Identify User Variants Based on User Behavior on Social Media

Haoran Xu and Yuqing Sun*
School of Computer Science and Technology
Engineering Research Center of Digital Media Technology, Ministry of Education of PRC
Shandong University, Jinan, China
Email: hr_xu1990@163.com and sun_yuqing@sdu.edu.cn

***Abstract*—In social media, users are allowed to express their opinions by commenting on an item or rating an item with scores. The collection of user reviews would generate a positive or negative influence to the media audience. Some malicious users may create multiple variant accounts on the same social media so as to influence or manipulate public opinions for business or criminal purposes. To maintain good social environment, it is necessary to find those fake users. In this paper, we investigate the user variants identification problem using both user behavior and item related information. We study the characteristics of user behaviors on social media and introduce two concepts $visibility$ and $distingushibility$ to preliminarily quantify whether a fake user can be identified. To better understand user intention and characteristics, we profile a user with apparent and implicit features, which are extracted from three aspects: User Generated Contents (UGC), user behavior context and item information. Based on these features, we propose the user Variants Identification Problem (VIP) and an identification algorithm, which finds the top-k similar variants in a social media. We evaluate our methods against two real datasets $MovieLens$ and $Amazon$ and make comparison on the effectiveness against different features in identifying user variants.**

***Index Terms*—user variants identification, interaction behavior, social media**

I. Introduction

Social media is now widely integrated into our daily life. Users are allowed to register on a media website with anonyms and share their ideas by comments or giving a thumb up on an item or other person's reviews. For example, on a video/music sharing websites like Youtube or Youku, users often write some reviews on a song or a film and rate it with scores. It is similar with the news websites, such as Yahoo News or Sina News, the shopping websites like Amazon, the community websites like Movielens and etc. The collection of user comments would bring positive or negative influence to the media audience. For example, a person may check the comments about a suit of clothes on Amazon website before buying it. If the comments are negative, he/she may not buy it. So, for some business or criminal purposes, malicious users may create multiple variant accounts on the same social media so as to influence or manipulate public opinions. For example, a vendor may hire some users to create a group of fake accounts, and then let them together boast their goods using different anonyms so as to defraud consumers. Another example is that, when a new movie is released, the publishers may organize a group of people boast their film so as to attract more audiences for a higher box office return. More seriously, some criminals may create multiple accounts to spread a rumor or to preach some fraud information. These deceptive behaviors do harm the interests of users on website. Hence, it is important and necessary to identity these variant accounts in social media.

Fake user identification is very related to the user mapping problem between two different social networks, which has been well investigated. They model a user based on user relationship[14], [5], [13], user attributes[4], [3], [2] and user generated contents(UGCs) [6], [15], [9] in social media. Then they compute the distance between users and find the most similar users to a target user. However, in many social network platform, user profile, attributes and user relationships are not available under privacy settings. Some users may leave attributes empty or fill in with misinformation. These methods can not be applied to such social media. Our work is also related to the user identification problem, which try to match an anonyminized user to an individual in real life [10], [11], [12]. The basic requirement for such methods is that an adversary need to have some background knowledge about the person in advance, such like some purchase history on amazon, the list of rating films on Netflex, although sometime it only requires a small amount of information. Unfortunately, such requirements can not be always satisfied.

Compared to the existing works, this work makes three contributions. The first is that we study a different problem, the variant identification problem (VIP), which finds the variants for an appointed user on the same social media website. We need not have any background knowledge about the target user in advance. The basic philosophy behind such identification is that user behaviors on items are intentional interaction and there must exist many hints of the similarity between two variants, such as the frequently used words, the time stamps of rating, the sort of reviewed items etc. To achieve their business or criminal purposes, the variants of the same user should

* corresponding author: sun_yuqing@sdu.edu.cn. This work is supported by the National Natural Science Foundation of China (61173140), Special Program on Independent Innovation & Achievements Transformation of Shandong Province (2014ZZCX03301) and Science & Technology Development Program of Shandong Province (2014GGX101046).

978-1-4673-8590-9/15/$31.00 ©2015 IEEE

have the same or similar attitude on the same item. In case a user intentionally performs differently using variants, this user could not generate large collective influence on the same item to the audience and it is not necessary to recognize him/her.

The second is that we use both user behaviors and the items information that a user ever reviewed as assistant information for identification. To the best of our knowledge, it is the first time to adopt the item information in recognizing a user. We perform a comprehensive study about the characteristics of user behaviors on social media, such as visits, comments, ratings etc. on item, and introduce two concepts $visibility$ and $distinguishability$ as the basic quantification on whether a fake user has perceptible malicious behaviors and can be identified. The users with few behavior could not have much influence on websites and are neglected. We also analyze different aspects of a user reviewed items, especially the collection of all users' reviews. Such information can help us to understand a fake user intention and the difference with others.

The third contribution is to propose several user models. To better understand user behaviors, we extract both explicit or implicit features about a user. With the advantage of the common knowledge, such as the ontology, we propose a series of models to profile a user, which are abstracted from the four aspects: User Generated Contents(UGCs), behavior context, item information and implicit characteristics. Finally, we perform a thorough experimental analysis on two real database $MovieLens$ and $Amazon$ to evaluate our models and study model combination. Some experiments also evaluate the influence of user behavior number on variants identification and the efficiency of algorithm.

