

基于文本内容的超链接分类研究

陈丽, 于浩, 郑德权, 赵铁军

(哈尔滨工业大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001)

摘要:针对目前网页分类以及相关研究的问题,提出了基于文本内容的超链接分类思想,为下一步进行信息抽取、话题追踪等互联网信息应用研究做了更好的准备.通过对特定领域内应用两种分类方法对其进行对比研究,取得了较好的效果.

关键词:网页分类;超链接;链接分类;文本内容;互联网

中图分类号:TP393

文献标识码:A

文章编号:1672-0946(2004)02-0153-04

Study on hyperlink classification based on text content

CHEN Li, YU Hao, ZHENG De-quan, ZHAO Tie-jun

(School of Computer Science & Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Considering the problems of current Web pages classification and the relevant researches, this paper puts forward the thought of hyperlink classification based on text content. It makes preparations for information extraction, topic tracking and other internet information researches. It also presents two methods in a special field in order to comparison and research. And they have gained better results.

Key words: Web page classification; hyperlink; link classification; text content; internet

随着 Internet/Intranet 的飞速发展,如何有效利用 Web 数据成为一个重要课题.搜索引擎为在 Web 中查找所需信息提供了快捷手段.但搜索引擎在精度、易用性等方面仍存在总体性能差、检索的查准率和查全率不高等诸多问题,使得其效果不能令人满意^[1].另外,搜索引擎找到的材料太多,无法分辨哪些是真正需要的.基于这一情况有必要对网上资料进行进一步的分类、整理,以尽快找到用户所需.

网上信息分类可以为用户查找信息提供一个大致范围.因此,国内外对网上信息进行分类的研究方法很多,大多数是对文本进行分类,即对于每个检索到的电子文档,根据文本的标题信息或文本内容自动判断出它与系统规定的各个文本类别之间的相关性,从而为每个文本指派一个类别^[2],

还有少量方法是根据 meta 对文本进行分类.根据文本的标题信息和该文本的内容进行分类的准确率比较高,但是时空开销较大,运行速度相对较慢.而仅基于文本的标题或者 meta 对文本进行分类的时空开销比较小,运行速度快,但准确率低.这些文本分类只是一些粗略的分类,类别涵盖的范围比较大.

本文对 Web 页面进行了进一步的研究,大多数的 Web 网上的信息是以树状结构分布的,相关信息节点是通过超级链接进行联接,这些超链接大多数是与 Web 页面文档内容有关的一些信息.如果我们对相关链接进行分类,并根据分类结果决定是否查看超链接文档的内容,或者是否对超链接文档进行信息抽取以及进行其他方面的应用,就可以加速对信息的查找,快速的获取用户所需的信息.

收稿日期:2003-12-06.

基金项目:国家自然科学基金(60302021);富士通研究开发中心有限公司和黑龙江省教育厅项目(105310660).

作者简介:陈丽(1976-),女,硕士研究生,研究方向:网络信息处理.

本文即是基于这样的一种情况提出来的。

提出了一种针对网页中的超链接的分类。考虑到网页中主体是文本 text, 文档的下面是与这篇文档相关联的多个超链接, 我们的目的就是根据文档的内容对与其相关的超链接进行分类。

图 1 为一个 Web 页形式, 可以将其从“相关文章”处分为两部分, 上半部分为一文档信息, 下半部

分是与该文档内容相关联的超链接。本文将根据上半部分的文档标题及内容对下半部分的超链接进行分类。初步的实验结果表明, 依据本文提出的基于文本内容的超链接分类能大大提高分类的准确率, 相比仅根据标题信息进行分类的效果好。当然这样将以牺牲时空开销为代价。

500 万像素的高级家用数码相机出世(多图)

2002 - 09 - 18 11:06

CASIO 卡西欧公司今日正式发布了他们的 500 万像素数码相机产品 - QV5700。QV5700 是一款基于原 QV4000 机身设计的相机。QV5700 充分保留了 QV4000 的机身外形与控制设计, 并在此基础上将 CCD 的分辨率提高到了 500 万, 同时还增加了相机的感光度 ISO 的范围。QV5700 手动 ISO 可以从 ISO50 - ISO800 之间进行选择。此外 QV5700 采用了佳能高亮度的 F2.0 大光圈, 3 倍光学变焦镜头, 保证了摄影者对光线的各方面需求。

这款卡西欧 QV5700 数码相机的问世, 标志着新一轮家用数码相机升级的开始, 在此之前, 500 万像素家用数码相机市场, 只有尼康和索尼两家在各领风骚。这次卡西欧携 QV5700 的杀人, 必将让这本已平静的市场, 再起波澜。(新浪科技)

【责任编辑: 杜二】

【相关文章】

- [卡西欧新型数码相机 最小最薄可听 MP3\(图\)](#) 2002 - 07 - 15 16:25
- [带我去潜水 卡西欧日前推出防水数码相机](#) 2002 - 06 - 10 11:29
- [携文胸装相机登陆 卡西欧中国搅局?\(图\)](#) 2002 - 06 - 07 16:30
- [全球最薄的液晶数码相机 - 卡西欧 EXILIM](#) 2002 - 05 - 17 10:41
- [卡西欧超薄数码相机 Exilim 亮相\(图\)](#) 2002 - 03 - 18 11:02

