基于 Web 挖掘的网页清洗技术

李嘉佑[1,2]　贾自艳[2]　何　清[2]　史忠植[2]
[1](中国科技大学,合肥 230027)
[2](中国科学院计算技术研究所智能信息处理实验室,北京 100080)
E-mail: lijiayou@csm.com.cn

摘　要　随着互联网上信息的大量增多,Web 挖掘技术越来越重要。而在 Web 挖掘过程中,基于 Web 的信息抽取的主要部分是如何去除网页中的噪音数据,它是 Web 数据的预处理的过程,这个预处理结果影响了 Web 挖掘的结果。在文中先分析了噪音数据的特点,然后根据实际观察提取规则并且用于模型统计的方法,去除噪音数据,抽取相关可利用的信息。

关键词　Web 数据　信息抽取　噪音数据

文章编号 1002-8331-(2006)25-0098-04　**文献标识码** A　**中图分类号** TP301

Web Page Cleaning Technology Based on Web Mining

LI Jia-you[1,2]　JIA Zi-yan[2]　HE Qing[2]　SHI Zhong-zhi[2]
[1](University of Science and Technology of China, Hefei 230027)
[2](Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

Abstract: With rapid expansion of information resources on the Internet increasingly, Web mining technology plays an important role.How to eliminate noisy information in web pages is a main part of information extraction based on Web mining.It is a preprocessing step in the Web mining.The result of Web mining lies on the step.In the paper, we firstly analyze the feature of noisy information.Then, based on our observation, using some extracting rules and statistic methods to eliminate noisy information and extract available information.

Keywords: Web data, information extraction, noisy information

1　引言

随着 WWW 中海量信息资源的飞速增长,Web 挖掘成为从网页中发现知识和信息的一项重要任务,因而,从网页数据源中抽取出可利用的信息资源越来越显出它的重要性。抽取 Web 信息是为一系列复杂的后续工作服务的,包括网页的自动分类、聚类、信息检索、事件跟踪(新闻和金融市场等)和电子商务等。而网页中有用的大量数据往往都被许多噪音数据所干扰,包括广告、导航条、版权说明等。尽管这些噪音数据对于在互联网上浏览的用户来说有一定的功能性作用。但是,它们也妨碍了网页数据的自动收集和挖掘,包括网页自动分类、聚类、信息抽取和信息检索等的准确性、效率和性能。

现在已经有许多关于网络信息采集的研究工作,并且大多数应用在搜索引擎中。它们一般采用常规的方法来实现,机械地从网上把信息采集起来,不进行过滤和评估,不具有智能,从而通常会造成大量重复或相似的冗余信息,进而使得索引和检索的效率降低,不能满足用户的需求。因此,作为前端的自动网页采集器在效率和性能上还有待进一步提高。Google 的成功在于其庞大的网页数据库以及采用了 PageRank 和超链文本分析等核心技术,然而其网页采集器采集回来的网页只在量上取胜,在质上却存在诸多问题,因而在其检索中仍然存在很多冗余信息以及答非所问的结果。因此,在本篇论文中,针对网页噪音数据对 Web 挖掘的干扰性,我们在 Web 挖掘前首先要对网页数据进行清洗,这是一个关键步骤,以提高 Web 挖掘的结果。

在本篇论文中,第一部分我们先介绍网页数据和 Html 文档的结构特点,以及分析噪音数据的特点,第二部分给出我们的方法,第三部分为实验结果分析,第四部分为结论。

2　噪音数据的分析

2.1　噪音数据的存在形式及其相关定义

从图 1 可以看出一个新闻网页(例如:新浪网)一般由下面几个部分组成:①最上方的导航链接,例如:"首页""娱乐";②"无处不在"的广告链接;③检索输入界面;④版权信息;⑤页面主题区。

基金项目:国家自然科学基金资助项目(编号:90104021)

