基于用户兴趣与主题相关的PageRank算法改进研究

王 冲 纪仙慧
(桂林电子科技大学计算机科学与工程学院 桂林 541004)

摘 要 针对传统的PageRank算法存在主题漂移、忽略用户兴趣等不足,提出一种基于用户兴趣与主题相关的PageRank改进算法——ITPR。为了更好地提高用户搜索质量,利用网页浏览时间与页面篇幅共同构建用户兴趣度因子,用线性拟合月点击量的方法预测用户兴趣度的升降,同时结合网页内容引入主题相关度因子,共同对网页PR值进行适当的修正,使其分配更为合理。仿真实验结果表明,在相同的实验环境下,改进的PageRank算法提升了网页排序质量、查准率以及用户搜索满意度。

关键词 PageRank,用户兴趣,线性拟合,兴趣度预测,主题相关度
中图法分类号 TP391.3 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.3.051

Improved PageRank Algorithm Based on User Interest and Topic

WANG Chong JI Xian-hui
(College of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract Aiming at the drifting theme and ignoring user interest of traditional PageRank algorithm, an improved algorithm based on user interest and topic(ITPR) was proposed. In order to satisfy the needs of user better, both browsing time and page length were used to build user interest factor, and its change was predicted by the linear fitting of the hits per month. Meanwhile, topic correlation factor based on page content was introduced, modifying the PR appropriately. The simulation experiment results show that the proposed algorithm achieves better page ranking quality, precision ratio and user's satisfaction.

Keywords PageRank, User interest, Linear fitting, Interest prediction, Topic relevant

1 引言

随着互联网技术的迅猛发展,网络已成为人们获取信息的重要途径和手段。中国互联网络信息中心(CNNIC)在《第34次中国互联网络发展状况统计报告》中指出[1]:"截至2014年6月,中国网民规模达6.32亿,其中,手机网民规模5.27亿,互联网普及率达到46.9%"。面对呈几何级增长的海量数据,如何快速准确地获取有效信息成为搜索引擎所面临的严峻挑战。网页排序算法作为搜索引擎的关键技术之一成为大家共同关注的焦点。

传统基于关键字匹配的网络搜索引擎在查询效果上不太理想,因为网页设计者只需在页面重复添加关键字便可以使网页搜索排序靠前[2]。基于链接结构的网页排序算法PageRank算法在搜索引擎Google中的成功运用,证明了该算法具有一定的应用价值和研究价值。随着国内外学者对PageRank算法研究的深入,该算法存在的不足之处逐渐凸显,许多学者针对其不足进行了算法改进。改进方案主要分为两大类,分别用于改进PageRank算法的搜索效率和排序效果[3],本文着重研究改进算法的排序效果。

2 算法分析

2.1 PageRank算法

斯坦福大学计算机学院研究生Lawrence Page和Sergey Brin借鉴引文分析思想于1998年提出一种基于网络链接分析的PageRank算法[4]。该算法的主要思想为:一个网页通过超链接链向其他网页,代表向链出网页投了一票,并将自身的PR值(PageRank值)平均分配给链出网页。网页的PR值越高,则权威性越高,网页搜索排名相应靠前。具体来说,网页PR值由3个因素决定[5]:(1)一个网页的链入网页越多,该网页排名越高;(2)一个网页链入网页的排名越高,该网页排名相应也高;(3)一个网页链入网页的出度越少,该网页排名越高。

综述以上3个因素,RageRank算法原始公式为:

$$PR(u)=\sum_{v\in I(u)}\frac{PR(v)}{Out(v)} \tag{1}$$

式中,$I(u)$为网页u的链入页面集合,即所有链向网页u的页面集合;$Out(v)$表示网页v的出度,即网页v的链出页面数目。

Web网页的PR值运用随机行走模型[6]计算,用户可以

到稿日期:2015-01-11 返修日期:2015-03-22 本文受广西可信软件重点实验室项目(PF14071X),桂林电子科技大学重点教改项目(ZJW07303),广西教改工程项目(2015JGA207)资助。

