基于社交关系和用户偏好的多样性图推荐方法

石进平[1,3] 李 劲[1,3] 和凤珍[2]
(云南大学软件学院 昆明 650091)[1] (云南大学旅游文化学院信科系 云南 丽江 674199)[2]
(云南省软件工程重点实验室 昆明 650091)[3]

摘 要 以协同过滤为代表的传统推荐算法能够为用户提供准确率较高的推荐列表,但忽略了推荐系统中另外一个重要的衡量标准:多样性。随着社交网络的日益发展,大量冗余和重复的信息充斥其间,信息过载使得快速、有效地发现用户的兴趣爱好变得更加困难。针对某个用户推荐最能满足其兴趣爱好的物品,需要具备显著的相关度且能覆盖用户广泛的兴趣爱好。因此,基于社交关系和用户偏好提出一种面向多样性和相关度的图排序框架。首先,引入社交关系图模型,综合考虑用户及物品之间的关系,以更好地建模它们的相关度;然后,利用线性模型融合多样性和相关性两个重要指标;最后,利用 Spark GraphX 并行图计算框架实现该算法,并在真实的数据集上通过实验验证所提方法的有效性和扩展性。

关键词 多样性,相关性,社交网络,个性化推荐系统,Spark GraphX
中图法分类号 TP391 **文献标识码** A

Diversity Recommendation Approach Based on Social Relationship and User Preference

SHI Jin-ping[1,3] LI Jin[1,3] HE Feng-zhen[2]
(School of Software, Yunnan University, Kunming 650091, China)[1]
(Department of Information and Science, College of Tourism and Culture, Yunnan University, Lijiang, Yunnan 674199, China)[2]
(Key Laboratory of Software Engineering of Yunnan Province, Kunming 650091, China)[3]

Abstract The traditional recommendation algorithm, represented by collaborative filtering, can provide users with a high recommended list with high accuracy, while ignoring another important measure which is diversity in the recommendation system. With the increasing development of social networks, with a lot of redundancy and duplication of information, the overload information makes it more difficult to find user interests quickly and effectively. For recommending the most content for users to meet their hobbies, user interests with a significant relevance and covering different aspects are needed. Therefore, based on social relations and user preferences, this paper proposed a sorting framework for diversity and relevance. Firstly, this paper introduced the social relations graph model, considering the relationship between users and items to better model their relevance. Then, this paper used a linear model to integrate the two important indexes of diversity and relevance. Finally, the algorithm was implemented by Spark GraphX parallel graph calculation framework, and experiments were carried on real dataset to verify the feasibility and scalability of the proposed algorithm.

Keywords Diversity, Relevance, Social network, Personalized recommendation system, Spark GraphX

推荐系统(Recommender Systems,RS)作为一种能够准确地过滤出用户最感兴趣物品的软件工具,吸引着许多研究者进行大量研究,并产生了一系列重要成果[1,8],极大地推进了推荐系统的发展进程。推荐系统的大量实际应用,如图书推荐、电影推荐、旅游路线推荐等,能有效帮助用户解决信息过载(Information Overload)问题,同时推荐系统具有极大的商业价值,这使得推荐系统成为学者们研究的热点话题。

近年来,学者们已经意识到传统推荐算法虽然通过准确预测用户对未知物品的评分,为用户提供精确的推荐结果;但是推荐结果的准确度不是衡量推荐结果好坏的唯一标准[5],推荐结果的多样化(Diversity)也扮演着重要的角色,更能迎合不同用户的广泛兴趣爱好[1]。同时,各学者针对多样性积极开展研究,并取得了一系列成果。例如:文献[4]提出了一种 CluDiv 重排序框架,利用传统的协同过滤方法对用户未接触的物品预测评分,并在预测评分的基础上基于聚类方法获得多样性推荐结果;类似地,文献[11-12]利用协同过滤方法预测用户的评分,进而利用优化的方式提高推荐结果的多样性;不同于文献[4,11],文献[12]将多样性和准确性同时融合到目标函数中进行优化,并通过贪心法求得近似解。文献[6]首先利用话题多样化来提高推荐列表的多样性,使得所推荐

