

社会化标签语义相似度的协同过滤算法

谌 颀

(广东技术师范学院天河学院 信息与传媒学院, 广东 广州 510540)

摘要: 为解决传统的协同过滤算法不能准确理解用户的喜好,影响推荐准确率和推荐效果,提出基于社会化标签语义相似度的协同过滤算法.算法以标签语义相似度为基础,将项目资源和相关标签的语义信息纳入,显著提高了推荐系统的预测性能.研究表明:与以具体评分数据为基础的算法相比,该算法较好地解决了词相似度和句子相似度计算问题,推荐准确度和性能较以往的协同过滤算法有明显提高,改善了推荐效果.

关键词: 协同过滤; 推荐系统; 社会化标签; 语义相似度; 预测性能

中图分类号: TP 301.6 **文献标志码:** A

网络信息迅猛增长带来了日益严重的信息过载问题^[1],个性化推荐系统可帮助用户在海量信息中有效搜索关心的资源.这些系统一般采用了基于内容、协同过滤或两种方法混合的技术^[1-2].虽然这些传统的推荐技术应用广泛,但是它们在理解用户喜好方面存在不足,因此推荐精确度和效果有较大的影响.标签系统为用户提供了另一种实现资源推荐的新方法,它是一种提供基于 Web 的服务,用户能使用简短的语言描述对网页信息资源进行分类的标签技术^[2].此外,标签还为内容相似性的比较提供了方法.考虑标签和项目资源的语义差别,本文提出了一种基于社会化标签语义相似度的协同过滤算法.

1 标签的概念模型

标签(tag)是用户为项目资源自由、随意添加的一组标记或注解,用户的这种自主行为具有重要的社会意义,它可以更好地帮助用户组织资源、浏览资源和推荐资源^[3].为了便于基于经验数据和模型的标签系统实现推荐行为,首先要建立一个通用可行的概念模型.考虑到通用性和有效性,研究选择采用三元组模型,包含用户集合(users)、标签集合(tags)和项目资源集合(items)等 3 个实体集合.

标签系统建立起了用户、标签和资源三者间的动态关系,如图 1 所示.用户集合区域是用户空间,包含了该系统中所有用户集合,每个元素代表一个用户;标签集合区域是标签空间,包含了该系统中全部标签集合,每个标签对应一个词(如“Avatar”)或一个短语(如“Forrest Gump”);项目资源集合区域则为项目资源空间,包括所有的项目集合,每个项目通常由一个唯一的编号标记.

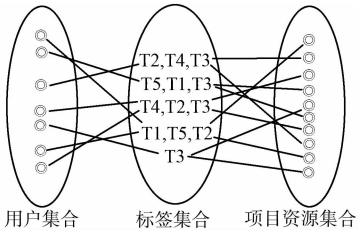


图 1 标签系统的动态关系

Fig. 1 Dynamic relationship of tag system

2 标签喜好预测

2.1 项目交互的预测方法

一种基于用户与具有相关标签的项目交互的预测方法,如图 2 所示.用户在访问资源时往往有两类行为:一是新增、查找、关注标签的行为;二是用户浏览项目资源时的交互行为,如点击、浏览、评分、收藏等^[4].通过对两类行为的分析可知用户感兴趣的内容.研究发现,如果将项目与标签之间的相关性作为

假定两个句子 X 和 Y , 其中 X 的长度为 m , Y 的长度为 n . 算法的主要步骤: 1) 分词; 2) 提取词干; 3) 词性标注; 4) 词义消歧; 5) 将分词结果组成词语义相似度矩阵 $S[m, n]$, 若是缩写词, 如 SCI (science citation index), 先查缩写词典, 再计算相似度值; 6) 采用 Hungarian 方法将句子语义相似度计算问题转化成两个图的最大匹配权重计算问题, 取句子 X 和 Y 中的词分别为两个图中的定点^[7]; 7) 把前面的

计算步骤合并计算,得到句子的相似度.

选用 Dice 相关系数计算句子的相似度(sim-sentences,SS),即

$$SS(X,Y)=\frac{2\times|X\cap Y|}{|X+Y|}.$$

(5)

首先,设一个参照值;然后,依次对所有数据对进行匹配,若匹配分值超过参照值,说明这两个词的语义相似,予以保留^[7];最终,通过式(5)可得匹配关系值即句子相似度值.例如,给定两个句子 S 和 T ,其长度分别为 5 和 3.根据图的匹配算法得到 $S[1]$ 与 $T[1]$ 进行匹配,其匹配分值是 0.8, $S[2]$ 与 $T[2]$ 的匹配分值为 0.7, $S[3]$ 与 $T[3]$ 的匹配分值为 0.75.使用 Dice 相关系数,设参照值为 0.5,可以看出,3 个分值都大于参照值,因此,选择这 3 个匹配对.由式(5)计算可得比值,即句子的相似度为 $2\times(1+1+1)/(5+3)=0.75$.

4 用户评分预测和资源推荐列表的计算

预测用户 u 对他未曾购买过的项目资源 i 的评分.一方面,从式(3)可推导出用户 u 对标签的喜好 $NTP(u,t)$,从而得到其感兴趣的 u_{pre} 标签集合.即设参照值 θ ,用户 u 添加标签 t ,如存在 $NTP(u,t)>\theta$,则说明用户对它感兴趣,就将它加入到 u_{pre} 标签集合中.另一方面,还要得到与项目相关度最高的 i_{rel} 标签集合.即设参照值 ω ,如果有 $w(i,t)>\omega$,则说明此标签与项目相关度高,就将它加入到 i_{rel} 标签集合中.然后,计算出 u_{pre},i_{rel} 两个标签集合的相似度值(SR),即 $SR(u,i)=SS(u_{pre},i_{rel})$.此值就是用户 u 对未购买项目 i 的预测评分.

