基于大众分类法的中文博客分类方法

Chinese Weblog Pages Classification Based on Folksonomy

丛 鲁 丽

(山东理工大学图书馆 淄博 255049)

摘 要 大众分类法(Folksonomy)的灵活性比传统的机器分类方法高,但是它不能处理大量的数据资源。为了解决这个问题,结合 Folksonomy 和传统机器学习算法的优点我们提出了一种新的算法 FSVMC (Folisonomy and Support Vector Machine Classifier)。在 FSVMC 中,支持向量机算法作为一个 TAG 代理,决定一个标签是否应该标注在某个资源上,而 Folksonomy 致力于网页文档的分类。此外还提出了一种创建可以标注网页标签数据库的方法。实验结果表明我们的方法比传统的机器学习方法更加有效和具有柔性。

关键词 文本分类 Folksonomy 支持向量机 标签代理 博客

中图分类号 G350 **文献标识码** A **文章编号** 1002 - 1965(2009)09 - 0050 - 03

1 背 景

根据 Goolge 网站提供的数据,目前 Google 目录中收录了 80 多亿个网址,而且每天以大约 150 万的速度增长。海量的数据,海量的用户。用户如何能找到自己所需的资料成为问题的关键。自动分类技术部分解决了这个问题,而且拥有一部分成熟的算法。如决策树、KNN、贝叶斯、C4.5 Quinlan、SVM 等[2~5]。特别是 Vapnik 在 1995 年提出的支持向量机算法[1],是目前最具有代表性的模式分类算法之一。

自动分类算法一个最主要的问题是很难及时、准确地对大规模数据进行分类,因为需要大量的时间和存储空间来对这些数据进行分类处理;另一个问题是自动分类算法很难从用户那边得到反馈。这样普通用户很难理解一些由专家预先定义好的类。Folksonomy 可以解决这些问题,但是分类精度不高。为了解决以上问题,我们提出了一种基于 Folksonomy 和 SVM 的混合分类算法 FSVMC(Folisonomy and Support Vector Machine Classifier)。

2 大众分类法(Folksonomy)

Folksonomy 是一个创造词,是由 Folk(或 Folks)与 Taxonomy 组合而成,Folks 在英文中是比较口语化的词,表示一群人、一伙人的意思。Taxonomy 则是指分类法,是信息组织中的一个重要组成部分。Folksonomy 的字面含义就是“一伙人的分类法”。Folksonomy 是由社会性书签服务中最具特色的自定义标签(Tag)功能衍生而来,举个例子:当一个博客在收藏 Sina.com 时,自定义了“门户”、“中国”、“新闻”这 3 个关键词作为标签,而其他人在收藏 Sina.com 时也自定义了自己的关键词作为标签,例如“中国”、“新闻”、“网站”。最后系统统计出使用“门户”、“中国”、“新闻”这个 3 个关键词来定义 Sina.com 的频率最高,那么这 3 个词就是用户对 Sina.com 的 Folksonomy 分类。由此可见,Folksonomy 就是由网络信息用户自发为某类信息定义一组标签进行描述,并最终根据标签被使用的频次选用高频标签作为

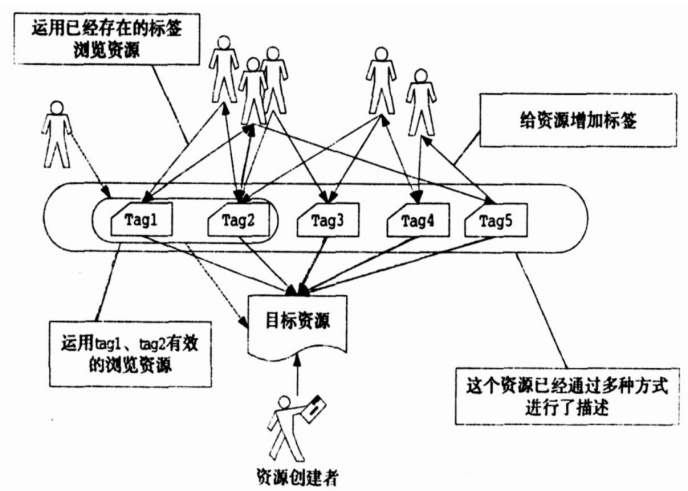

图 1 大众分类法

收稿日期:2008 - 12 - 25 修回日期:2009 - 05 - 25

作者简介:丛鲁丽(1962 -),女,副研究馆员,研究方向为数字图书馆。

该类信息类名的一种为网络信息分类的方法。其实质就是以词为类,但其类目却是平面的、非等级的。

自由分类法是对网络信息分类的一种新尝试,突破了传统的类目设置,充分体现了网络用户的信息需求特点。当然,自由分类法并非完美无缺,在对网络信息标引与检索方面同样存在着一些缺点:

a. 自由分类法适用的范围有限。自由分类法的类是由博客们创造的。然而却也并非全部博客网都提供为文章做标签并用标签进行检索的功能。即使提供了用标签进行检索的博客网站,检索到的也只是博客空间的信息,甚至只是该博客网的信息。这些信息只是网络信息的一小部分而已。

b. 作为自由分类的类目 Tag 缺乏控制。个人使用标签不够规范和统一,加上有许多同义、近义词的存在,使得同类信息被分散开来,例如:"十一"和"国庆节"、"超级女声"和"超女"等。这种由于 Tag 缺乏控制造成的信息分散也必然影响了信息的检全率。

c. 类目的平面非等级显示同样会隐藏重要信息不便浏览。自由分类不具有等级结构,不存在根结点,标识信息的 Tag 或者是字顺显示或者是随机罗列在页面上,尽管重要的、点击频次高的 Tag 通过特殊颜色或字体等被突出显示,仍难免被浩如烟海的信息所淹没。

d. 使用 Tag 检索的检索结果并不十分理想。

3　基于大众分类法的中文 Weblog 分类方法

3.1　**基本思想**　为了解决传统机器学习分类方法柔性不高,而大众分类法分类精度不高的问题。我们提出了基于大众分类法和支持向量机的中文 Weblog 分类方法。首先建立了待选标签数据库,然后采用支持向量机的方法作为标签代理,给需要分类的网页添加标签。

下面将对我们提出的算法进行详细的描述:

3.2　**预处理**　此部分主要是对 html 页面进行去燥、分词处理。我们采用中科院的开源中文分词系统 ICTCLAS 进行分词。仅选择切词结果中部分类型的关键词为候选特征项。比如仅保留名词、动词、形容词等词性的实词。

3.3　**创建候选标签数据库**　我们从 Blog 分类名称以及标签中选择出候选标签。这是因为大多数 Blog 系统允许用户自己进行分类和添加标签。需要注意的是这些分类和标签是根据用户的喜好和自己词汇的水平进行添加的。所以考虑到这些因素,我们在选择候选标签的时候应该注意标签的流行程度和对事物的描述能力,以及能够反映出新的分类和标签。