The rest of the paper is organized as follows. In Section II, we review existing works related to our research. Section III gives the formal definition of some concepts and the user variant identification problem. In Section IV, we propose several user modeling methods based on the user behaviors and item information. Then we present the top-k variant identification algorithm in Section V and experimental evaluation in Section VI. Finally, we conclude this paper and discuss future work.

II. Related Works

User Mapping across Social Networks. The most related work is the user mapping problem between two different social networks. The main idea of solving this problem is to model a user based on user relationship, user attributes and user generated contents(UGCs) in social media. Then they compute the distance between users and find the similar users to a target user. Long et al. [5] [13] utilize graph topologies to model a user and make comparison between two candidates in different networks. They rank candidate users with mapping possibilities so as to improve matching performance. The shortcoming of such methods is the high complexity with the size of network such that they are not suitable for large-scaled networks. Some works take advantage of user attributes to profile a user. Vosecky et al. [14] represent a user profile as a vector, consisting of individual profile fields. A user in one network can be recognized in another network if their similarity score reaches a certain threshold. Chung et al. [2] consider not only a user individual profile but also his/her friend profiles so as to boost the mapping accuracy. Cortis et al. [4] study the semantic relations between profile attributes(e.g. city vs. country). Liu et al. [6] also consider usernames as a reference in recognizing users in different networks. However, the profile attributes of users are not always available in practice.

Recently, many works focus on the user generated contents(UGCs) in solving the user mapping problem, such as tags, images, messages, etc. Correa et al. [3] [15] find behavioral patterns to determine if two users in different works belong to the same individual. Meo et al. [9] model users as the tag based profile or ontology-based profile and then identify a user according to their semantic distance. Liu et al. [8], [7] propose a heterogeneous behavior modeling method to analyze topical distribution, temporal behavior and behavior consistency across different platforms. Different from these work, we adopt user behavior and item information, which are always available on the social media website, like user comments or reviews. Besides, we extract both apparent and implicit features from these data to profile a user.

User Identification/De-anonymization Our work is also related to the user identification problem, which matches an anonyminized user to an individual in real life. Narayanan et al. [10] propose a de-anonymization algorithm and compute similar scores for each record as the matching candidates. They assume the adversary have an amount of background knowledge about an individual in advance for identification. Narayanan et al. [11] develop a re-identification algorithm targeting an anonymized social network. They assume that an attacker has some individual auxiliary information, such as some k-size node cliques on both the auxiliary and the target graphs. Then the de-anonymization is performed based on the social network topology. Payer et al.[12] try to identity authors of scientific publications based on the additional features derived from writing style and contents of the paper. From the aspect of assistant information for identification, these works are related to our work. But, the problem addressed in this paper is quite different from theirs. We do not need any background knowledge about individuals. Furthermore, we utilize the collections of user behaviors on items to enrich the information of both items and users. We also introduce the information of items to help us map the variants.

III. User variant Identification Problem

A. Dataset and Problem Definition

The dataset we considered in this paper is a set of user behaviors on a social media, as well as the media contents in details. For example, on the movie review website $MovieLens$, user behaviors include user ratings and reviews on movies. The media contents refer to the information about movies, such as movie genres, released year, directors, actors, movie video and etc. Another example is user comments on news website, like BBC. User behaviors refer to the

comments on each piece of news, as well as the behavior context, such as the time and IP address, where users submit the comments. News contents include title, keywords, author, time, and text etc.

A user in social media refers to a person in real life. Let U denote the user set on a social website and its size is $|U|$. An item refers to an object in a social media, such as a film on $Netflex$ or a news piece on BBC. Let V denote the item set in a system. Items are associated with multiple attributes. Each item has different attribute value and content. The set of items that user u ever reviewed in a system, is denoted as V_u.

Definition 1: **(User Behavior)**. Given a user $u \in U$ and an item $v \in V$, a user behavior refers to $u's$ once review behavior on v and is represented as a link $< (u, v), UGC, Cxt >$ between u and v, where UGC (User-Generated Content) refers to any form of contents created by u against v, Cxt means the context of behavior, such as the timestamp or IP address on this review. A typical UGC is the user's rating or comments on a movie. The set of user behaviors in the whole system is denoted as B.

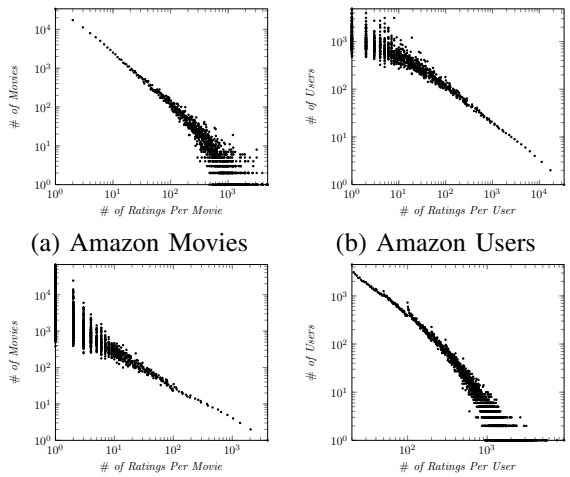
Definition 2: **(Item Feature)**. An item feature refers to a characteristic of an object, which is abstracted from the attributes or content of item. It also includes the collection of all users' generated contents on the item, such as the collective tag set of a web page on a collaborative tagging system, or the average rating value of a film on a video website.

