面向图书主题的爬虫算法研究

张莉婧 曾庆涛 李业丽 孙华艳 字云飞

(北京印刷学院信息科学技术学院 北京 102600)

摘 要 针对图书信息爬取结果中包含大量无用数据的问题,提出一种面向图书主题的爬虫算法。该算法主要由两部分组成:一部分是基于开放式分类目录系统(ODP)的动态关键词扩充的主题描述方法;另一部分是基于词项语义扩展度的向量空间模型(VSM)主题相关度算法。通过实验对新算法、基于关键词的VSM算法以及基于ODP的VSM算法进行了对比分析,结果表明新算法在图书主题爬虫中更具有优势。

关键词 主题爬虫,开放式分类目录系统,向量空间模型,语义扩展度

中图法分类号 TP302.1 **文献标识码** A

Research on Crawler Algorithm for Theme of Books

ZHANG Li-jing ZENG Qing-tao LI Ye-li SUN Hua-yan ZI Yun-fei

(School of Information Science and Technology, Beijing Institute of Graphic Communication, Beijing 102600, China)

Abstract Aiming at the problem that the information crawling result of a book contains a lot of useless data, a kind of crawler algorithm was proposed, which is based on the book topic. The algorithm mainly consists of two parts, one part is based on the ODP (Open Directory System) dynamic keyword expansion method to describe the subject, the other part is the semantic extension of lexical entry based on VSM (Vector Space Model) topic correlation algorithm. The new algorithm, the VSM algorithm based on keywords and VSM algorithm based on ODP were analyzed through experiment. The result indicates that the precision and the recall rate of the new algorithm are higher than that of other two algorithms.

Keywords Focused crawler, ODP, VSM, Semantic extension

1 引言

随着互联网的快速发展,图书信息资源正呈指数增长,庞大的图书信息资源已远远超出通用搜索引擎所能覆盖的范围,因此通用搜索引擎很难满足读者对固定领域图书检索的需求。面对该问题,产生了一种新的搜索引擎即垂直搜索引擎。垂直搜索引擎通过对网页库中的信息进行整合,将行业内所需信息分离出来并以一定的格式传送给用户,成为特定领域内的专业搜索引擎。主题爬虫是垂直搜索引擎中的关键部分,直接影响用户对搜索网页的满意程度。为此,主题爬虫成为当今的研究热点问题。主题爬虫是对通用爬虫的改进与提高,即在通用爬虫的基础上通过定制爬虫策略和相关算法尽可能多地过滤掉与主题不相关的网页,从而获得与主题相关的网页,最大限度地满足用户查询需求。

2 问题描述

主题爬虫就是根据给定主题爬取该主题范围内网页信息的过程。首先,根据给定主题对主题进行详细描述,构建主题特征词向量;其次,抓取网页的特征词向量;最后,计算网页特征词向量与主题特征词向量的相关性,即主题相关度计算。主题描述的准确性将直接影响主题相关度计算的结果,主题相关度的计算结果将直接影响主题网页获取的准确率,进而影响主题爬虫的性能。

2.1 主题描述方法

主题的描述方法[1]主要分为3种:基于关键词的主题描述、基于分类结构样本的主题描述和基于语义的主题描述。在某一特定主题领域的应用中,基于关键词的主题表示方法要求计算复杂度相对较低,具有一定的优势。但它假设每个关键词都是相互独立的,忽略了词语之间的关系。

为解决基于关键词主题描述的弊端,学者们提出了很多方法。文献[2]提出了一种新的主题描述方法,通过ODP分类树和主题描述文档共同创建主题关键词集合,虽然该方法考虑了主题关键词之间的相关性,细化了主题描述范围,但该方法也增加了爬虫的代价,需要爬取主题描述文档来建立主题向量。文献[3]通过将ODP分类中的不同位置节点赋予不同的权重来描述主题,该方法在一定程度上提高了主题爬虫

本文受北京市科技创新服务能力协同创新项目(PXM2016_014223_000025)资助。

张莉婧(1992-),女,硕士生,主要研究方向为出版方向数据分析,E-mail:2286913877@qq.com;**曾庆涛**(1982-),男,博士,讲师,主要研究方向为数字出版、大数据技术,E-mail:zengqingtao@bigc.edu.cn;**李业丽**(1962-),女,教授,主要研究方向为数据挖掘、信息处理技术、喷墨印刷控制,E-mail:liyl@bigc.edu.cn(通信作者);**孙华艳**(1990-),女,硕士生,主要研究方向为大数据处理;**字云飞**(1991-),男,硕士生,主要研究方向为大数据处理。