本文受国家自然科学基金项目(61562091),云南省应用基础研究计划面上项目(2016FB110),云南省软件工程重点实验室开放项目(2012SE303,2012SE205)资助。

石进平(1989-),男,硕士,主要研究方向为大数据分析与处理、机器学习;**李 劲**(1975-),男,博士,副教授,主要研究方向为大数据分析与处理、机器学习,E-mail:lijin@yun.edu.cn;**和凤珍**(1988-),女,讲师,主要研究方向为数据挖掘。

的新闻资讯来自于不同的新闻类别，但是该方法是以牺牲准确度为代价换取的多样性的提高；文献[7]通过将多样性问题建模为动态规划问题，利用优化效用函数的方式同时考虑相关性和多样性两个因素，贪心选择具有最大边际相关度(Maximal Marginal Relevance)的物品作为推荐结果。然而，这些解决多样性和相关性问题的方法忽略了用户之间的社交关系对相关性的影响，而且算法的时间复杂度和空间复杂度较高，可扩展性不高。

随着社交网络(Social Network)的日益发展，大量冗余和重复的信息充斥其中。许多学者针对基于社交网络的推荐系统进行研究，例如：文献[2]基于社交网络，利用过滤算法模型、概率模型两种模型来调整准确度和多样性之间的平衡；文献[8]基于社交网络提出了 RSboSN 推荐框架，该框架融入了社交网络中的“朋友”关系，过滤了与用户兴趣不太相关甚至相反的“朋友”关系，提高了推荐系统的准确度，而且社交网络中的“朋友”关系能够解决推荐系统中的数据稀疏性(Sparsity)问题和“冷”启动问题；文献[9]在文献[3]的基础上进行了扩展，该方法利用香依熵[9](Shannon Entropy)方法，通过引入社交关系，过滤掉一些“不相关”的推荐项，从而提高了推荐结果的准确度，但是此方法没有区分不同的社交关系，而是平等看待所有的社交关系。这些模型考虑了社交网络中用户之间的社交关系，但是现有的基于社交关系的推荐系统[3,8-9]多侧重于提高相关性而没有同时考虑多样性需求。

针对推荐系统的推荐结果，现有的基于社交网络的推荐系统中关注推荐结果相关度的模型忽略了推荐结果的多样性，而解决多样性问题的模型又没有考虑社交关系对推荐结果相关性的影响。受上述相关工作的启发，结合社交网络的快速发展和推荐结果的多样性进行综合考虑，提出了一种基于社交关系的多样性分布式推荐框架，以更高效地平衡多样性和相关性。

1 基于社交关系和用户偏好的多样性图推荐框架

多样性推荐是针对给定的用户，要求返回与此用户兴趣爱好相关度高、覆盖广泛类别的物品。以电影推荐系统为例，多样性推荐的任务就是返回与推荐用户兴趣相关性较高且覆盖多种电影类别而不是某一类别(如动作片)的电影。本节首先对多样性推荐问题进行详细的描述；接着详细描述基于社交网络和用户偏好的多样性图推荐框架(PSR-GRS)的构造，并对相关性、多样性进行建模；然后通过一个线性模型融合多样性和相关性两个重要指标对电影进行排序；最后给出该多样性推荐框架在 Spark GraphX 并行图计算平台上的分布式算法。

1.1 问题描述

推荐系统的多样性推荐针对类似电影网络的异质网络，该网络中同时存在观众和电影两种不同类型的对象。这两种对象之间存在着 3 种关系：观众与电影之间的观看关系、观众与观众之间的“朋友”关系以及电影与电影之间的相似关系。我们利用一个无向图来形式化定义一个电影网络 $G=(V,E,W)$，其中顶点 V 是两种不同类型对象的集合，包括观众 A 和电影 F 的集合；顶点之间通过带权(Weight)无向边 E 相连接，边的权重 W 被形式化定义为：

$$w(i,j)=\begin{cases} score(i,j), & \text{如果 } i\in A, j\in F \\ sim(i,j), & \text{如果 } i\in A, j\in A \text{ 或者 } i\in F, j\in F \end{cases}$$

