

使用分类改进标签推荐系统准确度的研究

谌 颀

(广东技术师范学院 天河学院计算机系, 广东 广州 510540)

摘 要: 由于标签的灵活性及其概念可理解性, 使用标签可以提高推荐系统的推荐性能. 协同标签系统在网络资源推荐服务中取得了巨大的成功. 分类为用户显示了不同的利益群体的不同喜好. 基于此, 提出了基于分类的标签推荐系统—TRSUC, 将它作为内分类推荐, 使分类标签成为全球用户和项目之间的中介实体. 通过对 MovieLens 中数据集进行实验, 结果表明, TRSUC 的推荐准确度明显优越于传统推荐算法.

关键词: 协同标签; 推荐系统; 分类; TRSUC; 准确度

中图分类号: TP319

文献标识码: A

文章编号: 1000-7180(2011)05-0093-04

Improving the Accuracy of Tagging Recommender System by Using Classification

CHEN Hang

(Department of Computer, Tianhe College of Guangdong Polytechnic Normal University, Guangzhou 510540, China)

Abstract: Collaborative tagging system has become more and more popular and recently achieved widespread success due to flexibility and conceptual comprehensibility of tagging systems. Recommender system has the access to adopt tagging systems to achieve better performance. In this paper we consider that the items can be categorized into different classifications in which users show different interests. Here we adopt a two-step recommender method called TRSUC (Tagging Recommender Systems by Using Classification) which can be described as Inner-Class Recommender or Global Recommender in which we use tag as the intermediary entity between user and item. The experiment using MovieLens as dataset shows that we acquire better results than the recommender algorithms without classifying the items.

Key words: collaborative tagging; recommender system; classification; TRSUC; accuracy

1 引言

社会标签是用户以关键词形式对文章、图片等信息资源进行分类并加以标注的方式, 它是 Web2.0 的重要技术之一. 作为一种更为灵活、有趣的信息分类方式, 标签已成为推荐系统的重要信息来源, 许多优秀的推荐系统都采用了标签做推荐, 如 Amazon, last.fm 等. 标签是用户特意标注的, 它含有用户信息和所标注的产品信息, 所以标签不仅代表有意义的概念, 而且反映了用户对产品的理解^[1]. 因此, 用户和产品的特性都可用标签来描述.

文中提出了基于分类的标签推荐系统 (TRSUC). 该算法通过反映用户对不同分类标签的兴趣程度来提高标签推荐系统的推荐准确度. 实验结果表明, 使用分类步骤的标签推荐系统的性能优于传统的推荐系统.

2 标签系统三元组模型

为了实现推荐, 需要建立一个通用的标签概念模型. 目前被广泛接受的是标签三元组模型, 该模型包含三个实体集合, 即用户集合 (users)、标签集合 (tags) 和被标注的项目资源集合 (items). 如图 1 所

收稿日期: 2010-09-13; 修回日期: 2010-11-28

基金项目: 广州市番禺区科技计划项目 (2010-z-37-1)

示,每一元组都形成独立空间,每一空间都由与边线相连的节点集组成.第一个是用户空间,包括标签系统的所有用户集合,其中每个节点表示一个用户;第二个是标签空间,包括标签系统中所有的标签集合,其中一个标签对应一个关键词或短语;第三个是项目资源空间,包括所有的项目集合.

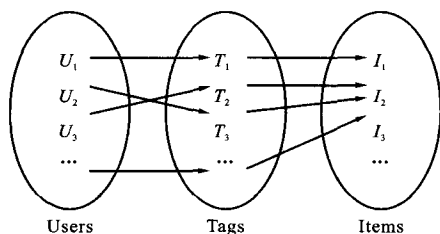


图1 标签系统三元组模型

标签受注于产品并能够描述产品的特性,同时代表了用户兴趣和产品特征,并可以反映用户对某一具体产品的兴趣.标签为帮助用户找到潜在产品起到重要的中介作用,而分类则为产品描述提供了更多的信息.实验结果表明:用户对某一分类的标签都有不同程度的喜好,用户的兴趣可以用产品分类信息来划分,进而发现用户对每一类产品的兴趣(喜好)及其程度.根据上述观点,标签分类的方法与基于用户^[2]和基于内容^[3]的协同过滤算法或是文献^[4]的SVD(奇异值分解)方法相比有更多优势,它不受产品特征和信息提取的局限,从很大程度上提高了标签推荐系统的准确度.

