基于区间型符号数据的群组推荐算法研究*

郭均鹏，宁 静，史志奇
（天津大学 管理与经济学部，天津 300072）

摘 要： 传统群组推荐算法基于点数据描述群组用户模型，存在着信息缺失、很难统筹考虑所有个体用户的需求等问题。针对该问题，对个体评分数据按照符号数据分析的思想进行“打包”，将群组成员的评分信息汇总为区间型符号数据。在 Hausdorff 距离基础上，采用区间内部点数据的描述统计量，提出了一种全新的区间数距离度量方法，并利用这种距离对区间型符号数据描述的群组实施 K-均值聚类，由此确定相似群组，最后通过最近邻的评分预测目标群组的评分。将这种全新的群组推荐算法与传统方法进行推荐精度与效率的对比实验，结果表明，在各种实验条件下，基于区间型符号数据的群组推荐算法均优于传统点数据的群组推荐算法。

关键词： 群组推荐；符号数据分析；聚类分析
中图分类号： TP391；TP301.6 **文献标志码：** A **文章编号：** 1001-3695(2013)01-0067-05
doi:10.3969/j.issn.1001-3695.2013.01.016

Group recommendation algorithm based on symbolic data analysis

GUO Jun-peng, NING Jing, SHI Zhi-qi
(*Dept. of Management & Economics, Tianjin University, Tianjin* 300072, *China*)

Abstract: The group user profile in traditional group recommendation is described by single-valued data. This results in the loss of data information and being difficult to meet the demands of all the memebers of the group. Aimed at this problem, this paper took the method of symbolic data analysis aggregating individual ratings of the group into interval symbolic data into account. It proposed a novel distance considering the descriptive statistics of individuals within the intervals. Based on the K-means clustering on the interval data of group ratings, it obtained the similar groups. Then it predicted the ratings of the target group by using the neighbors' ratings. It conducted a simulation study to evaluate the new method. The result shows that the new method based on interval symbolic data analysis is more accurate and efficient than the traditional item-based collaborative filtering algorithms for group recommendation.

Key words: group recommendation; symbolic data analysis(SDA); cluster analysis

0 引言

推荐系统是通过了解用户的喜好并向用户推荐符合其兴趣爱好对象的一种信息技术。作为一种重要的信息过滤技术，推荐系统已经成为各大网站不可或缺的个性化信息服务形式。相对于个体推荐，群组推荐的对象是一个由多个不同成员组成的群体，他们拥有某些共同的兴趣偏好或需求，如商业组织、爱好者协会等。如何为这些群体用户提供有效的个性化信息服务，带来了对群体推荐系统的极大需求。

近年来陆续有学者对群体推荐系统展开研究。国内学者对群组推荐算法方面的研究较少，文献[1]应用离子群算法构建领域项目分类模型，在此基础上建立群组用户的兴趣模型，然后采用最近邻方法完成推荐。国外研究群组推荐算法大多通过某种方法把群组视为个体，然后应用个体推荐的算法对群组进行推荐。文献[2]的 TV4M 系统通过计算各群组与目标群组的距离来确定相似群组，其中群组对推荐项目的矩阵用特征子集来表示。文献[3,4]在将个体特征表整合成群组特征表的过程中引入遗传算法，在考虑群组中个体间存在相互作用的同时提高了算法的精确度。文献[5]首先对群组中的个体进行推荐，然后将单个的推荐结果转换为对整个群组的推荐结果，最大限度地提高了群组个体的满意度，但耗时较多且效率较低。文献[6]通过将群组中喜好相同的用户合为一个用户对评分矩阵降维，同时将个体推荐结果整合为群组推荐结果，使推荐更精确有效。文献[7]将群组中个体的合作性因素和社会信任因素运用到群组推荐算法中，并证明当群组中合作性和信任度较高时推荐结果较为精确。

