·开发研究与工程应用· 文章编号：1000-3428(2014)08-0277-05 文献标识码：A 中图分类号：TP311

一种基于网站聚合和语义知识的电影推荐方法

周文乐，朱 明，陈天昊

（中国科学技术大学自动化系，合肥 230027）

摘 要： 针对传统个性化推荐方法中存在的稀疏性、冷启动、过度专业化且准确率低等问题，提出一种基于网站聚合和知识的电影推荐方法。利用网络爬虫聚合源网站对某部电影的相关推荐，得到待推荐电影集，使用电影知识构建基于本体论的电影模型，并在该模型的基础上给出一种学习用户偏好权重的算法，采用SimRank算法和加权平均值计算电影相似度，根据相似度高低向用户进行推荐。实验结果证明，该方法的推荐准确度在非实时推荐场景下较现有方法提高10%以上，且实时推荐的推荐质量有明显提高，在一定程度上解决了稀疏性、冷启动及过度专业化等问题。

关键词： 个性化推荐；网络爬虫；网站聚合；本体论；用户偏好；冷启动

A Film Recommendation Method Based on Website Aggregation and Semantic Knowledge

ZHOU Wen-le，ZHU Ming，CHEN Tian-hao

（Department of Automation，University of Science and Technology of China，Hefei 230027，China）

【Abstract】 To solve the shortcomings in traditional methods of personalized recommendation such as sparsity，cold-start，overspecialization and low accuracy problem，this paper proposes a recommendation method based on Website aggregation and knowledge. It gets a movie set to be recommended by Web crawler aggregating Websites，and also builds an ontology-based film model based on which that proposes an algorithm for learning the weights of user preference. It measures the similarity between movies using SimRank method and the weighted average to recommend to users according to the level of similarity. Experimental results show that the accuracy of this method is improved by about ten percent than the existing methods when it is used on non-real-time recommendation. And quality of recommendations is improved significantly on real-time recommendation. In some extent，sparsity，cold-start，overspecialization problem can be solved.

【Key words】 personalized recommendation；Web crawler；Website aggregation；ontology；user preference；cold-start

DOI： 10.3969/j.issn.1000-3428.2014.08.053

1 概述

互联网为广大用户提供了海量的影视资源，但也给用户获取真正感兴趣的资源带来困难。个性化推荐系统有针对性地向用户推荐，提升用户使用感受，提高资源利用率，成为当今研究的热点话题。主流推荐系统常用方法有基于内容的推荐、协同过滤推荐及2种组合的推荐等[1]。

基于内容的推荐技术是最早的推荐方法[2]。基于内容的推荐利用关键词向用户推荐历史记录的相似项目。这类方法是基于句法的，没有考虑语义级别的含义。所以只能推荐与用户历史记录包含的属性值完全相同的项目，导致了基于内容的推荐技术过度专业化。推荐结果缺乏新颖性，不能给用户带来新的惊喜。协同过滤推荐技术[3]是目前应用非常广泛的技术。基于内存的协同过滤用相似统计方法得到与目标用户有相似兴趣的邻居用户，并将最相似的用户历史点播作为推荐源，推荐目标用户未看过的电影。基于模型的方法根据用户历史记录得到一个模型，再用此模型进行预测。协同过滤存在评价稀疏，新资源或新用户加入的冷启动问题。且不同用户可能因为相同选择标准选择不同电影，或出于不同目的选择相同电影，没考虑用户的选择偏好，所以准确率较低。以上方法结合的[4]思路是，先根据协同过滤得到相似用户历史记录，再得到待推荐电影集。统计用户感兴趣的关键词，利用关键词过滤待推荐集，得到结果。这类方法准确度有所提高，但仍无法解决稀疏性、冷启动、过度专业化等问题。为克服以上问题，本文提出一种基于网站聚合和语义知识的推荐方法，提高准确度，并用于实时推荐。

