Future Generation Computer Systems 64 (2016) 186–197

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Binary sieves: Toward a semantic approach to user segmentation for behavioral targeting

Roberto Saia *, Ludovico Boratto, Salvatore Carta, Gianni Fenu

Dipartimento di Matematica e Informatica, Università di Cagliari, Via Ospedale 72, 09124 Cagliari, Italy

HIGHLIGHTS

- We propose a novel segmentation approach for user targeting, based on a semantic analysis of the items evaluated by a user.
- Through the semantic analysis we extend the ground truth, to generate non trivial segments.
- With respect to classic segmentation, the advertiser can introduce constraints and atomically model the user segments.

ARTICLE INFO

Article history:
Received 20 March 2015
Received in revised form
24 February 2016
Accepted 9 April 2016
Available online 19 April 2016

Keywords:
User segmentation
Semantic analysis
Behavioral targeting

ABSTRACT

Behavioral targeting is the process of addressing ads to a specific set of users. The set of target users is detected from a segmentation of the user set, based on their interactions with the website (pages visited, items purchased, etc.). Recently, in order to improve the segmentation process, the semantics behind the user behavior has been exploited, by analyzing the queries issued by the users. However, nearly half of the times users need to reformulate their queries in order to satisfy their information need. In this paper, we tackle the problem of semantic behavioral targeting considering *reliable* user preferences, by performing a semantic analysis on the descriptions of the items positively rated by the users. We also consider widely-known problems, such as the *interpretability* of a segment, and the fact that *user preferences are usually stable over time*, which could lead to a trivial segmentation. In order to overcome these issues, our approach allows an advertiser to automatically extract a user segment by specifying the interests that she/he wants to target, by means of a novel boolean algebra; the segments are composed of users whose evaluated items are semantically related to these interests. This leads to interpretable and non-trivial segments, built by using reliable information. Experimental results confirm the effectiveness of our approach at producing users segments.

1. Introduction

Behavioral targeting addresses ads to a set of users who share common properties. In order to choose the set of target users that will be advertised with a specific ad, a *segmentation* that partitions the users and identifies groups that are meaningful and different enough is first performed. In the literature it has been highlighted that classic approaches to segmentation (like *k*-means) cannot take into account the semantics of the user behavior [1]. Tu and Lu [2] proposed a user segmentation approach based on a semantic analysis of the queries issued by the users, while Gong et al. [1] proposed a LDA-based semantic segmentation that groups users with similar query and click behaviors.

When dealing with a semantic behavioral targeting approach, several problems remain open.

Reliability of a semantic query analysis. In the literature it has been highlighted that half of the time users need to reformulate their queries, in order to satisfy their information need [3–5]. Therefore, the semantic analysis of a query is not a reliable source of information, since it does not contain any information about whether or not a query led to what the user was really looking for. Moreover, performing a semantic analysis on the items evaluated by the users in order to filter them can increase the accuracy of a system [6–8]. Therefore, a possible way to overcome this issue would be to perform a semantic analysis on the description of the items a user positively evaluated through an explicitly given rating. However, another issue arises in cascade.

Preference stability. To complicate the previous scenario, there are domains like movies in which the preferences tend to be stable over time [9] (i.e., users tend to watch movies of the same genres

* Corresponding author.
E-mail addresses: roberto.saia@unica.it (R. Saia), ludovico.boratto@unica.it (L. Boratto), salvatore@unica.it (S. Carta), fenu@unica.it (G. Fenu).

http://dx.doi.org/10.1016/j.future.2016.04.006
0167-739X/© 2016 Elsevier B.V. All rights reserved.

or by the same director/actor). This is useful to maintain high-quality knowledge sources, but considering only the items a user evaluated leads to trivial sets of users that represent the target (this problem is known as *overspecialization* [10]).

Interpretability of the segments. The last open problem that has to be faced in this research area is the interpretability of a segment. Indeed, a recent survey on user segmentation (mostly focused on the library domain) [11], highlighted that, in order to create a proper segmentation of the users, it is important to *understand* them. On the one hand, easily interpretable approaches generate trivial segments, and even a partitioning with the *k*-means clustering algorithm has proven to be more effective than this method [12], while on the other hand, when a larger set of features is combined, the problem of properly understanding and interpreting results arises [13,14]. This is mostly due to the lack of guidance on how to interpret the results of a segmentation [15]. The fact that easily understandable approaches generate ineffective segments, and that more complex ones are accurate but not easy to use in practice, generates an important gap in this research area.

