Beyond Your Interests: Exploring the Information Behind User Tags

Weizhi Ma(✉), Min Zhang, Yiqun Liu, Shaoping Ma, and Lingfeng Chen

State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
mawz14@mails.tsinghua.edu.cn, {z-m,yiqunliu,msp}@mail.tsinghua.edu.cn,
clf0506@cs.duke.edu

Abstract. Tags have been used in different social medias, such as Delicious, Flickr, LinkedIn and Weibo. In previous work, considerable efforts have been made to make use of tags without identification of their different types. In this study, we argue that tags in user profile indicate three different types of information, say *the basics* (age, status, locality, etc), *interests* and *specialty* of a person. Based on this novel user tag taxonomy, we propose a tag classification approach in Weibo to conduct a clearer image of user profiles, which makes use of three categories of features: general statistics feature (including user links with followers and followings), content feature and syntax feature. Furthermore, different from many previous studies on tag which concentrate on user specialties, such as expert finding, we find that valuable information can be discovered with *the basics* and *interests* user tags. We show some interesting findings in two scenarios, including user profiling with people coming from different generations and area profiling with mass appeal, with large scale tag clustering and mining in over 6 million identical tags with 13 million users in Weibo data.

Keywords: Weibo · Tag classification · User group profiling

1 Introduction

Recent years, social tagging has become popular with the launch of sites like Delicious, Flickr, LinkedIn and Weibo. Since then, lots of social systems that support tagging of a variety of resources have been built. Tagging is a process in which a user assigns tags to an object. On Delicious, user can tag URL. On Flickr, user can assign tags to a photo. On LinkedIn and Weibo, user can add tags to themselves, named user tags. Take Weibo as an example, each user can add no more than ten tags to himself. The length of each tag is limited in 7 Chinese characters or 14 English letters, while the content of tag can be anything you want to describe yourself.

Due to the widely usage of tags, different techniques are employed to study various aspects of tagging [1]. The information behind tags is valuable in many

This work was supported by National Key Basic Research Program (2015CB358700) and Natural Science Foundation (61472206, 61073071) of China.

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 257–269, 2015.
DOI: 10.1007/978-3-319-25207-0_22

research areas. For example, Giannakidou et al. investigate to co-cluster tags with social data sources [2] and user interests discovering with tags [6]. Many researchers concentrate on using the information behind tags in Flickr to improve the performance of image retrieval [3–5]. Pennacchiotti et al. [7] propose a machine learning approach for twitter user classification based on hashtags. Many social tag prediction or recommendation work are conducted[8–10].

Tag studies based on Weibo come in many ways. Ghosh et al. [11] and Liang et al. [12] try to make use of user tags in Weibo to conduct expert finding studies. An automatic tag recommendation algorithm for Weibo is proposed in Wang et al.'s work [13]. To the best of our knowledge, in previous work, studies make use of different user tags indiscriminately. In fact, we find that user tags reflect user characteristics in three dimensions: 1) Tags show the attributes and status of a user, which are always ignored. 2) Tags indicate the topics that users are interested in. 3) Tags reveal the users's specialities. Therefore, we propose that user tags indicate three different kinds of information: *the basics*, *interests* and *specialty*. Moreover, we proposed a novel feature extraction method with the help of search engine. The analysis and classification will be introduced in Section 2.

As mentioned above, a lot of work in tag analysis concentrates on user specialties such as expert finding or user interests discovering, respectively. Several research efforts have been made for extracting profile information of a person [14,15]. Tang et al. [16] take user tags into account in user profiling. With the help of tag taxonomy, we find that valuable information can be discovered to profile user groups by considering both user *basics* and *interests*. We conduct user group profiling with different generations and area profiling with mass appeal by creating lists of keywords in Section 3 & 4.

