Web数据挖掘中的可信数据来源

鲍 宇[1,2,3] 曾国荪[1,3] 管红杰[2]

(同济大学计算机科学及技术系 上海 201804)[1] (中国矿业大学计算机科学与技术学院 徐州 221116)[2]

(嵌入式系统与服务计算教育部重点实验室 上海 201804)[3]

摘 要 从大量Web信息中获取有用的信息是Web数据挖掘的关键问题。如何评价Web信息是否可信,现在主要方法是通过BadRank算法进行内容评测,或是通过链接权重进行相关引用数计算。可信数据来源是数据挖掘的前提,在基于关键词的数据挖掘中,通过评价挖掘所涉及的不同数据域,以及数据域自身的可信性,对在不同域所获得的挖掘数据给以权重,从而对挖掘结果的序列产生影响,提高挖掘算法在获取可信结果方面的效率。并通过试验测试了可信域评价的效果。

关键词 Web数据挖掘,Web可信数据,数据挖掘

中图法分类号 TP338

Trusted Data Source in Web Data Mining

BAO Yu[1,2,3] ZENG Guo-sun[1,3] GUAN Hong-jie[2]

(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)[1]

(Department of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)[2]

(Embedded System and Service Computing Key Lab of Ministry of Education, Shanghai 201804, China)[3]

Abstract How to abstract the trusted information is a hot issue in Web data mining. The evaluation of information in Web is obtained by content evaluation, or by BadRank algorithm or by weighing the link of pages now. This essay from the viewpoint of data source gave a new approach to evaluate the trusted information by evaluating the domains that the search engine involved in, and we gave the domains different weight values. So the order of the search result sequence will be rearranged according to the weight. That will be improved the ability of the data mining algorithm in catching the trusted result. Experimental results show that proposed system can distinguish the trusted documents in trusted domain.

Keywords Web data mining, Web trusted data, Data mining

1 引言

网络的发展给信息的收集提供了极大便捷,而Web上的海量数据是复杂的,如何从中获取有效信息一直是数据挖掘研究的内容。Web在逻辑上是一个由文档节点和超链接构成的图,其上的数据挖掘主要有3种模式,即基于结构的挖掘、基于日志使用的挖掘和基于Web内容的挖掘。Web内容的随意性导致了数据的产生过程的多样性,而基于主题的Web内容挖掘的数据量过载要求必须强化信息过滤的规则。传统的Web内容挖掘获得的知识主要是通过主题词的查询获得的,这种使用关键词的检索方式很容易造成失配问题,信息来源和对同主题的不同见解也导致了Web信息的可信性问题,所以挖掘结果中含有大量失配和不可信的信息。许多工作在数据过滤方面展开,用于提高数据过滤的准确性、可用性和可信性。Zhou[1]使用了抽取本体产生threshold的数据过滤方法。Michael[2]从Page的链接出发,计算了不同的链接方式的权重,拓展了PageRank和HITS的方法,使用支持向量机对文本进行分类。Ridvan[3]尝试用模糊的权重进一步将分类精确化。这些工作都是从Page文本内容和文本的链接上进行分类,提高分类结果的可信性,从而提供数据挖掘的可用性。为了提高挖掘数据的可信性,一些挖掘工具采用在特定域搜索的方式[4,5],通过在特定的范围内的数据挖掘可以有效地提高挖掘结果的可信性。基于特定域的挖掘很明显限制了结果的数量,而采用关键词的方式却不能有效地保证数据来源的可靠性。本文的工作主要是:在基于关键词的数据挖掘中评价挖掘所涉及的不同域的可信性,对在不同域所获得的挖掘数据的给以权重,从而对挖掘结果的序列产生影响,提高挖掘算法在获取可信结果方面的效率。

2 Web内容挖掘的可信域规则

基于Web内容的挖掘主要以各种格式的文本文档为挖掘对象,常使用Web数据仓库和Agent分类。在挖掘过程中

到稿日期:2008-05-05 本文受863项目(2007AA01Z425),973计划前期研究专项(2007CB316502),国家自然基金项目(60673157),中国矿业大学青年基金(OD4544)资助。

