主题搜索引擎中特征模型技术的研究

文必龙[1],唐苏龙[1],张 浩[2]
(1. 东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318;
2. 天津理工大学 计算机科学与技术学院,天津 300191)

摘 要:主题搜索引擎的研究难点之一就是主题与网页信息之间的准确匹配。通过对网页的特征进行分析,提取网页特征中的主题特征词,并用提取的主题特征词表示网页主题信息,提出了利用网页特征及特征之间的关系来建立网页特征模型。该特征模型能准确地描述网页的内部特征和外部特征的主题表现力,有利于计算网页与主题之间的相似度。实验结果表明该特征模型能有效地表达网页的主题信息,并有助于提高主题搜索引擎的资源发现率和搜索准确率。

关键词:主题;主题搜索引擎;特征;特征模型;相关度计算
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2013)04-0087-04
doi:10.3969/j.issn.1673-629X.2013.04.021

Research on Feature Model Techniques of Focused Search Engine

WEN Bi-long[1], TANG Su-long[1], ZHANG Hao[2]
(1. School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China;
2. School of Computer Science and Technology, Tianjin University of Technology, Tianjin 300191, China)

Abstract: In focused search engine research, a difficult point is the precise match between theme and Web page. After analyzing the features, extract the theme feature words from Web pages, and take the theme feature words as Web page theme information, a method of using the Web page features to build Web feature mode is proposed. The feature model can accurately describe the subject of internal features and external features of the Web's expression, and is beneficial to calculate the similarity between Web pages and themes. The result of the experiment shows that Web feature model can effectively express the Web page theme, and contribute to improve the search resource discovery rate and accuracy rate of the focused search engine.

Key words: theme; focused search engine; feature; feature model; similarity calculation

0 引 言

主题搜索引擎是以某一特定专题或专门领域的网络信息资源库为目标,自动采集符合这一专题或领域的信息资源,满足特定用户对某一专题或领域的更深入的查询需求。主题搜索引擎在搜索过程中,只访问与主题相关的网页,而不用遍历整个网站。因此,如何让主题搜索引擎在采集数据时多抓取与主题相关的网页,而尽可能少抓取与主题无关的网页,提高检索效率,是主题搜索引擎研究的关键问题[1]。

目前,主题搜索引擎的研究主要是集中在对主题爬行器的研究,主题爬行器的爬行策略主要有基于网页内容评价的搜索策略和基于网页超链接评价的搜索策略两种[2]。基于网页内容评价的主题爬行策略以向量空间模型为基础,利用网页中的文本内容、超链接、锚文本等文字信息来评价网页超链接价值的高低,并以此指导主题爬行器的爬行。这种搜索策略的代表算法有 Best First Search[3] 和 Shark-Search[4] 等。基于网页内容评价的算法只是简单的利用网页的文字信息来判断网页的主题相关性,没有考虑到网页间的超链接对网页主题的影响。基于网页超链接评价的搜索策略通过分析网页间的相互引用关系来确定超链接的重要程度,从而决定超链接的访问顺序。基于网页超链接评价的搜索策略的代表算法有 PageRank 算法[5] 和 HITS 算法[6,7]。这类搜索算法过分强调超链接的作用,忽视了网页本身与主题的相关性,没有考虑用户的实际搜索请求,不适合特定主题的搜索。

文中通过对网站网页的特征进行分析,根据主题

收稿日期:2012-08-09;**修回日期**:2012-11-15
基金项目:国家科技重大专项(2011ZX05023-005-012);黑龙江省教育科学技术研究项目(11551018)
作者简介:文必龙(1967-),男,教授,博士,研究方向为软件工程与集成技术;唐苏龙(1988-),男,硕士研究生,研究方向为信息智能分析与处理。

特征词库提取网页特征中包含的主题特征词构建网页特征向量,并用网页特征及特征之间的关系来建立网页特征模型,从而充分利用网页的内部特征和外部特征的主题相关度来提高主题搜索引擎的资源发现率和搜索准确率。最后,通过实验证明所建立的特征模型能有效地表达网页的主题信息,并有助于提高主题搜索引擎的资源发现率和搜索准确率。

