

语义推荐算法研究综述

黄震华¹,张佳雯¹,张波²,喻剑¹,向阳¹,黄德双¹

(1. 同济大学电子与信息工程学院,上海 201804;2. 上海师范大学信息与机电工程学院,上海 200234)

摘要: 近年来,语义推荐技术已成为信息服务领域的一个研究热点和重点. 与传统的推荐算法相比,语义推荐算法在实时性、鲁棒性和推荐质量等方面具有显著的优势. 针对语义推荐算法的国内外研究现状、进展,从四个角度进行归纳和总结,即基于语义的内容推荐算法、基于语义的协同过滤推荐算法、基于语义的混合推荐算法以及基于语义的社会化推荐算法,旨在尽可能全面地对语义推荐算法进行细致的介绍与分析,为相关研究人员提供有价值的学术参考. 最后,立足于研究现状的分析与把握,对当前语义推荐算法所面临的挑战与发展趋势进行了展望.

关键词: 语义; 推荐算法; 内容推荐; 协同过滤推荐; 混合推荐; 社会化推荐

中图分类号: TN911. 23 **文献标识码:** A **文章编号:** 0372-2112 (2016)09-2262-14

电子学报 URL: <http://www.ejournal.org.cn>

DOI: 10.3969/j.issn.0372-2112.2016.09.035

Survey of Semantics-Based Recommendation Algorithms

HUANG Zhen-hua¹,ZHANG Jia-wen¹,ZHANG Bo²,YU Jian¹,XIANG Yang¹,HUANG De-shuang¹

(1. College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China;

2. College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China)

Abstract: Semantics-based recommendation technology has recently received a lot of attention in information services community. Compared with traditional recommendation algorithms, semantics-based recommendation algorithms have the marked advantages in the aspects of real-timing, robustness and recommendation quality. From the status and progress of domestic and foreign research, we summarize the following four aspects: semantics-based content recommendation algorithms, semantics-based collaborative filtering recommendation algorithms, semantics-based hybrid recommendation algorithms, and semantics-based social recommendation algorithms. And this paper is expected to provide a worthwhile reference for relevant researchers by detailedly analyzing semantics-based recommendation algorithms. Finally, we show readers the challenges and future research directions in this field.

Key words: semantics; recommendation algorithm; content recommendation; collaborative filtering recommendation; hybrid recommendation; social recommendation

1 引言

近年来,随着物联网、云计算和社会网络等技术的迅猛发展,网络空间中所蕴含的信息量将呈指数级增长^[1]. 例如, Facebook 每月上传的照片超过 10 亿张,每天生成 300TB 以上的日志数据; 淘宝网站每天有超过数千万笔交易, 单日数据产生量超过 50TB; YouTube 线上有数千万部电影, 每天要处理上千万个视频片段; AOL Music 在线音乐网站的音乐库包含有

250 万首歌曲和数千首音乐电视, 每天独立用户访问量达到 2500 万. 不难发现, 信息过载呈爆发趋势, 其结果导致了终端用户无法准确和高效地获取自己感兴趣的物品^[2,3].

目前, 推荐系统是解决信息超载问题最有效的工具之一^[4]. 推荐系统的概念是 AT&T 贝尔研究院的 Paul R 博士在 1997 年提出的, 通过建立用户与物品之间的二元关系, 利用用户的历史记录或物品之间的相似性关系, 挖掘每个用户潜在感兴趣的物品^[5,6]. 不难看出,

收稿日期: 2014-03-28; 修回日期: 2016-03-21; 责任编辑: 梅志强

基金项目: 国家自然科学基金 (No. 61272268); 上海市青年科技启明星计划 (No. 15QA1403900); 教育部新世纪优秀人才支持计划 (No. NCET-12-0413); 国家 973 课题 (No. 2014CB340404); 霍英东基金应用类课题 (No. 142002); 同济大学中央高校基本科研业务费专项资金

推荐系统由三个基本要素组成,即用户、产品和推荐算法,而推荐算法是推荐系统的核心部分,它决定着推荐系统性能的优劣^[7].传统的推荐算法可以归纳为3个大类,即基于内容的推荐算法、协同过滤推荐算法以及混合推荐算法^[8,9].我们将在第二节中,对传统推荐算法的国内外研究进展进行详细的分析.

2006年Loizou A博士在意大利特兰托市召开的推荐系统研讨会(ECAI 2006 Recommender Systems Workshop)上指出:传统的推荐算法由于没有考虑应用场景区的上下文语义,使得在实际应用中,这些算法在实时性、鲁棒性和推荐质量等方面存在严重的不足,并提出了语义推荐算法的概念,其核心思想是将语义知识融合到推荐过程中,来克服传统推荐算法的不足^[10].随后,许多学者开始将语义技术集成进传统各类推荐算法中,来提高总体推荐的效果.近年来,基于语义的推荐技术已成为信息服务领域的一个研究热点和重点^[11].

据我们调研发现,目前大多数的语义推荐算法均是在用户建模阶段利用基于知识的方法来表示用户信息,并且使用标准的网络本体语言例如OWL^[12]进行兴趣建模.这些工作主要通过两个不同的策略将语义技术融入推荐算法的用户兴趣建模中^[13-16]:

(1) 通过扩散激活的方法记录用户的兴趣偏好,并将用户资料表示为语义网络^[17],同时预先计算概念间关系的权重,将一个用户的兴趣偏好分数传播到其他用户的相关兴趣偏好分数计算之中.例如,文献[18]中提出一个种采用扩散激活方法的新闻推荐系统News@hand,该推荐系统能够动态地增加初始的用户长期兴趣偏好记录.

(2) 利用本体中的层次结构对用户的兴趣偏好做出推理.如果用户可能喜欢某个物品,则推断用户也会对该物品所属的一般类别的物品感兴趣,这种策略,我们称为上行扩散法.这种方法能够对用户长期的兴趣偏好推断新的知识,从而丰富用户模型.例如,文献[19]中提出的Quickstep是一个科学著作推荐系统,该系统采用了上行扩散法来构建用户的配置文档.

这两种策略将语义技术应用于推荐算法的用户建模模块中,另外一些语义推荐算法在内容相似度匹配阶段使用语义相似度来提高推荐算法的性能和效果^[20-23].

总的来说,目前推荐系统的国内外综述文献主要针对传统的推荐算法,而极少涉及语义推荐算法.鉴于语义推荐算法的重要研究意义和实用价值,我们有必要跟踪学习和总结该领域现阶段的研究成果,并深入

分析和预测其发展趋势,期望能够更好地指导未来的研究工作.

2 传统推荐算法

传统的推荐算法主要分为三类:基于内容的推荐算法、协同过滤推荐算法以及混合推荐算法.

基于内容的推荐算法^[24-28]根据用户喜好的物品,选择与该物品相似度较高的其他物品生成推荐结果,向该用户进行推荐.该推荐算法首先由系统隐式获取或是由用户显式给出对物品的兴趣偏好属性,然后通过计算和排序用户兴趣偏好与待预测物品描述文档间的相似度,来向用户推荐其可能感兴趣的项目.基于内容的推荐算法由于需要提取物品特征,自动化的特征提取方法很难应用于多媒体数据;同时,为用户推荐的内容仅限于与该用户曾经选择物品相似的物品,结果多样性较差.对于没有选择过任何物品的新用户,存在着新用户冷启动问题^[29].

协同过滤推荐算法分为两种:基于用户的协同过滤算法^[29-32]和基于物品的协同过滤算法^[33-36].基于用户的协同过滤算法是利用与当前用户相似的用户对某些物品的兴趣偏好记录,向当前用户推荐这些物品.基于物品的协同过滤算法是利用与当前物品相似的某些物品,向当前用户推荐这些物品.前者需要计算用户之间的相似度,后者需要计算物品之间的相似度,并为当前用户进行推荐.同样,协同过滤算法也存在冷启动问题,对于新物品和新用户无法进行推荐.另外,协同过滤需要获取用户大量的历史数据,存在数据稀疏问题^[37].

由于基于内容的推荐算法和协同过滤推荐算法各自存在不足之处,研究人员提出了混合推荐的方法^[38-40]来解决单一推荐技术的不足.可以按照不同的混合策略,如加权、特征组合、混合等将不同的推荐技术进行混合推荐.

传统的推荐算法各有优缺点,在不同的研究领域中有各自比较典型的推荐系统应用,具体比较和分析见表1.

从表1我们不难看出,传统的三类推荐算法均不考虑应用场景区的上下文语义,这将在很大程度上影响了推荐的效率和质量.在第3~6节中,我们详细讨论分析目前语义推荐算法的国内外研究进展,主要包括基于语义的内容推荐算法、基于语义的协同过滤推荐算法、基于语义的混合推荐算法以及基于语义的社会化推荐算法.

