搜索引擎中基于分类的网页更新方法研究

文坤梅　卢正鼎

（华中科技大学计算机科学与技术学院　武汉 430074）

摘　要　网络无限扩张的同时网页也在频繁地变化，搜索引擎必须定期更新它所检索的网页，耗费了大量时间和系统资源，因此提高更新效率是搜索引擎的关键技术。比较了目前存在的两种更新方法：统一更新方法和个体更新方法，指出其优劣所在。然后提出一种改进的基于分类的网页更新方法，并从理论上论证了其优化性。实验分析表明，分类更新方法很大程度上提高了网页更新效果。

关键词　搜索引擎，更新度，更新策略

Research on Classified Refresh Policy of Web Page in Search Engine

WEN Kun-Mei　LU Zheng-Ding

(Computer Science of Huazhong University of Science and Technology, Wuhan 430074)

Abstract　The Web is huge and the Web pages are updated frequently, the index maintained by a search engine has to refresh Web pages periodically. This is extremely time and resource consuming because the search engine needs to crawl the Web and download Web pages to refresh its index. So improving the refresh efficiency is the key technology of the search engine. We compare uniform refresh policy and proportional refresh policy, prefer the advantages and disadvantages of both policies. And then this paper presents a reformed method called classified refresh policy. Finally the paper demonstrates its optimization in theory. The experiments shows that the classified refresh policy improves the refresh effect obviously.

Keywords　Search engine, Freshness, Refresh policy

1　引言

随着网络的出现，信息供应过多，甚至阻碍了有用信息的发现，解决这一问题的技术大部分都来自于信息检索领域的技术，如搜索引擎[1]。搜索引擎依靠机器人[2](Crawler)浏览 Web 网，Crawler 被给定 url 开始集，并从网络中检索这些网页。Crawler 提取被检索网页中出现的 url，继续进行本地索引，直到没有新的 url 出现。Crawler 保存检索过的网页，形成一个大型的本地网页库。

网上的信息资源不断变化，Crawler 需要不断更新它所访问过的网页。而不同的网页其改变速度也不相同，因此，Crawler 必须判断重新访问哪些网页，以及用怎样的频率去访问，这些决定都会影响被检索网页集合的更新度。由于系统资源有限，Crawler 只能下载并更新有限的网页集合，网页更新方法在很大程度上决定了网页更新的效果。

2　网页更新的基本概念

2.1　更新度

一旦 Crawler 检索了某些重要的网页，它就要定期地更新这些页面。下面将定义更新度及网页“年龄”的概念。

$C=\{p_1,\cdots,p_n\}$是 n 张网页的本地集合。下面定义该集合的更新度[3]。

本地网页 P_i 在 t 时刻的更新度为：

$$U(P_i)=\begin{cases}1 & \text{如果 } P_i \text{ 在 } t \text{ 时刻更新了；}\\ 0 & \text{否则。}\end{cases}$$

更新是指本地网页中的内容等于真实世界中网页的内容，于是，本地集合 C 在 t 时刻的更新度是：

$$U(C)=(U(P_1)+\cdots+U(P_n))/n$$

$$=\frac{1}{n}\cdot\sum_{i=1}^{n}U(P_i) \qquad (1)$$

事实上，更新度就是本地集合中更新过的页面占集合中总页面的百分比。

2.2　网页更新方法

目前，已存在的方法[4]都可归为以下两类，不同的更新方法会得到不同的更新结果。

- 统一更新方法：Crawler 以相同频率 f 访问所有网页，不考虑网页的改变频率；
- 个体更新方法：不同网页其改变频率也不同，Crawler 根据个体页面的改变频率来重访各页面，