Definition 3: **(Corpus)**. A corpus is the collection of user set U, an item set V and a user behavior set B on a system, denoted as $\Gamma = (U, V, B)$.

From user review behavior on an item, we can have a further understanding on both items and users beyond the item information or user attributes themselves, which are called interaction effects. On one side, for some item, the collection of user comments or ratings can be used to analyze different aspects of the item. For example, in a collaborative tagging system, the tags on an item are regarded as the abstraction of the item content by different users. On another side, the collection of a user interaction behaviors on different items can help understand user characteristics, preferences, hobbies, etc. Although a user may register with different pseudonyms on the same social media, there must exist some hints of the similarities between these variants, such as the frequently used words and phrases, the attitude or core value on different things. This motivates us to identify user variants in the same social media by analyzing user behaviors.

User Variants Identification Problem(VIP): Given a corpus $\Gamma = (U, V, B)$ and a target user $\hat{u} \in U$, the user variants identification problem is to identify the variants of $\hat{u}$ from Γ.

Our goal is to identify user variants in the same social system. Different from previous works, we do not require any background about a user attributes, such as gender, age, address, etc. Furthermore, we profile a user on both explicit and implicit characteristics, which are abstracted from user behavior and items information rather than the traditional works that only adopt user generated contents. The consideration is based on the fact that user behaviors on items



(a) Amazon Movies (b) Amazon Users

(c) MovieLens Movies (d) MovieLens Users

Fig. 1. Rating Distribution

are intentional interaction. Users choose their interested or target items and make reviews. For instance, users may choose some interested products to give ratings on EC website for commercial purposes.

B. Visibility and Distinguishability

The statistics of user behaviors on website often follow a long tail phenomena[1]. The dataset we adopted in this paper has the same characteristic. Fig. 1 shows the statistics on two real data sets: $MovieLens$ and $Amazon$. $MovieLens$ is the movie data set published by the web-based recommender system MovieLens. It contains 10M ratings scale from 1 to 5 stars applied to 10,681 movies by 71,567 users and 95,580 tags made by 4,009 users on 7,601 movies. A user in this dataset may appear either in rating records or in tagging records, and some times a user may appear in both. $Amazon$ is another public dataset released by the retailer website Amazon. We choose its subset including the reviews of movies on $Amazon$. The data set contains 3.9M ratings and comments of 857,793 users on 149,852 movies. In this dataset, each review includes a rating and a comment on a movie.

We make a statistic on the number of behaviors for each user in these datasets. Fig. 1 (a) and (b) show the results on $MovieLens$ dataset, where the $x - axis$ is the number of movies that a user ever interacted (rating or comment) and the $y - axis$ gives the counting on users. Figure 1 (c) and (d) are the results on $Amazon$. From this statistics, we can see that only a few number of users has a large number of behaviors, while a large number of users have a few comments on movies. Since we do not have any background about a user attributes or characteristics in advance, a user is expected to be recognized only by his/her behaviors on a web site. It is easy to understand that users with few behavior are hard to be recognized. So the existence of unrecognizable users is an unavioded fact in solving the VIP problem. But considering the purpose of multiple user variants, such kind of users could not have much influence on websites.

Another problem in solving the VIP problem is the similarity between users. Two users must seem the same if they have exactly same behaviors. For example, in $MovieLens$, two users with exactly consistent ratings on same movies can not be distinguished. Hence, we introduce two concepts $visibility$ and $distingushability$ to illustrate where a target user and a dataset for analysis.

Definition 4: **(visibility)**. Given a corpus $\Gamma = (U, V, B)$, Γ is δ-visible if $\forall u \in U$, $|B_u| \geq \delta$.

Definition 5: **(distinguishability)**. Given a corpus $\Gamma = (U, V, B)$, two constant parameters $\epsilon, \theta \in [0..1]$, Γ is (ϵ, θ)-distinguishable if

$$Pr[Sim(u, u') > \epsilon \forall u \neq u' \in U] \leq \theta \tag{1}$$

, where $Sim(u, u')$ refers to the similarity function between u and u' based on B and V.

The $visibility$ checks a minimum threshold for collecting a user's behaviors on a system as the basic materials to analyze a user. $Distinguishability$ requires that there only are at most θ percent of users who are similar with each other in the dataset.

A special case is that if a user intentionally behaves in different manners, it does make our analysis difficult. Even in such case, the variants resolution problem is still meaningful. One consideration is that in practice, the purpose for a user (called malicious) to make multiple variants is to make high influence the audiaunce on social media, either positive or negative. So, a malicious user need to play in the same trend or as the same role. Otherwise, the effectiveness of different variants would be mutually exclusive. Another consideration is that there may be multiple users whose behaviors seem similar with a target user. Since the variant identification is an unsupervised process, it is more reasonable to recognize a small set of candidates rather than one exact user.

In the next section, we will discuss how to profile a user, based on which we propose the $top-K$ variants identification algorithm to solve the VIP problem in V.

IV. User Modeling

User profile modeling is a basis to analyze a user. We model a user with the help of the available information on a social media and some common knowledge. The former includes user generated content, user behavior context and the item information. For the later, we introduce the knowledge ontologies for the semantic analysis and a knowledge base for additional assistant information to items. To better understand user behaviors, we profile a user as both explicit and implicit features, which are extracted from four aspects: UGC, behavior context, item information and knowledge based description.