1 基于特征模型的主题搜索引擎设计

主题搜索引擎不同于传统的通用搜索引擎[8],它需要在对网页进行索引前进行主题相关度分析,尽可能多的获取与主题相关的网页而过滤掉与主题无关的网页,并对网页进行主题相关度计算,最终提高用户对特定主题搜索的效率。其结构设计如图 1 所示,与通用搜索引擎相比,基于特征模型的主题搜索引擎还包括对网页内部特征和外部特征的主题特征词提取和主题相关度分析,以及用于除掉与主题不相关的网页的过滤器。索引数据库中存放的是用特征模型描述的网页特征。搜索时,先将用户输入的关键词用向量空间模型表示,再与索引数据库中的数据进行匹配,最后将查询结果按相关度进行排序后输出。

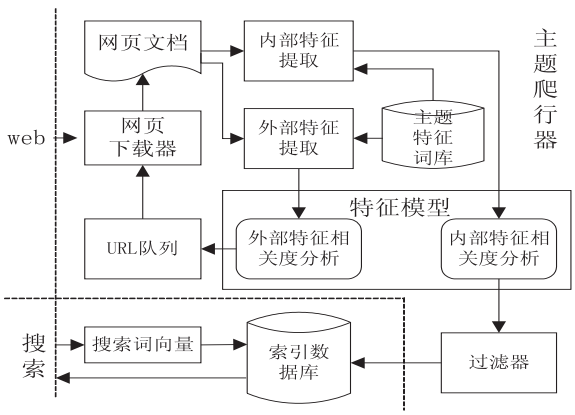

图 1 基于特征模型的主题搜索引擎结构设计

2 网页特征分析和提取

2.1 网页特征分析

特征是描述客观事物属性的概念或术语,体现了事物所具体的某种属性或特点。在网页中,通常用网页标题、网页正文、锚文本、超链接等属性来描述网页的信息,因此这些属性均可以称为网页特征。

在 Web 中,网页是使用超文本标记语言(HTML)来描述的。HTML 文档结构的数据不同于普通文本的数据,它是一种半结构化数据,其不同的文本信息分别用不同的 HTML 标记来描述,如网页标题用< TITLE > 标记,锚文本用<A>标记等。通过对大量中文网页的特征进行分析,发现网页特征与 HTML 标记有如下关系,如表 1 所示。

表 1 网页特征与 HTML 标记关系

特征	标记
网页标题	<TITLE>
网页描述文本	<META>
正文标题	<H1> ~ <H6>
强调性文本	<B>,<I>,<STRONG>
锚文本	<A>
网页的超链接	<A>中 HREF 属性
网页正文文本	除以上外的其它标记

通过上述分析,根据特征在网页中的作用,可以将网页的特征分为内部特征和外部特征两种:内部特征是网页本身所固有的特征,用来描述网页自身,能直接表达网页的主题信息,包括网页标题、正文标题、强调性文本和网页正文文本等;外部特征是实现网页本身与其它网页之间相互关联的特征,对网页主题信息具有预测作用,包括锚文本和网页的超链接等。

2.2 网页特征提取

特征提取是从网页特征中提取主题特征词条,排除那些被认为无关的词条的过程。词、词组和短语是网页特征文本的基本组成要素,并且在不同主题的网页中,各词条出现频率有一定的规律性,不同的主题特征词条可以区分不同主题的网页。因此可以对网页特征文本中的主题特征词条进行提取。

网页特征的主题特征词提取步骤如下:

① 建立主题特征词库。主题特征词库的建立一般是对给定一个与主题相关的样本网页集,提取网页中共同出现的特征词,将在样本网页集中都出现的特征词设为主题特征词,根据特征词出现的频率设定一个相应的权重,并在领域专家的参与下对主题特征词条进行审核评价后,添加到主题特征词库。

② 提取出网页特征文本中的主题特征词条。先对网页特征文本进行分词,并对停用词进行过滤,再根据主题特征词库提取主题特征词。最后获取网页特征中与主题相关的特征词条,得到主题特征词集,如图 2 所示。