Our main contributions are the following:

- Contrary to make use of different tags indiscriminately, we find that user tags indicate three different types of information, *the basics*, *interests* and *specialty*.
- We propose several novel tag feature extraction methods, which take features from tag links, user content and search engine for user tag classification.
- We find that valuable information can be explored in user tags with the help of tag taxonomy , such as user profiling with different generations and area profiling with mass appeal.

The remainder of this paper is organized as follows: In Section 2, we introduce work about user tag analysis and classification procedure. In Section 3, we present the result of user profiling with different generations. While in Section 4, we present our attempts in area profiling based on mass appeal and the method of characteristic tags extraction. We draw final conclusions and the outline of future work in Section 5.

2 User Tag Classification in Weibo

2.1 User Tag Taxonomy

In previous tag studies, researchers adopt identical data processing methods to different tags. In fact, tags may carry different types of information. For example,

Gloria Tang Tsz-kei, a famous Hong Kong singer, has a user tag list in Weibo which contains: "After 90's", "Musician", "Lively", "Singer", "Leo", "Like to amuse". We can see that "After 90's", "Leo" and "Lively" is *the basics* of hers, while user tag "Like to amuse" is her *interests*, "Singer" and "Musician" indicate her *specialty*. Based on analysis on large scale of user tag data, we propose user tag taxonomy as follows:

- ***The basics***: Tags which indicate a user's age, state, locality, constellation, blood type and other user basics, like "After 90's", "Libra", etc.
- ***Intersests***: Tags that show a user's interests. For instance: "singing", "sports", "traveling", etc.
- ***Specialty***: Tags which reveal a user's specialty, like "doctor", "teacher", etc.

2.2 Feature Extraction for Classification

We attempt to use an automatic method to conduct tag classification. The first step is user tag's feature extraction. We design some features for tags:

Statistical Features. We designed five features based on statistic as follows:

1. **Popularity**: the usage percentage of tag t in all users in the dataset.

$$Popularity(t) = \frac{|\{u|t \in tag(u)\}|}{|u|}$$

2. **Absolute position**: the average of tag t's rank position in user's tag list.

$$Absolute\text{-}Position(t) = \frac{\sum_{u \in \{v|t \in tag(v)\}} rank(t,u)}{|\{v|t \in tag(v)\}|}$$

($rank(t,u)$ means tag t's rank position in user's tag list u.)

3. **Relative position**: the average relative occurrence position of tag t in user's tag list.

$$Relative\text{-}Position(t) = \frac{\sum_{u \in \{v|t \in tag(v)\}} \frac{rank(t,u)}{|tag(u)|}}{|\{v|t \in tag(v)\}|}$$

4. **Co-occurrence percentage in followers**: the usage percentage of this tag in the followers of the user who has this tag.

$$Followers\text{-}Co(t) = \frac{\sum_{u \in \{v|t \in tag(v)\}} \frac{|\{w|u\text{->}w, t \in tag(w)\}|}{|\{w|u\text{->}w\}|}}{|\{v|t \in tag(v)\}|}$$

(u->w means u follows w.)

5. **Co-occurrence percentage in followings**: the usage percentage of this tag in the following of the user who already has this tag.

$$Following\text{-}Co(t) = \frac{\sum_{u \in \{v|t \in tag(v)\}} \frac{|\{w|w\text{->}u, t \in tag(w)\}|}{|\{w|w\text{->}u\}|}}{|\{v|t \in tag(v)\}|}$$

Feature 1 is based on the popularity of user tags, feature 2 & 3 are related to the user tag's position in tag list. We suppose that different types of tags have different popularity in Weibo, and the position in tag list could be a useful feature for the reason that users may tend to tag similar tags together. Different from feature 1, 2, 3, feature 4 & 5 are features extracted from user links(with follower and following relationship).