鲍 宇 博士生,E-mail: baoyucumt@126.com;曾国荪 博士,教授,博导,主要研究领域为网格计算、信息安全。

分类的规则包含关联、聚类等多种规则。本文将在 Web 挖掘中数据来源的可信增加规则。

2.1 挖掘的可信数据域

在 Web 的挖掘中，挖掘的结果通常是海量的，其结果序列按照搜索的关键词的关联程度和网页的链接指向的权重进行排序。在这些结果中，由于搜索的关键词的模糊性，必然产生了大量的失配信息，另外，一些 Web 文档利用基于关键词的挖掘工具的漏洞，将自己不可信的内容挤入挖掘结果中，甚至排在结果序列的前列。而对于基于链接的方式，目前的不可信文档通常采用互链的方式提高自己的链接数目，以串谋方式提高自己文档的权重。

如何有效地控制这些不可信的信息，一直是数据挖掘的难题之一。这其中涉及到 Web 挖掘的内容信任[6,7]，即根据挖掘的 Web 页面的内容判定文档内容的可信程度。但目前该类方法由于 Web 语义的研究进展缓慢，使其推进工作也受到了影响。文献[1-5]采用对关键词细分和链接的限制无疑提供了一些方法。基于 BadRank 的挖掘限制了恶意的链接，但对于不可信域的搜索却耗费了大量的时间。如果对挖掘所涉及的域进行可信评价，使不可信域内挖掘的数据限定在一定范围内，便可以有效地提高效率。Laender[8] 和 Lin[9] 根据网站内容的文本分析对网站进行了有效的分类，可以使网站的划分得到控制。根据网站的划分就可以有效地判断网站的偏好，从而可以限定挖掘的域。由于挖掘的最终结果是提供给用户使用，这里给出可信域的描述。

定义 1(Web 可信域)　在 Web 数据挖掘中，挖掘算法对挖掘源数据所在域包含特定的期望，按照挖掘规则所获取的挖掘结果符合期望的域，称为 Web 可信域。

通过可信域的描述，可以看出 Web 域的可信程度实际上是根据挖掘结果的主观判断，所以并不能判断一个域的绝对可信程度，只能获取相对的值。域的大小在 Web 挖掘中可以对应一个网站，也可以是网站的一部分，甚至包含几个网站。通过可信域的限定，在挖掘时，域的范围被限定在可信域的范围内，可信域同专业特定域[4,5]不同，不仅包含了特定域中的可信域，还要包含相关的可信域。

2.2 可信域判定规则

在 Web 挖掘中，可信域限定了数据挖掘的范围，属于数据挖掘的预处理。通过可信域的限定，对数据挖掘涉及的搜索范围和权重进行控制，从而控制挖掘的结果。因此，需要在挖掘中定义相应的可信域规则约束。

在定义挖掘的数据域时，目的是使数据挖掘的结果可信，所以限定规则必须服务于挖掘结果。然而，域的可信又是通过挖掘的结果来判定的，这就与在挖掘中使用限定可信域发生矛盾，因为挖掘域的限定需在挖掘开始就进行。另外，挖掘的域在使用过程中会由于挖掘的结果的反馈而发生修正。这些都必须在制定规则中考虑。

在首次挖掘时，挖掘的域是不受限制的，其中包含了不可信域。为更好地获取挖掘结果，应当指定相关规则，将可信概率比较高的数据域作为首次挖掘使用的可信域。

规则 1(权威域可信)　由于挖掘具有一定的目的性，具有相同偏好的相关研究信息的发布会集中在某些特定的域，该类域在初始化时作为可信域的初始值。

Web 数据挖掘在网络上获取可信数据的目标是获取权威所发布的信息，通过该信息获取挖掘的偏好的支持，该类信息比较容易识别，通常是由权威机构公布的官方消息或某个研究领域的权威消息。例如学术类可以选取“www. engineeringvillage2. org”域作为初始可信域之一。但是，有许多情况下初始域获取的信息并不能满足挖掘要求或者根本没有该类偏好的初始域，这时我们必须通过其它方式获取可信域。