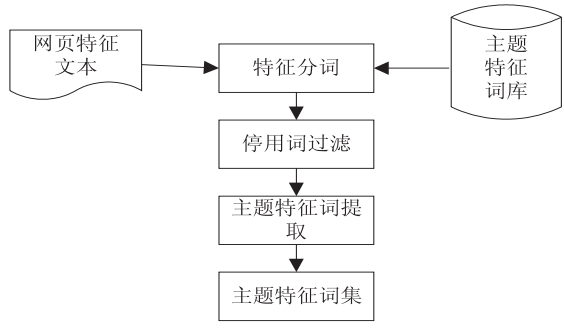

图 2 网页特征提取过程

③ 通过对所有网页特征的文本进行提取,得到一组规范统一、无噪声数据的网页特征。

3　网页的特征模型

3.1　特征模型的引入

网页的文本是用自然语言描述的[9],计算机无法理解其语义信息,因此必须将网页的信息表示成计算机能处理的形式。向量空间模型是最有效的文本表示方法之一,向量空间模型的基本思想是[10]:首先将文本分词处理,提取能代表文档的特征词条,将获取的词条数量作为向量的维数 n;同时,词条 t_i 在文档中出现次数作为该词条在文档中的权重 w_i;最后 Web 文档可表示为 $D=(t_1,w_1,t_2,w_2,\cdots,t_i,w_i,\cdots,t_n,w_n)$。

向量空间模型虽然能很好地用特征词来表示网页文本,但是在实际应用中,通常网页正文内容的是比较长的,导致表示文档的特征向量可能会达到数千维甚至上万维的大小,但其中有不少特征词是与网页主题无关的。其次,传统的向量空间模型将网页看作普通的文本,由于 HTML 文档与普通的文档在结构上有很大的不同,特征词在不同位置对文档的重要程度是有差别的。基于以上两点,借鉴向量空间模型的思想,本文利用网页的特征建立网页的特征模型来描述网页的主题信息。

3.2　网页的特征模型表示

特征模型是用来描述特征及特征之间的关系。由于网页的结构形式与普通文档有很大的不同,它不是一个结构化文档,不同的特征对网页主题表现的形式和价值是不同的。根据内部特征和外部特征对网页的作用,网页特征模型主要包括内部特征相关分析和外部特征相关度分析两个部分。

3.2.1　内部特征相关度分析

通过上述对网页特征的分析和提取,得到的网页内部特征都是由主题特征词条构成,可以通过如下步骤计算网页内部特征与主题的相关度:

① 构建主题特征向量。从主题特征词库获取主题特征词条,把主题特征词条的个数 n 作为向量空间的维数,用这个主题特征向量来表示主题,则主题可以表示为:

$$\boldsymbol{T}=(t_1,t_2,\cdots,t_i,\cdots,t_n)\quad(i=1,2,\cdots,n)$$

其中,t_i 为主题特征库中的主题特征词条。

显然,不同的主题特征词条的对主题表现力也是不同的,因此主题的特征词条权重向量可以表示为:

$$\boldsymbol{W}=(w_1,w_2,\cdots,w_i,\cdots,w_n)\quad(i=1,2,\cdots,n)$$

其中,w_i 为 $\boldsymbol{T}$ 中主题特征词条 t_i 对应的权重。

② 网页特征的向量空间模型表示。根据主题特征向量,将上述提取的网页主题特征词条构建成网页特征的向量空间模型 $\boldsymbol{F}$。

$$\boldsymbol{F}=(f_1m_1,f_2m_2,\cdots,f_im_i,\cdots,f_nm_n)\ (i=1,2,\cdots,n)$$

一般地,网页特征中包含部分主题特征词条,若特征 $\boldsymbol{F}$ 中包含主题特征词条 t_i,则 f_i 取值为 1;否则取值为0,表示该特征词不存在。m_i 为特征词 t_i 在特征 $\boldsymbol{F}$ 中出现的次数,出现次数越多,说明网页与该主题特征词条相关程度越大。

③ 网页特征向量的相关度分析。相关度分析是判断网页与主题是否相关的关键,可以用两个向量的夹角的余弦值来表示网页特征的主题相关度。

$$W_{(\boldsymbol{F},\boldsymbol{W})}=\cos(\boldsymbol{F},\boldsymbol{W})=\frac{\boldsymbol{F}\cdot\boldsymbol{W}}{|\boldsymbol{F}|\cdot|\boldsymbol{W}|}$$

$$=\frac{\sum_{i=1}^{n}f_i\cdot m_i\cdot w_i}{\sqrt{\sum_{i=1}^{n}f_i^2\cdot m_i^2\cdot w_i^2}\times\sqrt{\sum_{i=1}^{n}w_i^2}}$$