上述研究均基于点数据描述群体用户模型，进而采用传统的协同过滤等算法产生推荐，因而不可避免地存在着数据信息丢失、很难统筹考虑所有个体用户的需求等问题。符号数据分析(SDA)通过数据打包等技术，在不丢失信息的前提下，从全局上把握数据特征[8,9]，为群体推荐系统提供了一个全新的解决思路。目前，国外陆续有学者在符号数据推荐算法这方面展开了研究。文献[10]基于符号数据分析研究了群体推荐算法，并用分布式符号数据来表示群组原型和目标项目原型，研究结论表明了符号数据分析在群体推荐算法研究中的优势。文献[11]将符号数据分析运用于基于内容的推荐方法，利用分布式符号数据来表示项目，将用户评价过的项目信息汇总起来表示用户的兴趣模型，并定义了一种适用于符号数据的相似

收稿日期：2012-05-19；修回日期：2012-07-04 基金项目：国家自然科学基金资助项目(70701026,71271147)
作者简介：郭均鹏(1973-)，男，山东潍坊人，教授，博士，主要研究方向为数据分析、管理科学(guojp@tju.edu.cn)；宁静(1987-)，女，山西太原人，硕士，主要研究方向为数据分析、个性化推荐；史志奇(1985-)，男，江苏徐州人，硕士，主要研究方向为数据分析、个性化推荐.

性度量方法来计算项目与用户兴趣之间的相关性。文献[12]系统地介绍了基于符号数据分析的内容过滤推荐、协同过滤推荐和混合推荐算法,研究了如何通过符号数据分析的数据打包技术来建立用户的兴趣模型,并就各种推荐模式分别研究了相似性度量方法,进而产生推荐。所有这些文献,开创了将符号数据分析方法运用于推荐系统方案解决的研究,具有重要的理论和实践意义。但是研究内容主要为分布式符号数据在推荐系统中的应用,而对于区间型这种常用类型的符号数据并未涉及。

本文在现有研究的基础上,提出一般分布区间型符号数据的距离度量,以此为基础将区间型符号数据的 K-均值聚类分析应用到群组推荐中,为群组推荐算法提供了一种全新的解决方法。

1 一般分布区间型符号数据的距离——$\mu\sigma$ 距离

1.1 一般分布区间型符号数据的定义

定义 1[13] 若随机变量 X 服从某任意分布,且其观测值在 $[a,b]$ 取值,则称 X 为一般分布的区间型符号变量,简称一般分布区间变量或区间变量; $[a,b]$ 称为一般分布区间型符号数据,简称一般分布区间数。若 X 在 $[a,b]$ 服从均匀分布,则称 $[a,b]$ 为均匀分布区间数。

1.2 一般分布区间型符号数据的距离

本节将基于对传统 Hausdorff 距离的改进,提出了一般分布区间型符号数据的新距离。设 $A=[\underline{a},\overline{a}]$ 和 $B=[\underline{b},\overline{b}]$ 为两个区间数,可将其视为两个紧集,则 A、B 间的 Hausdorff 距离[14]为

$$H(A,B)=\max\{|\overline{a}-\overline{b}|,|\underline{a}-\underline{b}|\}=|c(A)-c(B)|+|r(A)-r(B)| \tag{1}$$

其中: $c(X)=\dfrac{(\underline{x}+\overline{x})}{2}$ 为区间数 $X=[\underline{x},\overline{x}]$ 的中点, $r(X)=\dfrac{(\overline{x}-\underline{x})}{2}$ 为区间数 $X=[\underline{x},\overline{x}]$ 的半径,此处 $X=A$ 或 B。显然,当两个区间数退化成两个实数时,式(1)就是这两个实数的绝对值距离。

对于均匀分布的区间数而言,式(1)中的中点 $c(X)$ 描述了区间数的集中位置,半径 $r(X)$ 描述了区间数的离散程度。而对于非均匀分布的区间数来说,中点和半径就无法准确描述区间数的集中位置和离散程度,此时当区间数内部点数据可获得时,对于一般分布的区间数可用样本均值表示集中位置、样本标准差描述离散程度。因此,对经典的 Hausdorff 距离加以改进,可以使之适用于一般分布的区间数。