基金项目：中国科学院基金资助重点项目“面向NGB的互联网视频访问控制应用示范子课题”（KGZD-EW-103-5(5)）。

作者简介：周文乐（1989-），女，硕士研究生，主研方向：数据挖掘；朱 明，教授、博士、博士生导师；陈天昊，硕士研究生。

收稿日期：2013-08-20 修回日期：2013-09-17 E-mail：wlzhou@mail.ustc.edu.cn

2 相关概念

2.1 网站聚合

聚合是指通过人工或技术方式，将互联网上的海量信息进行收集、挑选、分析、归类，从而将相关链接内容分类聚合，为网民提供更具针对性的信息。

2.2 本体

本体定义是“给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延的规则的定义”[5]。本体模型是概念化模型，目标是捕获相关领域的知识构造领域概念模型，明确描述领域涉及的概念、概念的含义及概念之间的关系，为简单的术语赋予明确的背景知识[6]。

2.3 SimRank 算法

为衡量结构性上下文相似性，基于“2 个对象是相似的，如果与它们关联对象相似”思想的 SimRank 算法被提出[7]，例如图 1 所示，有向箭头表示网页可链接至另一网页。教授 A、B 都与 $Univ$ 关联，则两者之间有一定的相似性。同理，Student$A$、B 也有一定相似性。文献[7]给出相似度计算公式式(1)。在有向图 G 中，节点 A 指向节点表示为 $O_i(A)$，节点 B 指向节点为 $O_j(B)$，C 为衰减因子，一般取为 0.8。

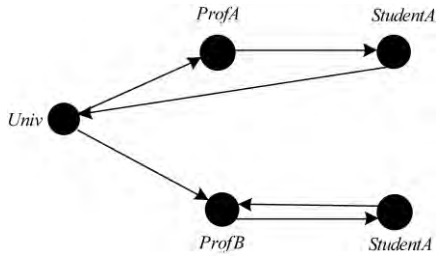

图 1 SimRank 关联示例图

$$s(A,B)=\frac{C}{|O(A)|\times|O(B)|}\times\sum_{i=1}^{|O(A)|}\sum_{j=1}^{|O(B)|}s(O_i(A),O_j(B)) \quad (1)$$

3 基于网站聚合和知识的推荐

3.1 源数据库建立

本文根据用户历史记录，利用网站聚合，获取用户历史项目中每部电影在其他网站上的所有相关推荐，作为待推荐集。本文利用网络爬虫采集数据，爬虫基于 Java 库 HTMLParser Libraries，这种方式比直接利用正则表达式提取网页信息有更强的复用性[8]。本文用到 HTMLParser Libraries 中的 htmlparser.jar 包。访问节点的方法采用 Filer 模式，对不同的网站分别设置过滤条件，对每个节点进行过滤，返回符合规则的节点列表，从而解析得到所需要的电影的相关信息，其中包括每部电影的推荐电影。最后，将在各个网站采集到的信息，利用电影名和上映日期 2 项作为关键字进行去重，将去重的结果存储到数据库中。通过网站聚合建立源数据库有以下优点：(1) 整合资源。很多网站都会向用户推荐电影，且推荐结果不尽相同。将每个网站推荐的资源汇总起来，可丰富源数据库。(2) 运算量少。聚合网站直接采集信息，相比协同过滤查找相似用户的计算量大为减少，也克服了协同过滤中评价矩阵稀疏和系统启动初期用户数量较少的问题。(3) 时效性强。可以根据被聚合网站的更新及时地将新电影信息更新至源数据库。

3.2 影视本体模型构建

基于本体论概念，抽象出用户普遍关注的电影信息并进行描述，构建电影的本体模型。在传统方法中，导演、演员等是不可分属性，如果两个演员不是同一个人，则相似度判断为 0。事实上，观众会认为某两个导演或演员有一定相似性。进而两部电影的导演或者主演之间有一定的相似性，则用户可能会因为喜欢一部电影而对另一部电影产生兴趣。本文分析观众对影人的相似度判断规律，将影人的性别、年龄、国籍等信息，以及主要作品类型和共同合作过的影人作为相似度判断标准。电影属性本体如表 1 所示。

