湖南大学硕士专业学位研究生

毕业（学位）论文选题报告

姓　　名	屠晓涵	学　号	S1510W0597	已修学分	33
所属学院	信息科学与工程学院		专业学位领域		计算机技术
指导教师	李仁发、周建国		选题时间		2016.12
研究方向	推荐系统				
论文题目	融合评分和评论的推荐系统研究				

一、课题来源与选题依据

1.研究背景

推荐系统在 90 年代中期成为一个独立的研究领域，研究人员开始关注的是评分方法的推荐问题。在大数据时代，信息存在过载问题，用户获得相关信息变成了一项艰巨的任务。一定程度上该问题可以由搜索引擎解决，但它们不提供信息的个性化展示，并且搜索引擎是主动的获取信息的行为，而推荐是用户被动获得信息。因此，为了进一步过滤信息，我们需要推荐系统。

推荐内容是许多信息系统中的重要任务。例如，诸如亚马逊的在线购物网站给每个客户提供用户可能感兴趣的产品的推荐。YouTube 视频门户向客户推荐电影。从多个维度分析用户信息，结合语义分析，并组合多种推荐技术，能够充分提升推荐系统的效率和准确率。在面对海量高维稀疏的数据时，更能快速分析定位用户的兴趣。如今单个推荐算法已经发挥了极致的作用，从语义分析角度集合用户行为，综合运用多个推荐算法更能提高推荐系统的效率。

推荐系统已经广泛应用于电子商务，电影，音乐，新闻，书籍，在线约会，社交网络等领域。诸如天猫、京东、今日头条、网易云音乐、新浪微博和豆瓣等互联网上的一些流行网站使用的各种 web 推荐系统。

推荐系统同时也遇到了许多挑战，例如物品信息的变化，用户偏好的变化，不可预测的项目，可扩展性问题，隐私保护，新鲜度，过度专门化等，其中最具挑战性的是冷启动和数据稀疏性问题。

2.研究目的

推荐系统通过向用户分配要使用的项目的建议来为此提供解决方案。推荐系统以自动化方式向用户提供“正确”项目以优化长期业务目标。通过算法实现自动化的过程。此外，使用推荐系统可以使企业可以更好地了解客户的购买行为，并制定有效的营销策略来吸引不同的客户。推荐系统目

标是对可能感兴趣的项目或产品的用户集合做出有用且合理的推荐。将语义分析融入到协同过滤等方法中，更深入的的分析用户和物品之间的潜在关系。同时结合现有的主流推荐方法，将评分和评论信息作出适当的权衡，得到超出一般推荐方法准确率的推荐算法。

3.研究意义

在大数据时代，获取信息的速度成为用户效率提升的重要保障之一，相比于搜索引擎，融入评分数据和评论信息的推荐系统更能发掘用户的兴趣爱好，然后向用户提供个性化的产品服务，成为有效获取信息的重要手段。本文所进行的研究在实际中具有一定的理论和实践意义:

1)理论意义

融合评分和评论的推荐系统可以应用到多个学科领域,有数据挖掘技术、信息检索方面、人工智能领域、机器学习及模式识别等,通过深入研究融合评分和评论的推荐技术,综合利用多学科领域的知识也可以在理论上加速各学科的发展。

2)实践意义

首先,为帮助用户快速找到满足自己要求的产品,基于用户评论的个性化产品推荐系统通过采集数据挖掘用户的真实兴趣,然后为目标用户推荐与其兴趣最相似的产品。其次,互联网中的产品量之多,用户在做决定是否购买某一个产品时通常依靠不同用户的对该产品的评价内容,但是在推荐系统对于结合用户评论的研究少之又少。本文的研究弥补了该不足。最后,由于融合评分和评论的推荐系统技术的研究起源比较晚,发展还不成熟,本文的研究工作也有助于对该技术的深入研究。本文针对产品进行推荐,实现从产品评分数据和用户评论信息的建模,采用基于内容和协同过滤的推荐方式为目标用户推荐最合适的产品。

4.国内外研究现状和发展动态

推荐系统生成推荐项的过程中有 2 个重要阶段：数据预处理阶段和推荐生成阶段。在数据预处理阶段，推荐系统需要从数据中获取用户偏好；推荐生成阶段，推荐系统根据用户偏好信息，利用推荐算法，从数据集中生成用户推荐项目。偏好获取技术是指通过跟踪、学习用户的兴趣、偏好以及性格特征等信息，实时、准确地发现不同用户对各种网络服务的需求，并对其变化做出适应和调整。传统的用户偏好获取技术通过显式或隐式的方式获取用户的偏好，主要分为启发式和建模两类。前者利用一些具有直观意义的启发式方法来获取用户需求，如最近邻算法、聚类（Kmeans 算法）、相似度计算等; 后者通过引入机器学习技术学习一个模型，如决策树归纳、贝叶斯分类、聚类等。针对用户偏好随时间迁移的问题，研究者使用一些自适应方法，如信息增补技术、遗传算法和神经网络技术，来解决此问题。

目前，主流的推荐技术主要包括以下几种：基于内容的推荐、协同过滤推荐、基于关联规则的推荐、基于效用的推荐、基于知识的推荐和混合推荐等。

4.1 基于内容的推荐

基于内容的推荐（Content-based Recommendation）是信息过滤技术的延续与发展，它是建立在项目的内容信息上作出推荐的，而不需要依据用户对项目的评价意见，更多地需要用机器学习的方法从关于内容的特征描述的事例中得到用户的兴趣资料。在基于内容的推荐系统中，项目或对象是通过相关的特征的属性来定义，系统基于用户评价对象的特征，学习用户的兴趣，考察用户资料与待预测项目的相匹配程度。用户的资料模型取决于所用学习方法，常用的有决策树、神经网络和基于向量的表示方法等。基于内容的用户资料是需要有用户的历史数据，用户资料模型可能随着用户的偏好改变而发生变化。