A. User Generated Content Modeling (UGC-based Model)

User generated content means any form of content created by users. It directly relates to a user subjective will. Currently, on popular social media there are two typical types of user review: textual contents such as comments or tags, and user attitudes such as rating score, like/dislike. Based on these available UGCs, we propose several models to profile a user.

Comment-based Model (U^c). Comments are the most popular way for users to express their opinions or ideas on an item. For each user $u \in U$, we create the bag of words and the distribution of the occurrences on words that u ever used in his/her comments.

Tag-based Model (U^t). Tags are another frequently used method for users to express their understanding about an item in social media. We collect all tags created or used by a user and model this user as the distribution of frequency on tags.

Attitude-based Model (U^a). User attitude, such as rating score, likes etc., as another subjective part of review, is also important in modeling a user. We propose an attitude-based model and profile a user as the relative distribution on scores. Let V_u be the collection of items that user u ever reviewed. $\forall v_i \in V_u$, user u's attitude(i.e. rating score) on v_i is denoted as $s_i(u)$. Suppose there are t values for scores, denoted as $\Sigma = \{\sigma_1, ...\sigma_t\}$. For a specific rating score $\sigma_j \in \Sigma$, the relative frequency $f(\sigma_j)$ is formalized as follow:

$$f(\sigma_j) = \frac{\sum_{v_i \in V_u} 1(s_i(u) = \sigma_j)}{|V_u|} \tag{2}$$

, where $1(\cdot)$ is an indicator. When the equation in the quote is satisfied, $1(\cdot)$ is 1. Otherwise, $1(\cdot)$ is 0.

B. Behavior Context Modeling (Cxt-based User Model, C^h)

Besides UGCs, the context of user behaviors is also an important user characteristic. For example, a user may always review movies at a certain time after work. Although this information is not subjective will of users, there exist some regular pattern in user behaviors according to the study of social behaviorsim. In our dataset, there is a timestamp on each behavior record. So we mainly study the temporal user patterns. To better understand user temporal characteristics, we divide user behaviors into 24 subsets according to the hour of timestamp and count user behaviors within each interval. Each user is modeled as the frequency distribution over 24 hours. The context based user model is denoted as C^h. In fact, this kind of user model is very effective in solving the VIP problem, which would be discussed in Section VI.

C. Item-based User Modeling

When users review on social media, they often consider item content and choose their interested items. Hence, item information can reflect user preference in some extent. We try to utilize this information to model a user. Generally, item information include two categories: an item itself attributes and the collective data generated by users on an item. We would model them in different ways.

Item Attribute-based Model. In social media, items attributes are provided. For example, a movie on $Movielens$ is associated with the title, genres, the released year, actors and etc. We model an item as the collection of its attribute values. For example, regarding the genre of a movie, we represent it as a genre vector, where each dimension is a certain genre class. To model a user $u \in U$, we can collect all the items

that u ever reviewed and represent u as a relative distribution on these attribute values.

Collection of User Interactions-based Model. The collection of user reviews on a certain item reflect the public understanding about the item such that they not only enrich the connotation of item but also reveal the behavior characteristics of related users. This information can help us understand user characteristics, preferences, hobbies, etc. To profile users by this information, we classify these data into two types: textual content and user attitude (i.e. rating score).

(1) *Collective Item Textual Content Model*(V^w/V^t). The collective item textual content refer to the collection of all user comments and tags on an item. For a given item v, it is represented as the distribution ϕ of words in the collective textual content on v. Based on this item representation, a user u is modeled as the distribution of the collective ϕ on items that u ever reviewed. The model reflects a user's interests on items. For clarity purpose, the collective comments based user model is denoted as V^w and the collective tags based user model is denoted as V^t, respectively.

(2) *Common Attitude-based User Model* (V^s). The common attitude on an item $v_i \in V$ is defined as the average score of all user rating scores, denoted by $\bar{s_i}$. Then user characteristic can be attained from the differences between a user attitude and the common attitude on each item. For a given user u, all the items that u ever reviewed is denoted as V_u. For $v_i \in V_u$, $s_i(u)$ is the rating score on v_i by u. Based on the common attitude, items in V_u are partitioned into t subsets against the score range $\Sigma = \{\sigma_1, ...\sigma_t\}$, as defined in the above. Each $V_u^\sigma \subset V_u$ represents the subset satisfying $\bar{s_i} = \sigma, \forall v_i \in V_u^\sigma$.

For each V_u^σ, we compute the distribution of user rating score $s_i(u)$ on $v_i \in V_u^\sigma$ over the score range Σ. Formally, for each V_u^σ, the frequency that u's rating score is σ' is computed as Equation 3. User u is then profiled as the collective distribution over V_u, say $\{f(\sigma', \sigma)|\sigma', \sigma \in \Sigma\}$.

$$f(\sigma', \sigma) = \frac{\sum_{v_i \in V_u^\sigma} 1(s_i(u) = \sigma')}{|V_u^\sigma|} \tag{3}$$

D. Implicit Characteristic Model (Semantic User Modeling)

Besides the analysis on user explicit features, we also investigate user implicit characteristics by semantic methods. We introduce a semantic tree extracted from $WordNet$, which is a lexical database for English. It groups English words into sets of synonyms called synsets, denoted as Syn. And we build a semantic tree Υ based on the relationships between synsets. Each synset contains one or more words and maps to a node in Υ. For a given level l, the set of nodes on l in Υ is denoted as $V(l)$. For a given node τ, its level in Υ is denoted as l_τ. To have a different grained analysis, a level l is allowed to specify in advance. A larger level means a better grained analysis.