文坤梅　博士，主要从事网上信息挖掘、数据库方面的研究工作。卢正鼎　教授，博士生导师，主要从事数据，主要从事数据库、分布式方面的研究工作。

网页的改变频率与访问频率其比率对任何个体网页来说都是相等的。

Crawler 更新网页集合 C，C 包括 p_1、p_2 页面。其中，p_1 每天改变五次，p_2 每天改变一次。假定 Crawler 一天只能更新一张网页，如果 p_2 在每天的中间时刻改变，而正好在 p_2 改变之后更新它，那么 p_2 将在半天保持为更新状态。考虑到 p_2 在上半天改变的概率是 1/2，因此，选择更新 p_2 时 C 的更新度是 $1/2\times1/2=1/4$ 天。同理，选择 p_1 其更新度为 $1/2\times1/10=1/20$ 天，由此看来，选择更新 p_2 会更有效。上例说明，在任何情况下改变太快的网页都不应该被频繁访问。这与个体更新方法恰恰相反，因此，统一更新方法总是优于或等于个体更新方法。

3 基于分类的网页更新方案

频繁访问改变太快的网页不能明显提高搜索效果，而应该把资源集中在改变速度适中的网页上。如此将会产生一个问题：改变频繁的网站长期得不到更新。因此，在统一更新方法的基础上提出一种改进方案，在此称之为分类更新方法。

3.1 基本思想

通过网页的改变历史来评估其改变频率，设定一个频率阈值，将网页分为两类：更新较快网页子集和更新较慢网页子集，然后以不同的频率访问这两类网页，这就是分类更新的基本思想。例如：在实际应用中，Crawler 除了每月统一更新所有网页之外，同时每星期还更新一次改变频率较大的网页子集。具体实现方法如下：

有 n 张网页，其改变频率的估值分别为 $f_1, f_2, \cdots f_n$；其统一更新频率为 f_0：若频率阈值为 f'，将网页分为两类：

$$\begin{cases} F_1=\{f_i \mid f_i \geqslant f'\} \\ F_2=\{f_i \mid f_i < f\} \end{cases};$$

则分类更新方法将以频率 f 访问 F_1 中网页，同时以统一更新频率 f_0 访问所有网页。

其中阈值 f'、频率 f 的设定与具体应用相关，若无特别指定，一般情况下，可取 $f'=2f_0$，$f=\frac{1}{k}\sum_{i=1}^{k},(f_i\in F_1)$。

3.2 方法的有效性

方法的有效性在于对网页改变频率的估算是否准确。设定一段监控时间 T，假设在 T 时间内网页改变 X 次，那么网页改变频率的估算值就为：X/T。

假定每隔时间 t 将访问网页 p，一共访问了 n 次。用 X(i)表示在第 i 次访问中网页变化与否，即：

$$X(i)=\begin{cases} 1 & \text{如果在第 i 次访问时网页改变了；} \\ 0 & \text{否则。} \end{cases}$$

于是，网页总改变次数 $X=\sum_{i=1}^{n}X_i$，总访问时间 T=nt=n/f，其中 f 是访问网页的频率。由于网页的改变是随机并彼此独立的，因此服从泊松分布，假定其改变频率为 f′，则网页的改变频率与访问网页的频率之比率 b=f′/f。

b 的估计值为 $\hat{b}=f/f=(1/f)(X/T)=X/n$。

事实上，b 的估计值总比其真实值小，因为监控所得的 X 值总比实际网页改变的次数少。可以通过计算 $\hat{b}$ 的期望 $E(\hat{b})$，来比较 $\hat{b}$ 和真实值 b 之间的差值。在时间 t 内，网页不改变的概率为：

$$p=P(X(ti+t)-x(ti)=0)=\frac{f'^0e^{-f't}}{0!}$$
$$=e^{-f't}=e^{-f'/f}=e^{-b}$$

即 X(i)为 0 的概率为 p，为 1 的概率为 1−p，其中 $P=e^{-b}$；

由上可得 $P(X=i)=C_n^i(1-P)^iP^{n-i}$；

$$E(\hat{b})=\sum_{i=0}^{n}\frac{i}{n}\cdot P(\hat{b}=\frac{i}{n})=\sum_{i=0}^{n}\frac{i}{n}\cdot P(X=i)$$
$$=\sum_{i=0}^{n}\frac{i}{n}\cdot C_n^i\cdot(1-P)^iP^{n-i}$$
$$=\sum_{i=1}^{n}C_{n-1}^{i-1}(1-P)^iP^{n-1}=(1-P)\cdot\sum_{i=1}^{n}C_{n-1}^{i-1}(1-P)^{i-1}P^{n-i}$$
$$=(1-p)((1-p)+p)^{n-1}=1-e^{-b}。$$

