基于评论挖掘的改进的协同过滤推荐算法

王全民,王 莉,曹建奇
(北京工业大学 计算机学院,北京 100124)

摘 要:随着因特网的飞速发展,电子商务网站为人们提供了越来越多的选择,随之而来的信息过载和信息迷失问题日益严重,个性化推荐系统的出现极大地改善了这一情况。协同过滤是目前主流的推荐算法,但随着用户物品数目的日益增多和系统规模的不断扩大,用户-物品评分矩阵存在着严重的稀疏性等问题,导致推荐系统的推荐质量严重下降。针对此问题,文中提出了一种改进的协同过滤推荐算法,将评论挖掘技术引入协同过滤算法中,量化物品在各个特征上的分数,然后结合物品特征和用户评分共同计算物品相似度,将得到的物品预测评分填充用户-物品评分矩阵,最后结合基于用户的协同过滤思想对用户产生推荐。实验结果表明,改进的协同过滤推荐算法提高了推荐结果的精确度。

关键词:评论;协同过滤;相似度;推荐算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2015)10-0024-05
doi:10.3969/j.issn.1673-629X.2015.10.005

Improved Collaborative Filtering Recommendation Algorithm Based on Comments Mining

WANG Quan-min, WANG Li, CAO Jian-qi
(College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract: With the rapid development of the Internet, electronic commerce website provide more choice for people, but information overload and information lost problems become increasingly serious, the personalized recommendation system has greatly improved this situation. Collaborative filtering recommendation algorithm is a popular recommendation algorithm, but with the increasing number of items and users and the continuous expansion of the system, the serious user-item rating matrix sparse problem leads to a lower recommended quality. In order to solve this problem, put forward an improved collaborative filtering recommendation algorithm, which introduces comments mining technology into collaborative filtering algorithm to get the item score on each feature, and then combine the feature of items and the user score to calculate the item similarity, fill the predicted rating score into the user-item rating matrix, finally recommend to the user based on the user-based collaborative filtering ideas. The experimental result shows that the improved collaborative filtering recommendation algorithm improves the precision of the recommendation results.

Key words: comments; collaborative filtering; similarity; recommendation algorithm

0 引 言

随着互联网的发展与普及,信息过载的问题日益突出,个性化推荐技术应运而生。个性化推荐系统的定义是 Resnick & Varian 在 1997 年提出的,“它是利用电子商务网站向客户提供商品信息和建议,帮助用户决定应该购买什么产品,模拟销售人员帮助客户完成购买过程”[1]。现在这个概念已经被广泛使用,并且推荐算法已经被广泛应用于电子商务网站、新闻网站、数字图书馆等众多系统中,为每个用户个性化定制内容,从而方便了用户且提高了网站的销售额,其中协同过滤推荐算法是主流的推荐算法之一[2]。

协同过滤算法认为有相同兴趣的用户可能会喜欢相似的物品,且用户可能对相似的物品表达相同的兴趣。与传统的文本过滤技术相比,协同过滤推荐技术有许多优点,比如能够过滤难以进行机器自动内容分析的信息,能够基于一些复杂且难以表达的概念进行过滤,并且具有推荐新信息的能力。协同过滤推荐算法主要分为两种:基于用户的协同过滤推荐算法和基

收稿日期:2015-01-14 **修回日期**:2015-04-16 **网络出版时间**:2015-09-23
基金项目:国家自然科学基金资助项目(61272500)
作者简介:王全民(1963-),男,副教授,博士,硕士生导师,CCF 高级会员,研究方向为网络与信息安全;王 莉(1989-),女,硕士研究生,研究方向为推荐系统、网络与信息安全。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150923.1503.002.html

于物品的协同过滤推荐算法[3]。如果给系统提供足够的用户偏好信息和历史评分数据,这两种协同过滤推荐算法都能够表现出很好的推荐精确度,但随着内容的复杂度和用户物品数目不断增加,站点结构不断复杂,系统规模不断扩大等因素,一些问题就会暴露出来。比如用户–物品评分矩阵存在着严重的稀疏性问题[4-5]、可扩展问题[6]、冷启动问题[7-8]等等,这些问题导致了推荐系统的推荐质量下降和用户满意度降低。很多研究人员对协同过滤算法进行了深入研究,并提出了许多改进方案。例如,Chen 等[9]通过建立影响集来提高推荐精确度;Liu 等[10]通过建立个性化相似模型来推荐物品;YK[11]通过扩展 K –means 算法和分类来改进传统的协同过滤;还有学者根据可信度[12-13]和项目重要性[14]进行推荐,等等。