目前网站分类有很多方法，基于 SVM、k-分类、神经网络、决策树、LLSF 和贝叶斯网的方法[8,9,12]，这些分类方法通过将网络域(以网站为单位)分类，根据域本身的结构和内容特征，可以分析出该类网站的专业特征(偏好取向)。通过这种手段，根据偏好获取网络域的划分，并且这些域的可信是可以判断或衡量的，那么从这些域挖掘出的数据显然更符合该类偏好，该类结果要优于普通结果。

定义 2(偏好)　对某种期望的序关系。

不同的网络域在内容选取上有不同的侧重，这种在内容的选择就造成了网络域的文档通过分类手段进行分类时，可以在分类的比例上进行排序，形成了其偏好的判断。而挖掘中则是挖掘关键字的序关系，如果挖掘关键字的分类和域偏好具有同类语义，则认为二者具有相同的偏好。

规则 2(专业域度量)　在提供的网络分类域中，如果具有和挖掘目标有相同偏好的域记为$\{D_1, D_2, \cdots, D_n\}$，并且 D_i 中和挖掘目标有相同偏好的文档数目记作 M_i，D_i 中挖掘到的文档总数目记为 S_i，则域对该类偏好的兴趣评价使用 $I_i = S_i/M_i$。

通过规则 2 可以获得域分类后的专业评价，如果一个域对某个专业具有很强的兴趣，那么从其中获取的该专业中的结果可信度要比低兴趣的域要高。但是这不是绝对的，因为挖掘要获取的是结果，而不是寻找高兴趣的伙伴。因此，对于结果的可信性还要进一步评价。

规则 3(判定内容可信的专业域优先)　在提供的网络分类域中，设具有和挖掘目标有相同偏好的域$\{D_1, D_2, \cdots, D_n\}$，其中 D_i 可以根据内容判定可信度，且内容可信度可以表示为序$[q_1, q_2, \cdots, q_n]$，则满足条件 $q_i > \varepsilon$ 的 D_i 称为内容可信域，其中 ε 表示给定的可信阈值，并且挖掘结果按照 q_i 给以不同的权重。

根据规则 3，D_i 的可信程度需要在挖掘之前进行判定。挖掘内容结果的可信需要从结果入手，这要求需要从结果的内容上判定可信。Yolanda G.[6] 给出了关于从语义上判定一个 Web content 是否可信的方式，因此，在获取到一定的结果之后，可以根据这些方法判定某个文档是否内容可信。在此，本文使用 Yolanda 的方法结合信任事实[13]的方法判断内容是否可信，使用 TD_j 表示抽取的第 j 篇文档可信度。另外，在域文档很多时(大多数情况)，不可能将域中的所有文档全部拿出评测。假设用 X_i($X_i << S_i$ 或 $X_i <= X_T$，X_T 表示抽取数量的最小值)表示从 D_i 域中抽取的文档数量，使用 T_i 表示 X_i 中判断可信的文档数量，记 $TD = \sum_{j=1}^{n} TD_j$，则

$$q_i = T_i * TD / X_i \tag{1}$$

挖掘服务最终是提供给用户的，因此可信评测的最终是由用户决定挖掘结果是否符合期望。而不同的用户，对同一结果很可能有不同的评价，而对于挖掘结果来说，如果用户认为可信，则该类结果认为可信的程度较高。所以，个性化的结果会影响可信度。如何评价一个用户对一个域可信，应当是

根据用户对该域的文档的评测情况，一个可信的结果集应当符合用户的偏好。本文使用 Bayes 的方法推断用户的可信偏好，对一个初始域的可信评价设为 $TU(TU\rightarrow[0,1])$，然后通过可信评价的数据积累推断域 D_i 的可信程度。

规则 4(用户个性化获取的域可信) 在挖掘所获取的结果中，如果通过用户使用情况获取的信息可以评测挖掘域 D_i 可信度，则将其记作 TU_i。对于满足 $TU_i>\sigma$ 的域 D_i，其中 σ 表示用户认为可信的阈值，称为用户个性化域可信，并根据 TU_i 不同给以权重。

