Portability in User Profiles: Single User – Multiple Devices

Saniya Zahoor ,Dr. Mangesh Bedekar ,Pranali Kosamkar
MIT, Dept. of Comp.Engg.,Pune

Abstract-The World Wide Web (WWW) has become the single largest source of information to mankind always accessible at his fingertips. The data content available on the web for the user to consume is enormous in proportions and at times makes it difficult to navigate to the right web-page. The web is what the user sees through the browser, which facilitates surfing and searching. Today's web browsers are smart applications in themselves and they can monitor user activity in near real time, assist the user in his routine tasks, have enhancements desired by the user etc. Today's search engines provide users with information relevant to users. Some common problems with search engines are Short queries and Word ambiguities. Users have a preferred browser which they usually use, and have a preferred set of sites for particular tasks. Users tend to repeat their search queries and return to the web-pages already visited. Users are ready to provide extra information for more personalized web experience [1].

Keywords-User profiling, WWW, Search engine.

INTRODUCTION:

Profiling a user is to understand all that the user does in his day-to-day tasks. It encompasses all of the user's details, his usage patterns, his behavioral characteristics, his likes and dislikes. More the time spent "understanding" the user, richer is the user profile. Most web-applications deployed today profile its users in various manners, known or unknown to the end user. It is omnipresent in most places, visible at times, hidden elsewhere. There are two ways to do profiling: Explicitly and Implicit way. In explicit profiling, we can get annoyed during building but it is easy to learn. The disadvantages are that these can hold wrong or incomplete information and we can learn only what is stated by the user. In implicit way, learning is transparent to build and once learned cannot be denied. Although it is difficult to learn also but we can learn almost everything about the user.

There is a need of User Profiling. Accurate information about the user's interests is collected and represented with minimal user intervention. User Profiling helps to address the information overload problem, disambiguates web search terms, deliver personalized web search results and make the web more relevant and personal. Also, Dynamism of user's behavior can be met and it helps in managing Privacy issues as well.

User may use, Laptop while at Home, Desktop in Office, Mobile while travelling and so on. His purpose of use and behavior will be different on different devices [hence different profiles]. There is a need to synchronize his behaviors [on different devices] and get the best combination of profile behaviors for him. It's the same user and they are his tasks which though disjoint are related and relevant. The more the user accesses different devices, the more detailed behaviors we will and synchronizing them will lead to better and richer profiles. More and relevant information can be given to a user. Even if one Device fails [is changed] the user still can get the relevant information from other device.

DATA FOR USER PROFILING:

Since we will be doing profiling implicitly [2], so we need to keep track of actions what user will perform every time he accesses the web. His course of actions that he performs are the important source of data for profiling the user [3]. Data is collected from the following elements:

a. The Accessed URL
b. Time of access of URL.
c. Total Time and Active Time spent on a particular web-page.
d. Bookmark a webpage.
e. Save a file.
f. Print a file.
g. Mouse movements while viewing a web page.
h. User clicks on browser controls.
i. clicking the "Back", "Next" button.
j. Amount and Speed of scrolling the web page.
k. number of times he copies text into the clipboard.

All these are the sources of data for user profiling. [4]

USER PROFILES - STRUCTURE & REPRESENTATION:

Following are the methods for structuring and representing user profiles [5, 6, 7]:

a. BAG-OF-WORDS:
Initially, bag is empty. As the User accesses the web, bag will be filled with words. Each word will have weight assigned to it and weight of a word will depend on the number of occurrence of that word. We can use Unigram (single word), bigram, trigram and N-grams.

b. CLICK BEHAVIORS:
User performs lot of actions on the webpage like he can click on the next button, back button, or he may click on the first line then scroll two lines n then click, so on. All his clicking actions on the webpage form a sort of pattern, which tells us about his interest and thus helps us to structure and represent his User profile well.

c. INCREMENTAL ALGORITHM:
Users are the dynamic Dynamism of User's behavior can be met well by incremental algorithm. Incremental algorithm represents the User profile as a vector whose dimensions (time they spent on some site, sites they visit

and so on) change over a period of time. Incremental algorithms do the incremental updates as the User's behavior

changes over the time.

d. LONGEST COMMON SUBSEQUENCE (LCS):
longest common subsequence is the most commonly used methods for recommendation systems. For example, there is a user u1 that visits facebook, yahoo, then gmail. And suppose another user u2 comes in the same System and he visits facebook, then the longest common subsequence will recommend him yahoo and gmail. This is how longest common subsequence works and thus helps to represent the user profiles well. Main focus is to get the subsequence that is common to subsequences present in the system.

