

基于联合物品搭配度的推荐算法框架

姚静天¹, 王永利¹, 侍秋艳¹, 董振江^{2,3}

(1. 南京理工大学 计算机科学与工程学院, 南京 210094; 2. 上海交通大学 计算机科学与工程系, 上海 200240;

3. 中兴通讯股份有限公司 云计算及 IT 研究院, 南京 210012)

摘要: 针对现有推荐系统大多基于物品(用户)相似度进行计算,其推荐结果无法兼顾推荐对象的搭配性特征的问题,提出了一种基于联合搭配度的推荐算法框架.该算法框架中的联合搭配度模型,结合了用户交互反馈、物品的文本和结构化知识3方面的信息,分别计算目标物品与候选物品的搭配程度,然后利用逻辑回归算法进行搭配度融合,可以得到与目标物品最相搭配的物品推荐列表.通过在淘宝真实数据集上的实验,该推荐算法框架相比于传统基于相似性的推荐算法,显著提高了搭配推荐的性能,同时在用户交互记录较少的情况下也能有较好的精确度.

关键词: 推荐系统; 物品搭配; 协同过滤; 专家知识; 相似度

中图分类号: TP 391 **文献标志码:** A

Joint Match Degree of Items for Recommendation Systems

YAO Jingtian¹, WANG Yongli¹, SHI Qiuyan¹, DONG Zhenjiang^{2,3}

(1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing

210094, China; 2. Department of Computer Science and Engineering, Shanghai Jiao Tong University,

Shanghai 200240, China; 3. Cloud & IT Institute, ZTE Corp., Nanjing 210012, China)

Abstract: Since most recommendation systems are based on the calculation of items' or users' similarity, the results can't give consideration to both the complementarity and similarity of recommended objects. An algorithm framework for calculating the joint match degree of items for recommendation systems was proposed. In the framework, combining with the informations of users' interaction feedback, items' textual knowledge and structural knowledge, the joint match degrees of the target item and those candidate items were calculated respectively. Integrating the match degrees by using logistic regression, a list of items matched with the target item was obtained. Through the experiments on a Taobao real data set, it is indicated that the model significantly improve the performance of recommendation collocation compared to the recommendation algorithm based on similarity only. Moreover, in the situation of fewer users' interaction record, the model can also have better accuracy.

收稿日期: 2016-10-09

基金项目: 国家自然科学基金资助项目(61170035, 61272420, 61502233); 2012年国家科技重大专项(2012ZX03002003); 江苏省科技成果转化专项资金项目(BA2013047); 江苏省六大人才高峰项目(WLW-004); 兵科院预研项目(62201070151); 中央高校基本科研业务费专项资金项目(30916011328)

第一作者: 姚静天(1992-), 女, 硕士研究生. 研究方向: 推荐系统、大数据分析、社交网络. E-mail: codingyjt@gmail.com

通信作者: 王永利(1974-), 男, 教授. 研究方向: 数据库与大数据分析、智能服务与云计算、模式识别等. E-mail: yongliwang@njjust.edu.cn

Keywords: *recommend system; item match; collaborative filtering; expert knowledge; similarity*

随着互联网和电子商务的发展,推荐系统已然成为工业界和学术界研究的一个重要课题.推荐系统通过抓取用户行为偏好、项目基本信息等,为用户提供个性化的推荐,帮助用户更快更好地找到自己需要的信息,以促进购买或点击行为的发生.

目前,推荐系统中常用的方法通常分为3类:协同过滤推荐算法、基于内容的推荐算法和基于专家知识的推荐算法.其中,协同过滤^[1]是目前研究和应用最为广泛的技术,可以通过研究用户对项目的评分,从基于群体用户的行为分析或兴趣偏好相似性的度量的角度,为用户推荐类似项目,而不需要对项目本身内容进行分析或需要专家知识的辅助.这种方法适用性很广,可以很方便地描述电影、音乐、图片等难以进行文本描述的对象,但也存在着诸多问题,比如数据稀疏性、冷启动问题等.基于内容的推荐算法^[2],则是利用用户过去喜欢的物品,抽取特征,为其推荐特征相似的物品.而基于专家知识的推荐算法^[3]则利用专家的先验知识,构建物品之间的关系,从而为用户进行推荐.后两种算法可以有效地弥补协同过滤算法带来的数据稀疏性和冷启动等问题,因此,也有一些推荐系统将这些方法按某些方式组合在一起,以达到更好的推荐效果.

通常的推荐系统,往往是通过用户对物品的评分数据,从物品或用户兴趣相似性的角度,给用户推荐类似的物品,而鲜有针对物品互补性搭配的推荐.而比起用户对物品的评分数据,用户的购买/点击行为的数据则更为丰富.此时,传统的基于物品相似度的推荐系统已经无法满足用户的实际购买需求.而后出现的服装风格推荐系统^[4-5],虽然考虑了物品之间的互补性和协调性,但却需要时装设计师的专家知识,人工形式化关键属性,建立它们之间的关系,这需要大量的人工操作,复杂而低效.

为了解决上述问题,本文提出了一种新的物品搭配度概念,从文本知识、结构化知识和用户交互反馈信息的角度对物品搭配度建模,然后提出了一个联合物品搭配度算法框架(joint match degree of items, JMDD),可以更准确地描述物品之间的搭配关系,为查询的每个物品生成一个相应的搭配物品推荐列表.它无须耗时复杂的人工标记或评分,而是利用了简单易得的物品本身的内容信息和用户的交互行为记录,再结合专家给出的物品搭配信息,为物

品之间生成了搭配程度的评分结果.实验证明,论文提出的算法显著提升了推荐结果的精确度、召回率和排序准确度,在系统冷启动过程中,也可以表现出较好的结果.同时,本文提出的模型也被证明在大数据集上也有较高的效率.

