基于 ODP 的上下文主题描述方法

吴 麒,陈兴蜀,朱 锴,王春晖
(四川大学计算机学院网络与可信计算研究所,四川成都 610065)

摘 要: 针对以往主题描述方法未充分考虑主题上下文的问题,提出了基于 ODP(开放式分类目录)的上下文主题描述方法.使用新的特征选择算法对主题特征进行了确定,并使用分类主题树的上下文对主题描述方法进行优化以提高主题爬行的性能.实验表明,该特征选择算法能够有效地提取出主题特征,并在保证正确率的基础上尽量减少特征维数以提高计算效率.同时,该主题描述算法充分考虑了主题上下文关系,且无论是在准确性还是在信息量总和上都有良好的性能.

关键词: 主题爬行; 下文相关; 特征选择; 主题描述
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2012)11-2320-04
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.11.028

A Novel Context-Sensitive Topic Description Method Based on ODP

WU Qi, CHEN Xing-shu, ZHU Kai, WANG Chun-hui
(1. *Network and Trusted Computing Institute, School of Computer Science, Sichuan University, Chengdu, Sichuan* 610065, *China*)

Abstract: It becomes an increasing important task to discover desirable information from tremendous web resources, and therefore the more accurately the focused crawler describes the users' interested topics, the more useful the information obtained by it will be. However, the neglect of context leads traditional approaches to the failure of achieving this goal. In order to overcome this shortcoming, a new context-sensitive topic description method based on ODP (Open Directory Project) is proposed in this paper. First of all, the best topic feature subset is determined by a new feature selection algorithm. And then, we optimize the topic description method to improve the performance by utilizing topic context. The experimental results show that this new approach extracts the features effectively with a better performance in both the precision and the sum of information while minimizing its dimension size significantly.

Key words: topical crawling; context-sensitive; feature selection; topic description

1 引言

随着互联网的迅速发展,Web 上的信息呈爆炸式地增长.面对如此海量的数据,在 Web 页面上发现有价值的信息就成为了一项重要的任务,于是主题爬虫应运而生.主题爬虫是由主题驱动的爬行程序,其目的是去获取与预先定义的主题相关的网页.由此可见,主题描述将决定主题爬虫的爬行内容.因此,如何准确地描述主题也就成为了主题爬行中的一个重要难题.

基于关键字的主题描述方法(TD_KW)使用一系列互相独立的词语组成的集合来对主题进行描述.文献[1]通过预先指定少量关键字的方式来定义主题.Niran 等[2]从相关网页的标题和锚文本中抽取频繁出现的关键字作为主题描述.文献[3]使用词频-逆文档频率(TF-IDF)从相关文档中提取特征关键字作为主题描述,并以此计算链接相似度.上述方法通过人为指定关键字或者简单提取的方式定义主题,准确率不高.

基于分类法的主题描述方法是通过使用树形结构的分类文档来对主题进行描述.Soumen 等[4]通过下载 ODP 目录来建立主题树,并选择其中的某个叶子节点作为主题.文献[5]将 ODP 目录中的叶子节点所包含的页面信息作为主题的描述信息.文献[6]从 ODP 目录中选择主题,并将其中信息作为正例文本,然后再收集一些不相关的数据作为反例文本.Chen[7]提出了基于上下文主题关键字的描述方法(TD_CTKW).这些方法对上下文信息的利用过于简单,从而造成相关页面的遗漏.

针对上述研究存在的缺陷,本文在充分考虑主题上下文的基础上,提出了一种新的基于 ODP 的上下文主题描述方法.

2 基于 ODP 的分类主题树

ODP 最早被用于提供目录导航服务,其组织结构为

收稿日期:2011-11-28;修回日期:2012-06-12
基金项目:国家科技支撑计划课题(No.2012BAH18B05);新媒体资源管理关键技术研究及原理系统

一棵树."DMOZ"[8]是目前最大且最全面的人工编辑的分类目录.本文使用"DMOZ"的目录结构建立分类主题树,并对同一个父亲节点的孩子节点按照名称的字母序进行排序.图 1 描述了该分类主题树的组织结构.

主题树的形式化定义如下:

定义 1(层数)　层数为主题树中节点所在的层数,表示为 i,根节点处于 0 层.

