推荐质量评估. 如图 1 所示.

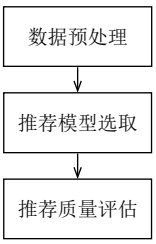

图 1 系统结构图

(1) 数据预处理模块

该模块根据推荐模型选取模块生成相应推荐模型所需的数据格式, 初始化实验参数. 读取 m 个用户对 n 个产品的评分文件, 用这些评分构成 m 行 n 列的评分矩阵, 提供给推荐模型.

(2) 推荐模型选取模块

根据 (1) 提供的数据, 用选取的推荐模型计算被推荐用户对产品的预测评分, 选择预测评分大于某阈值的产品推荐给该用户. 可以选择的模型有本文提出的基于差分隐私保护的模糊 C 均值聚类推荐模型, 或者其他典型的推荐模型, 如基于 K-means 聚类的协同过滤推荐模型、基于用户的协同过滤推荐模型等.

(3) 推荐质量评估模块

根据 (2) 得到的被推荐用户对产品的预测评分, 和该用户对产品的实际评分进行比较, 衡量所用推荐模型的推荐质量. 本实验通过准确度来衡量模型的推荐质量, 评价指标有均方根误差 (RMSE), 平均绝对偏差 (MAE) 和反映召回率和准确率情况的 F-measure. 通过计算以上三个评价指标, 来衡量所用推荐模型的推荐质量. 该模块主要为了衡量推荐质量, 通过选取本文提出的新模型和其他典型推荐模型, 分别进行实验, 通过推荐质量的比较, 表现新算法的有效性.

本实验使用 Java 语言在 Myeclipse 开发平台上实现以上的各个模块. 我们选择基于差分隐私保护的模糊 C 均值聚类推荐模型, 其实现流程如图 2 所示.

怎样将差分隐私和模糊 C 均值聚类融合是本实验的关键. 其具体流程如图 3 所示. 其核心是根据公式 (4) 得到聚类中心, 并添加满足差分隐私的 Laplace 噪声, 根据公式 (5) 计算隶属矩阵, 不断重复上述过程直到根据公式 (2) 得到本次迭代和上次迭代的价值函数的绝对差小于某个阈值为止.

实验系统的第三个模块用来验证上述算法的有效性, 主要工作如下:

(1) 选取合适数据集, 验证实验的有效性;

(2) 选取理想评价指标, 衡量推荐准确度;

(3) 为了使本文提出的新算法有比较理想的推荐效果, 通过给相关参数选取不同数值, 对比推荐结果的准确度, 从而选取理想参数;

(4) 为了验证新算法的有效性, 通过给图 1 中系统结构的推荐模型选择本文提出的新算法和其他相关典型算法, 分别进行实验, 对比结果.

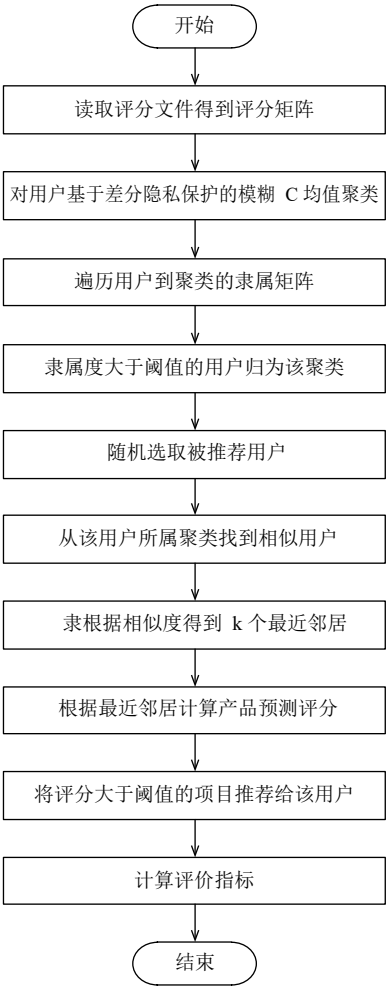

图 2 基于差分隐私保护的模糊 C 均值聚类推荐流程图

在以上设计中, 相关典型对比算法的选取是比较关键的问题. 由于本文提出的是基于差分隐私保护的模糊 C 均值聚类推荐算法, 为了验证模糊 C 均值聚类对传统硬聚类问题的改善, 设计其与典型的基于用户的 K-means 聚类推荐算法相比较; 由于新算法是对用户聚类再进行推荐, 为了验证聚类算法的推荐准确度不低于协同过滤算法, 将新算法、基于用户的 K-means

聚类推荐算法和基于用户的协同过滤推荐算法进行比较.