定义 2 设 A、B 为任意两个一般分布区间数,当区间数内部点数据可得或其服从分布已知时, A、B 间的 $\mu\sigma$ 距离可定义为

$$D(A,B)=|\overline{A}-\overline{B}|+\sqrt{3}\,|S_A-S_B| \tag{2}$$

其中: $\overline{A}$、$\overline{B}$分别表示区间 A、B 内点数据的均值; S_A、S_B 分别表示区间 A、B 内点数据的标准差。式(2)将标准差之差的系数设定为$\sqrt{3}$,是因为当区间数 A 和 B 内的点数据服从均匀分布(或仅知道区间数端点值)时:

$$\overline{X}=\frac{\underline{x}+\overline{x}}{2},S_X=\frac{\overline{x}-\underline{x}}{2\sqrt{3}}=\frac{r(X)}{\sqrt{3}} \tag{3}$$

则式(2)退化为区间数的 Hausdorff 距离,即式(1)。

易证得式(2)满足距离的非负性、对称性和三角不等性这三个基本条件。

对于区间向量的距离,假设

$$X=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}=([a_1,b_1],[a_2,b_2],\cdots,[a_p,b_p])^{\mathrm{T}}$$

$$Y=(y_1,y_2,\cdots,y_p)^{\mathrm{T}}=([c_1,d_1],[c_2,d_2],\cdots,[c_p,d_p])^{\mathrm{T}}$$

为两个 p 维的区间向量,将实数空间中的欧氏距离进行拓广,可以得到两个区间向量之间的距离定义:

$$d(X,Y)=\sqrt{\sum_{i=1}^{p}d(x_i,y_i)^2}=\sqrt{\sum_{i=1}^{p}[|\overline{x_i}-\overline{y_i}|+\sqrt{3}\,|S_{x_i}-S_{y_i}|]^2} \tag{4}$$

例 1 下面给出一个算例,分别计算 $\mu\sigma$ 距离和均匀分布假设下的 Hausdorff 距离,以对一般分布区间数的两类距离进行比较。给定三个区间数 $A=[0,4]$, $B_1=[2,6]$ 和 $B_2=[-2,2]$,其中 A 的内部点数据为{0,2,3,3,3,4,4}, B_1 和 B_2 服从均匀分布,包含的点数据分别为{2,3,4,5,6}和{-2,-1,0,1,2}。

通过图形角度可以更直观地表示出区间数 A、B_1、B_2 的位置关系。如图 1 所示,图中用不规则矩形表示区间数,矩形的左端表示区间的下限,右端表示区间的上限,高表示区间内部点数据的密度函数,即矩形的宽表示区间的范围,高表示区间内部点数据的分布情况,因此区间数 A 表示右端突起的不规则矩形, B_1 和 B_2 表示与 A 同宽的规则矩形。

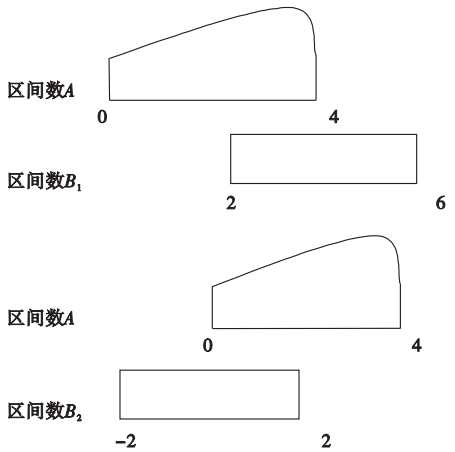

图1 区间数A、B_1、B_2的位置关系

从图 1 可以看出,区间数 A 内部点数据服从左偏态分布,使得偏向于 A 右侧的区间数 B_1 更靠近区间数 A 点数据密度较大的一侧,而偏向于 A 左侧的区间数 B_2 更靠近区间 A 点数据密度较小的一侧,因此区间数 B_1 到区间数 A 的距离应小于区间数 B_2 到区间数 A 的距离。

然而由式(1)计算区间数的 Hausdorff 距离,得 $D_H(A,B_1)=D_H(A,B_2)=2$;而由式(2)计算区间数的 $\mu\sigma$ 距离,得 $D_S(A,B_1)=1.6339<D_S(A,B_2)=3.0624$。计算结果表明,使用内部点数据信息的 $\mu\sigma$ 距离,会使计算出的距离更加符合实际情况。

由例 1 可知,对于一般分布区间数而言,基于均匀分布假设的 Hausdorff 距离忽略了区间内部点数据的信息,导致计算结果与现实情况相悖;而 $\mu\sigma$ 距离利用了区间内部点数据的描述统计量,会使计算结果更加有效可靠。

