基于WWW的聚类引擎研究

张 伟
(重庆教育学院计算机与现代教育技术系 重庆400067)

Study of Clustering Engine Based on WWW

ZHANG Wei
(Department of Computer & Modern Education Technology, Chongqing Education College, Chongqing 400067)

Abstract Today, search engine is the most commonly used tool for Web information retrieval, data mining may discover knowledge in large data. With the era of information and digital of media, Web data mining is becoming one of the hottest topics. By combining information retrieval technology with data mining technology, a prototype system of search engine is designed and implemented in this paper. It can group Web search results in a semantic, online and tree way, in order to help users find relevant Web information easier and faster.

Keywords Information Retrieval, Search Engine, Data Mining, Clustering

1 引言

WWW是为广大用户提供信息交换、共享而发展起来的一种Internet应用。WWW上的信息量一直呈爆炸性增长,据最新调查和统计,2003年WWW上的网页数目将达到80亿,全球互联网用户将达到3.6亿。人们上网的主要目的就是获取信息[1],然而网上99%的信息对99%的用户是没有价值的,因此Web信息检索(Web Information Retrieval)就成为一个很重要而又很困难的问题。

搜索引擎(search engine)如Yahoo、Infoseek、Altavista、Google、Excite等,是目前Web信息检索的主要工具,大约有85%的用户通过搜索引擎寻找自己需要的Web信息。大部分搜索引擎的工作原理是:通过机器人程序,在全球范围的Web服务器上定期搜索网页;将搜索到的网页进行分析、整理,提取关键词后,放入搜索引擎的索引数据库中;用户向搜索引擎的界面输入关键词进行查询。然而这些搜索引擎还存在很大局限性,据调查,有超过45%的用户表示使用WWW时遇到的最大问题就是找到自己真正需要的信息非常困难[2]。搜索引擎虽然对于WWW的发展起到了巨大的推动作用,但是随着因特网的发展,搜索引擎也暴露出以下一些不足:

1)网络资源的发展,使得搜索引擎能够检索的范围越来越小;

2)Web是一个动态增长的信息源,但是搜索引擎不能及时反映这种变化;

3)搜索引擎面对的用户多种多样,而用户的信息需求、知识背景和兴趣各不相同;

4)对于用户来说,用户检索到的有用信息经常淹没在众多的无用信息当中。

数据挖掘(Data Mining)旨在提取数据中隐含的、未知的、有用的、不一般的模式或知识,又称数据库中的知识发现(KDD:Knowledge Discovery in Databases)[3,4]。数据挖掘的数据源可以是一般的数据库或数据仓库,也可以是非数据库系统中的数据。数据挖掘的结果是概念化的知识,这些知识源于数据,但高于数据。随着社会的信息化和各种媒体的数字化,Web数据挖掘(Web Mining)逐渐成了一个研究热点。

聚类(clustering)是数据挖掘的基本形式之一。按照数据的相似性和差异性,将数据划分为若干组(组内还可以再分组),同组的尽量相似,不同组的尽量相异,这种对数据进行自动分组的方法称为聚类[5]。聚类与分类(classifying)不同:分类是一种监督学习(supervised learning),其类别是根据应用的需要事先确定的,根据表示事物特征的数据可以识别其类别;聚类是一种非监督学习(unsupervised learning),其类别不是人为指定的而是分析数据的结果,聚类完全由计算机自动进行,不需要人工干预。聚类通过比较数据的相似性和差异性,能发现数据的内在特征及分布规律,从而获得对数据更深刻的理解与认识。

信息检索技术和数据挖掘技术的结合,可望使搜索引擎上升到新的高度,使普通用户也能很容易地在WWW上找到自己真正需要的信息。把新的Web数据挖掘技术应用到搜索引擎中去,为Web信息的利用提出了新的解决方案,将会引起搜索引擎业界一场新的革命。因此,进行基于WWW的聚类引擎研究有着十分重要的意义。

2 基本思想

Web信息检索的研究是一个新的具有挑战性的前沿课题,探讨Web信息检索的精度及易用性,如何充分利用数据挖掘技术(本文采用聚类方法),使搜索引擎上升到新的高度,从而方便用户很容易地在WWW上找到自己真正需要的信息等问题,还需要进行深入细致的研究。

我们采取的基本思想是希望通过聚类方法自动组织搜索引擎的搜索结果,方便用户发现真正需要的Web信息,如图1所示。

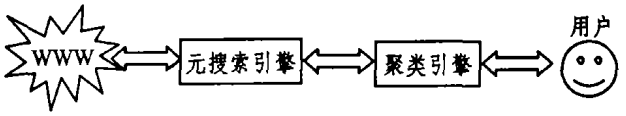

图1 基于WWW的聚类引擎

张 伟 博士,副教授。

我们对搜索引擎的搜索结果进行聚类分析，而没有对整个 WWW 数据进行聚类分析的原因是，因为 WWW 数据规模巨大而且质量参差不齐，导致信息与垃圾混杂、知识与谬误共存。由此可见，搜索引擎在这里实际上起到了对 WWW 数据进行过滤的作用。所以，对搜索引擎的搜索结果进行聚类更有效、更可行。可以说，我们是想把 Web 数据挖掘技术与搜索引擎相结合，从而构建一个基于数据挖掘的搜索引擎原型系统。该系统既可用于各种专业性的搜索引擎，也可扩展为通用搜索引擎。

