

文章编号:1000-5641(2013)03-0046-08

电子商务中的商品推荐系统

余文喆, 张蓉, 王立

(华东师范大学 软件学院, 上海 200062)

摘要: 对于迅速崛起的各种电子商务网站来说, 为了促进网站发展和增加经济效益, 吸引新客户并留住老客户是一种有效的手段. 设计和实现高效的推荐算法是各大网站最为关注的技术之一. 在电子商务网站中常见的一种推荐方式是以广告的形式在边栏推荐商品. 目前, 商品推荐系统根据推荐算法分为基于内容、协同过滤和混合的推荐系统. 然而, 现有推荐算法在电子商务网站的实际应用中正面临挑战, 包括推荐结果的多样化、个性化和智能化以及时效化. 现有算法需要不断改进来解决这些问题, 从而完善电子商务推荐系统.

关键词: 电子商务; 推荐系统; 多样化; 个性化; 智能化; 时效化

中图分类号: TP31 **文献标识码:** A **DOI:**10.3969/j.issn.1000-5641.2013.03.005

Recommendation in E-commerce

YU Wen-zhe, ZHANG Rong, WANG Li

(Software Engineering Institute, East China Normal University, Shanghai 200062, China)

Abstract: For e-commerce sites, in order to promote the development and win more benefits, attracting and keeping the customers becomes very important. One of the most useful technologies is recommendation algorithms. In e-commerce sites, sidebar advertising is a common form of recommendation, which can be divided into three main categories: content-based, collaborative filtering and hybrid recommendation algorithms. However, current recommendation algorithms are challenged by new application requirements, such as diversification, personalization, intelligentization and timeliness. It is urgent to design new algorithms to meet these requirements.

Key words: e-commerce; recommendation systems; diversification; personalization; intelligentization; timeliness

0 引言

近年来,随着电子商务的迅速发展,大量网络购物网站不断涌现,选择网上购物的用户数量日益增长,商家开始在电子商务网站投放广告. 在网络购物网站中,常见的广告表现形

收稿日期:2013-03

基金项目:国家自然科学基金重点项目(61232002);国家自然科学基金(61103039);武汉大学开放研究基金(SKLSE2012-09-16)

第一作者:余文喆,女,硕士研究生,研究方向为数据挖掘. E-mail: avecmanon@gmail.com.

通信作者:张蓉,女,副教授. E-mail: rzhang@sei.ecnu.edu.cn.

式是边栏广告,包括商品打折信息、与查询商品相似的商品信息和某些商品的评价信息等。当电子商务吸引了大量客户的时候,广告的盈利地位就变得非常稳固^[1]。广告帮助商家推销新商品、刺激商品流通以及树立品牌形象等。由此可见,广告作为一种营销工具对商家而言非常重要^[2]。

网络广告兴起的一个重要原因是,大家认为网络产生的信息可以给广告投放并盈利带来积极的影响。在电子商务网站上做好广告投放不仅可以吸引更多客户、增加网站访问量,还可以刺激消费,增大经济收益,从而促进电子商务的发展。电子商务网站的广告投放与其他媒体的广告投放主要有以下几方面的不同之处。

第一,用户目标明确。电子商务网站用户本身就有购物倾向或者需求,因而广告投放群体要区别于普通(新闻)网站的用户。

第二,用户数据丰富。电子商务网站拥有用户海量的购物行为(品牌、价位和风格偏好等)数据以及用户浏览历史数据(用户在不同时间浏览不同页面,把用户浏览电子商务网站的过程分为不同阶段,针对每个阶段投放合适的广告可以提高广告的有效性^[3]),利用这些数据可以更好地了解用户,从而支持精确锁定目标人群投放商品广告^[4]。