2　基于区间型符号数据的群组推荐算法

本文在一般分布区间型符号数据的 $\mu\sigma$ 距离基础上，对传统 K-均值聚类算法进行扩展，将区间型符号数据的 K-均值聚类算法运用于群组推荐算法中。该算法主要包括三个步骤: a) 用区间型符号数据描述群组对项目的评分; b) 确定目标群组的最近邻; c) 为目标群组预测评分并进行推荐。本章将分别对这三个步骤进行详细介绍。

2.1　群组评分的区间型符号数据描述

令群组 k 中的个体对项目 m 进行评分。假设个体 i 对项目 m 的评分为群组中的最低分 $a_{k,m}$，个体 j 对项目 m 的评分为群组中的最高分 $b_{k,m}$，此时可以得到群组 k 对项目 m 评分的区间型符号数据描述 $x_{k,m}$，即

$$x_{k,m} = [a_{k,m}, b_{k,m}] \tag{5}$$

设群组总体 R 中群组个数为 N，则群组总体对项目 m 的评分区间型符号数据描述可表示为

$$X = (x_{1,m}, x_{2,m}, \cdots, x_{N,m}) = ([a_{1,m}, b_{1,m}], [a_{2,m}, b_{2,m}], \cdots, [a_{N,m}, b_{N,m}]) \tag{6}$$

对项目总体中任意项目 m，群组总体中每个群组对其均有唯一的符号数据表示，则群组总体(个数为 N) 对项目总体(个数为 M) 的符号数据矩阵表示为

$$\begin{bmatrix} [a_{1,1}, b_{1,1}] \cdots [a_{1,M}, b_{1,M}] \\ [a_{N,1}, b_{N,1}] \cdots [a_{N,M}, b_{N,M}] \end{bmatrix}$$

以群组结伴看电影为例，设有三个群组 group 1、group 2 和 group 3，他们分别由四位成员组成。现令群组成员分别对电影 F_1、F_2、F_3、F_4 评分，评分等级为 1 ~ 5，评分越高，表示用户对电影越喜爱，∅表示用户对电影没有评分，三个群组对四部电影的评分数据如表 1 所示。

表 1　群组成员评分

群组	群组用户	电影			
		F_1	F_2	F_3	F_4
group 1	A_1	5	∅	3	2
	B_1	3	4	1	∅
	C_1	2	5	2	∅
	D_1	4	3	∅	4
group 2	A_2	1	∅	4	2
	B_2	2	3	5	∅
	C_2	3	5	∅	3
	D_2	2	4	3	4
group 3	A_3	∅	1	3	3
	B_3	5	3	∅	5
	C_3	3	4	1	∅
	D_3	4	2	∅	4

以 group 1 为例，在群组成员对电影 F_1 的评分中，最高分为 A_1 的评分 5，最低分为 C_1 的评分 2，因此按照式(5) ，群组 group 1 对电影 F_1 评分的符号数据描述表示为 [2,5]。依此类推，可以确定每一个群组对每部电影评分的符号数据描述，如表 2 所示。

表 2　群组整体对项目评分的区间型符号数据描述

群组	电影			
	F_1	F_2	F_3	F_4
group 1	[2,5]	[3,5]	[1,3]	[2,4]
group 2	[1,3]	[3,5]	[3,5]	[2,4]
group 3	[3,5]	[1,4]	[1,3]	[3,5]

2.2　确定目标群组的最近邻

本文对目标群组最近邻的选择标准以群组间的距离作为评判依据，即群组间距离越小，相似度越大。因此，在确定目标群组最近邻时，需要先计算群组间的距离，然后依据距离大小来确定目标群组的最近邻。

在计算群组间距离时，任取项目总体中的某一项目 m，按照区间型符号数据的 K-均值聚类分析方法，对表 2 中的数据实施聚类分析，便可得到与目标群组 k 关于项目 m 处于同一聚类的相似群组集合 P_m。然后按照这种方法依次得到目标群组 k 关于每一个项目的相似群组集合，共同形成目标群组 k 关于所有项目的相似群组集合$\{P_m\}(m = 1, \cdots, M)$。设某一群组 c 在此相似群组集合中，由式(4) 可得群组 c 关于项目总体到目标群组 k 的距离，如式(7) 所示。

$$D(k,c) = \sqrt{\sum_{m=1}^{M} D_m(k,c)^2} = \sqrt{\sum_{m=1}^{M} (|X_{k,m} - X_{c,m}| + \sqrt{3}\,|S_{k,m} - S_{c,m}|)^2} \tag{7}$$

