基于兴趣模型的 WEB 信息预测采集过滤方法

李振星[1] 徐泽平[2]

[1](北京航空航天大学机械工程及自动化学院 北京 100083)；

[2](中国科学院计算所 北京 100080)

E-mail: zhenxing_li@sina.com

摘要： Web 网上海量信息急速膨胀使得有效定向采集相关信息检索成为网上信息查询一个日益重要的研究方向。本文提出一种基于用户兴趣模型的 Web 文本信息预测采集过滤方法。这种方法根据正反集文本过滤方法，设计出一种用户兴趣模型，并在对 Web 站点结构进行分析的基础之上，通过对网页的相关度的预测来控制信息的采集。在保持定向采集精度的同时，缩短采集时间、减少存储、加快检索，节约了网络资源。

关键词: 信息采集; 兴趣模型; 文本过滤

引言

随着 Internet 和 WEB 的飞速发展，WEB 上的海量信息保持加速膨胀。基于 Internet 的各类信息检索得到了迅速的发展，迫切需要建立针对特定领域信息的专用检索系统，实践证明这种 WEB 检索系统是一个非常有用的信息检索工具[1]。

Web 检索系统通常由三个部分组成[2]：(1)在 Internet 网上搜索信息的信息采集工具 Robot；(2)把信息进行分类索引建立网页数据库的索引器；(3)通过 WEB 服务器为用户提供浏览器界面下进行信息查询的检索器。

有预测的定向采集相关信息，避免无关信息的采集，可以缩短采集时间、减少信息存储、加快检索时间，也节约了网络资源。信息采集工具采集到的网页质量直接关系到整个检索系统是否能够为该领域的用户提供良好的检索服务。

特定领域的 WEB 信息采集主要涉及到两个方面的问题：一是如何对网页进行过滤的文本过滤技术，判断其相关性；二是 Robot 的网页采集控制策略。由于 Internet 上大量的信息表现形式都是文本形式，所以文本过滤技术已成为当今信息技术领域讨论的热点[3][4]，文献[5]首先建立一种所谓的 CDT－判定树 (concept-based decision tree,基于概念的判定树)，然后进行概念扩充以便更好地表现用户的信息需求，计算待过滤文本的相似度，根据相似度阈值和匹配率阈值，最终将文本推送与其信息需求相符合的用户，这种过滤机制适合于多用户的信息分流。文献[6]提出一种进化式的信息过滤方法，从多个角度描述用户的信息需求，它们之间相互竞争又相互合作，使系统性能达到最优，它从一个新的角度对过滤器的训练作了新的尝试。传统的 Robot 的网页采集控制策略有两种：宽度优先策略和深度优先策略，文献[7]中提出了一种基于 URL 预测的优先级优先策略。文献中对于有预测的定向采集 Web 信息，没有给出完善的解决方法。已有的实现系统中，采集大都是给定限制全部下载信息，然后分析，过滤，处理。

本文在对 Web 站点结构进行分析后，在正反集文本过滤方法基础上，提出基于用户兴趣模型的特定领域 WEB 信息预测采集方法。这种方法首先设计出一种用户兴趣模型，通过对网页的相关度的预测来控制信息的采集。该算法得到了实际的测试运用。

2 Web 站点结构

Web 上的信息总体来说是无结构的、异质的、分布的、动态的[8]，但是对于 Web 上的信息组织仍然有一定的结构，这既包括由 URL 中目录层次反映出来的物理结构，也包括页面和页面间的链接构成的逻辑结构[9]。可以通过分析 Web 站点信息结构，初步判断信息的类别。

2.1 物理结构

一个完整的 URL 包括协议和路径两个部分[10]。在 Web 站点中使用的基本上都是 HTTP 协议，在此只对 HTTP 协议进行分析。对于 HTTP 协议，其 URL 的语法形式如下：

http://<host>:<port>/<path>?<searchpart>

其中<host>表示站点主机名(域名或 IP 地址)；<port>表示端口号；<path>表示页面的路径；<searchpart>表示 CGI 接口 GET 方法的参数表达式。对于一个站点来说，其主机名和端口号都一样，参数对于站点结构没有什么意义，能够用来表示站点结构的只有<path>部分。页面的路径和 Web 站点的文件系统是对应的，也是一种分层的树型结构，每个层之间通过“/”分开。