3 推荐算法

3.1 基于标签的推荐

我们采用了标签相关度和标签喜好度两个概念,但对二者的计算方法不同.

3.1.1 标签相关度算法

标签相关度指的是标签和产品之间关系程度.TF-IDF 是一种用于资讯检索与资讯探勘的常用加权技术,用以评估某一关键字词对于一个文件集或一个语料库中的其中一份文件的重要程度.字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降.TF 指的是词条在文档中出现的频率,IDF 常用于过滤一些非关键字(词),如 the, is.文献^[5]把标签分类为主观性、客观性及隐私性的;认为大部分标签都是客观类,其次是主观类,也就是说大部分标签都较真实地反映了与产品的相关性.文献^[6]得出重要结论:用户所标注的每一类标签在整个标签集中所占的比例相对稳定.这表明用户对某一具体产品

的喜好或对标注这种产品的标签的喜好呈相对稳定的趋势.

根据上述结论,用户所应用的标签很有可能与某一特定的产品更相关,因此 TF(词频)可以用来计算标签与产品之间的相关度^[7],且这些标签可以用来描述该产品的特性.由此得出公式:

$$\text{tag_rel}(i, t) = n_{t,i} / \sum_{k \in T_i} n_{k,i}$$

3.1.2 标签喜好度算法

标签喜好度是指用户对某一标签的喜爱程度.研究中我们根据用户的评分来预测他们对标签的喜好程度.如果用户倾向于对具有特殊标签的产品评分高,系统就能推算出该用户对该标签的喜好程度高.为了评估某一用户对某一单个标签的喜好,它可用来计算用户对电影评分的加权平均值.另外,标签相关度是另一个应考虑的因素.比如,在所使用的数据集中,用户用“violence”标注“Reservoir Dogs”7次,而“Planet of the Apes”仅标注过1次.显然用户将电影“Reservoir Dogs”分类为“violence”的比重更高.用 I_u 表示用户 u 对产品 i 的评分,则标签喜好度可计算为:

$$\text{tag_pref}(u, t) = \frac{\sum_{i \in I_u} r_{u,i} \cdot \text{tag_rel}(t, i)}{\sum_{i \in I_u} r_{u,i} \cdot \text{tag_rel}(t, i)}$$

注意: I 是用户 u 使用标签 t 所标注的产品集,分子和分母的总和都没有考虑用户未评分的产品,其中每个标签用户喜好的权重总和应为 1.0.

3.1.3 基于标签的推荐算法

使用标签相关度和标签喜好度来产生推荐的推荐算法源自于信息检索的一个算法,计算用户的概要文件向量和一个文档的词向量之间的相似度.在信息检索中,向量的每一个单元对应的是一个词,而在标签推荐算法中,向量的每一个单元对应的是标签.该算法通过计算用户对项目 i 的标签的喜好和标签 t 在项目 i 中的权重 $\text{tag_rel}(i, t)$ 来计算用户 u 对项目 i 的预测分数,计算公式可表示为

$$\text{predict_score}(u, i) = \sum_{t \in T_i} \text{tag_pref}(u, t) \cdot \text{tag_rel}(i, t)$$

3.2 使用分类的标签推荐算法

3.2.1 分类喜好

在现实生活中项目可以划分为不同的类.分类的主要算法有两种:分类算法和聚类算法.分类算法使用上述所说的产品的类来区分产品;聚类算法常常通过分析标签内容的特征来分类.K-Means 是一种常用的聚类算法.因为我们使用分类算法来划

分项目,为此提出来了一个新概念:分类喜好。

当用户浏览网站或选择某一项目时,分类信息往往是过滤掉用户不感兴趣的产品的第一步,如在浏览电影网站时,非常喜欢“动作”片的用户就不会对“爱情”类等电影过多注意,而喜欢“爱情”片的用户则不会对“动作”片过多关注。这意味着用户对每一种分类的喜好程度不一样。也就是说,用户的兴趣可以根据产品的分类信息将其分成多个部分。因此,用户对每一类项目的兴趣程度(如图2所示)可以使用不同的权重—分类喜好度—来反映,进而使用分类喜好度帮助用户顺利的选择分类产品。