基于内容的推荐所基于的基本假设是“一个用户可能会喜欢和他曾经喜欢过的物品相似的物品”。这里“曾经喜欢过的物品”就是利用该用户的历史记录计算出来的 Profile，作为该用户的 User Profile 来使用。具体总结为以下表格所示：

<table>
<tr><td colspan="2">基于内容的推荐算法会根据用户过去喜欢的物品，推荐元数据、描述、主题等类似的物品</td></tr>
<tr><td colspan="2">输入的内容：仅取决于物品及用户的内容/描述（但不包括使用数据）</td></tr>
<tr><td colspan="2">类型：
A.信息检索（比如 TF-IDF 技术等）
B.机器学习（决策树、支持向量机、朴素贝叶斯等）</td></tr>
<tr><td>优点</td><td>缺点</td></tr>
<tr><td>无需冷启动</td><td>物品内容需要让机器读懂</td></tr>
<tr><td>无需使用数据</td><td>容易局限化用户的分类</td></tr>
<tr><td>不存在流行度偏颇，可以根据罕见特性推荐物品</td><td>很难在实现时出现意外之喜</td></tr>
<tr><td>可以通过用户内容特性来提供解释</td><td>很难将多个物品的特征结合在一起</td></tr>
</table>

表格一：基于内容的推荐算法概览

4.2 协同过滤推荐

协同过滤推荐是指收集用户过去的行为以获得其对产品的显式或隐式信息，即根据用户对物品或者信息的偏好，发现物品或者内容本身的相关性、或用户的相关性，然后再基于这些关联性进行推荐。基于协同过滤的推荐可以分基于用户的推荐，基于物品的推荐，基于模型的推荐

（Model-based Recommendation）等子类。用户对物品的喜好或评分矩阵往往是一个很大的稀疏矩阵，为了减少计算量，可采用对物品或用户进行聚类的方法[28]，具体见下表所示：

<table>
<tr><td colspan="3">协同过滤推荐算法是在用户行为中寻找特定模式，来创建用户专属的推荐内容</td></tr>
<tr><td colspan="3">输入的内容：仅取决于使用数据（评分，购买，下载，用户偏好，评论等）</td></tr>
<tr><td colspan="3">类型：
基于相似类型的协同过滤（比如基于兴趣类似的用户或者基于类似的物品）
基于模型的协同过滤（ALS、SVD、贝叶斯网络等）</td></tr>
<tr><td>优点</td><td colspan="2">缺点</td></tr>
<tr><td>不需要对域知识有太深的了解</td><td colspan="2">冷启动的问题</td></tr>
<tr><td>不需要了解用户及物品的特性</td><td colspan="2">需要标准化的产品</td></tr>
<tr><td>大多情况结果都足够冷人满意</td><td colspan="2">对用户和物品的比率需求比较高：1：10</td></tr>
<tr><td></td><td>会受流行度的影响</td><td>很难提供解释</td></tr>
</table>

表格二：协同过滤推荐算法概览

4.3 基于关联规则推荐

基于关联规则的推荐（Association Rulebased Recommendation）是以关联规则为基础，把已购商品作为规则头，规则体为推荐对象。关联规则挖掘可以发现不同商品在销售过程中的相关性，在零 售业中已经得到了成功的应用。管理规则就是在一个交易数据库中统计购买了商品集 X 的交易中有多大比例的交易同时购买了商品集 Y，其直观的意义就是用户在购 买某些商品的时候有多大倾向去购买另外一些商品。

4.4 基于效用推荐

基于效用的推荐（Utilitybased Recommendation）是建立在对用户使用项目的效用情况上计算的，其核心问题是怎么样为每一个用户去创建一个效用函数，因此，用户资料模型很大 程度上是由系统所采用的效用函数决定的。基于效用推荐的好处是它能把非产品的属性，如提供商的可靠性（Vendor Reliability）和产品的可得性（Product Availability）等考虑到效用计算中。

4.5 基于知识推荐

基于知识的推荐（Knowledgebased Recommendation）在某种程度是可以看成是一种推理（Inference）技术，它不是建立在用户需要和偏好基础上推荐的。基于知识的方法因 它们所用的功能知识不同而有明显区别。效用知识（Functional Knowledge）是一种关于一个项目如何满足某一特

定用户的知识，因此能解释需要和推荐的关系，所以用户资料可以是任何能支持推理的知识结构，它可以 是用户已经规范化的查询，也可以是一个更详细的用户需要的表示。

4.6 混合推荐

由于各种推荐方法都有优缺点，所以在实际中，混合推荐（Hybrid Recommendation）经常被采用。研究和应用最多的是内容推荐和协同过滤推荐的组合。最简单的做法就是分别用基于内容的方法和协同过滤推荐方法去产生一个推荐预测结果，然后用某方法组合其结果。尽管从理论上有很多种推荐组合方法，但在某一具体问题中并不见得都有效，混合推荐一个最重要原则就是通过组合后要能避免或弥补各自推荐技术的弱点。

混合推荐系统是推荐系统的另一个研究热点，它是指将多种推荐技术进行混合相互弥补缺点，从而可以获得更好的推荐效果。最常见的是将协同过滤技术和其他技术相结合，以克服 Cold-Start 的问题。具体总结为以下表格所示：

