◦博士论坛◦

基于文本挖掘的网络新闻报道差异分析

阮光册[1,2]

(1.上海海关学院 基础部,上海 201204;2. 华东师范大学,上海 201204)

摘 要:运用文本挖掘技术发现网络新闻报道中潜在的、有价值的信息是情报研究的一个新尝试。笔者探讨了网络新闻的文本挖掘方法,以上海世博新闻媒体网络版报道为例,进行实证研究,并对报道差异进行对比分析。本文选取香港、台湾、境外媒体华语版、上海本地媒体对世博会相关报道,基于文本挖掘、特征提取对报道内容的差异进行阐述,并得出结论。

关键词:文本挖掘;网络新闻;特征提取; 上海世博

中图分类号:G350 **文献标识码**:A **文章编号**:1007-7634(2011)12-105-05

Analysis on Web Media Report Differences Based on Text Mining

RUAN Guang-ce[1,2]

(1.Department of Foundation, Shanghai Customs College, Shanghai 201204*, China;*
2.Huadong Normal University, Shanghai 201204*, China)*

Abstract: It is a new research on how to find potential but valued information in the web media reports based on text mining technology. This paper discusses the text mining methods of web media reports. In the case of web media reports on Shanghai Expo, the author has done some empirical study to analyze the differences among different web media. The paper selected the web media reports on Expo from Hong Kong, Tai Wan, overseas newspapers (Chinese version) and Shanghai, analyzed the differences among these different regions base on text mining and attribution extraction and drew some conclusions.

Keywords: text mining; web news; attribution extraction; Shanghai EXPO

1 引 言

网络新闻有狭义和广义之分,狭义的网络新闻是指以互联网为平台的新闻类的信息,包括传统媒体在网站上发布的新闻信息。由于采用网络作为平台,网络新闻具有海量性、即时性、交互性和超文本等特征。对于网络新闻的内容挖掘,可以发现其报道的内涵,通过对比分析又可发现其内容的差异性。本文就台湾、香港、境外媒体华语版、上海本地媒体对上海世博会相关报道进行分析,选取的样本包括上海、香港地区、台湾地区以及国外媒体华语版,共计30家中文主流媒体,研究的新闻文本量近29000篇。通过对样本数据规范化处理、属性抽取、文本挖掘,分析媒体报道的差异。

2 文本挖掘在网络新闻的应用

2.1 研究现状

文本挖掘是数据挖掘技术中日益流行的重要研究领域,运用文本挖掘的方法对网络新闻进行情报

收稿日期:2011-04-18

基金项目:2010上海市哲学社会科学规划课题一般项目(2010BTQ003)

作者简介:阮光册(1976-),男,浙江宁波人,讲师,主要从事信息分析与网络数据库应用研究.

分析国内外的研究还不多，笔者通过资料整理发现，国内在运用文本挖掘对网络新闻处理还仅应用于新闻分类和热门话题发现，取得的主要成果有基于向量空间模型[1]实现新闻文本分类，采用聚类算法实现话题发现[2]的技术应用，文献【3】采用概率推理模型对博客的新闻倾向性进行挖掘。国外在这方面的研究较为系统化，并有一些实证研究成果，如在网络聊天室文本流主题跟踪[4]、在线新闻实时监控[5]、以及新闻报道不同主题之间相互的影响[6]等方面。笔者发现这些研究基本属于计算机领域，并没有发现运用情报分析方法对媒体信息进行文本挖掘研究的文献。

2.2 文本挖掘对网络新闻的特征提取

文本挖掘可以发现潜在、有价值的信息。如果将网络新闻看作是文本的元数据，那么它的内容可以通过一些特征来描述，如：新闻的名称、日期、篇幅等，此外，除了这些基本特征以外，网络新闻还有它的语义特征，如：描述内容、态度、对某一事件的关注程度等。相对于新闻的基本特征，语义特征的提取较为困难，本文将在3.2中详细介绍新闻信息挖掘应用中语义处理的方法。网络新闻的特征提取可以从html、xml等规范提供的Web描述语言和框架入手，通过数据规范性处理，从网络新闻中抽取其主要信息，形成关于基本特征和语义特征的分析模型。