表 1 电影属性本体

类名	属性	数据类型	备注
Movie(电影)	M	字符串	电影编号
	$year$	数值	上映日期
	$TYPE=\{type_1,type_2,\cdots,type_n\}$	字符串集合	类型(可能有多个类型)
	won	布尔值	是否获过奖
	loc	字符串	制片国家
	$TAG=\{tag_1,tag_2,\cdots,tag_n\}$	字符串集合	标签集合(体现用户直观评价)
	$DCT=\{d_1,d_2,\cdots,d_n\}$	Celebrity 实例集合	导演列表
	$ACT=\{a_1,a_2,\cdots,a_5\}$	Celebrity 实例集合	主要演员列表
	$range$	数值	用户评分
Celebrity (影人，即演员、导演等)	C	字符串	编号
	$year$	数值	出生年份
	sex	Male/ Female	性别
	$bplace$	字符串	出生国家
	won	布尔值	是否获过奖
	$Work=\{M_1,M_2,\cdots,M_n\}$	电影编号属性集合	主要作品列表

3.3　带权重的电影相似度计算

在传统推荐算法中，相似度计算主要有 2 种方式：(1) 分别计算项目每个属性的相似度，再对各个属性相似度求算数平均数。(2) 先将项目的信息进行向量化，即用一个向量表征某个项目。对于 2 部电影中所有出现过的属性值，若某个项目含有此属性值，则向量中相应元素置为 1，否则为 0。最后对表征 2 部影片的 2 个向量进行余弦相似度计算。但是，在实际应用中，考虑到不同的用户对某一属性的偏好，即用户对某一属性的重视程度是不同的，而这 2 种传统方法存在的共同问题都是没有考虑到用户的偏好差异。

本文在计算 2 部电影之间的相似度时，考虑用户不同的偏好。首先分别计算每个属性的相似度，再对各个属性相似度求加权算数平均数。例如某一个用户在选择电影时，首先选择自己最喜欢的电影类型——动作片，则在计算相似度时，相应的 TYPE 属性权重应该较大。而某个用户最喜欢某个导演，他对这个导演或者跟这个导演风格相似的导演所拍摄的电影可能感兴趣，则对于这个用户，DCT 属性的权重应相应增大。以下讨论相似度计算公式和各属性权重的计算方法。

3.3.1　相似度计算公式

本文将 2 部电影分别表示为 $M_1, M_2, M_i = \{year_i, TYPE_i, won_i, loc_i, DCT_i, ACT_i, TAG_i, range_i\}$，$i = 1, 2$。用 $sim(property_j)$ 表示 2 部电影在第 j 个属性上的相似度，$w_j, j = 1, 2, \cdots, 8$ 代表每个属性的权重；$sim(x_1, x_2)$ 表示某个属性上，2 个属性值之间的相似度。则两部电影相似度计算公式为：

$$Sim(M_1, M_2) = \frac{\sum_{j=1}^{8} w_j \times sim(property_j)}{\sum_{j=1}^{8} w_j} \tag{2}$$

其中，每个属性的相似度计算方法如下：

(1) 某些用户喜欢最新上映的电影，而一些用户则偏爱老电影，所以电影上映时间也会成为用户选择的依据。设定如果 2 个电影年代相差不大(本文设定为不超过 5 年)，则认为完全相似，否则相差越大，相似性越小。

$$sim(year) = \begin{cases} 1 & \text{if } |year_1 - year_2| < 6 \\ \dfrac{5}{|year_1 - year_2|} & \text{else} \end{cases} \tag{3}$$

(2) 某些用户在选择电影时，其他用户对该电影的评分可能会影响他们对电影的选择，所以本文加入对评分相似度的计算。评分相似度 $sim(range)$ 计算利用式(4)：

$$sim(range) = \frac{1}{1 + |range_1 - range_2|} \tag{4}$$

(3) 求取 2 部电影类型相似度 $sim(TYPE)$ 或者标签 $sim(TAG)$ 的相似度，因为这 2 种属性都为字符串集合，所以相似度计算利用文本相似度的计算方法。令属性值集合分别为 A, B，$X = A \cap B$，$Y = A \cup B$，则 2 个集合相似度计算公式为：

$$sim(A, B) = \frac{|X|}{|Y|} \tag{5}$$

(4) 求取 2 部电影是否都获过国际奖项的相似度 $sim(won)$ 或者制片国家的相似度 $sim(loc)$，令 a，b 为相应属性值，则：

$$sim(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{else} \end{cases} \tag{6}$$