Content Features. Content features are extracted based on tag vector representation. The Word2vec algorithm provides an implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. It takes a text corpus as input and word vectors as output. We use an open-source package of Word2vec[1]. The user tag dataset is regarded as the input to Word2vec, which is formated as follows:

$$\begin{array}{llll} t_{i,1} & t_{i,2} & \ldots & t_{i,n_i} \quad (u_i\text{'s tag list.}) \\ & & & \ldots \\ t_{j,1} & t_{j,2} & \ldots & t_{j,n_j} \quad (u_j\text{'s tag list.}) \end{array}$$

As a result, each tag get a 200-dimension floating-point vectorized representation in the output of Word2vec.

Search Engine Based Syntax Features. We can extract syntax features by considering the co-occurrence frequency of the tags within certain sentences. We try to find the frequency with the help of a search engine. More specifically, we construct some sentences pattern using the user tags, put the sentences into Baidu[2] one by one in Chinese, and record the count of items and exact matching items returned by search engine. We propose 3 types of patterns including 9 instances, and the feature dimension we extracted from syntax is 18.

- ***The basics*** **related patterns:**
  I am __ (我是__, 我__.). E.g: I am 18 (我18岁). I am in after 90's (我是90 后).
  I am an Aries (我是白羊座).
  My __ (__的我). E.g: My passion (热情的我).
  I'm very __ (我很__). E.g: I'm very humorous (我很幽默).
- ***Intersest*** **related patterns**:
  I like __ (我喜欢__). E.g: I like traveling (我喜欢旅行).
  I love __ (我爱__). E.g: I love painting (我爱画画).
  (redundancy will be eliminated if there is existing "love" or "like" in the tag.)
- ***Specialty*** **related patterns**:
  I'm good at __ (我懂__). E.g: I'm good at design(我懂设计).
  __ expert (__专家, __家). E.g: political expert(政治家).

[1] https://code.google.com/p/word2vec/
[2] https://www.baidu.com, a popular search engine in China.

2.3 Dataset

The dataset we used in our work is a public Weibo dataset, provided by China Pameng[3]. The dataset is collected in Weibo from October 2012 to May 2013, which contains user's tag lists and user's follower & following relationships.

The number of user accounts in the dataset is 13,170,561. The sum of unique tags is 6,157,143. Each user has 123 followings and 6.21 tags on average.

It is necessary to preprocess the dataset due to the impact of noise brought by zombie users [17]. In this work, we filter out users whose followings or followers are fewer than 10. After filtering, we get 10,659,899 users with 6,156,993 unique tags.

2.4 Experiments and Results

In proposed taxonomy, a tag could belong to more than one types, for example: "Photography" and "music" can be classified into both *intersests* and *specialty*. At this time, the following labeling criteria can be adopted:

- If a user tag could be *the basics* & *interests*, or *the basics* & *specialty*, we tend to label it as *interests* or *specialty* but not *the basics* for the reason that we suppose *interests* and *specialty* tags convey more information.
- When it is really hard to judge whether the tag belongs to *interests* or *specialty*, we tent to classify it into a new type *interests* & *specialty*.
- If a user tag is meaningless, such as user tag "sser", "just", we take it as *noise*.

We choose the top 100 frequent tags and randomly select 1,086 tags from the whole tag set. We label these 1,186 tags according to the criteria manually. The distribution of tags is shown in Figure 1.

As we can see from Figure 1, more than 40% of user tags are classified into *interests*, which indicates that *interests* is a key component of user tags. The noise of user tags is about 20%, suggesting that a certain quantity of user tags is difficult to be understood.

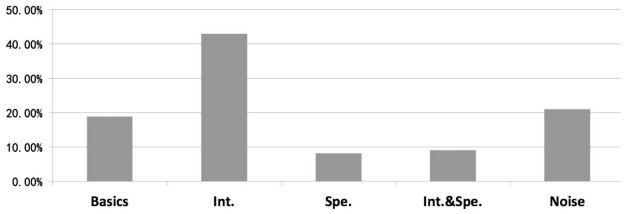

Fig. 1. Tag labeling result

[3] An organization which collects Weibo data. The official website is http://cnpameng.com.