2.3 文本挖掘对网路新闻的加权分类模型

文本分类是让机器学会一个分类函数或分类模型，该模型能把文本映射到已存在的多个类别中的某一类，根据已经训练过的文本集合，找到文本的特征和文本之间的关系模型，然后利用这些关系模型对文本集合进行分类识别。文本分类的数学表示方法为：T:Doc → C。

其中T表示为一个关系模型的概念，对于一组文本文档Doc，存在一个概念类C。Doc集合中的某一篇文本d，存在的函数关系，通过对训练过的文本模型进行函数评估得到分类的结果。然而，网络新闻由于其特殊性，往往一篇报道内包换有多个方面的主题，简单的采用分类的方法不能准确的描述文档的内涵。为此，在网络新闻研究中，对于存在的概念C进行权重的设定，用来区分权重较高概念作为分类的标准。如：新闻标题中出现的概念权值最高，新闻第一段出现的概念其权值一般要高于新闻正文其它段落出现的概念，对于非第一段出现的多个概念，可以频率统计作为测评标准。

2.4 文本挖掘对网络新闻的关联分析

关联分析是指从文档集合中找出不同词语之间的关系。关联分析在网络新闻中的应用可以分析出不同媒体对某一事件的关注模型，通过模型可以挖掘出特定网络媒体对特定事件的关注点的差异。文献【8】的研究认为，文本中出现的词都存在一定的关联性，并通过对网络中多层次网页进行关联规则的挖掘找出了不同词语之间的关系。网络新闻的关联分析具有其特殊性，除了要考虑词语间的关系，还需要结合文本分类，分析出新闻内容的差异，例如本文对一部分样本新闻进行关联分析后发现，同样关注上海世博安保问题的报道，但关注点确有所差异，有的新闻关注人流和高温，有的则关注安检等方面。此外，通过对网络新闻进行分类（如按地区分类），又可以发现不同类别的媒体在报道内容方面的差异。笔者认为，这种差异的分析可以在海量的新闻信息中发现有价值、隐藏的知识模型。

3 网络新闻的文本挖掘实施

3.1 实现流程

网络新闻的文本挖据采用图1所示的步骤进行。笔者使用上海图书馆慧科报纸数据库，以“上海世博”为检索词获取上海世博会新闻报道的专题集，在形成专题报道集的过程中，对文本信息进行相应规范化处理，针对规范化的文本信息，利用程序分别将新闻的不同信息入库（分别对应关系数据表中新闻标题、版次、正文等字段），并形成以地区为分类标准的不同实体集合。

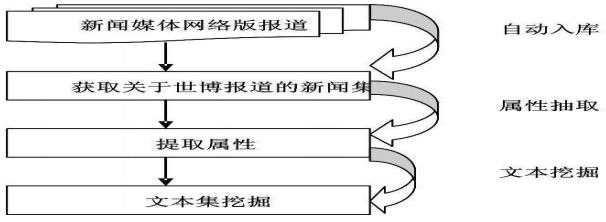

图1 新闻媒新闻文本挖掘流程

对于入库的文本信息，在特征提取时，首先要先创建一个特征集合。笔者在对网络新闻内容进行分析时发现，一篇新闻的内容可以用层次关系来描述。如：某篇新闻的内容是关于世博安全保障方面，然而对于安全保障又会有进一步的细化，如：安检、园区秩序。这里安全保障是该新闻的特征，安检和

全区秩序则为特征的详细描述。为了能够更准确的描述网络新闻的特征,本文定义了特征提取的模板:

{kNo, $Name_k$, $kList^k$}

其中,kNo为特征编号;$Name_k$为特征名,k=1,2,…,m,m为特征分类数;$kList^k$为特征k的详细描述,$kList^k=\{t_1^k, t_2^k, \ldots\ldots, t_n^k\}$,共有n个详细描述,$t_1^k$表示为特征k的一个详细描述。

3.2 网络新闻文本挖掘自然语言的处理方法

相对于关系数据库中的信息来说网络新闻属于非结构化信息,对非结构信息挖掘的难点之一是自然语言的处理。本文在实施文本挖掘特征提取时运用相似度的算法,用来匹配较为相近的内容。

本文首先定义了相似性抽取的模板:

Template={TNo, TName, News, KSet},

其中,TNo为模板的编号,TName为实体集的名称,取值为所选取的媒体名称,如:解放日报等;News为实体名称,取值为具体新闻名称,如:《站在历史的连接点上———写在上海世博会开幕之际》;KSet为实体描述的特征集合,能够反映某一News的报道内容。