第三,情境数据可得。电子商务网站通过查询、定位等交互手段获取与用户当前环境相关的信息,使用这些信息可以有效地甄选推荐广告。

第四,移植手机平台。电子商务网站因为智能终端设备(如手机)的广泛应用,在迎来新机遇的同时其广告业务受到新的挑战,即广告要高效^[5]。比如,由于手机屏幕的限制和手机作为移动设备具有即时性的特点,通常手机用户希望看到更为高效的广告投放(需要考虑时间、地点和商品这三个维度来计算),切实迎合个人兴趣,否则将会大大影响手机终端用户的体验效果,从而造成客户流失。

目前,各大知名电子商务网站都在积极地制订广告策略,以推销商品。在 Amazon、eBay 和 Taobao 上都有广告推荐栏目,比如在 Amazon 的商品信息页面上,有“经常一起购买的商品”(见图1)、“购买此商品的顾客也同时购买”和“看过此商品后顾客购买的其他商品”(见图2)等商品推荐栏目。这些商品推荐系统正是广告投放的一种表现形式,即根据用户的购买历史、个人信息,向用户推荐他们可能喜爱的商品^[6],是电子商务网站上最常用的一种广告模式。这么做,一方面,商家通过推荐商品吸引潜在客户、建立品牌忠诚度,在竞争激烈的电子商务领域博得一席之地;另一方面,用户通过推荐系统的个性化推荐,在海量商品中发现自己心仪的商品,节省了挑选商品的时间和精力。但当前广告(推荐)所使用的信息以及采用的方法还非常有限,因而其有效性还有待提高。

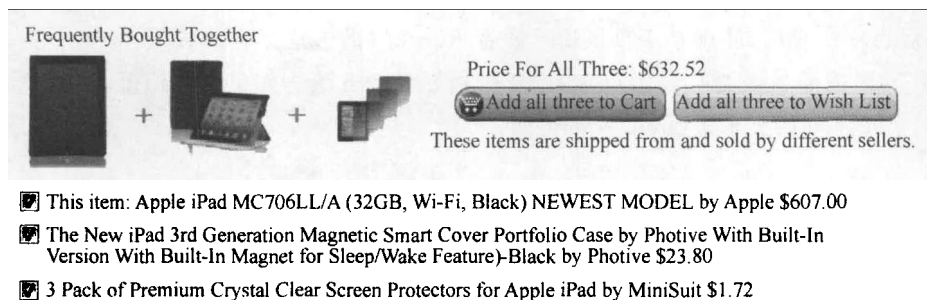


图1 “经常一起购买的商品”栏目

Fig. 1 “Frequently Bought Together”

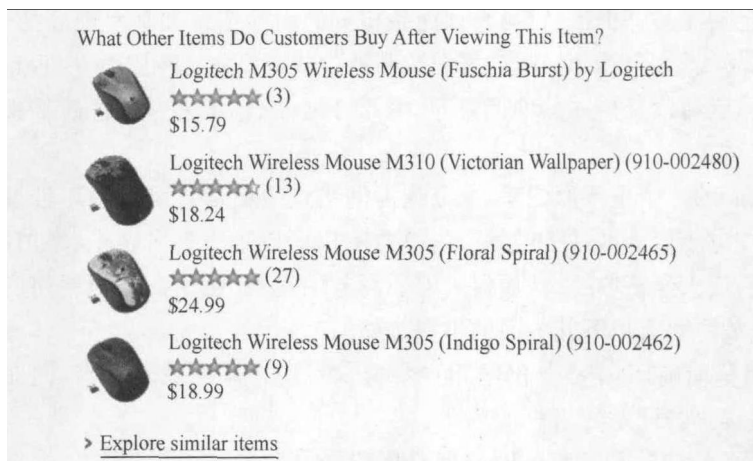


图2 “看过此商品后顾客购买的其他商品”栏目

Fig. 2 “What Other Items Do Customers Buy After Viewing This Item?”