因此,分别对每个内部特征的主题相关度进行分析,网页内部特征的主题的相关度可表示为:

$$W_{\text{in}}=\sum\lambda_i\times W_{(\boldsymbol{F},\boldsymbol{W})}$$

其中,λ_i 是根据特征 F_i 在网页中的主题表现力所设定的权重因子。

④ 设定一个阈值 r,当 W_{in} 的值大于或等于 r 的值时,可以认为网页与主题相关,当 W_{in} 的值小于 r 的值时,认为与主题无关。通常 r 值需要通过大量实验来确定,r 值越小,则能获取更多的网页。

3.2.2　外部特征相关度分析

为了提高获取主题相关的网页的效率,通常可以对网页的外部特征进行评价,预测网页的主题相关度。外部特征相关度分析包括锚文本的相关度分析和超链接的相关度分析两部分。

研究表明,锚文本包含的特征词与用户搜索该网页时所使用的搜索词具有一定的相似性[11,12]。因此,锚文本也有助于网页主题的预测。可以计算锚文本的主题相关度来提高网页的预测准确率。

①获取指向同一网页的锚文本,将所有的锚文本整合成一个锚文本文档,并提取锚文本文档中的主题特征词条,构建特征向量。

②计算特征向量与主题特征向量的相关度,得到该网页的锚文本主题相关度 W_{acr}。

根据网页的超链接结构特点,设计规范的网站通常会将具有相同主题的网页按照栏目进行归类,并将不同主题的网页分别放置在网站不同的目录下,且这些信息往往会反映在网页的 URL 字符串上,这些特点有助于网页的类型分类。通过比较两个 URL 字符串的相似度来计算两张网页的主题相关度。

$$W_{\text{url}}=\frac{\text{com}(\text{URL}_1,\text{URL}_2)}{\max(\text{len}_{\text{URL}_1},\text{len}_{\text{URL}_2})}$$

式中，$com(URL_1, URL_2)$ 为两个 URL 字符串从左自右相同的字符数，$max(len_{URL_1}, len_{URL_2})$ 为取两个 URL 字符串长度的最大值。W_{url} 值越大，则说明两个 URL 所链接的网页属于同一主题的概率也越大。

4 实验结果及分析

为了验证所提出的特征模型的有效性，文中通过对基于 Best First Search 算法和基于 PageRank 算法的爬行结果与基于特征模型方法的爬行结果进行对比分析，验证了文中建立的特征模型的有效性。

实验采用中国石油大庆油田部分下属单位的信息门户网站作为实验使用的网页数据集，包括 5 个单位的网站超过 7 万个网页实例。实验首先以石油百科中钻井工程为主题 300 多个网页作为主题训练样本，将其特征词提取，经过审核评价后得到 180 个常用的主题特征词，建立主题特征词库，并根据词在主题中的重要程度赋予不同的权重。

文中拟通过大量实验对比对选择一组最优的网页特征权值，具体方法如下：从网页数据集中随机挑选 1000 个网页作为样本，对各特征的权值以 0.50 为初值，分别在[0,1]范围内以步长 0.02 进行取值，取阈值 r 为 0.30，并对实验结果的前面 100 条记录和最后 100 条记录进行人为分析。通过实验验证发现，当各特征权值如表 2 时，实验结果比较符合人的主观判断。

表 2 网页特征权重取值

特征	权重
网页标题	0.42
网页描述文本	0.36
正文标题	0.64
强调性文本	0.56
网页正文文本	0.48
锚文本	0.58

对上述这组特征权重值进行分析：大小标题、强调性文本以及锚文本对主题的表现力较强；而网页标题、网页描述文本、网页正文对主题的表现力较差。观察网页的特征及 HTML 源文件发现，网页的大小标题、强调性文本和锚文本通常含有主题特征词，是网页要表达的概念或内容；而大量的网页标题非常简短，没有很好描述网页的内容，甚至许多网页的标题是编辑网页时自动生成的，另外是同一个站点下的网页，往往使用相同的标题，且该标题不含主题特征词，导致网页标题的主题表现力差。因此，上述的特征权重值是合理的。

