Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

基于文献共被引关系的协同过滤文献推荐系统*

□ 李琳娜 张志平 乔晓东 / 中国科学技术信息研究所 北京 100038
刘春霞 / 郑州铁路职业技术学院 郑州 450052

摘要：随着数字图书馆的文献数量和种类高速增长，数字图书馆用户迫切需要有效的个性化推荐工具来帮助其在众多文献中发现对其有价值的文献。协同过滤方法是推荐系统广泛采用的推荐技术，但数据稀疏性是影响其推荐效果的关键因素之一。在文献推荐领域，这一问题更加显著。文章提出了一个利用文献间共被引关系的协同过滤文献推荐方法。实验表明所提方法具有较好的推荐性能。

关键词：数字图书馆，个性化推荐，协同过滤，共被引关系
DOI：10.3772/j.issn.1673—2286.2012.03.008

1 引言

个性化推荐技术能够有效解决信息过载问题，多年来已经在研究上取得了丰富成果，特别是在电子商务领域已取得了很好的应用效果。根据具体的推荐策略，推荐系统一般分为基于内容的推荐系统、协同过滤推荐系统两类，后者是目前应用最广泛的推荐技术，其具体的推荐策略为：向用户推荐与其偏好相似的其他用户选择的项目。由于用户打过分的项目在整个项目空间中通常占少数，因此在发现相似用户群时，会出现相似度计算偏差，从而极大地影响推荐效果。故数据稀疏性（Sparseness）是协同过滤方法需要解决的核心问题[1]。

数字图书馆个性化推荐技术能够面向个体用户，提供符合其个人偏好的数字图书馆内容和服务，减少图书馆内容和个人信息需求之间的差距，已经成为数字图书馆技术发展的重要组成部分和前沿课题。作者面向文献推荐，利用文献计量学知识，提出了利用文献间共被引关系的协同过滤方法，有效地解决了数据稀疏性问题。具体地，基于一个用户的原始打分信息，利用文献间的共被引关系，预测用户对关系文献的打分，磨平原始打分矩阵，提升用户打分矩阵的密度，进而使用协同过滤技术完成推荐。

2 相关工作

数字图书馆个性化推荐系统研究已经存在一些研究成果。作者将这些研究成果大概归结为如下几类：

（1）基于网络结构的推荐技术。Sullivan等首次提出将激活-扩散模型应用到文献推荐领域[2]。文献[3]提出了基于图模型的文献推荐系统，具体的推荐策略是图搜索技术。Watanabe等开发了文献支持系统Papits[4]。该系统基于用户的浏览记录用scale-free网络构建用户模型，然后采用基于内容的推荐技术。文献[5]首先用FPT（Frequent-Pattern-Time）树发现用户的共同兴趣，然后基于神经网络的向后传播算法进行推荐。Gori和Pucci根据文献间的引用关系构建文献图，然后使用改进的pagerank算法进行推荐[6]。文献[7]提出了将本体和扩散激活模型融合的推荐技术。

（2）基于数据挖掘的推荐技术。Agarwal等提出

* 本文获科技部项目“面向外文科技文献信息的知识组织体系建设与应用示范”课题之一“信息资源自动处理、智能检索与STKOS应用服务集成”资金支持。

了基于子空间聚类算法的文献推荐方法[8]。文献[9]和文献[10]都使用基于关联规则的推荐技术。但前者采用蚁群算法对用户聚类；后者根据用户的背景知识使用自适应共振理论将用户聚类。

（3）基于本体的推荐技术。文献[11]提出了利用搜索主题的本体进行推荐的技术。Liao等人提出了文献推荐系统PORE[12-14]。该系统以图书馆本体基础，根据用户的浏览记录构建用户个性化本体，从而完成个性化推荐。Ferran等根据用户的使用记录建构个性化本体，然后基于该个性化本体进行推荐[15]。

（4）基于文献计量学的推荐技术。Citeseer系统利用文献间的引用关系发现相关文献[16]。McNee等提出了将引文网络和CF算法融合的推荐技术[17]。Strohman等认为单纯的基于文本的推荐方法或者基于引文的推荐方法都有各自的缺陷，提出了将二者融合的推荐技术[18]。