王 冲(1972—),男,副教授,硕士生导师,主要研究方向为信息检索技术、多媒体教育软件技术,E-mail:107892769@qq.com;纪仙慧(1990—),女,硕士,主要研究方向为信息检索技术,E-mail:2457653070@qq.com。

通过点击超链接浏览网页，也可以直接键入网址随机访问。于是引入阻尼因子 d（d 通常取 0.85），即网页冲浪者在任意时刻将以概率 d 从当前网页通过超链接均匀地访问其链出网页，以概率 $1-d$ 从整个 Web 网页中均匀选择一个网页进行浏览。因而 RageRank 算法的公式为：

$$PR(u)=(1-d)+d\sum_{v\in I(u)}\frac{PR(v)}{Out(v)} \tag{2}$$

2.2 Weighted PageRank 算法

传统的 PageRank 算法仅依据网页的链出结构，将 PR 值均匀分配于链出网页，使得权威性高的网页不能快速上浮。Xing 等通过对网页链接结构的深入分析，综合考虑网页的链入、链出结构，提出了加权算法（Weighted PageRank，WPR）[7]，即利用网页的链入链接个数和链出链接个数，对网页间链接的权重添加基于网页入度的权重因子和基于网页出度的权重因子，其计算公式为：

$$PR(u)=(1-d)+d\sum_{v\in I(u)}W^{in}_{(v,u)}W^{out}_{(v,u)}PR(v) \tag{3}$$

式中，$I(u)$表示网页 u 的链入网页集合。$W^{in}_{(v,u)}$ 为基于页面 u 入度和页面 v 链出网页入度之和的入链因子，$W^{out}_{(v,u)}$ 为基于页面 u 出度和页面 v 链出网页出度之和的出链因子。$W^{in}_{(v,u)}$ 与 $W^{out}_{(v,u)}$ 的计算公式分别如下：

$$W^{in}_{(v,u)}=I_u/\sum_{p\in O(v)}I_p \tag{4}$$

$$W^{out}_{(v,u)}=O_u/\sum_{p\in O(v)}O_p \tag{5}$$

其中，I_u 为网页 u 的入度，O_u 为网页 u 的出度，$O(v)$为网页 v 的链出网页集合。

WPR 算法改善了传统 PageRank 算法权值平均分配的不足。Xing 等的研究表明，WPR 算法的排序结果较传统的 PageRank 算法更为理想[8,9]，同时，加权因子的引入为后人对传统 PageRank 算法的改进提供了较为开放的研究思路。但 WPR 算法仍然仅考虑了网页的链接结构，与传统PageRank 算法同样存在以下不足：

(1)主题漂移。由于与主题查询词无关，查询到的网页可能权威值高但并非相关内容。

(2)偏重旧网页。旧网页存在时间长，获得链接的几率变大，PR 值也相应可能更高。

(3)忽略用户反馈。用户的浏览行为包含大量的价值信息，以上两种算法均未做分析。

2.3 相关改进算法

斯坦福大学的 Taher Haveliwala 利用对用户查询主题与网页主题的相似度进行计算的方法，提出了 Topic-Sensitive PageRank 算法。该改进算法依据 Open Directory Project (ODP)，提出 16 个基本分类标准，用户查询时，计算出用户查询主题与已知 16 个基本主题的相似度，并从中选择一个最接近的基本主题代替用户查询主题，在一定程度上改善了主题漂移问题[10]。李卫东[11]融合向量空间模型技术，王钟斐[12]结合锚文本相似度，也针对主题漂移现象分别对 PageRank 算法进行改进，其基本思想均为将向量空间相似度算法与 PageRank 算法相结合，但计算都相对复杂且均未考虑用户反馈信息。

