Control and Optimization in Applied Mathematics (COAM)
DOI. 10.30473/coam.2020.49118.1127
Vol. 4, No. 2, Autumn - Winter 2019 (39-48), ©2016 Payame Noor University, Iran

Discover Maximum Descriptive User Groups on the Social Web

Z. Abbasi[1], N. Akhoundi[2*]

[1,2]School of Mathematics and Computer Science, Damghan University, Damghan, Iran,

Received: December 12, 2019; **Accepted:** December 11, 2020.

Abstract. Product reviews in E-commerce websites such as restaurants, movies, E-commerce products, etc., are essential resources for consumers to make purchasing decisions on various items. In this paper, we model discovering groups with maximum descriptively from E-commerce website of the form $< i,u,s >$, where $i \in \mathcal{I}$ (the set of items or products), $u \in \mathcal{U}$ (the set of users) and s is the integer rating that user u has assigned to the item i. Labeled groups from user attributes are found by solving an optimization problem. The performance of the approach is examined by some experiments on real data-sets.

Keywords. Maximum descriptively, Optimization, User group discovery, Rating record.

MSC. 90C26.

* Corresponding author
zahra.abbasi@yahoo.com, akhoundi@du.ac.ir
http://mathco.journals.pnu.ac.ir

1 Introduction

Today, collaborative rating sites drive numerous decisions. For example, online shoppers rely on ratings on E-commerce websites to purchase a variety of goods. Typically, the number of ratings (such as user comments and star rating) associated with an item (a set of items) can easily achieve hundreds or thousands, thus deciding over such a huge amount of data can be cumbersome. Before making an informed decision, a user can either spend lots of time researching dozens of ratings and reviews or can be satisfied only on an average overall rating, associated with an item. There is no surprise that most users choose the second option because of lacking time. For example, Digikala is an E-commerce site that sells a large variety of products from electronic products such as mobiles, notebooks, etc., to apparel accessories. For each item, users can leave their reviews, feedback about the item, their rating, and their experience with that item. In the category browsing page, Digikala shows the number of reviews and average ratings that users assigned to that item (e.g., see the figure 1). Users should read all reviews or trust on average ratings while trying to decide which item is better to buy. Some useful reviewers' attributes such as gender, age, location, occupation, etc., can be useful for making a better decision. Analysis of such data enables innovative insights in various scenarios such as population studies [1], online recommendation [2], and targeted advertisement [3]. In this paper, we propose an algorithm that groups reviewers based on their attribute values, and the result will be shown by some description short sentences such as "Young female students rate this item 4". For this end, we define reviewer groups by users attributes, such as reviewer group $\{\langle \text{gender,female}\rangle, \langle \text{age,young}\rangle\}$. We aim to find reviewers group sets by maximum descriptively for a rating of records. We define the set of items by $\mathcal{I}$, and define the set of users (reviewers)

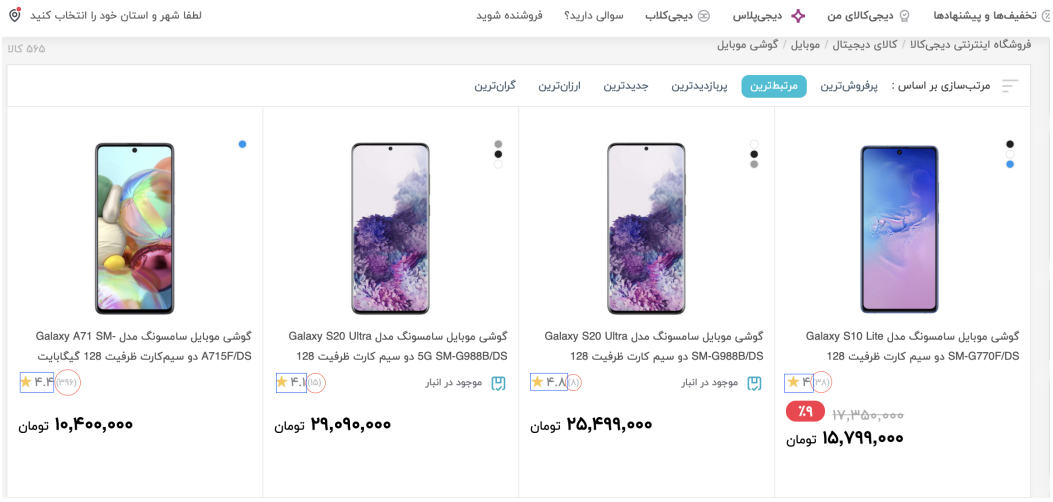

Figure 1: Mobile category browsing page, number of reviews are indicated by red circles, and the total average of ratings are indicated by blue rectangles.

