# Profiler for Smartphone Users Interests Using Modified Hierarchical Agglomerative Clustering Algorithm Based on Browsing History

Priagung Khusumanegara, Rischan Mafrur, and Deokjai Choi(✉)

Department of Electronics and Computer Engineering,
Chonnam National University, 77 Yongbong-Ro, Buk-Gu,
Gwangju, South Korea
{priagung.l23,rischanlab}@gmail.com, dchoi@cnu.ac.kr

**Abstract.** Nowadays, smartphone has been a life style for many people in the world and it has become an indispensable part of their live. Smartphone provides many applications to support human activity which one of the applications is web browser applications. People spend much time on browsing activity for finding useful information that they are interested on it. It is not easy to find the particular pieces of information that they interested on it. In this paper, user-profiler is presented as way of providing smartphone users with their interest based on their browsing history. In this study, we propose a Modified Hierarchical Agglomerative Clustering algorithm that uses filtering category groups on a server-based application to aid provides smartphone user profile for interests-focused based on browsing history automatically. Based on experimental results, the proposed algorithm can measure degree of smartphone user interest based on browsing history of web browser applications, provides smartphone users interests profile and also outperforms the C4.5 algorithm in execution time on all memory utilization.

**Keywords:** Smartphone · User interests · Modified Hierarchical Agglomerative Clustering

## 1 Introduction

Today, many vendors such as Google and Yahoo store historical data in a users' browser to understand the type page that user is visiting. This information is used to show ads that might appeal to users based on their inferred interest categories. For example, if a user browses many sport-related websites displaying AdSense ads or watch sport-related videos on YouTube, it means Google and Yahoo may associate a sport interest category with their history and show the user sport-related ads. Information about user interests is useful both for users and service providers, user can easily find the information that they needed so they do not spend time to find it and in point of view service providers, they can also easily to provide advertisement and recommendation to the users who use their service. It is not easy to find the particular pieces of information that users interested on


I. Khalil et al. (Eds.): ICT-EurAsia 2015 and CONFENIS 2015, LNCS 9357, pp. 89–96, 2015.
DOI: 10.1007/978-3-319-24315-3_9

it. In this paper, we implement a server-based application to provide smartphone user profile based on browsing history of web browser application. We propose a Modified Hierarchical Agglomerative Clustering algorithm that is inspired by Hierarchical Agglomerative Clustering algorithm. Our method can automatically provide smartphone user profile for interests-focused based on browsing history of web browser applications. First, we extract the useful information of historical web browser applications from smartphone users. Second, we use a distance function to calculate similarity distances between the extracted data. Third, we use Modified Hierarchical Agglomerative Clustering algorithm that use filtering category groups to provide smartphone user profile for interests-focused. The reminder of this paper is structured follows. The Sect. 2 describes the previous studies. The data extraction and user profiling algorithm is presented in Sect. 3. We then show the experimental results and evaluations of our work in Sect. 4. Finally, we conclude our findings and suggestions for future research in Sect. 5.

## 2 Related Work

In this section, we will review some existing works on web log data mining. Previous researchers have investigated how to generate user profile based on web server data logs using various data mining technique. Most of the approaches concerned on user classification (supervised) method and clustering (unsupervised) method based on useful information from web server data logs. In Jian's et al. work [1], classification (supervised) method is used to predict users' gender and age from web browsing behavior. Santra et al. [2] research about identification interested users using naïve Bayesian classification based on web log data and also comparison between decision tree algorithm C4.5 and Naïve Bayesian Classification algorithm for identifying interested user. JinHua Xu et al. [3] used KMeans algorithm for clustering web user based on web data logs. Xia Min-jie et al. [4] research using clustering technique based on web logs and users' browsing behavior to implement an ecommerce recommendation system. Neetu et al. [5] used classification technique to predict kid's behavior based on collected internet logs. Li et al. [6] focused on web log data processing to analyze and research the user's behavior. Shuqing et al. [7] provided novel algorithm to extract user's interest based on web log data and describes including long term interest and short term interest. Tsuyoshi et al. [8] described in his paper a method for clarifying user's interests based on an analysis of the site keyword graph. In this paper, we concern on how to provide smartphone user profile automatically using Modified Hierarchical Agglomerative Clustering Algorithm based on their historical logs of web browser applications in smartphone.

## 3 Data Extraction and User Profiling

In this section, we will describe about data extraction and each process to provide smartphone user profile for interest-focused based on browsing history of web browser applications.