(5) 2 部电影的导演和演员的相似度 $sim(DIR)$，$sim(ACT)$，利用 SimRank 公式有：

$$sim(O(A), O(B)) = \frac{C}{|O(A)| \times |O(B)|} \times \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} s(O_i(A), O_j(B)) \tag{7}$$

其中，$O(A)$，$O(B)$ 分别为 2 部电影的影人(导演或者演员)集合；$O_i(A)$，$O_i(B)$ 为集合中的元素，$C = 0.8$；$s(O_i(A), O_j(B))$ 为具体 2 个影人的相似度。计算规则如下：

通过统计电影人作品的类型和作品中涉及到的其他影人，可以得到这个电影人的擅长类型集合 ($CTYPE$) 以及合作 2 次以上的电影人集合 (PTN)。年龄、获奖情况、出生地以及主要作品的类型集合、合作人集合的相似度计算方法同电影的计算方法，最终得到公式：

$$s(O_i(A), O_j(B)) = \begin{cases} 1 & \text{if } O_i(A), O_j(B) \text{编号相同} \\ \dfrac{sim(year) + sim(TP) + sim(PTN) + sim(won) + sim(loc)}{5} \times 0.8 & \text{else} \end{cases} \tag{8}$$

最终，利用式(2)将以上各个属性相似度综合求取加权平均值，得到电影相似度。

3.3.2　权重计算方法

某用户历史记录中某属性的属性值越聚集于某一值，则说明此用户对该属性的关注程度越高，该属性所占权重越大；反之，某属性的属性值越发散、随机性越大，则说明用户对该属性的关注程度越低，该属性权重越低。假设用户历史记录中有 n 个电影，电影共有 p 个属性，某一个属性共出现 m_p 个属性值，则对于第 j 个属性的聚集程度由式(9)衡量，分子表示属性 j 的第 i 个属性值在历史记录中实际出现的次数，分母为点播电影总个数与属性值的个数

之比，表示每个属性值平均出现的次数。再利用式(10)得到每个属性权重。

$$score_j = MAX_{i=1,2,\cdots,m_p} \frac{count_i}{n/m_p},\ j = 1,2,\cdots,p \tag{9}$$

$$w_j = \frac{score_j}{\sum_{i=1}^{p} score_i} \tag{10}$$

3.4 推荐策略

根据实际应用场景，提出 2 种推荐策略。第 1 种是非实时推荐，当用户登录系统时，根据用户以往的历史记录，向用户推送推荐电影。第 2 种是实时推荐，当用户点开某部电影时，向用户推荐可能感兴趣的相似电影。

3.4.1 场景 1

考虑到用户的兴趣漂移[10]，用户最近观看过的电影能反映这个时间段的兴趣特征。仅选取用户最近观看过的电影集 $X=\{x_1,x_2,\cdots,x_n\}$，通过网站聚合得到候选集 $Y=\{y_1,y_2,\cdots,y_m\}$，并将系统最新收录的电影补充进去，利用式(10)得到用户偏好，再利用式(2)将 Y 中每个项目分别与 X 中所有项目计算相似度。并利用式(11)求平均值，平均值越大，说明此部电影越符合用户这段时间的兴趣特征，依据相似度大小进行推荐。

$$Sim(y_i,X) = \frac{\sum_{j=1}^{n} sim(y_i,x_j)}{n} \tag{11}$$

3.4.2 场景 2

当用户观看某一部电影时，聚合得到此电影候选集，将候选集中每部电影都与此电影计算相似度，按相似度大小进行推荐。基于协同过滤的推荐方法在整个系统启动的初始阶段，由于用户数量少、点播历史记录的缺乏，会出现冷启动问题。而对于其他网站聚合，不需要本系统中的相似用户。另外，针对新用户加入做如下处理：用户加入时，向用户随机推荐热门电影，当用户选择某部电影时，向用户随机推荐网站聚合得到的相关电源，随着观看数量增加，更新对于每个属性的偏好权重。这样可在一定程度上解决新用户问题。