by $\mathcal{U}$. For given datasets of rating records in the form $\langle i, u, s\rangle$, where $i \in \mathcal{I}$ (the set of items

Digikala is an Iranian E-commerce company based in Tehran

or products), $u \in \mathcal{U}$ (the set of users) and s is the integer rating that user u has assigned to item i. An user group is defined by conjunction of users' demographic attributes over rating records, such as *male teachers* or *young students who live in Tehran*. The problem of user group discovering is to find group set of G including user groups such that some objectives are optimized. In [5], the problem of user group discovery is modeled as the following constrained optimization:

$$\begin{array}{ll} \text{Min} & \text{error}(G), \\ \text{S.T.} & \text{covarage}(G) \geq \alpha, \\ & |G| \leq k. \end{array} \tag{1}$$

where G is taken over all user group sets. $\text{error}(G)$ is a function that computes the sum of total distance between mean scores in each group and mean scores of rating records. Function $\text{coverage}(G)$ computes the percentage of covering the rating record I, by G. In [4], in order to solve the problem of user group discovery, the constrained multi-objective optimization problem is defined as follows: for a given set of rating records R and integer constants σ and k, the problem is to identify all group-sets, such that each group-set G satisfies:

$$\begin{array}{ll} \text{Max} & \text{Coverage}(G), \\ \text{Max} & \text{Diversity}(G), \\ \text{Opt.} & \text{rDistb}(G), \\ \text{S.T.} & |G| \leq k, \\ & \forall g \in G : |g| \geq \sigma. \end{array} \tag{2}$$

where $\text{Diversity}(G)$ measures how distinct groups are in group-set G. The last constraint states that a group g should contain at least σ rating records, an application-defined threshold. Each group in G is a description of its attributes. For example, if a group G is $G = \{\langle \text{gender,female} \rangle, \langle \text{age,young} \rangle\}$, then G can be described as *young female* group. We would like to focus on discovering user groups with more accurate descriptions. To achieve this aim, our strategy is to find user groups with maximum descriptive attributes. In both optimization problems (1.1) and (1.2), the number of attributes of the user group is not considered. Sometimes, returned groups have their minimum number of attributes (only one attribute). However, groups with more attributes have better descriptions. So, it can be a good idea to find the optimal group with the maximum number of attributes. Maximizing the number attributes (maximum descriptively) should be considered as an objective of the user group discovery model. The rest of paper is organized as follows: Section 2 describes problem definitions. Section 3 encompasses the basic definitions and concepts of user group discovery problem, and our proposed algorithm. Finally some experiments are shown in section 4.

2 Problem Definitions

An E-commerce website like Digikala, comprises three main part. The first part is the set of items (the set of all products) denoted by $\mathcal{I}$. The second part is the set of users, we denote by

$\mathcal{U}$. We denote by $\mathcal{R}$ the set of rating records. It is the third part of E-commerce website. Each rating record $r \in \mathcal{R}$ is itself a triple $\langle i,u,s\rangle$, where $i \in \mathcal{I}$, $u \in \mathcal{U}$ and s is the integer rating that user u has associated to item i. In this paper this three parts are modeled as a triple $\langle \mathcal{I},\mathcal{U},\mathcal{R}\rangle$. The set of items $\mathcal{I}$ is associated with a set of attributes, denoted as $\mathcal{I}_A = \{ia_1, ia_2,\ldots\}$, where each item $i \in \mathcal{I}$ is a tuple with $\mathcal{I}_A$ as its schema. In other words, $i = \langle iv_1, iv_2,\ldots\rangle$, where each iv_j is a set of values for attribute ia_j. The schema for the reviewers is $\mathcal{U}_A = \{ua_1, ua_2,\ldots\}$, i.e., $u =< uv_1, uv_2,\ldots >\in \mathcal{U}$, where each uv_j is a value for attribute ua_j. As a result, the tuple for i, the tuple for u, and the numerical rating score s are joint by $r = \langle i,u,s\rangle$ which itself is a tuple in the form $\langle iv_1, iv_2,\ldots, uv_1, uv_2,\ldots, s\rangle$. The set of all attributes is denoted as $A = \{a_1, a_2,\ldots\}$.