最常见的混合方式的推荐结合协同过滤和基于内容过滤的两种方式，以利用某个算法的优点解决另一个的缺点
输入的内容：通过物品及用户的内容/描述及使用数据
类型： A.混合：将多种不同的推荐算法推荐出来的结果混合在一起，其难点是如何重排序。 B.加权融合：就是将多种推荐技术的计算结果加权混合产生推荐， 最简单的方式是基于感知器的线性混合，首先将协同过滤的推荐结果和基于内容的推荐结果赋予相同的权重值，然后比较用户对物品的评价与系统的预测是否相符，进而不断调整权值 C.切换：根据问题背景和实际情况采用不同的推荐技术。比如，使用基于内容推荐和协同过滤混合的方式，系统首先使用基于内容的推荐技术，如果它不能产生高可信度的推荐，然后再尝试使用协同过滤技术。因为需要各种情况比较转换标准，所以这种方法会增加算法的复杂度和参数化，当然这样做的好处是对各种推荐技术的优点和弱点比较敏感。 D.特性结合：将来自不同推荐数据源的特征组合起来，由另一种推荐技术采用。一般会将协同过滤的信息作为增加的特征向量，然后在这增加的数据集上采用基于内容的推荐技术。特征组合的混合方式使得系统不再仅仅考虑协同过滤的数据源，所以它降低了用户对物品评分数量的敏感度，相反，它允许系统拥有物品的内部相似信息，其对协同系统是不透明的。 E.级联：用后一个推荐方法优化前一个推荐方法。它是一个分阶段的过程，首先用一种推荐技术产生一个较为粗略的候选结果，在此基础上使用第二种推荐技术对其作出进一步精确地推荐。级联型允许系统对某些项避免在后面低优先级的推荐器中

被过滤掉，这些项可能是通过第一种推荐技术被较好的予以区分了的，或者是很少被用户评价从来都不会被推荐的物品。 因为级联型的第二步，仅仅是集中在需要另外判断的项上。

F.特征递增：前一个推荐方法的输出作为后一个推荐方法的输入，它与级联型的不同之处在于，这种方法上一级产生的并不是直接的推荐结果，而是为下一级的推荐提供某些特征。一个典型的例子是将聚类分析环节作为关联规则挖掘环节的预处理：聚类所提供的类别特征，被用于关联规则挖掘中，比如对每个聚类分别进行关联规则挖掘。

G.元层次混合：将不同的推荐模型在模型层面上进行深度的融合，而不仅仅是把一个输出结果作为另一个的输入。比如，User-Based 方法和 Item-Based 方法的一种组合方式是，先求目标物品的相似物品集，然后删掉所有其它的物品(在矩阵中对应的是列向量)，在目标物品的相似物品集上采用 User-Based 协同过滤算法。这种基于相似物品的邻居用户协同推荐方法，能很好地处理用户多兴趣下的个性化推荐问题，尤其是候选推荐物品的内容属性相差很大的时候，该方法性能会更好。

优点	缺点
比单独的某种算法要好	比较难
无冷启动问题	平衡时会出现很多问题
不存在流行度偏颇,可以根据罕见特性推荐物品	过程繁琐
可以实现多样化	技术要求高

表格三：混合方式的推荐算法概览

引用文献

[1] Sugiyama K, Hatano K, Yoshikawa M. Adaptive web search based on user profile constructed without any effort from users[C]//Proceedings of the 13th international conference on World Wide Web. ACM, 2004: 675-684.

[2] Wietsma R T A, Ricci F. Product Reviews in Mobile Decision Aid Systems[C]//PERMID. 2005: 15-18.

[3] Ricci F, Wietsma R T A. Product reviews in travel decision making[J]. Information and communication technologies in tourism 2006, 2006: 296-307.

[4] Aciar S, Zhang D, Simoff S, et al. Informed recommender: Basing recommendations on consumer product reviews[J]. IEEE Intelligent systems, 2007, 22(3): 39-47.

[5] Tintarev N, Masthoff J. A survey of explanations in recommender systems[C]//Data Engineering Workshop, 2007 IEEE 23rd International Conference on. IEEE, 2007: 801-810.

[6] S. Funk, Stochastic gradient descent. (2006) ttp://sifter.org/simon/journal/20061211.html

[Online; accessed 6-June-2012]

[7] A. Paterek, Improving regularized singular value decomposition for collaborative filtering, in Proceedings of 13th ACM International Conference on Knowledge Discovery and Data Mining—KDD'07 (2007), pp. 39–42

[8] Y. Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model, in Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'08 (2008), pp. 426–434

[9] Y. Koren, Collaborative filtering with temporal dynamics, in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'09 , pp. 447

[10] Zhou Y, Wilkinson D, Schreiber R, et al. Large-scale parallel collaborative filtering for the netflix prize[C]//International Conference on Algorithmic Applications in Management. Springer Berlin Heidelberg, 2008: 337-348.

[11] Kim C, Kim J. A recommendation algorithm using multi-level association rules[C]//Web Intelligence, 2003. WI 2003. Proceedings.IEEE/WIC International Conference on. IEEE, 2003: 524-527.

[12] 刘庆鹏，陈明锐. 优化稀疏数据集提高协同过滤推荐系统质量的方法[J].计算机应用,2012,32(4) : 1082 -1085.

[13] 薛福亮,张慧颖.应用 WUM 和 RBFN 补值的协同过滤推荐研究[J].计算机工程与应用, 1002-8331 (2012) 09-0022-05.

[14] 吕成戍，王维国，丁永健. 基于 KNN_SVM 的混合协同过滤推荐算法[J].计算机应用研究, 1001-3695(2012) 05-1707-03.

[15] Covington P, Adams J, Sargin E. Deep neural networks for youtube recommendations[C] //Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016: 191-198.