表 1 传统推荐算法比较

推荐算法	优点	缺点	典型应用
基于内容的推荐算法	<ol style="list-style-type: none"> 1. 不需要获取用户的历史数据; 2. 不存在数据稀疏性问题; 3. 不存在新物品冷启动问题; 4. 由成熟的分类学习算法作支撑. 	<ol style="list-style-type: none"> 1. 受到特征提取方法的限制, 不适用于多媒体数据; 2. 存在新用户问题; 3. 算法可扩展性较差; 4. 推荐结果多样性较差. 	<ol style="list-style-type: none"> 1. Pandora 个性化音乐网络电台^[41] 2. Youtube 个性化视频推荐系统^[42]
协同过滤推荐算法	<ol style="list-style-type: none"> 1. 可应用于多媒体领域; 2. 用户数量增多可以改善算法整体性能; 3. 易于发现用户的兴趣偏好; 4. 可利用历史行为给用户做推荐解释; 5. 推荐结果丰富. 	<ol style="list-style-type: none"> 1. 存在用户数据稀疏性问题; 2. 存在新用户和新物品冷启动问题; 3. 算法可扩展性较差. 	<ol style="list-style-type: none"> 1. 亚马逊个性化推荐列表^[43] 2. GroupLens 新闻推荐系统^[44] 3. MovieLens 电影推荐系统^[45]
混合推荐算法	<ol style="list-style-type: none"> 1. 算法可扩展性较强; 2. 解决数据稀疏性问题; 3. 改善冷启动问题; 4. 可应用于多媒体领域. 	<ol style="list-style-type: none"> 1. 推荐系统框架结构复杂; 2. 难以定义各个推荐算法的权重或推荐结果比例. 	<ol style="list-style-type: none"> 1. Newsweeder 新闻过滤系统^[46] 2. Web-based Cooking Assistant^[47]

3 基于语义的内容推荐算法

传统基于内容的推荐算法通过匹配用户兴趣偏好和待预测物品描述文档间的相似度来产生推荐结果, 这将导致推荐结果过度集中, 缺乏多样性, 并且对于新物品存在冷启动问题. 研究人员发现, 利用语义网络中的语义推理技术, 通过推理物品间的语义关系可以发现新的知识, 从而将增强推荐能力. 另外, 利用语义网络表示数据可以动态地将用户在特定领域的兴趣偏好情景化, 进而在建立用户模型过程中更好地表示用户信息, 以及能够对用户兴趣偏好的不完全信息进行推断^[48].

Victor C 等人^[49]提出一种领域无关的语义网络推荐算法, 利用语义增强技术提高了用户建模和内容匹配的性能. 在用户建模阶段, 算法收集标有语义注释物品的反馈信息, 并利用反馈信息获取用户与物品之间的关联关系, 同时算法赋予每个关联关系权重, 来表示用户对物品的偏好程度. 在内容匹配阶段, 算法对物品的描述是基于具体本体中所存在的概念, 本体概念以层次化结构进行组织, 因而, 层次结构的根结点定义物品类型的属性. 作者通过实验表明该推荐算法能够有效提高推荐的准确率, 并且对于拥有相对较小的兴趣偏好记录的用户则具有较好的推荐效果. Tao P 等人^[50]通过构造加权无向图 (WUG; Weighted Undirected Graph)^[51], 描述物品元数据之间的语义相似度. 每一个物品是无向图中的一个结点, 两个物品的元数据属性之间的相似度决定了结点之间边上的权重, 从而在语义关联图中找到与用户浏览记录相关联的物品, 并通过边上权重排序来产生最终的推荐结果. Pasquale L 等

人^[52]提出基于语义的内容推荐算法 FIRST, 利用机器学习策略来获取物品生产商提供的物品描述信息, 并将这些描述信息与物品标注的关键词相结合来表示用户的兴趣偏好. FIRST 算法在语义化物品描述信息的过程中, 利用 WSD (Word Sense Disambiguation) 自然语言处理技术和 WordNet^[53] 词库来避免物品信息的模糊和不确定性. 另外, FIRST 算法利用有监督学习方法从物品标注的关键词中学习用户兴趣偏好模型, 最后算法匹配用户兴趣偏好模型和物品语义信息, 从而产生准确的推荐结果. Sébastien H 等人^[54]提出了利用 RDF (Resource Description Framework) 知识库^[55]中的实例来定义语义相似度的方法. 该方法将 RDF 知识库映射成一张图, 其中顶点由类、实例和不同类型的数据组成, 边则表示类、实例和数据之间的相互关系. 同时用标签来表示实例, 并且用元组来表示从资源到类别需要遍历标签的路径, 然后采用欧几里得距离计算路径的长度. 基于 RDF 映射机制, 论文提出了一种改进的特征属性和实例, 便于计算语义匹配度. 此外, 该 RDF 映射机制使得领域专家能够显示定义实例, 从而保证了推荐的质量. Guo W 等人^[56]为每个物品创建语义图, 并通过倒频率策略来计算语义图中每一对物品之间的相似度, 最终将相似度高的物品推荐给用户. Fernández-Breis JT 等人^[57]基于语义网技术提出结合学习对象和学习目标的语义模型, 进而实现高效的推荐.

另一方面, 传统基于内容的推荐算法采用自动特征提取的方式, 因而很难应用于多媒体数据. 但是融合语义推理后, 则可以通过分析多媒体数据的语义特征来实现准确的推荐. 由于近年来数码相机性能的提升以及社交媒体网站 (例如 YouTube 和 Flickr) 的兴起,

自动视频编辑受到人们越来越多的关注,而为视频选择合适的背景音乐则是推广视频的关键因素之一。

Kuo F 等人^[58]提出基于音频/视频多通道潜在语义分析的背景音乐推荐算法。作者从 YouTube 网站收集配有背景音乐的视频,并使用音乐检测算法将音乐稀疏的视频过滤掉,提取能够表示视频与音频特征的词汇,然后通过多通道潜在语义分析技术学习表示视听特征单词之间的关联模型。给定一个视频,算法能够从关联模型中获取排序后的音乐推荐列表。另外,该推荐算法能够分析音乐节拍和视频镜头切换来计算被推荐的背景音乐和视频间的匹配度。Yoshida T 等人^[59]提出分别从视频标签和视听特征词中抽取语义和情感信息,并计算语义相似度和情感相似度,然后采用相似度融合函数^[60]得出两个视频之间的相似度,从而能够为用户推荐与他们曾经看过的语义相似的视频列表。由于单个特征无法充分捕捉所有用户的兴趣偏好,因此作者采用获取用户日志的方式,构建相似度融合函数,来提高推荐算法的推荐质量。

Web 2.0 技术的深入发展促生了用户生成内容(UGC, User Generated Content)^[61]标签的应用。UGC 标签系统是一种表示用户兴趣和物品语义的重要方式。当用户对一个物品打上一个标签,这个标签一方面描述了用户的兴趣,另一方面则表示了物品的语义,从而将用户和物品联系起来,从而在标签系统中对用户进行个性化的推荐。其中比较有代表性的是, Fan M 等人^[62]提出的基于内容语义的社会标签推荐算法 STR,算法提取资源描述和标签之间的关系,并可以缓解传统推荐算法的冷启动问题。当一个用户对他感兴趣的资源添加标签时,算法会推荐给该用户最有可能选择的标签列表,同时也能推荐描述中没有包含的语义标签。STR 算法遵循典型的“学习-预测”范式,主要分为以下三个阶段。

第 1 阶段:经验语义的获取。算法使用 IBM Model-1^[63]进行计算,其计算公式为:

$$P_r(T_r|W_r) = \sum_A P_r(T_r, A|W_r) \quad (1)$$

其中 $P_r(T_r|W_r)$ 表示从资源描述中提取的术语与标签之间的相似度(翻译模型);对于资源 r ,资源描述样本集 $W_r = \{w_i\}_{i=1}^{L_r}$ 和标签样本集 $T_r = \{t_x\}_{x=1}^{L_t}$ 的联系通过隐性变量 $A = \{a_z\}_{z=1}^{L_t}$,例如: $a_z = i$ 表示标签样本集中处于位置 z 的标签 t_z 匹配术语样本集中处于位置 i 的术语 w_i 。算法的目的是在 $\sum_r p(t|w) = 1$ 的约束下调整翻译概率,使 $P_r(T_r|W_r)$ 达到最大值,其中 $p(t|w)$ 是指术语 w 语义上关联标签 t 的概率。

第 2 阶段:构建基于内容的标签网。算法将翻译概率 $p(t|w)$ 表示为加权有向图: $(w, t, p(t|w))$ 。从顶点 v_i

到 v_j 的有向边表示为: $e_{ij} = \{\langle v_i, v_j \rangle, v_i \in w, v_j \in t\}$,边上的权重为 $w(e_{ij}) = p(v_i|v_j)$ 。

