●张玉峰　吴金红　王翠波

基于 Web 结构挖掘的网络动态竞争情报采集研究*

摘　要　通过挖掘蕴含在 Web 内部结构和网页中的关联信息与结构模式,Web 结构挖掘为企业实现多维度和多层面的竞争情报采集提供了一种有效途径。基于 Web 结构挖掘的网络动态竞争情报采集方法有:URL 挖掘、Web 内部结构挖掘和超链接挖掘。图 1。参考文献 15。

关键词　Web 结构挖掘　动态竞争情报　Web 挖掘技术　信息采集

分类号　G350

ABSTRACT　Through mining related information and structural patterns within the internal structures of Web and within web pages, Web structure mining provides an effective approach for the multidimensional acquisition of competitive intelligence for enterprises. Methods for the acquisition of dynamic competitive intelligence based on Web structure mining include URL mining, Web internal structure mining and hyperlink mining. 1 fig. 15 refs.

KEY WORDS　Web structure mining.　Dynamic competitive intelligence.　Web mining technology.　Information acquisition.

CLASS NUMBER　G350

1　引言

作为公共信息平台的互联网加快了信息新陈代谢的速度,促使商业局势瞬息万变,竞争情报的时效性也由此凸显,动态竞争情报日益成为企业获得成功的关键因素[1]。当前,企业情报工作者通常借助搜索引擎等工具从海量数据中寻找所需的信息。但是由于 Web 信息资源的分布性、动态性和异构性,加之技术条件的限制,现有搜索引擎普遍有更新周期长、搜索范围有限、搜索结果准确率较低和不能很好提供多媒体搜索服务等缺点,使检索结果不尽如人意。因此,如何从海量的网络信息源中挖掘出高质量动态情报是一个亟待解决的问题。

事实上,企业竞争情报工作通常是围绕一些主题如竞争对手、竞争环境和竞争策略等来组织展开的。巧合的是,Web 页面链接也表现出很强的主题特征,即 Linkage/Sibling Locality 特性[2] 和 hub 特性[3]。Linkage Locality 指页面具有链接到相关主题的其他页面的趋势;Sibling Localily 指页面的入链所对应的网页趋向于拥有相关主题的信息。另外根据 hub 特性,可以从相关主题页面中找出权威页面和中心页面,权威页面是指与主题相关的价值最高的页面,中心页面聚集了很多相关主题的权威页面的超链接。Web 链接的这些特征实质上对页面的内容起到了一种概括作用,它在一定程度上比超级链接页面作者所作的概括要更为客观、准确[4]。直接从 Web 页面链接结构中挖掘语义知识来指导和实现网络动态竞争情报的采集,是一条重要的有效途径。

2　Web 结构挖掘

Web 结构挖掘是通过分析 Web 页面的链接信息,推导链接之间的包含、引用或从属关系,挖掘潜在有用的知识。Web 结构挖掘的目的是获得主题高度相关的链接以及链接逻辑结构的语义知识,这些知识可以帮助人们从中找到有价值的权威页面、中心页面。Web 结构挖掘综合了人工智能、数据库、数据挖掘、计算机科学、信息学等多个领域的理论与技术,是 Web 挖掘的一个重要分支。

2.1　Web 挖掘

Web 挖掘是从互联网信息资源中挖掘有趣、潜在、有用的模式及深层知识的过程。它是将数据挖掘理论和技术应用于互联网信息资源挖掘的一个新兴研究领域。根据挖掘对象的不同大致可分为三类[5]:

(1)Web 内容挖掘。指从 Web 文档内容及其描述信息中获取潜在的、有价值的知识或模式的过程。

*本文系国家自然科学基金资助项目"基于数据挖掘的企业竞争情报智能采集机制研究"(70573082)成果之一。

Web 内容挖掘是从 Web 文档的内容中抽取有用信息，挖掘对象包括文字、图片、音频、视频或者结构记录比如列表或表格等。这个领域的研究经常包括其他学科的技术,如信息检索和自然语言处理。

(2)Web 使用挖掘。对用户访问 Web 时在服务器留下的访问记录进行挖掘,挖掘对象是服务器上的日志信息,也称为 Web 日志挖掘。典型的记录包括 Web 用户的 IP 地址,访问的页面和访问时间。Web 使用挖掘可以通过跟踪和理解 Web 用户在访问 Web 站点时的浏览行为,更好地提供基于 Web 的应用需求。