很显然，b 的估计值有偏差，且 $E(\hat{b})/b=(1-e^{-b})/b$，如图 1 所示，把 b 控制在(0，0.2)之间，即 $0<f'/f<0.2$ 时，误差小于 10%。

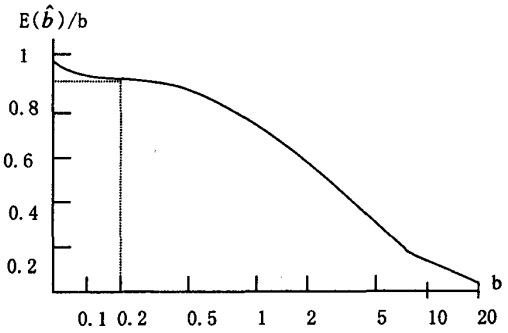

图 1 b 的估计值与真实值之间的关系图

4 实验与分析

选取新浪网中的十张网页，其改变频率如表 1 所示。

表 1 网页的改变频率

网页	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
改变频率(次/小时)	1	0.8	0.9	1.1	1.2	1/12	1/10	1/9	1/11	1/8

(下转第 16 页)

集。其设计过程是对用户输入切句切词(即调用搜引器设计过程中的语句分割模块和切词模块),再对切词结果依据知识库进行同义词扩展,得到一组用向量表示的概念集,该概念集与索引库中的网页向量进行向量内积运算得到相似度,以相似度大小的顺序排序输出给用户界面。

结论 本文给出了基于概念的智能搜索引擎的理论模型,并从编程与算法角度详细介绍了决定搜索引擎好坏的索引器的建立过程,整个设计中借助了知识库的各种知识,把词从表面含义提升到概念层次,提高了搜索引擎的查全率与查准率。但是该知识库有待进一步扩展与提高,这还要依靠人工智能、推理机、自然语言理解等诸多科技研究的结果,待知识库内容逐渐丰富壮大后,不仅能够提高查全率和查准率,还能从真正意义上使搜索引擎达到智能化,提高网络利用率。

参考文献

1 何绍义. 概念信息检索的理论与实践[J]. 情报学报,1995,14(2)

2 袁占亭,张爱民,张秋余. 基于概念的 Web 信息检索[J]. 计算机工程与应用,2003,39(23)

3 (美)Ayers D 等,著. 王辉,等译. Java 数据编程指南[M]. 北京:电子工业出版社,2002

4 (美)Heaton J 著. 童兆丰,等译. 网络机器人 Java 编程指南[M]. 北京:电子工业出版社,2002

5 (加拿大)Patterson L 著,徐征,等译. HTML 4 编程指南[M]. 杭州:浙江科学技术出版社,1999

(上接第 2 页)

其中,$P_1,P_2,\cdots,P_5$属于新闻类网页,因此更新速度较快,$P_6,P_7,\cdots,P_{10}$属于天气类网页,因此更新较慢。

若 Crawler 以 1/6(次/小时)统一更新所有网页,即 $f_0=1/6$(次/小时),$f'=1/3$(次/小时),根据分类更新方法,将网页分为两类 $F_1=\{P_1,P_2,P_3,P_4,P_5\}$,$F_2=\{P_6,P_7,P_8,P_9,P_{10}\}$,以 $f=1$(次/小时)访问 F1,同时以 $f_0=1/6$(次/小时)统一更新所有网页。

下面通过计算更新度来比较分类更新方法和统一更新方法的有效性。个体更新方法在此不做考虑,上文已说明统一更新方法优于个体更新方法。在网络容量巨大的情况下,系统无法实现个体更新方法。