第 3 阶段:候选标签排序。

候选标签排序的策略是基于 TextRank^[64]算法进行改进,其计算公式如下:

$$\text{score}(v_i) = (1 - \lambda) \cdot \text{offset} + \lambda \cdot \sum_{v_j \in \text{In}(v_i)} \frac{w(e_{ji})}{\sum_{v_k \in \text{Out}(v_j)} w(e_{jk})} \cdot \text{score}(v_j) \quad (2)$$

其中 $\lambda \in [0, 1]$ 为阻尼系数,代表从图中某一个特定顶点指向其它任意顶点的概率,两顶点之间的权重为 $w(e_{ji})$ 。顶点最后的分数不受给定初值的影响,顶点的初值只影响该算法达到收敛的迭代次数。根据最后的分数,分数最高的 m 个标签被推荐给目标资源。该公式表明如果其他重要的顶点指向某个顶点,那么该顶点更重要。

为了验证 STR 推荐算法的有效性,作者实施了仿真实验。实验选取两个数据集,其中一个数据集是通过豆瓣 API 在豆瓣网站上爬取著名的中文书籍的评论和标签,而另一个数据集是通过 Last.fm 的 API 获取的音乐和艺术家的资源内容和标签。然后选用词触发方法 WTM^[65]、TextRank 和 TFIDF^[66]三个相关的社会标签推荐算法来与 STR 算法进行对比。实验结果表明,基于内容语义的社会标签算法 STR 在准确率方面性能明显优于其他三种社会标签推荐算法。

从上面分析,我们可以看出,基于语义的内容推荐算法,不仅可以缓解传统基于内容的推荐算法所存在的新物品冷启动问题,而且可以应用于多媒体领域进行高质量的推荐。目前基于语义的内容推荐算法已广泛应用于背景音乐推荐、个性化物品推荐以及社会标签推荐等领域研究中。

4 基于语义的协同过滤推荐算法

由于传统协同过滤算法是基于大量用户历史数据集的,如用户对物品的评分或用户的浏览行为等,因而存在数据稀疏的问题,同时对于刚加入系统且没有进行任何行为的新用户、以及没有被浏览或评分的新物品都存在冷启动问题。近年来有许多研究人员则尝试将语义知识融合进协同过滤算法之中来解决上述问题。

将语义融合进协同过滤过程的一种方式是在计算用户间相似度与物品间关联时,利用语义知识来对用户兴趣偏好和物品内容进行语义描述,使得更高质量地预测用户对未评分物品的打分情况,从而能够解决冷启动和数据稀疏问题。

Hu B 等人^[67]提出基于内容语义相似度的协同过滤推荐算法,通过对描述物品的文本内容进行分词、去

停用词、提取标签、词义消歧等步骤,同时利用词义间最短路径方法来计算两个词之间的相似度,从而获得两个物品之间的相似度,并根据物品之间的相似度和物品已被用户打过的评分,来预测用户对未被评分物品的打分情况(即用户对该物品的喜好程度).实验表明使用内容语义相似度能够有效地改善协同过滤的数据稀疏性问题,从而提高了协同过滤推荐算法的性能. Abdelwahab A 等人^[68]提出基于降维的协同过滤推荐算法,使用消解奇异值潜在语义索引方法来减少数据表示的维度,并生成高质量的用户评分,从而缓解协同过滤算法的数据稀疏性问题,提高可扩展性,同时减少协同过滤算法对内存空间的使用率. Lemdani R 等人^[69]提出一个域独立的语义推荐算法,该算法包含三个功能互补的模块,其中协同过滤模块利用关联规则挖掘一组推荐的候选物品;语义模块利用本体知识检测物品特征,并通过用户的行为计算用户对物品特征的喜好程度;而频率模块考虑特征出现的概率以及物品特征之间的关联,以便辅助发掘刚加入系统的新物品.算法融合这三个模块能够有效对推荐结果进行解释,从而进行高质量的推荐. Ferrara F 等人^[70]计算维基百科中概念的语义关联度,并将语义关联度集成到基于用户的协同过滤推荐过程之中来提高推荐的准确性. Cantador I 等人^[71]提出将用户和物品配置文件中本体概念的权重传播到由领域本体语义关联的其他概念当中.这种语义传播策略是基于约束扩散激活技术,并且权重的衰减随扩散过程而逐渐增大.特别,文献[71]所给出的推荐算法能够通过相关概念的描述,从而解决传统协同过滤算法的冷启动和可扩展性问题.

将语义融合进协同过滤过程的另一种方式是采用加权的方式对协同过滤和语义相似度计算得出的推荐结果各自赋予一定的权重,以此来获得最终的推荐列表.其中比较有代表性的是,Shambour Q 等人^[72]所提出的混合多准则的语义增强协同过滤算法 Hybrid MC-

SeCF,它集成了基于物品的协同过滤与语义过滤策略,能够有效解决基于物品的协同过滤算法的稀疏性和冷启动问题. Hybrid MC-SeCF 算法以用户-物品的评分矩阵以及物品本体作为输入,并且输出用户-评分预测评分矩阵,其中物品本体以树的结构展示,只有两层结点,父节点是物品的分类,叶节点包括属于该分类的物品.图1给出了 Hybrid MC-SeCF 算法的模块结构图.从图1我们可以看出,Hybrid MC-SeCF 算法主要包含如下4个紧密相关的模块:

(1) 基于物品的多准则协同过滤模块.该模块根据给定的物品集合,根据基于物品的协同过滤算法计算物品之间的相似度.

(2) 基于物品的语义计算模块.为了能够利用物品之间的语义信息,需要以树的结构创建一个物品本体,并且物品位于叶子结点,父结点是物品所属的分类.在该模块中,Hybrid MC-SeCF 算法根据共同属于同一个分类的比率来计算基于物品的语义相似度.

(3) 选择邻域模块.分别按照基于物品的协同过滤策略和基于物品的语义相似度来计算与目标物品最为相似的物品集合,作为目标物品的邻域.

(4) 混合权重预测模块.根据对上述目标物品求得的两个邻域赋予不同的权重,以此来对用户还没有评分的目标物品进行评分预测.

为了验证 Hybrid MC-SeCF 算法的有效性,作者将该算法与基于物品的协同过滤算法进行性能比较,并采用平均绝对误差的方法 MAE^[73]作为评测标准.实验结果表明 Hybrid MC-SeCF 算法的平均绝对误差较小,在推荐准确度方面有了较大的改善.

此外,作者也验证了 Hybrid MC-SeCF 算法解决数据稀疏问题的能力.实验结果证明 Hybrid MC-SeCF 算法在面对稀疏的数据集时也表现了较好的性能,其推荐结果具有较小的平均绝对误差以及较高的覆盖率,因而具有较好的解决数据稀疏问题的能力.

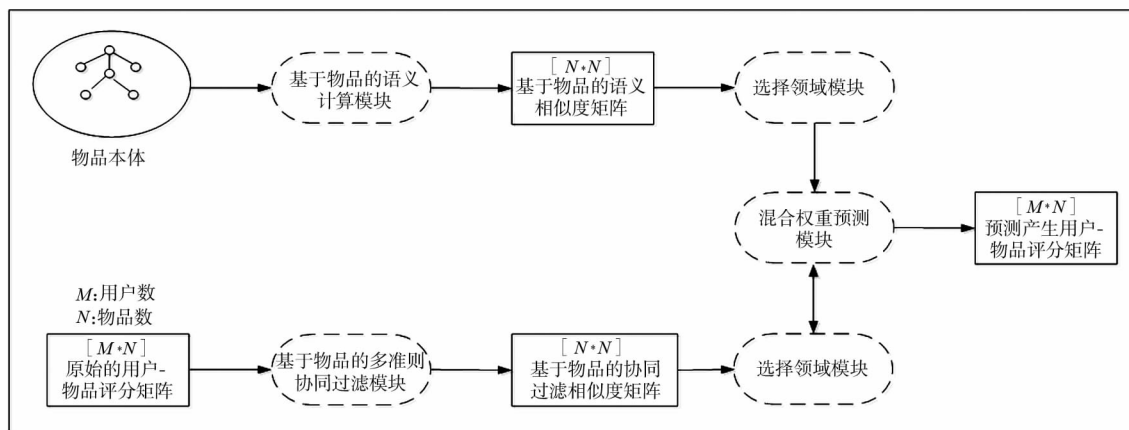


图1 Hybrid MC-SeCF算法模块结构

在社会标签系统中,许多学者也研究了如何将语义融入到协同过滤推荐的过程中. Chen H 等人^[74]提出一种新的基于标签语义相似度的协同过滤算法,该算法通过分析标签与物品之间的关联,使用 WordNet 提供的词汇结构对标签的语义相似度进行计算,以此来解决单词间相似度和语句间相似度的问题. 作者以 MovieLens 数据集为输入,通过仿真实验验证了该算法比传统协同过滤算法具有更高的效率以及更好的准确率. Kim H 等人^[75]提出基于语义标签的信息过滤方法,该方法首先通过给物品打标签来判断相似的用户,然后为每个用户挖掘语义上相关联的物品. 实验结果表明该方法提高了推荐质量,同时缓解了标签之间模糊性、同义性和互操作性的问题.