[16] Elkahky A M, Song Y, He X. A multi-view deep learning approach for cross domain user modeling in recommendation systems[C]//Proceedings of the 24th International Conference on World Wide Web. ACM, 2015: 278-288.

[17] Dziugaite G K, Roy D M. Neural Network Matrix Factorization[J]. arXiv preprint arXiv:1511.06443, 2015.

[18] Kim D, Park C, Oh J, et al. Convolutional Matrix Factorization for Document Context-Aware

Recommendation[C]//Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016: 233-240.

[19] Strub F, Mary J. Collaborative Filtering with Stacked Denoising AutoEncoders and Sparse Inputs[C]//NIPS Workshop on Machine Learning for eCommerce. 2015.

[20] Strub F, Gaudel R, Mary J. Hybrid Recommender System based on Autoencoders[C]//Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. ACM, 2016: 11-16.

[21] Wu Y, DuBois C, Zheng A X, et al. Collaborative denoising auto-encoders for top-n recommender systems[C]//Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. ACM, 2016: 153-162.

[22] Sedhain S, Menon A K, Sanner S, et al. Autorec: Autoencoders meet collaborative filtering[C]//Proceedings of the 24th International Conference on World Wide Web. ACM, 2015: 111-112.

[23] 曹毅. 基于内容和协同过滤的混合模式推荐技术研究[J]. 中南大学, 硕士学位论文, 2007, 20071122.

[24] Roy S, Guntuku S C. Latent Factor Representations for Cold-Start Video Recommendation[C]//Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016: 99-106.

[25] Shi Y, Karatzoglou A, Baltrunas L, et al. CLiMF: learning to maximize reciprocal rank with collaborative less-is-more filtering[C]//Proceedings of the sixth ACM conference on Recommender systems. ACM, 2012: 139-146.

[26] 项亮. 推荐系统实战[J]. 2012.

[27]N. Tintarev, J. Masthoff, A Survey of Explanations in Recommender Systems, Proceedings of the 2000 ACM conference on Computer supported cooperative work, CSCW 2000.

[28]M. O'Cornor, Jon Herlocker, Clustering Items for Collaborative Filtering, Proceedings of the ACM SI GIR Workshop, SIGIR 1999

[29]刘建国，周涛等，个性化推荐系统评价方法综述，复杂系统与复杂性科学第 6 卷第 3 期,2009

[30]Kim D, Park C, Oh J, et al. Convolutional Matrix Factorization for Document Context-Aware Recommendation[C]// ACM Conference on Recommender Systems. ACM, 2016.

二、主要研究内容及技术路线

1.研究内容

推荐系统遭受许多挑战，例如缺乏数据，数据的变化，用户偏好的变化，不可预测的项目，可扩展性，隐私保护。包括冷启动，隐私，过度专门化，可扩展性，稀疏性，新鲜度等。

由于用户数目的大量增长，而且用户之间选择存在差异性，使得用户的评分差别非常大。同时推荐对象的数量也大量增长，使得大量的推荐对象没有经过用户的评价。这些会导致部分用户无法获得推荐，部分推荐对象得不到推荐，这就是稀疏性问题。

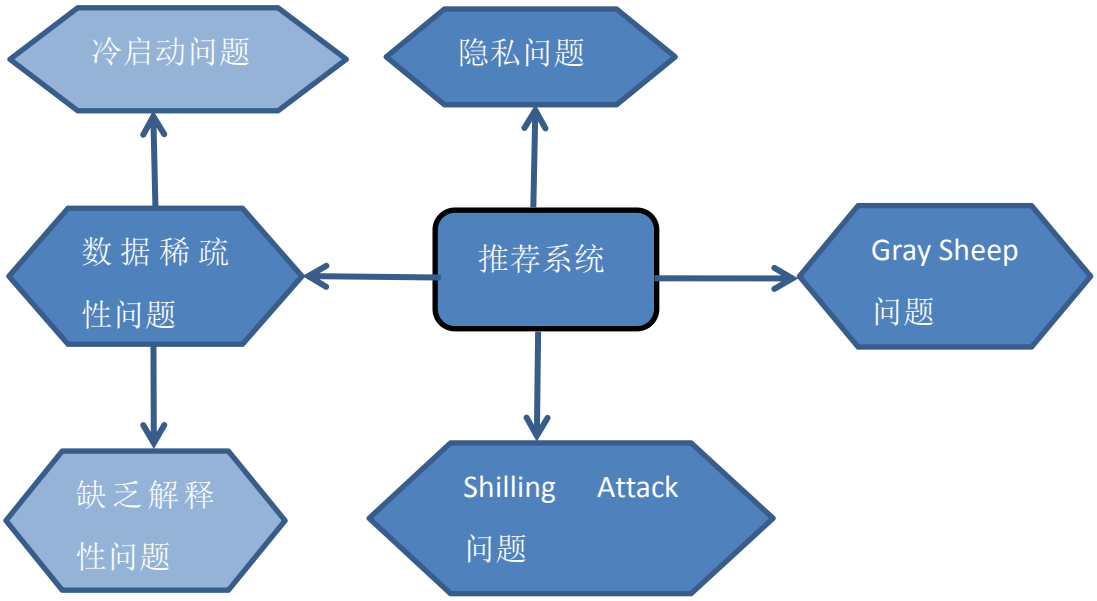

图 1 推荐系统面临的问题

数据稀疏性问题表现在：冷启动问题；Reduced Coverage 问题，rating 相对于 item 来说太少了；Neighbour Transitivity Problem，因为数据量过小，没有用户对相同的 item 进行过评价，因此没法计算他们之间的相似性（例如在电影推荐系统中，有很多电影只被小部分用户评级，而且这些电影会很少被推荐，即使那小部分用户给予很高评级。同样，对于那些有着不同品味的小众群体，找不到相同特定同口味的用户，也导致较差的推荐结果了）。目前数据稀疏性问题解决方法主要有以下几种方式：