自然语言相似度计算的公式如下所示:

$$Sim_T(TigerKey,\ s)=\frac{\left\|TigerKey\cap S\right\|}{\left\|TigerKey\right\|}$$

这里的TigerKey为特征提取模板$KList^k$的触发器$\{t_1,t_2,t_3,\cdots\cdots t_n\}$,考虑到自然语言的表述问题,在特征提取时需要对出现的词汇进行相似度判断。公式中,S表示为一个句子,在特征提取过程中,计算句子和TigerKey触发器的相似度[7],当大于一个阀值时,确定为某一KSet的内容。通过相似度的算法,在属性抽取过程中,程序将“安全检测”和“安全检查”视为同一属性描述,并进行提取。这样可以避免由于自然语言表述不同所造成的属性抽取错误。

3.3 挖掘算法的实现

为了更好的挖掘新闻的内容,需要对新闻所包含的各种描述属性进行挖掘。笔者在对网络新闻进行浏览时,发现一篇新闻报道虽然有某一方面的报道侧重点,但不可避免会涉及到多个主题,如报道世博服务为主题的新闻报道,还会涉及关于安全保障等方面的内容。因此,为了全面的反应相关内容,在特征提取时,本文考虑实体和特征一对多的关系,并设计了一对多的新闻实体模板。实体模板见表1所示。

表1 新闻实体模板

Nno	新闻编号	标时新闻的自动编号
nSet	新闻专题集	特定的新闻专题集,如:上海媒体
Nname	新闻名称	新闻标题
media	媒体名称	如:解放日报
kSet	特征集合	$\{KList^k \mid k=1,2,\cdots,n\}$
Rkeyword	相似性	

挖掘算法的实现如下伪代码所示,程序由C#开发:

算法:

myConn = new OleDbConnection(strConn);//连接数据库

KSet.Fill(ds, "news");//将新闻news填充到KSet集合中

KList.Fill(ds, "Keywords");//将特征集合填充到KList集合中

for(遍历KSet集合中的新闻全文)

{

for(遍历KList)

{

SearchKeyword(S,KList);//调用函数,在KSet的句子S中查找KList

if(not voaid)

{SetKeywork(KListk);//KList置于缓冲区内,}

Sim(S ,KList);//调用函数进行相似度计算

if(not voaid)

{SetKeywork(KListk);//KList置于缓冲区内,}

}

if(缓冲区不为空)

{按照新闻实体模板,填充新闻名称、属性集合、相似属性}

}

4 网络新闻报道差异分析

4.1 案例实施

本文选取了上海、香港、台湾、国外媒体华语版4-11月关于世博报道的29000篇新闻,这些报道来自于解放日报、新民晚报、大公报等共计30家中文版主流媒体。具体分布见表2所示。

表2 选取的媒体样本表

媒体地区	上海媒体	台湾媒体	香港媒体	国外媒体华语版
选取媒体数量	2家	8家	9家	11家

国外媒体华语版选取的媒体包括亚洲8家媒体，北美3家媒体，总报道量为3436篇。对选取的新闻文本，笔者对字数也进行了统计，媒体报道的平均字数见表3所示.