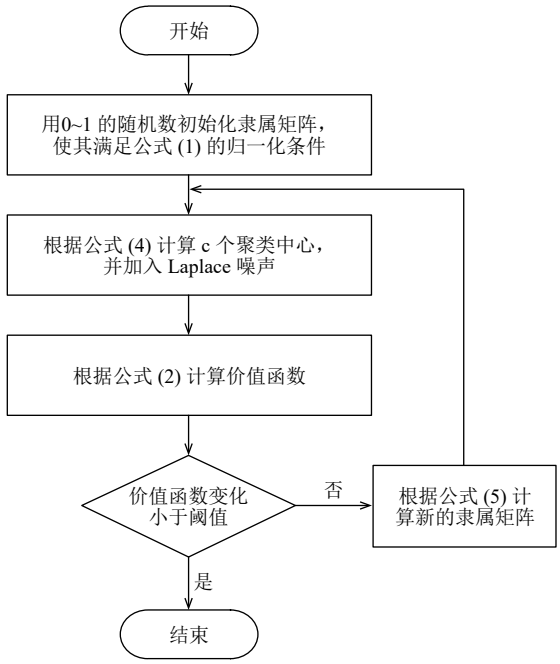

图 3 基于差分隐私保护的模糊 C 均值聚类处理流程

3 实验结果与分析

3.1 数据集

实验用到的数据集采用使用的最多的 MovieLens 数据集. 该数据集是 GroupLens Research 项目从 MovieLens 网站上获取的真实数据, 提供的数据量大, 具体真实. 其中, 用户信息主要有: 性别, 年龄, 职业; 电影信息主要有: 电影名称, 电影类别; 评分信息有: 用户 id, 电影 id, 评分, 评价时间, 用户至少对 20 部电影评价过, 评分范围为 1 到 5, 数越大代表喜欢程度越高. 本次实验数据采用 MovieLens 中的 100k 数据集, 包含 943 个用户对 1682 部电影的 100 000 个评分记录.

3.2 评价指标

为了证明存在光流扰动现象, 通过光流检测算法评价推荐系统的标准主要有统计精度度量 (prediction error)、决策支持精度度量 (IR metrics) 和排名度量方法 (ranking metrics) 三类[4]. 其中统计精度度量方法经常使用的评价指标有均方根误差 (RMSE), 平均绝对偏差 (MAE); 决策支持精度度量经常使用的评价指标是召回率 (recall) 和准确率 (precision)[24].

假设预测的用户评分集合表示为$\{p_1,p_2,\cdots,p_N\}$, 对应的实际用户评分集合为$\{q_1,q_2,\cdots,q_N\}$, N 代表评分个数.

(1) 均方根误差 (*RMSE*), 越小意味着推荐越准确. 定义为:

$$RMSE=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2/N} \tag{8}$$

(2) 平均绝对偏差 (*MAE*)[5]指标通过计算预测的用户评分与实际的户评分之间的偏差度量预测的准确性, 越小意味着推荐越准确, 定义为:

$$MAE=\sum_{i=1}^{n}|p_i-q_i|/N \tag{9}$$

(3) *F-measure* 指标[25]评价推荐的质量, 由召回率和准确率将两者结合组成, 其中召回率反映待推荐项目被推荐的比率, 准确率表示算法推荐成功的比率. *F-measure* 值越大推荐质量越高, 计算公式如下:

$$F-measue=2\times recall\times prescison/recall+presison \tag{10}$$

3.3 实验比较和分析

为了消除光流扰动效应, 避免在场景中没有运动目标将原始数据集按 70%/30% 比例随机分为训练数据集与测试数据集, 实验的结果是对所有结果取均值.

