基于向量空间模型的网页文本可信性分类方法

毛雪云[1,2],曾国荪[1,2],王 伟[1,2]
MAO Xue-yun[1,2],ZENG Guo-sun[1,2],WANG Wei[1,2]

1.同济大学 计算机科学与技术系,上海 201804
2.国家高性能计算机工程技术中心 同济分中心,上海 201804
1.Department of Computer Science and Engineering,Tongji University,Shanghai 201804,China
2.Tongji Branch,National Engineering & Technology Center of High Performance Computer,Shanghai 201804,China
E-mail:maoxueyun@163.com

MAO Xue-yun,ZENG Guo-sun,WANG Wei.Web text trustworthiness classification method based on VSM.Computer Engineering and Applications,2008,44(25):109-112.

Abstract: There are vast information documents in the open Internet environment.But how to judge their trustworthiness and security is a problem worthy of deep research.This paper introduces web text classification method,extracts many trust materials from documents,and establishes the trust eigenvectors of texts.Combining existing technique of feather selecting with the method of trustworthiness feather selection,this paper implements web text trustworthiness classification algorithm based on VSM,and achieves preferable effects.

Key words: content trust;text trustworthiness classification;trust eigenvector

摘 要: 开放网络环境下存在大量的信息文档,如何判断文档内容的可信性、安全性一直是一个值得深入研究的问题。论文研究了可信文本分类的方法,收集了体现文本可信性的点滴素材,建立了文本的信任特征向量,并结合已有的特征选择方法,实现了一个基于向量空间模型的文本可信性分类算法,实验表明该方法具有较好的分类效果。

关键词: 可信文本分类;信任特征向量;分类
DOI: 10.3778/j.issn.1002-8331.2008.25.033 **文章编号:** 1002-8331(2008)25-0109-04 **文献标识码:** A **中图分类号:** TP393

1 引言

开放网络环境下存在大量的 Web 文档, 如何高效地组织和管理这些文档以提高信息检索的效率具有重大的现实意义。传统的安全信任研究如名誉和认证等控制机制,本质上是实体信任的问题,只能反映实体身份的合法与否,不足以保证互联网中信息交换的内容本身的可信与否。因此,内容信任被认为是解决网络信息安全可信问题的措施之一, 成为近年的研究热点。

内容信任反映的是信息资源的一种本质特征[1],内容信任在一定程度上可归属于内容安全。美国南加州大学的 Yolanda Gil 和 Donovan Artz 在 2006 年 WWW 会议上列举了内容信任的应用场合,详细分析了影响内容信任的因素[2]。但其给出的内容信任影响因素较抽象,实现起来较困难。

作者经过大量研究发现,信息资源的内容本身蕴涵着能够反映实体可信程度的点滴素材,称为信任特征。信任特征是进行实体可信性判断的关键因素。本文利用信任特征构造了文本信任特征向量,并将各个特征项具体量化,结合广泛应用的向量空间模型(VSM),提出了一种可信文本的分类方法。通过可信文本分类,可以将不可信的文本过滤掉,从而大大提高文档的检索质量。本文详细分析了信任特征的提取方法,给出了分类的过程,描述分类过程中的关键问题,并用实验证明了该方法的有效性。

2 文本信任特征和可信文本分类

一般说来,文本可信性的判断是一个主观过程,然而我们能够找到一些相对客观的判断指导原则或标准。比如,可以考察文本的可理解性和表述性。一篇文本的可理解性和表述性越好,其可信性越好。通过对内容信任的研究,不失一般性,将文本分为以下几类:

可信任类:文本词条拼写正确,较少使用生僻词条,句子连贯通顺,易被用户接受;

不可信任类:文本内容有失偏颇,内容晦涩难懂;

基金项目: 国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z425);国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2007CB316502);国家自然科学基金(the National Natural Science Foundation of China under Grant No.90718015,No.60673157)。