从上面分析,我们可以看出,基于语义的协同过滤推荐算法,可以有效缓解传统的协同过滤算法所面临的新用户和新物品冷启动问题,以及数据稀缺性问题,并且能够提高算法的扩展性,高效预测用户对未评分物品的评分情况,从而判断是否向目标用户推荐该物品.

5 基于语义的混合推荐算法

混合推荐算法旨在解决单个推荐技术所面临的不足,例如基于内容的推荐算法存在的特征提取问题、协同过滤算法存在的冷启动和数据稀疏性等问题. 然而混合推荐算法依然存在一定的领域局限性,也缺乏可扩展性. 考虑将语义技术融入混合推荐过程之中,可以扩展应用领域,以及提高混合推荐的性能.

Hsu I 等人^[76]提出基于 XML 链接语言 XLink^[77]的语义推荐算法. 该推荐算法基于语义网技术,由 XLink 库、知识库、搜索引擎和推理机等部件组成,可以提供三个不同的表示链接知识的方法:分别是基于 XLink 的元数据、基于本体的推理以及基于规则的推理. 基于 XLink 的语义推荐算法广泛应用于学术论文推荐、数字图书馆等领域.

Liu L 等人^[78]依据软件服务需求提出一个有效的

语义混合推荐算法. 该算法通过 Web 服务的属性描述来计算语义相似度,同时考虑上下文信息和标准信息. 其中上下文信息表示为层次树的结构,能够在用户反馈数据稀疏的情况下帮助聚集用户评分;标准信息能够反映 Web 服务的特征信息. 算法按照如下思路将基于内容的语义推理和上下文相关的协同过滤过程耦合起来:初始当系统没有收集到足够的用户反馈数据时,算法首先调用基于内容的语义推荐模块,而随着收集到的用户评分越来越多后,则基于上下文信息和标准信息使用协同过滤推荐模块. 最终返回给用户的推荐结果是将上述两种计算模块所输出的推荐结果进行加权后产生.

Deng Y 等人^[79]提出了电影领域中基于本体和神经网络的混合推荐算法. 该算法通过神经网络的学习能力来感知用户的个性化信息,然后训练神经网络模拟用户的兴趣偏好来预测用户的选择. 另外,算法使用本体的语义描述能力建立用户模型和物品模型. 作者发现,运用神经网络作为算法基础可以有效解决新物品的冷启动问题,而使用本体相似度算法则能解决新用户的冷启动问题. 同时该算法由于不需要从其他用户获取信息因此也能保护用户的隐私问题.

Tath I 等人^[80]研究提出音乐特征降维,并在推荐过程中使用多领域信息. 为了降低音乐特征的代表维度,作者通过标签来表示音乐流派信息,标签信息主要包括文体起源、典型乐器、衍生形式、子类型和融合类型等信息. 另外,作者并从 Dbpedia 知识库^[81] (Wikipedia 的一种结构化形式)中提取层次结构,来分析降维后的用户与音乐组织结构.

Chen M 等人^[82]指出目前大多数推荐算法均是基于集中式架构,不适用于 P2P (peer-to-peer) 网络,为此,提出分布式语义 P2P 推荐算法. 该算法由于考虑到用户的兴趣偏好以及兴趣的多样性,因此能够为用户提供高质量的音乐搜索和推荐服务. 具体来说,该算法采用三种过滤策略来满足分布式语义叠加的需求,首

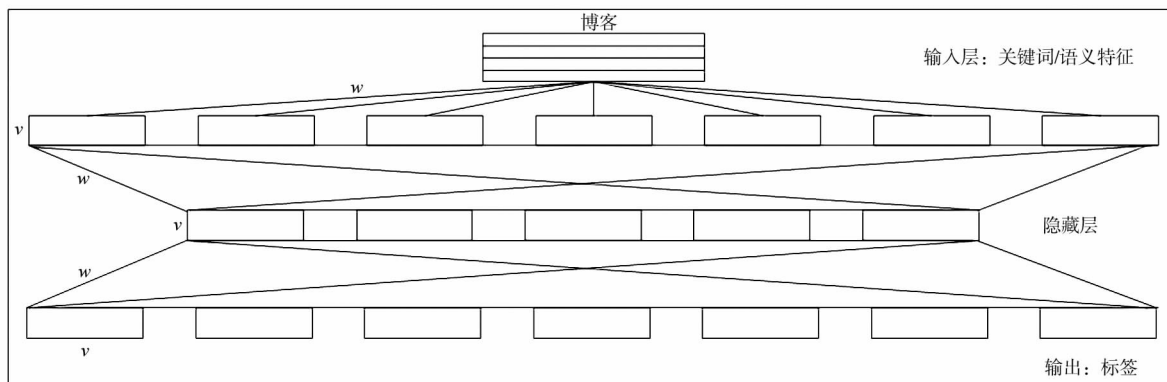


图2 神经网络学习结构图

先算法选择合适的元数据集表示一个音乐对象,并且使用元数据集构建基于特征向量的内容过滤 workflow;然后利用音乐对象特征向量中的一个属性来构建一个分布式节点的配置文件,并根据配置文件构建协同过滤 workflow;最后依据音乐之间的关联构建一个基于历史合作关系的过滤器,从而生成推荐结果。

博客是基于语义的混合推荐算法一个最为成功的应用领域.其中具有代表性的研究工作是, Lee S 等人^[83]提出面向博客的自动化标签推荐算法.该算法使用混合人工神经网络的策略,从 Web 2.0 协同标签中提取集体智慧 (collective intelligence)^[84]和单词的语义来学习如何设置最佳标签,并最终产生一个加权的标签列表,列表中的标签与给定的博客内容关联最为紧密.

从实施流程来看,算法包含两个主要阶段,即训练阶段和执行阶段.在训练过程中,算法首先使用网络爬虫获取已经人工打上标签的博客,并将这些博客分成训练集和测试集,然后使用测试集来学习博客内容和标签之间的关联.训练阶段主要分为三个步骤:关键词提取、语义处理以及人工神经网络学习.在关键词提取步骤中,算法首先使用 TFIDF 方法提取单个关键词,然后计算词与词之间共存的概率,并基于 WordNet 来判断连词.在语义处理步骤中,算法使用 WordNet 将步骤 1 中提取的关键词组成层次结构,然后选择顶层的同义词表示博客内容的语义信息.而在人工神经网络学习步骤中,算法构造图 2 所示的学习结构图,包括输入层、隐藏层和输出层,并学习关键词和标签语义特征之间的关联.由于不同特征会影响标签的选择,学习得到的权重决定了每个特征对选择标签的影响程度,因此需要反向传播进行学习.完成训练阶段后,系统进入执行阶段.在这个阶段中,用户提交博客内容,由算法自动提取关键词,进行语义分析,并调用人工神经网络产生的标签向用户进行推荐.

在仿真实验中,作者从英文博客中提取博客内容和标签,将其分为训练集和测试集,使用关键词提取方法从数据集中提取关键词并计算其频率,作为人工神经网络的输入数据集.将自动化标签推荐算法与 Yahoo! 中的协同过滤标签推荐算法^[84]进行比较,作者发现算法具有更丰富的语义信息,以及更好的推荐效果.

从上面分析,我们可以看出,基于语义的混合算法具有更加广泛的应用领域,例如学术资源、web 服务、电影/音乐等多媒体、P2P 网络以及社会标签系统等领域.同时,基于语义的混合算法具有更好的扩展性,可以灵活结合人工神经网络、层次结构树等技术进行推荐.

6 基于语义的社会化推荐算法

Facebook 和 Twitter 作为两种不同类型社会网络的

代表,是目前互联网界的新宠.国内以新浪微博为代表,也成为社会网络应用的热门平台.基于社会网络的推荐可以很好地模拟现实,好友的推荐可以增加推荐的信任度,并且通过获取用户的好友列表可以解决新用户的冷启动问题.但是社会化推荐也存在一些缺点,由于用户的好友关系不是基于共同兴趣产生的,因此用户好友的兴趣往往和用户的兴趣并不一致,导致推荐算法的准确率和召回率降低,而且社会化推荐存在数据稀疏问题,用户浏览和对物品的评分等行为的数据集较少.因此,近年来不少研究将语义技术融进社会化推荐中,旨在提高社会化推荐结果的质量.