为了得到良好的推荐效果, 首先将本文提出的基于差分隐私的模糊 C 均值聚类推荐算法设置不同参数找到较好的推荐效果. 设置聚类个数范围为 10–50, 分析不同聚类个数对推荐质量的影响, 实验结果如图 4; 设置最近邻个数为 10–120, 分析不同最近邻个数对推荐质量的影响, 实验结果如图 5.

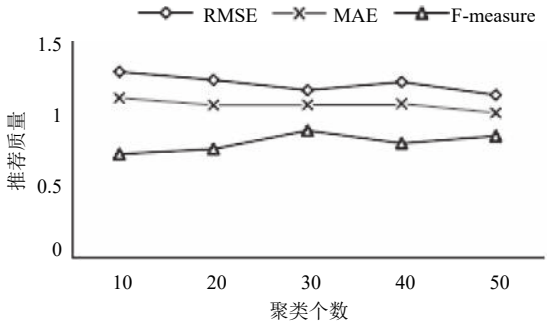

图 4 DPFCMC 聚类个数对推荐质量的影响

根据图 4 发现, 本文提出的基于差分隐私的模糊 C 均值聚类推荐算法在数据集上的聚类个数为 30 或者 50 时, RMSE、MAE 值相对较小, F-measure 值相对

© 中国科学院软件研究所 http://www.c-s-a.org.cn

较大, 聚类效果较好, 有比较理想的推荐准确度.

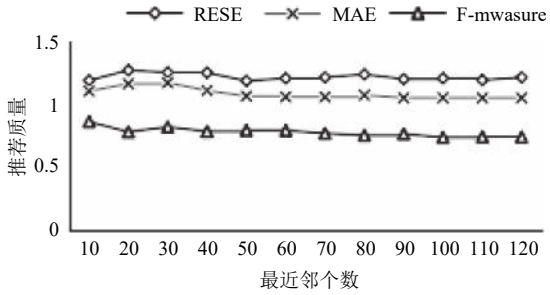

图 5 DPFCMC 最近邻个数对推荐质量的影响

根据图 5 发现, 本文提出的基于差分隐私的模糊 C 均值聚类推荐算法在数据集上的最近邻个数小于 50 时 RMSE、MAE 值相对较大, 同时 F-measure 值相对也较大; 最近邻个数大于 50 时, RMSE、MAE 值有小幅度的变化, 同时 F-measure 值有小幅度的减小趋势, 推荐质量没有明显变化趋势. 因此最近邻个数对推荐准确度影响不大.

根据以上试验发现聚类个数为 30 或 50 有较好的推荐质量, 最近邻个数对推荐质量没有太明显的影响. 因此在以下实验中, 将本文提出的新算法聚类个数取为 30, 分析在最近邻个数为 10-100 范围内, 本文提出的基于差分隐私的模糊 C 均值聚类推荐算法 (the Differential Privacy protection based Fuzzy C-Means Clustering recommendation, DPFCMC) 与基于用户的协同过滤的推荐算法 (User-Based Collaborative Filtering, UBCF)、基于 K-means 聚类的协同过滤推荐算法 (Collaborative Filtering based on K-Means clustering, CFKM) 的推荐准确度. 通过以上提到的对比实验, 来验证新算法的有效性. 实验结果如图 6, 图 7, 图 8 所示, 展示了以上提出的三种算法在 RMSE、MAE、F-measure 三个评价指标的比较结果.