本文将针对电子商务网站的特点,介绍电子商务网站上的主要广告形式,即商品推荐;并在介绍目前采用的主要推荐方法的基础上,总结在电子商务网站上做广告推荐的重点和难点.因此,本文内容安排是:第1节介绍现有推荐算法,即基于内容、协同过滤和混合的推荐算法,包括这些算法的核心思想、关键步骤和局限性;第2节综述现有推荐算法在实际应用中面临的问题和挑战;第3节总结全文.

1 现有推荐算法

目前,商品推荐系统根据推荐算法分为基于内容(Content-Based)、协同过滤(Collaborative Filtering)和混合(Hybrid)的推荐系统^[7].这些推荐算法广泛应用于学术界和工业界.然而,现有算法在实际应用中仍存在局限性.

1.1 基于内容的推荐算法

1.1.1 算法的核心思想和关键步骤

基于内容的推荐算法起源于信息检索^[8],通常用来推荐包含文本信息的文档、网页和新闻等,即向用户推荐与他们过去喜爱的商品相似的商品.该算法有两个关键步骤:

第一,表示商品和用户喜好.一件商品定义为一个商品档案,包含商品的特征属性;一个用户定义为一个用户档案,包含用户喜好,即该用户过去给予较高评分的商品的共同特征.商品、用户档案都由关键词组成,分别使用带权重的关键词向量来表示商品和用户,形式如 $\{\langle \text{key}, \text{value} \rangle\}$.通常,使用基于 TF-IDF 或者 Rocchio 的方法来计算权重.

第二,评估商品是否符合用户喜好.使用相似度量方法例如余弦相似度,计算商品、用户向量 \vec{w}_p 、 \vec{w}_u 之间的相似度 $\text{sim}(p, u)$:

$$\text{sim}(p, u) = \cos(\vec{w}_p, \vec{w}_u) = \frac{\vec{w}_p \cdot \vec{w}_u}{\|\vec{w}_p\| \times \|\vec{w}_u\|} = \frac{\sum_{i=1}^K w_{i,p} w_{i,u}}{\sqrt{\sum_{i=1}^K w_{i,p}^2} \sqrt{\sum_{i=1}^K w_{i,u}^2}},$$

其中 K 表示商品和用户档案中所有关键词的数量.商品、用户向量相似度越大商品越符合用户喜好.

除了使用上述信息检索算法,基于内容的推荐系统还会使用机器学习算法,例如利用聚类、决策树和贝叶斯分类器等进行推荐.其原理是根据商品的特征属性和用户评分(通常根据评分把商品分为用户喜爱、不喜爱两类,商品分类表示为 C_i),训练出一个用户喜好模型,使用该模型计算 $P(C_i | k_{i,j} \& \dots \& k_{n,j})$ ($k_{1,j}, \dots, k_{n,j}$ 表示商品关键词),来预测用户是否会喜爱某件未知商品^[9].

1.1.2 算法的局限性

基于内容的推荐算法的局限性有两点:

(1) 商品特征属性的提取.该算法需要提取足够多的商品特征属性来总结用户喜好.然而,对于多媒体数据,如图片、音频和视频等,提取特征属性比较困难.

(2) 推荐商品过分拟合.推荐商品的特征属性仅限于用户过去喜爱的商品的特征属性,导致推荐列表缺乏多样性,尤其是对新型商品而言,这种推荐算法完全失效.

1.2 协同过滤算法

1.2.1 算法的核心思想和关键步骤

从 Tapestry^[10]、GroupLens^[11] 和 Ringo^[12] 等早期推荐系统,到目前 Amazon^[13]、Netflix^[14] 等电子商务网站的商品推荐系统,都采用协同过滤算法:找到与用户有相似喜好的用户,向用户推荐相似用户过去喜爱的商品.该算法有两个关键步骤:

第一,生成相似用户.对一个用户,找到与该用户有相似喜好的一组用户,并且把这组用户评过分的商品作为候选的推荐商品.一个用户包含一组购买历史,即用户评过分的商品和相应评分,使用用户的商品评分向量来表示用户.通常,使用 Pearson 关联度量公式^[15],计算两两用户向量 \vec{r}_x, \vec{r}_y 之间的相似度 $\text{sim}(x, y)$ (用户 x, y 共同评过分的商品表示为 P_{xy}):

$$\text{sim}(x, y) = \frac{\sum_{p \in P_{xy}} (r_{x,p} - \bar{r}_x) (r_{y,p} - \bar{r}_y)}{\sqrt{\sum_{p \in P_{xy}} (r_{x,p} - \bar{r}_x)^2 \sum_{p \in P_{xy}} (r_{y,p} - \bar{r}_y)^2}}$$

对一个用户,选择与其相似度由高到低排名前 N 个用户作为相似用户.

第二,预测商品评分.对一个用户,根据相似用户给候选商品的评分,预测该用户给候选商品的评分.推荐系统中的协同过滤算法根据预测商品评分的不同方法分为两类:基于记忆和基于模型的算法^[7],前者计算一组相似用户 \hat{U} 给候选商品的加权评分和:

$$r_{u,p} = k \sum_{u' \in \hat{U}} \text{sim}(u, u') \times r_{u',p},$$

其中 k 表示归一化因子;后者使用概率模型建模,计算用户喜爱某件候选商品的概率^[16].

因为该算法计算了用户之间的相似度,所以又称基于用户的协同过滤算法.文献^[17]提出基于商品的协同过滤算法,计算了商品之间的相似度.

1.2.2 算法的局限性

协同过滤算法的局限性有两点:

(1) 用户评分稀疏性.该算法需要用户给足够多的商品评分来计算用户的相似度,然而,一方面,新用户、新商品的评分较少;另一方面,一些用户较为独特,他们评过分的商品可能没有其他用户的评分,导致这些用户很难找到相似的用户,影响了推荐的准确性.

(2) 算法的可扩展性.随着网上商品和选择网上购物的用户数量呈爆炸式增长,生成相

似用户的算法的可扩展性成为急需解决的问题.

1.3 混合的推荐算法

混合的推荐算法是融合了基于内容推荐和协同过滤这两种算法的推荐算法. 使用混合的推荐算法能够在一定程度上避免两种算法各自的局限性. 现有混合推荐算法主要分为三类: 第一类, 合并单独实现的两种推荐算法^[18]; 第二类, 在协同过滤算法中使用基于内容的推荐算法的特性, 即额外使用用户档案, 计算用户之间的相似度^[19], 这在一定程度上解决了协同过滤算法中存在的用户评分稀疏性问题; 第三类, 构建一个同时实现两种算法的模型^[20].

2 推荐算法面临的问题和挑战

随着商品、用户的增加, 用户需求的多样化, 在电子商务网站的实际应用中, 商品推荐系统正面临新的挑战, 主要包括推荐结果的多样化、个性化和智能终端推荐的时效化, 从而能最大程度地满足用户需求, 刺激经济增长.

和传统的推荐算法相比, 当前的推荐算法对动态信息的吸收和处理要求更高: 在静态的商品、用户等的描述信息基础上, 还要进一步考虑用户的购物行为历史, 以及当前的时间、地点, 甚至社会关系网络信息, 以获得我们所关注人的最新动态, 同时这些人的行为也会影响我们的个人行为. 个性化数据(Profile)可以由用户自己填写也可以通过之后的分析得到, 这部分数据是有效的推荐过滤条件. 丰富的购物历史可以准确细致地刻画用户的喜好. 时空数据目前在智能终端的推荐中广泛使用, 即利用当前时间和地理位置的特征进行推荐. 有效、综合地利用这些信息(图 3)将帮助商品推荐系统实现性能的扩展和质量提高.