其中，$score(i,j)$为观众 i 对电影 j 的打分值；$sim(i,j)$为电影 i,j 之间的相似度或观众 i,j 之间的相似度。因此，这个电影网络是由上述 3 种关系的子网络构成的。我们的目标是在这个网络中求得同时具有多样性和相关度的电影。

1.2 构建基于社交关系和用户偏好的用户行为图

正如前文所述，电影网络中存在着观众与电影之间的观看关系子网络、观众与观众之间的“朋友”关系子网络以及电影与电影之间的相似关系子网络。然而我们发现，这 3 种关系与推荐结果的相关度有着紧密的联系，应当综合考虑。正如实际生活中，当观看电影面临选择困难时，人们通常存在以下两种可能性：

可能性 1 让与自己趣味相投的朋友推荐；

可能性 2 选择观看一些与自己已经看过、非常喜欢的电影最相似的电影。

以上两种可能性在某种情况下会同时存在。例如，当面对重要决策时，人们通常会综合朋友的建议以及自己的想法作出理智的抉择。因此，我们的工作针对上述两种可能性同时存在的情况进行研究。下面通过一个例子来描述观众的轮廓图(Profile Graph)的构建过程。

电影网络中存在 $A_1,A_2,A_3,A_4,A_5\in A$ 5 个观众和 $F_1,F_2,F_3,F_4,F_5,F_6\in F$ 6 部电影，其中用户之间的社交网络如图 1 所示，该社交网络中存在如图 2 所示的“朋友”关系，1 代表观众之间是“朋友”；电影之间的相似关系如图 3 所示；图 4 描述了观众与电影之间的打分关系，0 表示该观众没有观看过此电影，其余的数值表示观众对观看过的电影的评分值。

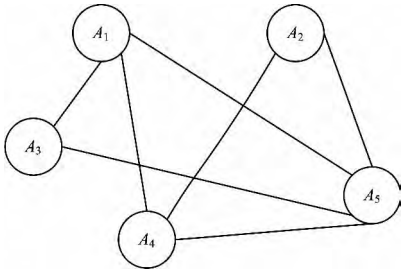

图 1 社交网络

	A_1	A_2	A_3	A_4	A_5
A_1	0	0	1	1	1
A_2	0	0	0	1	1
A_3	1	0	0	0	1
A_4	1	1	0	0	1
A_5	1	1	1	1	0

图 2 “朋友”关系

	F_1	F_2	F_3	F_4	F_5	F_6
F_1	1.0	0.5	0.4	0.3	0.3	0.6
F_2	0.5	1.0	0.3	0.2	0.6	0.7
F_3	0.4	0.3	1.0	0.4	0.5	0.3
F_4	0.3	0.2	0.4	1.0	0.3	0.1
F_5	0.3	0.6	0.5	0.3	1.0	0.2
F_6	0.8	0.7	0.3	0.1	0.2	1.0

图 3 电影相似关系

	F_1	F_2	F_3	F_4	F_5	F_6
A_1	5	5	0	0	0	0
A_2	0	0	4	2	0	3
A_3	1	0	0	3	0	5
A_4	0	1	4	4	0	0
A_5	4	4	0	0	4	0

图 4 观众与电影的观看关系

采用文献[3]和文献[9]的方法可以得到每个观众喜欢的

电影，这些电影的评分值均大于或等于该用户打分的平均值。因此，可以得到前面 5 个观众喜欢的电影，即 $P(A_1)=\{F_1,F_2\}$，$P(A_2)=\{F_3,F_6\}$，$P(A_3)=\{F_4,F_6\}$，$P(A_4)=\{F_3,F_4\}$，$P(A_5)=\{F_1,F_2,F_5\}$。对于 $P(A_1)$，可以得到图 5 中最左边的带权矩阵，用同样的方式处理后可以得到最终的带权邻接矩阵 M，如图 5 中最右边矩阵所示。