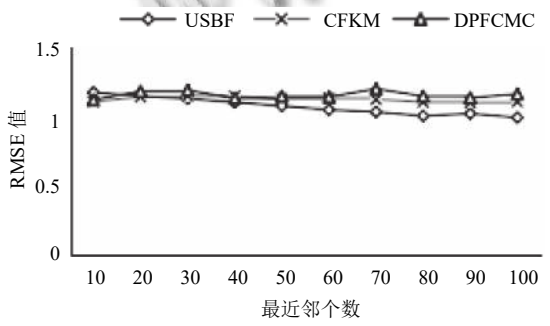

图 6 DPFCMC 与其他典型算法的 RMSE 值对比

根据图 6 结果可知, 在最近邻个数小于 60 范围内, 本文提出的基于差分隐私的模糊 C 均值聚类推荐算法与其他两种算法的 RMSE 值基本持平; 当最近邻个数大于 60, 新算法与基于 K-means 聚类的协同过滤推荐算法的 RMSE 值相差不大, 但这两种聚类算法的 RSME 值都比基于用户的协同过滤推荐算法略大. 因此最近邻个数在一定范围内, 本文提出的新算法与到其他两种算法相比 RSME 值基本持平, 准确度差距不大.

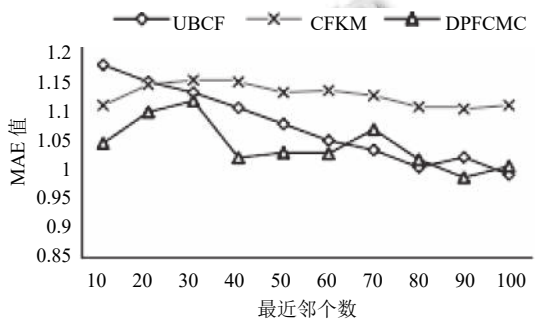

图 7 DPFCMC 与其他典型算法的 MAE 值对比

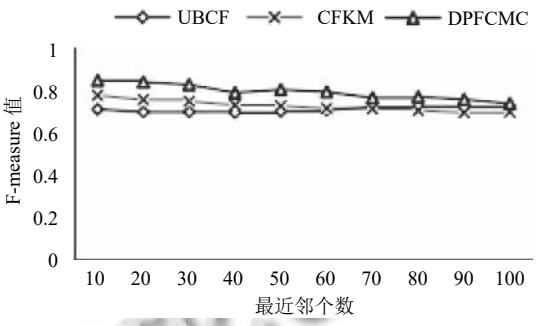

图 8 DPFCMC 与其他典型算法的 F-measure 值对比

根据图 7 结果可知, 本文提出的基于差分隐私的模糊 C 均值聚类推荐算法的 MAE 值大多数比其他两种算法的值小, 有更好的准确度.

根据图 8 结果可知, 本文提出的基于差分隐私的模糊 C 均值聚类推荐算法的 F-measure 值比其他两种算法都大, 有更好的准确度.

综合以上所有实验结果, 可知本文提出的基于差分隐私的模糊 C 均值聚类推荐算法的准确度比基于用户的协同过滤的推荐算法和基于 K-means 聚类的协同过滤推荐算法的准确度更好. 因此, 本文提出的新算法在保护隐私信息的同时保证了更好的准确度.

4 结束语

本文将差分隐私保护方法应用到推荐系统中, 并融合模糊 C 均值聚类, 提出了一种满足差分隐私保护

© 中国科学院软件研究所 http://www.c-s-a.org.cn

的模糊 C 均值聚类推荐算法. 通过获得隶属度函数解决传统硬聚类问题, 同时通过添加满足差分隐私保护的 Laplace 噪声对聚类过程中的聚类中心进行随机干扰. 通过新算法与现有相关典型推荐算法的对比试验证明, 本文提出新的基于差分隐私保护的模糊 C 均值聚类算法能够在保证一定推荐准确度的同时保护用户的隐私信息, 克服了传统聚类推荐算法中的硬聚类和隐私保护问题. 但在聚类数目和初始中心点的选取方面没有适当的算法进行优化, 在保护隐私信息和保证推荐质量之间难以寻找较理想的平衡, 这些将是之后需要继续深入研究的课题.

参考文献

1 Su XY, Khoshgoftaar TM. A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009, 2009: 421425.

2 Herlocker JL, Konstan JA, Terveen LG, *et al*. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 2004, 22(1): 5–53.

3 Goldberg D, Nichols D, Oki BM, *et al*. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992, 35(12): 61–70. [doi: 10.1145/138859.138867]

4 邓爱林, 朱扬勇, 施伯乐. 基于项目评分预测的协同过滤推荐算法. 软件学报, 2003, 14(9): 1621–1628.