（5）基于向量空间模型的推荐技术。文献[19]基于中图分类法和子网对向量的每个元素在歧义层、同义词层和上下位层进行扩展。文献[20]根据审稿人的发表记录，采用发表文献的标题、摘要、关键词及作者信息将其偏好表示为向量空间模型，然后采用基于内容的推荐技术。Gauch等人提出的推荐系统本质上是一个基于内容的推荐[21,22]。但是不再基于向量空间模型表示文献，而是基于概念树。文献[23]实现了文献推荐系统Scienstein，该系统是一个集成引文分析、作者分析、源分析、隐式打分、显式打分等多方面信息的混合推荐系统。文献[24]提出了将评审人的多种信息及论文的多种信息融合的推荐技术。

3 基于共被引关系磨平的协同过滤方法

本部分首先介绍协同过滤方法的一般框架，进而介绍共被引关系，以此为基础提出基于共被引关系的协同过滤文献推荐方法。

3.1 协同过滤一般框架

协同过滤推荐系统向用户推荐与其偏好相似的其他用户选择的项目。这里，用户偏好的表示不是基于项目内容，而是基于其对所有经验项目的打分向量。所有用户对所有项目的打分构成打分矩阵。一个打分矩阵的具体例子如表1所示。

表1 打分矩阵说明

	K-PAX	Life of Brian	Memento	Notorious
Alice	4	3	2	4
Bob	ϕ	4	5	5
Cindy	2	2	4	ϕ
David	3	ϕ	5	2

协同过滤推荐系统的算法可以分为两类[26]：基于记忆的和基于模型的算法。基于记忆的算法根据系统中所有被打过分的项目信息进行预测；基于模型的算法收集打分数据进行学习并推断用户行为模型，进而对某个项目进行预测打分。作者采用的是基于记忆的算法。

基于记忆的协同过滤推荐算法的具体思想为：用户u对未知项目s的预测打分$r_{u,s}$为：与u最相似的N个邻居对的实际打分的某种聚合。形式描述如下：

$$r_{c,s}=\underset{c'\in\hat{C}}{\text{aggr}}\, r_{c',s} \qquad (1)$$

这里。C'是整个用户空间中对项目s打过分并且与用户u最相似的N个用户形成的集合。其中N的范围可以是（1，$|U|$）内的整数。聚合函数aggr可以采用以下几种：

(a) $r_{c,s}=\frac{1}{N}\sum_{c'\in\hat{C}} r_{c',s}$

(b) $r_{c,s}=k\sum_{c'\in\hat{C}} sim(c,c')\times r_{c',s}$

(c) $r_{c,s}=\bar{r}_c+k\sum_{c'\in\hat{C}} sim(c,c')\times(r_{c',s}-\bar{r}_{c'})$ （2）

其中，公式2（a）（b）公式中的乘数k为规范化因子，公式（c）中的$\bar{r}_c$取值为用户打分的平均值。

许多计算两个用户间相似测度方法已经提出。大多数方法根据两个用户的共同打过分项目的打分向量计算相似度。最常见的两个测度为相关性测度和余弦相似度。本文使用余弦相似度作为相似度测度，即两个用户u与v的相似度为：

$$sim(\vec{u},\vec{v})=\frac{\vec{u}\cdot\vec{v}}{||\vec{u}||\times||\vec{v}||} \qquad (3)$$

其中，$\vec{u}$与$\vec{v}$分别表示用户u与v的打分向量。

3.2 共被引关系

文献间的共被引关系最早由Small和Marshakova于1973年分别提出[27, 28]。同时产生的共被引分析方法在科学计量学领域内被众多学者进行广泛的理论和实践研究。下面给出共被引关系及共被引强度的定义。

定义1 共被引：两篇文献被别的文献同时引用时，称这两篇文献有共被引关系，并以引用它们的文献数量作为共被引强度。

共被引矩阵[28]最早由Small提出，它是完全对称的矩阵，对角线选择默认值。具体形式如表2所示。其中，A, B, C, D表示文献，NO.(ij)表示文献i和文献j之间的共被引强度，NO.(ij)=NO.(ji)，对角线(即i=j)为默认值。

