National Seminar on
Control and Optimization
Nov 13-14, 2019
3th
سومین سمینار ملی
کنترل و بهینه‌سازی
۲۳-۲۲ آبان ۱۳۹۸

DISCOVER THE MAXIMUM DESCRIPTIVE USER GROUPS ON THE SOCIAL WEB

NASSER AKHOUNDI, AND ZAHRA ABBASI*

School of mathematics and computer science, Damghan university, Damghan, Iran
akhoundi@du.ac.ir; zahra.abbasi@yahoo.com

Abstract. Collaborative content sites have become essential resources for consumers to consult in order to make purchasing decisions on various items such as restaurants, movies, e-commerce products, etc. In this paper, we model discovering groups with maximum descriptively from collaborative social contents of the form $\langle i, u, s\rangle$, where $i \in \mathcal{I}$, $u \in \mathcal{U}$ and s is the integer rating that user u has assigned to item i. Labeled groups from user's attributes are found by solving a multi- objective optimization problem. The performance of the approach is examined by some experiments on real data-sets.

1. Introduction

A collaborative rating site $\mathcal{D}$ is modeled as a triple $\langle \mathcal{I}, \mathcal{U}, \mathcal{R}\rangle$, representing the sets of items, reviewers, and rating records respectively. 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. $\mathcal{I}$ is associated with a set of attributes, denoted as $\mathcal{I}_A = \{ia_1, ia_2, \ldots\}$, and 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., $u = \langle uv_1, uv_2, \ldots\rangle \in \mathcal{U}$, 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 1.1. a group g is a set of rating records $r = \langle u, i, s\rangle$ described by a set of attribute value pairs shared among the reviewers and the items of those rating records. The description of a group g is defined as $\{\langle a_1, v_1\rangle, \langle a_2, v_2\rangle, \cdots\}$ where each $a_i \in A$ (set of all attributes) and each v_i is a set of values for a_i.

2010 *Mathematics Subject Classification.* 46N10.
Key words and phrases. maximum descriptively, optimization, user group discovery, rating record.
* Speaker.

We denote the set of attributes g by $A(g)$, and the number of rating records contained in g is denoted by $|g|$. For example, in MOVIELENS datasets, the group $g = \{\langle gender, female\rangle, \langle location, DC\rangle, \langle genre, romance\rangle\}$ contains raing records for romance movies whose reviewers are all female in DC. We note that $A(g) = \{gender, location, genre\}$.

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 $g = \{\langle a_1, v_1\rangle, \langle a_2, v_2\rangle, \ldots, \langle a_n, v_n\rangle\}, n \geq k$, we say that g covers r, denoted as $r < g$, iff $\forall i \in [1, n], \exists r.v_j$ such that v_j is a set of values for attribute $g.a_i$ and $r.v_i \subseteq g.v_j$.

Definition 1.2. For two groups g_1 and g_2, we say that

(1) $g_1 < g_2$ iff $\forall r \in R$ such that $r < g_1$ we have $r < g_2$.
(2) g_1 similar g_2 denoted as $g_1 \sim g_2$ iff $A(g_1) = A(g_2)$.

For a group g and group set G, we say that $g < G$, iff $\forall h \in G, h \nsim g$, and $\exists \tilde{g} \in G$ such that $g < \tilde{g}$. We denote by $G_g^{-\tilde{g}}$ a group set that was constructed by replacing g with $\tilde{g}$ in G, i.e., $G_g^{-\tilde{g}} = G - \{\tilde{g}\} \cup \{g\}$.

For a given 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 called coverage, and it is defined as follow [1].

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

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 [1]

$$diversity(G, R) = \frac{1}{(1 + \sum_{g_1, g_2 \in G} |r \in R, r < g_1 \wedge r < g_2|)}.$$

We interest that number attributes of group set is maximal. We define number attributes of group set G as following,

$$Nattrs(G) = |\cup_{g \in G} A(g)|.$$

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

2. Maximum description optimization

We define our constraint 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:

- $Nattrs(G)$ is maximized;
- $coverage(G, R) \geq \alpha$;
- $diversity(G, R) \geq \beta$
- $|G| \leq k$.

In our model, we assume that a group set G satisfies two conditions. First $\forall g_1, g_2 \in G, g_1 \nsim g_2$, i.e., group set G does not contain two similar groups. Second there is no two groups, $g_1, g_2 \in G$ such that $g_1 < g_2$, which means that G has groups with maximal attributes.

Theorem 2.1. *The decision version of our problem is NP-Complete* [2].

Our algorithm is described in details in **Algorithm 1**.

Algorithm 1: Maximum description algorithm

Data: α, β, k, m
$G_{opt} \leftarrow \emptyset$
Initialization :
For $i = 1, \ldots, m$ Choose initial group sets $G_i^{(0)} = \{g_i\}$ where $coverage(g_1) \geq coverage(g_2) \geq \ldots \geq coverage(g_n)$.
Search Step :
for $i = 1, ..., m$ **do**
  for $k = 0, 1, 2, 3, ...$ **do**
    $G = G_i^{(k)}$
    opt = false
    for $h \in G$ *and* $\forall g$ *in lattice-based of* h **do**
      if $g < h$ *and* $coverage(G_g^{-\tilde{g}}) \geq \alpha$ *and* $diversity(G_g^{-\tilde{g}}) \geq \beta$ **then**
        $G_i^{(k+1)} \leftarrow G_g^{-\tilde{g}}$
        opt = true
        break
      else if $|G \cup \{g\}| \geq k$ *and* $coverage(G \cup \{g\}) \geq \alpha$ *and* $diversity(G \cup \{g\}) \geq \beta$ **then**
        $G_i^{(k+1)} \leftarrow G \cup \{g\}$
        opt = true
        break
    end
    if $opt == false$ **then**
      $G_{opt}.add(G_i^{(k)}, Nattrs(G_i^{(k)})$
      break
    end
  end
end
let $(G', Nattrs(G'))$ be the pair with maximum number attributes in G_{opt}
return G';

3. Experiments

Real datasets, MovieLens, has been used for our experiments. For each user, gender, age-group, occupation and zip code are provided. The MovieLens 1M datasets contains 100000 rating of 3952 movies by 6040 users. The attribute gender takes two distinct values: male or female. The numeric age is converted into categorical attribute value, namely teen-aged, young, middle-age 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 [2] (Table 2). In Table 1-2 the column **Cov**, **Natt**, and **Div** denote coverage, number attributes, and diversity respectively.

TABLE 1. Maximum description method

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. Formal programming

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. B. Omidvar-Tehrani, S. Amer-Yahia, P.F. Dutot, and D. Trystram, *Multi-objective group discovery on the social web*. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases Springer, Cham (2016) 296–312.
2. M. Das, S. Amer-Yahia, G. Das, and C. Yu, *MRI: Meaningful interpretations of collaborative ratings*, Proceeding of the VLDB Endowment, **4** (2011) 1063–1074.