Definition 1. We define a group g as a set of $\{\langle a_1, v_1\rangle, \langle a_2, v_2\rangle, \cdots\}$ where each $a_i \in A$ (set of all attributes) and each v_i is a set of values for a_i.

The set of attributes g are denoted by $A(g)$, and the number of rating records in g is denoted by $|g|$.

For example, in MovieLens datasets, the group

$$g = \{\langle \text{gender,female}\rangle, \langle \text{location,DC}\rangle, \langle \text{genre,romance}\rangle\}$$

contain rating records for romance movies whose reviewers are all female in DC. We note that $A(g) = \{\text{gender,location,genre}\}$.

Definition 2. Given a rating record $r = \langle v_1, v_2,\ldots, v_k, s\rangle$, where each v_i is a set of values for its corresponding attribute in the schema A, and a group

$$g = \{\langle a_1, v_1\rangle, \langle a_2, v_2\rangle,\ldots, \langle a_n, v_n\rangle\}, n \le k,$$

we say that g covers r, and denote by $r < g$, if and only if $\forall i \in [1,n], \exists r \cdot v_j$ such that $r \cdot v_j$ is a subset of values for attribute $g \cdot a_i$ i.e., $r \cdot v_j \subset g \cdot v_i$.

For example, based on Definition 2, the rating $\langle$female,DC,student,4$\rangle$ is covered by the group $\{\langle$gender,female$\rangle, \langle$location,DC$\rangle\}$. The set of all possible groups form a lattice where nodes correspond to groups and edges correspond to parent/child and ancestor/descendant relationships. In Figure 2 a partial lattice for rating records of the movie *Toy Story*(1995) is illustrated. We have four reviewer attributes **gender, age, location** (CA stands for California) and **occupation** to analysis. For simplicity, exactly one distinct value per attribute is shown. The complete lattice contains 15582 attribute-value combinations, see for example [4].

Definition 3. We say that two groups g_1 and g_2 are similar and denoted by $g_1 \sim g_2$, if and only if $A(g_1) = A(g_2)$.

We have the following two lemmas, proof of them is straightforward, so we omit them.

Lemma 1. Let g_1 and g_2 be two groups, define $B_1 = \{r \in R | r < g_1\}, B_2 = \{r \in R | r < g_2\}$. If $g_2 \subset g_1$, then we have $B_1 \subseteq B_2$.

Lemma 2. Let g_1 and g_2 be two groups and h is some arbitrary group differ from g_1 and g_2. Define $C_1 = \{r \in R | r < h \wedge r < g_1\}$, $C_2 = \{r \in R | r < h \wedge r < g_2\}$. If $g_2 \subset g_1$ then $C_1 \subseteq C_2$.

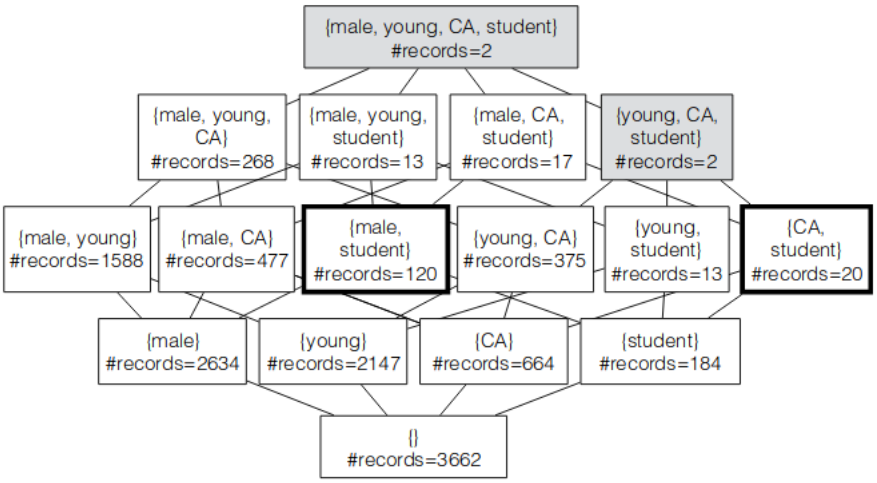

Figure 2: Partial lattice for movie *Toy Story*.