表2 对称共被引矩阵

	A	B	C	D
A		NO.(AB)	NO.(AC)	NO.(AD)
B	NO.(BA)		NO.(BC)	NO.(BD)
C	NO.(CA)	NO.(CB)		NO.(CD)
D	NO.(DA)	NO.(DB)	NO.(DC)	

可以如下构造共被引矩阵。用一个有向图表示引文索引，那么，其对应的入射矩阵即可以表示引文索引。在对应的有向图中，若D_i引用D_j或者D_j被D_i引用，那么节点D_i和节点D_j之间有一条有向边；在入射矩阵中，行D_i列D_j所对应位置的值为D_i引用D_j的次数。一篇文档被引用的次数成为引用频率，因此，引文矩阵C展示了引用关系，而C^T展示了被引用关系。共引矩阵和共被引矩阵都可以从引用矩阵计算得到。

3.3 基于共被引关系磨平的推荐方法

稀疏性是协同过滤系统的核心问题之一[1]。对于数字图书馆中海量科技文献，稀疏性体现在用户实际给出打分或者阅读的文献数量相对整个文献空间数量的差距。因此在发现与其相似的用户群时，会出现相似测度计算偏差问题，另外由于相似用户推荐的项目覆盖度不足，进而极大地影响推荐效果。面对该问题，作者提出了基于共被引关系磨平的协同过滤方法。该方法基于用户的原始打分信息，利用文献间的共被引关系，预测用户对关系文献的打分，磨平原始打分矩阵，提升用户打分矩阵的密度，进而使用基于项的协同过滤技术完成推荐。

具体的算法思想为：

1） 对于整个文献空间计算每个文献对之间的共被引强度。

2） 确定一个共被引强度阈值α。

3） 过滤掉小于阈值α的文献对，剩余文献对作为强关联对保留。

4） 在打分矩阵中，对于每个用户u有打分r_{ui}的文献i'，对与其有强关联对关系的每个文献赋予u对之的预估打分r_{ui}。当i'因为有多个强关联对关系存在而获得多个u对其的预估打分时，其预估打分为这多个打分的均值。

5） 在原始打分与预估打分形成的打分矩阵上，根据公式（1）与（2）计算用户u对未知文献的协同预测打分。

合理地，当α取值过小时，虽然打分矩阵密度会有提升，但会引入弱强度文献的磨平，从而造成推荐准确度的相对下降。具体实施时，应根据实验分析，权衡提升矩阵密度和推荐结果准确性，确定α的取值。

4 实验

为了全面评估所提方法的有效性，作者实验对比了所提方法（以下简称CCSCF）与单纯的协同过滤方法（以下简称CF）。

数据集来源于国家科技图书文献中心（以下简称为NSTL）。文献集合是具有共被引关系分析的1000篇文献。用户集是自2009年一年中有下载记录的所有单个用户，按照下载量分为活动频率高与低2组（以在上述文献数据集合中的下载量50为界限值）。整个数据打分以0-1二元打分为记分形式。实验数据划分原则为，全部用户2009年1-2月的下载信息为训练集。2009年3-12月的下载信息为测试集。对推荐文献数为20、30、50、70四种情况进行了实验分析。

评测测度使用决策支持准确度[29]。决策支持准确度评价的是用户是否采纳系统推荐的项。具体评价方法为：选择一个测试用户，将其对一些项的选择隐藏起

来，然后评价用户对系统推荐的项与实际情况的差异。在离线实验环境下，推荐项与隐藏项之间有表3所示四种关系：

表3 决策支持准确度

	Recomended	Not recommended
Used	True-Positive(tp)	False-Negative(fn)
Not used	False-Positive(fp)	True-Negative(tn)

定义如下指标：

$$Precision = \frac{\#tp}{\#tp+\#fp}$$

$$Recall(TruePositiveRate) = \frac{\#tp}{\#tp+\#fn}$$

按照上述实验设计，得到表4所示的实验结果。

基于上述结果，可知所提方法在准确率方面明显高于单纯的协同过滤方法。另外，还应注意到，随着推荐数量的增加，单纯协同过滤的Precision指标有明显下降，Recall指标并没有显著上升；而所提方法Precision指标没有明显下降，同时Recall指标显著上升，进一步说明了所提方法的稳定性和准确性优势。