其中: $X_{k,m}$ 和 $X_{c,m}$ 分别表示群组 k 和 c 内成员对项目 m 评分的均值; $S_{k,m}$ 和 $S_{c,m}$ 分别表示群组 k 和 c 内成员对项目 m 评分的标准差。

设定目标群组 k 的最近邻个数为 N_K，则选取在相似群组集合中与目标群组距离最小的前 N_K 个群组作为目标群组 k 的 N_K 个最近邻，若群组 c 在 N_K 个最近邻中，根据距离越小相似度越大的原则，群组 c 对于目标群组 k 的相似度权重如式(8) 所示。

$$w_c = \frac{\frac{1}{D(k,c)}}{\sum_{i=1}^{N_K} \frac{1}{D(k,i)}} = \frac{1}{D(k,c) \times \sum_{i=1}^{N_K} \frac{1}{D(k,i)}} \tag{8}$$

其中: $D(k,c)$ 为式(7) 中计算出的群组 k 与 c 之间的距离。

2.3　为目标群组进行推荐

为目标群组进行推荐首先要预测目标群组对项目 m 的评分，即通过目标群组的 N_K 个最近邻对项目 m 的评价值来预测目标群组 k 对 m 的评价值，预测公式如式(9) 所示。

$$\rho(k,m) = \sum_{c \in p_k} w_c \times g_{c,m} \tag{9}$$

其中: w_c 为群组 c 对于目标群组 k 的相似度的权重; $g_{c,m}$ 为群组 c 对于项目 m 的评价等级; p_k 为目标群组 k 的 N_K 个最近邻组成的集合。$g_{c,m}$ 的计算公式如式(10) 所示。

$$g_{c,m} = \frac{\sum_{i=1}^{n_c} q_{c,m}(i)}{n_c} \tag{10}$$

其中: n_c 为群组 c 中对项目 m 有评价的个体数量，$q_{c,m}(i)$ 为群组 c 中第 i 个个体对项目 m 的评分。

最后选取预测评分 $\rho(k,m)$ 最高的前 N 个项目，作为最终的 top-N 推荐集向目标群组进行推荐。

本算法中遍历一次群组评分，便可得到群组评分的区间型符号数描述，时间复杂度为 $O(c \times N)$。其中: N 为群组总数; c 为常数，表示群组中用户的数量。设目标群组 k 的相似群组集合中群组个数为 n，式(7) 中计算相似群组与目标群组 k 的距离需要经过 M 次循环，式(8) 取距离最小的前 N_K 个群组，分别计算他们对于目标群组 k 的相似度权重，最后的预测式(9) 中循环利用近邻群组对项目 m 的评分以及近邻群组对目标群组 k 的相似度权重得出最终的预测评分，其中近邻群组对 m 的评分等级是由式(10) 循环 n_c 次得到。因此算法的时间复杂度为

$O(N_K(n\times n\times M+n\times\lg n+n_c)+c\times N)$。其中,$n\times\lg n$、$n_c$ 和 $c\times N$ 均远小于 $n\times n\times M$,且 N_K 为常数,可用 c 表示,因此算法的时间复杂度为 $O(c\times n\wedge 2\times M)$。

3 实验结果与分析

3.1 实验数据来源

本文的实验数据来源于 movielens 数据库,该数据库包含了三个规模不同的数据集,由 GroupLens 实验室在 movielens 网站上对大量用户的电影评分收集得到,分值为 1~5。评分越高,意味着用户对该部电影越喜欢。本文选取数据库中的第二个数据集作为实验数据,它包含了 6 040 个用户对 3 900 部电影项目的约 100 万条评分数据。将此数据集分为训练集和测试集两个部分,在训练集上进行推荐算法模型的学习和参数调整,在测试集上对计算精确度和运行效率进行结果评价,从而达到评价与分析的目的。训练集和测试集分别占数据总量的 80% 和 20%。