$$\begin{array}{c} \\ F_1\\F_2\\F_3\\F_4\\F_5\\F_6 \end{array}\!\!\begin{array}{c} \begin{array}{cccccc} F_1 & F_2 & F_3 & F_4 & F_5 & F_6 \end{array}\\ \begin{pmatrix} 0&1&0&0&0&0\\1&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0 \end{pmatrix} \end{array} +\cdots+ \begin{array}{c} \begin{array}{cccccc} F_1 & F_2 & F_3 & F_4 & F_5 & F_6 \end{array}\\ \begin{pmatrix} 0&1&0&0&1&0\\1&0&0&0&1&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\1&1&0&0&0&0\\0&0&0&0&0&0 \end{pmatrix} \end{array} = \begin{array}{c} \begin{array}{cccccc} F_1 & F_2 & F_3 & F_4 & F_5 & F_6 \end{array}\\ \begin{pmatrix} 0&3&0&0&1&0\\3&0&0&0&1&0\\0&0&0&2&0&1\\0&0&2&0&1&1\\1&1&0&1&0&0\\0&0&1&1&0&0 \end{pmatrix} \end{array}$$

图 5　生成带权邻接矩阵 M 的过程

接下来需要构建个性化的用户轮廓图(Personalized User Profile Graph)，以 $P(A_1)$ 为例来描述观众 A_1 轮廓图的构建并展示本文方法与文献[9]中的 SR-GRS 的不同。由图 2 可以看出，A_3，A_4，A_5 是 A_1 的“朋友”，其中 A_5 与 A_1 具有相同的兴趣爱好(因为都喜欢电影 F_1，F_2)，尽管 A_3，A_4 也是 A_1 的“朋友”，然而他们与 A_1 的爱好没有交集，文献[9]中的 SR-GRS 模型平等地看待了 A_1 与“朋友”A_3，A_4，A_5 之间的社交关系，A_3，A_4 共同对 F_4 感兴趣。根据可能性 1，A_1 更愿意选择向“朋友”A_5 寻求推荐。根据可能性 2，由图 3 可得，F_6 与 F_1，F_2 之间的平均相似最大。综合考虑前面两种可能性，A_1 也可能对 F_5 和 F_6 感兴趣。前面分析得到观众 A_1 喜欢的电影 $P_(A_1)'=P(A_1)\cup\{F_5\}\cup\{F_6\}=\{F_1,F_2,F_5,F_6\}$，因此从带权邻接矩阵 M 中选择 F_1，F_2，F_5，F_6 所对应的行构成观众 A_1 的子矩阵(见图 7)和轮廓图 g(见图 8)。

$$\begin{array}{c} \\ F_1\\F_2\\F_4 \end{array}\!\!\begin{array}{c} \begin{array}{cccccc} F_1 & F_2 & F_3 & F_4 & F_5 & F_6 \end{array}\\ \begin{pmatrix} 0&3&0&0&1&0\\3&0&0&0&1&0\\0&0&2&0&1&1 \end{pmatrix} \end{array}$$

图 6　SR-GRS 中 A_1 的子矩阵

$$\begin{array}{c} \\ F_1\\F_2\\F_5\\F_6 \end{array}\!\!\begin{array}{c} \begin{array}{cccccc} F_1 & F_2 & F_3 & F_4 & F_5 & F_6 \end{array}\\ \begin{pmatrix} 0&3&0&0&1&0\\3&0&0&0&1&0\\1&1&0&1&0&0\\0&0&1&1&0&0 \end{pmatrix} \end{array}$$

图 7　PSRGraph 中 A_1 的子矩阵

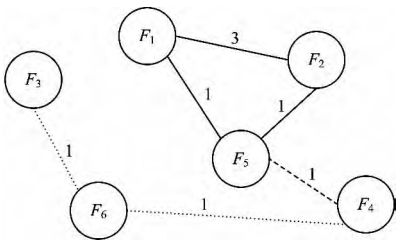

图 8　PSRGraph 中 A_1 的轮廓图

通过图 6 和图 7 的对比，展示了本文方法(PSRGraph)与 SR-GRS 的不同。通过图 7 生成个性化轮廓图 g，g 是带权邻接矩阵 M 对应图的一个子图，g 中的虚线边与图 7 中虚线框的边相对应，两种不同的虚线框代表前面提到的两种可能性。