针对偏重旧网页的现象，Philip. S Yu 等将时间作为一种权重因子融入 PageRank 计算过程，早在 2004 年就提出了 TimedPageRank 思想。北京大学计算机系[13]利用 HTTP 协议记录每个页面最近一次的修改时间，并将页面修改时间作为控制参数，对新修改的页面给予较高的权值，旧页面给予较低权值。段淮川和胡平[14]将与网页存在时间呈正比关系的搜索引擎服务器搜索到该页面的次数作为时间因子，使得 PR 值随时间的变化而浮动。

用户浏览网页时，可以被采集的反馈信息大体分为 4 类：链接点击行为(点击次数等)、鼠标行为(鼠标移动时间等)、激活行为(网页浏览时间等)和收藏行为(下载打印等)。Gyanendra Kumar 等[15]利用网页链接 $v\to u$ 的点击量占网页链接 $v\to p$(网页 p 为网页 v 的链出网页集合)总点击量的比重作为网页 u 传递 PR 值的依据。该算法利用链接点击行为对 PageRank 算法作出了改进，但链接点击是用户未浏览网页内容时仅依据网页标题或网页推荐等因素作出的初步选择，并不代表用户认可网页内容。彭聪等[16]在 Gyanendra Kumar 思想的基础上，用网页停留时间代替了网页链接点击量，更充分地体现了用户偏好，但并未考虑到网页篇幅对网页停留时间有着一定的影响。

方树峰[17]利用用户反馈信息，提出改进算法 FPR (PageRank based on Feedback)。该算法统计网页点击量，设计点击次数权重；通过计算用户及时搜索时间与页面最近一次被用户搜索点击的时差，设计点击时间权重；根据查询关键词出现的不同位置分配不同权重，权重累加之和代表网页内容权重。

3 改进的 PageRank 算法——ITPR 算法

在 ITPR 算法(PageRank based on Interest and Topic)中，借鉴 WPR 算法思想，引用入链因子和出链因子；融合网页停留时间、页面篇幅、用户正常阅读速度等因素，添加用户兴趣度因子；通过线性拟合的方法，预测未来一段时间内用户对某一网页兴趣的发展趋势，引入兴趣预测因子，对 PR 值进行修正调整；依据 Lucene 中采用的对网页内容和查询语句相关度打分的策略，设计主题相关度因子。ITPR 算法的计算公式如下：

$$PR(u)=(1-d)+d[\sum_{v\in I(u)}PD(u)(\alpha\, interest(u)+\beta W^{in}_{(v,u)}+\gamma W^{out}_{(v,u)})PR(v)+\eta TCR(u)] \tag{6}$$

式中，$PD(u)$是兴趣预测因子，$interest(u)$是用户兴趣度因子，$W^{in}_{(v,u)}$是入链因子，$W^{out}_{(v,u)}$是出链因子，$W^{in}_{(v,u)}$与 $W^{out}_{(v,u)}$的计算公式分别如式(4)和式(5)所示，α、β、γ为权重因子，分别表示用户兴趣度因子、入链因子、出链因子所占的比重，且满足 $\alpha+\beta+\gamma=1$；$TCR(u)$是主题相关度因子，η是调整主题相关因素对排序算法影响比重的权重因子。

3.1 用户兴趣度因子 *interest*(*u*)

用户在浏览网页时，对感兴趣的网页页面停留时间相对较长。考虑到不同网页的页面内容篇幅不同，用户浏览网页所需的时间不同，用平均停留在每个字上的浏览时间(简称字停留时间)来衡量用户对不同网页的兴趣程度，并对网页 v 所有链出网页的字停留时间进行归一化处理。用户兴趣度因子计算公式如下：