5 张光卫, 李德毅, 李鹏, 等. 基于云模型的协同过滤推荐算法. 软件学报, 2007, 18(10): 2403–2411.

6 Zhang F, Chang HY. A collaborative filtering algorithm embedded BP network to ameliorate sparsity issue. 2005 International Conference on Machine Learning and Cybernetics. Guangzhou, China. 2005. 1839–1844.

7 Massa P, Avesani P. Trust-aware collaborative filtering for recommender systems. In: Meersman R, Tari Z, eds. On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE. Berlin, Heidelberg. Springer. 2004. 492–508.

8 Konstan JA, Riedl J, Smyth B. Proceedings of the 2007 ACM conference on Recommender systems. ACM Conference on Recommender Systems. Minneapolis, MN, USA. 2007.

9 Chowdhury M, Thomo A, Wadge WW. Trust-based infinitesimals for enhanced collaborative filtering. International Conference on Management of Data. Mysore, India. 2010. 9–12.

10 曾春, 邢春晓, 周立柱. 个性化服务技术综述. 软件学报, 2002, 13(10): 1952–1961.

11 Borchers A, Herlocker J, Konstan J, *et al*. Ganging up on Information Overload. Computer, 1998, 31(4): 106–108. [doi: 10.1109/2.666847]

12 Kim TH, Park SI, Yang SB. Improving prediction quality in collaborative filtering based on clustering. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. Wi-Iat. Sydney, NSW, Australia. 2008. 704–710.

13 Berget I, Mevik BH, Næs T. New modifications and applications of fuzzy C-means methodology. Computational Statistics & Data Analysis, 2008, 52(5): 2403–2418.

14 Agrawal R, Srikant R. Privacy-preserving data mining. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York, NY, USA. 2000. 439–450.

15 Sweeney L. k-Anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557–570. [doi: 10.1142/S0218488502001648]

16 Chen R, Mohammed N, Fung BCM, *et al*. Publishing setvalued data via differential privacy. Proceedings of the VLDB Endowment, 2012, 4(4): 1087–1098.

17 Sarathy R, Muralidhar K. Some additional insights on applying differential privacy for numeric data. In: Domingo-Ferrer J, Magkos E, eds. Privacy in Statistical Databases. Berlin: Springer, 2010: 210–219.

18 彭慧丽, 张啸剑, 金凯忠. 基于差分隐私的社交推荐方法. 计算机科学, 2017, 44(S1): 395–398, 423.

19 何明, 常盟盟, 吴小飞. 一种基于差分隐私保护的协同过滤推荐方法. 计算机研究与发展, 2017, 54(7): 1439–1451.

20 Li X, Lu X, Tian J, *et al*. Application of fuzzy c-means clustering in data analysis of metabolomics. Analytical Chemistry, 2009, 81(11): 4468–4468. [doi: 10.1021/ac900353t]

21 Dembélé D, Kastner P. Fuzzy C-means method for clustering microarray data. Bioinformatics, 2003, 19(8): 973–980. [doi: 10.1093/bioinformatics/btg119]

22 Dwork C. Differential privacy: A survey of results. In: Agrawal M, Du D, Duan Z, *et al*, eds. Theory and Applications of Models of Computation. Berlin: Springer-Verlag, 2008. 1–19.

23 Dwork C, Mcsherry F, Nissim K, *et al*. Calibrating noise to sensitivity in private data analysis. Halevi S, Rabin T. Theory of Cryptography. Berlin: Springer-Verlag, 2006. 265–284.

24 马宏伟, 张光卫, 李鹏. 协同过滤推荐算法综述. 小型微型计算机系统, 2009, 30(7): 1282–1288.

25 Goldberg K, Roeder T, Gupta D, *et al*. Eigentaste: A constant time collaborative filtering algorithm. Information Retrieval, 2001, 4(2): 133–151. [doi: 10.1023/A:1011419012209]

© 中国科学院软件研究所 http://www.c-s-a.org.cn