目前应用最广泛的协同过滤推荐技术大多是以物品的整体评分这种粗粒度偏好特征进行研究[15]。文中提出从评论数据出发来挖掘用户的特殊偏好,物品的评论中蕴含着许多重要信息,可以帮助用户了解该物品或者服务机构在大众心目中的口碑,因此受到了许多消费者和商业网站的青睐,同时也引起了许多研究者的广泛关注。

针对协同过滤出现的问题,文中提出从物品特征这种细粒度的角度来分析用户偏好,采用评论挖掘等技术,从评论中提取偏好标签,分析情感倾向性,并结合实际评分数据共同产生推荐。实验结果表明,该方法可以显著提高推荐精确度。

1　相关概念和整体框架

1.1　相关概念

概念 1:物品特征 Ft:指的是物品的某个属性特征,如屏幕、体积等。

概念 2:情感词 Dr:用来修饰产品特征且表达了用户的喜好倾向,如挺好的、流畅、漂亮等。情感词表达了文本的情感色彩,从褒贬倾向的程度可以分为三类:正面、中立和负面。

概念 3:偏好标签:是从物品评论中抽取的,由物品特征和情感词组成的标签 Tag=(Ft,Dr)。

概念 4:物品–特征评分矩阵 $\boldsymbol{R}(m*n)$。如图 1 所示,m 行代表物品数,n 列代表特征数,矩阵项 r_{ij} 表示物品 i 在特征 j 的评分值。

	Ft_1	…	Ft_j	…	Ft_n
$Item_1$	r_{11}	…	r_{1j}	…	r_{1n}
…	…	…	…	…	…
$Item_i$	r_{i1}	…	r_{ij}	…	r_{in}
…	…	…	…	…	…
$Item_m$	r_{m1}	…	r_{mj}	…	r_{mn}

图 1　物品–特征评分矩阵

概念 5:用户–物品评分矩阵 $\boldsymbol{A}(x*m)$。其中,x 表示用户数目,矩阵项 a_{ij} 表示用户 i 对物品 j 的实际打分。矩阵中每一行代表了用户的评价向量,每一列代表了这个物品的被评价向量。

1.2　整体框架

基于评论挖掘的改进的协同过滤推荐算法以用户评论和实际评分数据作为数据源,分为以下 4 个模块,如图 2 所示。

(1)形成物品特征矩阵:首先进行评论预处理,从中提取物品特征,通过剪枝处理过滤掉不符合要求的物品特征,存入事务文件。接着根据句法结构模板提取情感词,评估情感词的情感倾向。然后进行物品特征聚类,量化物品在各个聚类后的特征上的分数,存入物品–特征矩阵中。

(2)计算物品相似度:根据改进的相似度计算公式计算物品相似度。

(3)填充用户物品评分矩阵:根据基于物品的协同过滤计算未评分物品的预测评分,填充到用户–物品评分矩阵中去。

(4)产生推荐:基于用户的协同过滤得到 TOP–N 推荐。

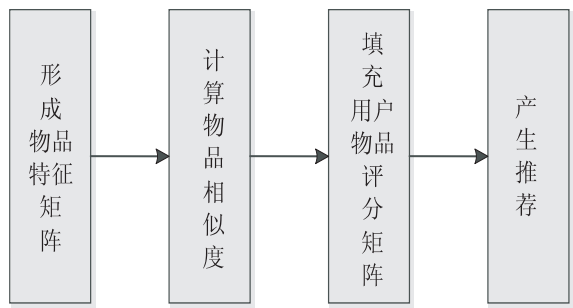

图 2　系统框架

2　改进的协同过滤推荐算法

2.1　形成物品–特征矩阵

Step 1:评论预处理。

评论预处理是关键的第一步,首先对评论数据进行分词和词性标注处理,提取名词,然后进行停用词过滤处理,存入事务文件,为下一步做准备。

中文分词有些人习惯称之为切词,指的是把汉字序列分成一个个单独的有意义的词条。英语因为单词之间有空格的存在使得程序很容易识别单词与单词之间的界限,而汉语是以字为基本的书写单位,词语之间没有明显的区分标志,因此中文分词是第一步也是关键的一步,分词的准确性影响着推荐的精确度。中文分词的难点主要在于词语切分、未登录词的识别和词