需要说明一下,如果把两个集合中的所有标签(单词、短语或句子)分别串起来就可以看成是两个“句子”^[8],那么,采用以上句子相似度算法计算两组标签集合的相似度是合理的^[9-10].

重复对用户 u 所有未购买过的项目进行 $SR(u,i)$ 的计算,并将全部预测评分结果从大到小排序,取前 N 个项目资源,即是 N 推荐列表.

5 实验与分析

实验选择 MovieLens(<http://www.grouplens.org>)第三组数据集作为实验数据集,该数据集含有 71 567 名用户对 10 681 部电影的 10 000 054 条评分数据及 95 580 个标签.

通过 N_{top} 方法计算出项目预测准确度.首先,将推荐产生的 N 个资源置于训练集中;然后,检测 N 个资源有多少存在于测试集中,量越多算法的准确度就越高,反之,就越低.

用户集中全部用户重复以上实验后,计算得出全体用户兴趣度的平均值.平均 N_{top} 准确度计算式为

$$N_{pre}=\frac{\sum_{U_i\in U}\frac{|R_{u_i}|}{|I_{u_i}|}}{|U|}.$$

(6)

式(6)中: I_u 为设用户 U_i 已标注过标签的电影集; U 为用户集合; R_{u_i} 是训练集中用户 U_i 的推荐列表; I_{u_i} 是用户 U_i 测试集中的电影集.如果 N_{pre} 的值越高,则说明对应推荐算法的推荐准确度越高.

基于 CFBTSS 算法做了 5 和 10 两组实验,其中,推荐列表长度分别取 5 和 10,并且分别用 Cosine-tag 和 Implicit-item 算法做了同样的实验,然后对三者进行比较,结果如表 1 所示.从表 1 可以看出:无论推荐列表长度是 5 还是 10,CFBTSS 算法结果皆优于另两种,其推荐准确度更高.

比较推荐列表的平均评分评价算法的有效性,分数越高,用户对推荐结果就越满意,算法也越有效.同样做了推荐列表长度分别为 5 和 10 的两组实验.当推荐列表长度为 5 时,CFBTSS 算法的平均分和准确度为 4.3,0.72;当推荐列表长度为 10 时,其平均分和准确度分别为 3.6,0.68.由此可知,推荐列表 5 的用户评分平均达到了 4.3 颗星,表明用户比较喜欢这些电影.从实验

表 1 不同算法的预测准确度比较

Tab.1 Compare prediction accuracy of different algorithms

算法	列表长度	
	5	10
Cosine-tag	0.625	0.550
Implicit-item	0.668	0.630
CFBTSS	0.732	0.690

结果看,提出的 CFBTSS 算法能够推荐更符合用户喜好的电影,表示这样的推荐更有意义.

以上两个实验结果表明:CFBTSS 算法的推荐精确度和性能与传统个性化推荐算法相比有了较大地改善.

6 结束语

文中算法以标签语义相似度为基础,与以具体评分数据为基础的算法相比,具有更好的理解意义和准确度,对改进个性化推荐系统具有重要意义.下一步工作是研究改进语义差别计算的方法及如何更加科学地描述项目资源的属性.

参考文献:

[1] 王国霞,刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用,2012,48(7):66-74.

[2] BEGELMAN G,KELLER P,SMADJA F. Automated tag clustering: Improving search and exploration in the tag space[C]//Proceedings of the 15th International Conference on World Wide Web. Edinburgh:ACM Press,2006:1-5.

[3] 谌颖. 使用分类改进标签推荐系统准确度的研究[J]. 微电子学与计算机,2011,28(5):96-93.

[4] 陈叶旺,李海波,余金山. 一种基于农业领域本体的语义检索模型[J]. 华侨大学学报(自然科学版),2012,33(1):27-32.

[5] 杨现民,余胜泉. 学习资源语义特征自动提取研究[J]. 中国电化教育,2013(11):74-80.

[6] 符征. 语义引擎的形成及其应用[J]. 自然辩证法研究,2013(11):21-25.

[7] 邓双义. 基于语义的标签推荐系统关键问题研究[D]. 上海:华东师范大学,2009:37-41.

[8] 许棣华,王志坚,林巧民,等. 一种基于偏好的个性化标签推荐系统[J]. 计算机应用研究,2011(7):2573-2579

[9] 崔林,宋瀚涛,陆玉昌. 基于语义相似性的资源协同过滤技术研究[J]. 北京理工大学学报,2005,25(5):402-405.

[10] 荀恩东,颜伟. 基于语义网计算英语词语相似度[J]. 情报学报,2006,25(1):43-48.

Collaborative Filtering Algorithm Based on Social Tags Semantic Similarity

CHEN Hang

(Information and Communication College, Tianhe College of Guangdong Polytechnic Normal University, Guangzhou 510540, China)

Abstract: In order to solve the traditional collaborative filtering algorithm can not accurately understand the user's preferences, affect the recommendation accuracy and recommendation effect, a collaborative filtering algorithm based on social tags semantic similarity is proposed. Based on the semantic similarity of tags, the semantic information of project resources and related tags is included, and the prediction performance of the recommendation system is significantly improved. Research results show that: compared with the algorithm based on the user rating, the proposed algorithm can solve the problem of word similarity and sentence similarity computation, and the recommendation accuracy and recommendation effect, as well as the performance of the proposed algorithm is significantly improved compared with the previous collaborative filtering algorithm.

Keywords: collaborative filtering; recommendation system; social tags; semantic similarity; prediction performance

(责任编辑: 黄晓楠 英文审校: 吴逢铁)