面向主题的WWW信息挖掘系统

余 晨 顾毓清
(中科院软件所 北京100080)

Topic-Driven Web Information Mining

YU Chen GU Yu-Qing
(Institute of Software, Chinese Academy of Sciences, Beijing 100080)

Abstract With the explosive growth of the World-Wide Web, it is becoming increasingly difficult for users to collect and analyze Web pages that are relevant to a particular topic. In this paper, Topic-Driven Web Information Gathering system is presented, which can efficiently collects Web pages for a topic in relatively limited hardware and network resources, and keeps the pages more up-to-date.

Keywords Search engine, Topic, Crawler, Authorities, Hubs

1 概述

WWW正以令人难以置信的速度飞速地发展,逐渐成为人们发布和获取信息的主要平台。虽然人们可以从WWW上获得大量信息,但由于WWW上的信息是无结构的、动态的、分散的,因此如何从WWW上高效地提取有用的信息仍是一个很有挑战性的课题。搜索引擎(如Excite、Google、AltaVista)的广泛应用,使人们检索信息的效率大大提高。搜索引擎的工作原理是:由一个爬行器(Crawler)尽可能多地收集WWW上的网页,按照每个网页的文本内容建立单词到网页的反向索引,用户在查询感兴趣的主题时,输入该主题的关键字作为查询条件,搜索引擎利用事先建立好的网页库和单词索引,检索出符合条件的网页。搜索引擎可以满足用户一般的查询需要,但有些用户只关心一个具体的主题,需要获得针对该主题比较详尽的信息,普通搜索引擎与用户在这一需求之间存在着很大的偏差:

·普通搜索引擎的目标是回答用户的所有查询,对于只关心某一具体主题的用户,这一目标显得过于浪费,覆盖一切的目标造成了技术上和资源上的极大压力,也降低了针对具体主题查询的查准率和查全率。

·使用普通搜索引擎用户往往用几个简单的词表达查询,而这样有时并不能准确地定义主题。

·用户的查询是在服务器的网页库和索引上进行的,从根本上说是静态的,由于WWW信息更新非常迅速,查询的结果很可能不准确,而且更新索引的代价非常大。

·普通搜索引擎仅通过对网页的内容的分析判断与用户查询符合与否检索出用户需要的网页,缺乏对检索得到的网页进行进一步挖掘和分析的能力,而检索结果网页(我们称之为主题相关网页)之间的链接结构,以及这些网页与其他主题无关网页之间的链接结构都蕴含着大量信息,普通搜索引擎不能对其进行利用。

对于关注具体主题的用户来说,他们更需要一个针对主题进行网页收集、管理和分析的工具,本文所描述的面向主题的WWW信息挖掘满足了这一需求,它与普通搜索引擎相比具有以下优势:

·对于特定的主题,具有更高的查准率和查全率。

·由于仅提取与主题相关的页面,大大降低了对网络流量和系统性能的要求,使其可以在普通的台式机上实现。

·不需要维护静态网页库和静态索引,实时性好。

·可以根据用户的需求准确地定义主题。

·对检索到的主题相关网页进行分析,提供多种衡量网页质量的参数,挖掘相关主题给用户提供有用的参考。

面向主题的WWW信息挖掘框架由网页收集部件、分析部件、反馈部件和用户接口组成。首先,在系统的初始化阶段,用户需要给出主题的种子网页。网页收集部件从种子网页开始,从WWW上收集与主题相关的网页。分析部件对收集到的主题相关网页集合进行分析,对网页之间的链接结构进行挖掘。当收集和分析结果返回给用户后,反馈部件使用相关度反馈机制,收集用户对哪些网页和自己的需求相关(及其相关的程度)、哪些不相关的反馈,通过多次交互逐步求精对网页检索部件中的参数进行修正。系统结构如图1所示。

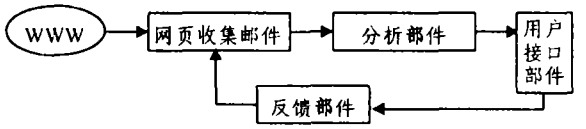

图1 系统结构

2 相关研究

文[4,5]描述了针对主题进行WWW聚焦检索的方法,我们的系统中采用类似的方法进行主题相关网页的收集,并采用启发式方法提高网页收集的效率。

文[1]介绍了基于WWW链接结构进行分析的HITS算法,文[2]采用该算法对检索结果网页进行排序,我们的系统中对HITS算法针对面向主题信息挖掘的特点进行了改进,将其用于主题相关网页集的链接分析。