To analyze user semantic characteristic, we first collect the text content, like comments and tags that a user ever used. Then we split these sentences or tags into words and regard them as user $u's$ vocabulary, denoted by Ψ_u. For a given level l, a user is modeled as the relative distribution over synsets τ on level l of Υ, which are calculated as follows. For each word $w \in \Psi_u$, we search a corresponding synset $\tau \in \Upsilon$ satisfying one of the three cases depending on a given level l: $w \in \tau \cap l_\tau = l$, τ is the ancestor of a node $\tau' \in \Upsilon, w \in \tau' \cap l_{\tau'} > l$, or a set of τ who are the successors of $\tau' \in \Upsilon, w \in \tau' \cap l_{\tau'} < l$.

In practice, user vocabulary may contain unstandard English words such that not every word can be mapped to a node in the semantic tree. To avoid of information loss, we adopt the *Comment Based Model* U^c to represent these words as the complement of semantic tree. This kind of user model is called *Semantic Fusion Model*, denoted as S^f.

V. Finding the Top-k Variants

In practice, since we do not have any background knowledge about a target user, it is reasonable to identify a set of the most similar users rather than to find an exact user as the variant. So we propose a top-k algorithm to solve the VIP problem, which will find the top-k similar users for a target user. In the following, we first discuss the similarity metrics to compare two user profiles and then present the algorithm.

A. Similarity Metrics

In the proposed methods, a user profile is modeled as a probability distribution function (PDF) over the selected features. There are many methods to compute the similarity between two PDFs, such as Euclidean distance, Manhattan distance and cosine similarity. In this paper, we adopt cosine similarity and Euclidean distance as the similarity metrics due to their low complexity and suitability for PDF vectors. The cosine similarity between two user profiles is denoted as $Sim(p(u_1), p(u_2))$. The similarity based on the Euclidean distance between them is denoted as $Dis(p(u_1), p(u_2))$. Suppose the dimension of user profile is m, the two similarity functions are defined as equation 4. For convenience, we adopt Ω as the set of the selected similarity metrics.

$$\begin{aligned} Sim(p(u_1), p(u_2)) &= \frac{\sum_{i=1}^m p_i(u_1) p_i(u_2)}{\sqrt{\sum_{i=1}^m p_i(u_1)^2} \sqrt{\sum_{i=1}^m p_i(u_2)^2}} \\ Dis(p(u_1), p(u_2)) &= \frac{1}{1 + \sqrt{\sum_{i=1}^m \left(p_i(u_1) - p_i(u_2)\right)^2}} \end{aligned} \tag{4}$$

B. Top-k Variants Identification

Definition 6: (Top-k Variants Identification) Given a corpus $\Gamma = (U, V, B)$, a similarity metric set Ω and a target user $\hat{u} \in U$, u is called k-identified if $u \in U_k$, where u is one of the real variants of target user $\hat{u}$ and U_k is the top-k similar users with u against Ω.

Algorithm 1 shows the top-k identification algorithm, where the inputs include a similarity metric ω, a corpus Γ, the size of likely list k, the target user $\hat{u}$ and an appointed variant u of $\hat{u}$. In Algorithm 1, line 2 is employed to model target user $\hat{u}$ as $p(\hat{u})$. Lines 3-6 are employed to calculate the similarities between candidate users' profile and $\hat{u}$'s, and add users into a priority list L which is sorted by his/her similarity with target user $\hat{u}$. We only remain top-k users in L and get the top-k

Algorithm 1 top-k variant identification algorithm

Require: a similarity metric $\omega \in \Omega$, a corpus $\Gamma = (U, V, B)$, where U is user set, V is item set and B is user behavior set, k, target user $\hat{u}$, and an appointed variant u of $\hat{u}$

1: $U_k \leftarrow \emptyset$, priority list $L \leftarrow \emptyset$
2: Model target user $\hat{u}$ as $p(\hat{u})$ according to his/her behavior $B_{\hat{u}}$
3: **for** each user $u_i \in U$ **do**
4: Model user u_i as $p(u_i)$
5: L.insert$(u_i, \omega(p(u_i), p(\hat{u})))$
6: **end for**
7: $U_k \leftarrow \{u_i | u_i \in L(k)\}$
8: **if** $u \in U_k$ **then**
9: $\hat{u}$ is k-identified
10: **else**
11: $\hat{u}$ is no-identified
12: **end if**

candidate user subset in line 7. Lines 8-12 are used to check if the top-k likely users contain the truly variant u. If so, $\hat{u}$ is k-identified. Especially, if u is the top-1 in the likely list, we say that $\hat{u}$ is exact matched.

VI. Experiments

A. Data Preprocess

Our experiments are conducted on two real data sets: $MovieLens$ and $Amazon$ and the details are given in section III. Before we perform the experiments, we will execute some data preparation and preprocessing.