作家简介:李嘉佑(1977-),女,硕士研究生,研究方向:数据挖掘、搜索引擎。贾自艳(1971-),女,博士研究生,研究方向为机器学习、数据挖掘、人工智能。何清(1965-),男,副研究员,博士后,主要研究方向:模糊集理论、人工智能、数据挖掘、机器学习。史忠植(1941-),男,博士生导师,主要研究方向为人工智能、机器学习和分布式人工智能等。

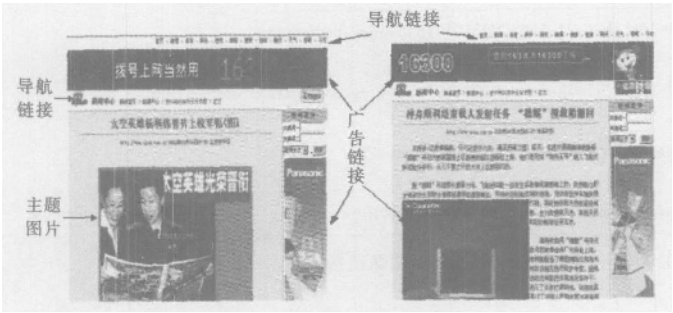

图1 含有噪音数据的网页

需要特别注意的一个现象:现在许多站点的收入是来自于广告,而且随着因特网的普及,这个现象会越来越明显。设计者在设计广告时考虑的首要问题是如何吸引用户的“眼球”。因为图像相比文字具有更强的表现力,所以通常网页中的广告都是以图像的形式嵌入到网页中。如果用户对广告感兴趣,则点击图片就可以进入广告自己的站点。实际上,人们对广告信息不感兴趣,而且图像占据了网页下载的大多数时间。如果事先知道一个链接是广告,那么可以在广告图像下载之前,将该链接去掉,这样可以节约大量的网络流量,提高信息采集的速度,同时可以提高后续图像检索的精度。

根据噪音数据的粒度,一般将其分为两大类:

(1)全局噪音(Global Noise):这种噪音具有极大的粒度,通常不小于一个网页。一般包括数据镜像站点、合法/不合法的冗余网页等。

(2)局部噪音(Local(intra-page) Noise):指网页内的噪音,这些数据通常伴随着网页的主要内容。例如:广告、导航信息等。

在本文中,噪音数据是就局部噪音而言的,这里包括:图1所示的导航链接、广告链接和版权信息等链接及其所指向的网页内容(文本或者图像)。

2.2 噪音数据的特点及其影响

通过上面对噪音数据的分析,可以得出下面的结论(假设):

假设1 根据商业模式,企业希望每个页面都有对自己广告的推荐(链接),这样用户可以容易访问到。

假设2 对于网站作者而言,为了方便用户的阅读,每个页面尽可能包含对该网站重要功能的推荐(链接),例如:“首页”、“新闻”、“娱乐”等(导航假设)。

假设3 广告链接都是以图像的形式嵌入网页中,而且占据的比重较大。

根据前面的分析可知,这是由于广告的本质决定的,就象电视广告,设计者就是千方百计去吸引观众的注意力,甚至不惜重金邀请明星参加。同时它不是单独存在的,一般需要穿插在一些收视率高的节目中间。

根据上述假设,总结噪音数据的特点是:噪音数据一般具有高冗余性。因为网页是根据模板自动生成的,所以一般每个网页都会有相同的导航链接和广告链接,这点在图1中也可以看出。我们可以根据这个特征设计噪音去除模型。

噪音数据带来的问题有下面几点:

(1)噪音数据影响了信息采集的速度和专题采集的精度,并且浪费大量的带宽。这主要是因为广告链接是图片造成的。

(2)噪音数据影响了搜索引擎的性能,加大了索引表的规模。

(3)噪音数据使得数据挖掘算法精度降低,例如文本分类、聚类。文献证明了噪音数据去除后可以大大提高数据挖掘的性能。

既然噪音数据具有上述危害,我们必须研究相应的算法将其去除,下面对噪音去除策略进行详细介绍。

3 噪音去除策略

我们在多主体开发环境MAGE的基础上,构建了并行、分布式的智能网络信息采集的多主体模型;基于该模型实现的信息采集器(Spider)是以智能主体为内核,具有主体的智能性、自治性、自适应性,各主体之间相互协作、合理搭配,对环境的适应性强,效率高,能够快速高效地采集所需要的网页。由于主体的扩展性好,可以在其知识、推理能力及功能上进行不断的升级与修改,能够满足更复杂、智能性更高的应用需求。一般地,Spider从一定的初始URL地址出发,根据URL中指向其它URL地址的超链接而跳到其它的URL,进而不断地深入和扩展,基本上遍历整个网络。我们的Spider分为四个主体内容:URL主体、抓取主体GA、解析主体PA、解析主体PA,本文的内容集中于解析主体与噪音去除主体部分。

3.1 建立噪音去除策略进行网页数据清洗

3.1.1 基于HTML标记信息的噪音判别模型

我们可以总结锚文本(anchor text)的规律,如“首页”、“新闻”、“娱乐”等词;或者基于HTML标记的规律;或者基于统计规律将其判断为噪音数据。这种判断方法在判断广告链接上都会用到,我们将在下面详述。

现在我们的信息采集不仅要处理文本数据而且也要处理图像数据。这就出现了一个问题,如何判断广告图像和主题图像?根据前面的分析,我们在图像下载之前如果能够判断其为广告图像,这样就可以避免浪费网络资源。那么如何判断一个图像链接是广告链接呢?这需要分析广告链接的特点,而不是图像自身的特点,例如:颜色、纹理。这些特点需要图片下载后才能分析出来,这样已经浪费了网络资源。

图2和图3给出了两个网页中含有的图像链接信息。其中A1,A2,A3为与主题相关的图像,而其它为广告图像。一般来说图像链接相关的信息包括:

(1)在<A>标记之间的每个图像都是广告链接的候选者。

(2)一般广告图片是与下列信息相关:图片出现的源网页(U_S),锚文字(Anchor text),锚指向的链接地址(U_D)(可以没有),图片地址(U_{img})。

(3)一般图像都有三个数据属性:Height、Width和Aspect

```
A1  <img src=http://image2.sina.com.cn/dy/c/2003-10-16/1_1-1-21-180_20031016141652.jpg border=1>
                                        ...
B1  <img src=http://image2.sina.com.cn/sms/change/zhenwen080703.jpg width=190 height=85>
                                        ...
C1  <img src=http://image2.sina.com.cn/sms/change/zhenwen080705.jpg width=199 height=108>
```

图2 网页http://news.sina.com.cn/c/2003-10-16/14161933575.shtml图像链接

A2	<img src=http://image2.sina.com.cn/dy/c/2003-10-16/1_1-1-21-166_20031016155843.jpg border=1>
	…
A3	<img src=http://image2.sina.com.cn/dy/c/2003-10-16/1_1-1-23-166_20031016155844.jpg border=1>
	…
B2	<img src=http://image2.sina.com.cn/sms/change/zhenwen080703.jpg width=190 height=85>
	…
C2	<img src=http://image2.sina.com.cn/sms/change/zhenwen080705.jpg width=199 height=108>

图 3　网页 http://news.sina.com.cn/c/2003-10-16/15581934136.shtml 图像链接

Ratio。

(4) 图像标签中可能含有“Alt”标记文本。

通过对广告链接的分析，我们可以根据广告链接的特点直接将某个链接直接判断为广告链接；另外，在信息采集过程中，只要是图像链接和对应的目的链接我们都不进行抓取，而仅仅是记录它相关的信息，以及被链接的次数，然后根据相应的计算模型判断这些链接是否是广告链接，并进行标记；对那些不是广告链接的 URL 进行采集。