Golbeck J 等人^[85]提出电影推荐算法 FilmTrust,该算法融合了基于语义网的社会网络,利用社会网络中的信任评分机制来计算用户间相似度. Cantado I 等人^[86]对用户共同兴趣偏好的不同表示方法建立一个多层语义社会网络模型,从大量基于本体的用户配置文件中提取用户共同的兴趣偏好,然后自动在域概念空间中聚簇并构建多层社会网络,进而在多个语义偏好层种计算用户之间的相似度,并根据用户之间的相似度使用协同过滤算法向用户进行推荐. Ghita S 等人^[87]探讨了在分布式社会网络中交换和共享基于 RDF 图的推荐结构,并使用扩展的 FOAF 词汇^[88]来指定兴趣分组,以及组内的成员共享推荐结果的方式. Martin-Vicente M 等人^[89]提出了个性化分发优惠券的推荐算法,并利用社会网络来为系统添加信息并获取更多的新用户.

用户隐私保护对于社交网站是一个至关重要的问题^[90]. Li Q 等人^[91]提出了一个基于语义的智能隐私配置算法 SPAC,可以为社会网络用户自动推荐隐私设置. SPAC 算法学习用户的隐私配置模式,利用机器学习技术依据用户配置文件和用户隐私设置的历史信息,来预测用户对未标记或新出现的朋友进行隐私保护设置的策略.针对不同用户的内容,该算法采用语义知识进行推断,提出语义增强的 k 近邻分类方法,从而在很大程度上改善了隐私推荐的准确性.在 SPAC 算法中,作者通过为用户建立用户配置文件,来记录用户的隐私信息,同时为用户在社会网络中所拥有的不同资源建立数据配置文件,记录用户的个人信息、用户在社会网络中上传的照片、撰写的博客日志、上传的视频、音频等资源. SPAC 算法的体系结构如图 3 所示,其输入主要由 4 部分构成:(1) 用户允许部分好友访问的资源;(2) 用户好友的特征;(3) 用户隐私配置的历史记录,作为算法的训练集以及;(4) 好友特征的本体知识;而 SPAC 算法输出未标记或新出现好友的隐私保护设置.另外,我们不难看出,SPAC 算法本质上其实是一个分类系统,其目的是找到一个分类函数 $M: \mathbf{F} \rightarrow \{0, 1\}$.

1},其中 F 是用户好友的特征向量,0 和 1 表示用户决定允许或拒绝该好友访问用户所对应的个人信息和资源. 由于传统的分类方法如决策树、基于规则的分类等难以处理文字无关但语义相关的向量,因此作者提出了一种语义增强的 k 近邻分类算法,利用域本体来计算用户间的相似度,计算公式如公式(3)所示. 其中 simF 函数用来测量两个特征的值之间的相似度. 而用户的特征均被映射到本体图的结点上,如图 4 所示,其中相同的概念以同一个结点表示,且按照边的数量来计算两个概念之间的最短路径,同时考虑结点在图中所处的深度.

$$\text{SIM}(x, y) = \frac{\sum_{i=0}^n w_i \cdot \text{simF}(\text{feature}_i \text{ of } x, \text{feature}_i \text{ of } y)}{\sum_{i=0}^n w_i} \quad (3)$$

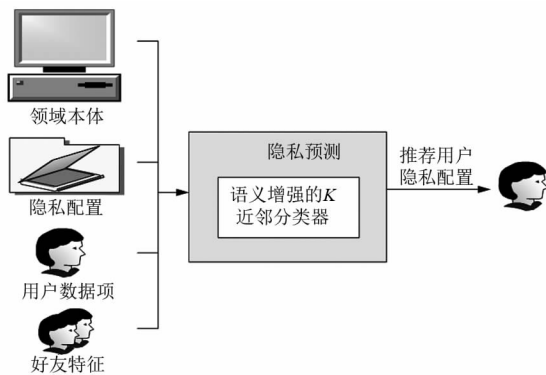


图3 SPAC算法体系结构图

此外,论文通过仿真实验验证了 SPAC 算法的准确度以及 k 值对算法性能的影响. 作者采用常用的 n 次交叉验证方法^[92]进行实验,将语义增强的 k 近邻分类算法与其他三种分类算法进行比较:(1)基础算法,即用

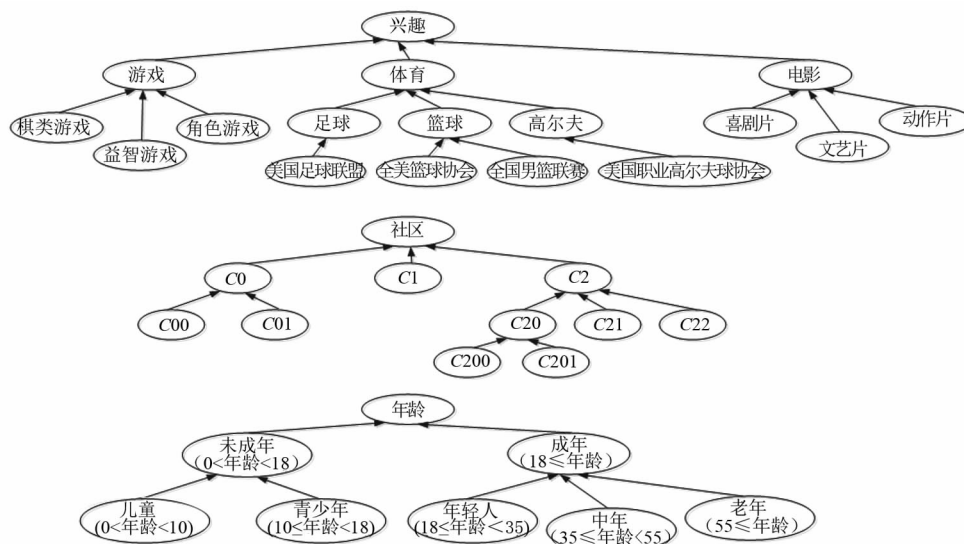


图4 算法所使用的本体图

户对他的一些朋友打标签,剩下的朋友将使用默认值;(2)决策树算法;(3)典型的 k 近邻分类算法.

实验结果表明,基于相同的已经打标签的伙伴数,语义增强的 k 近邻分类算法比其他三种算法具有更高的准确度;当已经打标签的伙伴数较少时,语义增强的 k 近邻分类算法的学习能力比其他三种算法效果更好.

从上面的分析中我们可以看出,基于语义的社会化推荐算法可以有效解决数据稀疏性问题,并且提高推荐算法的精度(准确率和召回率),避免社会网络中用户好友的兴趣和用户兴趣不一致的问题.

7 讨论与挑战

与传统的推荐算法相比,语义推荐算法可以有效的解决数据稀疏问题,以及新用户和新物品的冷启动

问题,从而能够改善推荐方案的可扩展性,并提高推荐结果的精度和多样性. 语义推荐算法已经被广泛应用于多媒体、学术科研、社会标签系统、社会网络应用、Web 服务等各个领域. 作为一个具有较好发展前景的研究领域,能够深入研究并且取得更进一步突破的方向主要包括:

(1) 语义相似度的计算方法

语义相似度是一个主观性很强的概念,没有明确的客观标准可以衡量. 脱离具体的应用讨论语义相似度,很难得到一个统一的定义. 语义相似度计算的两种基本方法是基于本体知识的方法,以及基于统计的上下文向量空间模型方法.

基于本体知识计算语义相似度的方法,利用语义词典如 WordNet、HowNet 中的同义词或义原组成的树

状层次体系结构,通过计算两个概念间的信息熵或语义距离,来获取概念间语义相似度.值得一提的是,利用本体知识来计算语义相似度,是基于两个词之间具有一定的语义相关性,(即当它们在概念间的结构层次网络图中存在一条通路),而且该方法容易忽略不同本体概念间的相似度,对于不同本体难以有效划分权重来确定对整体概念间相似度的贡献^[93].基于统计的上下文向量空间模型方法将上下文信息的概率分布作为词汇语义相似度计算的参数,利用相关熵或采用平均互信息量来计算语义相似度.基于统计的定量分析方法能够对词汇间的语义相似度进行比较精确的度量,但是这种方法较为依赖于训练所用的语料库,计算量大,计算方法复杂,并且受数据稀疏和数据噪声的干扰较大.提高语义相似度能够直接提高语义推荐算法的推荐结果质量,因此如何改善语义相似度的计算方法正是值得关注的研究方向.

(2) 个人推荐 vs 群体推荐

目前推荐算法研究领域几乎都是考虑向单个用户推荐物品、信息以及好友等,但是很少有向群体用户提供推荐服务的推荐算法.个人推荐算法在支持个人决策行为时非常有效,但是却并不能为群体决策提供推荐策略,这主要是因为它不包括群体用户评分,因此不能有效地为群体产生推荐意见.

