# On Understanding User Interests through Heterogeneous Data Sources

Samamon Khemmarat[1], Sabyasachi Saha[2], Han Hee Song[2], Mario Baldi[2], and Lixin Gao[1]

[1] University of Massachusetts Amherst, MA, USA
[2] Narus Inc., CA, USA

**Abstract.** User interests can be learned from multiple sources, each of them presenting only partial facets. We propose an approach to merge user information from disparate data sources to enable a more complete, enriched view of user interests. Using our approach, we show that merging different sources results in three times of more interest categories in user profiles than with each single source and that merged profiles can capture much more common interests among a group of users, which is key to group profiling.

## 1 Introduction

User interest profiles allow businesses and service providers to customize their services and products to better suit users' needs and likings. User "footprints" left in cyberspace, spread across different services, contain a large amount of information about them. While many research works focused on joining user data across various services of the same type (*e.g.*, online social networks) [2], aggregating users' interests at various social networks or websites can only capture a very specialized, partial view of the user, the persona that user wants the world to see. A more comprehensive user profile can be captured by combining user information from different types of services. However, it is not trivial to do so because of each service having its own representation of user data. In the last few years, Internet users are increasingly interactive and form groups with shared interests (*e.g.*, meetup.com, Google Hangouts, etc.). Understanding the common interests of groups of users allows services to be tailored to groups [3]. However, such group profiling requires finding commonality in information from different users, which needs to be done at a semantic level.

The goal of this research work is to represent user interests as they can be learned from different data sources in a single format that can be easily explained, compared, and combined. We propose a generalized method that flexibly joins user interests from heterogeneous sources of data. Using the proposed approach, we create user profiles from two representative data sets, online social network (OSN) profiles and web browsing traces collected from a Cellular Service Provider (CSP) and combine them. We show that our approach (i) can create a richer user profile from heterogeneous information sources, and (ii) can create more effective group profile by finding more common interests among users, compared to using a single information source.

## 2 Reconstructing User Interest

We construct a profile $P_{ur}$ of a user, $u$, analyzing raw data from a single information source, $r$. To allow comparing and merging interests across different sources and users (or to group interests of users), interests from each source are mapped on a category hierarchy $\mathcal{H}$. Then we create the unified user profile $\mathcal{P}_u$ combining the interest categories in all $P_{ur}$. In particular, the process includes the following key steps.

**1. Interest Item Extraction.** We define an *interest item* as a unit of data that provides information about coherent topics of interests, e.g., a URL requested by a user in browsing activity logs. We built specific parsers and noise filters, for each data source, to extract a set of interest items $\mathcal{I}_u$ for user $u$.

**2. Interest Item Enhancement.** In this step we create a vector , $\boldsymbol{V_k}$, of terms, $t_{kj}$, that enrich the semantics of each interest item, $i_k \in \mathcal{I}_u$, using additional resources and processes, e.g., using synonyms of words or metadata of URLs. $\boldsymbol{V_k}$, is used to aid interest item categorization (next step).

**3. Interest-to-Category Mapping.** Each interest item, $i_k$, is mapped into an interest category hierarchy $\mathcal{H}$. Using Machine Learning techniques, we categorize $i_k$ to one or few interest categories $\{h_s\}(\in \mathcal{H})$.

**4. User Profile Creation.** A user's ($u$) interest profile, $P_{ur}$, can be created by aggregating all of his interest categories, represented as a single vector of interest categories along with the frequencies $\{(h_s, f_s)\}$ with which interest items map on them. We, then, create the unified user profile $\mathcal{P}_u$ combining all $P_{ur}$ of the user.

## 3 Experimental results

Our dataset contains data of 15,428 users. The association between the browsing traces and OSN's ID of a user was done with the Mosaic system [4].The browsing traces, T1 and T2 are 5-day long and were collected from a backbone router of a major CSP in North America. The categories from the ODP directory [1] are used as reference interest categories, to which the extracted interest items from different data sources are mapped to.