定义 2(主题根)　主题根为主题树的根节点,表示为 N_{01}.

定义 3(主题节点)　主题节点为主题树中非叶子节点,表示为 N_{ij},其中 i 为该主题节点所处的层数(根节点为 0 层),j 为该节点处于其兄弟节点集合中的位置.每个主题节点代表一个主题.

定义 4(预设主题)　预设主题为用户感兴趣的主题节点,用以引导主题爬虫爬行,表示为 S.

定义 5(主题子树)　主题子树是以某个主题节点 N_{ij}为根,并与其所有后代节点构成的主题树,表示为 ST_{ij}.

定义 6(主题路径)　主题路径为从根节点 N_{01}到用户预设主题 N_{ij}之间的主题节点序列,表示为 N_{ij}. $Path = \{ N_{01} = P_0, N_{1a} = P_1, \cdots, N_{(i-1)z} = P_{(i-1)}, N_{ij} = P_i \}$.在图 1 中,预设主题 N_{32}的主题路径为实线所表示路径.

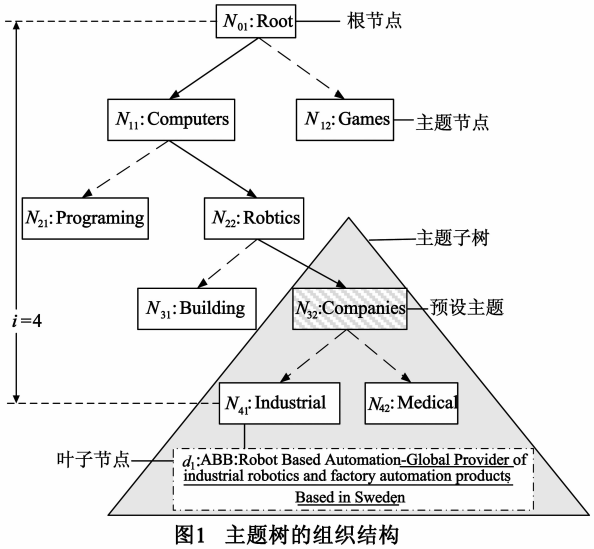

图1　主题树的组织结构

定义 7(主题文档集合)　主题文档集合是某棵主题子树 ST_{ij}所包含的叶子节点的集合,表示为 D_{ij}.该集合中的描述性文档以 d 表示,也就是说主题文档是对该主题进行描述的所有描述性文档的集合.

定义 8(主题特征)　主题特征是从某个主题节点 N_{ij}的主题文档中提取出的有区分度的特征词集合,表示为 $F_{ij}(k)$,其中 k 表示特征词的个数.

在实际应用中,通过下载 ODP 的 RDF[8]文件来构建分类主题树.

3　特征选择

在文本分类研究中,进行合理的特征选择[9,10]是十分重要的.文献[3]使用传统的词频-逆文档频率(TF-IDF)来对关键字的得分进行评估,并作为关键字的权重依据,从而完成特征提取.虽然 TF-IDF 计算简单,且易于使用,但其简单地将特征项在分类文档中出现的次数作为其重要度的依据,因此并不能将具有代表性的特征提取出来.

假设存在主题节点 $N_{(i-1)h}$,其儿子主题节点集合为$\{N_{i1}, N_{i2}, \cdots N_{i(l-1)}, N_{il}\}$,预设主题节点为 $S = N_{ij}(1 \leqslant j \leqslant l)$,其主题文档为集合 $D_{ij} = \{d_1, d_2, \cdots, d_n\}$,和预设主题节点处于同一个父亲节点下的任意一个兄弟主题节点被表示为 $N_{im}(1 \leqslant m \leqslant l, m \neq j)$,且其主题文档分别为集合 D_{im},主题节点 N_{ij}中的其中一个特征项为 t,那么 t 的区分度得分(Disc)定义如下:

$$\mathrm{Disc}(t) = \frac{\frac{1}{l-1}\sum_{D_{im}, m \neq j}(u_{D_{im},t} - u_{D_{ij},t})^2}{\frac{1}{n}\sum_{d \in D_{ij}}(x_{d,t} - u_{D_{ij},t})^2} \times u_{D_{ij},t} \tag{1}$$

$$u_{D_{ij},t} = \frac{1}{n}\sum_{d \in D_{ij}} x_{d,t} \tag{2}$$

其中,i 代表主题节点的层数,h,j,m 代表主题节点在其兄弟节点集合中的位置,l 代表 $N_{(i-1)h}$的儿子节点的个数,n 代表文档集合中文档个数,d 为 D_{ij}中的任意一个文档,$x_{d,t}$代表特征项 t 在文档 d 中出现的次数,而 $u_{D_{im},t}$与 $u_{D_{ij},t}$(公式 2)的计算方式类似.