对于一个设计比较规范的站点，内容的组织都是按照栏目进行的，每个栏目包含某个主题的内容。一个站点设置若干个栏目，内容较多的栏目则要设置子栏目，每个栏目或者子栏目的文件分目录进行存放。这样，站点物理结构相同的页面就属于同一栏目，它们有相同或相似的主题，可以利用站点的物理结构来进行信息的采集。

2.2 逻辑结构

站点物理结构反映的是页面的存储方式，Web 页面内容之间的主要联系是通过页面间的链接来进行的。如图 1 所示，目前对于页面链接的分类没有一定的标准，参考国内外一些研究成果[4][6][11]，将页面之间的链接主要包括以下 5 种类型：

- DOWNWARD —— 下行链，目标页面是当前页面的下级页面。
- UPWARD —— 上行链，与 DOWNWARD 相反，目标页面是当前页面的上级页面。
- HORIZONTAL —— 水平链，目标页面和当前页面处于同一目录。
- CROSSWISE —— 交叉链，目标页面和当前页面不在同一路径上。
- OUTWARD —— 外向链，目标页面和当前页面不在同一站点。

通常情况下，下行链的目标页面是对当前页面详细描述；上行链的目标页面是对当前页面的概括；水平链的目标页面和当前页面属于同一领域内容；对于交叉链和外向链只是内容相关，具体属于什么关系很难确定。

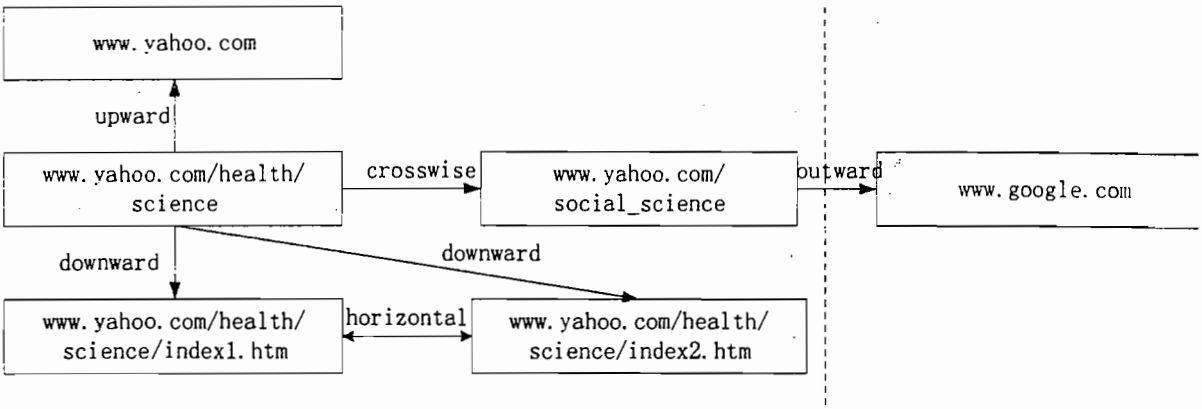

图 1　链接类型图

3 用户兴趣模型及页面相关度

在 Web 上对某个特定领域的信息进行采集，首先需要让系统了解这个领域的精确描述，通过正反集文本过滤方法，可以构建用户兴趣模型来表示用户所感兴趣的领域。通过用户兴趣模型计算页面的相关度，是决定是否需要采集该页面的重要指标。