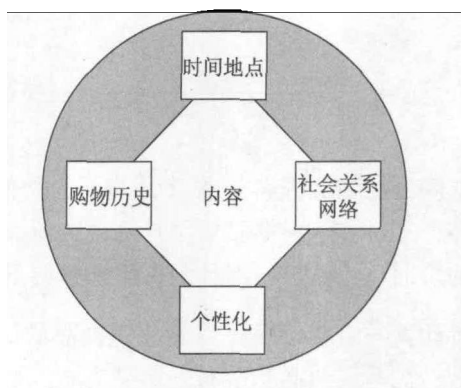


图 3 影响推荐结果的相关维度

Fig. 3 Dimensions affecting recommendation results

2.1 推荐列表多样化(Diversification)

和传统的搜索引擎系统相似, 商品查询可能会遇到查询词少、查询目标不明确, 以及查询结果(类型)繁多等问题. 解决这类问题的一个可行办法是多样化技术, 即最大程度地满足多数用户的需求; 尽量使每个用户都可以在第一页的结果中找到实现查询目标的所谓“最佳匹配”(Best Guess). 目前在搜索引擎上已经开始使用多样化技术^[21], 但该技术商品推荐系统中还没有被广泛采用. 现有商品推荐系统通常直接把预测评分(根据某种相似度值)由高到低排序的前 N 件商品推荐给用户, 导致推荐列表商品类型单一, 从而大大影响用户的

满意度。所以,商品推荐系统需要同时考虑推荐列表的准确性和多样性,达到两者的平衡。推荐列表多样化的目的,是在保证推荐商品符合用户喜好的情况下使推荐商品的类型各不相同,即最大程度地覆盖多数用户的喜好。同时,在手机用户增加的情况下,良好的多样化机制也将大大提高手机用户的购物体验,从而刺激消费。

在商品推荐系统中进行推荐列表多样化有两个难点:第一,如何实现推荐商品类型的多样化,以及如何定义和区分商品类型。通常根据商品之间特征属性值的相似度进行商品分类,选择不同类型的商品进行推荐,实现推荐列表的多样化^[22]。然而,一方面,该方法需要足够多的特征属性来计算商品之间的相似度;另一方面,两件有相同特征值的商品可能不是同类商品,需要采取更为细致的辨别措施,但这样会增加计算复杂度。第二,如何平衡推荐列表的准确性和多样性,来提高推荐列表的总体质量。通常赋予每件商品一个权重,选择权重和最大的一组商品作为推荐商品;或者使用阈值来限制由于商品多样性上升造成的准确性下降^[23]。然而,使用权重和阈值一方面存在如何取权重和阈值的值的问题;另一方面,根据电子商务网站的海量数据计算权重和阈值不仅耗时,而且一旦数据集有所变化,就需要重新计算^[24]。

2.2 个性化推荐(Personalization)

商品推荐系统需要考虑个性化推荐,比如根据用户的购物习惯、个人喜好等向用户推荐他们可能喜爱的商品。个性化推荐系统的目的是为新老客户准时、准确、合理地推荐他们可能会感兴趣的物品,其中一项重要任务就是为新产品的推广选择合适的用户,以及为新用户推荐合适的产品。通过良好的推荐手段,推荐系统可以提高用户满意度,从而留住网站的客户。

根据目前商品推荐系统使用的推荐算法和电子商务网站的数据特点,在电子商务网站中进行个性化推荐有三个难点:第一,使用基于内容的推荐算法时,如何自动生成用户档案(profile),覆盖用户所有的关键信息^[25],包括用户的个人信息、购物历史,甚至社会关系网络中的重要信息;第二,电子商务网站的数据具有稀疏性的特点,比如新用户、新商品的可用信息较少,使用协同过滤算法时,会影响生成相似用户的准确性,因而需要解决 Cold Start 等由数据稀疏性引起的问题;第三,电子商务网站的数据量大,而且仍在呈爆炸式增长,这就需要推荐算法具有可扩展性,而保证其高性能也成为急需解决的问题。