(3)Web 结构挖掘。主要是从 Web 组织结构和链接关系中推导信息和知识,挖掘对象是 Web 文档内部的结构信息和外部的超链接信息。

对于 Web 结构挖掘,我们可以把这个领域再划分为 3 个类别,即超链接挖掘、内部结构挖掘和 URL(Uniform Resource Location)挖掘。超链接挖掘揭示 Web 文档之间的逻辑联系,内部结构挖掘获取文档内部或 Web 组织内部的结构框架,而 URL 挖掘则对相关的 URL 地址进行分析和聚合。

2.2　Web 结构挖掘的处理步骤

Web 结构挖掘的过程不是很复杂,但数据准备阶段很重要。Web 链接信息虽有一定的结构性,但因为自述层次的存在和复杂的相互关联,所以是一种非结构化的数据。我们在将数据挖掘技术引入结构挖掘的时候,要做大量的预处理工作。按照每个步骤产生的数据集特征,可以将 Web 结构挖掘的基本过程分为 4 个阶段。

(1)生成种子页面集。种子页面集是 Web 结构挖掘的基础,数量依具体问题而定。这些页面之间的链接要求不是特别紧密,但在种子页面集中至少会有些链接可以找到主题相关的页面。

(2)生成种子 URL 集。选择好种子页面集后,将其页面上的出链和相关信息提取出来,形成种子 URL 集供爬行器选择。

(3)生成候选集。通过 Web 上的协议,顺链搜索获取种子 URL 所指向的页面,经过页面分析,将提取的链接存入数据库里,形成候选链接集。页面分析通常包括链接的提取和页面内容的分析,这里主要是指链接的提取。为保证挖掘结果准确,需要对链接进行甄别,排除干扰链。按照链接的功能,网页上的链接可分为参考链接、网络功能链接和广告链接;只有参考链接在内容上有主题相关性,后两种链接大多数是出于某种需要而加入的链接,在 Web 结构挖掘中会形成干扰,应把它们排除在外。

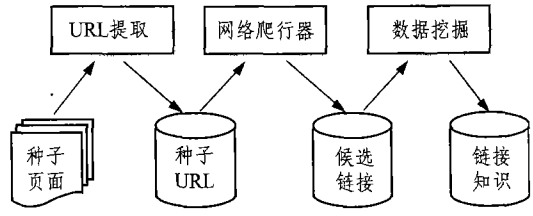

图 1　Web 结构挖掘的处理步骤

(4)生成链接知识库。这是 Web 结构挖掘最重要的部分。按照一定算法,运用数据挖掘技术对候选集中的链接数据进行分析,从中挖掘出有用的知识,如生成网络拓扑图、找出中心网页和权威网页等。Web 结构挖掘的算法很多,有基于随机漫游模型的,比如 PageRank,Reputation 算法,基于 Hub 和 Authority 相互加强模型的,如 HITS 及其变种,基于概率模型的,如 SALSA,PHITS,基于贝叶斯模型的,如贝叶斯算法及其简化版本等等[6]。

2.3　影响 Web 结构挖掘质量的因素

衡量 Web 结构挖掘结果的一个重要标准是主题相关性,对挖掘结果产生显著影响的因素主要包括种子页面和相关度算法。

(1)种子页面集[7]。种子页面集是决定 Web 结构挖掘质量的关键,好的种子页面集可以大大减少爬行器搜索页面的数量,有效提高 Web 结构挖掘的聚集速度和效率,种子页面集应具有较高的主题相关性。有 3 种方法可以生成种子页面集:①人工指定,即由专家给出相关的种子页面集。这种方法生成的种子页面集质量高,但容易产生"过训练"的现象。②自动生成,用搜索引擎利用关键词的方法搜索网络,从返回结果中抽取前 N 个页面作为种子页面。由于搜索引擎自身的特点,造成种子页面集的主题聚集度不够。③混合模式,即自动生成与人工指定相结合,首先利用通用搜索引擎获得部分相关页面,然后经人工筛选、过滤、合并、评价,形成一个能充分反映主题特征的种子页面集。但这 3 种方法也有不足。应增加系统的学习能力,通过建立学科主题种子页面库,在对检索历史、用户反馈信息分析的基础上,动态优化相关主题的种子页面集,为新主题定制提供默认种子页面,并为用户进行种子页面选择及评价提供参考。