的性能，但由于节点的权重是人为给定的，因此带有一定的主观色彩。文献[4]提出一种基于关键词动态扩充的主题关键词描述方法，在初始状态下，需要通过人工定制与自动抽取相结合的方式构建一个初始关键词集合。如果主题范围太大，人工定制的代价就会很大。

本文将 ODP 与关键词动态扩充相结合，充分利用两者的优点，提出一种 ODPE 主题描述方法。

2.2　主题相关度计算

为了控制抓取的网页与主题是相关的，常见的解决思路有 4 种[5]，即行业搜索、链接描述方法、网页链接评分方法及基于内容的相关度判定。4 种方法各有利弊。其中基于内容的相关度判定在图书主题爬虫中略胜一筹。VSM 是基于内容相关度判定中常用的方法。VSM 首先将文本的内容转化为简单易于理解的向量[6]，然后通过计算向量间的相似度来表达文本的语义相似度。目前，该模型广泛应用于主题爬虫中的两个文本之间的相关度计算中。VSM 在提高文本相似度的同时表现出了弱点，即没有涉及文字的语义，提取特征项时将特征项一视同仁，权重的计算采用了最基本的 TF-IDF 算法。

为了解决 VSM 中存在的问题，研究者们提出了各种改进措施。文献[7]通过引入领域本体知识来增加向量空间模型中的语义信息，提高了页面相关性的判定精度；但该方法中本体的构建需要领域专家来完成，实现比较困难。文献[8]应用词共现模型将共现度较高的词组添加到特征词空间中，在增加语义的同时增加了计算量。文献[9]采用基于语义的方法先对特征集合进行聚类分析，将达到某一特定阈值的几项合并成一项并代替原有项，从而降低了矩阵向量的维度。

在以上方案的基础上，对 VSM 模型进行了进一步的优化，提出语义扩展度的概念，即将语义相关度与语义相似度相结合来共同决定网页的取舍。

3　数学模型

3.1　基于 ODP 主题初始关键词集合的构建[10]

ODP 是由来自全球的 91826 名志愿编辑者共同维护管理的一种非商业性和非盈利性的主题目录，它为学者和商人提供了最具权威性的主题目录；现在已经扩展为 103 万多个类目，成为最大的目录系统；共有 90 种语言版本，应用于全球范围内。ODP 专案为互联网上最大、最普遍的搜索引擎和门户网站提供主要的目录服务，包括 Netscape，AOL，Google，Lycos，HotBot，DirectHit 等在内的成百上千个网站。

基于 ODP 主题初始关键词集合的构建的定义如下。

定义 1（主题概念）　将给定的主题词映射到 ODP 分类树上的节点称为主题概念。

定义 2（主题子树）　以主题节点为根的子树。

定义 3（主题相关概念）　主题概念节点的祖先或子孙节点构成主题相关概念，从而由主题相关概念集合构建主题初始关键词集合。

基于 ODP 主题初始关键词集合构建的步骤如下。

（1）确定主题概念 T。将给定的主题关键词映射到 ODP 分类树上，并确定其对应的主题概念。

（2）确定主题初始关键词集合 $TKCS$。因为 ODP 中英文版本的分类树相对中文分类树覆盖面更全，所以选择英文版本分类树。主题概念的父节点相对主题概念表达的范围更广，同时为了降低计算的复杂度，选择分类树中向下四代子孙节点（C_{down}）作为概念集合的一部分。因此，主题初始关键词集合由主题概念和其子孙节点构成，即 $TKCS=C_{down}\cup T=\{W_1,W_2,\cdots,W_n\}$。

（3）确定主题初始关键词的权重 Q。将主题初始关键词权重表示为 $Q=\{q_1,q_2,\cdots,q_n\}$。

$$q_i=\frac{N_i}{N_{\max}} \tag{1}$$

其中，N_i 为第 i 个关键词在整个集合中出现的频次，$N_{\max}$ 为所有关键词中出现的最高频次。最后按 q_i 的大小进行排序，由于样本有限，只将一定阈值范围内的词语保存成初始关键词集合。