1 相关概念与问题定义

本文将首次提出一些术语,然后明确要解决的问题.

1.1 物品搭配度

本文关注的重点在于物品之间的搭配程度,相搭配的物品之间必定有某些特征相似或互补的关系,才能使得他们搭配起来和谐统一.而人们的审美各不相同,从单个用户的角度量化搭配程度显然不是一种客观的评价方式.因此,本文试图根据大多数用户以及专家达人的角度,对物品之间的搭配程度进行量化定义,即“物品搭配度”:

定义 1 物品搭配度 $D(i, j)$, 表示物品 i 和物品 j 之间的搭配程度, 搭配度越高, 表示两者搭配起来相协调的可能性越高, 反之则更不协调.

为了统一量化标准,物品搭配度的值需要进行离差标准化,最终取值范围为 $[0, 1]$.物品搭配度可以用来产生每个物品的搭配物品列表,从而推荐给用户,帮助用户进行物品的挑选.

1.2 用户交互反馈

本文研究的目的是挖掘物品之间潜在的关系,而用户的交互行为可以从侧面反映出物品之间的联系.假设有 M 个用户、 N 个物品,行为观察的时间段为 T ,定义用户行为反馈为一个三阶的张量 $\mathbf{R} \in \mathbb{R}^{M \times N \times T}$.

$$R_{k,j,\eta} = \begin{cases} a, & \text{用户 } k \text{ 在 } \eta \text{ 时刻对} \\ & \text{物品 } j \text{ 产生了交互行为} \\ 0, & \text{否则} \end{cases} \quad (1)$$

式中, a 表示在时间点 η 时,用户 k 对物品 j 产生了交互行为的权重 $a(a \in \mathbb{R}^+)$.例如:在某个时刻,用户点击浏览了某个物品,或者生成了该物品的订单,则该交互行为的权重就为某个正实数.需要注意的是, $\mathbf{R}$ 中值越大的情况并不一定意味着用户 k 对物品 j 越有兴趣,也有可能浏览或下单之后发现自己并不喜欢它;而值为 0 也并不一定说明用户 k 对物品 j

不感兴趣,有可能只是用户 k 还未注意到该物品.

1.3 搭配知识库

事实上,除了上文所提及的用户行为反馈,与物品直接相关的各种知识也可以从另一方面帮助推荐系统挖掘物品之间的联系,提高其推荐质量.因此,在推荐系统中,每个物品都可以一一映射到知识库中描述该物品的一系列实体(例如,一件物品通常可以映射到由物品类别、物品图片等信息组成的描述该物品的实体上),而这些实体,在本文中就称为“物品实体”.

具体而言,综合知识库中包含的物品实体信息根据其特点一般可分为3种类型.

定义2 文本知识. 对于一个物品实体,例如知识库中的电影或物品,通常会用文本概述来表示文本知识,它往往是电影的主题梗概或物品的标题.

定义3 结构化知识. 这种知识是一种可以包含多种类型实体或实体间关系的网状结构. 例如对于物品推荐来说,实体通常包括物品本身和物品类别(如服饰)、地点等属性,以及描述这些实体间关系的链接(如产地关系、从属关系、搭配关系等).

定义4 图像知识. 除了上述文本型描述,知识库中还可以用图像来代表一个实体. 例如,用展示图代表一个物品或是用海报图代表一部电影.

结构化知识与上文所述的用户交互反馈作为物

品的结构特征,文本知识和图像知识则是内容特征,它们共同构建了物品搭配度的模型. 考虑到图像知识中包含的特征与文本知识有一定的重合,且其数据量远远大于文本数据,处理过程复杂耗时,本文暂时不考虑图像知识的处理.

1.4 问题定义

本文所研究的问题为:给定一个物品的集合,提供其文本、结构化知识库,以及用户交互反馈信息,挖掘出物品之间的搭配关系,为每个待测物品生成一个搭配列表用于推荐给用户.

2 算法框架与模型

现介绍如何从物品实体的文本知识、结构化知识和用户交互反馈信息中提取物品搭配度关系模型,从而为物品对生成统一的搭配度指标.

2.1 算法框架

具体而言,本文提出的基于联物品搭配度的推荐算法框架主要分为2个阶段:基于内容/结构特征的物品搭配度建模和协同模型融合过程,如图1所示.