3.1 用户兴趣模型

假设所有文档集为 D，人为的将其分为 m 的领域，每个领域的文档集为 D_i(i=1，2，…，m)。对于每一文档的目标表示模型有多种[12]，常用的有布尔模型、概率模型、向量空间模型，近年来应用较多且效果较好的是向量空间模型(vector space model，VSM)。在 VSM 中，将文档看作是一组词条$(T_1, T_2, \cdots, T_n)$构成，对于每一个词条 T_i，都根据其在文档中的重要程度赋以一定的权值 W_i，这样每个文档就可以用词条特征向量$(T_1, W_1, T_2, W_2, \cdots, T_n, W_n)$表示。如果用户对某个特定领域例如 D_1 感兴趣，其它领域 D_i(i=2，3，…，m)对于该用户来说为不感兴趣。可以将用户兴趣模型定义为：

$$P = \alpha U + \beta V + \gamma W \tag{1}$$

其中，P 表示用户兴趣，它是一个词条特征向量；U 表示感兴趣文档集的词条特征向量；V 表示不感兴趣文档集的词条特征向量；W 表示普通词汇的词条特征向量；α、β、γ 分别为感兴趣、不感兴趣、普通词汇词条特征向量的经验系数。

在构造用户兴趣模型的时候，只需要用户提供两类文档集的训练样本：感兴趣文档集样本 D 和不感兴趣文档集样本 D'。感兴趣文档集的词条特征向量 U、不感兴趣文档集的词条特征向量 V 以及普通词汇的词条特征向量 W 的计算方法如下：

(1) 分别对 D 和 D'中的每一文档预处理(对 HTML 文档需要去除 TAG 标志)之后进行分词

(2) 对于分词结果的每一个词条利用 TFIDF 词条权值作为评价函数：

$$W_{ik} = \frac{tf_{ik} \log(\frac{N}{n_k})}{\sqrt{\sum_{k=1}^{n} (tf_{ik})^2 \bullet \log^2(\frac{N}{n_k})}} \tag{2}$$

其中 tf_{ik} 表示词条 T_k 词条在文档 D_i 中出现的频数，N 表示全部样本文档总数，n_k 表示词条 T_k 的文档频数。

(3) 分别统计 D 和 D' 中每一词条的总词条权值：

$$Sum(W_k) = \sum_{i=1}^{N} W_{ik} \qquad (3)$$

假设对文档集 D 的运算结果为特征向量 V1，其中对文档集 D′的运算结果为特征向量 V2。特征向量的长度如果不进行限制就是分词词典中词条数，由于词频数较小的词条不具有统计价值而且增加了系统的负担，可以限制分词词典中词的数量，例如为 m。如果某一词条在 V1 和 V2 中都不出现（或者都出现），那么认为该词条不能代表目标信息将舍去。然后对 V1 和 V2 进行归一化，用 V1 表示感兴趣文档集的词条特征向量 U，用 V2 表示不感兴趣文档集的词条特征向量 V。对于普通词汇的词条特征向量 W(T1，W1，T2，W2，…，Tm，Wm)这样进行构造：

当 Ti 在向量 U 和向量 V 中出现则 $W_i=0$，否则 $W_i=\dfrac{1}{m-2n}$

目前在计算机信息处理中对词库中的词通常都不考虑语义的相关性，而是认为各词都相互独立，这样向量 U、V、W 为相互正交的 3 个单位向量。

3.2 页面相关度

页面相关度按照 VSM 中向量之间的夹角来度量，假设用户兴趣模型为 P，未知文档为 S，那么两者之间的相关度为：

$$Rev(S, P) = Cos(S, P) = Cos(S, (\alpha U + \beta V + \gamma W)) \qquad (4)$$

对于经验系数 α、β、γ 的具体取值根据不同的领域不同的专业分词词典而不同。

由于搜索引擎采集的对象一般都是 Web 上的 HTML 文档，HTML 文档中提供了许多标记信息，这些标记信息往往能够反映出内容的重要程度，因此可设置 CofTitle，CofMeta，CofLinkText 等针对 HTML 文档中的{<Title>，</Title>}、{<Meta>，</Meta>}、{<A>，</A>}等域文本的加权系数，对出现在这些域中的词条赋予加权系数，具体的加权系统可按情况而定，但是必须大于 1，一般情况下取大于 2 小于 10 为佳。