HTML 链接标记信息对链接所指页面的主题有一定的预测作用。如果这些信息能够表明一个链接是噪音链接，我们将可以直接将一个链接除去。例如：如果我们认为链接地址中出现“sms”为噪音数据，那么根据该知识可以将 B1、B2、C1、C2 直接判断为噪音链接。

基于上述假设，我们可以基于链接标记信息设计相关性计算模型以实现预测 URL 的噪音相关性。基于 HTML 标记文字的噪音相关性计算模型 $NRel_{ml}$ 可以描述如下：①给定采集的 URL 种子集合 URs，噪音数据的向量模型 $Noise_m$（用户直接输入，或者由训练文件学习得到）。②对于任意属于 URs 中的一个 u，从中解析出链接图像地址 U_{img} 和目的地址 U_D；③从 u 中解析得出对 U_{img} 和 U_D 的描述信息（包括 URL 信息、链接文本信息、链接中的 Title 信息、Alt 文本信息等），得到描述信息向量 U_{Nm}；④根据公式(1)判断其是否于属于广告（噪音）链接，属于则加入噪音相关的链接集合，否则进行下一步的判断。

$$P(NRel_{ml}(U_{img}))=P(U_{Nm}=Noise_m)=\frac{\sum_{w\in U_{Nm}\cap Noise_m} weight(w,Noise_m)}{|U_{Nm}\cap Noise_m|} \tag{1}$$

$$N\varpi=\frac{\sum_{w\in Noise_m} weight(w,Noise_m)}{|Noise_m|} \tag{2}$$

其中，$weight(w,Noise_m)$ 为词 w 在噪音向量模型 $Noise_m$ 中的权重。

如果 $p(NRel_{ml}(U_{img}))>N\varpi$，则链接 U_{img} 与噪音相关，这个判断的准确率很高，但是 $p(NRel_{ml}(U_{img}))\leq N\varpi$ 则判断 U_{img} 与噪音无关的准确率不高。用户很容易给定他们需要的专题信息特征，但是对于噪音数据用户不容易给出，他们必须分析大量的数据，才能给出噪音数据的模型，所以该方法对于噪音数据不是很适用，但是如果能够知道噪音数据的特点，那么这种方式不失为一种可行的方法。

3.1.2　基于数据冗余的噪音判别模型

根据假设 1 和 2 可知，广告和导航这些噪音数据希望在每个网页都出现，同时为了便于站点维护以及增大规模，大多数站点都是基于模板方式自动生成网页。可以设想，对于同一个网站，不同网页指向同一个广告（导航）的链接地址是相同的。例如：图 2 中的广告链接 B1，C1 分别和图 3 中的广告链接地址 B2，C2 相同。可以设想我们通过统计图像地址被链接的次数，可以知道链接次数高的地址可以认为是噪音地址。所以可以设计如下简单计算模型，实现判断任意一图像链接 u 是否为噪音链接。

$$p(NRel_{m2}(u))=\frac{|Prev(u)|}{|\{p|Host(p)=Host(u),p\in URLs\}|} \tag{3}$$

如果 $p(NRel_{m2}(u))>\Theta$，那么则认为 u 为噪音数据。其中 URLs 为本次采集得到的 URL 集合，Prev(u) 为指向地址 u 的所有网页，Host(p) 为地址 p 对应的主机地址。

3.1.3　基于标记文本信息熵的噪音判别模型

基于 HTML 标记信息的噪音数据判断模型需要事先定义好噪音数据的向量空间，而基于数据冗余的噪音数据判断模型处理的数据粒度是整个 URL 地址，需要基于链接统计。如果一个广告链接的链接地址变化了，则该模型的准确性受到影响。通过从链接的锚文本（Anchor Text）中抽取词来标记链接的重要度。如果一个词在多个链接的锚文本中多次出现，那么它对用户来说含有较少的信息，相反如果一个词只在较少的链接锚文本中出现，那么它对用户来讲带有较多的信息。因此，本文也借鉴该方法实现对噪音链接数据的判断：通过从锚文本中抽取词，并且使用信息熵来表示词的信息程度。将香农的信息熵应用到词—链接矩阵，可以用公式(4)来描述词的熵。