作者简介: 毛雪云(1984-),女,硕士生,主要研究领域为分布式计算、信息检索;曾国荪(1964-),男,教授,博士生导师,主要研究领域为网格计算,信息安全;王伟(1979-),男,博士生,主要研究方向为信任管理、机器学习、信息检索。

收稿日期: 2008-01-24 **修回日期:** 2008-07-14

中立类:文本内容无法确定是否可信。

本文首先介绍文本的信任特征,然后介绍通过这些信任特征进行分类的过程。

2.1 信任特征

基于向量空间模型的可信文本分类区别于常规的文本分类的最大特色在于文本信任特征向量的构造。常规方法中,特征向量中的特征项取自文本中的字、词或短语,特征项的权值则由特征项在文本中出现频率、位置等信息计算得到。这种特征项的选取方法是非常直观的。比如文本中出现词语"姚明",那么此文本很可能是体育类的;进一步,如果"姚明"在文本标题出现或在正文中多次出现,那么此文本分到体育类的可能性是非常大的。可信文本分类与此不同,文本中多次出现词语"相信",即使是词语"十分可信",也不能说明这个文本是可信的。可见,文本中孤立的字、词、短语等都不能衡量文本的可信程度。

文献[2]认为文本的可信性和文本的权威性,广泛性,信誉度,上下文环境息息相关。其中对一些标准的量化已经得到了解决办法。例如,广泛性可以利用计算指向文本的链接数目,信誉度可以利用其评价文本[1],这些方法是通过第三方来评判文本的可信性,而对于如何根据文本本身判断信任性,尚且没有一个可行的方案。本文针对文本自身内容,从文本的可理解性和表述性两个方面考察了其可信性。

定义 1(可理解性(Understandability)) 指文本信息是否能够表达得有序清楚,易被用户理解,可读性强。可理解性反映了文本信息的可用性,可理解性较好的文本可信性较好。用户对文本的理解程度取决于很多方面,通过大量观察发现,可以从以下 3 个文本特征衡量文本的可理解性:

(1)文本中常用词条比例(Fraction of Popular Words,FPW)

由于文本中不可避免地会用到常用词汇,因此根据数据集中词条出现频率,定义人们日常生活中经常使用的词条集合,通过考察这些常用词条在文本中的比例,判断文本是否能够方便被用户理解。

(2)非标记文本比例(Fraction of No-mark Text,FNT)

由于在 Web 文本中,非标记文本是对用户真正有用的信息,这部分信息的比例影响到用户理解网页的程度。

(3)锚文本数量(Amount of Anchor Text,AAT)

网页中通常含有锚文本,可以作为所在页面内容的评估。一般来说,页面中的链接都会和页面本身的内容有一定的关系,例如,网页 A 中含有锚文本"Computer",则认为 A 是关于"Computer"的。同时锚文本也是目标页面内容的精确描述,若该锚文本链接指向网页 B,那么就可以认为网页 B 也是关于"Computer"的内容。如果用户不满足于网页 A 的内容,那么可以通过连接到 B 得到更详细的信息。

定义 2(表述性(Presentation)) 指文本的词条是否正确,句子和段落是否连贯通顺。可信文本一般表述性较好,而不可信文本的表述性较差。因此,为考察文本的表述性,考察以下 4 个文本特征:

(1)标题词条数量(Amount of Words in Title,AWT)

标题是文本内容的体现,许多分类技术对标题给予特别的考虑。通过研究大量的网页文本发现,文本标题在 7~20 个词条以内。长度超过 23 个词条时,标题通常是不通顺的。

(2)词条平均长度(Average Length of Words,ALW)

根据[3]的统计规律,文本词条的平均长度在 3~7 个之间,超过 8 个的文本含有较多合成词,例如 freepictures,downloadvedio,freemp3,这些拼写错误的合成词影响了文本的表述性。

(3)连贯性(The Consistency of Words,CW)