As can be seen from the distribution of user tags, the number of user tags in different types varies widely. Thus, before conducting classification experiment, we carry out over sampling to make the count of tags in different types balanced. Specially, in our experiments, tags in *interests* & *specialty* are regard as both *interests* and *specialty* tag in classification experiments.

After feature normalization, we leveraged different classification algorithms, such as SVM, Multilayer Perceptron, Naive Bayes and Decision Tree. The performance in Decision Tree is the best. Hence we choose Decision Tree as the classification algorithm. We use the statistic, content and syntax features separately for classification experiments. In 10-fold cross validation, the classification results are listed in Table 1.

Table 1. Classification Results in Using Different Features

Feature	Basics			Interests			Specialty		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Statistics	81.6%	81.6%	81.6%	65.9%	65.4%	65.2%	83.7%	83.0%	82.9%
Content	87.4%	87.1%	87.1%	77.8%	77.2%	77.1%	89.8%	89.3%	89.3%
Syntax	83.2%	82.9%	82.8%	74.7%	74.6%	74.6%	86.7%	86.6%	86.6%
Statistics & Content	**88.4%**	**88.1%**	**88.0%**	79.9%	79.3%	79.2%	89.3%	88.7%	88.6%
Statistics & Syntax	85.9%	85.4%	85.3%	77.2%	77.0%	76.9%	89.7%	89.5%	89.4%
Content & Syntax	87.7%	87.3%	87.3%	78.1%	77.7%	77.6%	89.3%	88.8%	88.8%
All	87.9%	87.4%	87.3%	**81.4%**	**80.8%**	**80.7%**	**90.1**%	**89.7**%	**89.6**%

Table 1 indicates that the performance in *specialty* type classification is better than others and the performance in *interests* tags classification is the worst, which may be caused by the variety of *interests* tags. The performance of using content features individually is pretty good, which shows that Word2vec is really effective. Furthermore, comparing the classification results in using statistics & syntax features with content features, we find that the precision, recall and F1-Measure values are nearly the same, showing that the statistics & syntax features are effective.

Moreover, we find that the syntax features are not always helpful. In identifying *interests* and *specialty* tag tasks, the performance will be better when syntax features are applied, but worse in *the basics* identification task. It can be attributed to that tags in *the basics* are easier to be identified, the result is good enough without syntax features, while as *interests* or *specialty* tags, they are more difficult to be classified correctly with only statistic & content feature for their variety.

3 User Profiling with Different Generations

Researchers concentrate on expert finding and recommendation in some special domains by utilizing user tags in previous studies. However, we find that after user tag classification, extra valuable information can be discovered. In fact, we can divide users into different user groups according to their *basics* tags, for example, age, constellation, location, etc. In this section, we focus on user profiling

with different generations based on user age. The result indicates that with the help of tag taxonomy, some extra information can be found in *interests* tags.

3.1 User Generation Distribution

In this section, our work focus on user group's interest distribution modeling and user group profiling. First, users are divided into different generation groups according to their tag lists. We set three generation groups: "After 70s", "After 80s" or "After 90s" (People who were born between 1970 and 1979 belong to "After 70s", and others are defined in a similar way.). For the reason that the number of "After 60s" and "After 00s" users is smaller than 1,000 based on user tags, we don't take them into consideration. If user's tags don't reveal his/her age, we will ignore this user.

The dataset has been introduced in Section 2.3, and the distributions of users in different generations are as follows: There are 19,821 users in user group "After 70's", 1,703,438 users in user group "After 80's", and 1,681,892 users in user group "After 90's". We find that many users don't have age tags, which can be our future work.

