimize the content search on search engines.

4. Reporting structures can be developed to give a monthly or a weekly consumption analysis to a group of users.

5. Various other algorithms can be implemented on the content tracked to get different inference of the tracked and recorded data.

6. REFERENCES

[1] "TrackingSystem"http://en.wikipedia.org/wiki/Tracking_system, (October 2007).

[2] Nicholas Jackson, The Next Online Privacy Battle: Powerful Super cookies, ATLANTIC (Aug. 18, 2011, 10:31 AM), http://www.theatlantic.com/technology/archive/2011/08/the-next-online-privacy-battle-powerful-supercookies/243800.

[3] Minh Tran, Xinshu Dong, Zhenkai Liang, and Xuxian Jiang" Tracking the Trackers: Fast and Scalable Dynamic Analysis of Web Content for Privacy Violations," Department of Computer Science, North Carolina State University, School of Computing, National University of Singapore.

[4] Shingo Otsuka, Masashi Toyoda, Jun Hirai, Masaru Kitsuregawa," Extracting User Behavior by Web Communities Technology on Global Web Logs", Database and Expert Systems Applications ,Lecture Notes in Computer Science Volume 3180, 2004, pp 957-968

[5] Erwin A. Alampay, Ma. Regina M. Hechanova "MONITORING EMPLOYEE USE OF THE INTERNET IN PHILIPPINE ORGANIZATIONS" EJISDC (2010) 40, 5, 1-20.

[6] Yilu Zhou, Edna Reid, Jialun Qin, Hsinchun Chen and Guanpi Lai "U.S. Domestic Extremist Groups on the Web: Link and Content Analysis".

[7] CNN, U.S. Hate Groups Hard to Track, July 1999.

[8] P. B. Gerstenfeld, D.R. Grant, C. Chiang, "Hate Online: a Content Analysis of Extremist Internet Sites," Analysis of Social Issues and Public Policy, vol. 3, 1:29-44.

[9] R. Kay, "Web Harvesting," Computer World, June 21, 2004 http://www.computerworld.com. Accessed January, 15, 2005.

[10] D. Gibson, J. Kleinberg, P. Raghavan, "Inferring Web Communities from Link Topology," Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, ACM, 1998.

[11] E. O. F. Reid, "Identifying a Company's Non-Customer Online Communities: a Proto-typology," Proceedings of the 36th Hawaii International Conference on System Sciences, Springler, 2003. http://e-business.fhbb.ch/eb/publications.nsf/id/214. Accessed June 18, 2004.

[12] Omer Tene and Jules Polonetsky "To Track or "Do Not Track": Advancing Transparency and Individual Control in Online Behavioral Advertising", 7TENE POLONETSKY FINAL_JAD,(2/28/2012,11:25 AM).

[13] Jonathan R. Mayer and John C. Mitchell" Third-Party Web Tracking: Policy and Technology", Stanford University Stanford, CA., 2012 IEEE Symposium on Security and Privacy.

[14] Christopher Butler," Unlimited vs. Limited Web Tracking",http://www.newfangled.com/unlimited_vs_limited_web_tracking",September,2010.

[15] John O'Rourke "Automating a user defined Categorization of the Web", Senior Research Proposal Draft 3.

[16] Giuseppe Attardi, Antonio Gulli and Fabrizio Sebastiani "Automatic Web page Categorization by Link and Context Analysis",(1999).

[17] Ittner,D.D.,Lewis,D.D.,Ahm,D.:"Text categorization of low quality images", Proceedings of SDAIR-95,4th Annual Symposium on Document Analysis and Information Retrieval,Las Vegas,US,301-315,1995.

[18] Lewis,D.D.,Schapire,R.E.,Callan,J.P.,Papka,R.:"Training algorithms for linear text classifiers", Proceedings of SIGIR-96,19th ACM International Conference on Research and Development in Information Retrieval,Zurich,CH,298-306,1996.

[19] Ng,H.T.,Goh,W.B.,Low,K.L.:"Feature selection, perception learning, and a usability case study for text categorization", Proceedings of SIGIR-97,20th ACM International Conference on Research and Development in Information Retrieval,Philadelphia,US,67-73,1997.

[20] Schutze,H.,Hull,O.A.,Pedersen,J.O.:"A comparison of classifiers and document representations for the routing problem", Proceedings of SIGIR-95,18th ACM International Conference on Research and Development in Information Retrieval,Seattle,US,229-237,1995.

[21] Yang, Y.: "Expert Network: effective and efficient learning from human decisions in text categorization and retrieval", Proceedings of SIGIR-94,17th ACM International Conference on Research and Development in Information Retrieval, Dublin, IE, 13-22, 1994.

[22] Yang, Y.: "An evaluation of statistical approaches to text categorization", Technical Report CMU-CS-97-127, School of Computer Science, Carnegie Mellon University, Pittsburgh, US, 1997. Forthcoming on the Information Retrieval Journal.

[23] Hao Chen and Susan Dumais "Bringing Order to the Web: Automatically Categorizing Search Results", Computer Human Interaction - CHI, pp. 145-152, 2000.

[24] Marchionini, G., Plaisant, C., and Komlodi, A. Interfaces and tools for the Library of Congress national digital library program. Information Processing and Management, 34, 535-555, 1998.

[25] . Pratt, W. Dynamic organization of search results using the umls. In American Medical Informatics Association Fall Symposium, 1997.

[26] Allen, R. B., Two digital library interfaces that exploit hierarchical structure. In Proceedings of DAGS95: Electronic Publishing and the Information Superhighway (1995).

[27] Maarek, Y., Jacovi, M., Shtalhaim, M., Ur, S., Zernik, D., and Ben Shaul, I.Z. Web Cutter: a system for dynamic and tailor able site mapping. In Proceedings of the 6th International World Wide Web Conference (Santa-Clara CA, April 1997).