1.3　基于社交关系和用户偏好的多样性和相关性

上节描述了基于社交关系和用户偏好构建个性化的用户轮廓图，下面详细描述基于个性化的用户轮廓图的多样性和相关性。

1.3.1　多样性和相关性

对于每个个性化的用户轮廓图，利用式(1)计算图中每个顶点的香依信息熵[9]。信息熵描述信息量的大小，顶点的熵越大，其包含的信息量就越大；信息熵越大，多样性也越丰富。

$$H(X)=-\sum_{i}^{n}P(x_i)\log_b P(x_i) \tag{1}$$

其中，X 是图中的顶点；x_i 是与 X 直接相连的顶点；n 是与 X 直接相连的顶点的数目；n 是对数的基，通常取 2，e 或 10；$P(x_i)$是顶点 x_i 的概率质量函数，根据文献[3]和文献[9]，用 $P(x_i)$作为顶点 X 与 x_i 之间的边的权重。由于与每个顶点直接相连的邻居个数不同，因此有必要进行归一化处理。以图 8 中的 F_1 为例，信息熵的计算过程如下：与 F_1 的邻居$\{F_2,F_5\}$相对应的边的权重为$\{3,1\}$，$H(F_1)=-((\frac{3}{4}\log_2\frac{3}{4})+(\frac{1}{3}\log_2\frac{1}{3}))=0.8396$。

图 8 中，图 g 中的顶点所拥有的度(Degree)越大，表示其被共同喜欢的用户越多，这些顶点与目标用户的兴趣相关度越高。因此，我们用顶点的度作为该顶点与目标用户兴趣爱好的相关度，形式化定义为等式(2)：

$$Relevance(X)=Degree(X) \tag{2}$$

文献[3-4,9]先获得相关性结果，再在相关性结果中选择具有多样性的电影作为推荐结果。不同于上述方法，为了获得更好的推荐结果，利用等式(3)中简单的线性模型 DivRel 融合相关性和多样性。

$$DivRel=Relevance(X)*H(X) \tag{3}$$

1.4　PSR-GRS 多样性推荐框架

在介绍了信息熵的多样性模型和线性模型后，当给定一个目标用户时，PSR-GRS 多样性推荐框架的工作方式如下：

1)根据用户对电影的评分生成用户喜欢的电影评分图；

2)分析社交网络中用户的“朋友”关系、电影相似关系和评分关系，并利用这些关系生成个性化的用户轮廓图；

3)利用算法计算同时具有多样性和相关性的顶点并降序排列；

4)按照文献[4]的方法调整类别权重，并从第 3)步的结果中从高到低选择电影，将所选电影的类别权重减 1，直到调整后的类别权重都为 0 为止。

通过这 4 个步骤，最终得到同时具有相关性和多样性的电影推荐列表。

1.5　基于 Spark Graphx 的并行算法

在 PSR-GRS 推荐框架中，推荐的核心是计算每个物品的多样性。物品之间彼此相互独立；计算多样性之前需要生成每个用户的个性化爱好，用户之间也是相互独立的，允许进

行并行计算。因此，可以基于 Spark 的并行图计算组件 GraphX 实现并行化 PSR-GRS 算法(简写为 PPSR-GRS)。

Spark 通过 RDD(Resilient Distributed Dataset)转换可以方便地实现并行化操作，同时利用 GraphX 的并行消息传递机制实现顶点和边的多样性技术，极大地提高了算法的运行效率。PPSR-GRS 具体如算法 1 所示。