文本连贯性是指句子部分与部分之间的连续性。连贯性是其信任性的重要标志,一直是内容信任领域研究的一个热门问题。理论上可以通过分析文本的语法,最后观察其内容的语义正确性来判断,但是这种自然语言理解(NLP)的方法时间和空间代价较大,因此,采用了处理垃圾文本时常用的方法,详见小节 3.1。

(4)压缩率(Compression Ratio,CR)

通过观察大量的数据集发现,有些文本的大段内容是重复的,这部分内容是由文本创建者从其他部分拷贝来的。对于这种有冗余的文本,采用压缩率来衡量其冗余度(Redundancy),详见 3.1 节。

2.2 文本分类过程

图 1 说明了进行可信文本分类的过程。整个过程可以分为 4 部分:

(1)收集网页。

(2)对网页进行人工标注,分为可信类、不可相信类和中立类。

(3)训练:在训练模块中,首先将训练文本集向量化,得到信任特征的集合,并计算信任特征的权值,类别信任特征向量生成器得到每个类别的中心向量。

(4)测试:在测试模块中,首先将待分类文本用信任特征向量表示,再经过分类器分类,得到所属的类别。

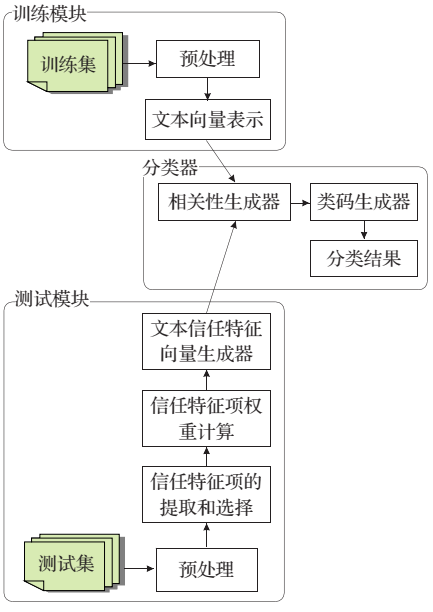

图 1 可信文本分类流程图

常规的文本自动分类方法将词条作为文本的特征向量,无法体现影响文本信任性的因素。与传统方法不同的是,本文利用影响信任性的各种文本特征,构建了文本信任特征向量,以此实现了信任分类。

3 基于向量空间模型的可信分类

根据上述分类过程,实现了一个基于向量空间模型(VSM)

的文本分类方法,VSM 是目前文本分类中使用较多[4],效果较好的一种文本表示方法。根据这种方法,一篇文档用特征项集合 $v(T_1,T_2,\cdots,T_N)$来表示,其中 $T_k(1\leqslant k\leqslant N)$代表一个特征项,每一个特征项被赋予一定权值 W_k,得到一个带有权值的特征项集合 $v(T_1,W_1;T_2,W_2;\cdots T_N,W_N)$,简单记做 $v(W_1,W_2,\cdots,W_N)$。

3.1　信任特征的抽取

如第 2 章所述,信任特征项集合确定后,需要确定每一特征项在这一文本类中的权值。一般的分类采用 TF*IDF(Term Frequency Times Inverse Document Frequency)方法,这种方法的依据是词条的出现频率,而对于以上抽取的 7 个特征项,无法根据词频来计算。因此,采用计算信任特征项的具体值来构建文本信任特征向量。下面介绍了各个特征项的权值计算方法。

(1)标题词条数量(AWT)

AWT=标题中包含的词条数量

(2)词条平均长度(ALW)

词条平均长度指的是词条所包含的平均字母个数,即

$$ALW=\frac{\text{文本总字母数}}{\text{文本中总词条数}}$$

(3)最常用词条比例(FPW)

按照词条在数据集中的出现频率定义了前 N 个最常用的词条,计算每一篇文本含有这些词条的比例,即

$$FPW=\frac{\text{文本中出现的最常用词条数}}{\text{文本总词条数}}$$

N 取 500 时,如果文本 A 含有 10 个词条,其中 4 个属于这 500 个常用词,则 FPW=0.4。

(4)非标记文本比例(FNT)