4 实验及结果分析

本文实验数据来自豆瓣电影网，用户可以对看过的电影进行 1 ~ 5 的评分。本文电影信息数据库采集豆瓣电影上的十万部电影。被聚合的网站选取用户数量较多、信息比较全面的热门影视资料网站或者在线视频播放网站，如百度影视宝库、时光网、乐视网、电影网、爱奇艺等。

4.1 实验 1

本实验用于验证算法在应用场景 1 时的有效性。随机抽取豆瓣电影中 100 位用户，用户历史记录条目大于 100。选取用户评分大于 2 的电影，构成实验数据集。每部电影通过聚合网站聚合可得到约 30 部 ~ 40 部相关电影。随机抽取数据集 80% 作为训练集，20% 作为测试集。在训练集中分别抽取在时间上连续的 α 部电影构成某段时间观看的电影集，按照 3.4.1 节所述策略计算相似度，依据相似度大小依次推荐 10 部电影。对照算法采用协同过滤及基于内容的推荐方法结合的算法，首先利用协同过滤得到 15 个相似用户[11]，相似用户计算采用余弦相似度[12]。从而得到待推荐电影集，再选取以某个时间点开始的一段连续时间内用户感兴趣的 α 部电影进行基于内容的推荐。算法有效性利用推荐准确度来衡量：

$$\text{推荐准确度} = \frac{\text{用户感兴趣的电影数}}{\text{推荐电影集数}}$$

在 α 取 8 个不同值时(5，10，15，20，25，30，35，40)，实验结果如图 2 所示。

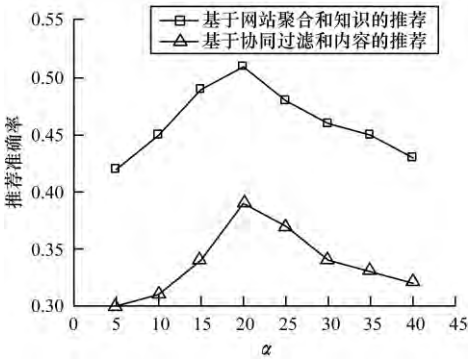

图 2 推荐准确度对比

实验结果显示，比较基于协同过滤和内容的推荐，本文提出的方法平均准确度提高了 10% 以上。另外，从图中可以看出：2 种方法准确率都随着 α 值的改变而变化，且都在 α 取 20 左右时，有较高的推荐准确度。这是因为，在 α 值较小时，用户的观看记录较小，从统计学意义上讲比较随机，不能准确地表现出用户某段时期的兴趣特征，而当 α 值较高时，在这个时间段内，用户可能已经产生了兴趣漂移，推荐准确度会下降。

4.2 实验 2

本实验验证推荐算法在应用场景 2 中的有效性。构建一个实时推荐的模拟系统，招募 30 位电影爱好者作为测试用户。根据实验 1 结果，系统首先令用户输入自己最近看过的、认为比较好的 15 部 ~ 25 部电影。在得到用户历史记录后，算出偏好权重。在此基础上每当用户点开一个电影，系统界面展示 2 组推荐结果，每组 20 部。第 1 组推荐结果由本文算法给出，第 2 组由协同过滤和基于内容的推荐算

法给出。每当得到 2 组推荐，系统要求用户分别对推荐结果的推荐质量做出评价，评价标准为是否相似且感兴趣。质量等级分为好、一般、差。为每位用户进行 30 次实时推荐后，统计得到如图 3 所示的结果。结果显示，本文方法在推荐质量上要优于基于协同过滤和内容的推荐方法。