算法 1 PPSR-GRS

输入：数据集 dataset，推荐用户 user；推荐数量 N，类别阈值 threshold

输出：推荐结果集 result

```
1. rawRatingsRDD←sc. textFile(dataset);//读取数据集
2. linesRDD←rawRatingsRDD. map(line). persist();//分割数据内容
3. clusterRDD←getMovieCluster(linesRDD);// 聚类
4. usersRDD←构造用户属性顶点;
5. itemsRDD←构造物品属性顶点;
6. verticesRDD←VertexRDD(usersRDD++ itemsRDD);//合并用户、物品两种异质顶点
7. edgesRDD←rawRatingsRDD. map(line=>Edge(line._1,line._2,line._3));//根据用户评分构建边集 RDD;
8. g←Graph(verticesRDD,edgesRDD). persist();//生成包含用户、物品的异质图 g;
9. totalScoreRDD←g. aggreateMessages(triplet. sendToSrc(triplet. attr),_+_,TripletFields. EdgeOnly);//通过消息聚合计算每个用户评分的总和
10. outdegreeRDD←g. outDegrees;//计算每个顶点的出度
11. avgScoreRDD←totalScoreRDD. innerJoin(outdegreeRDD);//计算用户评分的平均分
12. positiveEdgeRDD←PositiveRDD(avgScoreRDD,edgesRDD);//选出用户喜欢的物品集合
13. usersPersonalizedPositiveEdgesRDD←PersonalizedPositeItem(posi-tiveEdgeRDD);
    //根据评分生成所有用户喜爱的电影结合
14. personalEdgesRDD←PersonlizedPreferenceItem(usersPersonal-izedPositiveEdgesRDD);//根据可能性 1 和可能性 2 计算用户也可能喜欢的物品集合
15. PersonalizedGraph←Graph(itemsRDD,personalEdgesRDD);//生成所有物品构成顶点属性图
16. userPensonalizedProfile←PernonalizedSubGraph(user,usersPer-sonalizedPositiveEdgesRDD,PersonalizedGraph);//抽取用户 user 的个性化子图
17. div←CacuDiversity(userPensonalizedProfile);//根据等式(3)计算多样性
18. result←Recomendation(div,N,threshold);//推荐 N 个物品，同类别中物品被推荐的数量上限为 threshold
return result;
```

2 实验结果

2.1 评价标准

Aytekin[4] 描述了一种能有效度量特定用户推荐列表多样性的方法。该方法计算用户推荐列表中两两物品相异度的平均值，定义如下：

设 $I=\{i_1,i_2,\cdots,i_n\}$ 为所有物品的集合，$U=\{u_1,u_2,\cdots u_m\}$ 是全部用户的集合，那么用户 $u(u\in U)$ 的推荐列表 $R_u\subseteq I$ 的多样性 Div 为：

$$Div=\frac{1}{N(N-1)}\sum_{i\in R(u)}\sum_{j\in R(u),j\neq i}(1-Sim_i(j)) \tag{4}$$

其中，$Sim(i,j)$ 是物品 $i,j\in I$ 的相似度。式(4)仅考虑推荐列表的不相似度，不能更完整地反映推荐结果的多样性，比如，尚未考虑推荐结果的类别覆盖度。受此启发，我们提出一种新颖的多样性度量方法 DisCoverDiv，其被形式化表示为：

$$DisCoverDiv=Dis(R_u)*Cover(R_u) \tag{5}$$

其中，$Dis(R_u)$ 为推荐结果 R_u 中两两物品之间的不相似度，可以通过式(4)获得，$Cover$ 是推荐结果 R_u 所覆盖的类别数目 $|C(R_u)|$ 占推荐列表长度 $|R_u|$ 的比例，被描述为：

$$Cover=\frac{|C(R_u)|}{|R_u|}\times 100\% \tag{6}$$

准确率(Precision)作为衡量结果准确度的标准，在推荐系统[1]等领域有着广泛的应用。本文利用 Precision 来评价 top-N 推荐列表的准确度，其定义如下：

$$Precision=\frac{|R_u\cap T|}{|R_u|}\times 100\% \tag{7}$$

其中，R_u 为用户 $u\in U$ 的推荐结果列表，$T\subseteq I$ 为用户 $u\in U$ 评过 5 分的电影集合。在准确度度量中，假设用户对物品的评分越高，该物品与用户兴趣的相关度也就越高。基于此，我们仅选择用户评过 5 分的电影进行测试。正如文献[4]中提到，这种做法往往会比实际的准确度低。