Firstly, we utilize movie titles in the dataset and obtain the detailed information for each movie by the outside information OMDb API[1]. Due to the limitation of OMDb database and the misspellings on some movie titles, we capture the additional information of only a part of movie set. For MovieLens dataset, we obtain detailed information of 7,180 movies out of 10,681 movies in rating records, as well as 5,175 movies out of 7,601 movies in tag records. For Amazon dataset, we acquire detailed information of 22,846 movies out of 149,852 movies. To guarantee an unbiased experimental results, we remain the review records whose movie information has been obtained. After this filtering, we get several datasets with detailed movie information. The basic statistics of these two datasets are shown in Table I.

TABLE I
BASIC STATISTICS OF THE TWO DATASETS

	Amazon	MovieLens Ratings	MovieLens Tags
# of users	532,212	69,878	3,415
# of movies	22,846	7,180	5,175
# of ratings	1,154,213	7,258,169	$\emptyset$
# of comments		$\emptyset$	66,832

Secondly, the users who do not satisfy the properties of $visibility$ and $distinguishability$ are removed. A target user is randomly chosen from the filtered dataset. To verify the effectiveness of variant user identification, we partition the behavior set of a selected user into two parts with equal size so as to simulate the variant behaviors. One part is chosen as the target user and the other as the variant. Additionally, we adopt the $k\text{-}fold$ test for this random partition. We repeat this random partition for 20 times and compute their average as the experimental result.

At last, we perform the variant identification according to Algorithm 1. We choose n users to repeat the process of variant user identification and adopt the average result to evaluate the experiments.

In this paper, the items in the two dataset are both movies and the item attributes can be listed as $director$, $editor$, $actor$, $genre$, and $release\ year$. For simple illustration, we donate these attributes as Ad(director), Ae(editor), Aa(actor), Ag(genre), and Ay(release year). We make a year division and generate a class for every ten years from 1900 to 2010. The user profile in *Item Attribute-based Model* can be represented as frequency distribution over the generated features. In the following experiments, we adopt the brief form discussed in sectionIV as the selected model, say V^{Ad}, V^{Ae} ,V^{Aa}, V^{Ag} and V^{Ay}. Take Ag as an example, the genre-based user profile is represented as the relative frequency of genre counted from user u's interacted items.

B. Evaluation Metrics and Environment

We adopt the $Accuracy$ metric to evaluate the effectiveness of our method. For a given user u, the $Accuracy(u)$ is defined as the fraction of top-k variant identification cases in m times of random partition. So, for a certain user profiling method M, the accuracy of M is formalized as:

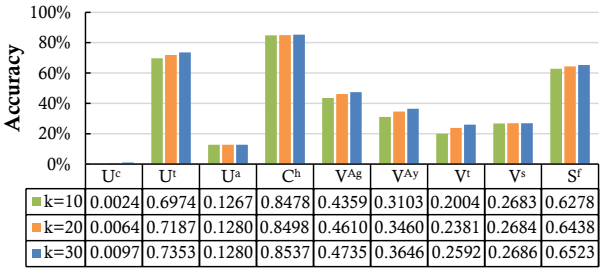
$$Accuracy(M) = \frac{1}{|U_n|} \sum_{u \in U_n} Accuracy(u) \tag{5}$$

The proposed algorithm is implemented in Java. All the experiments are performed on desktop PC with Intel Core i5 2.90GHz processor, 16GB RAM and operating system Windows 7.

C. Experiments Results and Analysis

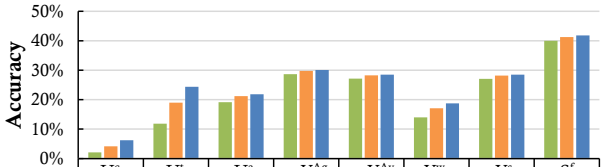
Effectiveness of Model. This experiments evaluate which model is more effective in identifying user variants. The results are shown in Fig. 2. We perform the $top-k$ variant identification algorithm based on different user models against two datasets. From this figure, we can find that there is an obvious different performance on two datasets. The accuracy on $Movielens$ dataset is higher than that on $Amazon$ dataset in general since it has fewer users than $Amazon$. There is an interesting phenomenon that *Cxt-based User Model* (T^h) has the highest accuracy over 80%, which illustrates that the temporal pattern of user behaviors is very helpful in profiling a user. Since the record $time$ in $Amazon$ dataset has been generalized as a fix value, we cannot profile a user as a *Cxt-based User Model*.

[1]http://www.omdbapi.com/



	U^c	U^t	U^a	C^h	V^{Ag}	V^{Ay}	V^t	V^s	S^f
k=10	0.0024	0.6974	0.1267	0.8478	0.4359	0.3103	0.2004	0.2683	0.6278
k=20	0.0064	0.7187	0.1280	0.8498	0.4610	0.3460	0.2381	0.2684	0.6438
k=30	0.0097	0.7353	0.1280	0.8537	0.4735	0.3646	0.2592	0.2686	0.6523