非标记文本比例指的是正文词条数和总文本词条数的比例,即

$$FNT=\frac{\text{正文词条数}}{\text{总文本词条数}}$$

其中总文本词条数包括正文文本和各种标记文本,以及网页中的各种脚本。

(5)锚文本数量(AAT)

锚文本数量指的是文本中锚文本的个数,即

AAT=锚文本个数

(6)词条连贯性(CW)

如前面所述,使用 NLP 技术分析词条连贯性,时间和空间代价是十分昂贵的,尤其当数据集达到几十万个时,几乎是不可计算的。因此,采用处理垃圾网页时的统计方法,通过考察文本局部词条来判断整篇文本的连贯性[3]。具体来说,定义 N 个连续出现的词条为一个词条组,其连贯性计算如下:

$$P(W_{i+1}\cdots W_{i+n})=\frac{\text{词条组在文本中出现的频率}}{\text{文本共划分出的词条组数目}}$$

如果数据集中的某篇文本可划分为 K 个词条组,则这篇文本含有 $K+N-1$ 个词条,通过计算 K 个词条组的几何平均作为该文本的 CW 值。即

$$CW=\sqrt[K]{\prod_{i=0}^{K-1}P(W_{i+1}\cdots W_{i+n})}$$

该统计方法是基于词条组之间相互独立的假设,但是,这种假设在实际中是难以满足的。例如,当 N=3 时,第 1 个词条组(含有文本中第 1,2,3 个词条),第 2 个词条组(含第 2,3,4 个词条),第 3 个词条组(含第 3,4,5 个词条),都覆盖了第 3 个词条,因此,提出了一种改进的方法,通过计算词条组出现的条件概率,提高计算精度。定义

$$P(W_n|W_{i+1}\cdots W_{i+n-1})=P(W_{i+1}\cdots W_{i+n})/P(W_{i+1}\cdots W_{i+n-1})$$

类似地

$$CondCW=\sqrt[K]{\prod_{i=0}^{K-1}P(W_n|W_{i+1}\cdots W_{i+n-1})}$$

由于实际中 P 值很小,为避免 K 个 P 值相乘后下溢,利用上述公式的 log 值衡量文本的连贯性,定义如下:

$$CondCW=-\frac{1}{K}\sum_{i=0}^{K-1}\log P(W_n|W_{i+1}\cdots W_{i+n-1})$$

P 值越大,网页文本连贯性越高,CondCW 值越小。

(7)压缩率(CR)

对于这种有冗余的网页通常的做法是观测每一个词条在文本中的分布情况[5],或者使用 Shingling-Based 技术。但是,这些做法适用中等数据集的情况,当数据集达到几千个时,时间和空间耗费庞大,因此,使用 GZIP 算法[6]将文本压缩,并用 CR 值衡量网页冗余度,定义

$$CR=\frac{\text{原文本大小}}{\text{压缩后文本大小}}$$

压缩率描述文件压缩后的效果,CR 越大,压缩后的文本越小,原文本的冗余度就越大,所含信息量就越小。

3.2　分类算法

每一类的信任特征向量表示完成后,即完成向量空间模型建立。待分类文档用同样的方法得到其特征向量,文本转化为向量形式之后便可以进行分类了。分类算法是分类技术的核心,常用的文本分类方法有朴素贝叶斯分类法,K-最邻近参照分类法。本文实现的可信文本分类过程具体描述如下:

3.2.1　训练阶段

(1)定义类别集合 $C=\{C_1,C_2,C_3\}$;

(2)经过人工标记和预处理后的训练网页文本集 $U=\{U_1,U_2,\cdots,U_n\}$;

(3)计算训练文本集的信任特征向量 $V=(T_{awt},T_{alw},T_{fpw},T_{fnt},T_{aat},T_{cr},T_{cw})$,各个特征项的含义和取值如上一节所述;