### 3.1 Data Extraction

In this work, we use browsing history data of 30 smartphone users that is collected during one month. In this study, we develop an android application that can be used to collect browsing history from all browser applications and then we install that application on each user's smartphone. The structure of collected data from user's smartphone is shown in Table 1.

**Table 1.** The example of collected data

| User ID | Visit time | URL |
|---|---|---|
| 10 | 1399652396.55 | http://www.kakao.com/fightingkorea |
| 7 | 1399809440.79 | http://cyber.kepco.co.kr/ckepco/ |
| 1 | 1400251354.06 | http://asked.kr/ask.php?id=1927949 |
| 5 | 1399553574.34 | http://m.winixcorp.com/ |
| 5 | 1399637818.62 | http://www.dalkomm.com/ |

Based on the Table 1, each row of collected data represents the URLs that the user visits. Attributes of the data include user ID, visit time, and URL data. A URL (Uniform Resource Locator) is the unique address of documents and other resources on the World Wide Web. The first part of URL structure is called a protocol identifier which indicates what protocol that is used, and the second part is called a resource name which specifies the IP address or domain name where the resource is located. In our work, we extract collected data to derive a resource name part of URL structure which is useful information to analyze user interests and after that the Modified Hierarchical Agglomerative Clustering algorithm is assigned to provide smartphone user profile for interest-focused.

### 3.2 Modified Hierarchical Agglomerative Clustering Algorithm

In this study, we have Modified Hierarchical Agglomerative Clustering to aid in providing smartphone user profile for interest-focused. We use *Levenshtein* distance function to measure minimum distance between two extracted URL data. We use *Levenshtein* distance to measure minimum distance between two extracted URL data. *Levenshtein* distance between two extracted URL data $url_1$, $url_2$ is given by $dist_{url_1,url_2}(|url_1|, |url_2|)$, where,

$$dist_{url_1,url_2}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} dist_{url_1,url_2}(i-1,j) + 1 \\ dist_{url_1,url_2}(i,j-1) + 1 \\ dist_{url_1,url_2}(i-1,j-1) + 1_{(url_i \neq url_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Where, $1_{(url_i \neq url_j)}$ is the indicator function equal to 0 when $url_i = url_j$ and equal to 1 otherwise.

The Modified Hierarchical Agglomerative Clustering algorithm that is implemented in our work is following below.

**Algorithm.** Modified Hierarchical Agglomerative Clustering

Input:

$A\ set\ X\ of\ extracted\ URL\ data\ \{URL_1, \dots, URL_n\}$

$$URL\ Filtering\ Categories\ C_{4xn} = \begin{bmatrix} key_{1,1} & key_{1,2} & \cdots & key_{1,n} \\ key_{2,1} & key_{2,2} & \dots & key_{2,n} \\ key_{3,1} & key_{3,2} & \dots & key_{3,n} \\ key_{4,1} & key_{4,2} & \dots & key_{4,1} \end{bmatrix}$$

$A\ Lavenshtein\ distance\ function\ dist(c_1, c_2)$

***output***:

$Degree\ of\ Interest, user\ interests$

1: $C_{group_1} = []$
2: $C_{group_2} = []$
3: $C_{group_3} = []$
4: $C_{group_4} = []$
5: ***for*** $i = 1\ to\ n$
6: $c_i = \{URL_i\}$
7: $end\ for$
8: $C = \{c_1, \dots, k\}$
9: $l = n + 1$
10: ***While*** $C.size > 1\ do$
11: $(c_{min1}, c_{min2}) = min.distance\ (c_i, c_j)\ for\ all\ c_i, c_j\ in\ C$
12: $Remove\ c_{min1}\ and\ c_{min2}\ from\ C$
13: $Add\ \{c_{min1}, c_{min2}\}\ to\ C$
14: $If\ \{c_{min1}, c_{min2}\}\ any\ [key_{1,n}], where\ n = 1,2,..,n:$
15: $Add\ \{c_{min1}, c_{min2}\}\ to\ C_{group_1}$
16: $elif\ \{c_{min1}, c_{min2}\}\ any\ [key_{2,n}], where\ n = 1,2,..,n:$
17: $Add\ \{c_{min1}, c_{min2}\}\ to\ C_{group_2}$
18: $elif\ \{c_{min1}, c_{min2}\}\ any\ [key_{3,n}], where\ n = 1,2,..,n:$
19: $Add\ \{c_{min1}, c_{min2}\}\ to\ C_{group_3}$
20: $elif\ \{c_{min1}, c_{min2}\}\ any\ [key_{4,n}], where\ n = 1,2,..,n:$
21: $Add\ \{c_{min1}, c_{min2}\}\ to\ C_{group_4}$
22: $l = l + 1$
23: ***end while***
24: $Degree\ of\ Interest\ C_{group_1} = \frac{length.C_{group_1}}{\sum_{i=1}^{4} length.C_{group_i}}$
25: $Degree\ of\ Interest\ C_{group_2} = \frac{length.C_{group_2}}{\sum_{i=1}^{4} length.C_{group_i}}$
26: $Degree\ of\ Interest\ C_{group_3} = \frac{length.C_{group_3}}{\sum_{i=1}^{4} length.C_{group_i}}$
27: $Degree\ of\ Interest\ C_{group_4} = \frac{length.C_{group_4}}{\sum_{i=1}^{4} length.C_{group_i}}$
28: set $C_{group_i}$ which has max $(Degree\ of\ Interest\ )\ as\ user\ interests$