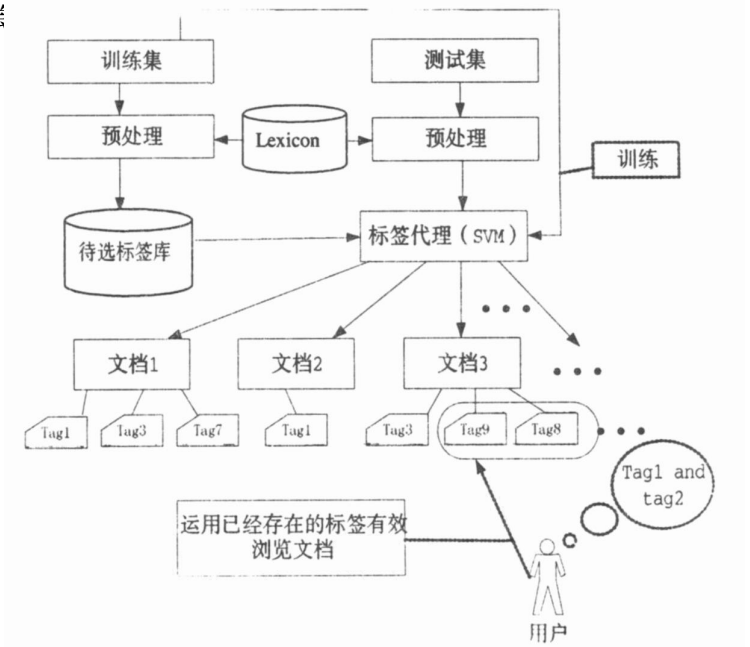

图 2　FSVMC

首先,我们通过有多少个 Blog 网站和 Bloger 包含同一个标签来判断一个标签的流行程度。也就是说,如果一个标签非常流行那么就会有很多人都在用它。我们将设定一个阈值来判定。其次,我们从选出的流行标签中计算它的描述能力。一个标签如果很流行但是他的描述能力差的话,也不是我们所要的标签。我们为每一个标签建立了一个 SVM 分类器,如果 SVM 分类精度不高的话,则说明这个标签描述能力不高。在决定一个标签的描述能力中有两个因素需要考虑:一个是在候选标签中不合适的标签要尽量少;另一个是在判断过程中运用的数据量要尽可能的少。为了满足这两个条件,我们应当在小数据量的条件下得出合适的分类精度。然后我们可以认为这个合适的分类精度是一个概率事件,可以定义 p 为这个分类的合适精度,这个事件的概率为 p。对这样的分布由于我们没有先验经验,我们自然地认为它满足二项分布。通过中心极限定理,可以得到,如果 n 不是太小的话这个分布可以近似地看作是正态分布。这样我们可以计算置信区间。如果我们设定其值为 99.5 %,那么置信区间的下界可以通过以下公式计算出来:

当置信区间的下界值超过 80 %(这个阈值是在 5.2中选出的)时,我们就认为这个标签是我们所需要的。

3.4　**创建标签代理**　标签代理是一个标签分类器阵列,由它来决定是否将一个标签添加到某篇文章(如图 3 所示)。

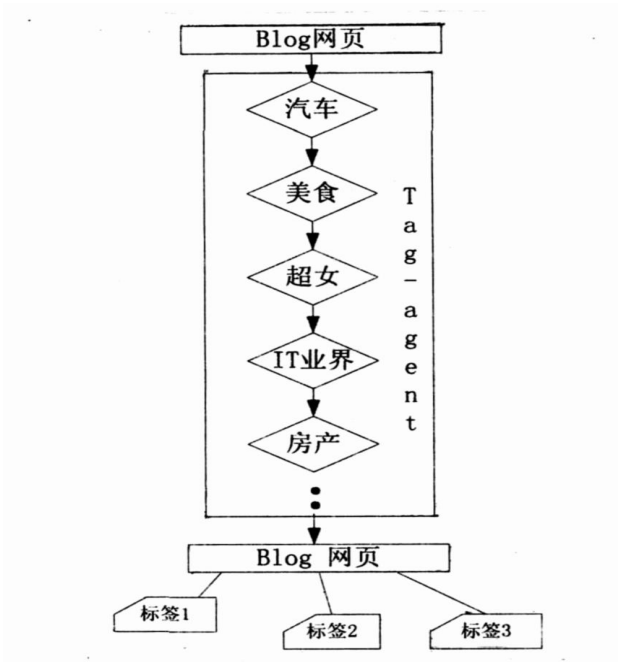

图 3 标签代理

标签分类器需要周期的进行训练,可以从训练数据中选择最新的数据来进行训练,这样可以获得互联网上一些新的分类和词汇。为了达到每篇文章有多个标签的效果,标签分类器应满足以下几个方面的要求:分类速度要快;消耗空间(内存和硬盘)要少;避免过度拟和情况的发生;要有高的分类精度。综合以上几个因素,我们比较了目前几种流行的机器分类算法,如 K - NN、Naive Bayes、决策树、SVM 等,最终选定 SVM 作为标签分类器的分类算法。

为了训练每一个标签分类器,我们从训练数据集中选择最近的文章来进行训练。作者们认为是"A"分类的文章作为正数据集,为了加快训练速度,如果这些文章超过 2 000 篇的话我们只取 2 000 篇。当然如果数据量越大训练效果越好,但是时间消耗也大。然后从其它不是"A"分类的文章中随机选取 2 000 篇,作为负数据集。为了处理有些分类名称和词汇改变比较快的情况,我们需要定期地对标签分类器进行训练,如一天一次。

4 实 验

4.1 **数据集** 由于 Blog 文章都有作者自己加的标签和分类名称,我们从互联网上选择真实的数据。我们做了一个爬行器从新浪(http://blog.sina.com.cn)上下载了 14 910 篇文章,从搜狐上下载了(http://blog.sohu.com/)74 339 392 篇文章。为了体现分类器的通用性,我们将这两个数据集分别作为训练集和测试集。有些固定的分类如新浪里面的"我的所有文章"就会把它们排除在我们的训练集外,因为这样的分类名称几乎在每一个 Blog 中都会存在。

4.2 **分类精度** 在这个实验中,随着数据量的增加分类精度也随着增加。图 4 为一部分实验数据,图中可以明显地看出在分类精度上有一个分界线大概为 0.8,当低于 0.8 时,如日记、日常生活等虽然用的人很多,但不能很好地表现文章内容。相反一些别的标签如旅游、音乐等就能很好表现文章的内容,也能得到高的分类精度。