其中,物品搭配度的建模包含3个部分:基于文本知识(textual knowledge)的物品搭配度模型(JMDI(T))、基于结构化知识(structural

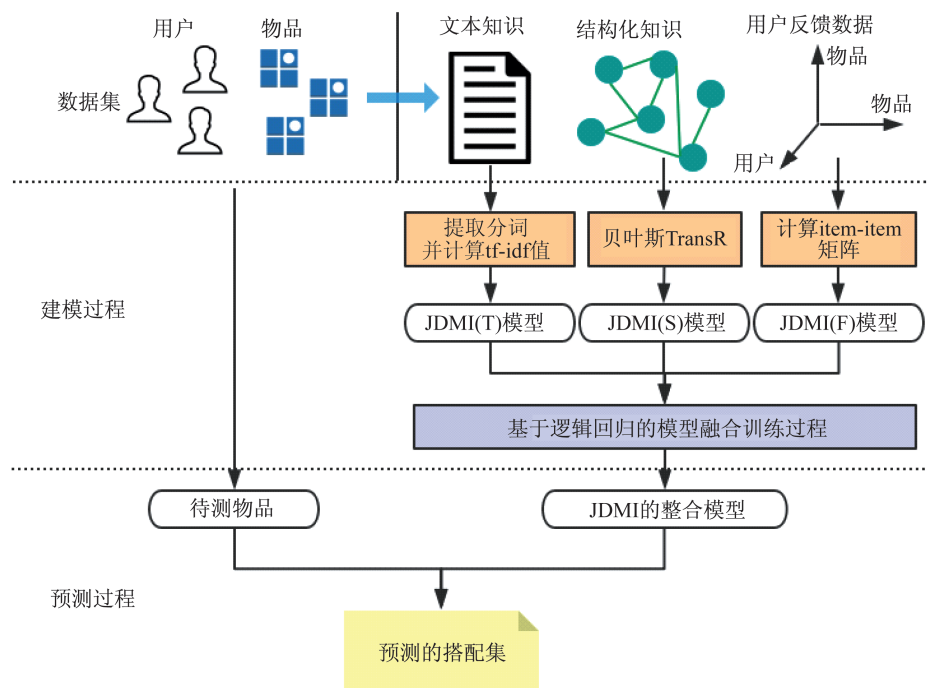


图1 基于联物品搭配度的推荐算法框架 JMDI 的基本流程

Fig.1 Basic process of the framework JMDI based on the combination of items' joint match degree

knowledge)的物品搭配度模型(JMDI(S))和基于用户交互反馈(user interaction feedback)的物品搭配度模型(JMDI(F)).

物品搭配度建模过程是从物品实体的文本知识、结构化知识以及用户交互反馈信息出发,对数据进行预处理后,分别构建物品搭配度模型.数据预处理阶段将分别对文本知识、结构化知识和用户交互反馈数据进行预处理.对文本知识首先用分词算法进行分词并映射到整数集上,然后计算每个词的tf-idf值,筛选出关键词的tf-idf向量;对结构化知识进行降维处理,将物品实体空间转化为物品间联系的空间,从而利用相关实体预测算法得到预测的相搭配商品及其损失函数;对用户交互反馈数据进行处理,生成物品对偏序关系矩阵.在此基础上,本文将分别构建出物品搭配度模型.

而搭配度的协同模型融合过程则是利用上一步的子搭配度模型,通过逻辑回归方法,学习融合参数,构建融合后的搭配模型.下面将详细介绍这个过程.

2.2 物品搭配度模型

为了构建物品搭配度模型,本文从以下3个角度分别提取特征表示,构建子搭配度模型.

假设在推荐系统中,有一个物品集 $I = \{i_1, i_2, \dots, i_N\}$,其中 i_j 表示单个物品,由物品编号、类别和代表词组(由分词算法在物品文本描述上处理而得)组成的元组 (δ_j, c_j, τ_j) 表示, I 包含了其他所有数据集中出现的所有物品;一个用户交互行为集 $B = \{b_1, b_2, \dots, b_M\}$,其中 b_k 表示单次交互行为,由用户编号、物品编号、交互行为类型和行为发生时间组成的元组 $(u_k, \delta_k, \phi_k, \eta_k)$ 描述,表示用户 u_k 在 η_k 时刻对物品 δ_k 产生了交互行为 ϕ_k ;专家搭配知识集 S ,其构成的集合为 $S = \{s_1, s_2, \dots, s_L\}$,其中, s_j 由套餐编号和搭配套餐物品列表组成的元组 $(Y_{j1}, Y_{j1}, \dots)$ 描述. B 和 S 中 δ 所代表的物品都属于物品集 I .本文算法根据上述信息建立模型,然后对于给定物品集 I 中的每一个物品,预测最有可能与其搭配的 $Top-K$ 个物品.因此构造一个目标函数 $D(i, j), i \neq j, j \in [1, N]$,表示待测物品 i 和候选物品 j 之间的搭配度,而构建的模型应该能使搭配度最高的 $Top-K$ 个物品更加符合实际情况,即能更好地预测专家搭配结果.

2.2.1 基于文本知识(模型1)

基于文本知识的搭配度模型(JMDI(T))根据物品本身的文本知识评估物品之间的搭配程度.

首先需要对文本降维处理,利用相关技术^[6]提取文本关键词,对文本知识中出现的词语计算tf-idf值,取值最大的 m 个分词作为该物品实体的词向量.具体而言,物品 p 的一个分词 ϵ_p 的tf-idf值为词频 $tf(\epsilon_p)$ 和逆文档频率 $idf(\epsilon_p)$ 的乘积,某个词的重要性越高,则tf-idf值就越大.其中,

$$tf(\epsilon_p) = \frac{r(\epsilon_p, \tau_p)}{l(\tau_p)} \quad (2)$$

$$idf(\epsilon_p) = \ln\left(\frac{d(c_p, \cdot)}{d(c_p, \epsilon_p) + 1}\right) \quad (3)$$

式中: $l(\tau_p)$ 表示物品实体 p 的分词集合 τ_p 的大小; $r(\epsilon_p, \tau_p)$ 表示 τ_p 中分词 ϵ_p 出现的次数; $d(c_p, \cdot)$ 表示与 p 类别相同的物品个数; $d(c_p, \epsilon_p)$ 表示与 p 类别相同且包含分词 ϵ_p 的物品个数.为了防止式(3)造成某些分词重要性偏高或偏低,因此需要对 $idf(\epsilon_p)$ 进行平滑操作 $H(\cdot)$.因此, $tfidf(\epsilon_p) = tf(\epsilon_p)H(idf(\epsilon_p))$,这里 $H(\cdot)$ 可为一阶线性平滑函数,即

$$H(idf(\epsilon_p)) = \alpha idf(\epsilon_p) + \beta \quad (4)$$

根据tf-idf值可以计算出每个物品实体的分词向量及其tf-idf向量,由此可以进一步构建物品之间的搭配度模型.