1)降维技术

通过奇异值分解（Singular Value Decomposition）来降低稀疏矩阵的维度，为原始矩阵求的最好的低维近似，但是存在大数据量运算成本及对效果的影响等问题（因为经过了 SVD 变化之后，一些 insignificant 用户或者物品被扔掉了，对这类用户或者物品的推荐效果就要打折扣，小众群体的存在体现不出来）。Funk[6]首先将基于 SVD 的矩阵分解应用到推荐算法中，也叫做 SGD 算法

（Basic SVD：Funk SVD)。Paterek[7]通过将矩阵分解与基线估计相结合来提高 SGD。Koren [8]提出了 SVD ++，使用偏置的 SVD，基于平均评分，用户电影偏置，与矩阵分解组合的基线估计。 使用隐式评分扩展了偏置的 SVD,并且在矩阵分解期间仅考虑相关邻域项。Koren [9]提出了 time-SVD ++，他扩展了以前的 SVD ++模型，将时间信息考虑在内。这些矩阵分解方法都是基于 SVD 并且使用梯度下降方法来求解这个问题。不同的是 Zhou 等[10]使用交替最小二乘法与正则化（ALSWR）算法来求解矩阵分解问题。ALS 算法与 SVD 不同的是，SVD 共享很多的变量，只能应用于较小规模或者大容量的共享内存中求解，而 ALS 算法能够很好地扩展到大规模集群计算，利用分布式并行计算来求解，大大加快了求解速度，并且适应当今大数据的批量计算与云计算的架构模型。

2)填值法

通过对矩阵中已知偏好的研究，对缺少项进行填充，从而降低用户偏好矩阵的稀疏性。如可以通过多层次关联规则挖掘方法对用户-项目矩阵进行填充[11]，从而从根本上解决协同过滤算法的数据稀疏性问题。或者采用综合均值优化填充来解决数据稀疏性[12]。或者通过 web 数据挖掘，获得隐性数据来填充评分矩阵，然后采用径向基函数进行平滑处理，来解决数据稀疏性[13]。再者使用 k 邻近方法对数据填充，后利用支持向量机交叉验证分类，解决数据稀疏性等问题[14]。

3)深度学习

近些年来，深度神经网络成为研究的热点，深度学习方法结合协同过滤方法成为解决数据稀疏性问题的重要方法。深度神经网络可以很大程度上替代某些特征工程的过程，自动化地完成特征的提取，解决推荐系统面临数据不足的问题。YouTube 代表了现有最大规模和最复杂的工业推荐系统之一，但依然注重由深度学习带来的性能改进[15]。

[16]提出一个基于内容的推荐系统，以解决推荐质量和系统的可扩展性，根据他们的网络浏览历史和搜索查询使用丰富的功能集来表示用户。并使用深度学习方法将用户和项目映射到潜在空间，其中用户及其首选项目之间的相似性被最大化，通过引入多视图深度学习模型来扩展模型，以共同学习来自不同领域和用户特征的项目的特征，解决了数据稀疏性问题。矩阵因式分解通过简单的固定函数，即作为对于对应的行和列的潜在特征向量的内积来近似矩阵的条目。[17]使用一个任意函数替代内积，从数据中学习在学习潜在特征向量的同时，用多层前馈神经网络替换内积，并通过交替优化固定潜在特征的网络和优化固定网络的潜在特征来学习，稀疏性问题得到解决。[18]提出了卷积矩阵分解（ConvMF)，一种上下文感知推荐模型，将卷积神经网络（CNN）集成到概率矩阵分解中（PMF)，以此来捕获文档的上下文信息，解决了数据稀疏性问题。

虽然稀疏输入受到很少的关注，但仍然是神经网络中一个非常具有挑战性的问题。[19]介绍了

一个神经网络架构，从稀疏评级输入计算非线性矩阵分解。采用基于自动编码器的架构，这是很有前景的一种方法。[20]通过使用输入具有缺失值的数据的损失函数，以及通过并入边信息来增强自动编码器架构。实验表明边信息略微改善了对所有用户/项目的平均测试误差，但它对冷用户/项目具有更大的影响。[21]提出协作去噪自动编码器（CDAE）方法，使用去噪自动编码器架构，得到 Top-N 推荐。[22]提出了 AutoRec，它是一种采用协同过滤（CF）算法的新型自动编码器框架。在线服务在很大程度上依赖于自动个性化来向大量用户推荐相关内容，这要求系统可以迅速扩展以适应第一次访问在线服务的新用户流。AutoRec 是一种紧凑和高效的可训练模型，能很好的解决数据稀疏性问题。

1.2 冷启动问题

数据稀疏性问题的一具体表现是冷启动问题。这里单独列出进行研究。冷启动有新用户和新项目两种情况，当项目新进入系统时，没有或很少用户评价，系统很难推荐这个项目，从而导致恶性循环。当新用户进入系统，未产生行为数据，系统无法根据其历史行为进行推荐。目前解决方案研究主要有以下两种：

(1)对于新项目问题，主要使用组合基于内容和协同过滤进行推荐。协同过滤推荐算法忽略了项目本身的特征信息，不能推荐没有用户评分的项目也不能对没有评分的新用户推荐项目，如果充分利用各种信息，联合基于内容的协同过滤算法，产生的推荐结果会更加精确[23]。