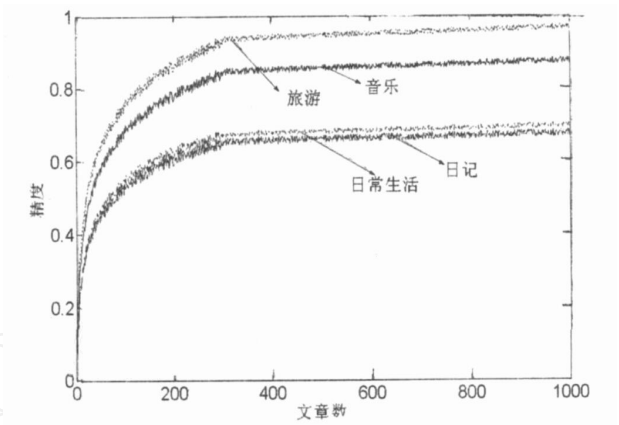

图 4 部分标签分类精度

4.3 **创建标签库** 采用上面所提到的程序来选择标签,然后与手工选择的一些标签进行比较。结果如表 1 所示。结果显示的分类的标签选择算法是有效的,即使当数据量不大,比如只有 100 篇文章的时候分类精度还是相当高的。

表 1 候选标签选择评价

文章数	*tp*	*fp*	*fn*	*tn*	正确率	精度
> = 100	91	3	113	135	66.0	96.8
> = 200	76	3	67	107	72.3	96.2
> = 400	41	2	24	62	79.8	95.3
> = 800	27	1	5	34	91.0	96.4
> = 1600	11	1	0	16	96.4	91.6

tp 表示机器和人都认为可以作为候选标签数;*fp* 表示机器认为可以作为候选标签但人认为不可以作为候选标签数;*fn* 表示人认为可以作为候选标签但机器认为不可以作为候选标签数;*tn* 表示机器和人都认为不可以作为候选标签数;正确率为 $(tp + tn)/(tp + tn + fp + fn)$;精度为 $tp/(tp + fp)$。

4.4 **标签代理** 根据上面所述,部分实验结果如图 5 所示。从图中我们可以看出大部分的标签分类器的分类精度很高,但是有部分分类器的召回率却很低。但是大部分用户是根据我们给他们的分类去浏览自己想要的内容,即使召回率不高,但是分类精度高也可以是用户满意。

4.5 **实验环境** 我们的硬件试验环境为至强 3.0G

(下转第 40 页)

7 Law J, Whittaker J. Mapping Acidification Research: a Test of the co - word Method[J]. Scientometrics, 1992, 23(3): 417 - 461

8 Neal Coulter, Ira Monarch, Suresh Konda. Software Engineering as Seen Through Its Research Literature: a Study in co - word Analysis[J]. Journal of the American Society for Information Science, 1998, 49(13): 1206 - 1223

9 Coulter N, Monarch I, Konda S, et al. Ada and the Evolution of Software Engineering[J]. In G. Engle (Ed.), Proceedings of TRI - Ada '95: Solutions for a Changing, Complex World, 1995: 56 - 71

10 Coulter N, Monarch I, Konda S, et al. An Evolutionary Perspective of Software Engineering Research Through co - word Analysis [J](Tech. Rep. No. CMU/SEI - 96 - TR - 019). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 1996

11 Law, et al. Policy and the Mapping of Scientific Change: a co - word Analysis of Research Into Environment Acidification[J]. Scientometrics, 1988, 14(3 - 4): 251 - 264

12 Qin He. Component Study of Co - word Analysis[D]. Urbana - Champaign: University of Illinois, 2001

13 冯 璐, 冷伏海. 共词分析方法理论进展[J]. 中国图书馆学报, 2006, 32(2): 88 - 92

14 Tomas Cahlik. Comparison of the Maps of Science[J]. Scientometrics, 2000, 49(3): 373, 387

15 杨立英. 科技论文共现理论研究与应用[D]. 北京: 中国科学院研究生院, 2007

16 Chaomei Chen. Mapping Scientific Frontiers: the Quest for Knowledge Visualization[M]. Springer - Verlag, 2003

17 Katy Börner. Cybertools that Support the Study of Science[C]. 在中科院文献情报中心讲座资料, 2008

18 Xiaoguang Wang, Feicheng Ma, Juncheng Wang, et al. The "small - world" Characteristic of Author co - words Network[EB/OL]. [2008 - 04 - 23]. http://ieeexplore.ieee.org/iel5/4339774/4339775/04340694.pdf