表4 实验结果

	CF Precision	CCSCF Precision	CF Recall	CCSCF Recall
20	0.36	0.43	0.09	0.15
30	0.35	0.41	0.12	0.19
50	0.28	0.37	0.14	0.21
70	0.17	0.36	0.14	0.29

5 结语

作者提出了一个利用文献间共被引关系的协同过滤文献推荐方法，有效解决了稀疏性这一协同过滤核心问题。实验表明较之于当前单纯的协同过滤方法推荐性能有较大提升。目前作者正在基于此研究成果，着手为NSTL开发个性化推荐系统，以此来提升NSTL网络服务系统的服务能力。

在将来的工作中，作者将引入更大量的数据集合，评测所提方法准确性性能优势的统计充分性；同时引入新颖性与多样性测度，评测所提方法的全面性能优势。

参考文献

[1] SULLIVAN D O, SMYTH B, WILSON D. Preserving Recommender Accuracy and Diversity in Sparse Datasets [J]. International Journal on Artificial Intelligence Tools, 2004, 13(1): 219-236.
[2] WOODRUFF A, GOSSWEILER R, PITKOW J, et al. Enhancing a Digital Book with a Reading Recommender [C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. NY, USA. ACM, 2000: 153-160.
[3] HUANG Z, CHUNG W Y, ONG T H, et al. A Graph-based Recommender System for Digital Library [C]// Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries. NY, USA. ACM, 2002: 65-73.
[4] WATANABE S, ITO T, OZONO T, et al. A Paper Recommendation Mechanism for the Research Support System Papits [C]// 2005 International Workshop on Data Engineering Issues in E-Commerce, Tokyo. Japan. IEEE, 2005: 71-80.
[5] GAO K, WANG Y C, WANG Z Q. Similar Interest Clustering and Partial Back-Propagation-based Recommendation in Digital Library [J]. Library Hih Tech, 2005, 23(4): 587-597.
[6] GORI M, PUCCI A. Research Paper Recommender Systems: A Random-Walk based Approach [C]// Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. Washington, USA. IEEE Computer Society, 2006: 778-781.
[7] WENG S S, CHANG H L. Using Ontology Network Analysis for Research Document Recommendation [J]. Expert Systems with Applications, 2008, 34(3): 1857-1869.
[8] AGARWAL N, HAQUE E, LIU H, et al. Research Paper Recommender Systems: A Subspace Clustering Approach [C]// International Conference on Web-Age Information Management. Springer, 2005, 3739: 475-491.
[9] CHEN C C, CHEN A P. Using Data Mining Technology to Provide a Recommendation Service in the Digital Library [J]. Electronic Library: Library and Information Studies, 2007, 25(6): 711-724.
[10] TSAI C S, CHEN M Y. Using Adaptive Resonance Theory and Data-Mining Techniques for Materials Recommendation based on the E-Library Environment [J]. Electronic Library: Library and Information Studies, 2008, 26(3): 287-302.
[11] MIDDLETON S E, SHADBOLT N R, DE ROURE D C. Ontological User Profiling in Recommender Systems [J]. ACM Transactions on Information Systems, 2004, 22(1): 54-88.