For each individual user, we study interest items overlap between profiles. Figure 1 plots quantities of interest categories that overlap between the profiles from the two sources. We contrast this result with Figure 2, which plots the same quantities for two browsing profiles created from two periods of time, T1 and T2. The smaller overlap in Figure 1 suggests that a richer profile can be created by combining data from disparate sources. The average number of interest categories per user increases by up to 3 times when the profiles are combined.

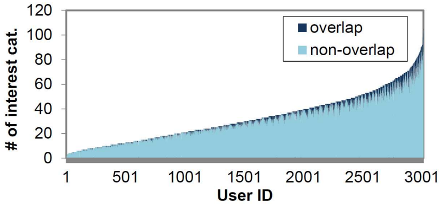

**Fig. 1:** Overlaps between browsing and OSN profiles

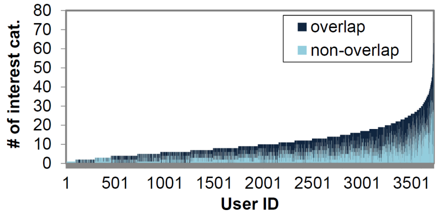

**Fig. 2:** Overlaps between the browsing profiles from T1 and T2

| Profile type | | 10 user group | | 50 user group | |
|---|---|---|---|---|---|
| | | grp int | top-1 cov. | grp int | top-1 cov. |
| Uncat. | Browse | 0 | 0.27 | 1 | 0.46 |
| | OSN | 0 | 0.01 | 0 | 0.00 |
| Cat. | Browse | 4 | 0.86 | 5 | 0.80 |
| | OSN | 6 | 0.92 | 7 | 0.93 |
| Cat.& Merged | | 13 | 0.97 | 14 | 0.96 |

**Table 1:** Effectiveness of group profiling

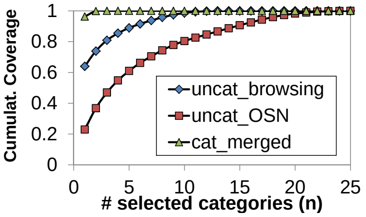

**Fig. 3:** Group coverage

**Group Profile.** Now, we show the effectiveness of merging profiles when we want to discover interests commonly shared among a group of users, *e.g.*, gathered in a coffee shop. The effectiveness is measured as (i) the number of *group interests*, interests shared by more than 50% of users, and (ii) the fraction of users in the group that have the most popular interest, referred to as *top-1 coverage*. The comparison is performed between three types of profiles, original OSN and browsing profiles with no categorization, categorized OSN and browsing profiles, and merged categorized profiles. We generate 50 groups of 10 and 50 randomly selected users from our dataset. Table 1 shows that using the categorized & merged profiles results in the highest number of group interests as well as the best top-1 coverage. Furthermore, we define *coverage* for a set of categories to be the proportion of users for whom at least one of his interests can be found in the set. In Figure 3 evaluating the number of interest categories required to satisfy users in the group, we observe that the categorized & merged profiles require only two interest categories to satisfy all members, whereas the uncategorized profiles require 25 categories to be picked to cover interests of all members.

With our results, we illustrated that combining interests from multiple sources leads to increased availability of user data and higher utility in profiling a group of users.

## References

1. Open directory project. `http://www.dmoz.org`.
2. A. Malhotra, L. C. Totti, W. M. Jr., P. Kumaraguru, and V. Almeida. Studying user footprints in different online social networks. *CoRR*, abs/1301.6870, 2013.
3. L. Tang, X. Wang, and H. Liu. Group profiling for understanding social structures. *ACM Transactions on Intelligent Systems and Technology*, 3(1):15, 2011.
4. N. Xia, H. H. Song, Y. Liao, M. Iliofotou, A. Nucci, Z.-L. Zhang, and A. Kuzmanovic. Mosaic: Quantifying privacy leakage in mobile networks. In *ACM SIGCOMM*, 2013.