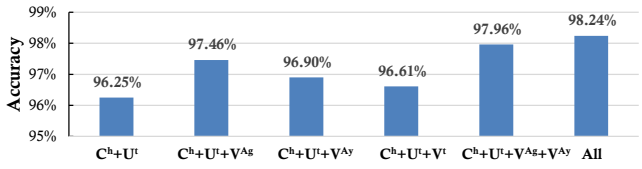
(a) Movielens



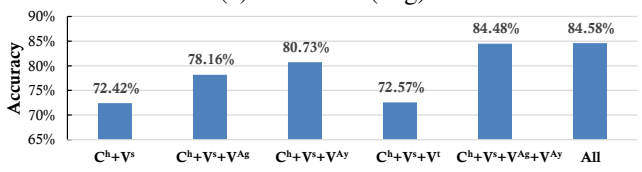
	U^c	U^t	U^a	V^{Ag}	V^{Ay}	V^w	V^s	S^f
k=100	0.0209	0.1183	0.1917	0.2869	0.2719	0.1398	0.2711	0.4000
k=200	0.0420	0.1902	0.2117	0.2974	0.2825	0.1709	0.2820	0.4125
k=300	0.0623	0.2437	0.2183	0.3005	0.2848	0.1877	0.2848	0.4183

(b) Amazon

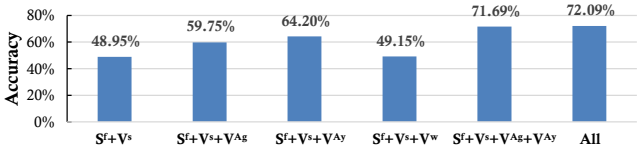
Fig. 2. Accuracy on Modeling



(a) Movielens(Tag)



(b) Movielens(Rating)



(c) Amazon

Fig. 3. Model Combination

The *Tag-based Model* (U^t) performs very well on $Movielens$ dataset. This confirms that the semantic information of tags is very useful in distinguishing users. But this method is not applicable for the $Amazon$ dataset since there is no available tags. So we have to extract some tags from the comments, which weaken the advantages of the *Tag-based Model* (U^t). The *Implicit Characteristic Model* (S^f) also has a highlighted performance, especially in $Amazon$ dataset, which shows the high effectiveness of implicit characteristic. The *Item Attribute-based Model* (e.g. V^{Ag} and V^{Ay}) performs relatively stable and fair on both datasets.

Effectiveness of Model Combination. Beside the independent models, we also evaluate the effectiveness of model combination. Based on the observation of previous experiments, we combine the models with good performance. For any two models m_i and m_2, their combination model is represented as $\widetilde{M}_{m_1+m_2}$. For example, $\widetilde{M}_{S^f+V^s+V^{Ay}}$ means a combination of *Implicit Characteristic Model* (S^f), *Common Attitude-based User Model* (V^s) and *Item Attribute-based Models* (V^{Ay}). User similarity is then evaluated as the sum of similarities against each separate model. Formally,

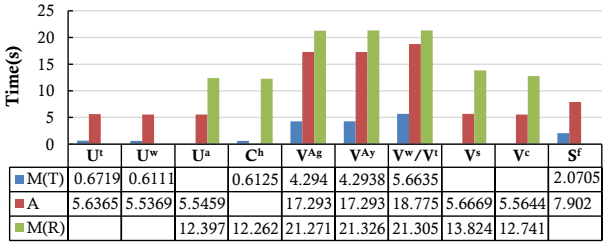
$$Sim_{\widetilde{M}} = \sum_{m \in \widetilde{M}} \omega(p_m(u_1), p_m(u_2)) \tag{6}$$

, where $p_m(u)$ is the profile based on the model m.

The results of combination models on $MovieLens(tag)$ dataset is shown in Fig.3(a), where $x-axis$ is the combined models and $y-axis$ is the corresponding accuracy on $k = 10$. Comparing to the results of each separate model, we can see that the combination significantly improve the accuracy. Moreover, the combination of models with high accuracy remains a good performance. For example, the accuracy of model V^{Ag} is higher than model V^{Ay}, and the combination $\widetilde{M}_{C^h+U^t+V^{Ag}}$ is also higher than $\widetilde{M}_{C^h+U^t+V^{Ay}}$. Besides, the more models used for combination, the higher accuracy. There are similar results on $MovieLens(rating)$ dataset and $Amazon$ datasets, shown in Fig.3(b) and Fig.3(c).

Influence of Quantity. The experiment studies how the number of user behaviors influences the identification. The results on both datasets are shown in Fig. 5, where $x-axis$ is the record number and the $y-axis$ is the accuracy under a given model. To clearly claim the impact of the record number, we draw a regression line for the data points in each scatter. These figures demonstrate that the number of records do have important effect on identification accuracy. Further more, the trends of these regression lines are also consistent with the results of previous experiments. Although there may not exist an exact number that guarantees 100% identification accuracy, we can still conclude that the user with over 500 interaction records has more possibility to be identified.

Efficiency. Finally, we study the efficiency of each model and show the results in Fig. 4, where $y-axis$ shows the average time of identifying a specific user and $x-axis$ denotes different models. Since the number of records in dataset $Movielens(Rating)$ (i.e. $M(R)$) are much more than other two datasets, the models always take much more time. The user set in dataset $Movielens(Tag)$ (i.e. $M(T)$) is relative smaller, so it always takes less time. From this figure, we



	U^t	U^w	U^a	C^h	V^{Ag}	V^{Ay}	V^w/V^t	V^s	V^c	S^f
M(T)	0.6719	0.6111		0.6125	4.294	4.2938	5.6635			2.0705
A	5.6365	5.5369	5.5459		17.293	17.293	18.775	5.6669	5.5644	7.902
M(R)			12.397	12.262	21.271	21.326	21.305	13.824	12.741	