3.2 *Interests* Tag Clustering for User Profiling

Over 1.3 million different user tags are tagged in these users. The frequency of user tags follows a power law distribution. Considering that the quantity of user tags is extremely large, we filter out the tags whose frequency is less than 500. After filtering, the number of user tags has reduced to 1,733, while the frequency of these tags accounted for more than 80% of all tag's frequency.

It's hard to identify user's interests distribution in a large scale, so we try to cluster the tags into several tag sets. The method we chose to conduct a tag clustering experiment is K-means. We assigned the k value with 15 and put all the 1,733 user tags into tag clustering experiment. Figure 2 shows the percentages of users in each tag set. Users may have more than one tag, so the sum of percentages in each set is not 100%. We can find that the distributions of After 80's and After 90's are very similar. It's hard to get useful information from this clustering result.

Thus, before clustering the user tags, we try to filter out tags that do not belong to *interests* type. We conduct an automatic classification which is introduced in Section 2.4 and get 907 user tags that belong to *interests* type.

Clustering experiment is conducted on the 907 user tags. The result is that each tag set can be assigned with a keyword as its feature at this time, which has better performance than using all user tags. For example, design, comic, art, etc. But it's unsatisfactory in that one of the sets contains too many user tags, so we labeled this tag set and divided it into 5 tag sets manually. Moreover, famous stars from different countries are clustered into 3 sets, we combined the three sets into a big set. At last, the user tags are clustered into 17 sets. The keywords of each set are shown in Table 2. The keyword "Others" means that this set is mixed by a variety of *interests* tags without a keyword.

Table 2. Interests Clustering Results

Set	Key Word	Set	Key Word	Set	Key Word	Set	Key Word	Set	Key Word	Set	Key Word
0	Design	**1**	Art	**2**	Music	**3**	Stars	**4**	Social Science	**5**	Technology
6	Reading	**7**	Others	**8**	Housing	**9**	Geek	**10**	Comic	**11**	Fashion
12	Travel	**13**	e-commerce	**14**	Finance	**15**	Food	**16**	Sports		

3.3 Result and Analysis

We calculate the tag amount of each clustering set in user groups of different generations. The statistical results show the differences of interests in different generations, which are drawn in Figure 3. We find that the five types, Design, Music, Art, Stars and Technology, contain much more users than other groups. It indicates that these are social common interests. To see the results of other interest sets more clearly, we removed 6 tag sets and draw Figure 4.

In Figure 4, it is apparent that the popularity of different Interests varies in each group. Users in the group of "After 70's" are keen on Reading and Housing, which is reasonable as people in this age are more concerned about living a better life and personal finance. Young people, aged in 15-35, show more interest in finance than "After 70's". Moreover, we find that "After 70's" even show more interests in sports than "After 80's" and "After 90's". It indicates that young people's enthusiasm in sports is relatively low, which is an ominous sign.

We can find many other interesting information from the results. In fact, this work implies that valuable information can be mined in user tags after tag taxonomy. Moreover, these analyses will be helpful in tracing the transformation and evolution of social common interests.

4 Area Profiling with Mass Appeal

In Section 3, we introduced the findings in user profiling with different generations based on user tag taxonomy. In this section, we focus on area profiling with mass appeal, which is based on user groups from different provinces/cities.