3.2 实验对比方法

本文选取基于项目的点数据群组协同过滤算法[15] 作为对比方法,其基本思想是运用平均值法将群组成员的个体评分转换为群组评分,然后把群组对相似项目的评分进行加权来预测群组对目标项目的评分。其推荐过程主要分为以下三个步骤:

a) 形成群组评分矩阵。将群组中用户的评分转换为群组整体对项目的评分。该转换式如式(10)所示,由此便可得到点数据表示的群组评价矩阵。

b) 寻找最近邻。通过计算目标项目 m 与其他项目之间的相似性,按照相似度的大小进行排名,最终选择出目标项目 m 的近邻集合 $N=\{N_1,N_2,\cdots,N_k\}$,其中 $m\notin N$,$\mathrm{sim}(m,N_i)>\mathrm{sim}(m,N_{i+1})$,$i=1,\cdots,k-1$。本文的对比方法使用修正的余弦相似度公式作为项目之间的相似性度量。

c) 产生推荐集。根据群组对近邻项目的评分来预测群组对目标项目的评分,已知项目 m 和对项目 m 已评分的群组集 I_m,则群组 $k(k\notin I_m)$ 对项目 m 的预测评分可表示为

$$p_{k,m}=\frac{\sum\limits_{n\in\{\text{all similat items with }m\}}\mathrm{sim}(m,n)\,g_{k,n}}{\sum\limits_{n\in\{\text{all similat items with }m\}}|\mathrm{sim}(m,n)|} \tag{11}$$

其中:$\mathrm{sim}(m,n)$ 是项目 m 和最近邻项目 n 的相似度;$g_{k,n}$ 表示群组 k 对项目 n 的评分,可由式(10)计算得出。取预测值最高的前 N 个项目作为 top-N 推荐集推荐给群组用户。

3.3 评价标准

推荐系统的性能指标一般有推荐的精度和推荐的效率这两种,使用的指标有 mean absolute error(MAE)、root mean squared error(RMSE)等。本文选用 RSME 作为推荐精度的检验标准,将计算机完成预测评分所需的时间作为推荐效率的评价标准。

RSME 表示测试集中项目的预测评价值与实际评价值的均方根误差,即

$$\mathrm{RSME}=\frac{\sum\limits_m\sqrt{\dfrac{\sum\limits_i(r_{k,m}(i)-\rho_{k,m})^2}{n_k-1}}}{M}$$
$$r_{k,m}(i)\neq 0 \tag{12}$$

其中:$r_{k,m}(i)$ 为群组 k 中个体 i 对项目 m 的实际评价值;$\rho_{k,m}$ 为目标群组 k 对项目 m 的预测评分;n_k 为群组 k 中对项目 m 有评分的个体数量;M 为测试集中项目的个数。RSME 越小,表明推荐精度越高。

3.4 实验过程及结果分析

为评价算法的精度,本文从数据集的稀疏程度对算法的影响、稳定状况下(即数据最不稀疏)邻居数量对算法的影响以及推荐算法的效率这三个方面设计实验。

1) 稀疏度实验

本部分实验的目的是在实验数据稀疏度变化的情况下,比较不同等级的稀疏度对点数据协同过滤群组推荐算法与基于区间型符号数据的群组推荐算法在推荐精度上的影响。实验从测试集中随机选取了用户评分个数为 [50,80]、[80,110]、[110,140]、[140 以上] 的四个不同稀疏等级的数据集作为实验数据,在每个数据集中随机生成 30 个群组(每个群组中最少有一名成员)。表 3 列出了不同数据稀疏度情况下两种算法的 RSME 值。

表3 不同数据稀疏度情况下两种算法的 RSME 值

用户评分个数	均方根误差 RSME 值	
	基于区间型符号数据的推荐算法	基于项目的点数据推荐算法
[50,80]	1.271 2	1.387 9
[80,110]	1.162 3	1.351 3
[110,140]	1.115 6	1.305 8
[140 以上]	1.064 2	1.283 0