使用公式(1)对预设主题节点 $S = N_{ij}(1 \leqslant j \leqslant l)$内部的特征项进行得分评估,并按分值从大到小选择 k 个特征词,并形成其主题特征 $F_{ij}(k)$.主题特征 $F_{ij}(k)$既要能包含大多数特征,也要能排除绝大多数的噪音数据.后面将通过实验对最优 k 值进行确定.

4　上下文相关的主题描述算法

针对以往主题描述方法未充分考虑主题上下文的问题,本文提出了一种上下文相关的主题描述算法(TD_CS).

以预设主题节点 $S = N_{ij}(1 \leqslant j \leqslant l)$为例,假设存在查询文本 q.在主题爬行中,主要就是根据查询文本对未爬行网页进行预测,也就是计算查询文本与预设主题的相关度.使用如下公式计算其查询文本相关度:

$$Pr(N_{ij} \mid q) = \frac{Pr(N_{ij})\,Pr(q \mid N_{ij})}{\sum_{c=1}^{l} Pr(N_{ic})\,Pr(q \mid N_{ic})} \tag{3}$$

$$Pr(N_{ij}) = \frac{\mid D_{ij} \mid}{\sum_{c=1}^{l} \mid D_{ic} \mid} \tag{4}$$

$|D_{ij}|$表示主题文档集合 D_{ij}中的文档数目.

基于朴素贝叶斯模型,可以得知:

$$Pr(q|N_{ij}) = (|q|!)\prod_{c=1}^{|V|}\frac{Pr(w_c|N_{ij})^{f_c}}{f_c!} \tag{5}$$

其中$|V|$表示查询文本 q 中不同词汇的个数,$|q|$表示查询文本中词汇的个数,w_c 表示词汇集中的单词. f_c 表示单词 w_c 在查询文本 q 中出现的次数.

$$Pr(w_c|N_{ij}) = \frac{1 + w_{c,D_{ij}}}{|V_{F_{ij}(k)}| + \sum_{b=1}^{|V_{F_{ij}(k)}|} w_{b,D_{ij}}} \tag{6}$$

公式(6)表示词汇 w_c 在主题文档集合中出现的概率.其中$|V_{F_{ij}(k)}|$表示主题特征 $F_{ij}(k)$中的特征词数,也就是 k. $w_{c,D_{ij}}$表示主题文档集合 D_{ij}中词汇 w_c 出现的次数.为了解决不经常出现的单词的零概率问题,在公式(6)中使用了拉普拉斯平滑技术.

结合提出的特征选择技术,公式(6)中词汇 w_c 必须为主题特征 $F_{ij}(k)$中出现的词汇.在公式(6)中,如果词汇 $w_c \in F_{ij}(k)$,那么 $w_{c,D_{ij}}$表示主题文档集合 D_{ij}中词汇 w_c 出现的次数;如果 $w_c \notin F_{ij}(k)$,那么 $w_{c,D_{ij}}$将等于零.

为了在主题爬行中覆盖更多的相关网页,将主题树中节点的上下文关系引入描述算法中.假设其预设主题 $S = N_{ij}(1 \leqslant j \leqslant l)$,其主题路径 $N_{ij}.Path = \{N_{01} = P_0, N_{1a} = P_1, \cdots, N_{(i-1)z} = P_{(i-1)}, N_{ij} = P_i\}$,那么查询文本 q 与预设主题间的相关度得分为:

$$\mathrm{Score}(q) = \frac{\sum_{c=1}^{c=i}\alpha_c Pr(P_c|q)}{\sum_{c=1}^{c=i}\alpha_c} \tag{7}$$

其中 $\alpha_c = \frac{1}{i-c+1}$,表示预设主题节点 N_{ij}与路径中的主题节点 P_c 之间层数距离,也代表主题节点 P_c 所占权重,a 到 z 为各个主题节点在其兄弟节点中的位置.