性标注三个方面。目前国内的分词方法很多,文中使用中科院的 ICTCLAS 来对评论数据进行分词处理,ICTCLAS 系统是国内比较权威的词法分析软件,可以区分 40 种词性。停用词是指在语言中会经常出现,虽然它们是构成句子的成分词语,却无法表达文本的意思,比如“了”、“我们”、“至于“等等。

Step 2:物品特征的提取。

物品特征包括直观特征和隐藏特征。比如“这款手机尺寸合适,但是待机时间有点短”,其中”尺寸”就是直观特征,而”电源容量”是这句话的隐藏特征,需要通过语义理解分析才能够得到。提取隐藏特征的技术目前还不成熟,文中只针对手机的直观特征做出了抽取,暂时没考虑隐藏特征,将其作为下一步的研究方向。

从评论语料库中抽取物品特征,物品特征往往是提到的频率最高的名词或者名词短语。ICTCLAS 词性标注只标注了名词,没有标注名词短语。可以根据表 1 的规则来识别名词短语,单独的一个名词也可以看做是名词短语。然后利用关联规则 Apriori 算法来提取满足最小支持度的频繁项集作为候选特征词,文中只考虑了 3 项以及 3 项以下的频繁项集。提取出来频繁项之后再进行下一步剪枝过滤处理。

表 1 名词短语模板

名词短语模板	示例
名词	系统
形容词+名词短语	不错的手机
名词短语+名词短语	音效屏幕摄像头
名词短语+助词+名词短语	屏幕的材质
名词短语+连词+名词短语	屏幕和摄像头

Step 3:物品特征剪枝过滤。

剪枝处理 1:提取出来的物品特征有些在评论中出现的频率很高,但是和物品特征毫无关系。为此,通过常见名词却非物品特征词词表,来过滤候选特征集中的非物品特征频繁项。

剪枝处理 2:利用与种子特征集合的 PMI 剪枝处理。

确定一个种子特征集合,通过计算候选特征与种子特征集合的点互信息过滤特征。

首先人工从候选频繁项集中选出具有代表性的特征词组成种子特征集合,人工定义一组种子词汇 Seeds = {内存,速度,屏幕,服务,系统,性价比,电池,按键,价格,待机时间,音效,外观,散热,电池容量,分辨率……},通过计算和种子词汇的相似度来对物品特征过滤。

PMI 的计算公式如下:

$$\mathrm{PMI}(w_1)=\sum_{w\in \mathrm{Seeds}}\log_2\frac{\mathrm{hits}(w_1,w)}{\mathrm{hits}(w_1)*\mathrm{hits}(w)}$$

其中,$\mathrm{hits}(w_1,w)$ 为候选特征词 w_1 和种子特征词 w 在评论语料库中共同出现的次数;$\mathrm{hits}(w)$ 为词 w 单独出现的次数。

高的点互信息值 PMI 意味着词语之间的关联性强,若 PMI 值大于等于指定阈值,则该词为特征词,否则过滤掉。

Step 4:情感词的提取。

由于情感词大多数情况下都是形容词、副词,采用 3 词模型来确定用户针对某一特征的情感词。

在评论中抽取和物品特征相关联的情感词,情感词往往与特征词距离比较近且存在一定的句法规则。

句法结构模板如表 2 所示。

表 2 句法结构模板

句法结构模板	示例
名词短语+形容词	外观漂亮
名词短语+副词+形容词	外观不好看
形容词+名词短语	漂亮的外观
形容词+助词+名词短语	流畅的运行

Step 5:评估情感词的倾向。

情感倾向分析的粒度主要分为三个级别:篇章级、句子级和词语级。文中将从词语级来分析偏好标签的情感倾向。借助 HowNet 建立情感词典,计算情感词的情感倾向。有些评论句情感词会带有否定词修饰,如“这个手机不好看”,好看是褒义词,但是和否定副词“不”在一起就是贬义词,所以需要整理出常用的否定词词表一起对情感极性做出判断,如没,无,不,别,甭,莫,休,未,勿,否,非,没有,不用,不必,不要,不可,未必,不是,并非,等等。