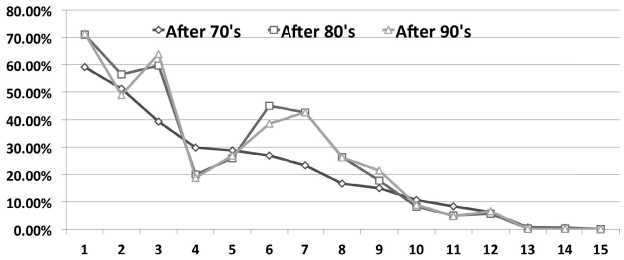

Fig. 2. User Interests Distribution without tag filtering

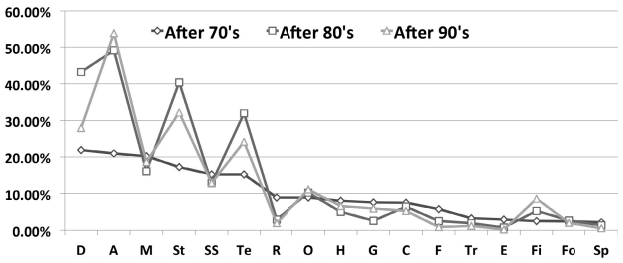

Fig. 3. User Interests Distribution

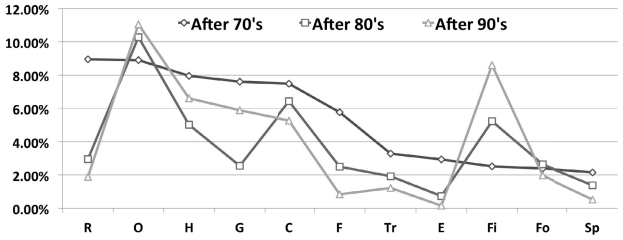

Fig. 4. User Interests Distribution in Eleven Sets

Users' area tags indicate the location of the users in usual cases. Thus we can infer the location information of users from their tag lists. According to this, users can be classified into different province/city user groups. Each group has a tag list which is unioned by the tags owned by the users who belong to the group. Then, we can perform area profiling experiments to find the characteristic tags in each user group to describe the province/city. Through comparing the differences in filtering out and not filtering out *the basics* user tag, we find that user tag taxonomy is helpful in characteristic tag finding. Moreover, the result is evaluated by a labeling task on Zhongbao, a Chinese crowdsourcing platform[4].

4.1 Area Characteristic Tags Extraction

For the reason that most of Weibo users are Chinese, many user tags about location are provinces or cities in China. Considering that if we choose the city tag appearing in a user's tag list to classify users into different groups directly, the number of user groups will be large and the amount of user in each city will be small. So we merged the user groups according to the affiliations of the areas. For example, Guangzhou is a city of Guangdong province in China, if a user tagged himself/herself with "Guangzhou", we put him/her into the user group of Guangdong province. Furthermore, people who have no location tags are filtered out in this experiment. For the reason that there are 34 provinces, autonomous regions, municipalities and special administrative regions in China,

[4] http://www.chinacrowds.com

we construct 34 user groups. The dataset has been introduced in Section 2.3. As a result, we get more than 360,000 users from the 34 provinces/cities.

Inspired by relative entropy, we propose a feature extraction method names tag entropy. Basic symbol notations are defined in Table 3. The tag list of each area is the combination of user's tag list in the group.

Table 3. Basic Symbols Notation

Symbol	Definition
A	The set of the 34 areas $\{a_1, a_2, ..., a_{34}\}$.
B	Complementary set of the 34 area $\{b_1, b_2, ..., b_{34}\}$, b_i represents the union set of $\{a_1, a_2, ..., a_{i-1}, a_{i+1}, ..., a_{34}\}$.
$\|a_i\|$, $\|b_i\|$	The number of users in set a_i, b_i.
$TagA(x, i)$	The frequency of tag x in a_i's tag list.
$TagB(x, i)$	The frequency of tag x in b_i's tag list.
$TE_{x,k}$	The tag entropy of tag x in a_k.

The formulations to calculate tag entropy of tag x in a_k are as following:

$$P_{x,i} = \begin{cases} \frac{TagA_{x,i}}{|a_i|} & x \text{ in } a_i \\ \frac{1}{|a_i|} & x \text{ not in } a_i \end{cases} \qquad (4-1)$$

$$Q_{x,i} = \begin{cases} \frac{TagB_{x,i}}{|b_i|} & x \text{ in } b_i \\ \frac{1}{|b_i|} & x \text{ not in } b_i \end{cases} \qquad (4-2)$$

$$TE_{x,k} = P_{x,k} * log(\frac{P_{x,k}}{Q_{x,k}}) \qquad (4-3)$$

We can get the tag entropy of each tag in different user groups. In our approach, we use the top N tags in $TE_{x,k}$ value as the characteristic tag of each province/city, and we ignore the order of the top N tags.