5 性能评估

5.1 主题特征 K 值的确定

为确定最合理的特征数 K,以达到较优的性能,从 ODP 目录下选择 Science 分类下的"Science/Astronomy"、"Science/Biology"、"Science/Computer Science"、"Science/Physics"四个分类以及 Society 分类下的"Society/Military"、"Society/People"、"Society/Social Science"、"Society/History"四个分类进行文本分类性能测试.在测试时,随机在各个子分类的 1000 个文档中选取 600 个作为该分类训练数据集,剩余 400 个文档作为测试数据集.

图 2 和图 3 描述了文本分类的准确率与 K 值的关系.对于 Disc 算法,分类准确率随着 K 值的增加,呈现出轻微的"抖动",这是由于"过拟合"引起的.在图 2 中,当 K 值为 180 时,分类准确率出现了一个"小峰值",为 91.3%.随着 K 值的增加,分类准确率约有下降,然后再次上升并在 400 达到峰值,但与 K 值为 180 时相比,上升幅度很小.在图 3 中,当 K 值为 175 时,准确率到达峰值,以后随着 K 值的增加,准确率反而下降.综上所述,设定最佳 K 值为 175.

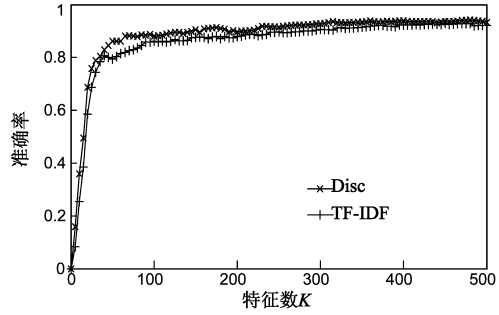

图2 "Science"子类特征选择性能

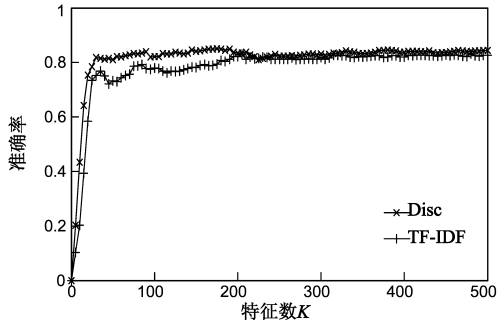

图3 "Society"子类特征选择性能

5.2 主题描述算法性能

在主题爬行过程中,使用链接锚文本与预设主题的相关度来对链接进行排序.本文利用相关页面数和信息量总和[7]来对算法的性能进行测试.

本文对上下文相关的主题描述算法(TD_CS,公式(7))以及基于上下文主题词的描述算法(TD_CTKW)[7]进行了性能对比测试.在实验过程中,对从分类主题树种选取了 2 个不同的主题(Science/Astronomy, Arts/Music/Instruments/Winds),其种子 URLs 见表 1 所示,并且设置爬行深度为 3,爬行最大 URL 数为 5000.

图 4 和图 5 描述了 TD_CTKW 和 TD_CS 算法在两个测试主题上的平均性能.在爬行初期,TD_CTKW 考虑的上下文信息比 TD_CS 少,也就是说 TD_CTKW 在主题范围的选择上更依赖所选择的主题本身,因此其能在初期获取更多的相关页面,而信息量总和也更大.随着爬行页面数的增加,TD_CS 由于充分考虑了主题的上下文关系,因此能更容易从看似"不相关的页面"中发现出相关页面,所以其获取的相关网页数和信息量总和增长更快.因此从总体上看,TD_CS 比 TD_

CTKW 的算法性能更加优越.