3.2　关键词的动态扩充[11]

网页内容经过预处理之后得到文本词项集合，在文本词项集合中获得能够准确表达网页主题的特征向量即称为网页的特征提取。实际上，网页的特征提取等价于关键词的抽取。目前关键词抽取的算法主要由有监督关键词抽取和无监督关键词抽取两类组成，这两种方法各有利弊。本文采用文献[12]提出的关键词抽取算法，其计算公式如下：

$$S(v_i)=(1-d)w(v_i)+d*w(v_i)*\sum_{v_i\in In(v_i)}\frac{w_{ji}}{\sum_{v_k\in Out(v_j)}w_{jk}}S(v_j) \tag{2}$$

该抽取算法将两种方法相结合并加入了 G1 赋权法对权重进行优化，在一定程度上提高了关键词的抽取效率。最后，将抽取的关键词作为网页特征向量，并表示为 $K=\{k_1,k_2,\cdots,k_n\}$。

主题库的构建是主题爬虫中最基本且最重要的环节。常见的主题库分为两类：固定的主题库和动态的主题库。其中固定的主题库需要高质量的定义，一旦定义不准确就会出现严重的隧道问题。动态主题库是在爬取网页的过程中通过引入某种算法不断扩大主题库。动态主题库的构建对初始关键词要求相对较低，随着主题库的不断扩大，主题表达的范围不断扩大，抓取主题网页的准确率也逐渐提高。

已知主题向量为 Q，页面向量为 K，两者的语义扩展度为 exp，如果 exp 大于阈值 u，就将页面向量 P 中的前 m 个项 $\{W_1,W_2,\cdots,W_m\}$ 保存到缓冲区 m-set 集合中；否则直接删除。在爬行一定数量的网页之后，遍历 m-set 中的每项，删除重复项，最后将无重复项加入关键词集合中，当总关键词项达到最大数目 χ 时，动态关键词扩充结束。

3.3　基于词项的语义扩展度计算

Word2vec[13] 是神经网络概率语言模型的实现，它是一种将人工神经网络与概率模型相结合的新型算法，其在自然语言处理领域得到了广泛的应用。它将文本集作为输入，通过训练模型生成每个词对应的词向量，并将这些词向量作为词的特征来计算两个词的相似度。其采用一个三层的神经网络：输入层-隐层-输出层。根据词频使用 Huffman 编码，使所有词频相似的词隐藏层激活的内容基本一致，出现频率越高

的词语，激活的隐藏层数目越少，从而有效地降低了计算的复杂度。Word2vec 利用了词的上下文，语义信息更加丰富，计算词语间的相似度相对更加准确，因此本文选择 Word2vec 方法计算词语间的相似度[14]。

假设通过 Word2vec 训练得到两个词的词向量分别为 $X=\{x_1,x_2,x_3,\cdots,x_n\}$，$Y=\{y_1,y_2,y_3,\cdots,y_n\}$，则词语间的相似度表示为：

$$sem(x_i,y_i)=\frac{\sum_{i=1}^{n}(x_i*y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}*\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{3}$$

语义相似性与语义的相关性存在一定的依存关系，语义相似度是语义相关度的一种特例。如果语义相关，则语义必然相似；但如果语义相似，无法判断语义是否相关。因此为了更加准确地计算两个词语之间的关系，提出了语义相关度的计算[15]。哈尔滨工业大学信息检索实验室在《同义词词林》的基础上，更新完成了一部具有汉语大词表的《哈工大信息检索研究室同义词词林扩展版》。其词义词典主要由 5 层结构构成，其中第 1 层用大写字母表示；第 2—5 层分别由小写英文字母、二位十进制整数、大写英文字母、二位十进制整数表示。按由左到右的顺序进行编码，编码排位越靠右表明词语间越相似。其中第 8 位标记有“=(同义)”“#(相关)”“@(独立)”3 种不同表示形式。因此可以根据词项在《同义词词林》语义词典中的编码来判断语义的相关性。如果 $code_1=code_2$ 且第 8 位为“=”，则相关度为 0.85；如果 $code_1=code_2$ 且第 8 位为“#”，则相关度为 1。