随着在线社区活动的增加,人们对群体推荐算法的需求变得越来越强烈.群体决策存在社会效应,群体成员的决策过程及结果或多或少地会受到群体中其他成员的行为、想法等的影响,同时需要考虑不同用户身份在群体中所占的比重.例如领袖对群体决策过程通常产生更具决定性的影响,因此群体决策比个人决策的过程更为复杂.我们考虑群体决策的过程,如何全面考虑群体中不同用户不同兴趣偏好的冲突从而对群体提供整体的推荐,并能够针对不同的群体设计相应的群体用户推荐算法,也是语义推荐算法将来的发展方向之一.

(3) 基于地域信息的推荐

不同地方的用户兴趣存在着很大的差别,比如在饮食文化上四川人口味偏辣,而南方人口味偏甜等,因此在向不同地域的用户推荐饮食时则要考虑用户的地域信息才能给用户推荐符合其兴趣偏好的物品.另外,一个用户往往在一定范围内活动,通过 Foursquare 数据分析发现,45% 的用户其活动范围半径不超过 10 英里,而 75% 的用户活动半径不超过 50 英里.因此,在语义推荐算法中需要考虑推荐地点和用户当前所处地点的距离,不能推荐太远的地方.这就需要结合用户定位技术,将用户的地域信息融入到推荐算法中.

(4) 推荐算法的人机交互

目前大多数语义推荐算法都是通过示例方法建立

用户模型,即需要新用户提供初始化信息建立初始的用户模型,随着用户的浏览记录和行为的增多,学习并修正用户模型从而提供更丰富的推荐结果.但是示例用户兴趣建模会间接打击用户浏览的积极性,并且初始化生成的用户模型不一定准确.因此如何让用户参与推荐结果的反馈中去,使得算法能够根据用户反馈进行自适应改进,需要设计合理的人机交互策略,吸引用户参与到推荐结果的反馈中,从而更准确地获取用户的兴趣偏好,提高推荐的性能.

(5) 推荐算法去噪并进行实时性推荐

语义推荐算法在采集用户信息以及浏览记录时,不仅存在用户数据稀疏性问题,还可能会存在大量噪音数据和无意义数据,例如用户在系统中浏览物品时,由于某些原因离开了,但是系统一直在记录用户的浏览时间,并认为用户对该物品有很大的兴趣,从而向用户推荐该物品,这样的推荐结果明显不是非常的精确.算法中可以采用设定阈值^[93]或加入参数来控制采集信息的方法进行去除噪声数据,这些方法对于推荐结果的准确度具有很大的作用.

同时,由于用户的暂时兴趣偏好是一个随时间推移而动态变化的参数^[94],因此如何根据用户的兴趣变化来及时地修正用户的配置文件,从而实时地向用户进行推荐,也是一个影响推荐结果精确度的因素.

(6) 大数据环境下语义推荐算法

近些来,随着物联网、云计算和社交网络等技术的迅猛发展,大数据越发显现 4V (Volume, Velocity, Variety, Veracity) 特性.根据国际数据公司 IDC 的测算,2011 年网络空间将产生 1800EB 的数据,2012 年会增长 40%,达到 2500EB,截至 2020 年,将会达到 35000EB,超出了目前磁盘空间的存储能力.因而,在大数据环境下,语义推荐算法在自适应性、实时性和鲁棒性方面都面临着巨大的挑战.

底层数据的有效建模是解决所面临上述挑战的理论基础^[94].近来有研究人员提出采用物理学中张量的概念来表示大数据^[95].由多阶张量表示大数据结构不仅可以对数据特征向量进行降维处理,同时也易于提取数据的语义信息,并且能够避免学习算法中的“过拟合问题”.根据大数据的不同模态特征,将大数据表示为多阶张量,在训练过程中能够有效地利用未标记样本分布信息,更好地刻画整个样本空间上的数据特性.然而,现有工作所建立的大数据张量模型均是静态的,没有考虑如何建立大数据去噪声和动态性的张量模型,这方面将是将来语义推荐算法的一个重要研究方向.

另外,在物理实施大数据环境下语义推荐算法时,目前主流的策略是采用分布式并行框架(如 Hadoop 平

台)^[96]. 然而我们发现当数据规模和模态个数逐渐增大时, Hadoop 平台的推荐实时性是存在严重的缺陷. 近年来, 学者尝试在 Hadoop YARN 上构建 Spark 分布式内存计算框架^[97]. Spark 框架拥有 Hadoop 平台所具有的优点, 而且可以将平台的中间输出和结果可以保存在分布式内存阵列中, 从而不再需要读写 HDFS, 从而能较好地适用于大数据环境下语义推荐算法. 因此, 我们预计, 将 Spark 框架融合进语义推荐算法是将来的一个重要研究方向.

8 结语

随着语义网及其相关理论的出现使得推荐算法的发展迎来了新的机遇, 从而产生了语义推荐算法. 语义推荐算法的研究蕴含着巨大的应用价值和理论研究意义, 近年来一直吸引着众多的研究者展开相关工作. 本文在分析传统三类推荐算法所存在问题的基础上, 介绍和分析了语义推荐算法的研究现状和进展, 主要包括基于语义的内容推荐算法、基于语义的协同过滤算法、基于语义的混合推荐算法以及基于语义的社会推荐算法, 并讨论了今后的发展方向, 希望能对相关领域的研究人员和工程技术人员提供有益的帮助.

参考文献

- [1] Zhou Z, Liu M, Zhang F, et al. A data processing framework for IoT based online monitoring system [A]. Proc of IEEE 17th International Conference on Computer Supported Cooperative Work in Design [C]. Whistler; IEEE, 2013. 686 - 691.
- [2] Chen W, Niu Z, Zhao X, et al. A hybrid recommendation algorithm adapted in e-learning environments [J]. World Wide Web, 2014, 17(2): 271 - 284.
- [3] Crespo RG, Martínez OS, Lovelle JMC, et al. Recommendation system based on user interaction data applied to intelligent electronic books [J]. Computers in Human Behavior, 2011, 27(4): 1445 - 1449.
- [4] Li L, Zheng L, Yang F, et al. Modeling and broadening temporal user interest in personalized news recommendation [J]. Expert Systems with Applications, 2014, 41(7): 3168 - 3177.
- [5] Resnick P, Varian H R. Recommender systems [J]. Communications of the ACM, 1997, 40(3): 56 - 58.
- [6] Rashid AM, Ling K, Tassone RD, et al. Motivating participation by displaying the value of contribution [A]. Proc of the SIGCHI Conference on Human Factors in Computing Systems [C]. Montreal; ACM, 2006. 955 - 958.
- [7] Qiu T, Han TY, Zhong LX, et al. Redundant correlation effect on personalized recommendation [J]. Computer Physics Communications, 2014, 185(2): 489 - 494.
- [8] 周军锋, 汤显, 郭景峰. 一种优化的协同过滤推荐算法 [J]. 计算机研究与发展, 2004, 41(10): 1842 - 1847.
Zhou Junfeng, Tang Xian, Guo Jingfeng. An optimized collaborative filtering recommendation algorithm [J]. Journal of Computer Research and Development, 2004, 41(10): 1842 - 1847. (in Chinese)
- [9] Wu ML, Chang CH, Liu RZ. Integrating content-based filtering with collaborative filtering using co-clustering with augmented matrices [J]. Expert Systems with Applications, 2014, 41(6): 2754 - 2761.
- [10] Loizou A, Dasmahapatra S. Recommender systems for the semantic web [A]. Proc of Recommender Systems Workshop [C]. Trento; ECAI, 2006. 1 - 5.
- [11] Lika B, Kolomvatsos K, Hadjiefthymiades S. Facing the cold start problem in recommender systems [J]. Expert Systems with Applications, 2014, 41(4): 2065 - 2073.
- [12] Sarigecili MI, Roy U, Rachuri S. Interpreting the semantics of GD&T specifications of a product for tolerance analysis [J]. Computer-Aided Design, 2014, 47: 72 - 84.
- [13] Janowicz K, Raubal M, Kuhn W. The semantics of similarity in geographic information retrieval [J]. Journal of Spatial Information Science, 2014(2): 29 - 57.
- [14] Sabucedo L, Gago JMS, Iglesias MJF. A tensor-based approach for semantic recommenders in e-government [J]. Advanced Science Letters, 2014, 20(1): 317 - 320.
- [15] Chen LC, Kuo PJ, Liao IE. Ontology-based library recommender system using MapReduce [J]. Cluster Computing, 2015, 18(1): 113 - 121.
- [16] Adomavicius G, Jannach D. Preface to the special issue on context-aware recommender systems [J]. User Modeling and User-Adapted Interaction, 2014, 24(1): 1 - 5.
- [17] Hu B, Zhou Y. Content semantic similarity boosted collaborative Filtering [A]. Proc of International Conference on Computational Intelligence and Security [C]. Suzhou: IEEE, 2008. 7 - 11.
- [18] Middleton SE, De Roure DC, Shadbolt NR. Capturing knowledge of user preferences: ontologies in recommender systems [A]. Proc of the 1st International Conference on Knowledge Capture [C]. New York: ACM, 2001. 100 - 107.
- [19] 张波, 向阳, 黄震华. 一种基于语义的决策服务协作自组织方法 [J]. 控制与决策, 2013, 28(9): 1414 - 1418.
Zhang Bo, Xiang Yang, Huang Zhenhua. Semantic based self-organization method for decision-making service collaboration [J]. Control and Decision, 2013, 28(9): 1414 - 1418. (in Chinese)
- [20] Kim H, Kim HJ. A framework for tag-aware recommender systems [J]. Expert Systems with Applications, 2014, 41(8): 4000 - 4009.