表3 各地媒体世博报道的平均字数统计表

媒体地区	上海媒体	台湾媒体	香港媒体	国外媒体华语版
平均报道字数	803字	689字	916字	714字

从统计的平均报道字数上来看，各地媒体对世博的报道还是较为重视，报道的篇幅也较长，报道较为全面。

对选取的文本采用文本挖掘软件WordSmith4对其进行挖掘统计，图2是部分挖掘结果。

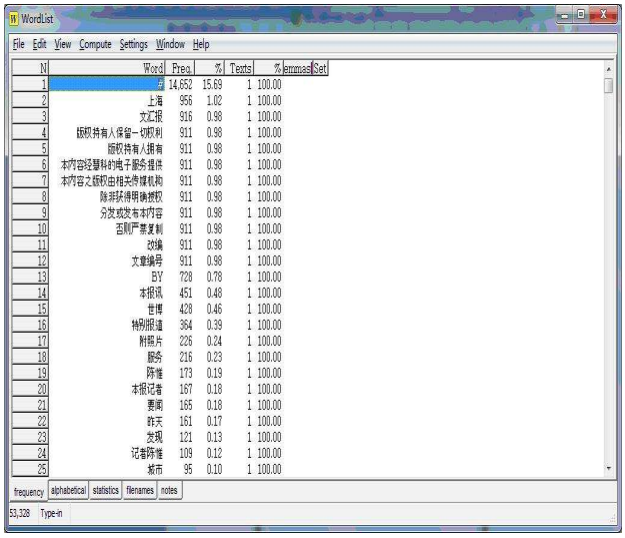

图2 相关特征的提取结果

左侧为特征的描述，中间的“%”为特征的权值。该软件还能对特征出现的位置进行统计，进而判断其重要性。

4.2 媒体报道差异分析

根据文本挖掘、特征相似匹配、以及挖掘算法的应用，笔者对29000篇来自香港、台湾、上海、国外媒体华语版的的新闻报道进行处理，可以大致的发现，这些报道更多的集中在对世博会社会氛围、安保工作、科技创新绿色生活方面的报道，相关的统计的结果见图3所示。

笔者对各地媒体报道关注度差异进行了对比分析，发现不同地区对这三大主题的关注度有一定细微的差异。从图4笔者发现，在三大主要宣传主题中，上海媒体报道量较为平均，分别为24%、26%和29%，而其香港、台湾、国外媒体华语版的报道则相对有其侧重点，其中，香港和台湾媒体更关注世博会的主题，对科技创新绿色生活专题的报道较多，分别为其报道了的39%和41%，国外媒体话语版则更多的关注世博会的安保工作，共有34%的报道量反应这方面的问题。

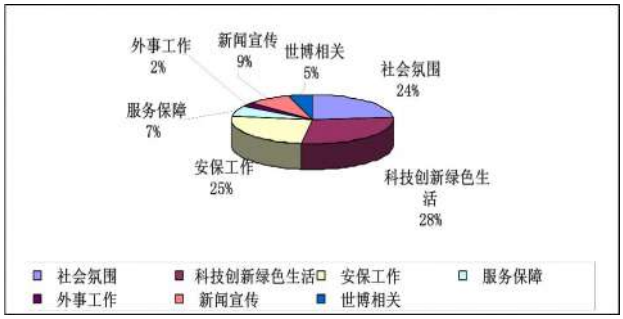

图3 报道关注度的分类统计

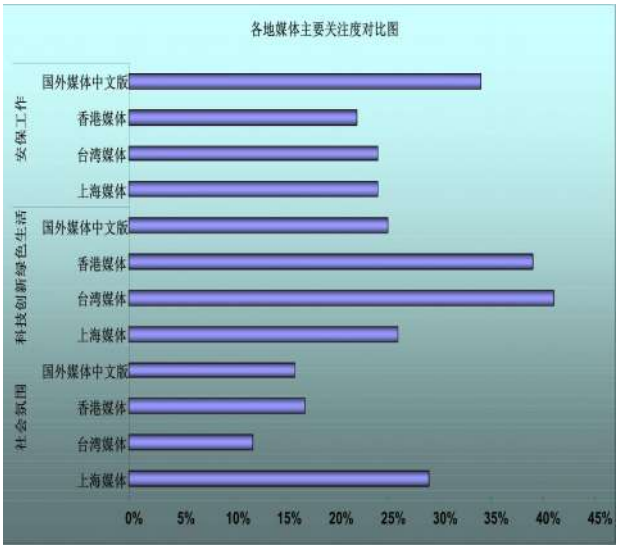

图4 媒体关注度差异对比

此外，在差异对比中，笔者发现各地媒体所关注的细节也有所不同，本文选取了关注的前五项进行对比分析，见表4所示。

表4 媒体关注细节的差异对比

媒体地区	报道关注的细节（按关注度从高到低）
上海媒体	参观人数
	科技创新
	志愿者（小白菜）
	和谐社会
	排队
台湾媒体	科技创新
	绿色环保
	排队
	参观人数
	预约
香港媒体	科技创新
	和谐社会
	参观人数
	绿色环保
	排队
国外媒体华语版	排队
	科技创新
	参观人数
	人流、秩序
	绿色环保

从表4中可以发现，各地媒体共同关注的内容

主要集中在科技创新、绿色环保以及参观人数方面，说明世博主题和参观人数是各媒体的关注重点，然而各地媒体在报道中也有一些差异，如：上海媒体在志愿者（小白菜）这方面有较多的报道，但其它地区媒体则关注较少，说明对于世博的社会效应关注不够；另外国外媒体华语版的对世博园区内人流、排队秩序方面关注较多，说明其对世博的安全运营较为关注。

