SCIENCE CHINA
Information Sciences

• LETTER •

February 2016, Vol. 59 029101:1–029101:3
doi: 10.1007/s11432-015-5466-8

A novel user behavioral aggregation method based on synonym groups in online video systems

Tingting FENG*, Yuchun GUO & Yishuai CHEN

School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044*, China*

Received September 20, 2015; accepted October 13, 2015; published online January 5, 2016

Citation Feng T T, Guo Y C, Chen Y S. A novel user behavioral aggregation method based on synonym groups in online video systems. Sci China Inf Sci, 2016, 59(2): 029101, doi: 10.1007/s11432-015-5466-8

Dear editor,

Huge commercial interests have arisen as online video service surges, and user demographic information is precious and critical to predict user preferences, make recommendations and increase advertisement income. For instance, lady perfume advertisers usually target their advertisement to the females. Thus, many studies predict user demographic information such as gender based on their viewing behaviors in terms of videos [1–3]. However, with the enhancing of privacy protection awareness, it is increasingly difficult to get user demographic information [4]. Moreover, the lack of user viewing behavior records leads to serious data sparsity problem. To handle this problem, in our early work [5], we proposed a user behavioral aggregation method based on tags and keywords of videos, and further handled data sparsity problem via commonly-used dimension-reduced process. The limitation of this method is that such a process loses some original information of features inevitably. To solve this problem, in this paper, we propose a novel user behavioral aggregation method based on synonym groups (SGs), and then predict user gender with their behaviors, i.e., watching or not. This method not only solves data sparsity problem, but also preserves topic information and improves gender inference accuracy simultaneously.

User Behavioral Aggregation. Users often search their interested topics revealed in video titles, and choose videos most related to their interests to view. Since video titles disclose different user groups' interests, we can infer user gender based on revealed words. In this paper, we combine words with the same meaning into a single topic group as a new feature to greatly reduce information losing via commonly-used dimension reduction. We define the semantic features as SGs based on WordNet which is a large lexical database of English as well as a useful tool for computational linguistics and natural language processing. WordNet groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms, which are interlinked by means of conceptual-semantic and lexical relations. Our behavioral aggregation method has three main advantages. (i) It decreases data sparsity by combining topics with the same meanings into a single feature. Since each SG covers several topics, and each topic covers several videos, it is evident that the number of features deceases largely. (ii) Different from general dimension reduction methods (e.g., principal component analysis (PCA)) which weight features and lose the original information inevitably, our new method aggregates user behaviors via combing synonyms with almost no loss of information. (iii) It significantly improves the gender inference accuracy,

* Corresponding author (email: tt06274031@126.com)

The authors declare that they have no conflict of interest.