矩阵分解的潜在因子表示携带用户和项目的情感因子的有价值的信息。从视频情感建模的角度建立有效的推荐系统来研究这些潜在因子。[24]提出了一种基于建模用户和项目之间的情感联系的潜在学习因子表示视频的新方法。并提出了情感建模方法，潜在因子表示视频内容的情感建模的功效，visual-CLiMF[24]方法用于基于隐式反馈，解决冷启动问题。Visual-CLiMF 是基于流行的 CLiMF[25]的改进方法，表明 Item 的情感上下文可以用作辅助信息，以提高 MRR 性能。这是一种将辅助内容信息与协同信息结合起来解决视频推荐中的冷启动问题的新方法。

(2)对于新用户问题，一种解决方法是通过用户注册时填写的人口统计学信息给用户提供粗粒度的个性化推荐。另一种方法是在新用户第一次访问推荐系统时，不立即给用户展示推荐结果，而是给用户提供一些物品，让用户反馈他们对这些物品的兴趣，然后根据用户反馈提供推荐[26]。

1.3 Gray Sheep 问题

问题表现为有些人的偏好与任何人都不同（Black Sheep 是那些偏好与正常人完全相反，根本没有办法向他们推荐的人群，因为在现实中也很难解决这个问题，因此这一般被认为是（acceptable failure）。通常采用 Hybrid（结合 Content-based 和 CF）方法解决该问题。

1.4 Shilling Attack 问题

实际上是 AntiSpam 的问题，有些人对于自己的东西或者对自己有利的东西打高分，竞争对手的东西打低分，这会影响协同过滤算法的正常工作。被动的解决办法可以是采用 Item-Based（在 shilling attack 这个问题上，Item-based 的效果要比 User-based 好，因为作弊者总是较少数，在计算 Item 相似度的时候影响较小），Hybrid（能够部分的解决（bias injection problem）；主要的解决办法是采用 AntiSpam 的技术识别和去除作弊者的影响。

1.5 可解释性

推荐系统作为人工智能领域的一个重要产品，要想受到广泛接受和应用，最核心的是要提高推荐结果的合理性，也就要求推荐结果具有较好的可解释性，虽然这一点人们早就意识到，但是针对可解释性的专门研究还很欠缺[27]，当前的研究中，对推荐算法可解释性讨论一般是算法评估之后的选作环节。随着用户要求的提高，“推荐理由”的研究在产业界和学术界都受到越来越高的重视。融合评分数据和评论信息的推荐系统进行语义分析可以增强推荐系统的可解释性。

2.融合评分和评论的推荐系统技术栈

Hadoop 是 Apache 软件基金会旗下的一个开源大数据基础平台，核心组成部分包括 YARN（Yet Another Resource Negotiator）资源管理框架，HDFS（Hadoop Distributed File System）分布式文件系统和 MapReduce 计算框架。Hadoop 为用户提供了系统底层细节透明的分布式基础架构。HDFS 的高容错性、高伸缩性等优点允许用户将 Hadoop 部署在低廉的硬件上，形成分布式系统；MapReduce 分布式编程模型允许用户在不了解分布式系统底层细节的情况下开发并行应用程序。所以用户可以利用 Hadoop 轻松地组织计算机资源，从而搭建自己的分布式计算平台，并且可以充分利用集群的计算和存储能力，完成海量数据的处理。Hadoop 利用这些优势，得以在大数据处理应用中广泛应用得益于其自身在数据提取、变形和加载（ETL）方面上的天然优势。Hadoop 的分布式架构，将大数据处理引擎尽可能的靠近存储，对例如像 ETL 这样的批处理操作相对合适，因为类似这样操作的批处理结果可以直接走向存储。Hadoop 的 MapReduce 功能实现了将单个任务打碎，并将碎片任务（Map）发送到多个节点上，之后再以单个数据集的形式加载（Reduce）到数据仓库里。

Spark 是 Apache 软件基金会旗下的一个基于分布式内存计算平台，是 UC Berkeley AMP lab 所开源的类 Hadoop MapReduce 的通用并行框架，Spark 拥有 Hadoop MapReduce 所具有的优点；但不同于 MapReduce 的是 Job 中间输出结果可以保存在内存中，而不需要读写 HDFS，因此 Spark 能更好地适用于数据挖掘与机器学习等需要迭代的 MapReduce 的算法。Spark 基于 RDD（Resilient Distributed Datasets）的思想构建了一体化和多元化的大数据处理体系，提出一站式的解决方案，通过 SparkSQL、

Spark Streaming、MLlib 和 GraphX，SparkR 组件解决了大数据中批处理（Batch Processing）、流式处理（Streaming Processing）和即席查询（Adhoc Query）等三大核心问题。并且在各组件之间无缝地共享数据和操作，Spark 已经在流处理、图计算、机器学习和结构化数据查询等方面取得了重要成果，成为当今大数据领域最热门的计算框架。