$$E(t_i)=-\sum_{j=1}^{n} w_{ij}\log_n w_{ij} \tag{4}$$

其中，w_{ij} 表示词 t_i 在链接标记文本 j 中的权重，$w_{ij}=\frac{tf_{ij}}{\sum_{k=1}^{n} tf_{ik}}$；$tf_{ij}$ 表示词 t_i 在链接标记文本 j 中出现的频率。n 是 URL 的总数。

定义链接的熵为 HTML 标记所含的所有词的平均熵：

$$p(NRel_{m3}(u))=E(U_i)=\frac{\sum_{j=1}^{k} E(t_j)}{k} \tag{5}$$

其中，$t_1,t_2,\cdots,t_k$ 为链接标记信息中出现的词。

我们使用该方法计算图 2 和图 3 中图像链接标记文本的熵。为了减少计算量，我们在对链接地址抽取词时采用 TF*IDF 的策略。这里仅仅描述这种方法的思想，所以我们对链接描述文字的特征抽取仅仅抽取文件名字，得到词对链接的矩阵见表 1。

表 1　图 2 和 3 中图像链接对应的“词对链接”的矩阵(TU Matrix)

		A1	B1	C1	A2	A3	B2	C2
t_0	1_1-1-21-180_20031016141652	1	0	0	0	0	0	0
t_1	zhenwen080703	0	1	0	0	0	1	0
t_3	zhenwen080705	0	0	1	0	0	0	1
t_4	1_1-1-21-166_20031016155843	0	0	0	1	0	0	0
t_5	1_1-1-23-166_20031016155844	0	0	0	0	1	0	0

根据公式(4)词的熵计算如下:

$$E(t_0)=E(t_4)=E(t_5)=-\sum_{i=1}^{7} w_{ij}\log_7 w_{ij}=-1\log_7 1=0$$

$$E(t_1)=E(t_3)=-\sum_{i=1}^{7} w_{ij}\log_7 w_{ij}=-0.5\log_7 0.5-0.5\log_7 0.5=0.356\,2$$

根据公式(5)得到图像链接的熵值如下:

$E(A1)=E(A2)=E(A3)=0$

$E(B1)=E(B2)=E(C1)=E(C2)=0.356\,2$

如果一个链接的熵值越大,那么它是噪音数据的可能性就越大。所以从计算结果看来 B1, B2, C1 和 C2 为广告噪音数据。因为描述的方便,我们仅仅给出简单的例子,该例子仅仅考虑到图像链接和链接地址中的文字信息。实际上,该方法可以推广到任意的链接。可以看出该方法和基于 $p(NRel_{m2}(u))$ 的计算模型的结果是一致的,也是基于统计得到的,但是 $p(NRel_{m2}(u))$ 仅仅考虑链接地址,而 $p(NRel_{m3}(u))$ 除了考虑链接地址中的信息外,还考虑其它相关的标记信息。

4　实验结果分析

我们采用信息检索的评价体系来评价噪音去除的性能,即:

$$NPrecision=\frac{\text{过滤掉的噪音链接数}}{\text{过滤掉的所有噪音链接数}}$$

$$NRecall=\frac{\text{过滤掉的噪音链接数}}{\text{所有噪音链接数}}$$

在三种相关性计算模型中,Rel_{m1} 的计算代价最小,所以速度最快,它的采集准确率在开始时很高,随着时间采集准确率很快降低。Rel_{m2} 仅仅是对 HTML 标记信息的比较,所以计算代价较小。Rel_{m3} 需要从内容语义上进行计算,所以代价较大,如果在采集过程中同时进行语义相似度计算,那么采集的速度会很慢,所以一般这个功能由噪音去除主体完成。