对于访问的新域的用户可信度设定为 $TU_i=\sigma$，将对域 D_i 的访问中用户对该域所获取的第 K 个文档 F_k 的评价可以使用 TF_k 表示。如果函数 f 的参数由该域文档的时间 t_1、用户使用该域的次数 t_2 和用户对文档信息的其它评价指数 t_n[6,13]组成，则 $TF_k=f(t_1,t_2,\cdots,t_n)$。通过 TF 计算 D_i 可信度 $TU_i=TF'/\sum_i\sum_k TF_k+TF_k$，得出 D_i 可信度，其中 $TF'=\sum_k TF_k, k=1,2,\cdots$ 且 $TF_k>=\sigma$。根据连续参数的 Bayes 公式计算用户评价的第 k 个文档的可信度 TF_k，推断该域下一次获取文档的用户推断可信度 q_i' 为：

$$q_i'=P(TU_i/TF)=P(TF/TU_i)P(TU_i)/\int P(TU_i\mid p=p)P(TF/TU_i)\mathrm{d}p=\frac{TF'+TF_k}{\sum_i\sum_k TF_k+TF_k} \tag{2}$$

则一个域 D_i 的可信度 Q_i 可以使用如下公式计算得到：

$$Q_i=\omega_0 I_i+\omega_1 q_i+\omega_2 q_i' \tag{3}$$

其中，ω_x 表示各项参数的权重。

3 基于可信域规则的数据挖掘

3.1 系统模型

模型研究的重点是结合用户与挖掘结果内容的实际情况，对数据来源通过可信域规则进行限定。模型的框架基于现有的挖掘模型，通过加入可信域规则的限定，达到数据来源可信过滤的目的。可信域规则是在数据挖掘搜索域的评价和处理，因此，数据可信域规则的处理放置在数据预处理阶段。如图 1 所示，通过可信域的处理，获取数据域可信的规则，存入规则库。同时根据所得的域可信规则限定数据搜索引擎的搜索范围，从而减少不必要的搜索。另外，数据可信域的结果将作为新的分类规则，在分类算法中限定结果的排序。

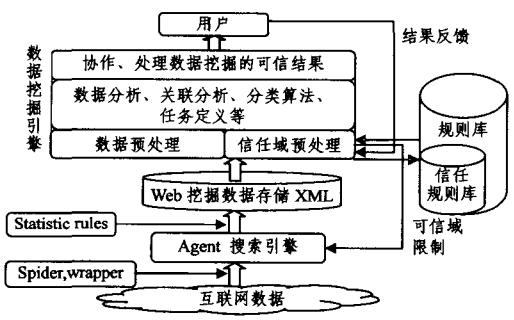

图 1 可信数据挖掘引擎

对于模型中数据域的限定主要体现在数据搜索引擎中，但在初始阶段，搜索引擎只是根据初始规则(规则 1)进行挖掘，其他规则需要通过训练或学习的手段强化。规则 2 的分类规则可以有效地判定专业域。然后，对专业域结果进行内容可信评测，判断专业域的可信度。下一步，收集用户评测的反馈结果，形成新的可信规则判定，如图 2 所示。信任域预处理模块主要功能如下：

1)控制和协调模块：控制和协调各模块之间数据传输，包括对模块外部的数据传输的控制。获取各项评测结果，并通过式(3)结算最终域可信评测结果 Q_i。

2)内容可信评测：主要通过文档结构和语义抽取文档内容各项属性，评测挖掘内容是否符合可信规范。

3)专业域可信评测：根据内容评测的结果获得 TD，根据式(1)计算 q_i。

4)用户个性化评测：获取用户反馈的信息和评测数据，通过式(2)计算 q_i'。

5)域评测数据收集和处理：提供各项接口、包括用户评测接口，数据、规则存取接口等。并通过接口将 Q_i 传输到数据搜索引擎，将 Q_i 作为一个属性数据添加到数据挖掘结果中。