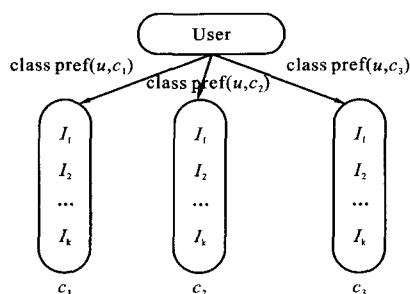


图2 用户对项目分类的映射图

那么,如何估算用户对每一类产品的喜好度呢?用户评分可做为分类喜好度的计算方法。如果用户给一特定分类的产品评分高,就表示用户对该类产品很感兴趣。通过观察各分类的项目空间和整个项目空间,将各个分类的平均评分值和所有产品的平均评分值进行比较得出

$$w(u, c_k) = \frac{\sum_{i \in c_k \cap i \in i_u} r_{u,i}}{\sum_{i \in i_u} r_{u,i}}$$

因为指数函数使权重更自然的接近1.0,所以权值公式可修正为

$$\text{class_pref}(u, c_k) = \exp(w(u, c_k) - 1).$$

3.2.2 使用分类的标签推荐

如上所述,在某一分类域内,用户常常先过滤掉其不关注的信息,然后选择其感兴趣的产品。而使用分类的推荐算法正好相反,其算法中也可分为两步(图3)。

(1) 内分类推荐过程

内分类推荐过程是指分别在各个分类内进行推荐。在此,我们使用标签喜好度和标签相关度分别描述用户与项目,然后使用标签推荐法进行推荐并在各分类中产生推荐列表。因此内分类推荐算法可以表示为

$$\text{Inner_Score}(u, i) = \sum_{t \in T_i} \text{tag_pref}(u, t) \cdot \text{tag_rel}(i, t).$$

(2) 全局推荐过程

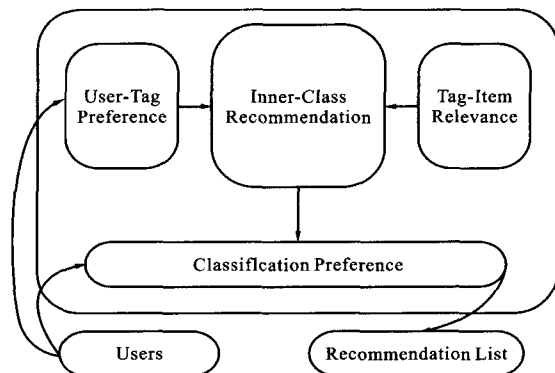


图3 分类推荐系统模型

全局推荐过程是模拟用户行为寻找他们更感兴趣的项目的一种方法。最终的推荐列表是由计算 $\text{Item_Rank}(u, i)$ 产生:

$$\text{Item_Rank}(u, i) = \max\{\text{class_pref}(u, c_k) \cdot \text{Inner_Score}(u, i)\}$$

(某些项目可能落入几个分类中,这里我们使用的是具有最高预测分值的项目)

4 实验结果与评价

文中做了两个实验来比较 TRSUC 算法和未使用分类信息的标签推荐算法的性能。实验采用 MovieLens 作为算法应用的数据集,此数据集包括 7 1567 名用户对 10 681 部影片的 10 000 054 条评价数据(5 分制)和 95 580 个标签,在分析过程中将数据集中的电影分为 18 类:Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, and Western.

实验一将显示并验证上述分类喜好度的计算方法。从数据集中随机抽选三个用户,如图4所示,用户对每一分类表现出不同的喜好程度。如用户 69 388 对 Animation, Film-Noir, IMAX 和 Western 表现出较高的兴趣,而对 Horror 的电影兴趣不高。因此当系统做推荐时,对用户 69 388 不应过多推荐“Horror”类电影;其他用户对不同分类也表现出不同的喜好,这让我们更有信心使用分类信息来修正推荐结果。

实验二保留了用户先前标注的电影,所以产生了 top-N 推荐列表。并记录每一被保留的电影是否出现。对于每一个活跃的用户 u , I 表示用户 u 已标注过的电影集, R_{ui} 是推荐列表中的电影集。因此平均准确度公式定义如下:

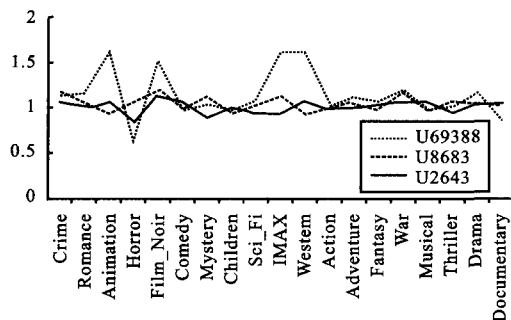


图4 用户对不同分类喜好程度的分布图

$$\text{top_n_precision} = (\sum_{u \in U} \frac{|R_u|}{|I_u|}) / |U|$$

我们为用户空间中的每一个用户做同样的实验,并计算出所有用户兴趣预测准确度的平均值。如果 top-n-precision 的值越高,则说明对应的推荐算法的推荐准确度越高。实验中,我们观察每一个活跃用户,且每次都查找 top-5, top-10 和 top-15 推荐列表,TRSUC 与未使用分类的标签推荐系统的推荐准确度结果比较如图 5 所示。从图 5 中可知,分类标签推荐系统的推荐准确度更高,尤其在 top-5 中表现更为明显,这一实验结果验证了该算法为活跃用户找到相关性电影的性能。

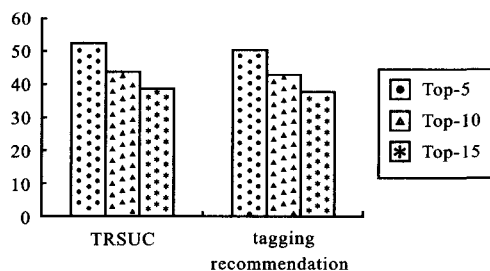


图5 TRSUC 与传统标签推荐系统的准确度比较

最后,还详细分析了 TRSUC 算法的时间复杂度,并且对比了该算法与传统标签推荐算法的性能。虽然分类喜好度的计算要占用较多 CPU 处理时间,但是众所周知,用户对每一类产品的喜好度相对稳定,因此我们可以在离线时完成分类喜好度的计算,并周期性地重复该过程以更新喜好度系数。内分类推荐过程只对每一分类中的项目进行筛选,将注意力资源分散在不同类别的项目中,并生成每个类对应的推荐列表,而传统标签推荐系统则要在整个项目空间中进行筛选。因此,在时间和注意力分配方面,该方法对在线实时推荐有重要意义。同时由于用户集群喜好的频率低,因此认为这并不是该算法的瓶颈。

5 结束语

标签既代表用户特征又代表项目特征,通过标签相关度和标签喜好可以找到用户与项目之间的关系。此外,用户也常常用分类信息的方法来过滤其不感兴趣的项目。而分类标签推荐过程与用户查找项目的过程恰好相反,在产生最终推荐列表前,对不同的分类用相应的权重来修正不同类中的推荐列表。实验结果表明,基于分类的标签推荐算法提高了标签推荐系统的准确度。在今后的研究中,将着重研究如何得到更准确的分类喜好度,除用户评分外,还应考虑用户标注行为因素。

参考文献:

- [1] Sen S, Vig J, Riedl J. Learning to recognize quality tags [C]// Proceedings of IUI, 2009:87-96.
- [2] Sen S, Vig J, Riedl J. Tagommenderers: Connecting users to items through tags [C]// International World Wide Web Conference. Chiba: IW3C2, 2009:671-680.
- [3] 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.
- [4] Funk S. Netflix update: Try this at home [EB/OL]. [2010-03-02]. [http:// Sifter. Org/simon/journal/20061211. html](http://Sifter.Org/simon/journal/20061211.html).
- [5] Sen S, Lam S K, Rashid A M, et al. Tagging, communities, vocabulary, evolution [C]// Proceedings of the ACM 2006 Conference on CSCW. Najing: ACM, 2006: 181-190.
- [6] Scott A Golder, Bernardo A Huberman. The structure of collaborative tagging systems [EB/OL]. [2010-03-02]. [http://www. hpl. hp. com/research/idl/papers/tags/](http://www.hpl.hp.com/research/idl/papers/tags/).
- [7] Sen S, Harper F M, LaPitz A, et al. The quest for quality tags [C]// Proceedings of the 2007 international ACM conference on Supporting group work. Sanibel Island: ACM, 2007:361-370.

作者简介:

湛 颀 男, (1980-), 硕士研究生, 讲师, 高级工程师。研究方向为软件开发、数据挖掘、个性化推荐算法、网络信息安全。