因此词语之间的相关度计算公式为：

$$rel(c_1,c_2)\begin{cases}1, & \text{if } code_1=code_2 \text{ and } F_1=F_2=\text{‘\#’}\\ 0.85, & \text{if } code_1=code_2 \text{ and } F_1=F_2=\text{‘=’}\\ 0, & \text{else}\end{cases} \tag{4}$$

其中，F_1，F_2 分别为词项 c_1，c_2 在《同义词词林》语义词典中编码的第 8 位符号。

将语义相似度与语义相关相结合定义为语义的扩展度，语义扩展度能更好地表示词项之间的关系。语义扩展度的计算公式如下：

$$exp(c_1,c_2)=\lambda\times sem(c_1,c_2)+(1-\lambda)\times rel(c_1,c_2) \tag{5}$$

其中，λ 为调节参数，本实验中取 0.5。

4 图书主题爬虫的设计

图书主题爬虫区别于普通主题爬虫[16]，因为它不存在链入其他网站的问题，只是单纯地在自家数据库系统中进行信息的查找，所以它对主题搜索策略的要求较低，对主题描述和主题相关度计算的要求较高[17]。图书主题爬虫系统的设计可以分为主题描述模块和主题相关度计算两大模块。首先通过 ODP 构建主题初始关键词集合，并将主题核心词链接作为初始 URL，得到相关商品链接列表，通过下载链接列表，得到网页内容；进而对网页内容进行预处理并通过特征提取得到网页的特征向量；最后，将主题特征向量与网页特征向量进行扩展度的计算，如果满足阈值条件则将前 m 个关键词保存到缓冲区，同时保存链接，否则，忽略该链接。其总流程如图 1 所示。

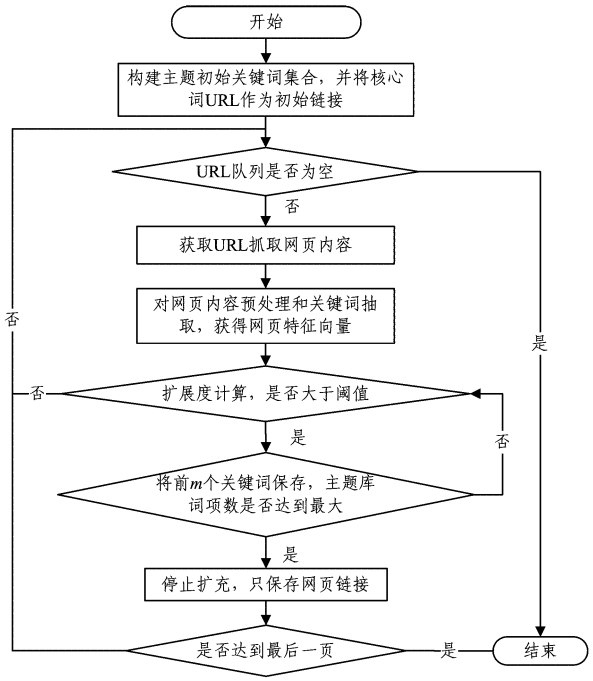

图 1 图书主题爬虫流程图

图书主题爬虫的具体步骤如下：

(1)根据主题，借助 ODP 构建主题初始关键词集合 TKCS，并根据式(1)计算集合 TKCS 中词项的权重，最后得到初始关键词集合的权重向量 $Q=\{q_1,q_2,\cdots,q_n\}$。

(2)将初始关键词集合中权重最大的词项构成链接，作为搜索引擎中的初始 URL，爬取图书列表页，每页均会出现图书链接列表。

(3)下载图书链接，获得图书详情页，爬取图书详情页中的图书标题和图书简介内容，将爬取的内容进行分词、词性标注、停用词过滤等预处理步骤，最后通过式(2)进行网页特征向量的抽取，将网页特征向量表示为 $K=\{k_1,k_2,\cdots,k_n\}$。

(4)根据式(3)计算 $Q=\{q_1,q_2,\cdots,q_n\}$ 与 $K=\{k_1,k_2,\cdots,k_n\}$ 的语义相似度 sem，根据式(4)计算两者的相关度 rel，最后根据式(5)计算语义扩展度 exp。