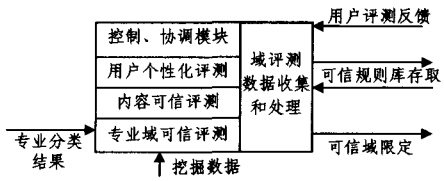

图 2 信任域预处理

3.2 系统实现与结果

系统的实现程序使用的是 Java。本文使用了搜索引擎 Nutch[11]，并在其基础上实现了信任预处理模块。Nutch 是 Apache 的一部分，主要分为两个部分：爬虫 crawler 和查询 searcher。Crawler 主要用于从网络上抓取网页并为这些网页建立索引。Searcher 主要利用这些索引检索用户的查找关键词来产生查找结果。以同济大学校园网站 http://www.tongji.edu.cn 为挖掘数据源，并对其通过实例进行评测，将其中选取 25 个子域作为评测域，每个域设置 100 个网页评测。测试结果显示子域的可信度前 4 个排列如表 1 所列：

表 1 http://www.tongji.edu.cn 部分子域的可信度

序号	域	可信度
1	deanoffi.tongji.edu.cn	1
2	jwc.tongji.edu.cn	0.98645641
3	gs.tongji.edu.cn	0.94615467
4	see.tongji.edu.cn	0.92956304

该结果有效地体现了用户特征的专业偏好和域可信效果。在开始，赋予 deanoffi 域为可信权威域，其它域通过相应的计算，用户使用时对 gs 域结果选取较多且评测较好，所以该域可信度靠前。在使用时，第 25 名由于其中包含的死链影响和文法结构不严谨导致可信度较低。

在局域网中增加测试域 testtrusteddomain.tongji.edu.cn，该域内容由 gs.tongji.edu.cn 原有内容上设置了不可信内容而生成的新域，并在搜索引擎所使用的域名服务器中增加该域。测试结果如图 3 所示，不可信网页(同一个)在最终结果的排名中由于域不可信的程度增加，排名逐步靠后。并且随着不可信内容比例的增加(0.05,0.1,0.2)，在搜索反馈后，不可信内容比例大的体现反馈效果更为迅速。

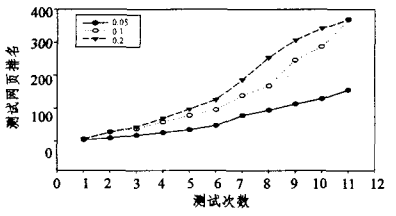

图3 可信排名效果

结束语 本文通过Web挖掘的数据域的可信评测，给出了一个Web挖掘可信域评测系统原型。利用Web域专业分类，评测域的专业度，并利用Web内容可信判定现有的结果和用户的主观评测反馈结果评测域的内容可信度，从而达到评测一个域挖掘的结果的可信度，从而获得Web挖掘的可信数据源，影响最终结果，以求在显示中将可信结果排列在前。在实现方法上，设计了信任域预处理，使之与挖掘引擎紧密结合，共同完成Web上的数据挖掘。本文下一步工作将进一步研究现有的信任评测和实现结果排序，以求更换推断的可信结果。

参考文献

[1] Zhou X, Li Y, et al. Using Information Filtering in Web Data Mining Process // 2007 IEEE/WIC/ACM International Conference on Web Intelligence. 2007:163-169

[2] Michael C, Hsinchun C. A machine learning approach to web page filtering using content and structure analysi. Decision Support Systems, 2008(44):482-494

[3] Ridvan S, Kemal T, Novruz A. A new approach on search for similar documents with multiple categories using fuzzy clustering. Expert Systems with Applications, 2008(34):2545-2554

[4] Manber U, Smit M, Gopal B. WebGlimpse: combining browsing and searching // Proceedings of the USENIX 1997. 1997,1

[5] Sun A, Lim E P. Performance measurement framework for hierarchical text classification. Journal of the American Society for Information Science and Technology, 2003, 54(11):1014-1028

[6] Yolanda G, Donovan A. Towards content trust of web resources. Journal of web semantics, 2007(11):337-358

[7] Kristin R E. Behind the Web site: An inside look at the production of Web-based textual government information. Government Information Quarterly, 2004(21):337-358

[8] Laender A H, Berthier R N, Altigran S. DEByE-date extraction by example. Data & Knowledge Engineering, 2002, 40(2):121-154

[9] Lin S H, Ho J M. Discovering informative content blocks from Web documents // Proceedings of the Eighth ACM International Conference on Knowledge Discovery and Data Mining. ACM Press, 2002:588-593