根据页面相关度可以判断页面是否属于该领域，设定兴趣阈值为 ε，$0 \leq \varepsilon < 1$ 如果相关度 $Rev(S, P) > \varepsilon$ 则属于该领域，否则不属于该领域。通常设置 ε=0。

4 页面相关度预测及采集控制策略

目前大多数的 Web 检索系统，都是首先将所有页面采集下来，然后进行分析处理，存储相关的信息，进行检索。但是对于检索特定领域信息的系统来说，它的 Robot 采集回来的大部分是其它领域不相关信息，这样不但浪费网络带宽而且非常的耗时。为此，需要根据站点结构，利用页面相关度预测来控制页面的采集。

4.1 页面相关度预测

由于页面之间的链接可以一定程度上反映页面内容之间的联系，所以可以通过指向未知页面 h 的所有页面 P 的相关度来对该页面的相关度进行预测[13]：

$$V_{sim}(h) = \frac{\sum_{p \in P} V_{sim}(p)}{|P|} \tag{5}$$

由于站点的物理结构也能一定程度上反映页面内容之间的联系，所以我们对 5 种类型的链接都赋予不同的加权系数，假设 P 以下行链、上行链、水平链、交叉链、外向链指向 h 的页面分别 P_1、P_2、P_3、P_4 和 P_5。那么加权后的页面相似度预测公式为：

$$V_{sim}(h) = \frac{\alpha \sum_{p \in P_1} V_{sim}(p) + \beta \sum_{p \in P_2} V_{sim}(p) + \delta \sum_{p \in P_3} V_{sim}(p) + \gamma \sum_{p \in P_4} V_{sim}(p) + \lambda \sum_{p \in P_5} V_{sim}(p)}{|P|} \tag{6}$$

其中 $V_{sim}(p)$表示页面 P 的相关度；α、β、δ、γ、λ 分别表示对下行链、上行链、水平链、交叉链、外向链的加权系数。由于水平链的两页面属于同一栏目内容，所以水平链的加权系数最大；其余的依次是下行链、上行链、、交叉链和外向链。实际系统中只有水平链的加权系数比较有实际意义。

4.2 页面采集控制策略

目前大多数的 Robot 使用的搜索算法都是宽度优先或者深度优先算法，利用页面相关度预测就可以设计一种指定范围内的相关度优先算法。在这种算法中，根据页面相关度的预测结果从中选取出相关度最大的作为下一次采集的候选，具体算法如下：

1) 设定 Robot 需要采集的一个或者多个范围，对于每一个范围设定资源限制(时间或者页面数量)并设定采集的起始页面；
2) 对于每一个采集范围，首先采集起始页面 P_0，将其内部的所有链接提取出去，利用相关度预测算法进行相关度预测，然后提取出预测相关度最大的链接进行下一次采集；
3) 采集到页面文本，进行兴趣模型的相关度判断，根据相关度对此页面所有链接进行加权、累计。
4) 从待采集链接库中，提取出预测相关度最大的链接进行下一次采集，循环到（3）这个操作，直到已经达到资源限制的上限。

5 系统测试

依据上述思想，开发出一个初步的特定领域的 Web 信息采集器原型，系统运行的环境为 Sun Ultra10+Solaris 2.6。首先人工从人民日报站点 http://www.peopledaily.com.cn 上采集到 3000 余篇体育类文档和 3000 余篇军事类文档，利用这些语料进行用户兴趣模型训练，对兴趣模型 $P = \alpha U + \beta V + \gamma W$ 中的向量长度选取为 400，对国中网站点 http://www.china.com/zh_ch 军事栏目和体育栏目进行测试，设定各 1000 篇文档，对于经验系数 α、β、γ 选取的值可得到系统查全率和系统精度如表 1 所示：