- [21] Bouza A, Bernstein A. User preference similarity as classification-based model similarity [J]. *Semantic Web*, 2014, 5(1):47–64.
- [22] Liu L, Lecue F, Mehandjiev N. Semantic content-based recommendation of software services using context [J]. *ACM Transactions on the Web*, 2013, 7(3):17.
- [23] Janowicz K, Raubal M, Kuhn W. The semantics of similarity in geographic information retrieval [J]. *Journal of Spatial Information Science*, 2011, 2(2):29–57.
- [24] Mooney RJ, Roy L. Content-based book recommending using learning for text categorization [A]. *Proc of the Fifth ACM Conference on Digital Libraries* [C]. San Antonio: ACM, 2000. 195–204.
- [25] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems [J]. *Computer*, 2009, 42(8):30–37.
- [26] Martínez L, Pérez L G, Barranco M. A multigranular linguistic content-based recommendation model [J]. *International Journal of Intelligent Systems*, 2007, 22(5):419–434.
- [27] Sun F, Shi Y, Wang W. Content-based recommendation system based on vague sets [A]. *Proc of International Conference on Intelligent Human-Machine Systems and Cybernetics* [C]. Nanjing: IEEE, 2013. 294–297.
- [28] Khan M, Nair S. Survey of content based recommendation systems in a nutshell [J]. *International Journal of Advanced Research in Computer Science and Electronics Engineering*, 2014, 3(1):24–30.
- [29] Balabanović M, Shoham Y. Fab: content-based, collaborative recommendation [J]. *Communications of the ACM*, 1997, 40(3):66–72.
- [30] Sandvig JJ, Mobasher B, Burke R. Robustness of collaborative recommendation based on association rule mining [A]. *Proc of ACM Conference on Recommender Systems*. [C] Silicon Valley: ACM, 2007. 105–112.
- [31] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(6):734–749.
- [32] Moscato V, Picariello A. Towards a user based recommendation strategy for digital ecosystems [J]. *Knowledge-Based Systems*, 2013, 37(2):165–175.
- [33] Huang Y, Gao X, Gu S. UARR: A novel similarity measure for collaborative filtering recommendation [J]. *Cybernetics and Information Technologies*, 2013, 13(Special Issue):122–130.
- [34] Choi K, Yoo D, Kim G, et al. A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis [J]. *Electronic Commerce Research and Applications*, 2012, 11(4):309–317.
- [35] Zheng VW, Zheng Y, Xie X, et al. Towards mobile intelligence: Learning from GPS history data for collaborative recommendation [J]. *Artificial Intelligence*, 2012, 184–185(2):17–37.
- [36] Liu Q, Chen E, Xiong H, et al. Enhancing collaborative filtering by user interest expansion via personalized ranking [J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2012, 42(1):218–233.
- [37] Zheng Z, Ma H, Lyu MR, et al. QoS-aware web service recommendation by collaborative filtering [J]. *IEEE Transactions on Services Computing*, 2011, 4(2):140–152.
- [38] Zhang J, Peng Q, Sun S, et al. Collaborative filtering recommendation algorithm based on user preference derived from item domain features [J]. *Physica A Statistical Mechanics & Its Applications* 2014, 396(2):66–76.
- [39] Li Y, Zhai CX, Chen Y. Exploiting rich user information for one-class collaborative filtering [J]. *Knowledge and Information Systems*, 2014, 38(2):277–301.
- [40] Liang T, Liang Y, Fan J, et al. A hybrid recommendation model based on estimation of distribution algorithms [J]. *Journal of Computational Information Systems*, 2014, 10(2):781–788.
- [41] Dai N, Qi X, Davison B D. Enhancing web search with entity intent [A]. *Proc of the 20th International Conference Companion on World Wide Web* [C]. Hyderabad: ACM, 2011. 29–30.
- [42] Burgess J, Green J. *YouTube: Online Video and Participatory Culture* [M]. New York: John Wiley & Sons, 2013.
- [43] Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering [J]. *IEEE Internet Computing*, 2003, 7(1):76–80.
- [44] Konstan JA, Miller BN, Maltz D, et al. GroupLens: applying collaborative filtering to usenet news [J]. *Communications of the ACM*, 1997, 40(3):77–87.
- [45] Miller BN, Albert I, Lam SK, et al. MovieLens unplugged: experiences with an occasionally connected recommender system [A]. *Proc of the 8th International Conference on Intelligent User Interfaces* [C]. Miami: ACM, 2003. 263–266.
- [46] Bomhardt C. Newsrec, a svm-driven personal recommendation system for news websites [A]. *Proc of IEEE / WIC/ACM International Conference on Web Intelligence* [C]. Beijing: IEEE, 2004. 545–548.
- [47] Sobocki J, Babiak E, et al. Application of hybrid recommendation in web-based cooking assistant [A]. *Proc of Knowledge-Based Intelligent Information and Engineering*

- Systems [C]. Bournemouth; Springer, 2006. 797 – 804.
- [48] Xu H, Zhang R, Lin C, et al. Construction of E-commerce recommendation system based on semantic annotation of ontology and user preference [J]. TELKOMNIKA Indonesian Journal of Electrical Engineering, 2014, 12(3): 2028 – 2035.
- [49] Victor C, Luigi C. Taking advantage of semantics in recommendation systems [A]. Proc of International Conference of the Catalan Association for Artificial Intelligence [C]. Tarragona; IOS Press, 2010. 163 – 172.
- [50] Peng T, Wang W, Gong XY, et al. A graph indexing approach for content-Based recommendation system [A]. Proc of International Conference on Multimedia and Information Technology [C]. Hong Kong; IEEE, 2010. 93 – 97.
- [51] Canhasi E, Kononenko I. Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization [J]. Expert Systems with Applications, 2014, 41(2): 535 – 543.
- [52] Lops P, de Gemmis M, Semeraro G, et al. Content-based filtering with tags: The first system [A]. Proc of International Conference on Intelligent Systems Design and Applications [C]. Pisa; IEEE, 2009. 255 – 260.
- [53] Preotiuc-Pietro D, Hristea F. Unsupervised word sense disambiguation with N-gram features [J]. Artificial Intelligence Review, 2014, 41(2): 241 – 260.
- [54] Harispe S, Ranwez S, Janaqi S, et al. Semantic measures based on RDF projections: application to content-based recommendation systems [A]. International Conference on the Move to Meaningful Internet Systems [C]. Graz; Springer, 2013. 606 – 615.
- [55] 贾哲, 黄志球, 王珊珊, 等. 支持本体推理的 P3P 隐私策略冲突检测研究 [J]. 计算机科学与探索, 2013, 7(1): 74 – 82.
Jia Zhe, Huang Zhiqiu, Wang Shanshan, et al. Detecting P3P privacy conflicts based on ontology [J]. Journal of Frontiers of Computer Science and Technology, 2013, 7(1): 74 – 82. (in Chinese)
- [56] Guo W, Kraines SB. Semantic Content-Based Recommendations Using Semantic Graphs [M]. Berlin; Springer, 2010. 653 – 659.
- [57] Fernández-Breis JT, Frutos-Morales F, Gil AE, et al. Recommendation of Personalized Learning Contents Supported by Semantic Web Technologies [M]. Berlin; Springer, 2013. 540 – 545.
- [58] Kuo FF, Shan MK, Lee SY. Background music recommendation for video based on multimodal latent semantic analysis [A]. Proc of IEEE International Conference on Multimedia and Expo [C]. California; IEEE, 2013. 1 – 6.
- [59] Yoshida T, Irie G, Arai H, et al. Towards semantic and affective content-based video recommendation [A]. Proc of IEEE International Conference on Multimedia and Expo Workshops [C]. California; IEEE, 2013. 1 – 6.
- [60] 刘敏华, 萧德云. 基于相似度的多传感器数据融合 [J]. 控制与决策, 2004, 19(5): 534 – 537.
Liu Minhua, Xiao Deyun. Multi-sensor data fusion based on similitude degree [J]. Control and Decision, 2004, 19(5): 534 – 537. (in Chinese)
- [61] Van Dijck J. Users like you? Theorizing agency in user-generated content [J]. Media, Culture, and Society, 2009, 31(1): 41.
- [62] Fan M, Zhou Q, Zheng TF. Content-based semantic tag ranking for recommendation [A]. Proc of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology [C]. Atlanta; IEEE, 2012. 292 – 296.
- [63] Liang Y, Yan Z, Sun B, et al. Cardiovascular risk factor profiles for peripheral artery disease and carotid atherosclerosis among Chinese older people: a population-based study [J]. PloS One, 2014, 9(1): 136 – 148.
- [64] Park J, Kim J, Lee J H. Keyword extraction for blogs based on content richness [J]. Journal of Information Science, 2014, 40(1): 38 – 49.
- [65] Liu Z, Chen X, Sun M. A simple word trigger method for social tag suggestion [A]. Proc of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics [C]. Bournemouth; Springer, 2011. 1577 – 1588.
- [66] Sun H. The application of fuzzy set theory in modern information retrieval [J]. Advances in Services Science and Services Information Technology (Set), 2014, 27(1): 177 – 186.
- [67] Hu B, Zhou Y. Content semantic similarity boosted collaborative filtering [A]. Proc of International Conference on Computational Intelligence and Security [C]. Guangzhou; IEEE, 2008. 7 – 11.
- [68] Abdelwahab A, Sekiya H, Matsuba I, et al. An efficient collaborative filtering algorithm using SVD-free latent semantic indexing and particle swarm optimization [A]. Proc of International Conference on Natural Language Processing and Knowledge Engineering [C]. Dalian; IEEE, 2009. 1 – 4.
- [69] Lemdani R, Bennacer N, Polailon G, et al. A collaborative and semantic-based approach for recommender systems [A]. Proc of International Conference on Intelligent Systems Design and Applications [C]. Cairo; IEEE, 2010. 469 – 476.
- [70] Ferrara F, Tasso C. Integrating semantic relatedness in a