3 网页收集部件

网页收集部件负责由种子网页开始从WWW上收集与主题相关的网页,它的核心是一个聚焦爬行器。

3.1 基本原理

余 晨 硕士生,主要研究领域为Web技术,软件工程。顾毓清 研究员,博士生导师,主要研究领域为软件工程。

网页收集部件负责由种子网页开始从 WWW 上收集与主题相关的网页，它由负责收集网页的爬行器和负责判断网页与主题相关性的分类器组成。该爬行器与普通搜索引擎的爬行器不同，被称为聚焦爬行器(Focused Crawler)。普通搜索引擎爬行器的目标是覆盖尽可能多的网页，所以在扩展 URL 时采用的广度优先的方法，跟踪当前页面中的每一个 URL，下载到本地网页库。而聚焦爬行器的目的是尽可能多地找到主题相关的网页并减少不相关网页的下载，所以采用 Best-First 算法扩展 URL。

如果与同一主题相关的网页在 WWW 上是平均分布的，那么聚焦爬行器就无法实现，然而，经研究，WWW 上的网页分布具有主题聚合性，即如果一个网页是与主题相关的，那么它的邻居，即它使用超级链接指向的网页的主题相关性远大于在 WWW 上随机选出的网页。聚焦爬行器利用 WWW 的这一特性，优先对主题相关网页中的超级链接进行跟踪。

3.2 基本实现

首先由用户给出种子 URL(也可以由用户给出简单查询，用普通的搜索引擎搜索到的结果作为种子 URL)，种子 URL 所指向的网页作为分类器的训练集。分类器将训练集中的文档连接起来，形成主题特征文档，在这一过程中，需要删除常用但无意义的单词，比如 and，or 等，如果文档过长，可以只取每个文档的前500个单词。采用 TF-IDF(Term Frequency Inverse Document Frequency)表示法将特征文档转化为向量，向量的每一维对应一个特征词。当分析网页与主题的相关性时，也将网页转化为基于特征词的向量表示，然后利用公式(1)计算网页 p 与主题 t 的相关程度[3]：

$$rel(p,t)=\frac{\sum_{k\in FT}V_{kp}V_{kt}}{\sqrt{\sum_{k\in FT}V_{kp}^2\sum_{k\in FT}V_{kt}^2}} \quad (1)$$

其中，$V_{kp}=freq_{kp}*IDF_k$，$V_{kt}=freq_{kt}*IDF_k$，$freq_{kp}$为特征词 k 在网页 p 中出现的频率，$freq_{kt}$为特征词 k 在主题 t 的特征文档中出现的频率，IDF_k 为特征词 k 在所有 WWW 网页中出现的概率的倒数，当然这只能是一个估计值。

爬行器维护一个待下载的 URL 队列(称为 Crawl Frontier)，开始时 Crawl Frontier 仅包括种子网页的 URL，爬行器下载队首 URL 所指向的网页，使用分类器计算网页与主题的相关程度，如果相关程度大于阈值，则将其作为主题相关网页保存起来，并将该网页中的超级链接加入 Crawl Frontier，重复这一过程，直到 Crawl Frontier 为空，或达到其他停止条件为止。

3.3 提高效率

下载网页的时间和全文计算网页与主题的相关程度是爬行器执行效率的瓶颈，所以需要尽量减少不相关网页的下载，这就需要在不下载网页的前提下，通过父网页的信息预测 URL 所指向网页的主题相关性，使 Crawl Frontier 中的 URL 按预测的相关性排序，这样爬行器就可以优先下载和分析"期望值"高的链接，提高检索效率。

我们采用以下启发式方法预测链接的主题相关性：

(1)通过父网页上该超级链接的元数据进行计算。元数据是指在父网页上为其通过超级链接指向的子网页提供的说明信息。经研究，超级链接的 AnchorText、URL 文本以及链接相邻的文字等元数据对该链接所指向的网页内容起很重要的说明作用，所以我们运用于判断文档相似性同样的方法，利用公式(1)计算三者的 rel 值，并计算加权和。

(2)网页的主题相关性具有传递性，即网页与主题相关，那么它用超级链接指向的子网页也很有可能与主题相关，子网页的相关性也可以用父网页相关性乘一个衰减因子来估计。

(3)组织良好的 WWW 站点的目录结构一般都具有主题聚合性，即同一目录中的网页往往是与同一主题相关的，利用这一特点，认为如果一个目录中发现了若干个与主题相关的网页，那么很有可能在这个目录中有更多的相关网页，反之如果一个目录目前访问过的网页中大部分都与主题无关，那么目录中剩下的网页也很有可能与主题无关，因此一个网页的主题相关性还可以根据它所在的目录中已经分析过的网页的主题相关性来估计。