First is start by assigning each extracted URL data to a cluster, if we have $n$ URLs, it means we have $n$ clusters. Second, we compute the minimum distance between each cluster using Levenshtein distance function. Third, we find the closest (most similar) pair of clusters and merge them into a single cluster. Forth, we filter element of clusters using URL filtering categorizes. URL filtering categorizes will filter clusters based on keywords of users' interests. Fifth, we compute distances between the new cluster and each of the old clusters. We repeat steps 3, 4 and 5 until all extracted URL data has been clustered into a category of users' interests. After that we calculate interest degree for each category groups.

**Table 2.** URL filtering categories

| Category group | Category type |
|---|---|
| Business | Business/Economy, Job Search/Careers, real estate, and shopping |
| Communications and search | Blog/Web Communication, social networks, email, and search engines/portals |
| General | Computer/Internet, education, news/media, and reference |
| Lifestyle | Entertainment, games, arts, humor, religion, restaurants/food, and travel |

In our real work, we classify URL filtering categories into four main categories. The categories and their category type are shown on Table 2. Filtering categories consists of business category group, communications and search category group, general category group, and life style category group. The matrix $C$ of size $4 \times n$ to represent filtering categories can denoted as

$$C_{4xn} = \begin{bmatrix} key_{1,1} & key_{1,2} & \dots & key_{1,n} \\ key_{2,1} & key_{2,2} & \dots & key_{2,n} \\ key_{3,1} & key_{3,2} & \dots & key_{3,n} \\ key_{4,1} & key_{4,2} & \dots & key_{4,n} \end{bmatrix}$$

Where rows represents category group of users' interests and columns represent keywords on each category group. We categorize the clustered results into category group based on keywords on matrix of filtering categories.

## 4 Experimental Results

In our study, we collected browsing history data of 30 smartphone users during one month continuously. Browsing history data was tested on log files stored by the server. We extracted collected data and then use proposed algorithm which is called by Modified Hierarchical Agglomerative Clustering to provide smartphone user profile for interests-focused. In our experiment, we compare the performance between our method and C4.5 algorithm.

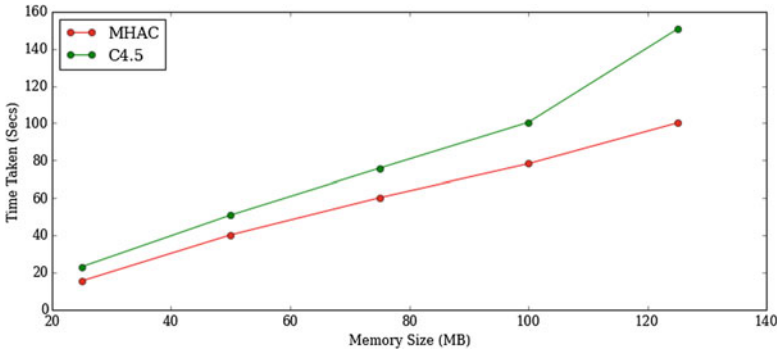


**Fig. 1.** Execution time comparison with C4.5 algorithm

Figure 1 presents the execution time results of Modified Hierarchical Agglomerative Clustering algorithm (MHAC) and C4.5 algorithm. Our method consistently outperforms the C4.5 algorithm on all memory utilization in execution time. The results of degree of smartphone users' interests for each category are shown on Fig. 2. Finally, the results of smartphone user profile for interests-focused based on the highest degree for each user is shown on Fig. 3.