- collaborative filtering system [A]. Proc of Mensch & Computer Workshopband [C]. Shanghai: IEEE, 2012. 75 – 82.
- [71] Cantador I, Castells P, Bellogín A. An enhanced semantic layer for hybrid recommender systems [J]. International Journal on Semantic Web & Information Systems, 2011, 7 (1): 44 – 78.
- [72] Shambour Q, Lu J. A hybrid multi-criteria semantic-enhanced collaborative filtering approach for personalized recommendations [A]. Proc of IEEE/WIC/ ACM International Conference on Web Intelligence and Intelligent Agent Technology [C]. Warsaw: IEEE, 2011. 71 – 78.
- [73] Herlocker J, Konstan J A, Riedl J. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms [J]. Information Retrieval, 2002, 5 (4): 287 – 310.
- [74] Hang C, Meifang Z. Improve tagging recommender system based on tags semantic similarity [A]. Proc of International Conference on Communication Software and Networks [C]. Xi'an: IEEE, 2011. 94 – 98.
- [75] Kim HN, Rocznik A, Lévy P, et al. Social media filtering based on collaborative tagging in semantic space [J]. Multimedia Tools and Applications, 2012, 56 (1): 63 – 89.
- [76] Hsu I. SXRS: an XLink-based recommender system using semantic web technologies [J]. Expert Systems with Applications, 2009, 36 (2): 3795 – 3804.
- [77] Hsu I, Ting D H, Hsueh N L. MDA-based visual modeling approach for resources link relationships using UML profile [J]. Computer Standards & Interfaces, 2014, 36 (3): 648 – 656.
- [78] Liu L, Lecue F, Mehandjiev N. A hybrid approach to recommending semantic software services [A]. Proc of IEEE International Conference on Web Services [C]. Washington: IEEE, 2011. 379 – 386.
- [79] Deng Y, Wu Z, Tang C, et al. A hybrid movie recommender based on ontology and neural networks [A]. Proc of IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing [C]. Hangzhou: IEEE, 2010. 846 – 851.
- [80] Tatli I, Birturk A. A tag-based hybrid music recommendation system using semantic relations and multi-domain information [A]. Proc of IEEE 11th International Conference on Data Mining Workshops [C]. Vancouver: IEEE, 2011. 548 – 554.
- [81] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia-A crystallization point for the Web of data [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7 (3): 154 – 165.
- [82] Chen MH, Lin KCJ, Kung CC, et al. On the design of the semantic p2p system for music recommendation [A]. Proc of International Symposium on Parallel and Distributed Processing with Applications [C]. Taipei: IEEE, 2010. 442 – 448.
- [83] Lee S, Chun A. Automatic tag recommendation for the web 2.0 blogosphere using collaborative tagging and hybrid ANN semantic structures [A]. Proc of International Conference on Wseas International Conference on Applied Computer Science [C]. Wisconsin: WSEAS, 2007. 88 – 93.
- [84] Woolley A W, Chabris C F, Pentland A, et al. Evidence for a collective intelligence factor in the performance of human groups [J]. Science, 2010, 330 (6004): 686 – 688.
- [85] Golbeck J, Hendler J. Filmtrust: Movie recommendations using trust in web-based social networks [A]. Proceedings of the IEEE Consumer Communications and Networking Conference [C]. Las Vegas, NV: IEEE, 2006. 282 – 286.
- [86] Cantador I, Castells P. Multilayered Semantic Social Network Modeling by Ontology-Based User Profiles Clustering: Application to Collaborative Filtering [M]. Berlin: Springer, 2006. 334 – 349.
- [87] Ghita S, Nejd W, Paiu R. Semantically Rich Recommendations in Social Networks for Sharing, Exchanging and Ranking Semantic Context [M]. Berlin: Springer, 2005. 293 – 307.
- [88] Graves M, Constabaris A, Brickley D. Foaf: Connecting people on the semantic web [J]. Cataloging & Classification Quarterly, 2007, 43 (3 – 4): 191 – 202.
- [89] Martin-Vicente M I, Gil-Solla A, Ramos-Cabrer M, et al. Semantics-driven recommendation of coupons through digital TV: Exploiting synergies with social networks [A]. Proc of IEEE International Conference on Consumer Electronics [C]. Las Vegas: IEEE, 2012. 564 – 565.
- [90] Squicciarini A C, Paci F, Sundareswaran S. PriMa: a comprehensive approach to privacy protection in social network sites [J]. Annals of Telecommunications-Annales Des Télécommunications, 2014, 69 (1 – 2): 21 – 36.
- [91] Li Q, Li J, Wang H, et al. Semantics-enhanced privacy recommendation for social networking sites [A]. Proc of IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications [C]. Liverpool: IEEE, 2011. 226 – 233.
- [92] Schaffer C. Selecting a classification method by cross-validation [J]. Machine Learning, 1993, 13 (1): 135 – 143.
- [93] Zhang J, Peng Q, Sun S, et al. Collaborative filtering recommendation algorithm based on user preference derived from item domain features [J]. Physica A: Statistical Mechanics and its Applications, 2014: 66 – 76.
- [94] Li L, Zheng L, Yang F, et al. Modeling and broadening

temporal user interest in personalized news recommendation[J]. Expert Systems with Applications, 2014, 41(7): 3168 – 3177.

- [95] O'Leary D E. Artificial intelligence and big data[J]. IEEE Intelligent Systems, 2013, 28(2): 0096 – 99.
- [96] Leverich J, Kozyrakis C. On the energy (in) efficiency of hadoop clusters [J]. ACM SIGOPS Operating Systems Review, 2010, 44(1): 61 – 65.
- [97] Liu Y, Emoto K, Matsuzaki K, et al. Accumulative Computation on MapReduce [J]. IPSJ Online Transactions, 2014, 7(0): 33 – 42.

作者简介



黄震华 男, 1980 年 9 月出生, 福建仙游人. 博士、副教授. 研究方向为信息服务、数据挖掘和大数据分析等.

E-mail: huangzhenhua@tongji.edu.cn



张佳雯 女, 1991 年 11 月出生, 上海金山人. 现为硕士研究生在读, 研究方向为推荐算法、数据挖掘等.

E-mail: smilezjw@qq.com



张波 男, 1978 年 11 月出生, 江苏常州人. 博士、副教授. 研究方向为信息论、语义计算和模式识别等.



喻剑 男, 1974 年 6 月出生, 浙江义乌人. 研究方向为数据挖掘、信息服务等.



向阳 男, 1962 年 12 月出生, 江苏徐州人. 教授、博士生导师. 研究方向为数据挖掘、语义计算、大数据分析等.



黄德双 男, 1964 年 1 月出生, 安徽合肥人. 教授、博士生导师. 研究方向为模式识别、数据挖掘、智能计算等.