(5)判断扩展度是否满足阈值 u，如果满足，则保存网页链接，并将 K 中的前 m 个关键词放入缓冲区中；否则，忽略该网页链接。

(6)遍历缓冲区词项，删除重复词项，最终将词项放入向量 Q 中，随即更新向量 Q。

(7)判断 Q 中关键词词项是否达到最大值 χ，如果达到最大值，则关键词动态扩充结束；否则，继续扩充。

5 实验分析

5.1 实验数据

京东图书商城是国内较有影响力的图书平台，涵盖较大的数据库系统，因此选择京东图书商城作为爬虫平台。

5.2 实验环境

本文提出的图书主题爬虫采用 python 编程实现，使用实验室的 PC 机，内存为 2G，操作系统为 Microsoft Windows 7 系统。

5.3　实验数据介绍

因为图书的种类非常多，当想找一本书时很难准确地定位该书属于哪一类，所以我们忽略图书的种类。以“数据挖掘”为主题核心词，即将数字出版作为搜索关键词在搜索框内输入，得到初始 url“https://search.jd.com/Search?keyword=%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98&enc=utf-8&wq=shujuwajue”，从而根据设定的各种算法，以图 2 的形式抓取信息。

出版社	标题	内容简介
人民邮电出版社	数据挖掘导论 完整版 intro	《数据挖掘导论（完整版）》全面介绍了数据挖掘，涵
机械工业出版社	数据挖掘与数据化运营实战	阿里巴巴资深数据分析专家卢辉掷写，多年数据挖掘应用
机械工业出版社	大数据挖掘：系统方法与实	本书是大数据挖掘领域的经典之作，由全球科学计算领域的
机械工业出版社	数据挖掘：实用机器学习工	weka系统的主要开发者书写的研究、商业应用和教学
电子工业出版社	数据挖掘：你必须知道的37	本书是为广大数据分析师量身定制的入门读物，它旨在帮助
人民邮电出版社	程序员的数学1+2+3 数学原	程序员的数学三书包括《程序员的数学》《程序员的数学2：

图 2　爬取信息格式

可以根据内容简介来判断每本书是否与主题相关，以验证每种算法的有效性。

5.4　实验结果

按照上述方式，通过传统的爬虫共爬取 899 本书，而通过本文提出的 ODP2EVSM 算法共爬取了 501 本书，通过传统的基于关键词的 VSM 算法共爬取了 625 本书，通过基于 ODP 的 VSM 算法共爬取了 575 本书。借鉴文献[18]提出的查准率、召回率及 *F-measure* 对 3 种主题爬虫效果进行比较。查准率公式为：

$$Precision=\frac{relevant\_books}{downloaded\_books} \tag{6}$$

其中，*relevant_books* 表示爬取的与主题相关的图书数量，*downloaded_books* 表示爬取的总图书数量。

召回率公式为：

$$Recall=\frac{relevant\_books}{web\_relevant\_books} \tag{7}$$

其中，*relevant_books* 表示与主题相关的图书数量，*web_relevant_books* 表示整个网络中与主题相关的图书数量。

准确率与召回率的综合评价指标为 *F-measure*，其计算公式如下：

$$F\text{-}measure=\frac{2\times P\times R}{P+R} \tag{8}$$

本实验将 3 种算法爬取的前 500 本书作为数据源进行 3 个指标的计算。其中 500 本书就是爬取的 500 个网页，从第 50 个网页开始，以步长 50 增加，分别计算 3 个指标的数值，并将计算结果绘制成图表形式。图 3 给出了 3 种算法的查准率结果对比图。