图 1 网页样例

1 算法设计

由于网上的信息量巨大, 涉及范围广, 对各个领域文档的超链接进行分类难度非常大, 因此我们先讨论限定领域的超链接分类。

我们选定 IT 领域的新产品类文档, 通过对大量 IT 新产品类文档进行分析发现, 大多数的文档与其相关联的超链接中含有某公司某产品的名称, 我们可以抽出文档中介绍的产品名称及该产品所属的公司名称, 然后再分别抽出每一个超链接中的产品名称及所属公司的名称, 再用文档中抽出的产品名称和公司名称与每一个超链接中抽出的产品名称及公司名称进行比较, 从而确定每一个超链接所属的类别。

根据这类文档的特点, 我们定义了九个类别: 同公司同类产品、不同公司同类产品、同公司不同类产品、不同公司不同产品、同公司类、同产品类、不同公司类、不同产品类、其他类。具体定义如下。

1) 同公司同类产品

文档中抽取出的公司名称及产品名称与超链

接中抽取出的公司名称及产品名称相同。

2) 不同公司同类产品

文档中抽取出的公司名称与超链接中抽取出的公司名称不同, 而文档中抽取出的产品名称与超链接中抽取出的产品名称相同。

3) 同公司不同类产品

文档中抽取出的公司名称与超链接中抽取出的公司名称相同, 而文档中抽取出的产品名称与超链接中抽取出的产品名称不同。

4) 不同公司不同产品

文档中抽取出的公司名称及产品名称与超链接中抽取出的公司名称及产品名称都不相同。

5) 同公司类

文档中抽取出的公司名称与超链接中抽取出的公司名称相同, 并且文档中未找到产品名称或者超链接中未找到产品名称或均未找到。

6) 同产品类

文档中抽取出的产品名称与超链接中抽取出的产品名称相同, 并且文档中未找到公司名称或者超链接中未找到公司名称或均未找到。

7) 不同公司类

文档中抽取出的公司名称与超链接中抽取出的公司名称不同,并且文档中未找到产品名称或者超链接中未找到产品名称或均未找到。

8) 不同产品类

文档中抽取出的产品名称与超链接中抽取出的产品名称不同,并且文档中未找到公司名称或者超链接中未找到公司名称或均未找到。

9) 其他类

文档中未抽取出的公司名称及产品名称并且超链接中也未抽取出的公司名称及产品名称;文档中抽取出的公司名称而超链接中未抽取出的公司名称;或者文档中未抽取出的公司名称而超链接中抽取出的公司名称,并且文档中抽取出的产品名称而超链接中未抽取出的产品名称,或者文档中未抽取出的产品名称而超链接中抽取出的产品名称。

为此,我们建立两个简单的词表:产品名称词表 and 公司名称词表。产品名称词表收录了一些 IT 新产品,公司名称表收录了与产品相对应的公司名称。利用这两个词表可以抽取网页上文档中及与其相关的超链接中的公司名称和产品信息,然后进行比较分类。

1.1 文档特点

1) 通过对大量这类文档进行的研究,我们发现这类文档中想要介绍的产品名称及其公司名称多为文档中第一次出现的公司名称及产品名称。

2) 标题中出现的公司名称(产品名称)多为文档中想要介绍的产品名称及其公司名称。

3) 公司名称(产品名称)出现的频率对信息的抽取有一定的影响。

1.2 存在的问题

1) 在一篇文档中可能出现多个公司名称及多种产品名称,而且同一公司名称或者产品名称又可能出现多次。但单纯根据出现的次数的多少来确定这篇文档介绍的是某一公司的某种产品也会产生不正确的结果。

2) 公司名称有的用英文名称,或者有多个译音。如英特尔、因特尔、INTEL。

3) 产品名称有用英文字母表示的。如掌上电脑表示为 PDA。

1.3 解决方法

针对上述问题的 2、3,可以确定一个名称,以这个名称为标准,如遇其余名称则进行抽取转换,即确定英特尔为公司名称,如遇到因特尔或者 IN-TEL 则抽取出的公司名称仍为英特尔。确定产品

为掌上电脑,如遇 PDA 则抽取出的公司名称为掌上电脑。这就需要在建立公司名称词表和产品信息词表时,加入一些信息。如掌上电脑表示为 < 掌上电脑:PDA; >,即采用一种查询扩展的方法进行信息的抽取。这样可以减小由于词的不匹配现象导致的一些相同意义的信息不能被准确的抽取出来而降低下一步的分类结果的准确性^[3]。

1.4 具体实现流程

- 1) 建立公司名称词表与产品名称词表;
- 2) 读入将要分类的文本;
- 3) 进行匹配处理抽取公司名称及产品名称;
- 4) 读入超链接;
- 5) 进行匹配处理抽取超链接中的公司名称及产品名称;
- 6) 进行分类处理;
- 7) 返回 4,直到处理完与文本相关联的所有超链接;
- 8) 返回 2,直到处理完所有的文本。