各网页在不同更新方法下所获得的更新度如表 2 所示。

表 2 各网页在不同更新方法下所获得的更新度

网页	统一更新方法	分类更新方法
P_1	1×(1/6)	1×(1/2)+1×(1/6)
P_2	(1/0.8)×(1/6)	(1/0.8)×(1/2)+(1/0.8)×(1/6)
P_3	(1/0.9)×(1/6)	(1/0.9)×(1/2)+(1/0.9)×(1/6)
P_4	(1/1.1)×(1/6)	(1/1.1)×(1/2)+(1/1.1)×(1/6)
P_5	(1/1.2)×(1/6)	(1/1.2)×(1/2)+(1/1.2)×(1/6)
P_6	(1/12)×(1/6)	(1/12)×(1/6)
P_7	(1/10)×(1/6)	(1/10)×(1/6)
P_8	(1/9)×(1/6)	(1/9)×(1/6)
P_9	(1/11)×(1/6)	(1/11)×(1/6)
P_{10}	(1/8)×(1/6)	(1/8)×(1/6)

由更新度计算公式(1)可得:

统一更新法所得更新度 $U_1=U_1(P_1)+U_1(P_2)+\cdots+U_1(P_{10})=0.7875$

分类更新法所得更新度 $U_2=U_2(P_1)+U_2(P_2)+\cdots+U_2(P_{10})=3.3393>0.7875$

可知 $U_2>U_1$,因此,分类更新方法能获得比统一更新方法更大的更新度。

结束语 分类更新方法结合前两种方法的优点,不会将系统资源耗费在过度更新改变过于频繁的网页上,也不会过多访问改变缓慢的网页,而是均衡的分配系统资源。实际上,分类更新方法的基本思想可以继续扩展,在实际应用中,网络容量日益膨胀,网页改变速度各不相同,统一更新方法不能适应用户对信息更新度的要求,因此,可以根据网页的改变速度把网络化分为不同子集,再根据各子集的改变频率来更新网页集合,这也可以满足用户对某些网页集合的特殊要求,如实时更新。

参考文献

1 宋聚平,王永成,尹中航,滕伟. 面向主题的网页搜索系统. 上海交通大学学报,2003,37(3):401~403

2 张领,叶允明,等. 一种高性能、分布式 WEB CRAWLER 的设计与实现. 上海交通大学学报,2004,38(1):59~61

3 Cho J, Garcia-Molina H. Synchronizing a database to Improve Freshness. In: Proc. of 2000 ACM Intl. Conf. on Management of Data (SIGMOD) Conf. Dallas, Texas, United States. 2000. 117~128

4 Arasu A, Cho J, Garcia-Molina H, Paepcke A, et al. Searching the Web. ACM Trans. on Internet Technology, 2001, 1(1): 2~43

搜索引擎中基于分类的网页更新方法研究

作者： 文坤梅， 卢正鼎， WEN Kun-Mei， LU Zheng-Ding
作者单位： 华中科技大学计算机科学与技术学院,武汉,430074
刊名： 计算机科学 ISTIC PKU
英文刊名： COMPUTER SCIENCE
年，卷(期)： 2004,31(z1)
被引用次数： 1次

参考文献(4条)

1. Arasu A;Cho J;Garcia-Molina H;Paepcke A Searching the Web 2001(01)

2. Cho J;Garcia-Molina H Synchronizing a database to Improve Freshness 2000

3. 张领;叶允明 一种高性能、分布式WEBCRAWLER的设计与实现[期刊论文]-上海交通大学学报 2004(01)

4. 宋聚平;王永成;尹中航;藤伟 面向主题的网页搜索系统[期刊论文]-上海交通大学学报 2003(03)

引证文献(1条)

1. 陈晓志.董守斌.张凌.张元丰 基于URL类型和网页链接变化的信息采集更新算法[期刊论文]-郑州大学学报（理学版） 2007(2)