最后，对同样的种子 URL 分别用上述三种算法采集数据，并用相同的查询词分别查询 10 次，分析前 100 条查询结果的主题相关度，实验结果如表 3 所示。

表 3 实验结果对照

测试算法	主题查准率(%)	资源查全率(%)
Best First Search 算法	28	47
PageRank 算法	39	66
基于特征模型方法	72	78

从表 3 中可以看出，基于特征模型方法在主题查准率和资源查全率比 Best First Search 算法、PageRank 算法有明显的优势，更有助于提高主题搜索引擎的搜索准确率。

5 结束语

网页信息与主题的相关度分析是主题搜索引擎的关键组成部分，文中针对已有算法在主题资源发现率和搜索准确率方面的不足，提出的利用网页特征及特征之间的关系来建立网页特征模型，并用特征模型来表达网页主题的方法，充分考虑了网页的内部特征和外部特征对网页主题的不同作用，有效地提高主题搜索引擎的资源发现率和搜索准确率。在实际应用中，特征模型的参数选择需要结合实际网页和通过大量实验来选取最优值。

参考文献：

[1] Phalp K T, Henderson P, Walters R J, et al. RolEnact: role-based enact able models of business processes[J]. Information and Software Technology, 1998, 40: 123-133.

[2] 刘金红，陆余良. 主题网络爬虫研究综述[J]. 计算机应用研究，2007，24(10)：26-29.

[3] Cho J, Garcia-Molina H, Page L. Efficient crawling through URL ordering[C]//Proc of the Seventh World-Wide Web Conference. New York: ACM, 1998: 161-172.

[4] Hersonvici M, Jacovi M, Yells S. The shark-search algorithm-An application tailored web site mapping[C]//Proc of the Seventh World-Wide Web Conference. New York: ACM, 1998: 317-326.

[5] Brin S, Page L. The anatomy of a large-scale hypertexual web search engine[C]//Proc of the 7th World-Wide Web Conference. New York: ACM, 1998.

[6] Kleinberg J. Authoritative sources in a hyperlinked environment[C]//Proc of the 9th ACM-SIAM Symposium on Discrete Algorithms. New York: ACM, 1998: 668-677.

[7] 罗林波，陈 绮，吴清秀. 基于 Shark-Search 和 Hits 算法的主题爬虫研究[J]. 计算机技术与发展，2010，20(11)：76-79.

[8] 姜 琨. 主题搜索引擎中的爬取技术研究[D]. 长沙：国防科技大学，2011.

[9] 黄萱菁，夏迎炬，吴立德. 基于向量空间模型的文本过滤系统[J]. 软件学报，2003，14(3)：435-442.

(下转第 95 页)

致的,但更新的处理步骤比重建要繁琐。而后两者的时间耗费逐渐接近,随着事务数据库越来越大,更新的优势越来越明显。这是因为总的事务数量不断增多,重建 TD-FP-Tree 时要两次扫描大量事务,时间耗费也越来越大,而每次新增事务集的数量则是固定的,时间耗费也趋于稳定。

实验二设置最小支持度为 80%,原始事务数据库 2000 个事务,新增事务集 500 个事务。实验二模拟实际应用中增量挖掘的全过程,比较的是基于不同 Tree 结构的效率,前者是 PFU-TD-FP 同 TD-FP-Growth 的组合,后者是 FUFP 同 FP-Growth 的组合。最后的执行时间包含三部分,一是原始数据库挖掘时间,二是更新 Tree 结构时间,三是更新后数据库挖掘时间。图 4 显示了这些时间。