With the method introduced above, we get the characteristic tags of each province/city of China. For example, the top five characteristic tags of Qinghai province is: Qinghai Lake, Qinghai-Tibet Plateau, The Origin Of Three Rivers, Tibetan, Xia Du (an alias of Xining.); The result of Qinghai is consistent with generally acknowledged. However, in fact, it's challenging to evaluate the result of province/city characteristic tags extracted by the experiments directly, because the impression of an area is usually based on people's background. So we proposed a labeling task to evaluate the results.

4.2 Evaluation and Analysis

We design a labeling task to evaluate the results: We choose the top 20 tag entropy tags and 180 tags selected by using the method of multistage stratified

sampling in the province's tag list. So that each province/city has 200 tags. Then, the 200 tags of each province/city will be evaluated whether the tag is a characteristic tag of the province/city with 3 level labeling: "Relevant", "partially relevant" or "irrelevant". Each tag is labeled by three users and the labeling of the tag is depended on the majority opinion. If the labels of a tag given by the three annotators are different with each other, this tag will be labeled as "partially relevant".

The labeling task is released on Zhongbao crowdsourcing platform. In order to compare the differences between raw province's tag list and province's tag list after *the basics* tag filtering, we calculate the precision, recall, F1-Measure on the two data sets by considering top k characteristic tag result as the right answer. In evaluation, we attempt two method: One is regarding the tags labeled with "relevant" as characteristic tags, the other one is regarding the tags labeled with "relevant" or "partially relevant" as characteristic tags. Table 4 shows the results of the evaluation.

Table 4. "relevant" and "relevant & partially relevant" label counts

Result		**Raw data**			**Filtered data**		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Relevant	**Top5**	10.85%	44.12%	17.42%	11.43%	46.47%	18.35%
	Top10	23.44%	47.65%	31.42%	23.88%	48.53%	32.01%
	Top20	47.90%	48.68%	48.29%	**48.34%**	49.12%	48.73%
Relevant & Partially Relevant	**Top5**	11.79%	89.41%	20.83%	12.10%	**91.76%**	21.38%
	Top10	23.12%	87.65%	36.59%	23.27%	88.24%	36.83%
	Top20	45.31%	85.88%	59.32%	45.85%	86.91%	**60.03%**

We find that the performance on filtered data is better than raw data in Table 4, which indicates that after tag filtering, we can get better results. For the reason that only considering "relevant" tag is more strict than using both "relevant" and "partially relevant", the recall is obviously lower than the latter. The precision of the data increases with the increment of k, as the right answer set is expanded with the increment of k.

The highest precision, 48.34% , is achieved in "relevant" tags with top 20 on filtered data. The highest recall, 91.76%, is found in "relevant" and "partially relevant" tags with top 5 on filtered data. For the reason that the value of k restricts the precision and recall, the highest F1 only achieved in 60.03% with Top 20 "relevant" and "partially relevant" tags on filtered data.

In this part, we use tag entropy to conduct city profiling with mass appeal experiments. Our labeling task shows that tag taxonomy is useful in area profiling, which helps get better results.

5 Conclusion and Future Work

In this paper, we presented our work in user tags on Weibo. Firstly, we showed that tags in user profile indicate three different kinds of information: *basics*,

interests and *specialty* of a person. We introduced our analysis about user tags in Section 2.1. With the help of search engine, we proposed a novel user tag feature extraction method. We conduct experiments to classify the user tags into different types with Decision Tree. The classification results show that the statistic and syntax features we extracted are effective.