19 Kevin W. Boyack, Richard Klavans, Katy Börner. Mapping the Backbone of Science[J]. Scientometrics, 2005, 64(3): 351 - 374

20 Ying Ding, Gobinda G. Chowdhury, et al. Bibliography of Information Retrieval Research by Using co - word Analysis[J]. Information Processing & Management, 2000(37): 817 - 842

21 王建芳. 基于计量的科技知识演化关系分析方法研究[D]. 北京: 中国科学院研究生院, 2007

22 Henry Small. Tracking and Predicting Growth Areas in Science[J]. Scientometrics, 2006, 68(3): 595 - 610

23 祁建清. 基因算法在密码学中的应用分析[J]. 信息工程大学学报, 2003, 4(1): 93 - 95

24 秦春秀, 赵捧未, 刘怀亮. 词语相似度计算研究[J]. 情报理论与实践, 2007, 30(1): 105 - 108

25 章成敏, 鞠海燕. 基于混合策略的中文查询串相似度计算[J]. 情报杂志, 2005(11): 101 - 105

26 孙 爽, 章 勇. 一种基于语义相似度的文本聚类算法[J]. 南京航空航天大学学报, 2006, 38(6): 712 - 716

27 张晓孪, 王西锋. FCA 中的概念语义相似度计算[J]. 现代图书情报技术, 2007(3): 51 - 54

28 Sujit Bhattacharya, Prajit K. Basu. Mapping a Research Area at the Micro Level Using co - word Analysis[J]. Scientometrics, 1998, 43(3): 359 - 372

29 贺德方等. 数字时代情报学理论与实践 - 从信息服务走向知识服务[M]. 北京: 科学技术文献出版社, 2006

(责编: 王平军)

(上接第 52 页)

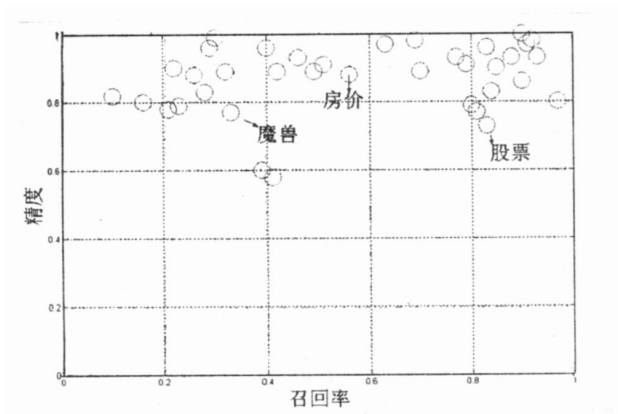

图 5 标签代理

的双 CPU, 2.0G 的内存, 1G 的网卡, 1TB 的存储容量。软件环境为 Linux 操作系统 Debian 版本, Apache 服务器, PHP 语言开发, PostgresSql 数据库服务器。

5 结 论

我们提出了一种基于 Folksonomy 和 SVM 的非常灵活的分类方法。实验证明, 通过我们提出的方法, 用户可以很容易地找到他们所感兴趣的信息。但是与人工添加标签的 Folksonomy 算法相比, 我们的算法还是不够灵活, 会给某篇文章添加太多的标签, 让用户难以决定。如何选择出更加符合文章内容的候选标签以及如何提高召回率是以后的研究重点。

参 考 文 献

1 Vapnik, V N. The Nature of Statistical Learning Theory[M]. New York: Springer - Verlag, 1995

2 Apte C, Damerau F, Weiss S M. Text Mining with Decision Trees and Decision Rule[A]. Proceeding of the Automated Learning and Discovery Conference. Carnegie - Mellon University, 1998: 99 - 103

3 Gunn S R. Support Vector Machines for Classification and Regression[R]. ISIS Technical Report. Image Speech and Intelligent Systems Group of University of Southampton, 1998: 31 - 36

4 Tan S. Neighbor - Weighted K - nearest Neighbor for Unbalanced text Corpus[J]. Expert Systems with Applications, 2005: 1 - 5

5 R E Schapire, Y Singer. Boostexter: A Boosting - Based System for Text Categorization[J]. Machine Learning, 2000: 135 - 168

(责编: 刘影梅)