Fig. 4. Experimental Procedure

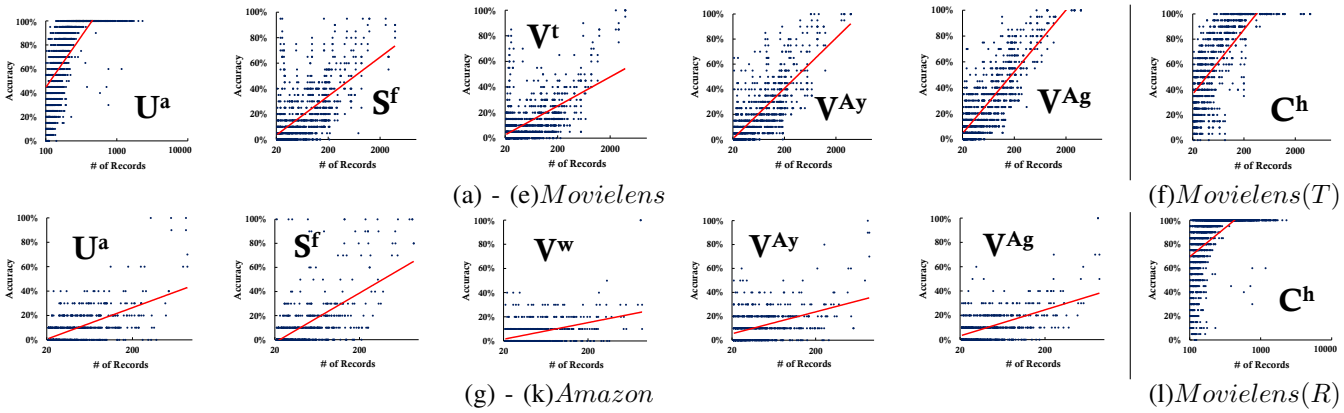


Fig. 5. Impact of Record Number

can find that the *Item-based User Models*(i.e. V^{Ag}, V^{Ay}, V^s) always take the longest time for three datasets. This is because we have to organize records data for each items before we generate the item-based profile for each user. As we known that model *Cxt-based User Model* (C^h) has the highest accuracy, and it takes less time than other models, it should be the best choice for identification. The *Tag-based Model* (U^t) performs also as pretty well as the *Common Attitude-based User Model* (V^s). Both of them cost less time and have a high accuracy. Besides, although the *Implicit Characteristic Model* (S^f) takes much more time because of the building and mapping of semantic tree, it is still a nice choice w.r.t its high accuracy.

VII. Conclusion and Future Works

In this paper, we investigate the user variants identification problem using both user behavior and item related information. We study the characteristics of user behaviors on social media and introduce two concepts $visibility$ and $distingushibility$ to preliminarily quantify whether a fake user can be identified. To better understand user intention and characteristics, we profile a user with apparent and implicit features. Based on these features, we propose the user Variants Identification Problem (VIP) and an identification algorithm, which finds the top-k similar variants in a social media. Experiments on two real datasets $MovieLens$ and $Amazon$ show that the proposed methods are effective against different features in identifying user variants.

In future, we will explore the social relationships among users and items as the assistant information for identification. By analyze the relationship between nodes, we will try to comprehensively analyze the similarity between users so as to improve the identification accuracy. Also we would like to study other identification problems and to provide an efficient way to protect users from being identified.

References

[1] C. Anderson. *The long tail: how endless choice is creating unlimited demand*. Random House, 2007.

[2] C. T. Chung, C. J. Lin, C. H. Lin, and P. J. Cheng. Person identification between different online social networks. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*, pages 94–101. IEEE Computer Society, 2014.

[3] D. Correa, A. Sureka, and R. Sethi. Whacky!-what anyone could know about you from twitter. In *Privacy, Security and Trust (PST), 2012 Tenth Annual International Conference on*, pages 43–50. IEEE, 2012.

[4] K. Cortis, S. Scerri, I. Rivera, and S. Handschuh. An ontology-based technique for online profile resolution. In *Social Informatics*, pages 284–298. Springer, 2013.

[5] N. Hoang Long and J. J. Jung. Privacy-aware framework for matching online social identities in multiple social networking services. *Cybernetics and Systems*, 46(1-2):69–83, 2015.

[6] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon. What's in a name?: an unsupervised approach to link users across communities. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 495–504. ACM, 2013.

[7] S. Liu, S. Wang, and F. Zhu. Structured learning from heterogeneous behavior for social identity linkage. 2015.

[8] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 51–62. ACM, 2014.

[9] P. d. Meo, E. Ferrara, F. Abel, L. Aroyo, and G.-J. Houben. Analyzing user behavior across social sharing environments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):14, 2013.

[10] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[11] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 173–187. IEEE, 2009.

[12] M. Payer, L. Huang, N. Z. Gong, K. Borgolte, and M. Frank. What you submit is who you are: A multimodal approach for deanonymizing scientific publications. *Information Forensics and Security, IEEE Transactions on*, 10(1):200–212, 2015.

[13] S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, and C. Chen. Mapping users across networks by manifold alignment on hypergraph. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[14] J. Vosecky, D. Hong, and V. Y. Shen. User identification across multiple social networks. In *Networked Digital Technologies, 2009. NDT'09. First International Conference on*, pages 360–365. IEEE, 2009.

[15] R. Zafarani and H. Liu. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 41–49. ACM, 2013.