2.2 实验环境

在两个不同大小的 MovieLens[1)] 数据集上对本文提出的方法进行实验验证。实验环境为 Intel Core i7 处理器，8Cores，主频 3.4GHz，Ubuntu 15.04 系统，16GB 内存。程序基于 Spark Graphx[2)] 分布式图计算框架，采用 Scala2.10 实现。MovieLens 数据集的详细信息如表 1 所列。

表 1 数据集的详细信息统计

	用户数	电影数	结点数	边数
MovieLens1M	6000	4000	10000	1000000
MovieLens10M	72000	10000	82000	10000000

2.3 基于社交关系和用户偏好模型的有效性

首先验证基于社交关系和用户偏好模型的有效性。在 MovieLens1M 数据集上，从多样性和准确性两方面对比了所提方法 PSR-GRS 与 SR-GRS，GRS，CluDiv 模型的差异，实验结果如图 9、图 10 所示。从图中可以看出，多样性随着准确度的提高而下降，准确度也随着多样性的提高而降低，这客观地反映了多样性与准确度之间的博弈关系。此外，CluDiv 采用协同过滤算法进行评分预测，并利用其进行聚类，从不同类别中选择评分较高的项目作为推荐结果，其能够覆盖广泛的类别，在多样性方面，表现优于其他 3 种方法，但是该方法没有考虑用户的偏好和社交关系，在准确度方面的表现较差。然而，PSR-GRS 和 SR-GRS，GRS 3 种基于图模型的方法均能在多样性和准确度之间取得较好的折中。但是，SR-GRS 和 GRS 侧重于提高准确度，在优化准确度后进一步选择更具新颖性的推荐结果；不同于 SR-GRS，GRS 方法，本文提出的基于社交关系和用户偏好的模型 PSR-GRS 在推荐过程中同时

1) https://grouplens.org/datasets/movielens/

2) http://spark.apache.org

融合了多样性和准确性，因此其在多样性和准确度上均优于 SR-GRS 和 GRS 方法。

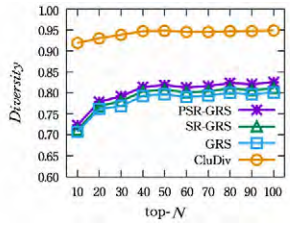

图 9　不同算法推荐结果的多样性对比

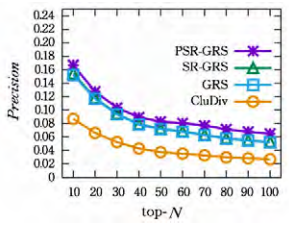

图 10　不同算法推荐结果的准确度对比

2.4　PPSR-GRS 算法的执行效率

该实验的目的主要是对比串行的 PSR-GRS 算法与并行的 PPSR-GRS 算法执行所耗费的时间。对比两种算法在不同规模的数据集上分别为 1000，2000，3000，4000，5000 个用户推荐 top-10 部具有多样性电影所用的时间。PPSR-GRS 算法采用 8 Cores 并行执行，实验结果如图 11 和图 12 所示。

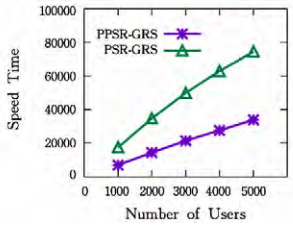

图 11　PPSR-GRS 与 PSR-GRS 执行时间的对比（MovieLens1M）

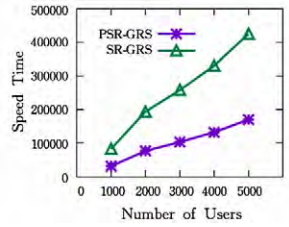

图 12　PPSR-GRS 与 PSR-GRS 执行时间的对比（MovieLens10M）

由实验结果可知，PPSR-GRS 并行算法的效率明显优于 PSR-GRS 串行算法。从 MovieLens1M 到 MovieLens10M，网络规模越大，并行算法的执行效率优势越明显。