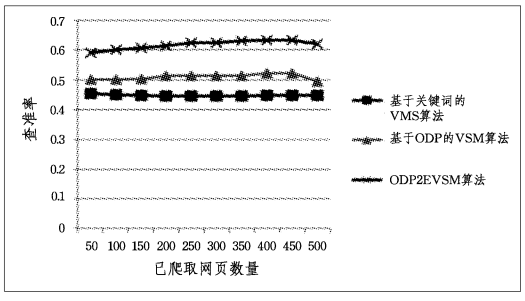

图 3　3 种算法查准率结果对比图

图 4 给出 3 种算法的召回率结果对比图。

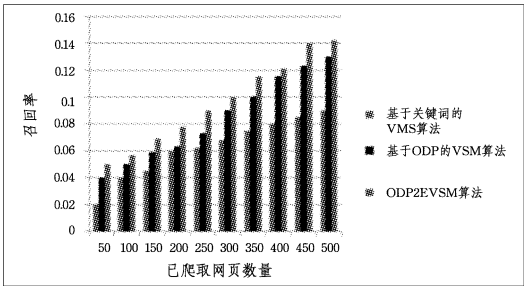

图 4　3 种算法的召回率结果对比图

通过以上数据可获得 *F-measure* 的值，如图 5 所示。

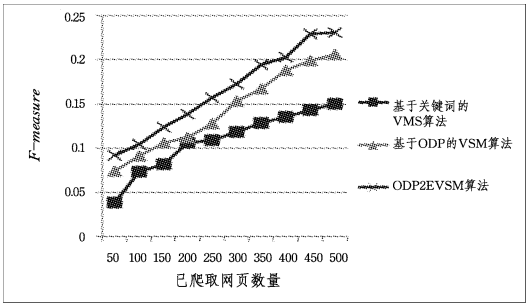

图 5　3 种算法的 *F-measure* 结果对比图

通过对以上 3 个指标的比较可知，本文提出的主题爬虫算法明显优于基于关键词的 VSM 算法和基于 ODP 的 VSM 算法。

结束语　针对图书主题爬虫设计了一种新的主题爬虫算法，即 ODP2EVSM 算法。首先，借助 ODP 进行关键词动态扩充来细化主题描述；然后，通过计算主题特征向量与网页特征向量的扩展度判断网页与主题是否相关。该算法在提高主题图书的获取数量的基础上增加了计算量，有待进一步改善。

参 考 文 献

[1] 王聪睿. 主题爬虫关键技术研究[D]. 石家庄：石家庄铁道大学，2015.

[2] 王良伟. 面向垂直搜索引擎的主题爬虫方法研究[D]. 重庆：重庆大学，2013.

[3] 邱伟林. 面向领域的垂直搜索引擎的研究与实现[D]. 大连：大连海事大学，2011.

[4] 刘建明. 垂直搜索引擎中的主题爬虫技术研究[D]. 广州：广东工业大学，2013.

[5] 杜娟娟. 主题爬虫算法的研究与实现[D]. 兰州：兰州交通大学，2013.

[6] 罗路天. 垂直搜索引擎中主题网络爬虫算法的设计与研究[D]. 广州：广东工业大学，2016.

[7] LIU W J，DU Y J. An Improved Topic-specific Crawling Approach Based on Semantic Similarity Vector Space Model[J]. Journal of Computational Information System，2012，8(20)：8605-8612.

[8] 张燕平，刘超，曲永花. WCBVSM 与 SACA 结合的文本分类模型[J]. 计算机工程与应用，2012，48(11)：137-142.

[9] 苏喻，郑诚，马忠杰. 基于语义的 VSM 模型改进[J]. 计算机应用与软件，2011，8(28)：158-161.

[10] 吴麒，陈兴蜀，朱锴，等. 基于 ODP 的上下文主题描述方法[J]. 电子学报，2012，11(40)：2320-2323.

（下转第 469 页）

仍然能够取得最优的推荐性能。

表 4　几种经典推荐算法的性能对比/%

算法	准确率@10	召回率@10	准确率@20	召回率@20	准确率@30	召回率@30
ItemCF	15.29	17.14	9.71	21.76	7.32	24.61
BPR	15.64	17.53	10.79	24.18	8.40	28.27
WRMF	15.50	17.38	10.74	24.08	8.42	28.32
TSPR-0(不重复)	16.19	18.14	10.93	24.50	8.59	28.89
SMART-2(不重复)	16.68	18.69	11.24	25.20	8.82	29.64
TSPR-0(可重复)	30.50	34.19	21.55	48.31	16.52	55.54
SMART-2(可重复)	31.82	35.67	22.23	49.82	17.01	57.20

综上分析,本文提出的 SMART 算法确实能够捕捉到电商平台用户对快速消费商品的购买兴趣和购买习惯,从而在一定程度上提升了推荐性能,特别是针对用户重复购买商品的场景,准确率和召回率都达到了较高的水平。