Step 6:物品特征聚类。

有许多物品特征词表达着相同的主题。汇总产品特征,保证在语义空间里独立。首先通过同义词词典显示去重,然后使用 K – means 聚类算法对抽取出的特征词进行聚类,每个分组表述的是同一个特征,使用频率最大的特征词代表该分组,从而得到独立的 K 个偏好标签特征组。

Step 7:评估偏好标签的情感强度。

接着量化物品在各个聚类后的特征上的分数。在提取物品特征和情感词对之后,评估情感强度,作为物品在该维特征值上的得分。文中采用五分制,score(Dr)褒义的评分为 5 分,贬义为 1 分,中性为中间值 3 分。

情感强度计算公式为:

$$R_{ij}=\frac{\sum \mathrm{score}(\mathrm{Dr})}{\mathrm{num}}$$

物品 Item_i 在该特征 Ft_j 上的平均情感强度等于分值之和除以该特征上的偏好标签的数目。用平均分来表示用户对物品某个特征的喜好,该评分综合了所有用户的评分,更具有说服力。

2.2　物品的相似度计算

构建物品–特征矩阵,离线计算物品的相似度。相比于基于用户的协同过滤,基于物品的协同过滤计算量要少得多。基于用户的协同过滤适用于用户规模较小的情况,而基于项目的协同过滤适用于物品规模较大的情况。在电子商务领域,用户数量不论是从总体规模还是增长速度上都比物品要快得多,因此在电子商务领域中多采用基于项目的协同过滤算法。

传统的协同过滤一般采用评分来计算相似度,但是存在着数据稀疏性等问题,隐式反馈等可以弥补显示反馈数据的缺少带来的不足。文中将物品的评分相似度和物品的特征相似度结合起来计算物品相似度,将物品的评分相似度和特征相似度采用适当的权值进行结合,权值大小表示物品的评分及其特征对产生最近邻居的影响程度,然后根据项目的相似度阈值获取项目的最近邻居,再依据项目最近邻居产生物品推荐结果。

$$\text{SIM}(\text{Item}_i,\text{Item}_j)=\beta\times\text{Sim}_{\text{score}}(\text{Item}_i,\text{Item}_j)+(1-\beta)\times\text{Sim}_{\text{Ft}}(\text{Item}_i,\text{Item}_j)$$

其中,$\text{Sim}_{\text{score}}(\text{Item}_i,\text{Item}_j)$ 表示实际打分得到的物品相似度;$\text{Sim}_{\text{Ft}}(\text{Item}_i,\text{Item}_j)$ 表示由物品–特征矩阵得到的物品相似度;β 的取值范围为[0,1]。实验得到 $\beta=0.7$ 的时候,算法的准确度较高,即当实际打分得到的物品相似度占 70%,物品–特征矩阵得到的物品相似度占 30%,前者占主导作用,后者修正相似度。

用余弦相似度[16]计算物品之间的相似度,夹角越小说明越相似。

$$\text{Sim}(\text{Item}_i,\text{Item}_j)=\cos(\text{Item}_i,\text{Item}_j)=\frac{\overrightarrow{\text{Item}_i}\times\overrightarrow{\text{Item}_j}}{\|\overrightarrow{\text{Item}_i}\|\times\|\overrightarrow{\text{Item}_j}\|}$$

找出和用户喜欢的物品相似的物品推荐给用户。

2.3　填充用户–物品评分矩阵

基于物品的协同过滤预测用户 user 对未评分项目 Item 的评分公式:

$$\text{Pre}(\text{user},\text{Item})=\frac{\sum_{i\in\text{ratedItems(user)}}\text{SIM}(\text{Item},\text{Item}_i)\times\text{score}_i}{\sum_{i\in\text{ratedItems(user)}}\text{SIM}(\text{Item},\text{Item}_i)}$$