如图 2 是融合评分和评论的推荐系统技术栈，将 Hadoop HDFS、Spark、推荐算法紧密结合在一起。

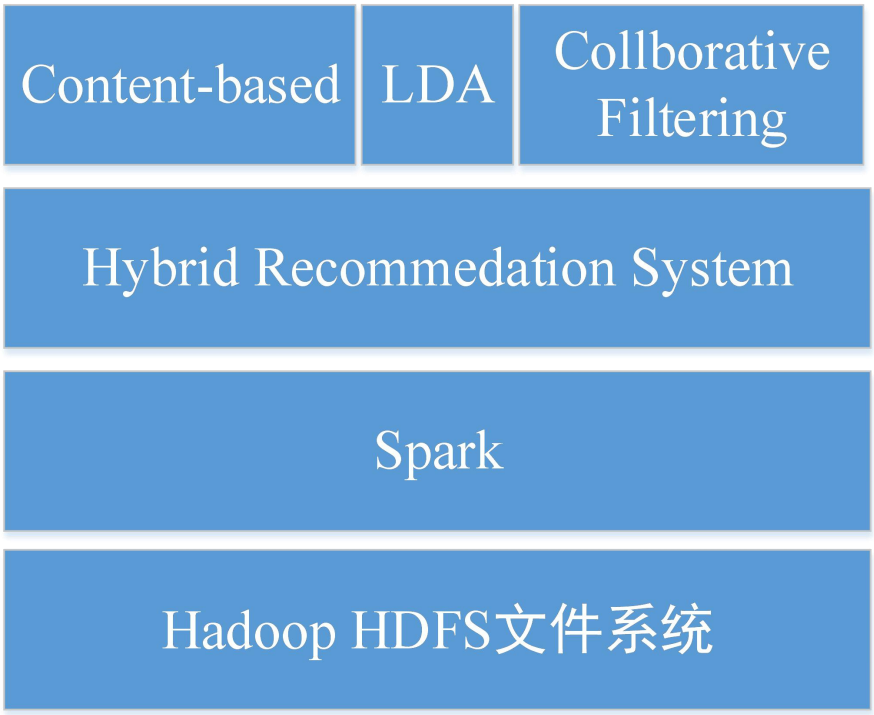

图 2 融合评分和评论的推荐系统技术栈

3.研究方法

综合采用实证研究法和定量分析法进行研究和实验。

（1）通过实证法将设计融合评分和评论的推荐算法，在 Spark 平台实现，并结合实际的数据集进行实验。整个过程涉及到模型及算法的设计和实现。

（2）通过定量分析融合评分和评论的推荐模型的评估参数，判断推荐的准确率和效率。推荐系统的评价指标主要划分为 4 个方面，分别是准确度、基于排序加权、覆盖率以及多样性和新颖性。其中，主要的准确度指标有平均绝对误差（MAE，Mean Absolute Error）、平均正确率均值（MAP，Mean Average Precision）、平均百分比排序（MPR，Mean Percentage Ranking）、均方根误差（RMSE，Root Mean Square Error）、AUC（Area under Curve）指标、平均排序分、准确率（Precison）、召回率（Recall）。主要的加权排序指标有折扣累积利润（DCG，Discounted Cumulative Gain）和排序偏差

准确率。主要的覆盖率指标有预测覆盖率、推荐覆盖率和种类覆盖率。主要的新颖性指标有 UE（unexpectedness），多样性没有较为统一的评价指标。对于融合评分和评论的推荐，有准确性指标如平均 Accuracy、RMSE、GMAE（Group MAE）等。

4.技术路线

如图 3 所示，融合评分和评论的推荐系统架构图。定义了源数据层，数据预处理层，存储层，计算层，接口层和控制层。元数据层收集日志数据和数据库数据。数据预处理层对源数据进行降维和去噪工作。存储层负责将预处理后的数据存储在 HDFS 或者 Hive 数据仓库中，数据文件的存储格式可以定义为 Parquet 或者 ORC 格式。计算层采用 Spark 进行分布式并行计算。接口层用于查询和可视化推荐的结果。管理层实现推荐程序的管理和执行，实现整个大数据分析和处理流程的自动化和直观。