结束语　本文针对电商平台上用户对快速消费品的购买习惯和购买兴趣,提出一个基于"用户-商品-种类"三部图的推荐算法,通过将用户对商品和商品种类的偏好融入节点间的转移概率中,实现了带有兴趣倾向性的随机游走,使得稳态时的游走概率能够在一定程度上体现用户的购买习惯和兴趣,从而提升算法的推荐性能;此外,文中还针对用户对商家的兴趣偏好,提出一个游走概率调整策略,使商品推荐更倾向于该用户感兴趣的商家。本文利用真实的京东生鲜类商品的评论数据集对所提的推荐算法进行了性能评价,结果显示 SMART 算法确实能够捕捉到电商平台用户对快速消费商品的购买兴趣和购买习惯,特别是针对用户重复购买商品的场景,准确率和召回率都达到了较高的水平。但是电商平台上的商品推荐还应该考虑用户对于重复购买同类商品的时间间隔,这些工作都将在后续研究中完成。另外,本文提出的图推荐算法不仅适用于快速消费品的推荐,同样适用于餐馆推荐或者菜品推荐,因此后续研究的另一个工作是将算法推广到餐馆、菜品推荐等其他场景中。

参 考 文 献

[1] BELLOGIN A, PARAPAR J. Using graph partitioning techniques for neighbor selection in user-based collaborative filtering [C] // Proceedings of the Sixth ACM Conference on Recommender Systems. Dublin, Ireland, 2012: 213-216.

[2] LINDEN G, SMITH B, YORK J. Amazon. com Recommendations: Item-to-Item Collaborative Filtering [J]. IEEE Internet-Computing, 2003, 7(1): 76-80.

[3] MORADI P, AHMADIAN S, AKHLAGHIAN F. An effective trust-based recommendation method using a novel graph clustering algorithm[J]. Physica A: Statistical Mechanics and its Applications, 2015, 436: 462-481.

[4] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009(8): 30-37.

[5] SU X, KHOSGOFTAAR T M. Collaborative filtering for multi-class data using belief nets algorithms [C]// 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI). 2006.

[6] HAVELIWALA T H. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 784-796.

[7] XIANG L, YUAN Q, ZHAO S, et al. Temporal recommendation on graphs via long- and short-term preference fusion [C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 723-732.

[8] LEE S, PARK S, KAHNG M, et al. PathRank: Ranking nodes on a heterogeneous graph for flexible hybrid recommender systems [J]. Expert Systems with Applications, 2013, 40(2): 684-697.

[9] CHEN B S, WANG J D, HUANG Q H, et al. Personalized Video Recommendation Through Tripartite Graph Propagation[C]// Proceedings of the 20th ACM International Conference on Multimedia. Nara, Japan, 2012: 1133-1136.

[10] 刘梦娟,王巍,李杨曦,等. AttentionRank$^+$:一种基于关注关系与多用户行为的图推荐算法[J]. 计算机学报, 2017, 40(3): 634-648.

[11] LibRec 工具箱[OL]. http://www.librec.net.

[12] DIAZ-AVILES E, DRUMOND L, SCHMIDT-THIEME L, et al. Real-time top-n recommendation in social streams[C]// ACM Conference on Recommender Systems. ACM, 2012: 59-66.

[13] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: Bayesian personalized ranking from implicit feedback [C]// Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. Chicago, USA, 2009: 452-461.

(上接第 463 页)

[11] 刘燕兵,谭建龙,郭莉. 可动态增删关键词的串匹配算法[J]. 计算机工程与应用, 2005, 41(35): 138-141.

[12] 张莉婧,李业丽,曾庆涛,等. 基于改进 TextRank 的关键词抽取算法[J]. 北京印刷学院学报, 2016, 24(4): 51-55.

[13] ZHU L, WANG G J, ZOU X C. A Study of Chinese Document Representation and Classification with Word2vec[C]// 9th International Symposium on Computational Intelligence and Design(ISCID). 2016: 298-302.

[14] LIU C H, LIU Q, LEE C H. Valence-arousal ratings prediction of Chinese words using similarity measures based on Word2Vec [C]// International Conference on Asian Language Processing (IALP). 2016: 317-319.

[15] 游博. 词语语义相关度计算研究[D]. 武汉:华中师范大学, 2013.

[16] 李璐,张国印,等. 基于 SVM 的主题爬虫技术研究[J]. 计算机科学, 2015, 42(2): 118-122.

[17] 蒋华荣,郁雪. 应用遗传算法优化子空间的 SVM 分类算法分类算法[J]. 计算机科学, 2013, 40(11): 255-260, 275.

[18] 王良芳. 文本挖掘关键词提取算法的研究[D]. 杭州:浙江工业大学, 2013.