Before formalizing the mining problem, quality dimensions should be defined for groups. For a set of rating records $R \subseteq \mathcal{R}$ and a group-set G, the percentage of rating records in R contained in groups in G is measured by a quality dimension called coverage. Coverage is a value between 0 and 1 and it is defined as follows.

$$\text{coverage}(G,R) = \frac{|\cup_{g\in G}\{r \in R, r < g\}|}{|R|}. \tag{3}$$

Another quality dimension is called diversity. Diversity of G is a value between 0 and 1 that measures how distinct groups in group-set G are from each other, is defined as follows

$$\text{diversity}(G,R) = \frac{1}{1 + \sum\limits_{g_1,g_2\in G} |\{r \in R, r < g_1 \wedge r < g_2\}|}. \tag{4}$$

For a group set G, we define the number of attributes as following,

$$\text{attributes}(G) = |\cup_{g\in G} A(g)|, \tag{5}$$

for example number of attributes $G = \{g_1, g_2\}$ where

$$g_1 = \{\langle \text{gender,male} \rangle\}, g_2 = \{\langle \text{gender,male} \rangle, \langle \text{occupation,student} \rangle\},$$

is

$$|\{\text{gender}\} \cup \{\text{gender,occupation}\}| = 2.$$

Group sets that have more attributes provide users with more information to make their decisions.

3 Maximum Description Optimization

We define our constrained optimization problem as follows: for a given set of rating records R, the problem is to identify all group sets, such that each group set satisfies:

$$\begin{array}{ll} \text{Max} & \text{attributes}(G), \\ \text{S.T.} & \text{coverage}(G,R) \geq \alpha \\ & \text{diversity}(G,R) \geq \beta \\ & |G| \geq k. \end{array} \tag{6}$$

Definition 4. Let g is a group and G is a group set, we say $g < G$, if and only if $\forall h \in G, h \nprec g$, and there are $\tilde{g} \in G$ such that $\tilde{g} \subset g$. we denote by $G_g^{-\tilde{g}}$ a group set that was constructed by replacing g with $\tilde{g}$ in G, i.e., $G_g^{-\tilde{g}} = G - \{\tilde{g}\} \cup \{g\}$.

Theorem 1. If a group set G has the following two properties

$$\forall g_1, g_2 \in G, \quad g_1 \nprec g_2, \tag{7}$$

$$\text{There is no two groups,} \quad g_1, g_2 \in G \quad \text{such that} \quad g_1 \subset g_2, \tag{8}$$

then for some group g such that $g < G$, the following statements are holds.

1. $\text{attributes}(G_g^{-\tilde{g}}) > \text{attributes}(G)$.
2. $\text{coverage}(G_g^{-\tilde{g}}) \leq \text{coverage}(G)$.
3. $\text{diversity}(G_g^{-\tilde{g}}) \geq \text{diversity}(G)$.

Proof. By definition 4 there exist, $\tilde{g} \in G$ such that $\tilde{g} \subset g$. We have $A(\tilde{g}) \subset A(g)$, hence $\bigcup_{h \in G} A(h) \subset \bigcup_{h \in G_g^{-\tilde{g}}} A(h)$ hence

$$\text{attributes}(G_g^{-\tilde{g}}) = |\bigcup_{h \in G_g^{-\tilde{g}}} A(h)| > |\bigcup_{h \in G} A(h)| = \text{attributes}(G).$$

This is complete the proof of the first part. For section 2, by lemma 1 $\bigcup_{h \in G} \{r \in R | r < h\} \subseteq \bigcup_{h \in G_g^{-\tilde{g}}} \{r \in R | r < h\}$ hence $\text{coverage}(G_g^{-\tilde{g}}) \leq \text{coverage}(G)$. Lastly by lemma 2 we have

$$\sum_{g_1, g_2 \in G_g^{-\tilde{g}}} |\{r \in R, r < g_1 \wedge r < g_2\}| \leq \sum_{g_1, g_2 \in G} |\{r \in R, r < g_1 \wedge r < g_2\}|,$$

hence

$$\begin{aligned} \text{diversity}(G_g^{-\tilde{g}}) &= \frac{1}{1 + \sum_{g_1, g_2 \in G_g^{-\tilde{g}}} |\{r \in R, r < g_1 \wedge r < g_2\}|} \\ &\geq \frac{1}{1 + \sum_{g_1, g_2 \in G} |\{r \in R, r < g_1 \wedge r < g_2\}|} = \text{diversity}(G). \end{aligned}$$

□

Remark 1. The group sets that contain one group with one attribute satisfies in (7) and (8).