(2)相关度算法。通常我们采用 HITs 算法和 PageRank 算法来挖掘权威网页和中心网页,但这两种算法都有缺陷,如对页面的质量和重要性不加区分,仅

仅根据网页的入链和出链的数量;给每个链接都给予相同的权值,不能区分链接之间的重要程度;对重复链接的情况没有予以考虑;对链接的时间也没有考虑,导致新生成的Web页面不能获得好的排名。这些因素都可能使挖掘出现"主题污染"或"主题漂移"现象。例如,当查询"电影奖"时,得到的结果却是许多电影公司的主页。这是因为和"电影奖"有关的网页通常会链接向电影公司的主页,由于电影公司主页的商业性,大量的链接会发生在这些公司主页之间,错误地诱导结构挖掘算法。

通过引入页面内容的语义分析可以解决提高链接的主题相关性。目前已有不少研究者对这些算法进行改进,如IBM Almaden实验室的CIEVER系统[8]、Compaq系统研究中心的Web Archaeology项目[9]以及STED算法[10],这些算法在实际应用中都结合传统的内容分析技术进行了优化,取得了不错效果。

3 基于Web结构挖掘的网络动态竞争情报采集

Web链接结构挖掘应用于面向Web的企业竞争情报系统,能够帮助企业获取有关竞争对手、竞争环境的最相关的链接,通过这些链接,对相关信息源结构进行挖掘,可揭示权威网页之间的关联,揭示蕴含在这些文档结构信息中的有用模式,有助于从多个维度和层面挖掘竞争情报。

3.1 URL挖掘,集成高质量的竞争情报采集源

URL挖掘,也就是从Web页面中获取相关主题的URL集合的过程,它是Web结构挖掘的一个重要研究内容。URL挖掘为企业开展竞争情报工作发现情报源提供了一种便利手段,在企业的情报工作中,可以利用Web结构挖掘方法挖掘出与竞争情报需求相关的URL,作为企业采集网络竞争情报的入口。这些竞争情报采集源具有较高的主题相关性,有助于选择性地搜索有限的网络空间,发现、下载与主题相关的信息,提高竞争情报采集速度。法国图书馆的"网络文献采集项目"(BnF)就利用了Web结构挖掘的URL发现功能。它首先利用Web挖掘技术,获得包含相关主题的网络资源的一系列网址,经过分析处理,BnF把这些网址发送给有关专家,以评估是否进行采集[11]。

3.2 Web内部结构挖掘,获取竞争对手的最新动向

企业站点的信息资源直接产生于公司企业内部各生产、销售、服务、管理部门和环节,是了解竞争对手最有价值的第一手情报源,它是获得竞争对手情报的主要途径之一。Liwen Vaughan等根据信息产业部公布2002年度中国IT行业100强名单,对这些企业的Web站点进行数据采集,汇总相关的链接信息后,通过进一步统计分析发现:企业商务站点的链接数量和企业的年收入、利润、研发经费等经营活动有着密切联系,在一定程度上代表了企业的业绩[12]。

Web结构挖掘为企业实时监视竞争对手动态提供了可能,而且从情报分析的角度来说,它是一种更加隐蔽的办法。Web结构挖掘只是对相关联的网页跟踪分析,不会引起被调查对象的注意,其结果相对来说更加高效、准确[13]。市场上已经出现了这类产品,如C-4-U、TrackEngine、ChangeDetect等系统可以提供网站监视功能,自动跟踪用户想监视的竞争对手的网站,当竞争对手的网站有变动时自动通过Email等方式通知用户[14]。

3.3 超链接挖掘,获得行业发展的动态趋势

目前对Web结构挖掘讨论最多的是利用超链接挖掘找出权威页面和中心页面,并用来评价网站的重要程度以及在搜索引擎中的排名次序。实际上,超链接挖掘在竞争情报采集中也有非常重要的作用。Web结构挖掘能够快速准确地获得相关主题的权威页面和中心页面,这为企业网络竞争情报的采集和质量评价提供了极大帮助。

从页面的作用来看,中心页面是相关信息的链接"集市",通过它很容易找到大批与竞争情报需求相关的链接。通过这些链接,企业可以成批获得零售商、中间商、合作商和竞争对手的信息,减少了工作人员搜索网页的时间,降低了信息遗漏的几率。而权威页面在竞争情报工作中的作用更大:通过浏览权威页面,企业可以了解本行业的最新动态信息,了解本行业内一些著名的大型企业的发展动态,获得企业发展所必需的竞争环境情报;权威页面的内容是本行业内最可信赖的情报来源,从其他来源获得的情报,可以通过与权威页面的内容进行比较分析,辨别情报的真伪,确认情报的价值[15]。