由表 3 可以看出,基于区间型符号数据的群组推荐算法的 RSME 值在各个稀疏度下均明显小于对比方法,通过分析可以得出如下实验结论:

a) 随着用户评分个数的增加,两种算法的 RSME 值均有明显下降,说明数据稀疏度越小,两种群组推荐算法的精确性越高。

b) 用户评分个数相同,即数据稀疏度相同的情况下,基于区间型符号数据的群组推荐算法要优于基于点数据的群组推荐算法,即区间型符号数据的群组推荐算法精度更高。

2) 稳定状况下推荐算法对比实验

稳定状况指数据最不稀疏时的状况,在这种状况下,主要通过改变电影数量和目标群组的邻居数量来进行对比实验。本部分实验将从推荐精度和推荐效率两个角度对推荐算法进行评价。

(1) 推荐精度对比实验

本部分实验分别在 30、50、70 部电影和 4、6、8、10 个近邻的情况下进行,分别采用两种算法在相应实验条件下对目标群组进行推荐,得到的均方根误差如表 4、5 所示,其中 N 为选取的电影数量,h 为近邻项目数。

对比表 4 和 5 可以得出以下结论:

a) 相同电影数量的情况下,邻居数量越多,RSME 值越小,算法的精度越高,推荐结果越准确。

b) 相同邻居个数的情况下,电影数量越多,RSME 值也越小,算法的精度也越高,推荐结果也越准确。

c) 基于区间型符号数据的群组推荐算法在各种实验设计条件下,RSME 值均小于对比方法,即推荐精度均优于基于项目的点数据群组协同过滤算法。

表 4　基于区间型符号数据的群组推荐算法的 RSME 值

近邻/个	电影/部		
	$N=30$	$N=50$	$N=70$
$h=4$	1.381 4	1.274 6	1.142 7
$h=6$	1.314 6	1.184 2	1.007 3
$h=8$	1.178 3	1.063 5	0.923 6
$h=10$	1.052 4	0.989 3	0.875 3

表 5　基于项目的点数据群组协同过滤算法的 RSME 值

近邻/个	电影/部		
	$N=30$	$N=50$	$N=70$
$h=4$	2.123 8	1.880 4	1.554 3
$h=6$	1.975 2	1.448 4	1.432 6
$h=8$	1.670 2	1.324 5	1.287 2
$h=10$	1.476 1	1.212 7	1.171 3

(2) 推荐效率对比实验

评价推荐算法的推荐效率是以计算机完成预测评分所需的时间为标准的,所需时间越小,表明推荐效率越高。在与推荐精度对比实验相同的实验条件下,两种推荐算法所需的计算时间如表 6 和 7 所示,其中 N 为选取的电影数量,h 为近邻项目数。

表 6　基于区间型符号数据的群组推荐算法的完成时间

近邻/个	电影/部		
	$N=30$	$N=50$	$N=70$
$h=4$	345 ms	416 ms	489 ms
$h=6$	367 ms	476 ms	506 ms
$h=8$	398 ms	523 ms	572 ms
$h=10$	452 ms	593 ms	598 ms

表 7　基于项目的点数据群组协同过滤算法的完成时间

近邻/个	电影/部		
	$N=30$	$N=50$	$N=70$
$h=4$	2 785 ms	3 289 ms	3 367 ms
$h=6$	3 068 ms	3 567 ms	3 895 ms
$h=8$	3 683 ms	3 840 ms	4 061 ms
$h=10$	4 179 ms	4 369 ms	4 509 ms

对比表 6 和 7 可以得出以下结论:

a) 相同电影数量的情况下,邻居数量越多,实验完成所需的时间越多,效率越低。

b) 相同邻居个数的情况下,电影数量越多,实验完成所需的时间也越多,效率也越低。

c) 基于区间型符号数据的群组推荐算法在各种实验设计条件下,计算时间均小于对比方法,即推荐效率均优于基于项目的点数据群组协同过滤算法。

综合上述实验结论可以认为,基于区间型符号数据的群组推荐算法不论在推荐精度上还是推荐效率上,均优于基于项目的点数据群组协同过滤算法,这也证明了在群组推荐算法中使用群组成员的评分信息有助于提高算法的精度和效率。基于区间型符号数据的群组推荐算法用数据的区间型组织方式来体现群组的特性,充分利用了群组内部的评分信息,使推荐结果更加可靠准确,更具有效性。