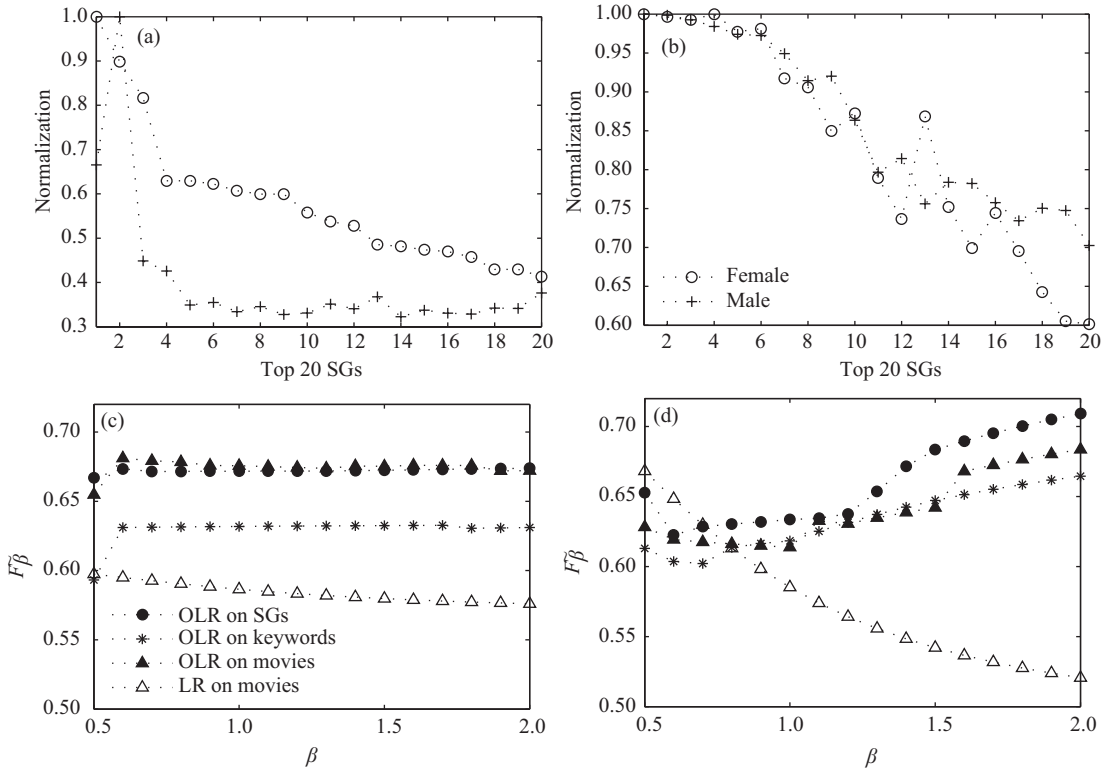

Figure 1 The viewer normalized percentage based on Top 20 popular SGs in (a) Flixster and (b) MovieLens datasets. $\tilde{F_\beta}$ values based on three types of features in (c) Flixster and (d) MovieLens datasets.

since SGs involve more user behavior information than keywords or movies with the same number do. The steps of aggregating user behaviors are as follows.

Step 1. We traverse all movie titles in the movie library of online video systems, and cut movie titles into individual words according to the delimiters, e.g., space, semicolon, comma, etc.

Step 2. We remove stop words from all of individual words according to the stop word corpus of WordNet. The stop word corpus includes words like "the", "to".

Step 3. We get meaning list for each individual word via WordNet's synsets, and denote every word with the most common explanation, i.e., the first meaning shown by WordNet's synsets. For instance, "drama" has several meanings and we denote it with the first meaning, i.e., "play".

Step 4. We build word groups with synonyms based on similatities. For example, word group "girl" covers several synonyms, i.e., "fille", "girl", "miss" and plural forms of those words. In addition, a common use of WordNet is to determine the similarity between words, and we require similarity equal to 1 to get perfect SGs.

Step 5. We aggregate user behaviors and build user behavior matrix based on SGs.

Datasets and Basic Statistic Analysis. Flixster is an American social movie site for discovering new movies. We take the dataset including users' registering in September 2009 and their corresponding records between Sep. 1st and Sep. 30th 2009. The dataset contains 912137 anonymous ratings of 38283 movies given by 20503 users. The female percent is 55.19%, and the male percent is 44.81%. MovieLens dataset is from a project of GroupLens Research about personalized recommendation systems. The dataset contains 1000209 anonymous ratings of approximately 3900 movies given by 6040 users who joined MovieLens in 2000. Only 28.29% of its users are female. In addition, video titles of MovieLens dataset include the information of video genres, and video titles of Flixster dataset include the related description of video content, such as actors and directors. We list the representing words of top 10 popular SGs in Table S1. The bold fonts distinguish the different representing words, and the underlines distinguish the same words with same order in both female and male lists in two datasets. And the frequencies of female and male viewers attracted by top 20 popular SGs for each dataset are shown respectively in Figure 1(a) and (b). It is obvious that different gender groups have different preferences of interest SGs.