[12] LIAO I E, LIAO S C, KAO K F, et al. A Personal Ontology Model for Library Recommendation System [J]. Digital Libraries: Achievements, Challenges and Opportunities, 2006, 4312: 173-182.
[13] LIAO S C, KAO K F, LIAO I E, et al. Pore: a Personal Ontology Recommender System for Digital Libraries [J]. Electronic Library: Library and Information Studies, 2009, 27(3): 496-508.
[14] LIAO I E, HSU W C, CHENG M S, et al. A Library Recommender System based on a Personal Ontology Model and Collaborative Filtering Technique for English Collections [J]. Electronic Library: Library and Information Studies, 2010, 28(3): 386-400.
[15] NURIA F F, ENRIC M P, JULIA M A. Towards Personalization in Digital Libraries through Ontologies [J]. Library Management, 2005, 26(4/5): 206-217.
[16] GOODRUM A. Scholarly Publishing in the Internet Age: a Citation Analysis of Computer Science Literature [J]. Information Processing & Management, 2001, 37(5): 661-675.
[17] MCNEE S M, ALBERT I, COSLEY D, et al. On the Recommending of Citations for Research Papers [C]// Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, NY, USA. ACM, 2002: 116-125.
[18] STROHMAN T, CROFT W B, JENSEN D. Recommending Citations for Academic Papers [C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, NY, USA. ACM, 2007: 705-706.
[19] YU Z T, ZHENG Z Y, GAO S X, et al. Personalized Information Recommendation in Digital Library Domain Based on Ontology [C]// IEEE International Symposium on Communications and Information Technology, 2005: 1249-1252.
[20] SUN Y H, NI W J, MEN R. A Personalized Paper Recommendation Approach based on Web Paper Mining and Reviewer's Interest Modeling [C]// Proceedings of the 2009 International Conference on Research Challenges in Computer Science, Washington, DC, USA. IEEE Computer Society, 2009: 49-52.
[21] CHANDRASEKARAN K, GAUCH S S, LAKKARAJU P, et al. Concept-based Document Recommendations for Citeseer Authors [C]// Proceedings of the 5th International Conference on Adaptive Hypermedia and Adaptive Web-based Systems, Berlin, Heidelberg. Springer-Verlag, 2008: 83-92.
[22] PUDHIYAVEETIL A K, GAUCH S, LUONG H, et al. Conceptual Recommender System for CiteseerX [C]// Proceedings of the Third ACM Conference on Recommender Systems, NY, USA. ACM, 2009: 241-244.
[23] GIPP B, BEEL J, HENTSCHEL C. Scienstein: a Research Paper Recommender System [C]// Proceedings of the International Conference on Emerging Trends in Computing, Virudhunagar, India. IEEE, 2009: 309-315.
[24] BASU C, HIRSH H, COHEN W W, et al. Technical Paper Recommendation: a Study in Combining Multiple Information Sources [J]. Artificial Intelligence Research, 2001, 14(1): 231-252.
[25] ADOMAVICIUS G, TUZHILIN A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions [J]. IEEE Transaction on Knowledge and Data Engineering, 2005, 17: 734-749.
[27] MARSHAKOVA I V. System of Document Connectionism based on References [J]. Nauchno-Teknicheskaya Informatsiya, 1973(2): 2-6.
[28] SMALL H. Co-citation in the Scientific Literature: a new Measure of the Relationship between two Documents [J]. American Society for Information Science, 1973, 24(4): 265-269.
[29] HERLOCKER J, KONSTAN J, TERVEEN L, et al. Evaluating Collaborative Filtering Recommender Systems [J]. ACM Transactions on Information Systems, 2004, 22(1): 5-53.

作者简介

李琳娜（1981-），博士后。研究方向：数字图书馆，推荐系统。E-mail: liln@istic.ac.cn
张志平（1963-），研究员，研究方向：智能信息检索，数字图书馆技术。
乔晓东（1965-），研究员，研究方向：知识服务技术，数字图书馆和资源管理。
刘春霞（1979-），硕士，研究方向：推荐系统，数据挖掘。

Research Paper Recommender Systems: A Collaborative Filtering Approach Based on Co-citation

Li Linna, Zhang Zhiping, Qiao Xiaodong, Liu Chunxia / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: With the growth of the number and variety of literature in digital libraries, users urgently demand effective personalized recommendation to find out useful publications related to their research interest. Collaborative Filtering is a widely used technique in recommender systems. Focusing on the sparseness which is the major challenge of collaborative filtering, a research paper recommender system which adopted collaborative filtering is presented in this paper. The co-citation relationship is employed simultaneously. Experimental results demonstrated that the proposed method could achieve better recommendation than pure collaborative filtering.

Keyword: Digital libraries, Personalized recommendation, Collaborative filtering, Co-citation

（收稿日期：2011-08-25）