4　结束语

本文利用区间型符号数据内部点数据的均值和标准差,提出了一种全新的针对一般分布区间型符号数据的距离度量——$\mu\sigma$ 距离。通过将群组用户的评分表示为区间型符号数据,并将 $\mu\sigma$ 距离应用于区间型符号数据的 K-均值聚类方法,得到目标群组的最近邻,最后通过将相似群组的评分加权得到目标群组的预测评分,达到群组推荐的效果。通过与基于项目的点数据群组协同过滤算法进行对比实验,证明了将区间型符号数据运用于群组推荐的方法在推荐精度和推荐效率上均优于传统的点数据群组推荐算法。由于目前基于符号数据的群组推荐算法是一个新的研究领域,本文提出的基于区间型符号数据的群组推荐算法虽然在精度上有明显优势,但推荐算法中的冷启动这一固有问题还有待进一步研究解决; 另外,在推荐系统中,依靠大部分群组用户对某个项目进行精确打分比较难以实现,这就需要系统能根据用户的行为自动转换为用户对该项目的评分,从而有效地降低稀疏性问题,为用户提供更优质的推荐。因此,利用隐式评分完成群组推荐也是将来的研究方向。

参考文献:

[1] 李岱峰,覃正. 一种基于资源多属性分类的群组推荐模型[J]. 统计与决策,2010(8):153-155.

[2] YU Zhi-wen, ZHOU Xing-she, HAO Yan-bin, *et al*. TV program recommendation for multiple viewers based on user profile merging[J]. User Modeling and User-Adapted Interaction, 2006, 16(1): 63-82.

[3] CHEN Y L, CHENG Li-chen, CHUANG C N. A group recommendation system with consideration of interactions among group members[J]. Expert Systems with Applications, 2008, 34(3): 2082-2090.

[4] McCARTHY K, SALAMÓ M, COYLE L. Group recommender systems: a critiquing based approach[C]//Proc of the 11th International Conference on Intelligent User Interfaces. New York: ACM Press, 2006: 267-269.

[5] GARCIA I, SEBASTIA L, ONAINDIA E. On the design of individual and group recommender systems for tourism[J]. Expert Systems with Applications, 2011, 38(6): 7683-7692.

[6] ROY S B, CHAWLA A, DAS G, *et al*. Space efficiency in group recommendation[J]. International Journal on Very Large Data Bases, 2010, 19(6): 877-900.

[7] QUIJANO-SÁNCHEZ L, RECIO-GARCÍ J, DÍAZ-AGUDO B, *et al*. Personality and social trust in group recommendations[C]//Proc of the 22nd IEEE International Conference on Tools with Artificial Intelligence. Washington DC: IEEE Computer Society, 2010: 121-126.

[8] BOCK H H, DIDAY E. Analysis of symbolic data[M]. Berlin: Springer-Verlag, 2000.

[9] 胡艳,王惠文. 一种海量数据的分析技术—符号数据分析及应用[J]. 北京航空航天大学学报: 社会科学版,2004,17(2):40-44.

[10] QUEIROZ S R M, De CARVALHO F A T. Making collaborative group recommendations based on modal symbolic data[C]//Lecture Notes in Computer Science, vol 3171. 2004: 121-153.

[11] BEZERRA B L D, De CARVALHO F A T. A symbolic approach for content-based information filtering[J]. Information Processing Letters, 2004, 92(1): 45-52.

[12] BEZERRA B L D, De CARVALHO F A T. Symbolic data analysis tools for recommendation systems[J]. Knowledge and Information Systems, 2010, 26(3): 385-418.

[13] 郭均鹏,李汶华,高峰. 一般分布区间型符号数据的描述统计与分析[J]. 系统工程理论与实践,2011,31(12):2367-2372.

[14] De CARVALHO F A T, De SOUZA R M C R, CHAVENT M, *et al*. Adaptive Hausdorff distances and dynamic clustering of symbolic interval data[J]. Pattern Recognition, 2006, 27(3): 167-179.

[15] ARDISSONO L, GOY A, PETRONE G, *et al*. Intrigue: personalized recommendation of tourist attractions for desktop and handset devices[J]. Applied AI, 2003, 17(8-9): 687-714.