5 结　语

随着网络和信息技术的不断发展，网络上的文本信息呈几何数增长，面对海量的信息，人工对其进行相关分析已变得不可能。因此，借助于文本挖掘技术发现潜在的有价值的信息是情报分析研究的一个重要应用。本文以上海世博会媒体网络报道为例，通过文本挖掘，发现了新闻报道中存在的差异，并对差异进行了比较研究。

参考文献

1 魏　程,刘　鲁,翟　铭.一种四维向量空间模型的Web新闻文本分类方法[J].微计算机应用,2010,(3):58-62.
2 高　妮,周明全.基于文本挖掘的话题发现技术[J].计算机工程,2009,35(19): 36-38.
3 廖祥文,曹冬林.基于概率推理模型的博客倾向性检索研究[J].计算机研究与发展,2009,46(9):1530-1536.
4 E.Bingham. Topic identification in dynamical text by extracting minimum complexity time components [C]. ICA: Proceedings of ICA2001,2001.
5 Montes - y - Gómez ,A. Gelbukh &A. López - López. Discovering ephemeral associations among news topics [C] . USA: In : Proceedings of IJCAI - 2001 Workshop on Adaptive Text Extraction and Mining. 2001.
6 Montes - y - Gómez , etc. Information retrieval with conceptual graph matching [C]. England:In: Proc. DEXA-2000, 11th International Conference and Workshop on Database and Expert Systems Applications, Greenwich,, September 4-8, Lecture Notes in Computer Science, Springer,2000.
7 [美] Christopher D.Manning,[德]Hinrich Schutze.统计自然语言处理基础[M].苑春法,李庆忠,译.北京：电子工业出版社,2005:183-188.
8 Brin S. Extracting Patterns and Relations from the World Wide Web [C].Valencia:Proc of Web DB Workshop,1998.

（实习编辑：赵红颖）

（上接第89页）

时俱进的能力的服务与管理队伍。大学图书馆的文化竞争力还涵盖图书馆及时掌握并适应用户需求变化的趋势把握，包括经过长期精心培育建立起来的独特的差别优势，并能增强图书馆在大学教学、科研、学科建设、管理中竞争实力的关键能力。

通过以上论述我们不难得出这样一个结论：大学图书馆因其在大学中独特的功能和定位，决定了其人物合一、动静结合、点面和谐的校园文化静态标识性；而大学图书馆管理与服务的稳定性、发展性、前瞻性成就了大学图书馆的校园文化动态标识性。大学图书馆动、静结合所形成的特有的文化竞争力铸就了大学图书馆成为大学文化的重要表征，是大学文化的标识性机构，在大学及其大学文化建设中有着不可替代的地位和作用。

参考文献

1 金旭东.21世纪美国大学图书馆运作的理论与实践[M].北京:北京图书馆出版社,2007:1-7.
2 金旭东.21世纪美国大学图书馆运作的理论与实践[M].北京:北京图书馆出版社,2007:19.
3 ACRC.Standards for Libraries in Higher Education[EB/OL]. http://www.ala.org/ala/acrl/acrlstandards,2007-03-12.
4 [EB/OL]. http://www.bvtc.edu.cn/libraryweb/train/liblaw.htm, 2007-03-12.
5 李东来,刘锦山.城市图书馆新馆建设[M].北京:北京图书馆出版社,2006:I.
6 刘文波.浅析作为文化象征性的图书馆存在[J].图书馆理论与实践,2008,(4):12-13,79.
7 [EB/OL].http://www.360doc.com/content/11/0121/14/3114071_88078303.shtml,2007-03-12.
8 付立宏,袁　琳.图书馆管理教程[M].武汉:武汉大学出版社,2005:32-58.
9 [美]佛里蒙特·E,卡斯特;组织与管理[M].李柱流,译.北京:中国社会科学出版社,1985:19.
10 刘文波.浅析作为文化象征性的图书馆存在[J].图书馆理论与实践,2008,(4):12-13,79.
11 [英]坎贝尔.核心能力战略[M].严　勇,译.大连:东北财经大学出版社,1999:74-82.

（实习编辑：赵红颖）