(4)确定每个类别的类特征向量 $V(C_i)$,计算方法为该类别所有网页信任特征向量的算术平均。

3.2.2　分类阶段

(1)对于测试网页文本集合 $D=(D_1,D_2,\cdots,D_n)$每个待分类文本 D_k,计算其信任特征向量 $V(D_k)$与每个类特征向量 $V(C_i)$的相似度 $Sim(D_k,C_i)$,计算公式为

$$Sim(D_k,C_i)=\frac{V(D_k)*V(C_i)}{\|V(D_k)\|*\|V(C_i)\|}$$

(2)选取相似度最大的一个类别 $\text{argmax}_{ci}\in_c\text{sim}(D_k,C_i)$作为 D_k 的类别。

4　实验结果

4.1　实验数据集说明

利用 Heritrix 从互联网上抓取了 2 000 张网页,将这些网页进行人工标注后分成上述 3 类:可信任类,不可信任类,无法判断类。其中可信任类占数据集 61.78%,不可信任类占 22.08%,无法判断类占 16.14%。根据第 3 章所述构建了训练集中部分网页的信任特征向量,如表 1 所示。

表 1 网页的信任特征向量实例

类别	网页来源(部分)	信任特征向量
可信任类	http://lifestyle.msn.com/messageboards/Article.aspx?cp-documentid=347767	(4,4,0.14,0.59,69,4.24,6.47)
	http://men.msn.com/articlemh.aspx?cp-documentid=5843983	(6,4.5,0.16,0.59,21,6.90,1.14)
不可信任类	http://pages.cs.wisc.edu/~blbowers/spam-message.txt	(4,3,0.17,0.22,1,1.25,2.36)
	http://www.gly.bris.ac.uk/www/comp/e-mail/spam_message.html	(2,7,0.68,0.48,3,3.7,1.4)
无法判断类	http://lifestyle.msn.com/default.aspx	(2,5,0.21,0.57,9,4.41,1.24)
	http://autos.msn.com/default.aspx,	(5,5.02,0.34,0.47,138,6.06,2.11)

4.2 实验结果及其分析

为了验证上面提取的文本特征对信任分类是否有效,在 Windows XP 系统下采用 Weka 环境进行了实验。实验中,分别使用简单向量距离分类法、KNN、和 Naïve Bayes 算法进行分类实验。3 种分类方法下实验结果如图 2 所示。

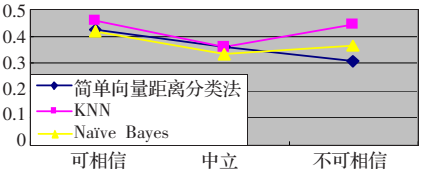

图 2 几种分类方法的 $F1$ 值比较

从图 2 可以看出,KNN 算法效果优于另外两种, 因此,采用 KNN 算法对测试集的网页进行分类。

首先,通过计算每个类别所有网页信任特征向量的算术平均,得到 3 个类特征向量分别是

$V(C_1)$=(4,4.17,0.32,0.55,58,4.89,2.35)

$V(C_2)$=(11,8,0.45,0.28,58,1.02,5.78)

$V(C_3)$=(5,5.48,0.21,0.45,125,5.01,2.14)

然后,使用 KNN 算法,利用文本信任特征项进行分类,在 K 值不同情况下计算出每个类别的 $F1$ 值,根据 $F1$ 值的大小评估分类效果,$F1$ 值越大分类效果越好,结果如图 3 所示。

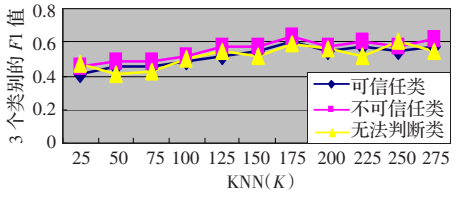

图 3 网页文本可信性分类的实验结果

如图 3 所示,$F1$ 值一般在[0.4 0.7]之间。当 N 取 175 时,$F1$ 最大,分类效果最好,达到 65%。

分析认为,要评价文本的信任性,除了上述信任特征外,文本的权威性,广泛性,客观性,上下文环境等都是影响文本信任性的因素,需要更精确的提取方式;另一方面,人工分类不可避免带有感情色彩,这使得分类结果有一定偏差。