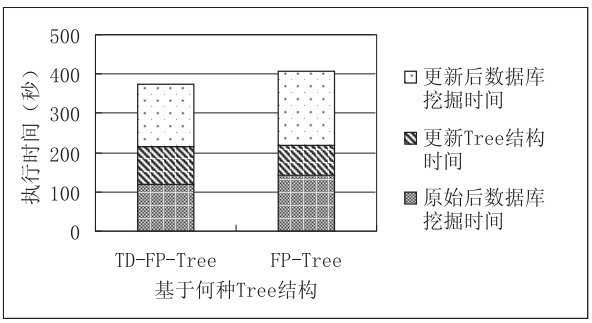

图 4　基于不同 Tree 结构的增量挖掘比较

从图 4 中观察到在更新 Tree 结构时 TD-FP-Tree 可能花费更多的时间,这是因为两者处理 Rescan_Transactions 的策略不同,文中提出的 PFU-TD-FP 算法需要先删除原 Tree 中表示相同事务的路径,而 FUFP 则只要在 Tree 的末端添加新节点即可。但是由于 TD-FP-Growth 的挖掘速度更快,所以基于 TD-FP-Tree 的增量挖掘在总时间上比基于 FP-Tree 的要少。

4　结束语

文中基于 FUP 的思想提出快速更新 TD-FP-Tree 结构的算法,根据项在原始数据库和新增事务集中是否频繁将其分类处理,可以减少重新扫描原始数据库的次数,并减少重排序事务中项时依赖项的个数,从而达到快速更新 TD-FP-Tree 的目的。文中针对 TD-FP-Tree 的结构特点提出了可行的更新方案,并采用并行处理的方法进一步提高效率。实验表明,文中提出的算法不仅可以快速更新 TD-FP-Tree,而且在同基于 FP-Tree 结构的增量挖掘中也有更好的表现。今后将继续在 TD-FP-Tree 快速更新算法上的研究,研究方向是将其应用到实际系统中,如个性化推荐、元搜索引擎等。

参考文献:

[1] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Dallas, TX: [s. n.], 2000: 1-12.

[2] Wang K, Tang L, Han J, et al. Top Down FP-Growth for Association Rule Mining[C]//Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Taipei, Taiwan: [s. n.], 2002: 334-340.

[3] Cheung D W, Han J, Ng V T, et al. Maintenance of discovered association rules in large databases: An incremental updating approach[C]//The Twelfth IEEE International Conference on Data Engineering. New Orleans, La: [s. n.], 1996: 106-114.

[4] 杨学兵,安红梅. 一种高效的关联规则增量式更新算法[J]. 计算机技术与发展,2007,17(1):108-113.

[5] Hong T P, Lin J W, Wu Y L. A fast updated frequent pattern tree[C]//The IEEE international conference on systems, man, and cybernetics. Taiwan: [s. n.], 2006: 2167-2172.

[6] Lin C W, Hong T P, Lu W H. The Pre-FUFP Algorithm for Incremental Mining[J]. Expert Systems with Applications, 2009, 36(5): 9498-9505.

[7] Li C X, Zhao L. Improved Incremental Mining Algorithm[J]. Computer Engineering, 2010(24): 42-44.

[8] Zou H, Zhu S H. Research on Incremental Mining Algorithm Based on HFUFP-Tree[J]. Computer Applications and Software, 2011(9): 102-105.

[9] 邹　海,朱四红. 基于 HFUFP-tree 的增量挖掘算法研究[D]. 合肥:安徽大学,2010.

[10] 钱　峰,张　蕾. 一种以 FP-tree 为基础的增量式挖掘算法的研究[J]. 科技信息,2009(24):386-389.

[11] Leung C K, Khan Q I, Li Z, et al. CanTree: a canonical-order tree for incremental frequent-pattern mining[J]. Knowledge and Information Systems, 2007, 11(3): 287-311.

[12] 邹力鹍,张其善. 基于 CAN-树的高效关联规则增量挖掘算法[J]. 计算机工程,2008(3):29-31.

(上接第 90 页)

[10] 陈飞宏. 基于向量空间模型的中文文本相似度算法研究[D]. 成都:电子科技大学,2011.

[11] 周　博,刘奕群,张　敏,等. 锚文本检索有效性分析[J]. 软件学报,2010,22(8):1714-1724.

[12] Eiron N, Ks M. Analysis of anchor text for Web search[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada: [s. n.], 2003.

主题搜索引擎中特征模型技术的研究

作者： 文必龙， 唐苏龙， 张浩
作者单位： 文必龙,唐苏龙(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)， 张浩(天津理工大学 计算机科学与技术学院,天津 300191)

刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(4)