[10] Alberto D, Antonio G, Pablo G. User-centred versus system-centred evaluation of a personalization system. Information Processing and Management, 2008(44):1293-1307

[11] http://lucene.apache.org/nutch/index.html

[12] 高克宁，王波，等. WWW网站分类体系包装器WCSW[J]. 东北大学学报：自然科学版，2007，21(1)：44-48

[13] 王伟，张东启. 基于Bayes网络的内容信任[R]. 上海：同济大学嵌入式系统与服务计算教育部重点实验室，2008

(上接第194页)

现基本功能的基础之上，在用户层实现策略支持。这样就使系统具有良好的可扩充性、通用性、灵活性以及可维护性。

此外，通过proc文件系统，本检查点系统提供了一个简单的接口供用户以命令行的形式对进程进行动态操作，方便灵活。采取检查点时使用命令行：flag==checkpoint，需要恢复、重启进程时，使用命令行：flag==recovery，程序将加载检查点文件，回滚到最近的检查点状态重新执行。

结束语 本文简要介绍了用户级进程检查点系统与系统级进程检查点系统各自的优缺点，并在此基础上利用Linux内核模块设计实现了基于Linux内核的进程检查点系统。介绍了此系统的设计思路、工作原理以及工作流程，并且给出了几个主要函数的原型及各自的功能实现，最后对系统的性能进行整体评价，此检查点与恢复系统结构设计具有灵活性、通用性、良好的可扩充性以及可维护性优点。

参考文献

[1] Jose C S, Petrini F, Davis K, et al. Current Practice and a Direction Forward in Checkpoint/Restart Implementation for Fault Tolerance // Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium, 2005. IPDPS'05. April 2005:19

[2] Elnozahy M, Alvisi L, Wang Y M. A Survey of Rollback-Recovery Protocols in Message-Passing Systems. ACM Computing Surveys, 2002, 34(3):375-408

[3] 汪东升，沈美明，郑纬民，等. 一种基于检查点的卷回恢复与进程迁移系统[J]. 软件学报，1999，10(1)：68-73

[4] 魏晓辉，鞠九滨. 分布式系统中的检查点算法[J]. 计算机学报，1998，21(4)：367-375

[5] Sancho J C, Petrini F, Johnson G, et al. On the Feasibility of Incremental Checkpointing for Scientific Computing // Proceedings of the 18th International Parallel & Distributed Processing Symposium, 2004. IPDPS'04. April 2005:58

[6] Luís M S, Joao G S. System-level versus User-Defined Checkpointing // Proceedings Seventeenth IEEE Symposium on Reliable Distributed Systems, 1998. ISRDS'98. October 1998:68

[7] Meyer N. User and Kernel Level Checkpointing // Proceedings of the Sun Microsystems HPC Consortium Meeting, 2003. April 2003:15

[8] Tannenbaum T, Litzkow M. The Condor distributed processing system. Dr. Dobb's Journal, 1995, 25(2):40-48

[9] Plank J S, Micah B, Gerry K, et al. Libckpt: Transparent checkpointing under unix // Usenix Winter Technical Conference. NewOrleans, Louisiana, USA, 1995

[10] Sankaran S, Jeffrey M S, Barrett B, et al. The LAM/MPI Checkpoint/Restart Framework: System-Initiated Checkpointing // Proceedings of the LACSI Symposium, 2005. LACSI'05. October 2003:479

[11] Gioiosa R, Jose C S, Song Jiang, et al. Transparent, Incremental Checkpointing at Kernel Level: a Foundation for Fault Tolerance for Parallel Computers // Proceedings of the 2005 ACM/IEEE SC|05 Conference, 2005. SC'05. 2005:9

[12] Paul H H, Jason C D. Berkeley lab checkpoint/restart (BLCR) for Linux clusters. Journal of Physics, 2006, 46(3):494-499

[13] Zhong Hua, Nieh J. CRAK: Linux Checkpoint / Restart as a Kernel Module. Technical Report CUCS-014-01. Department of Computer Science, Columbia University, New York, November 2001