Based on theorem 1, we can develop an algorithm with smart search to find local maximum of optimization problem (6). For example in partial lattice of Toy Story movie (see figure 2), we can set $G = \{g_1\}$ as an initial solution, where $g_1 = \{\langle \text{gender,male} \rangle\}$ (g_1 is a group with one attributes and maximum coverage). Three parent nodes of g_1, are as following:

- $h_1 = \{\langle \text{gender,male} \rangle, \langle \text{age,young} \rangle\}$
- $h_2 = \{\langle \text{gender,male} \rangle, \langle \text{location,CA} \rangle\}$
- $h_3 = \{\langle \text{gender,male} \rangle, \langle \text{occuption,student} \rangle\}$.

We see that $h_i < G$ for $i = 1, 2, 3$. Hence based on Theorem 1, for each h_i where $\text{coverage}(G_{h_i}^{-g_1}) \geq \alpha$, new group set $G_{h_i}^{-g_1}$ is better than G. If there is no such group, then G is a local optimal solution of optimization problem 6, otherwise we can choose $h_k = \arg\max_i \{\text{coverage}(G_{h_i}^{-g_1})\}$ and substitute G with a better solution $G_{h_k}^{-g_1}$.

3.1 Algorithm

The algorithm is started with initial groups that have one attribute. These groups are chosen based on the best coverage. Let the groups by one attribute (penultimate level of the lattice, e.g., see Fig. 2) are ordered as follows

$$\text{coverage}(g_1) \geq \text{coverage}(g_2) \geq \ldots \geq \text{coverage}(g_n). \tag{9}$$

Based on Theorem 1, we search in parent lattice of groups to increase the number of attributes as long as coverage condition is satisfied. Our algorithm is described in details in **Algorithm 1**. The pseudo-code of algorithm 1 works as follows: In line 1, the parameters of α, m are given. In line 3, the initial group sets are generated as defined in (9). In lines 4-15, the search step is performed over the m initial group sets to find the local optimal solution. We know that in group set G, if we let h is a parent of some group $g \in G$, then it is easy to show that $h < G$. The search procedure is based on parent searching because of several reasons. The first reason, since the initial group set G satisfies the requirements of (7) and (8), and h is a parent of $g \in G$, the group set G_h^{-g} satisfies these requirements too. Secondly based on Theorem 1, we see that $attributes(G_h^{-g}) > attributes(G)$, so if $\text{coverage}(G_h^{-g}) \geq \alpha$ (Line 9 algorithm), then G_h^{-g} is better solution than G. Finally, we can describe the parent-based search procedure in detail, as follows. The search procedure was performed for each initial solution in the for-loop in line 4. For each initial solution, until convergence occur, we find a group h in parent lattice-based of $G_i^{(k)}$ in line 8 that satisfies in $\text{coverage}(G_h^{-g}) \geq \alpha$. If there is no such group, then $G_i^{(k)}$ is a local optimal solution, and in line 14 we add it into the local optimal solution set G_{opt}. Otherwise the solution $G_i^{(k)}$ is replaced by the current better solution G_h^{-g}. Finally, in line 16, the best local optimal solution inside G_{opt} is selected and return in line 17.

Algorithm 1 Lattice search algorithm

Data: α, m
$G_{opt} \leftarrow \emptyset$;
Initialization :
For $i = 1, \ldots, m$ Choose initial group sets $G_i^{(0)} = \{g_i\}$ (as defined in (12));
Search Step :
for $i = 1, ..., m$ **do**
  for $k = 0, 1, 2, 3, ...$ **do**
    $G = G_i^{(k)}$;
    opt = true;
    for $g \in G$ *and* $\forall h$ *in parent lattice-based of* g **do**
      if $coverage(G_h^{-g}) \geq \alpha$ **then**
        `// Replace` $G_i^{(k)}$ `with better group set` G_h^{-g}
        $G_i^{(k+1)} \leftarrow G_h^{-g}$;
        `//` G `isn't local optimal solution`
        opt = false;
        break;
    `// Check if` $G_i^{(k)}$ `is optimal solution`
    if *opt* == *true* **then**
      $G_{opt}.add(G_i^{(k)}, attributes(G_i^{(k)}))$;
      break;
let $(G', attributes(G'))$ be the pair with maximum number attributes in G_{opt};
return G';