2.3 智能终端的推荐

随着智能手机的普及,商品推荐系统迎来新的挑战。智能终端的商品推荐系统的目的是随时随地向用户提供商品推荐;同时,由于手机屏幕的限制和即时性的特点,手机用户也希望看到更为高效的物品推荐。

根据智能终端的用户需求和数据特点,在智能终端上进行商品推荐有三个难点:一,智能终端的数据更为丰富,如用户当前所处环境的各种信息,包括时间、地理位置和天气等,需要考虑如何使用这些信息来体现用户所处环境,满足用户的特定需求,提高推荐质量^[26];二,智能终端的数据实时更新,如用户的地理位置信息等,因而需要实时更新推荐列表,这就要求推荐算法具有较高的计算性能;三,智能终端的数据噪声大,由于手机信号或者设备问题导致收集到的数据不够精确,如何降低噪声来减少对推荐的影响也是一个难题。

3 结束语

随着电子商务的迅速发展,商家开始在电子商务网站投放广告。商品推荐系统作为电子

商务网站上最常用的一种广告模式,在帮助商家推销商品,增大经济收益的同时,也节省了用户挑选商品的时间和精力.本文介绍了目前商品推荐系统使用的推荐算法:基于内容、协同过滤和混合的推荐算法,并在此基础上总结了现有推荐算法在电子商务网站的实际应用中面临的问题和挑战,即推荐结果的多样化、个性化和智能终端推荐的时效化.未来的商品推荐系统将不断改进这些问题,从而完善系统,最大程度地满足用户需求,促进电子商务蓬勃发展,达到双赢的目的.

[参 考 文 献]

- [1] WEN H J, CHEN H G, HWANG H G. E-commerce web site design: strategies and models[J]. *Information Management & Computer Security*, 2001, 9(1): 5-12.
- [2] ARENS W F. 当代广告学[M]. 8 版. 北京:人民邮电出版社, 2006.
- [3] EBAY. Online Retail Media[EB/OL]. 2012. <http://www2.ebayadvertising.com/uk/online-retail-media>.
- [4] HABEGGER J. Why Amazon is about to Become a Force in Online Advertising[EB/OL]. 2011. <http://www.commercialalert.org/issues/culture/internet-socialmedia/why-amazon-is-about-to-become-a-force-in-online-advertising>.
- [5] O'REILLY T. The Convergence of Advertising and E-Commerce[EB/OL]. 2010. <http://radar.oreilly.com/2010/02/convergence-advertising-mobile-ecommerce.html>.
- [6] SCHAFER J B, KONSTAN J, RIEDI J. Recommender systems in e-commerce[C]. *Proceedings of the 1st ACM conference on Electronic commerce*, 1999: 158-166.
- [7] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. *Knowledge and Data Engineering*, 2005, 17(6): 734-749.
- [8] BAEZA-YATES R, RIBEIRO-NETO B. *Modern Information Retrieval*[M]. [S. l.]: Addison-Wesley, 1999.
- [9] MOONEY R J, ROY L. Content-based book recommending using learning for text categorization[C]. *Proceedings of the fifth ACM conference on Digital libraries*, 2000: 195-204.
- [10] GOLDBERG D, NICHOLS D, OKI B M, et al. Using collaborative filtering to weave an information tapestry[J]. *Communications of the ACM*, 1992, 35(12): 61-70.
- [11] RESNICK P, IACOVOU N, SUCHAK M, et al. GroupLens: an open architecture for collaborative filtering of netnews[C]. *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 1994: 175-186.
- [12] SHARDANAND U, MAES P. Social information filtering: algorithms for automating "word of mouth"[C]. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1995: 210-217.
- [13] LINDEN G, SMITH B, YORK J. Amazon.com recommendations: item-to-item collaborative filtering[J]. *Internet Computing*, IEEE, 2003, 7(1): 76-80.
- [14] NETFLIX. The Netflix Prize[EB/OL]. 2009. <http://www.netflixprize.com>.
- [15] RODGERS J L, NICEWANDER W A. Thirteen ways to look at the correlation coefficient[J]. *American Statistician*, 1988, 42(1): 59-66.
- [16] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering [C]. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998: 43-52.
- [17] DESHPANDE M, KARYPIS G. Item-based top-N recommendation algorithms[J]. *ACM Transactions on Information Systems*, 2004, 22(1): 143-177.
- [18] CLAYPOOL M, GOKHALE A, MIRANDA T, et al. Combining content-based and collaborative filters in an on-line newspaper[C]. *Proceedings of ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation*, 1999.
- [19] BALABANOVIC M, SHOHAM Y. Fab: content-based, collaborative recommendation[J]. *Communications of the ACM*, 1997, 40(3): 66-72.