表 1　种子 URLs

主题	种子 URLs
Science/Astronomy	http://astronomyonline.org/
	http://en.citizendium.org/wiki/Astronomy
	http://www.astronomytoday.com/
	http://www.astronomy.net/
Arts/Music/Instruments/Winds	http://www.windplayer.com/
	http://www.contrabass.com/
	http://www.musiciansnews.com/windbrass/

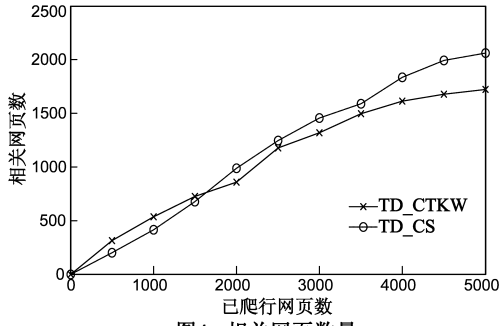

图4　相关网页数量

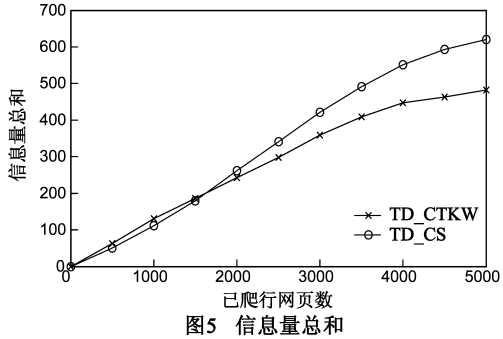

图5　信息量总和

6　结论

主题爬行是最近几年的研究热点,对于垂直搜索引擎等具有重要意义.本文通过对分类主题树的形式化定义,并在充分考虑上下文的条件下,提出了基于 ODP 的主题描述方法.测试结果表明,本文提出的方法在相关网页数量和信息量总和这两项性能指标上有较大地提高.

参考文献

[1] AHMED P, NIKITA S. Application of structured document parsing to focused web crawling[J]. Computer Standards & Interfaces, 2011, 33(3): 325 – 331.

[2] NIRAN A, ARNON R. Learnable crawling: an efficient approach to topic-specific web resource discovery[A]. The 2002 International Symposium on Communication and Information Technology[C]. Songkla: IEEE Press, 2002. 1034 – 2038.

[3] WANG C, GUAN Z Y, CHEN C, et al. On-line topical importance estimation: an effective focused crawling algorithm combining link and content analysis[J]. Journal of Zhejiang University SCIENCE A, 2009, 10(8): 1114 – 1124.

[4] SOUMEN C, KUNAL P, MALLEL A S. Accelerated focused crawling through online relevance feedback[A]. The 7th International World Wide Web Conference[C]. Brisbane: Elsevier Science B.V. Press, 2002. 336 – 348.

[5] LIU H Y, JEANNETTE J, EVANGELOS M. Using HMM to learn user browsing patterns for focused Web crawling[J]. Data & Knowledge Engineering, 2006, 59(2): 270 – 291.

[6] SELMA A. A Web page classification systems based on a genetic algorithm using tagged-terms as features[J]. Expert Systems with Application. 2011, 38(4): 3407 – 3415.

[7] CHEN Z M, MA J, LEI J S, et al. A cross-language focused crawling algorithm based multiple relevance prediction strategies[J]. Computers and Mathematics with Applications. 2009, 57(6): 1057 – 1072.

[8] NETSCAPE. Open Directory Project[EB/OL]. http://www.dmoz.org/, 2011-5-18.

[9] 刘桃,等.领域术语自动抽取及其在文本分类中的应用[J].电子学报,2007,35(2):328 – 332.
Liu Tao, et al. Automatic domain-specific extraction and its application in text classification [J]. Acta Electronica Sinica, 2007, 35(2): 328 – 332. (in Chinese)

[10] 戴新宇,等.一种基于潜在语义分析和直推式谱图算法的文本分类方法 LSASGT.电子学报,2005,36(8):1626 – 1630.
Dai Xin-yu, et al. LSASGT: an approach to text categorization based on latent semantic analysis and spectral graph transducer [J]. Acta Electronica Sinica, 2005, 36(8): 1626 – 1630. (in Chinese)

作者简介

吴　麒　男,1985 年生于四川,四川大学计算机学院博士研究生.研究方向为信息安全与 Web 数据挖掘.

E-mail: ocean_shark@sina.com.cn

陈兴蜀(通讯作者)　女,1968 年生于四川,四川大学计算机学院教授、博士生导师.研究方向为信息安全.

E-mail: chenxsh@scu.edu.cn