其中,ratedItems(user)为 user 做出过打分的物品集合。

得到物品之间的相似度之后,选择与目标物品相似度最大的邻居集合进行评分预测,将得到的物品预测评分填充用户–物品评分矩阵的空矩阵项,来降低用户–物品矩阵的稀疏性。如果没有实际评分,则用户–物品评分矩阵的矩阵项填充评分预测值 Pre(user,Item),来提高推荐质量。

2.4　产生 TOP–N 推荐

TOP–N 推荐是指由最近邻居集合产生的候选推荐物品集合中选出前 N 个评分值最大的物品进行推荐。

首先根据余弦相似度公式得出目标用户的前 K 个最近邻居,然后基于用户的协同过滤重新预测评分,公式如下:

$$\text{Pre}(\text{user},p)=\overline{\text{score}_{\text{user}}}+\frac{\sum_{u\in\text{neighbor(user)}}\text{sim}(\text{user},u)\times(\text{score}_{u,p}-\overline{\text{score}_u})}{\sum_{u\in\text{neighbor(user)}}\text{sim}(\text{user},u)}$$

将得到的物品预测评分按照从大到小的顺序排序,将前 N 个物品推荐给目标用户。

3　实　验

3.1　实验数据集和评价标准

通过爬虫程序从互联网上获取了 286 个用户对所购买手机的 2 000 多条评论数据和 3 000 多条评分数据。其中 80% 作为训练集,20% 为测试集。

采用平均绝对误差(Mean Absolute Error,MAE)作为预测评分的评价标准,将预测得到的预测评分与实际评分之间的绝对平均误差来判断预测的准确性,MAE 越小表示推荐质量越高。设预测的用户评分集合为 $\{p_1,p_2,\cdots,p_N\}$,对应的实际用户评分集合为 $\{q_1,q_2,\cdots,q_N\}$,MAE 定义为:

$$\text{MAE}=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N}$$

3.2　实验结果

相似度度量标准比较如图 3 所示。

实验采用的目标用户邻居数目从 5 增加到 40,间隔为 5。实验结果表明,传统的基于用户的协同过滤由于其评分矩阵稀疏等原因,精确度较低;当最近邻居数目较少时,基于物品的协同过滤接近改进后的算法,但随着邻居数目的增加,差距拉大;改进后的算法要优于基于物品的协同过滤和基于用户的协同过滤,且随着用户邻居数目的增加,算法的推荐精确度提高。

4　结束语

协同过滤推荐算法目前面临许多问题,而用户评论中蕴含着大量有用的信息,文中提出了一种新的算

法来提高推荐的精确度,利用自然语言处理技术从评论中提取物品特征并量化偏好得分,改进相似度得到新的预测评分填充用户-物品评分矩阵,缓解了协同过滤中的矩阵稀疏性问题,得到的邻居用户集合更加准确。实验结果表明,改进的协同过滤推荐算法显著地提高了推荐质量。

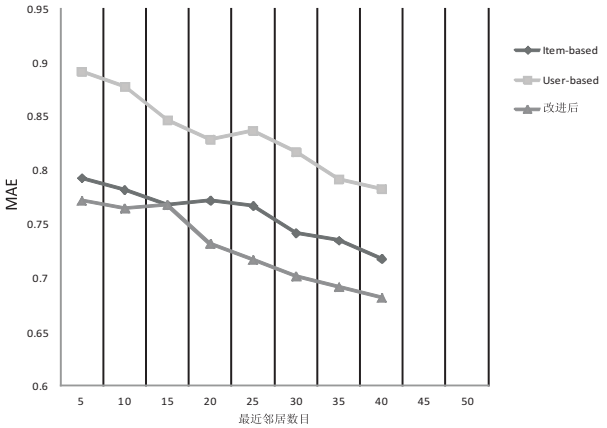

图 3 相似度度量标准比较

参考文献:

[1] Zhou Z, Ezeife C I. A low-scan incremental association rule maintenance method based on the apriori property[C]//Proc of advances in artificial intelligence. Ottawa, Canada: Springer, 2001: 26-35.

[2] 王卫平,杨金侠. 个性化信息服务中基于 Tag 的用户兴趣模型[J]. 计算机系统应用,2011,20(2):80-84.