Furthermore, we find that valuable information can be discovered with the *basics* and *interests* user tags. We present some interesting findings in Section 3 and Section 4: User profiling with different generations and area profiling with mass appeal. In user profiling with different generations, we take both user age and user interests into consideration. The result shows the interests distribution of users in different generations in using *interests* user tags. In area profiling with mass appeal, we proposed a method to extract area characteristic tags with tag entropy. Furthermore, we designed a labeling task to verify that if the tag classification is helpful. In fact, we find many valuable information behind the user tags by considering the *basics* and *interests* at the same time. We believe that valuable information can be found in other platforms. Not only user tags can be classified into different types, tags in other platforms indicate different information. But in other platforms, the taxonomy of tags may be different.

Future work include user *basics*, *interests* and *specialties* finding with the combination of the implicit user information in Weibo content, the explicit information behind hashtags and user tag information. Moreover, besides profiling users with generation and area basics, further profiling work in users with other basics, such as status and personality, will be conducted.

References

1. Gupta, M., Li, R., Yin, Z., et al.: Survey on Social Tagging Techniques. ACM Sigkdd Explorations Newsletter **12**(1), 58–72 (2010)
2. Giannakidou, E., Koutsonikola, V., Vakali, A., et al.: Co-clustering tags and social data sources. In: The Ninth International Conference on Web Age Information Management, WAIM 2008, pp. 317–324 (2008)
3. Li, X., Snoek, C.G.M., Worring, M.: Unsupervised multi-feature tag relevance learning for social image retrieval. In: Conference on Image and Video Retrieval, pp. 10–17 (2010)
4. Xiao, J., Zhou, W., Tian, Q.: Exploring tag relevance for image tag re-ranking. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1069–1070. ACM (2012)
5. Zhu, X., Nejdl, W., Georgescu, M.: An adaptive teleportation random walk model for learning social tag relevance. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information retrieval, pp. 223–232. ACM (2014)
6. Giannakidou, E., Koutsonikola, V., Vakali, A., et al.: In & out zooming on time-aware user/tag clusters. Journal of Intelligent Information Systems **38**(3), 685–708 (2012)
7. Pennacchiotti, M., Popescu, A.M.: A Machine Learning Approach to Twitter User Classification. ICWSM **11**, 281–288 (2011)

8. Heymann, P., Ramage, D., Garcia-Molina, H.: Social tag prediction. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 531–538. ACM (2008)
9. Sigurbjörnsson, B., Zwol, R.V.: Flickr tag recommendation based on collective knowledge. In: Www 2008 Proc of International Conference on World Wide Web pp. 327–336 (2008)
10. Seitlinger, P., Kowald, D., Trattner, C., et al.: Recommending tags with a model of human categorization. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 2381–2386. ACM (2013)
11. Ghosh, S., Sharma, N., Benevenuto, F., et al.: Cognos: crowdsourcing search for topic experts in microblogs. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 575–590. ACM (2012)
12. Liang, C., Liu, Z., Sun, M.: Expert finding for microblog misinformation identification. In: COLING (Posters), pp. 703–712 (2012)
13. Wang, X., Li, S., Zou, X., et al.: An automatic tag recommendation algorithm for micro-blogging users. In: 2013 International Conference on Computer Sciences and Applications (CSA), pp. 398–401. IEEE (2013)
14. Cunningham, H., Maynard, D., Bontcheva, K., et al.: GATE: an architecture for development of robust HLT applications. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 168–175 (2002)
15. Yu, K., Guan, G., Zhou, M.: Resume information extraction with cascaded hybrid model. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 499–506 (2005)
16. Tang, J., Yao, L., Zhang, D., et al.: A Combination Approach to Web User Profiling. ACM Transactions on Knowledge Discovery from Data **5**(1), 293–302 (2010)
17. Binlin, C., Jianming, F., Jingwei, H.: Detecting zombie followers in sina microblog based on the number of common friends. International Journal of Advancements in Computing Technology **5**(2) (2013)