考虑到物品之间若要搭配,它们之间势必有一些共性元素,例如品牌、色调、季节等,这样才能协调而不至于搭配突兀,这种信息一般都会作为关键词出现在物品标题中,以使用户能够很容易地搜索到.由此可以认为,经过类别过滤的物品之间的搭配度与其相似度成正相关.因此,基于文本知识的搭配度可以由分词向量的相似度 $w(p, p')$ 表示,即

$$D_1(p, p') = w(p, p') \quad (5)$$

式中, $w(p, p')$ 为物品 p 和 p' 分词的tf-idf向量的余弦相似度^[7].

考虑到两个物品的标题分词长度和分词集合不一定相同,且物品 p 与 p' 的相似度应是相对于 p 而言的,因此将 $w(p, p')$ 优化为

$$D_1(p, p') = w(p, p') = \frac{\langle x, y \rangle}{\|x\|} = \frac{\sum_{\epsilon_p = \epsilon_{p'}} (tfidf(\epsilon_p) tfidf(\epsilon_{p'}))}{\sqrt{\sum tfidf^2(\epsilon_p)}} \quad (6)$$

另一方面,为了减少后续计算量,考虑到在搭配问题中,相似的同类物品一般都是可替代产品,而非搭配物品,而某些类别对一般不会出现在同一组搭配中(例如羽绒服和热裤).即相搭配的物品一般

都不属于同一个类别,且某些类别势必与另一些类别不相搭配.因此 $D_1(p, p')$ 可以定义为

$$D_1(p, p') = \begin{cases} \frac{\sum_{\epsilon_p = \epsilon_{p'}} (tfidf(\epsilon_p) tfidf(\epsilon_{p'}))}{\sqrt{\sum tfidf^2(\epsilon_p)}}, & p \text{ 与 } p' \text{ 类目搭配} \\ 0, & \text{否则} \end{cases} \quad (7)$$

2.2.2 基于结构化知识(模型 2)

复杂网络中通常隐含着大量实体及其相互关系的结构信息,一般可以用知识图谱来表示.知识图谱是一种图结构,可以由若干条 (v_h, r, v_t) 的三元组构成,表示头实体 v_h 与尾实体 v_t 之间存在关系 r , h, t 表示实体.这种表示方式存在着计算效率低和数据稀疏性问题.因此,本文试图将网络结构的知识嵌入到低维向量空间中,并尽量保留网络中的某些信息. TransR 模型^[8]是当前最为先进的复杂网络嵌入方法之一.基于结构化知识的搭配度模型(JMDI(S))将 TransR 改进并应用于搭配场景下.下面介绍将 TransR 模型应用到结构化知识上,生成关系 r 的向量空间的方法.

首先,用一个无向图 $G(V, E)$ 表示结构化知识,其中 $V = \{v_1, v_2, \dots, v_{|V|}\}$ 为图的点集,表示知识中不同物品实体的集合, E 为代表不同实体间各种关系的边集.

TransR 模型与其他类似模型不同, TransR 模型将实体和关系表示在不同语义空间中,用特定关系的投影矩阵联系起来.具体而言,如图 2 所示,将每个三元组 (v_h, r, v_t) 中的实体嵌入到向量 $v_h, v_t \in \mathbb{R}^k$ 中,关系 r 嵌入到向量 $r \in \mathbb{R}^d$ 中.对于每个关系 r ,定义投影矩阵 $M_r \in \mathbb{R}^{k \times d}$,将实体向量投影到关系 r 的子空间,实体的向量可以定义为

$$v_h^r = v_h M_r, v_t^r = v_t M_r \quad (8)$$

然后,通过不断调整 v_h, v_t, r 使得 $v_t \approx v_h + r$, 其损失函数定义为

$$f_r(v_h, v_t) = \|v_h^r + r - v_t^r\|_2^2 \quad (9)$$

有别于普通 TransR 模型基于边际的目标函数,本文采用 sigmoid 函数来计算成对三元组的排序概率,然后将 TransR 扩展为贝叶斯的版本,其算法步骤如下:

- 对每个物品实体 v , 作 $v \sim N(0, \lambda_v^{-1} \mathbf{I})$;
- 对每个关系 r , 作 $r \sim N(0, \lambda_r^{-1} \mathbf{I})$ 且 $M_r \sim N(0, \lambda_M^{-1} \mathbf{I})$;
- 对任意四元组 $(v_h, r, v_t, v_{t'}) \in \pi$, 计算 $\sigma(f_r(v_h, v_t) - f_r(v_h, v_{t'}))$. 其中 π 为满足 $(v_h,$

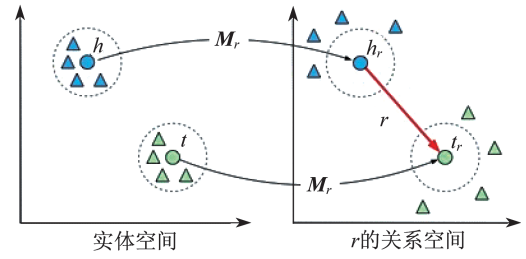


图 2 TransR 的简单示例

Fig.2 Simple example of TransR

$r, v_t)$ 是正确三元组、而 $(v_h, r, v_{t'})$ 为错误三元组的四元组的集合. $\sigma(x) = \frac{1}{1 + e^{-x}}$ 为 logistic sigmoid 函数.