e. WEB ACCESS PATTERN TREE:
Also known as WAP tree. From the log of the user, access pattern of the user is obtained and from that user's interests are predicted well.

f. K-MEANS CLUSTERING ALGORITHM:
K- Means clustering algorithm is one the methods of structuring and representing User Profile of a user. In this, k is defined first and within this clusters are formed such that the intra- cluster distance is minimized and inter-cluster distance is maximized. We can then correlation between the user's behavior and cluster same users in the same cluster. For example, most recommendations systems say movie/music/item based recommenders use this algorithm to form correlation among users. A user belonging to some category get the recommendations related to the taste of that category. Thus, user to user correlation is done by this algorithm.

g. K-NEAREST NEIGHBOR (KNN) COLLABORATIVE FILTERING:
k-nearest neighbor collaborative filtering is used to form association between same attributes like in movie recommendation systems, movies belonging to the same genre are grouped together by k-nearest neighbor collaborative filtering algorithm.

STORAGE AND ACCESS METHODS:

First we need to collect the data, store it somewhere, make patterns out of it, and use the patterns to predict user's interest well and recommend him accurately[8, 9]. Thus we need to do two tasks:

- Store the data
- Access the data

a. RELATIONAL DATABASES:
Relational Database can be used for:

1. Classification: we use the various methods for classification like Decision Trees, Neural Networks, Support Vector Machines (SVM).
2. Association Rules: Association rules are derived from the patterns created by the User in his active time spent surfing the web.
3. Clustering: Clustering is the unsupervised method for grouping similar tasks performed by the user and helps us to store and access user profiles well.

b. FLAT FILES: Flat files have been largely used as the data source of information used by data mining algorithms. Even when operational data is stored in a DBMS, many systems extract a portion of them to be used during the mining process, working only with data that is in main memory, which limits the total amount of information to be used. Implementations using flat files offer some advantages related to performance results. We store data in files having comma separated value format, no need to look for any particular database, like relational database, here we can store.

c. XML: We can also use XML pages to store and access data. Two types of
XML pages can be used, we can also use XML to store and access database. In XML mining can be done using association rule, clustering. There are two types of XML pages:

1. Static (ASSOCIATION RULE EXTRACTION ALGORITHM).
2. Dynamic (WEIGHTED-FP GROWTH ALGORITHM).

There is a difference between static XML pages and dynamic XML pages in the preprocessing step. Dynamic XML needs preprocessing and static XML does not need preprocessing step.

PROPOSED SYSTEM:

User usually use multiple devices, mobile, laptop, desktop, tablet. He uses them for a varying purpose and behaves differently on these devices. We will thus have different profiles of the same user on different devices. All of his profiles need to be synchronized with each other to get the best out of his user profile. This leads to the need of Portability of the user profiles, comprising of the following headings:
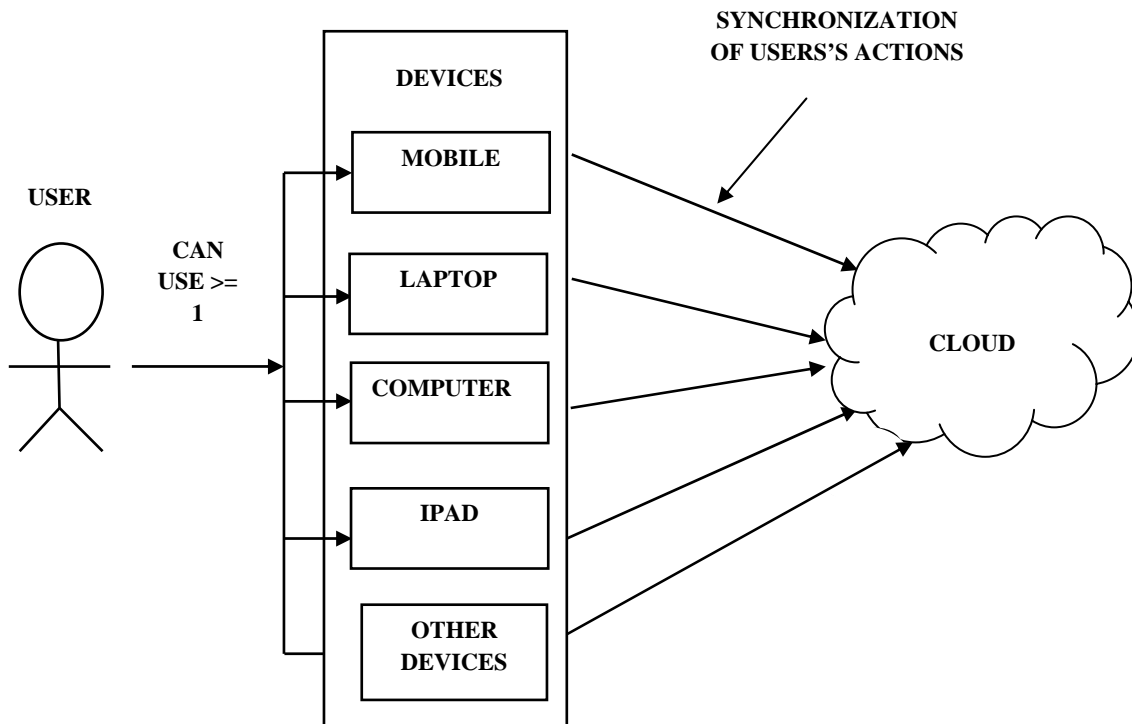
a. Single User
b. Profiling of a single user on multiple devices
c. Model a single user by many micro profiles that best represent the user
d. Multiple Devices (Desktop @ Office, Laptop @ Home, Phone @ Travel)
e. Synchronization of a single user's profile on multiple devices through the use of a common point (CLOUD). C2DM [Cloud 2 Device Messaging] technology can be easily used for this purpose.