Our contributions. In this paper, we have moved the item analysis process from the canonical deterministic space model (i.e., that based on strict mathematical criteria) to a more flexible semantic space model that allows us to extend the analysis capability, which in the literature has been highlighted as a challenging topic [16,17]. In particular, we tackle the problem of *defining a semantic user segmentation approach, such that the sources of information used to build it are reliable, the generated segmentation is not trivial and it is easily interpretable*.

The proposed approach is based on a semantic analysis of the description of the items positively evaluated by the users. The choice to start from items with a positive score was made since it is necessary to start from a knowledge-base that accurately describes what the users like, so that our approach can employ the semantics to detect latent information and avoid preference stability.

The approach first defines a binary filter (called *semantic binary sieve*) for each class of items that, by analyzing the description of the items classified with the class, defines which words characterize it. In order to detect more complex targets, we are going to define an algorithm that takes as input a set of classes that characterize the ads that have to be proposed to the users and a set of boolean operators. The algorithm combines the classes with the operators by means of a boolean algebra, and creates the binary filters that characterize the combined classes. Then we consider the words (that as we will explain later, are actually particular semantic entities named *synsets*) that describe the items evaluated by a user, and use the previously created filters to evaluate a *relevance score* that indicates how relevant is each class of items for the user. The relevance scores of each user are filtered by the segmentation algorithm, in order to return all the users characterized by a specified class or set of classes.

By selecting segments of users who are semantically related to the classes specified by the advertisers, we avoid considering only the users who evaluated items of that class; this allows our approach to overcome the open problems previously mentioned, related to preference stability and to the triviality of a segmentation generated by considering the evaluated items. Moreover, by defining the semantic binary sieves that characterize each class and the relevance scores that characterize each user, we avoid the interpretability issues that usually affect the user segmentation; indeed, each class of items is described by thousands of features (i.e., the words that characterize it), but this complexity is hidden to the advertiser, which is only required to specify the users she/he wants to target (e.g., those whose models are characterized by *comedy AND romantic* movies).

Considering that the evaluation of the users for the items offered in a context of e-commerce are usually thousands or millions, the proposed approach represents an efficient strategy to model in a compact way the information related to these big amounts of data.

The scientific contributions of our proposal are now recapped:

- we introduce a novel data structure, called *semantic binary sieve*, to semantically characterize each class of items;
- we present a semantic user segmentation approach based on reliable sources of information; with respect to the state-of-the-art approaches that are based on the semantic analysis of the queries issued by the users, we perform a semantic analysis on the description of the items positively evaluated by the users;
- we solve the overspecialization issues caused by preference stability by building a model for each user that considers her/him as interested in a class of items, if the items she/he evaluated are semantically related with the words that characterize that class;
- we present a boolean algebra that allows us to specify, in a simple but punctual way, the interests that the segment should cover; this algebra, along with the built models, avoids the interpretability issues that usually characterize the segmentations built with several features;
- we perform five sets of experiments on a real-world dataset, with the aim to validate our proposal by analyzing the different ways in which the classes can be combined through boolean operations. The generated segments will be evaluated by comparing them with the topic-based segmentation (as several state-of-the-art approaches do), based on the real choices of the users.

The rest of the paper is organized as follows: we first present the works in the literature related with our approach (Section 2), then we provide a background on the concepts handled by our proposal and the formal definition of the tackled problem (Section 3), we continue with the implementation details (Section 4) and the description of the performed experiments (Section 5), ending with some concluding remarks (Section 6).

2. Related work

In this section we are going to explore the main works in the literature related to the open problems highlighted in the Introduction.

Behavioral targeting. A high variety of behavioral targeting approaches has been designed by the industry and developed as working products. Google's *AdWords*[1] performs different types of targeting to present ads to users; the closest to our proposal is the "Topic targeting", in which the system groups and reaches the users interested in a specific topic. *DoubleClick*[2] is another system employed by Google that exploits features such as browser information and the monitoring of the browsing sessions. In order to reach segments that contain similar users, Facebook offers *Core Audiences*,[3] a tool that allows advertisers to target users with similar location, demographic, interests, or behaviors; in particular, the interest-based segmentation allows advertisers to choose a topic and target a segment of users interested by it. Among its user targeting strategies, Amazon offers the so-called *Interest-based ads policy*,[4] a service that detects and targets segments of users with similar interests, based on what the users purchased, visited, and by monitoring different forms of interaction with the

[1] https://support.google.com/adwords/answer/1704368?hl=en.

[2] https://www.google.com/doubleclick/.

[3] https://www.facebook.com/business/news/Core-Audiences.

[4] http://www.amazon.com/b?node=5160028011.