$$interest(u)=\frac{S(v\to u)}{\sum_{p\in O(v)}S(v\to p)} \tag{7}$$

式中，$S(v\to u)$为通过点击链接 $v\to u$ 浏览网页 u 时，页面 u 的字停留时间。$O(v)$为网页 v 链出网页的集合。其中：

$$S(v\to u)=\frac{scan(v\to u)}{click(v\to u)N(u)} \tag{8}$$

$$N(u)=C_w(u)+C_p(u)\times 50+C_v(u)\times 100 \quad (9)$$

式(8)中，$scan(v\to u)$为通过点击链接 $v\to u$ 浏览网页 u 时页面 u 的停留时间之和，即页面总停留时间。$click(v\to u)$为通过点击链接 $v\to u$ 浏览网页 u 的总点击量。$N(u)$为页面总字数。式(9)中，$C_w(u)$为网页文字个数，$C_p(u)$为网页图片个数，$C_v(u)$为网页视频个数。为了方便计算，将图片和视频分别转换为 50 和 100 个文字。页面停留时间 T 的计算公式如下所示：

$$T=\begin{cases}0, & t<\tau\\ t, & \tau\leqslant t<2\tau\\ 2\tau, & t\geqslant 2\tau\end{cases} \quad (10)$$

假设用户对网页 u 的正常阅读时间为 τ，当页面停留时间小于 τ 时，表示用户对此页面不感兴趣，页面点开后未经正常阅读就被关闭，称此次点击无效，且不计入总点击量，页面停留时间设为 0；当页面停留时间大于 τ 时，称此次点击有效，并将之计入总点击量，页面停留时间计入页面总停留时间。为了避免一个网页被遗忘关闭或是其他原因造成的长时间被打开状态，页面停留时间大于 2τ 时按照 2τ 计入页面总停留时间。$\tau=N(p)/280$，因为正常人一般阅读速度为 280 字/分[18]。

3.2 兴趣度预测因子 $PD(u)$

网页月点击量的变化在一定程度上反映着用户对该网页兴趣度的变化，通过分析基于时间序列的网页月点击量的变化情况，预测在未来一段时间内用户对该网页兴趣度的升降情况。假设 D 为某一网页创建至今前 $n(n=1,2,3,\cdots)$个月的网页月点击量之和的数据集，$D=\{d_1,d_2,d_3,\cdots,d_n\}$，将 D 中每个元素以 d_1 为单位进行单位化(d_i/d_1，$0<i<n+1$)，并构建集合 $M=\{m_1,m_2,m_3,\cdots,m_n\}$，其中 $m_i=d_i/d_1$，对基于 n 的 m_i 进行线性拟合，拟合直线的斜率代表了网页月点击量的变化趋势：斜率大于 1，表示该网页的月点击量增加，用户对该网页的兴趣呈上升趋势；斜率小于 1，表示该网页的月点击量降低，用户对该网页的兴趣呈下降趋势。网页的月点击量拟合直线如图 1 所示。

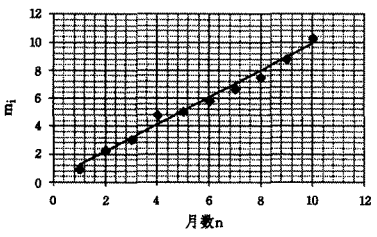

图 1 网页的月点击量拟合直线

依据网页月点击量的变化趋势，引入兴趣度预测因子 $PD(u)$(即拟合直线的斜率 k)预判用户兴趣度。由式(6)可见，一个网页从链入网页传递的 PR 值越高，受兴趣度预测因子的影响就越大，可使得用户兴趣减弱的价值网页(我们定义 $[\alpha\, interest(u)+\beta W^{in}_{(v,u)}+\gamma W^{out}_{(v,u)}]PR(v)$值高的为价值网页，否则为普通网页)迅速下沉，用户兴趣上升的普通网页相对上浮。兴趣度预测因子的计算公式如下：