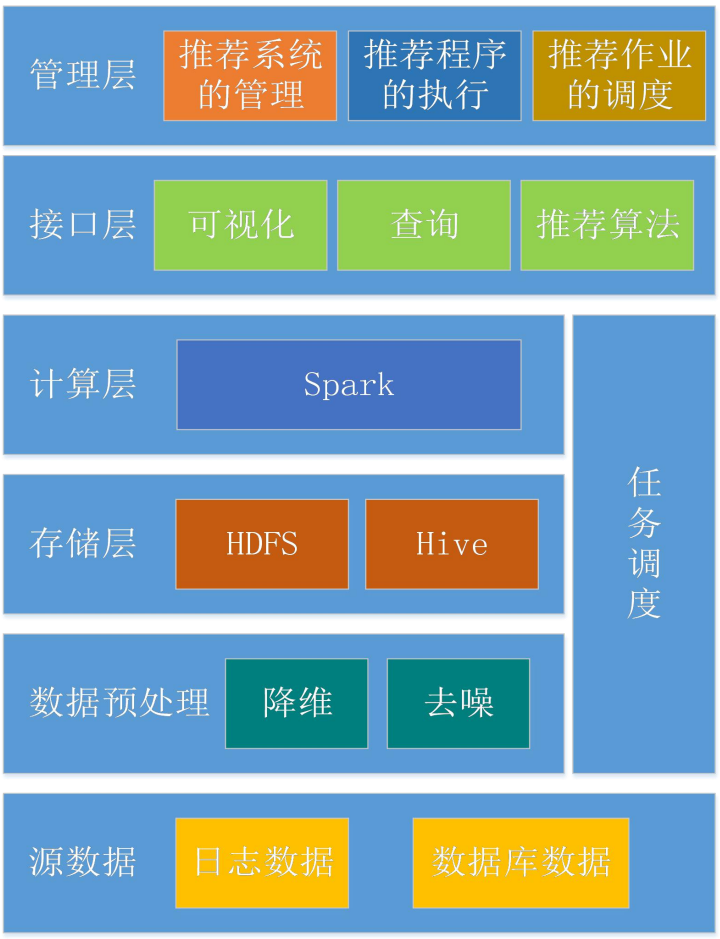

图 3 融合评分和评论的推荐系统架构图

5.实施方案

部署 5 个节点的 Spark 集群，利用 Hadoop HDFS 文件系统存储实验数据。部署结构如图 4 所示。

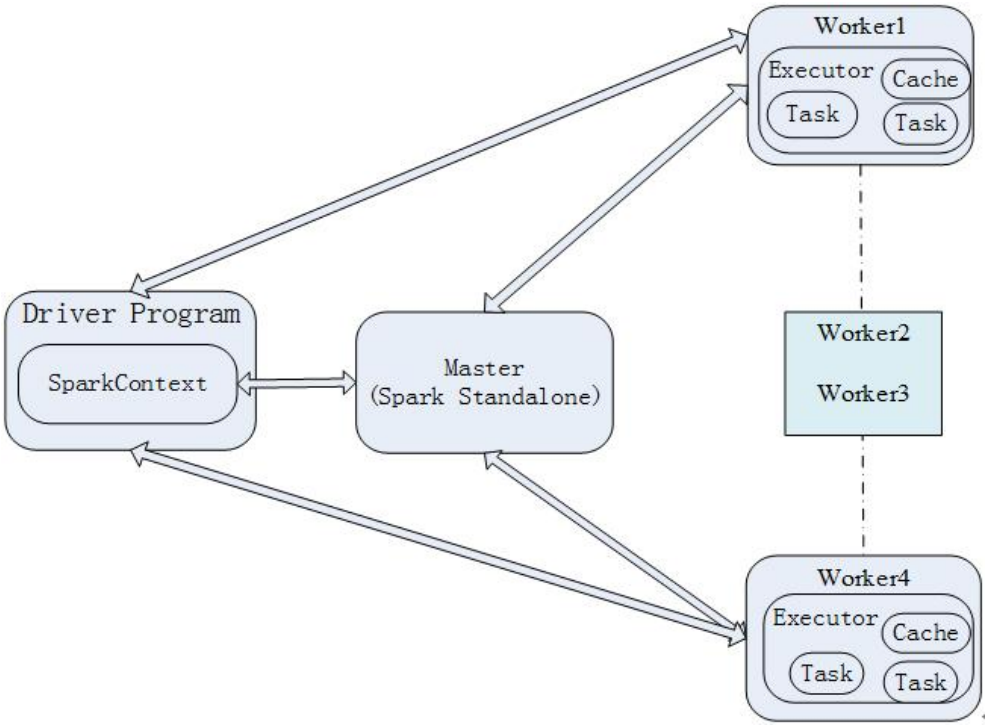
Worker1
Executor
Cache
Task
Task
Driver Program
SparkContext
Master
(Spark Standalone)
Worker2
Worker3
Worker4
Executor
Cache
Task
Task

图 4 实验环境流程图

三、研发基础及进度安排

1.研发基础

根据研究的问题，有针对性的收集资料和阅读文献，对大数据实时处理和推荐系统领域的发展现状有一个初步的了解和掌握，并着重关注该融合评分和评论的推荐存在的问题及挑战，以及解决这些问题的前沿方法和技术。

2.进度安排

1.文献阅读

（1）阅读前沿学术论文，撰写开题报告与答辩 PPT，完成开题答辩

（2）阅读推荐系统原理相关书籍及国内外相关前沿研究动态

2.熟悉和搭建大数据处理平台

（1）学习 Hadoop 大数据平台和 Spark 分布式计算框架

（2）搭建 Hadoop HDFS 分布式文件存储平台和 Spark 分布式计算平台

（3）熟悉 Spark MLlib 基础库和 RDD 编程模型，实现基于 LDA 算法和协同过滤推荐算法等

3.研究数据的预处理，融合评分和评论的推荐模型的建立和推荐算法的改进

（1）数据预处理，结合基于 LDA 和协同过滤等算法的改进及实验

（2）建立评分、评论聚合框架，既可处理评分数据，也可处理评论信息；融合评分和评论的推荐算法性能和效果的分析和评估

（3）采用 Java 开发技术和大数据处理技术，完成基于 web 的融合评分和评论的推荐程序的管理，推荐程序的执行，推荐作业的调度

4.撰写小论文

5.完成学位论文，毕业答辩

四、预期研发成果及创新点

1.预期研究成果

预期成果主要有以下几个部分：

1） 学位论文

2） 实现融入语义分析和评分数据的多模型的推荐算法

3） 发表 1 篇学术论文

4） 完成基于 web 的融合评分和评论的推荐程序的管理，推荐程序的执行，推荐作业的调度，并申请软件著作权

2.创新点

1）建立多模型混合的推荐引擎，运用机器学习无监督聚类算法进行语义分析，解决数据稀疏性问题

2）改进协同过滤推荐算法，有效缓解用户冷启动问题

3）基于 Spark 平台实现推荐算法的并行化，提高推荐的准确率和效率

指导教师意见	指导教师签名：_____ 二〇一 年 月 日
评议小组名单	由学院主管领导确定 3~5 名具有高级职称的教师或校外专家为评议小组成员 组长：_____ 组员：_____、_____、_____、_____、_____ （公章） 主管领导签名：_____ 二〇一 年 月 日
评议小组意见	1、选题价值：☐有理论意义；☐有工程背景；☐有实用价值；☐意义不大。 2、选题难度：☐偏高；☐适当；☐偏低。 3、工作量：☐偏大；☐适当；☐偏小。 4、实施方案的可行性：☐好；☐较好；☐一般；☐不可行。 5、研究生在论文选题报告中反映出的综合能力和表达能力：☐好；☐较好；☐一般；☐较差。 7、研究生在论文选题报告中反映出的创新能力：☐好；☐较好；☐一般；☐较差。 8、对论文选题报告的总体评价：☐好；☐较好；☐一般；☐较差。 （在相应的方块内作记号“√”）
评议结论	评议小组组长签名：_____ 二〇一 年 月 日