从而可以训练得到每个物品的表示结构化含义的向量.然后,根据式(9)计算出每个待测物品与候选物品的损失函数,用以表示物品之间的搭配程度.由于其搭配度与损失函数值呈负相关,因此

$$D_2(p, p') = \rho - f_r(p, p') \quad (10)$$

式中, ρ 为参数,可以取为所有 f_r 的中位数.

2.2.3 基于用户交互反馈(模型 3)

在基于用户交互反馈的搭配度模型(JMDI(F))中,主要通过协同过滤的思想分析用户交互反馈,得到物品之间的搭配关系.如上文所述,用户交互反馈是关于用户-物品-时间的三维不连续空间 $\mathbf{R}$, 每个点表示一次交互行为代表的兴趣值,其值与行为类型相关.例如,通常购买行为的兴趣值大于点击行为.

通常的协同过滤算法很少会考虑历史行为中的时序信息,但同一时间窗内购买的物品之间很有可能就隐含着彼此的关系.因此,在某个时间窗 T_c 的约束条件下将 $\mathbf{R}$ 转换到一个低维矩阵 $\mathbf{Q} \in \mathbb{R}^{M \times N}$, 表示用户对物品的兴趣值的和,矩阵中的值 $Q_{i,j}$ 表示在 T_c 时间间隔内,用户 i 对物品 j 产生交互行为的总兴趣值.

对 $\mathbf{Q}$ 处理,将用户-物品矩阵转换为物品-物品的稀疏矩阵 $\mathbf{Z}$,使得

$$Z_{i,j} = \sum_{k=1}^{|M|} \varphi(|Q_{k,i} - Q_{k,j}|) \quad (11)$$

其中, $|Q_{k,i} - Q_{k,j}|$ 表示同一用户对物品实体 i, j 产生兴趣时,兴趣值的差异;当 $x > 0$ 时, $\varphi(x)$ 为一个随着 x 递增而递减的函数, $x = 0$ 时, $\varphi(x)$ 为 1.

然后对于待测物品 p 和搭配物品 p' ,构造搭配度 $D_3(p, p')$,在一个合理的时间窗 T_c 内,考虑到以下几种情况:

a. 若有同一用户对 p 与 p' 产生了交互行为, 则 p 与 p' 更有可能相搭配, 其搭配度与 p 和 p' 被同一用户产生交互行为的概率成正比, 即 $D_3(p, p')$ 与 $P_{cnt}(p, p')$ 成正比;

b. 若有同一用户对 p 与 p' 产生了交互行为, 则 p 与 p' 更有可能相搭配, 且兴趣值越接近, 其搭配度越高, 即 $D_3(p, p')$ 与 $Z_{p, p'}$ 成正比;

c. 若 p 与 p' 从未被同一用户发现过, 考虑到可能存在可替代物品的情况, 因此寻找物品 q 满足 $Z_{q, p} > 0, Z_{q, p'} > 0$, 且 q 与 p 同类, 与 p' 异类, 其搭配度和 p 与 q 的相似度 $w(p, q)$ 成正比, 与 $D_3(q, p')$ 成正比。

综合上述逻辑, 当有同一用户在时间窗 T_c 内对 p 与 p' 产生了交互行为时, 则

$$D_3(p, p') = P_{cnt}(p, p') Z_{p, p'} \quad (12)$$

否则

$$D_3(p, p') = D_3(q, p') w(p, q) \quad (13)$$

其中, $P_{cnt}(a, b) = cnt(a, b) / (cnt(a) + cnt(b))$, 表示 a, b 在时间窗 T_c 内被同一用户产生交互行为的概率。

2.3 搭配度融合排序方法

根据上文所述, 本文分别从文本知识、结构化知识和用户交互反馈信息中提取出了物品搭配度模型。本节中, 为了将上述 3 种物品搭配度模型整合起来, 在 JMDI 框架中引入了联合学习过程。

由上述 3 个子搭配度模型可以初步得到每个待测物品的候选搭配列表及其子搭配度, 然后需要对每个候选物品的搭配度进行联合评估排序, 从而选出最可能搭配的 $Top-K$ 个物品。

给定专家标记的搭配套餐, 可以得到物品之间偏序关系的正负反馈。具体来说, 若物品 i, j 搭配而 i, j' 不搭配, 则将 (i, j) 对标记为 +1, 将 (i, j') 对标记为 -1。随后利用逻辑回归模型^[9] (logic regression, LR) 训练子搭配度的融合参数。将子模型输出的候选物品编号及其与待测物品的子搭配度构成的向量分别作为训练融合模型的样本, 记为 $X = \{\delta_i, \mathbf{x}_i, i \in [1, 100] \cap \mathbb{Z}\}$, 其中 $\mathbf{x}_i = (D_{i1}, D_{i2}, D_{i3})$ 。接着利用专家搭配套餐获得样本的正负标记 $Y \in \{-1, 1\}$, 从而根据逻辑回归模型的原理, 假设

$$P(y = 1 | \mathbf{x}; \boldsymbol{\theta}) = g(\boldsymbol{\theta}^T \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}} \quad (14)$$