5 结束语

本文提出了按照文本信任性进行分类的新方法,探讨了分类的主要过程和具体实现方法,结合实验结果分析了使用新的信任特征提取技术下的分类性能。文中提供的信任性划分依据和人工分类文档为后续研究也将有很大帮助, 并为信息检索,机器学习,网络安全等研究提供了可借鉴的范例。

今后将分析其他各种因素对信任性分类的影响,进一步完善文本信任特征向量,提高分类精度。

参考文献:

[1] 谷华楠,曾国荪,王伟.基于信任素材的信息文档内容信任评估[J].计算机科学,2007,34(11A):127-130.

[2] Gil Y,Artz D.Towards content trust of webresources[C]//Proceedings of the 15th International World Wide Web Conference,Aug 2006.

[3] Ntoulas A,Najork M,Manasse M,et al.Detecting spam Web pages through Content Analysis[C]//WWW'2006.

[4] 陈治纲,何丕廉,孙越恒,等.基于向量空间模型的文本分类系统的研究与实现[J].中文信息学报,1997,20(1):37-41.

[5] Ferrterly D,Manasse M,Najork M.Detecting phrase-level duplication on the World Wide Web[C]//28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,Aug 2005.

[6] GZIP[EB/OL].http://www.gzip.org/.

[7] Castillo C,Donato D,Becchett L.A reference collection for Web spam[J].SIGIR Forum,2006,40(2):11-24.

[8] Witten I,Frank H.Data mining practical machine learning tools and techniques with Java implementation[M].[S.l.]:Morgan Kaufmann,2007.

[9] 刘丽珍,宋瀚涛.文本分类中的特征选取[J].计算机工程,2004,30(4):14-16.

(上接 108 页)

[4] Karagiannis T,Rodriguez P,Papagiannaki D.Should internet service providers fear peer-assisted content distribution[C]//Internet Measurement Conference(IMC),Berkeley,CA,USA,2005.

[5] Eugene Ng T S,Chu Yang-hua,Rao S G.Kunwadee Sripanidkulchai,Hui Zhang.Measurement-based optimization techniques for bandwidth-demanding peer-to-peer systems[C]//IEEE INFOCOM'03,Orlando,Florida,USA,2003.

[6] Bernstein D S.Adaptive peer selection[C]//2nd International Workshop on Peer-to-Peer Systems,Berkeley,CA,USA,2003.

[7] Bindal R.Improving traffic locality in bittorrent via biased neighbor selection[C]//IEEE International Conference on Distributed Computing Systems(ICDCS),Lisboa,Portugal,July 2006.

[8] 谢勇均,闫涛,郑婕,等.Tracker 中一种具有拓扑意识的结点选择算法(TAPS)[J].微电子学与计算机,2007,24(1).

[9] Francis P,Jamin S,Jin C,et al.IDMaps:A global Internet host distance estimation service[J].IEEE/ACM Trans on Networking,2001,10.

[10] 程久军,于魁飞,吕晓鹏,等.一种基于 P2P 文件共享应用的片段选择算法[J].高技术通讯,2006,16(1).

[11] Guo Lei,Chen Songqing,Xiao Zhen,et al.Measurements,analysis,and modeling of bittorrent-like systems[J].IEEE Journal on Selected Areas in Communications,2007,25(1).

[12] Huang Kun,Wang Li'e,Zhang Dafang,et al.A dynamic quota-based peer selection strategy in BitTorrent[C]//The Sixth International Conference on Grid and Cooperative Computing(GCC 2007),2007.