HYPOTHESIS:

Weight assigned to a Device is directly proportional to the amount of time spent on the device (Example - Mobile (40%), Desktop(30%), Laptop(20%), Tablet (10%)) and Weights assigned to all User Activities (sites he visits, files downloaded, web-pages bookmarked, printed, time spent reading them, submitting a search for querying, clicking the browser "Back" , "Next" button,…). Websites will be arranged according to Weight(W) given by the following relation,

Weight (W) = Weight assigned to Device * Weight assigned to Activities

∑W(DEVICES)=100%

Higher the weight, higher its relevance to the user. User might be an active [Heavy] user of the Web. The size of the User Profile [on each Device] may grow to an alarming size. Users Behavior changes over time. Sites [Web Pages] highly relevant at a particular time might be less relevant over a period of time. Time Decay Mechanism is introduced into user profiling to maintain an optimum balance between, Time v/s Size [Space] v/s Relevancy tradeoffs. Frecency handles the drift in user profile easily. It helps in Learning, Un-Learning and Re-learning of users changes in behavior automatically without user interaction. These, thus, pruned user profiles are then synchronized with other profiles of the same user over the Web, in an incremental manner. Incremental updates use optimum combination of time and data required to be transferred to-and-fro from the device to the cloud. Cloud time-stamping ensured "Old" behavior patterns are "updated" with the "New" ones.

Case Study:
    Users:student,C.E.O,common man.
    Devices:Mobile,Laptop,Pc,Ipad.
    a.   User :Student
        Devices: mobile, laptop.
Student will have mobile and laptop. So initially when our application will run of these two devices, these will have 50% each weightage. Depending on the usage, weights will gradually change and if the device is not used, its weight will gradually decrease to zero.
    b.   User: C.E.O
        Devices: mobile, laptop, pc, ipad.
C.E.O may use mobile, laptop, ipad, pc. So initially when our application will run on these four devices, weights will

be 25% each on these four. Then depending on usage, weights will increase or decrease.
    c. User: Common man
        Devices: mobile.
So initially weight will be 0. As the usage increases, weight will increase to 100%. If at any point he uses some other devices, then weight will be 0 and gradually increase compared to his usage on mobile. Assumption is sum of weights of all devices has to be 100%.

Weight assumptions:

DEVICE COMBINATIONS	WEIGHTAGE
Single Device	100%
Two Devices	50% Each
Three Devices	34% One And 33% Other Two
Four Devices	25% Each
Five Devices	20% Each

If there are n number of devices we will assign weight 100% to the first device and then share the weights accordingly between n number of devices. As the second device comes into play, the weight of first device will decrease from 100% and the share will go to second device depending upon how much it is used in comparison with the first.