然后利用梯度上升方法对参数 $\boldsymbol{\theta}^T$ 求解, 由于式(14)是非线性的, 因此对其进行 logit 变换得到

$$g(x) = \log\left(\frac{P}{1-P}\right) = \theta_0 +$$

$$\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \quad (15)$$

$g(x)$ 即最终的搭配度 D , 对其进行排序, 通过对专家搭配套餐物品对的类目进行统计, 去掉从未出现过的类目搭配, 从而选出搭配度 $Top-K$ 的候选物品作为与物品 p 相搭配的物品列表。

3 算法描述与分析

根据上文描述, 本文的方法需要将数据集划分为训练集 Trainset 和测试集 TestItems。对于训练集, 由于模型融合过程有监督学习, 因此需要将训练集再划分为两部分, 分别用于生成子模型结果和训练模型融合参数。算法的主要流程如下:

a. 数据预处理, 计算每个物品文本信息的 tf, idf 值, 统计并计算 $P_{cnt}(p, p')$;

b. 根据 2.2.1 节的模型计算待测物品的相似度和物品搭配度 D_1 ;

c. 根据 2.2.2 节训练模型并获得待测物品的相搭配物品实体及其搭配度 D_2 ;

d. 根据 2.2.3 节计算时间窗 T_c 内的 Q 矩阵, 进一步处理得到矩阵 Z , 计算待测物品的物品搭配度 D_3 ;

e. 利用训练集对上述子搭配度的融合参数进行训练, 然后融合候选物品与待测物品的搭配度, 进行离差标准化处理, 最后取 $Top-K$ 个作为最终结果。

由于并非所有的用户都会有齐全的各类数据, 而本算法框架融合了物品本身文本知识、结构化知识和用户行为反馈, 使得它可以解决现实中经常遇到的数据缺失的问题, 一定程度上扩大了算法的适用面。

另一方面, 本文提出的 JMDI 框架模块化程度较高, 模块内的计算也可以很容易地并行化处理, 因此本算法也适用于大规模数据的处理。

4 实验结果及对比分析

4.1 实验数据集及评价方法

本文采用阿里天池大数据实验室提供的 Taobao_Clothes_Matching^[10] 数据集。该数据集包括 3 部分, 分别为物品基本信息数据(文本、图像)、用户历史行为数据和搭配套餐数据, 总共上万套餐, 十万级物品及图像, 百万级用户、千万级行为的数据。为了测试算法的性能, 将搭配套餐数据的 80% 作为训练集, 其余作为测试集, 并利用所有搭配数据生成

测试集的答案. 物品集 I 和专家搭配套餐集 S 、用户历史行为数据集 B , 以及待预测商品集 $TEST$ 中的物品关系如图 3 所示.

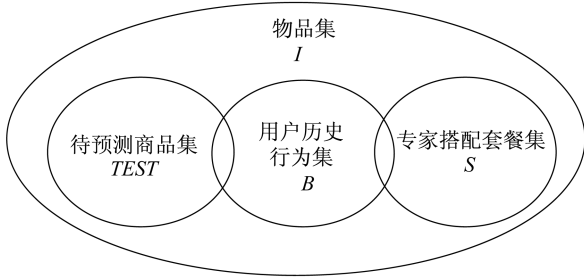


图 3 数据集中物品来源的关系

Fig.3 Schematic diagram of relationship between sources of items in data sets

为了评价算法的效果, 本文对于每个物品采用 $MAP@K$ ^[11] (mean average precision) 和 $Recall@K$ 作为评价指标, 其值趋于 0~1 之间, 数值越大越理想. 其中 $MAP@K$ 具体计算方法如下.

对每个物品, 其 $ap_i@K$ 表示为

$$ap_i@K = \frac{\sum_{k=1}^K \frac{1}{1 - \ln p(k)} \Delta(k)}{n} \quad (16)$$

式中: n 表示答案集中物品的数量; $p(k)$ 表示在 k 截断之前的预测准确率; 当第 k 个物品在答案集中 $\Delta(k)$ 为 1, 否则为 0. 对每个物品的 $ap_i@K$ 在待预测物品集合下求平均值得到最终评测值为

$$MAP@K = \frac{\sum_i ap_i@200}{|I|} \quad (17)$$

式中, $|I|$ 表示待预测物品集中物品个数.

这种评测指标可以反映出预测搭配的命中率和排序准确率. 因此, 需要本算法能够尽量减少搭配集物品数目而提高命中率, 同时搭配集的物品排序要尽可能贴近真实情况.

4.2 实验设计与结果分析

首先在数据集上进行抽样实验, 调整每个子模型的参数值, 观察对 $MAP@K$, $Recall@K$ 的影响, 设定评价指标中的 K 为 50, 100, 150, 200, 确定模型中各个参数如下:

- 在式(4)中, $\alpha = 0.0085$, $\beta = 0.07$;
- 模型 3 中, 时间窗 T_c 取 30 d.

然后在正式数据集上运行本文算法框架, 根据基于文本知识、结构化知识和用户交互反馈的搭配度模型, 分别计算得到搭配结果, 与其他算法比较评估其结果. 最后将完整框架的处理结果与其他算法

比较.