3 具体实现

基于 WWW 的聚类引擎研究的主要难点在于 WWW 环境的复杂性(分布、自治、异构、动态)和 WWW 数据的复杂化(不规则、不精确、不稳定、半结构化或非结构化)。

基于 WWW 的聚类引擎应该是：语义的、在线的和树型的。我们研究内容的核心是一种语义的、在线的、树型的 Web 信息聚类方法。

语义的聚类是指：根据信息的内容(语义)进行聚类，使之能够直接汇聚具有指定特征或含有特定内容的信息。聚类后赋予每个类一个关键词组来简明且准确地概括类内信息的共同主题(topic)，使用户看到关键词组就能了解这个类的内容。类与类之间应该允许重叠，即若某条信息涉及多个主题则可以属于多个类。

在线的聚类是指：即时(just-in-time)响应用户的聚类要求，迅速提供最新的聚类结果。树型的聚类是指：按照类与类之间的联系将所有的类组织成一棵树(tree)的形式。树中的每个结点都对应一个类，一个结点(类)的子结点(子类)就是在其主题范围内的那些结点(类)。

我们把这种新的 Web 信息聚类方法简称为 SOTC (Semantic Online Tree Clustering)。基于 WWW 的聚类引擎系统结构如图2所示。

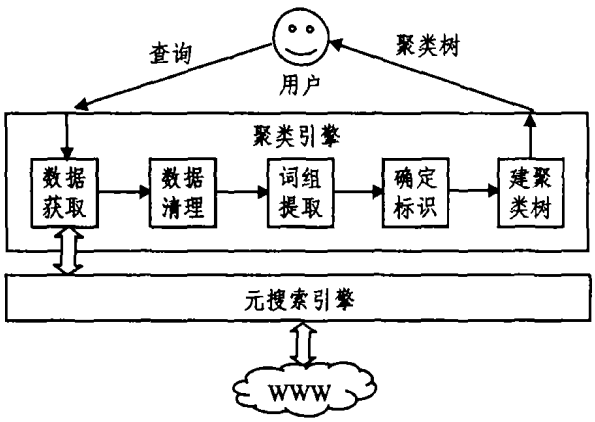

图2 基于 WWW 的聚类引擎系统结构

系统的输入是用户提出的查询(query)，其输出是聚类之后的相关 Web 信息，中间经过以下几个步骤：数据获取、数据清理、词组提取、确定基类和建聚类树。具体步骤如下：(1)数据获取：系统通过元搜索(meta-search)进行数据获取，即将用户提出的查询转交给多个搜索引擎，依靠搜索引擎得到相关 Web 信息的列表；(2)数据清理：搜索引擎的搜索结果中一般包括相关 Web 信息的 URL、标题和摘要等，所谓数据清理就是将 Web 信息的标题和摘要合在一起作为一个“文档”，以这个“文档”来表示 Web 信息的主要内容；(3)词组提取：即是对“文档”进行关键词组提取，这是“语义的”准则体现；(4)确定标识：标识也称基类，就是以词组为中心，将与词组在语义上的相似度高于某个阈值的所有文档聚集在一起所组成的文档类；(5)建聚类树：目的是为了进一步方便用户浏览与查找各个文档的类，以缩短 Web 信息的浏览距离。

SOTC 的 Web 信息聚类方法是语义的、在线的和树型的。其中，“语义的”准则体现为：以关键词组作为文档特征、基于语义进行文档聚类和以关键词组来标识类；“在线的”准则体现为：多线程两级并行元搜索、以 Web 信息的标题和摘要作为文档和利用后缀数组进行关键词组提取；“树型的”准则体现为：将所有的类组建成一棵树的形式。

小结 我们结合信息检索技术和数据挖掘技术，设计并实现了基于 WWW 的聚类引擎系统，使其能够以语义的、在线的和树型的方式对搜索引擎的搜索结果进行聚类。Web 信息检索和数据挖掘在未来还会继续迅速发展，如 XML 格式[6]Web 信息检索与挖掘、Web 信息自动问答系统[7]等等都是很有意义的研究课题，我们将在后续工作中予以报道。

参 考 文 献

1 中国互联网信息中心. 中国互联网络发展状况统计报告(2001/1)，2001. http://www.cnnic.net.cn/develst/cnnic200101.shtml

2 Kehoe C, Pitkow J. Tenth WWW Survey Report. The Graphics, Visualization & Usability Center, Georgia Institute of Technology, 1999

3 Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Inc., 2000

4 Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, Inc., 1999

5 王能斌. 数据库系统原理. 电子工业出版社, 2000

6 World Wide Web Consortium. Extensible Markup Language (XML) 1.0 Recommendation. February 1998. http://www.W3.org/TR/REC-xml

7 Kwok C, Etzioni O, Weld D S. Scaling Question Answering to the web. In: Proc. of the 10th Intl. World Wide Web Conf. (WWW10), Hong Kong, May 2001