4 Experiments

Real datasets, MovieLens, have been used for our experiments. For each user, gender, age-group, occupation, and zip code are provided. The MovieLens 1M datasets contain 100000 ratings of 3952 movies by 6040 users. The attribute of gender takes two distinct values: male or female. The numeric age is converted into categorical attribute values, namely teen-aged, young, middle-aged, and old. 21 occupations such as student, doctor, lawyer, etc are also listed. Finally, zip codes are converted into the USA states (http://zip.usps.com). Thus, 52 distinct values can be taken for the attribute location [3]. Five items are selected randomly and then, the groups are provided by our algorithm (Table 1) that we assume $\alpha = 0.8, \beta = 0.8, k = m = 2$ and DEM method [5] (Table 2). In Table 1-2 the column **Cov**, **Natt**, and **Div** denote coverage, number attributes, and diversity respectively. The algorithm was written in PHP and Laravel. The algorithm is freely available as a Laravel package in https://github.com/NARooshnavand/user-group-discovery.

Table 1: Our Algorithm

Id	Cov	Natt	Div	Optimal group set
73	0.804	4	1	First group={ young student women in California} Second group={men}
200	0.801	3	1	First group={ young student women} Second group={men}
500	0.806	3	1	First group={ young student women} Second group={men}
600	1	4	1	First group={ old administer men in California} Second group={old educator men in Seattle }
821	0.818	4	1	First group={ old educator men in Texas } Second group={ young }

Table 2: DEM Algorithm

Id	Cov	Natt	Div	Optimal group set
73	0.812	4	1	First group={young women in California} Second group={men}
200	0.908	2	1	First group={men} Second group={young women}
500	1	1	1	First group={men} Second group={women}
600	1	4	1	First group={old administer men in California} Second group={old educator men in Seattle}
821	0.812	3	1	First group={middle-aged men} Second group={young}

References

[1] Amer-Yahia, S., Kleisarchaki, S., Kolloju, N. K., Lakshmanan, L. V., Zamar, R. H. (2017). "Exploring rated datasets with rating maps", In Proceedings of the 26th International Conference on World Wide Web (pp. 1411-1419).

[2] Amer-Yahia, S., Omidvar-Tehrani, B., Basu, S., Shabib, N. (2015). "Group recommendation with temporal affinities", EDBT.

[3] Omidvar-Tehrani, B., Amer-Yahia, S., Termier, A. (2015). "Interactive user group analysis", In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 403-412).

[4] Omidvar-Tehrani B., Amer-Yahia S., Dutot P.F., Trystram D. (2016). "Multi-Objective Group Discovery on the Social Web" ECML PKDD 2016, part I,LNAI 9851, 296-312

[5] Das, M., Amer-Yahia, S., Das, G., Yu, C. (2011). “Mri: Meaningful interpretations of collaborative ratings”, Proceedings of the VLDB Endowment, Vol. 4, No. 11.

کشف گروه‌های کاربری با بیشترین توصیف‌پذیری در شبکه‌های اجتماعی

عباسی، ز.
ایران- دامغان – دانشگاه دامغان- دانشکده ریاضی و علوم کامپیوتر،
zahra.abbasi@yahoo.com

آخوندی روشناوند، ن. - نویسنده مسئول
ایران – دامغان – دانشگاه دامغان- دانشکده ریاضی و علوم کامپیوتر
akhoundi@du.ac.ir

تاریخ دریافت: ۲۱ آذر ۱۳۹۸ تاریخ پذیرش: ۲۱ آذر ۱۳۹۹

چکیده

نظرات و قضاوت کاربران در وب‌سایت‌های تجارت الکترونیک مانند وب‌سایت‌های رستوران‌ها، فروش فیلم، فروش محصولات و غیره، برای مشتری‌ها در خصوص اخذ تصمیم خرید بسیار اهمیت دارند. در این مقاله، ما کشف گروه‌های کاربری با بیشترین توصیف‌پذیری را از وب‌سایت‌های تجارت الکترونیک به صورت $\langle i,u,s\rangle$، که $i \in I$ (مجموعه آیتم‌های محصولات)، $u \in U$ (مجموعه کاربران) و s عدد صحیح که امتیاز کاربر u به آیتم i اختصاص داده است. گروه‌های برچسب‌دار از صفات کاربری با حل مساله بهینه‌سازی به دست می‌آیند. کارایی روش با برخی آزمون‌ها بر روی مجموعه داده‌های واقعی مورد ارزیابی قرار می‌گیرد.

کلمات کلیدی

بیشترین توصیف‌پذیری، بهینه‌سازی، کشف گروه کاربری، داده‌های ارزیابی.