为了验证本文提出的物品搭配度模型的有效性, 设计对比实验, 将本文模型与下列实验的结果对比.

a. Item-based CF(T)^[12]: 将基于物品文本相似度的协同过滤算法应用于文本知识数据集, 相似度采用余弦相似度, 对训练集中标注的搭配物品对, 寻找与待测物品最相似的几个物品, 获得与它们相搭配的物品及其近似物品作为算法输出.

b. BPRMF^[13] + TransE^[14]: 将基于贝叶斯个性化排序的矩阵分解算法^[2]结合 TransE 算法^[3], 应用于结构化知识数据集, 生成候选搭配列表.

c. Item-based CF(F): 将基于物品文本相似度的协同过滤算法应用于用户交互反馈数据集, 根据用户同时感兴趣的物品对, 寻找与待测物品最相似的几个物品, 获得与它们相搭配的物品及其近似物品作为算法输出.

d. 混合 CF: 将基于相似度的协同思想应用于全部数据集上, 生成候选搭配列表.

图 4(a), 5(a) 表示 JMDI 中的基于文本知识的搭配度模型与 Item-based CF(T) 算法的效果比较; 图 4(b), 5(b) 表示 JMDI 中的基于结构化知识的搭配度模型与 BPRMF + TransE 算法的对比; 图 4(c), 5(c) 表示 JMDI 中的基于结构化知识的搭配度模型与混合 CF 算法的对比.

可以发现, 本文提出的 JMDI 框架中的每个子模型的表现都优于其对比算法. 而基于结构化知识的搭配度模型由于引入了 TransR 模型, 相较于其对比算法, 效果提升显著, 能更好地应对复杂关系网络, 充分挖掘网络知识中隐含的信息. 而比较 3 个子模型可以看出, 基于文本知识的子模型对于搭配预测的贡献最大, 但另外两个子模型的贡献也很重要. 这是由于物品的文本描述本身包含了物品的关键特征的描述, 而一般为了协调, 相搭配的物品大多都有类似的特征, 因此文本描述的信息准确地刻画了物品特征, 便可以更好地预测搭配物品; 而基于结构化知识和用户交互反馈的模型, 由于主要基于客观事实或行为, 其隐含特征的挖掘效果相对不如文本知识, 但它更具有可靠性, 在样本数据足够多的情况下表现稳定, 不会受人为设定文本描述信息的干扰.

由图 6 可以看出, 3 个子模型的使用显著提升了算法的整体效果, 证明本文的算法框架充分挖掘了各类数据中的隐含信息, 并且将其有效融合为一体. 与混合 CF 算法的对比显示, 本文将文本、结构

化知识与协同过滤结合到一起,有效提升了搭配预测的效果.此外,通过将3个模型结合起来,可以缓解协同过滤的数据稀疏性和冷启动问题,在没有用

户有相应交互行为的时候,可以通过利用物品相关知识,很好地弥补协同过滤的不足,使推荐系统的整体效果得到提升.