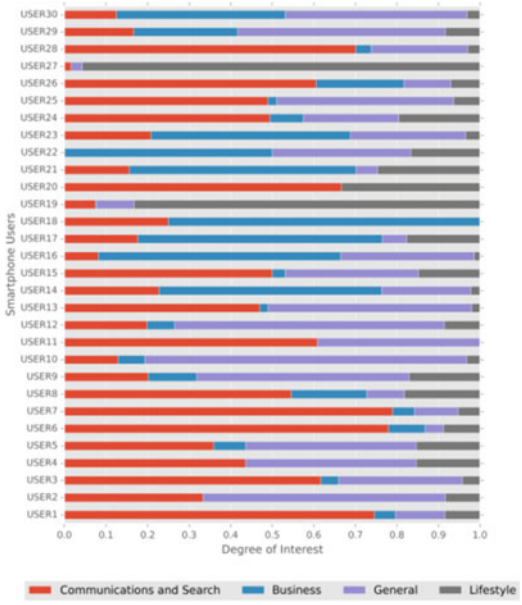


**Fig. 2.** Degree of smartphone user interests

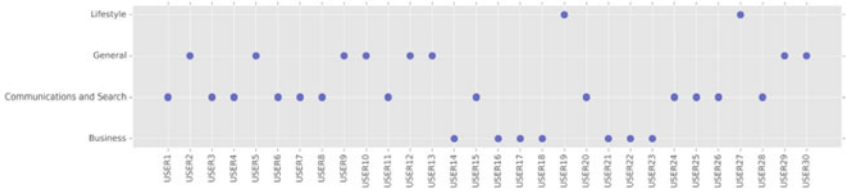


**Fig. 3.** Results of smartphone user profile for interests-focused

## 5 Conclusion and Future Work

In this paper, we have implemented a server-based application that can be used to provide user profile for interests-focused based on browsing history of web browser applications. In our approach, we propose a Modified Hierarchical Agglomerative Clustering to cluster extracted data which can automatically provide an interests profile of smartphone users. Based on experimental results, the proposed method can measure degree of users' interests based on browsing history of web browser applications, inferring particular pieces of information that they interested on it, and outperforms the C4.5 algorithm in execution time on all memory utilization. Because amount of data that will be processed is increased, so in the future we need to implement Map-Reduce algorithm on Modified Hierarchical Agglomerative Clustering to enhance performance of clustering algorithm.

**Acknowledgements.** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2012R1A1A2007014).

## References

1. Hu, J., Zeng, H.-J., Li, H., Niu, C., Chen, Z.: Demographic prediction based on user's browsing behavior. In: International World Wide Web Conference, Beijing (2007)
2. Santra, A.K., Jayasudha, S.: Classification of web log data to identify interested users using Naïve Bayesian classification. IJCSI Int. J. Comput. Sci. Issues **IX**(1), 381–387 (2012)
3. Xu, J.H., Liu, H.: Web user clustering analysis based on KMans algorithm. In: International Conference on Infrmation, Networking and Automation (ICINA), HangZhou (2010)
4. Min-jie, X., Jin-ge, Z.: Research on personalized recommendation system for e-commerce based on web log mining and user browsing behaviors. In: International Conference on Computer Application and System Modeling, ZhengZhou (2010)
5. Anand, N.: Effective prediction of kid's behaviour based on internet use. Int. J. Inf. Comput. Technol. **IV**(2), 183–188 (2014)
6. Li, J.: Research of analysis of user behavior based on web log. In: International Conference on Computational and Information Sciences, Anshan (2013)
7. Wang, S., She, L., Liu, Z., Fu, Y.: Algorithm research on user interests extracting via web log data. In: International Conference on Web Information Systems and Mining, Chengdu (2009)

8. Murata, T., Saito, K.: Extracting users' interests from web log data. In: International Conference on Web Intelligence, Tokyo (2006)
9. McKinney, W.: Python for Data Analysis. O'Reilly Media Inc, Sebastopol (2012)
10. Russell, M.A.: Mining the Social Web. O'Reilly Media Inc, Sebastopol (2011)
11. Rossant, C.: Learning IPython for Interactive Computing and Data Visualization. Packt Publishing Ltd, Birmingham (2013)
12. Vaingast, S.: Beginning Python Visualization. Springer, New York (2009)