综上所述，子网页的主题相关性估计值为：

$$Meta\_Rel=\delta_1\times relmeta(url,page,t)$$

$$Tran\_Rel=\delta_2\times rel(page,t)\times decayfactor$$

$$Dir\_Rel=\delta_3\times F(dir)$$

$$rel(url,t)=Meta\_Rel+Tran\_Rel+Dir\_Rel \quad (2)$$

其中，Meta_Rel 为根据链接元数据计算的主题相关性，Tran_Rel 为传递相关性估计值，DirRel 为根据目录其他网页估计的主题相关性，F(dir)为目录 dir 中已处理过的网页相关性的函数。

算法如下：

```
enqueue(frontier_queue,seed_url,1)
while(not terminal){
    url = dequeue(frontier_queue);
    page = crawl_page(url);
    if(rel(page,topic)> x){ //x 为阈值
        index_page(url);
        url_list = extract_urls(page);
        for each u in url_list {
            record_linkinfo(u);
            if(!Crawled(u)&& !StopURL(u))
                enqueue(frontier_queue,u,rel(u,topic));
        }
    }
}
```

其中，enqueue(queue,element,rel)将 element 加入 queue，queue 为 Crawl frontier 的 URL 队列，并根据 element 所指向的网页的主题相关性 rel 进行排序；dequeue(queue)从 queue 中取出队首 URL；index_page(url)将 url 所指向的主题相关网页保存下来；record_linkinfo(url)记录 url 的链接信息。

算法中需要特别解释的是 StopURL 函数，它的作用是判断 URL 是否在"StopList"中，StopURL 是一个 URL 列表，它包括大部分网页都经常链接的 URL，比如，www.amazon.com，这些 URL 经常被其他网页引用，但往往不属于我们搜索的目标。另外还包括那些经常出现页面错误或服务器错误的网站，这一部分需要爬行器在执行算法的过程中动态地学习，我们认为如果一个网站无法链接，或已有大于20个网页被访问过，其中90%不存在，这样的网站都应被加入 StopList。

4 分析部件

网页收集部件收集到主题相关的网页，与一个主题相关的网页数量可能很多，我们的系统并不是简单地将这些网页返回给用户，而是由分析部件进行分析，使用户能更好地利用收集到的网页。分析部件对网页的分析包括链接结构分析和相关主题分析。

4.1 链接结构分析

网页收集部件对网页的内容进行分析，除了网页的内容之外，一个网页指向另一个网页的链接也包含着很有用的信息，分析部件的作用之一就是对网页之间的链接结构进行分析。

当一个网页的作者建立指向另一个网页的链接时，说明

作者对另一个网页的认可。一个网页被其他网页引用的次数越多，其来自不同作者的认可也越多，说明该网页越重要，这类网页被称为权威(Authority)网页。另一方面，还有一类重要的网页，称为HUB。HUB是指这样一类网页，它提供了指向权威网页的集合。不仅网页内容与主题的相关性，链接结构也能提供很宝贵的信息，权威网页和HUB网页的挖掘对用户也很有意义，所以对于每个检索到的网页，本系统给用户提供三个衡量网页质量的参数，一为该网页与主题的相关程度，它是由网页收集部件给出的，二为该网页的权威值，三为该网页的HUB值，并能够分别按三个参数对检索到的网页进行排序。

4.1.1 *基本算法* 如果将整个WWW视为一个顶点为网页、边为网页间链接的图，那么网页检索部件找到的所有与主题相关的网页和链接就组成了该图的一个子图，我们称其为主题子图。分析部件对主题子图的拓扑结构进行分析，分析网页的权威值和HUB值，找出对该主题的权威性网页和HUB网页。我们采用HITS算法对链接结构进行分析。通常好的HUB网页指向许多好的权威网页，好的权威网页有许多好的HUB网页指向，这种HUB网页与权威网页之间的相互作用是HITS算法的基础，根据这一特点HITS算法采用迭代的方法计算各网页的权威值和HUB值。原始的HITS算法描述如下：

1. 构建Web子图：由查询得到的结果集(比如由普通搜索引擎返回的结果集)构建一个大约有200个网页的根集，将根集扩展更大的网页集，其中包含了根集网页所指向的网页和指向根集网页的网页，这样得到的网页集构成了Web子图的顶点，网页之间的链接(不包括站点内部各网页间的链接)构成了Web子图的边。

2. 将各顶点的权威值和HUB值初始化为1。

3. 进行迭代，使用以下公式计算每个顶点的权威值 a_u 和HUB值 h_v，其中E为Web子图的顶点集合，u、v其中的顶点，即，主题相关的网页。

$$a_u=\sum_{(v,u)\in E} h_v \tag{4}$$

$$h_v=\sum_{(v,u)\in E} a_u \tag{5}$$

以上等式反映了若一个网页由很多网页引用，其权威值会增加，若一个网页指向多个权威性网页，则其HUB值也会相应增加。每一个迭代之后，对各节点的权威值和HUB值进行规范化处理，使所有的权威值和HUB值的平方和均为1。