当并行核数设置为 1，2，4，6，8 时，PPSR-GRS 算法为 5000 个用户进行推荐的加速比和并行效率如图 13、图 14 所示。其中，加速比 S_m 描述了多核并行执行算法与对应单核执行算法耗时的比率，计算公式为 $S_m = T_m / T_1$，T_m 表示有 m 个处理器并行执行算法所消耗的时间。此外，并行效率 E_m 描述并行算法在通信开销情况下参与计算的处理器的利用率，计算公式为 $E_m = S_m / m$。由图 13、图 14 可以看出，随着并行核数的增加，并行算法的加速比在增加；随着并行数量的增加，并行效率在逐渐降低。

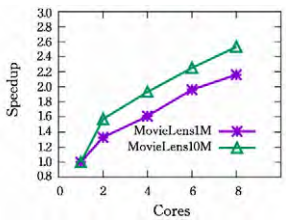

图 13　PPSR-GRS 算法的并行加速比

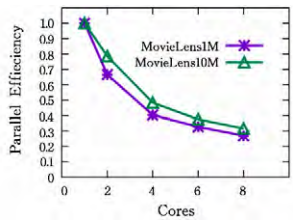

图 14　PPSR-GRS 算法的并行效率

结束语　多样性是目前推荐系统研究的一个重要内容。针对用户可能利用社交关系或用户偏好进行决策，且社交关系和用户偏好可能同时影响用户抉择的情况，提出基于社交关系和用户偏好的多样性图推荐模型，并给出了相应的推荐算法。在真实的数据集上验证算法，实验结果表明了文中所提方法的有效性。在本文研究的基础上，未来可以考虑如何通过图抽样的方法来降低图节点的计算量，从而进一步提高推荐的效率。

参 考 文 献

[1] KUNAVER M，POŽRL T. Diversity in recommender systems— A survey[J]. Knowledge-Based Systems，2017，123：154-162.

[2] JAVARI A，IZADI M，JALILI M. Recommender Systems for Social Networks Analysis and Mining：Precision Versus Diversity[J]. Understanding Complex Systems，2016，73：423-438.

[3] LEE K，LEE K. Escaping your comfort zone：A graph-based recommender system for finding novel recommendations among relevant items[J]. Expert Systems with Applications，2015，42(10)：4851-4858.

[4] AYTEKIN T，KARAKAYA M Ö. Clustering-based diversity improvement in top-N recommendation[J]. Journal of Intelligent Information Systems，2014，42(1)：1-18.

[5] MCNEE S M，RIEDL J，KONSTAN J A. Being accurate is not enough：how accuracy metrics have hurt recommender systems[C]// CHI '06 Extended Abstracts on Human Factors in Computing Systems. ACM，2006：1097-1101.

[6] ZIEGLER C，MCNEE S M，KONSTAN J A，et al. Improving recommendation lists through topic diversication[C]// International Conference on World Wide Web. 2005：22-32.

[7] HURLEY N，ZHANG M. Novelty and Diversity in Top-N Recommendation— Analysis and Evaluation[J]. ACM Transactions on Internet Technology，2011，10(4)：1-30.

[8] SUN Z，HAN L，HUANG W，et al. Recommender systems based on social networks[J]. Journal of Systems and Software，2015，99(C)：109-119.

[9] LIU R，JIN Z. An Improved Graph-based Recommender System for Finding Novel Recommendations among Relevant Items[C]// International Conference on Mechatronics，Materials，Chemistry and Computer Engineering. 2015.

[10] SHANNON C E. A mathematical theory of communication[J]. ACM Sigmobile Mobile Computing & Communications Review，2001，5(1)：3-55.

[11] ANTIKACIOGLU A，RAVI R. Post Processing Recommender Systems for Diversity[C]// The ACM SIGKDD International Conference. ACM，2017：707-716.

[12] LEE S C，KIM S W，PARK S，et al. A Single-Step Approach to Recommendation Diversification[C]// 26th International Conference on World Wide Web Companion. 2017.