- [20] BASU C, HIRSH H, COHEN W. Recommendation as classification: using social and content-based information in recommendation[C]. Proceedings of the 15th national/10th conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence, 1998: 714-720.
- [21] AGRAWAL R, GOLLAPUDI S, HALVERSON A, et al. Diversifying search results[C]. Proceedings of the Second ACM International Conference on Web Search and Data Mining, 2009.
- [22] ZIEGLER C N, MCNEE S M, KONSTAN J A, et al. Improving recommendation lists through topic diversification[C]. Proceedings of the 14th international conference on World Wide Web, 2005: 22-32.
- [23] YU C, LAKSHMANAN L, AMER-YAHIA S. It takes variety to make a world: diversification in recommender systems[C]. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, 2009: 368-378.
- [24] BOIM R, MILO T, NOVGORODOV S. DiRec: Diversified recommendations for semantic-less Collaborative Filtering[C]. Proceedings of the 2011 IEEE 27th International Conference on Data Engineering, 2011: 1312-1315.
- [25] ADOMAVICIUS G, TUZHILIN A. Personalization technologies: a process-oriented perspective[J]. Communications of the ACM, 2005, 48(10): 83-90.
- [26] VAN DER HEIJDEN H, KOTSIS G, KRONSTEINER R. Mobile Recommendation Systems for Decision Making 'On the Go'[C]. Proceedings of the International Conference on Mobile Business, 2005: 137-143.

(上接第36页)

- [3] HRISTIDIS V, KOUDAS N, PAPAKONSTANTINOY Y. Prefer: A system for the efficient execution of multi-parametric ranked queries[C]//Proc SIGMOD, 2001: 259-270.
- [4] XIN D, CHEN C, HAN J. Towards robust indexing for ranked queries[C]//Proc VLDB, 2006: 235-246.
- [5] FAGIN R. Combining Fuzzy Information from Multiple Systems[C]//Proc 15th PODS, 1996: 216-226.
- [6] LI C K. Enabling data retrieval: by ranking and beyond[D]. Illincis: University of Illinois at Urbana-Champaign, 2007.
- [7] VIACHOU A, DOULKERIDIS C, KOTIDIS Y, et al. Reverse Top-k Queries[C]//ICDE Conference, 2010: 365-376.
- [8] ILYAS I F, BESKALES G, SOLIMAN M A. A survey of top-k query processing techniques in relational database systems[J]. ACM Comput Surv, 2008, 40(4): 11.
- [9] VLACHOU A, DOULKERIDIS C, KOTIDIS Y, et al. Monochromatic and Bichromatic Reverse Top-k Queries [J]. IEEE Trans Knowl Data Eng, 2011, 23(8): 1215-1229.
- [10] XIN D, CHEN C, HAN J W. Towards Robust Indexing for Ranked Queries[C]//VLDB, 2006: 235-246.
- [11] LIU Q, GAO Y J, CHEN G, et al. On Efficient Reverse k-Skyband Query Processing[C]//DASFAA(1), 2012: 544-559.