1.5 具体分类算法

本文的文本分类研究主要采用向量空间模型方法,在向量空间模型中,文本被表示为一个高维向量,向量的每一维表示一个特征,通常是一个字或词,而其取值则是相应的权值。研究者在这个模型基础之上应用了多种分类技术,这些技术大多可归为统计识别方法和机器学习方法,比如, K 近邻方法、Bayes 方法、决策树方法、神经网络方法、符号学习算法等^[4-8]。

根据以上分析,本文设计了 2 种算法。

1) 首次出现算法

这是一种比较简单直观的算法,是根据这类文档的特点设计的。即将文档(包括标题)中第一次出现的公司名称(产品名称)作为我们选定的公司名称(产品名称),然后根据文本中和超链接中各自抽取出的公司名称(产品名称)进行分类。这种算法能够快速的抽取用户需要的信息,但抽取信息的准确率相对低一些。

2) 加权算法

这种算法是针对首次出现算法的缺点设计的,即将找到的公司名称(产品名称),根据公司名称(产品名称)在标题中是否出现以及在文档中出现的次序、频率信息,对该公司名称(产品名称)赋一个权值,然后根据权值大小确定一个公司名称(产品名称)为选定公司名称(产品名称),并根据选定公司名称(产品名称)进行分类,该算法的准确率会

极大提高,但算法的时空开销会比首次出现算法大。

2 实验结果及分析

我们在 <http://www.sohu.com>、<http://zaobao.com>、<http://www.sina.com>、<http://21cn.com>、<http://163.net> 等多个网站上,收集了 103 篇 IT 新产品类的文档,包含 702 个超链接,应用上述两种算法进行了测试分析,两种算法的准确率和召回率见表 1。

表 1 测试结果

算法名称	准确率	召回率
首次出现算法	95.89%	86.46%
加权算法	96.05%	86.61%

从表 1 的结果可以看出,加权算法的准确率和召回率稍好于首次出现算法的准确率和召回率,而首次出现算法的时空开销明显小于加权算法,在测试范围内大约为加权算法的 50%。针对图 1 所示的网页,利用以上两种算法进行的分类结果如图 2 所示。

图 2 所示的分类结果中,“携文胸装相机登陆卡西欧中国搅局?(图)”被分为同公司类,这是因为相机没有被收录到产品名称词表中,其余四个超链接分类正确。从以上分析中可以看出,公司名称词表中表示的公司名称及产品名称词表中表示的产品名称个数对分类的结果有很大的影响,由于公司名称词表中和产品名称词表中表示的公司名称和产品名称的数量有限,致使分类的准确率和召回率降低。

500 万像素的高级家用数码相机出世(多图).txt
 卡西欧新型数码相机 最小最薄可听 MP3(图)//同公司同类产品
 带我去潜水 卡西欧日前推出防水数码相机//同公司同类产品
 携文胸装相机登陆 卡西欧中国搅局?(图)//同公司类
 全球最薄的液晶数码相机——卡西欧 EXILIM//同公司同类产品
 卡西欧超薄数码相机 Exilim 亮相(图)//同公司同类产品

图 2 分类结果

3 结论

提出了基于文本内容的超链接分类方法,并进行了实验,取得了较好的结果,为在 Web 上快速、准确查找出所需要的信息打下了基础。下一步可以根据对超链接的分类结果进行信息抽取、话题追踪等互联网信息应用研究。文中提到的算法只要对分类以及词表略加改动即可应用在不同的领域。

另外,本文提到的两种算法还有待于进一步完善,对于链接指向的页面还需要进一步分析,以得出更为精确的结果,建立词表的方法上也有待改善。

参考文献:

[1] 张开舟. 万维网信息检索系统开发技术[J]. 情报学报, 2002, 21(1):42-47.

- [2] 李国臣. 文本分类中基于对数似然比测试的特征词选择方法[J]. 中文信息学报, 1999, 13(4):16-21.
- [3] 贺宏朝. 一种基于上下文的中文信息检索查询扩展[J]. 中文信息学报, 2002, 16(6):32-37.
- [4] 黄科, 马少平. 基于统计分词的中文网页分类[J]. 中文信息学报, 2002, 16(6):25-31.
- [5] KJERSTY A, LINE E. Text Categorisation: A Survey[M]. Oslo, Norway: Norwegian Computing Center, 1999.
- [6] WIENER E, PEDERSEN J O, ANDREAS S, et al. A Neural Network Approach to Topic Spotting[A]. Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval[C]. Las Vegas, NV, USA: Univ. of Nevada, 1995:317-332.
- [7] YANG Yiming. An Evaluation of Statistical Approaches to Text Categorization[J]. Journal of Information Retrieval, 1999, 1(1/2):67-88.
- [8] LEAH S, LARKEY, BRUCE C W. Combining Classifiers in Text Categorization[A]. Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval[C]. New York, USA: ACM Press, 1996:289-297.