4.1.2 *针对主题相关网页集合的改进* 原始的HITS算法对于挖掘主题相关的权威网页和HUB网页存在以下问题：

·没有考虑网页的主题相关性，对所有的网页都赋予相同的权重，有可能造成主题的偏移。

·仅从网页的角度而没有从网站的角度去考虑，没有考虑重复链接的问题。即经常出现这样的情况，一个网站的很多网页都指向另一个网站的某一个网页，这样会提高该网页的权威值，我们假设一个网站所有网页都由同一作者制作，所以同一网站多个网页指向一个网页只能说明同一作者对这一网页的认可，因此在计算该网页的权威值时不应将来自同一作者的认可简单叠加，我们采用降低同一网站内多重链接的权重的方法。同样，如果一个网页引用多个网页属于同一个网站，那么计算它的HUB值时，也应考虑减少这些网页的权重。

因此将算法改为：

1. 构建Web子图：网页检索部件检索到的主题相关的网页集合以及它们之间的链接(不包括站点内部各网页间的链接)作为待分析的Web子图。

2. 将各节点的权威值和HUB值初始化为相应网页的主题相关值。

3. 迭代计算各节点的权威值和HUB值，在一次迭代过程中，在计算网页的权威值时，将指向它的网页的HUB值都乘一个权值authorityweight(v,u)，如果一个网站只有一个网页指向该网页，那么该权值为1，如果一个网站同时有多个网页指向该网页，则适当降低每个网页的authorityweight(v,u)值；对网页HUB值进行计算时，也进行相似的处理。计算顶点的权威值 a_u 和HUB值 h_v 的公式变为：

$$a_u=\sum_{(v,u)\in E} h_v\times authorityweight(v,u) \tag{6}$$

$$h_v=\sum_{(v,u)\in E} a_u\times hubweight(v,u) \tag{7}$$

4.2 相关主题分析

分析部件的另一功能是相关主题分析，对主题进行扩展。在主题相关网页的超级链接中，一部分指向主题相关的网页，这些网页已经返回给用户，另一部分指向与主题无关或相关性不强的网页，其中可能很多描述了与该主题有密切联系的主题，对这部分链接进行分析，也可以挖掘出对该主题感兴趣的作者同时对哪些主题感兴趣，这些主题可以被认为是该主题的相关主题。

具体的方法是：对主题相关网页中那些在网页检索过程中被标注为主题无关的超级链接进行分析，这些超级链接包含指向相关主题的链接，对它们进行分析，这些相关的主题用关键词的形式给出。为了提高效率，我们使用链接的元数据(Metadata)而不是链接所指向的网页本身进行分析。

使用最大模式挖掘算法，每个待分析链接的元数据作为一个事务，元数据中的有效词作为事务项，所谓有效词是元数据中的词去掉主题特征词和StopWord以后的词。在这些事务中找出最大频繁集，即待分析链接中的最经常出现的词的集合，该集合对主题的扩展具有很好的启发作用。

总结 本文描述了在WWW上针对一个具体主题进行信息收集、管理和分析的框架，它在针对主题聚焦搜索的基础上，实时高效地进行主题相关网页的收集，使用改进的HITS算法对主题相关网页集合内部的链接结构进行挖掘，找到权威网页和HUB网页，使用最大模式挖掘算法，对主题相关网页集合外部的链接结构进行分析，挖掘相关主题，对主题进行扩展。

参考文献

1 Kleinberg J. Authoritative Sources in a Hyperlinked Environment. The 9th ACM-SIAM Symposium on Discrete Algorithms, May 1998

2 Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: The 7th Intl. World Wide Web Conf. (WWW7), April 1998

3 Salton G, McGill M. Introduction to Modern Information Retrieval. McGraw-Hill, 1983

4 Chakrabarti S, van den Berg M, Dom B. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. In: The 8th Intl. World Wide Web Conf. (WWW8), May 1999

5 Ester M, Gro M, Kriegel H P. Focused Web Crawling: A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies. In: The 27th Int. Conf. on Very Large Databases (VLDB '01), Rom, Italien, 2001

6 Davison B D. Topical Locality in the Web: Experiments and Observations: [Technical Report DCS-TR-414]. Department of Computer Science, Rutgers University, 2000

7 Han Jiawei, Kamber M. 数据挖掘概念与技术. 机械工业出版社, 2001