图 4 给出了三种噪音判别模型的效率,$Nrel_{m1}$ 模型的准确率也较高。但是召回率较低,这是因为人们对噪音数据模型的描述不是很准确,同时网页中链接标记数据(特别是图像链接)信息较少造成的。这种计算模型需要人工指定噪音模型,所以一般较少使用。但是如果噪音数据的标记特征很准确,则该方法的计算代价最小。$Nrel_{m2}$ 和 $Nrel_{m3}$ 噪音过滤效果都较好,它们召回率和准确率都与阈值的选择有关。相比而言 $Nrel_{m2}$ 的计算代价较少,而 $Nrel_{m3}$ 由于考虑的信息较多,计算量较大,但是召回率较高,而且使用条件宽一些。

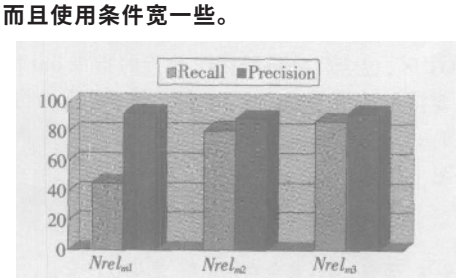

图 4　噪音数据判断模型试验结果

图 5 是采集“伊朗地震”专题时,没有执行噪音去除算法的图像结果(共有 1 461 幅),图 6 是执行噪音去除后的采集的图像结果(剩余 295 幅)。

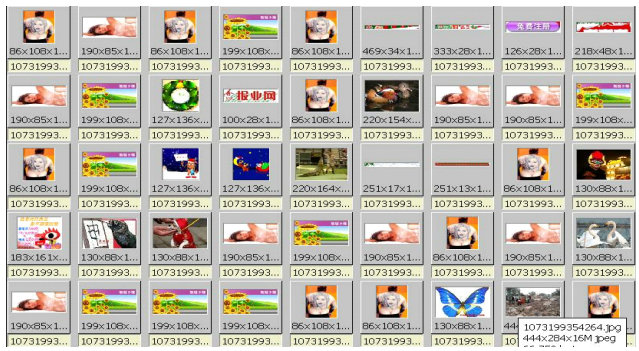

图 5　未经噪音去除的伊朗地震专题图像结果示意

图 6　噪音过滤后伊朗地震专题图像结果示意

5　结论

统计结果表明 99%的 Web 信息对于 99%的用户是无用的,也就说每个用户和企业单位只对特定领域的信息感兴趣,与自己兴趣无关的信息就是噪音。该模式需要解决的主要问题是噪音数据判别模型。实验结果表明我们给出的噪音数据判别模型的可行性。后面的工作是进一步提高信息采集,特别是专题跟踪采集模式的效率,同时改善噪音去除算法的性能。

(收稿日期:2006 年 3 月)

参考文献

1.Junghoo Cho.CRAWLING THE WEB: DISCOVERY AND MAINTENANCE OF LARGE-SCALE WEB DATA[D].Ph D Dissertation.2001

2.Steve Lawrence, C Lee Giles.Searching the World Wide Web[J]. Science, 1998; 280(5360)

3.Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology[C].In: Pazienza, Maria Teresa Pazienza eds.volume 1299 of Lecture Notes in Artificial Intelligence, Springer, International Summer School, SCIE-97, Frascati, Italy, 1997

4.N Kushmerick.Cleaning the web[J].IEEE Intelligent System, 1999; 14(2): 20~22

5.S Soderland.Learning information extraction rules for semi-structured and free text[J].Machine Learning, 1999; 34: 233~272

6.D Freigat.Information extraction from html: application of a general learning approach[C].In: proceedings of the fifteenth conference on artifical intelligence AAAI-98, 1998: 517~523

7.C Hsu, M Dung.Generating finite-state transducers for semi-structured data extraction from the web[J].Journal of Information Systems, 1998; 23(8): 521~538