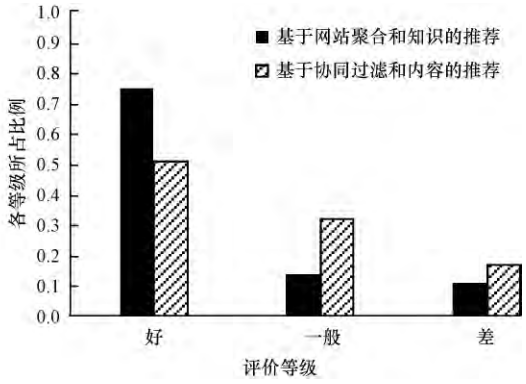

图 3　实时推荐质量对比

5　结束语

本文提出一种基于网站聚合和知识的推荐方法。新系统在建立源数据库时不依靠其他用户群，而是聚合其他权威网站得到相关推荐项目。构建包含影人信息的领域本体模型，可防止过度专业化。并通过属性的聚集程度得到用户偏好差异。实验结果表明，本文方法的推荐准确率和推荐质量较传统方法有明显提高，且可解决冷启动、稀疏性等问题。下一步工作将深入挖掘电影领域本体，例如对电影社会化标签、简介关键词等方面的语义关联挖掘，以改善推荐效果。

参考文献

[1] 许海玲，吴　潇，李晓东，等. 互联网推荐系统比较研究[J]. 软件学报，2009，20(2)：350-362.

[2] Zimmerm J，Kurapati K，Buczak A L，et al. TV Personalization System [D]. Pittsburgh，USA：Carnegie Mellon University，2004.

[3] 易　明. 基于 Web 挖掘的个性化信息推荐[M]. 北京：科学出版社，2010.

[4] 李忠俊，周启海，帅青红. 一种基于内容和协同过滤同构化整合的推荐系统模型[J]. 计算机科学，2009，36(12)：142-145.

[5] 邓志鸿，唐世渭，张　铭，等. Ontology 研究综述[J]. 北京大学学报：自然科学版，2002，38(5)：730-738.

[6] 李　宁，王子磊，吴　刚，等. 个性化影片推荐系统中用户模型研究[J]. 计算机应用与软件，2010，27(12)：51-54.

[7] Jeh G，Widom J. SimRank：A Measure of Structural-context Similarity [C]//Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton，Canada：ACM Press，2002：538-543.

[8] 邱　哲，符滔滔. 开发自己的搜索引擎：Lucene 2.0 + Heritrix [M]. 北京：人民邮电出版社，2007.

[9] 杨卓俊. 基于语义词典的电子商务商品推荐系统研究[D]. 杭州：浙江工业大学，2012.

[10] 涂金龙，涂风华. 一种综合标签和时间因素的个性化推荐方法[J]. 计算机应用研究，2013，30(4)：1044-1047.

[11] 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[J]. 软件学报，2003，14(9)：1621-1628.

[12] 徐　翔，王煦法. 协同过滤算法中的相似度优化方法[J]. 计算机工程，2010，36(6)：52-54.

编辑　顾逸斐

(上接第 276 页)

[4] 王颖颖，张　赟，胡乃静. 在线新事件检测系统中的性能提升策略[J]. 计算机工程，2008，34(15)：72-74.

[5] Linguistic Data Consortium. ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Events [EB/OL]. (2005-05-09). https://www.ldc.upenn.edu/Projects/ACE.

[6] 付剑锋，刘宗田，刘　炜，等. 基于特征加权的事件要素识别[J]. 计算机科学，2010，37(3)：239-241.

[7] 将德良. 基于规则匹配的突发事件结果信息抽取研究[J]. 计算机工程与设计，2010，31(14)：3294-3297.

[8] 姜吉发. 一种跨语句汉语事件信息抽取方法[J]. 计算机工程，2005，31(2)：27-29.

[9] 李　潇，罗军勇，尹美娟. 基于邮件通联关系的邮箱用户权威别名评估[J]. 计算机应用与软件，2011，28(4)：271-273.

[10] 王昭龙，李　霞，许瑞芳. 多关键词查询中 LCA 剪枝概念数的查询扩展技术[J]. 计算机科学，2010，37(4)：132-162.

[11] 汪　洋，帅建梅. 基于语义扩展模型的中文网页关键词抽取[J]. 计算机工程，2012，38(22)：163-166.

[12] 杜金洋，易　河，杨　春. 基于关键词语义扩展的检索策略[J]. 计算机应用，2009，35(6)：1575-1577.

编辑　顾逸斐