Optimal Logistic Regression. In traditional Logistic regression (LR), a user $u \in N$ is predicted as male, if the probability $p_u > p_0$, $p_0 = 0.5$ and vice versa. Since the most commonly-used threshold value $p_0 = 0.5$ is not always optimal, we proposed the optimal LR (OLR) in early work [5], which makes p_0 equal to the point maximizing the value of optimization objective for all users in training dataset.

Evaluation Metric. Considering gender imbalance in online video systems, we take the integrated F-measure $\tilde{F}_\beta = \frac{(\beta^2+1)\tilde{\rho}\tilde{\gamma}}{\beta^2\tilde{\rho}+\tilde{\gamma}}$ for all classes as evaluation metric, where β is a positive real parameter to tradeoff the precision and recall. $\tilde{\rho}$ and $\tilde{\gamma}$ are the balanced precision and the balanced recall [5].

Results and Discussion. The basic statistic analysis and experiment results show that SGs can reduce data sparsity, involve more users, and improve the performance of OLR. In addition, the ratio between the number of users in training dataset and that in testing dataset is 8:2.

(i) Since each SG covers many individual words, and each word associates with lots of different video titles, the new semantic features largely aggregate user behaviors. In Flixster dataset, there are 38283 movies, and 16231 SGs. The population sparseness value of the movie rating matrix, i.e., the ratio of zero elements in the behavior matrix, is 99.89%, and this value based on viewing matrix of SGs is 99.17%. In MovieLens dataset, there are 3883 movies, and 2687 SGs. And the population sparseness values are 95.81% and 90.83%, respectively.

(ii) To show that our user behavioural aggregation method can improve the viewer coverage largely, we take top 500 popular features for each type, i.e., movies, keywords and SGs. Since MovieLens dataset requires that each viewer rates at least 20 movies, each type of features can cover all viewers. And in Flixster dataset, movies only cover 77.44% of viewers, and keywords [5] cover 85.68% of viewers. While SGs cover 93.81% of viewers. The viewer coverage of top 500 SGs improves 21.14% and 9.49% compared with top 500 movies and top 500 keywords respectively.

(iii) To show that our new features can further improve the performance of OLR, we compare evaluation results based on three types of features respectively, and take LR based on movie ratings as baseline. To make the features explainable and keep original information, we directly input user behavior matrix to gender inferers without any transformation. In Figure 1 (c) and (d), the triangle, the star and the circle dotted lines represent the $\tilde{F}_\beta$ values based on top 500 movies' ratings, keywords and SGs respectively. Overall, OLR based on SGs performs best. Although, in Flixster dataset, results of OLR based on SGs and movies are similar, the former one involves much more viewers, which improves 21.14% than the later one. In addition, OLR based on keywords does perform much better than LR based on movie ratings as analyzed in our early work. Obviously, according to the results in Figure 1, SG has absolute advantage. It improves the OLR largely and involves almost all the users.

Conclusion and Future Work. We propose a novel user behavioral aggregation method which not only solves the data sparsity problem, but also preserves topic information and improves gender inference accuracy. Since some words may adopt different explanations according to the practical circumstances, we will improve our user behavioral aggregation method to solve this problem in future work.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61271199, 61301082, 61572071).

Supporting information Table S1. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Weinsberg U, Bhagat S, Ioannidis S, et al. BlurMe: inferring and obfuscating user gender based on ratings. In: Proceedings of the 6th ACM Conference on Recommender Systems, New York, 2012. 195–202
2 Salamatian S, Zhang A, Calmon F D P, et al. How to hide the elephant-or the donkey-in the room: practical privacy against statistical inference for large data. In: Proceedings of IEEE Global Conference on Signal and Information Processing, Austin, 2013. 269–272
3 Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. In: Proceedings of the National Academy of Sciences, Berkeley, 2013. 5802–5805
4 Bruckman A. Gender swapping on the Internet. High Noon on the Electronic Frontier: Conceptual Issues in Cyberspace, 1996. 317–326
5 Feng T, Guo Y, Chen Y, et al. Tags and titles of videos you watched tell your gender. In: Proceedings of the IEEE International Conference on Communications, Sydney, 2014. 1837–1842