表1 系统测试结果对比表

α	β	γ	系统采集率	系统精度
1	−1	−0.3	72.4%	100%
1	−1	−0.2	89.3%	92.4%
1	−1	−0.1	93.4%	82.1%
1	−1	0	100%	67.4%

由于用户兴趣模型训练的文档集与测试的文档集都是体育类和军事类文档，如果换成其它类的测试文档集系统性能可能会有一些降低。总体看来，通过预测和兴趣模型的控制，可以达到优先定向采集的目的，在预测优先采集的基础上，能保证一定的采集精度。在具体的参数设置上针对不同的情况还需要不同的调整和优化。

6 结束语

在Internet信息膨胀的今天，快速而有效地采集到Web上有用的资源对于特定领域信息专用检索系统来说是至关重要的。本文在对Web站点结构进行分析后，根据正反集文本过滤方法，构造出用户兴趣模型，提出了页面的相关度的预测和页面采集控制信息策略。有效的缩短采集时间、减少信息存储冗余、加快检索时间，也节约了可贵的网络资源。

参考文献:

[1] 张卫丰等，Web搜索引擎框架研究，计算机研究与发展 2000年3月

[2] 姚国祥等，网上信息搜索技术与搜索引擎，计算机科学 2000年7月

[3] Yan T W, Molina HG, SIFT—A tool for wide-area information dissemination, In: Proc of 1995 USENIX Technical Conf. Http://www-db.standford.edu/pub/yan

[4] Yan T W, Molina HG, Distributed selectvie dissemination of information, In: Proc of 3rd Internaional Conference on Parallel and Distributed Information System. Austin, Texas, 1994

[5] 林鸿飞等，中文文本过滤的信息分流机制，计算机研究与发展 2000年4月

[6] 田范江等，进化式信息过滤方法研究，软件学报 2000年4月

[7] TIAN Fan-jiang, Wang Xi-dong, WANG Ding-xing, Efficient Word-Wide-Web Information Gathering, Journal of Software Jan 2001

[8] 王继成等，基于Internet的信息资源发现技术与实现，计算机研究与发展 1999年11月

[9] 余智华，WWW站点的分析与分类，硕士学位论文 中科院计算所 1999年6月

[10] Berner-Lee, T., Masinter, L., and M. McCahill, Uniform Resource Locators, RFC1738 December 1994

[11] Ellen Spertus, ParaSite:Mining Structural Information on the Web, The Sixth Internatinal World Wide Web Conference, April 1997

[12] 邹涛等，WWW上的信息挖掘技术及实现，计算机研究与发展 1999年8月

[13] Dunlop, M. D., Rijsbergen, C. J. van, Hypermedia and free text retrieval, Information Processing and Management, March 1993

致谢 中科院计算所的唐卫清研究员对本文的工作给予了细心的指导，并对本文的完成提出了很多有益的建议，在此表示感谢.

作者简介: 李振星（1972），男，北京航空航天大学机械工程及自动化学院博士研究生。在中科院计算所与唐卫清研究员、硕士研究生徐泽平合作，做有关web信息定向智能采集，大规模web信息的处理及web信息智能查询方面的研究，重点针对中文web信息的采集、处理、查询。

Forecast and Filter Method for Web page Gathering Based on Interested Model

Li Zhen-xing[1] Xu Ze-ping[2]

[1](School of Mechanical Engineering & Automation, BUAA, Beijing 100083, China);
[2](Institute of Computing Technology, CAS, Beijing 100080, China)
E-mail: zhenxing_li@sina.com

Abstract: Following rapid expansion of huge information on Web, the efficient Web information gathering on specified fields becomes more important in information retrieval research. Based on the interested model of user, this paper presents the Forecast and Filter Method for Web page Gathering. The method applies text filter with plus and minus sets provided by user to design the interested model. Forecast for the relativity of Web page controlled the gathering, based on the analysis of Website structure. Gathering time shortened, storage decreased, retrieval speeded, net resources saved.

Key words: Information Gathering; Interested Model; Text Filtering