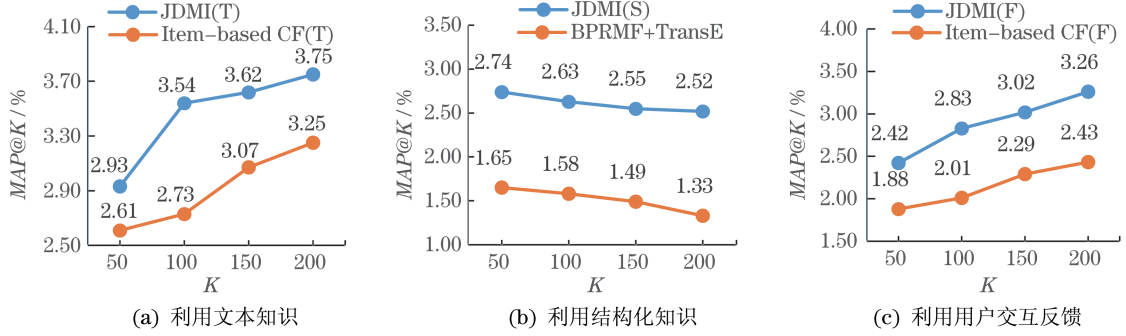


图4 JMDI 算法框架子模型与对比算法的 $MAP@K$ 比较结果

Fig.4 Comparison of contrast algorithm and each submodel in JMDI on $MAP@K$

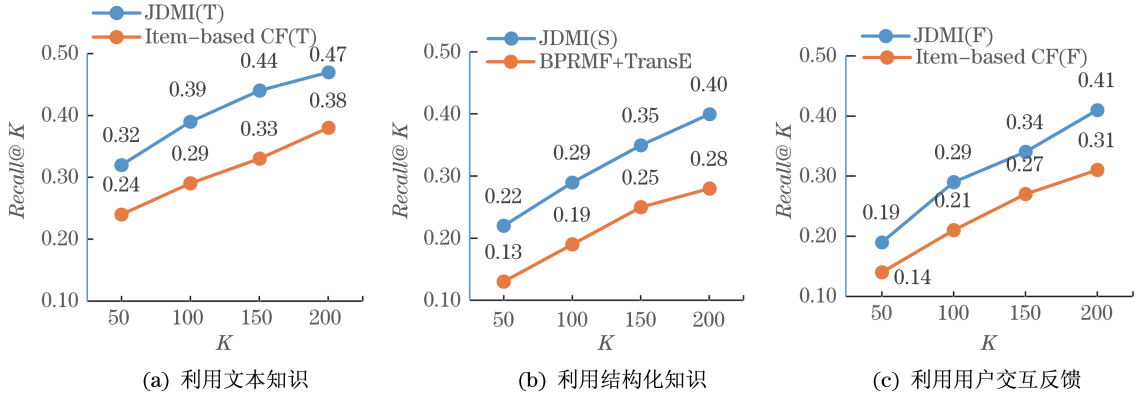


图5 JMDI 算法框架子模型与对比算法的 $Recall@K$ 比较结果

Fig.5 Comparison of contrast algorithm and each submodel in JMDI on $Recall@K$

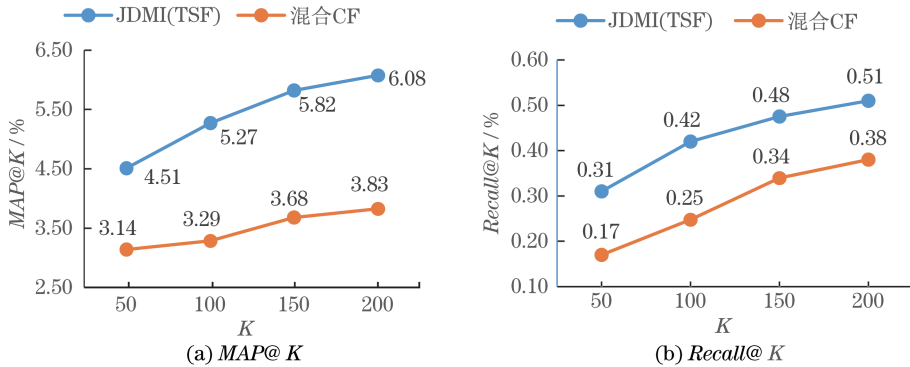


图6 JMDI 算法框架与对比算法的比较结果

Fig.6 Comparison results of contrast algorithm and framework JMDI

5 结论

本文针对传统的基于相似度度量的推荐算法无法表示物品之间搭配关系的问题,提出了一个基于联合物品搭配度的推荐算法框架 JMDI,整合

了文本知识、结构化知识和用户交互反馈等信息用于搭配推荐.框架中构建了3个物品搭配度子模型并通过逻辑回归方法进行搭配度融合,生成与目标物品相搭配的物品推荐列表.在 Taobao_Clothes_Matching 数据集上的实验表明,本文提出的基于联合物品搭配度(JMDI)的推荐算法框架可

以提供有效的搭配物品推荐. 由于该算法的效果还依赖于不同模型生成的结果的融合方法, 因此下一步工作是研究如何更好地将子模型的结果进行融合. 此外, 今后还将研究图像知识的处理, 将图像特征融入到算法框架中, 以期得到更好的结果.

参考文献:

- [1] LIU Q, CHEN E H, 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.
- [2] PAZZANI M J, BILLSUS D. Content-based recommendation systems [M] // BRUSILOVSKY P, KOBASA A, NEJDL W. The Adaptive Web. Berlin Heidelberg: Springer, 2007: 325 – 341.
- [3] BURKE R. Knowledge-based recommender systems [J]. Encyclopedia of Library and Information Systems, 2000, 69(32), 180 – 200.
- [4] HU Y, YI X, DAVIS L S. Collaborative fashion recommendation: a functional tensor factorization approach [C] // Proceedings of the 23rd ACM International Conference on Multimedia. Brisbane, Australia: ACM, 2015: 129 – 138.
- [5] WONG W K, ZENG X H, AU W M R, et al. A fashion mix-and-match expert system for fashion retailers using fuzzy screening approach [J]. Expert Systems with Applications, 2009, 36(2): 1750 – 1764.
- [6] ZHOU L. Exploration of the working principle and application of word2vec [J]. Sci-Tech Information Development & Economy, 2015, 25(2): 145 – 148.
- [7] YE J. Cosine similarity measures for intuitionistic fuzzy sets and their applications [J]. Mathematical and Computer Modelling, 2011, 53(1/2): 91 – 97.
- [8] LIN Y K, LIU Z Y, SUN M S, et al. Learning entity and relation embeddings for knowledge graph completion [C] // Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas: AAAI, 2015: 2181 – 2187.
- [9] RUCZINSKI I, KOOPERBERG C, LEBLANC M. Logic regression [J]. Journal of Computational and Graphical Statistics, 2003, 12(3): 475 – 511.
- [10] Taobao_clothes_matching [EB/OL]. [2015]. <https://tianchi.shuju.aliyun.com/datalab/dataSet.htm?spm=5176.100073.888.29.DatQOr&id=13>.
- [11] VAN DEN OORD A, DIELEMAN S, SCHRAUWEN B. Deep content-based music recommendation [C] // Advances in Neural Information Processing Systems 26. South Lake Tahoe, NV, USA: MIT Press, 2013: 2643 – 2651.
- [12] SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms [C] // Proceedings of the 10th International Conference on World Wide Web. Hong Kong, China: ACM, 2001: 285 – 295.
- [13] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: Bayesian personalized ranking from implicit feedback [C] // Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Montreal, Quebec, Canada: AUAI Press, 2009: 452 – 461.
- [14] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data [C] // Advances in Neural Information Processing Systems 26. South Lake Tahoe, United States: MIT Press, 2013: 2787 – 2795.

(编辑: 丁红艺)