4 结束语

网络竞争情报源的多样性、动态性、综合性,使得竞争情报的采集必须采用先进的信息技术来满足企业对竞争情报的急切需求。Web结构挖掘通过分析Web网页之间的语义关联,揭示蕴含在Web网站内部结构中的有用模式,为企业提供了实时高质量的动态竞争情报;通过挖掘竞争对手、竞争环(下转第95页)

的应用(III):实证研究. 中国图书馆学报,2006(5)
14 温有奎. 基于知识元语义网格平台的知识发现研究. 计算机工程与应用,2006(4)
15,24 温有奎,徐国华. 知识元链接理论. 情报学报,2003(6)
16,25 温有奎,赖伯年. 网格技术将推动知识管理革命. 情报学报,2004(1)
17 温有奎等. 基于创新点的知识元挖掘. 情报学报,2005(6)
19 温有奎. 基于"知识元"的知识组织与检索. 计算机工程与应用,2005(1)
20 温有奎. 知识元挖掘. 西安:西安电子科技大学出版社,2004
21 曾民族. 向知识标引进军——阅读《知识元挖掘》的体会. 情报学报,2006(2)
22 朱庆华.《知识元挖掘》评介——兼议情报学的理论研究. 情报科学,2006(12)
23 马炳厚.《知识元挖掘》评介. 图书馆理论与实践,2006(5)
26 张静等. 论中小学多媒体知识元库的建设. 现代教育技术, 2005(5)

文庭孝 湘潭大学管理学院副教授,中国科学院科技政策与管理科学研究所博士后。通讯地址:湘潭大学管理学院。邮编411105。

侯经川 湘潭大学管理学院教授,浙江大学公共管理学院博士后。通讯地址同上。

龚蛟腾 湘潭大学管理学院讲师,武汉大学信息管理学院博士。通讯地址同上。

刘晓英 湘潭大学图书馆馆员。通讯地址同上。

汪全莉 湘潭大学管理学院讲师。通讯地址同上。

(来稿时间:2007-05-18)

(上接第64页)境的最相关的链接信息,为企业竞争情报工作提供了高质量的采集源;通过实时监控竞争对手网站,全方位地获得竞争对手最新的发展动态。还可以将Web结构挖掘方法与内容分析技术相结合,快速、多维、多层面地采集动态竞争情报。

参考文献

1 张玉峰,部先永,晏创业. 动态竞争情报及其采集基础. 中国图书馆学报,2006(12)
2 C. Aggarwal. F. Al-Garawi and P. Yu. "Intelligent Crawling on the World Wide Web with Arbitrary Predicates". In Proceedings of the 10th International WWW Conference, Hong Kong, May 2001
3 Kleinberg J. Authoritative sources in a hyperlinked environment. In: Tarjan RE, et al. , eds. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms. New Orleans: ACM Press, 1997
4 李卫,刘建毅,何华灿等. 基于主题的智能信息采集系统的研究与实现. 计算机应用研究,2006(2)
5 Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, SIGKDD Expirations, ACM SIGKDD, July 2000
6 朱炜,王超,李俊,潘金贵. Web超链分析算法研究. 计算机科学,2003(9)
7 李春旺. Web信息主题采集技术研究. 图书情报工作,2005(4)
8 Chakrabarti S, Dom B, Gibson D, Kumar S, Raghavan P, Rajagopalan S, Tomkins A. Experiments in topic distillation. In: Proceedings of the ACM SIGIR workshop on Hypertext Information Retrieval on the Web. Melbourne: ACM Press, 1998
9 Bharat K, Henzinger M. Improved algorithms for topic distillation in a hyperlinked environment. In: Voorhees E, et al. , eds. Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval. Melbourne: ACM Press, 1998
10 王晓宇,周傲英. 万维网的链接结构分析及其应用综述. 软件学报,2003(10)
11 Serge Abiteboul, Mihai Preda, Gregory Cobena, Adaptive online page importance computation, Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, May 20-24, 2003
12 Liwen Vaughan and Guozhu Wu. Links to commercial websites as a source of business information. Scientometrics, 2004(4)
13 杨光. 链接分析在企业竞争情报活动中的应用. 图书情报工作,2005(1)
14 吴伟. 国外竞争情报软件研究. 情报理论与实践,2004(1)
15 陈萍丽. Web挖掘及其在竞争情报系统的应用. 情报科学,2003(9)

张玉峰 武汉大学信息管理学院教授,博士生导师。通讯地址:武汉。邮编430072。

吴金红　王翠波 武汉大学信息管理学院05级情报学博士研究生。通讯地址同上。

(来稿时间:2007-02-05)