Assigning weights to the devices is a balance. We need to go for comparative usages of all devices, the more used device will get higher share of weight out of 100%, and rest will also get the share according to their usage with respect to the rest of the devices. This is about assigning weights to devices.

Now how to assign weights to the actions that user performs on web page. The hypothesis remains the same i.e; total weight of the actions of a particular user has to be equal to 100%. And weights change according to how recent the actions are and how frequent the user is. Example: Printing a file means that the file is more relevant to the user compared to if we bookmark that webpage only. Thus, all actions can be prioritized and assigned weights according to the actions they perform. Likewise, links that they search frequently are more relevant than they don't search so frequently. Thus, basic principle remains the same while assigning weights to user's actions.

CONCLUSION:

The proposed system will lead to increased Relevancy by personalizing the web, Pre-fetching web-pages already read, providing search term suggestions [Personalized Auto-Complete], Re-ordering search results already visited, recommending web pages already visited, recommending search terms already queried. Also, Identification of Users physical Location from combination of devices adds physical context to the entire operation. It leads to easy Inter-operability i.e; new devices can be easily and automatically synchronized if any device(s) fail [11, 12,13].

BIBLIOGRAPHY:

[1] Josna Jojo, Dr. N. Sugana, C. Lee Giles," User profile creation based on navigation pattern for modeling user behavior with personalized search", in Current Trends in Engineering and Technology (ICCTET), IEEE, pp.371-374, 3 July2013.

[2] Ugur Cetintemel, Michael J. Franklin, C. Lee Giles, "Self-Adaptive User Profiles for Large-Scale Data Delivery", in Data Engineering, 2000. Proceedings of the 16th International Conference, pp. 622-633, 2000.

[3] Jie Yu, Fangfang Liu,"Mining user context based on interactive computing for personalized Web search", in Computer Engineering and Technology (ICCET), 2010 2nd International Conference, pp. 16-18 April 2010.

[4] Saniya Zahoor, Dr. Mangesh Bedekar, Implicit client side user profiling for improving relevancy of search results, CCSEIT-2014, in Proceedings of Fourth International Conference on Computational Science, Engineering and Information, Technology, Army Institute of Technology, Pune, India, August, 8-9, 2014.

[5] Hawalah, A.Fasli, "A hybrid re-ranking algorithm based on ontological user profiles", in Computer Science and Electronic Engineering Conference (CEEC), pp. 50 – 55, 13-14 July 2011.

[6] Bangzuo Zhang , Yu Guan, Haichao Sun , Qingchao Liu , Jun Kong," Survey of user behaviors as implicit feedback", in Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010 International Conference, pp. 345 – 348, 24-26 Aug. 2010.

[7] Muhammad Ali Zeb, Maria Fasli. "Interest Aware Recommendations based on Adaptive User Profiling" Published in: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on (Volume:3) pp 357-360, 22-27 Aug. 2011.

[8] Shahabi,Zarkesh, A.M. , Adibi, J. , Shah, V," Knowledge discovery from users Web-page navigation " in Research Issues in Data Engineering, 1997. Proceedings. Seventh International Workshop, pp. 20 – 29, 7-8 Apr 1997.

[9] Udechukwu,Barker, K.,Alhajj, R., "Mining user navigational patterns in dynamically changing environments" in Information Technology: Coding and Computing, 2004. Proceedings. ITCC-2004. International Conference on (Volume:2),pp. 372 - 377 Vol.2, 5-7 April 2004.

[10] Xu Qi. "Research on User Profiling Technology for Personalized Demands" In Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on (Volume 3) pp 198 – 201, 11-12 May 2010.

[11] Wei Wang,Dongyan Zhao, Haining Luo, Xin Wang. "Mining User Interests in Web Logs of an Online News Service Based on Memory Model" Published in: Networking, Architecture and Storage (NAS), 2013 IEEE Eighth International Conference, pp 151- 155, 17-19 July 2013.

[12] Jie Yu, Fangfang Liu. "Mining User Context based on Interactive Computing for Personalized Web Search" Published in: Computer Engineering and Technology (ICCET), 2010 2nd International Conference on (Volume:2), pp 209-214, 16-18 April 2010.

[13] Viviani, Bennani, N.Egyed-Zsigmond, E., "A Survey on User Modeling in Multiapplication Environments, in Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services (CENTRIC), 2010 Third International Conference, pp.111 – 116, 22-27 Aug. 2010.