[3] 许海玲,吴 潇,李晓东,等. 互联网推荐系统比较研究[J]. 软件学报,2009,20(2):350-362.

[4] 黄创光,印 鉴,汪 静,等. 不确定近邻的协同过滤推荐算法[J]. 计算机学报,2010,33(8):1369-1377.

[5] Sarwar B M, Karypis G, Konstan J A, et al. Application of dimensionality reduction in recommender system-a case study [C]//Proceedings of the ACM WebKDD web mining for e-commerce workshop. Boston, MA, United States: ACM, 2000: 82-90.

[6] 赵银春,付关友,朱征宇. 基于 Web 浏览内容和行为相结合的用户兴趣挖掘[J]. 计算机工程,2005,31(12):93-94.

[7] Bobadilla J, Ortega F, Hernando A, et al. A collaborative filtering approach to mitigate the new user cold start problem[J]. Knowledge-based Systems, 2012, 26: 225-238.

[8] Schafer J B, Frankowski D, Herlocker J, et al. Collaborative filtering recommender systems[J]. The Adaptive Web, 2007, 4321: 291-324.

[9] Chen Jian, Yin Jian. A collaborative filtering recommendation algorithm based on influence sets[J]. Journal of Software, 2007, 18(7): 1685-1694.

[10] Liu X, Data A, Rzadca K, et al. Stereo trust: a group based personalized trust model[C]//Proceedings of the 18th ACM conference on information and knowledge management. Hong Kong, China: ACM, 2009: 7-16.

[11] Wu Y K, Tang Z H. Collaborative filtering system based on classification and extended k-means algorithm[J]. Advances in Information Sciences and Service Sciences, 2011, 3(7): 187-194.

[12] Jamali M, Ester M. Trust walker: a random walk model for combining trust-based and item-based recommendation [C]//Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. Paris, France: ACM, 2009: 397-406.

[13] Jeong B, Lee J, Cho H. User credit-based collaborative filtering[J]. Expert Systems with Application, 2009, 36(3): 7309-7312.

[14] Bobadilla J, Hernando A, Ortega F, et al. Collaborative filtering based on significances[J]. Information Sciences, 2012, 185: 1-17.

[15] Lee J S, Jun C H, Lee J, et al. Classification-based collaborative filtering using market basket data[J]. Expert System with Application, 2005, 29(3): 700-704.

[16] Ahn H J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]. Information Science, 2008, 178(1): 37-51.

(上接第 23 页)

检测算法[J]. 微计算机信息,2006,22(1):265-268.

[6] Qiao Yu, Ong S H. Arc-based evaluation and detection of ellipse[J]. Pattern Recognition, 2007, 40(7): 1990-2003.

[7] 范 怡,傅继武. 基于中点提取的椭圆检测算法[J]. 计算机应用,2011,31(10):2705-2707.

[8] 袁 理,叶 露,贾建禄. 基于 Hough 变换的椭圆检测算法[J]. 中国光学与应用光学,2010,3(4):379-384.

[9] Barwick D S. Very fast best-fit circular and elliptical boundaries by chord data[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(6): 1147-1152.

[10] Nguyen T, Ahuja S. A real-time ellipse detection based on edge grouping[C]//Proceedings of the 2009 IEEE International conference on systems, man and cybernetics. Piscataway, NJ: IEEE, 2009: 2793-2795.

[11] 李楠楠,卢荣胜,李 帅,等. 基于边界曲线弧分割的多椭圆检测[J]. 计算机应用,2011,31(7):1853-1855.

[12] Mai F, Hung Y. A hierarchical approach for fast and robust ellipse extraction[J]. Pattern Recognition, 2008, 41(8): 2512-2524.

[13] 吴 倩,邹 伟,徐 德,等. 面向惯性约束聚变实验靶图像的快速椭圆检测[J]. 中国图象图形学报,2014,19(1):76-84.

[14] Zhang T Y, Suen C Y. A fast parallel algorithm for thinning digital patterns[J]. Communications of the ACM, 1984, 27(3): 236-239.

基于评论挖掘的改进的协同过滤推荐算法

作者： 王全民， 王莉， 曹建奇， WANG Quan-min， WANG Li， CAO Jian-qi
作者单位： 北京工业大学 计算机学院,北京,100124
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(10)