$$PD(u)=\begin{cases}1, & n=1\\ k, & n>1\end{cases} \quad (11)$$

3.3 主题相关度因子 $TCR(u)$

传统的 PageRank 算法和 WPR 算法存在主题漂移问题，主要由于两种算法仅依据链接结构，均未考虑网页内容与主题的相关性。ITPR 算法依据 Lucene 中应用的对网页内容和查询语句相关度的打分策略来获得主题与网页内容的相关度因子 $score(q,u)$，即 $TCR(u)$。TF-IDF 模型作为一种加权策略，广泛应用于信息检索、搜索引擎和数据挖掘等方面。$TCR(u)$是一种基于 TF-IDF 的评分方式，受多个因素共同决定：(1)一条查询可能包含多个词项，对命中查询语句词项个数多的网页给予较高评分；(2)网页内容中出现查询语句词项频率高的网页的评分较高；(3)由于网页篇幅影响词项出现频率，且词项被较短文本命中更能体现该词项重要性，因此网页篇幅被考虑；(4)如果某个词项在所有网页内容中出现频率较高，该词项可能是普遍词，则得分降低。其打分公式如下[19,20]：

$$score(q,u)=coord(q,u)*queryNorm(q)*\sum_{t\,in\,q}[tf(t\ in\ u)*idf(t)^2*t.getBoost()*norm(t,u)] \quad (12)$$

式中，t 为包含域信息的词项，q 为用户所提交的查询语句；打分 $score(q,u)$为查询语句 q 中每个词项 t 与网页 u 内容的匹配值之和；$coord(q,u)$为协调因子，其基于文档中所包含的被查询词项的个数；$queryNorm(q)$为查询的归一化值，指查询语句中各词项权重的平方和；$tf(t\ in\ u)$为词项 t 在文档 u 中出现的词频；$idf(t)$为词项 t 在倒排文档中出现的频率；$t.getBoost()$为词项 t 在查询语句中的权重；$norm(t,u)$为标准化因子。

主题相关度因子利用 Lucene 打分机制，使内容相关度高的网页获得较高的分值，为了避免垃圾网页重复关键字获取高分，权重因子 η 不宜过大。

4 实验仿真

4.1 实验步骤

为验证 ITPR 算法的可行性与有效性，从本校国际学院网站抓取网页，作为实验数据源，利用 myeclipse、Heritrix、Lucene 工具搭建仿真实验环境，并对算法进行仿真实验对比，实验过程如下：

(1)通过网络爬虫工具 Heritrix 从本校国际学院网站抓取约 5000 个网页，并将它们过滤去噪后做为实验样本存入数据库。

(2)在 myeclipse 平台上，创建新项目并添加 Lucene 3.0.jar 包，利用 Lucene 提供的接口，提取数据库中的网页信息并建立索引。

(3)利用 Java 语言分别实现传统 PageRank 算法、WPR 算法、FPR 算法以及 ITPR 算法。在 ITPR 算法中，取 α 为 0.3，β 为 0.4，γ 为 0.3，η 为 0.3，代入式(6)。

(4)在搭建好的实验平台上，分别采用传统的 PageRank 算法、WPR 算法、FPR 算法和 ITPR 算法对仿真数据进行主题查询，对比分析查询后页面的排序质量。

4.2 实验结果

分别采用“留学项目”、“招生简章”、“留学德国”、“留学法国”、“留学英国”作为查询关键词，在搭建好的仿真实验平台上进行 5 组对比实验。首先以“留学项目”为查询关键词对比分析 4 种算法的页面排序质量。查询结果如图 2－图 5 所示。

次序	网页标题	PR值	创建时间	k
1	英国赫瑞·瓦特大学项目	1.6032664	2008-10...	1.00351...
2	意大利公立大学免学费留学2012年招生简章	0.811113	2012-04...	0.69425...
3	留学德国—灵感与创新的国度	0.7915958	2009-10...	0.98246...
4	加拿大渥太华大学	0.7915958	2010-06...	0.94726...
5	高中生直通美国项目	0.7847648	2010-06...	1.00910...
6	法国大学无正式文凭 国家文凭具同等价值	0.7847648	2009-09...	0.76316...
7	意大利公立大学免学费留学2014年招生简章	0.745242...	2014-05...	1.16574...
8	德国名校--德累斯顿工业大学（免学费）	0.745242...	2013-11...	0.97647...
9	英国考文垂大学留学项目	0.745242...	2011-08...	1.12654...
10	留学项目	0.723285...	2009-04...	1.23125...

图 2　传统的 PageRank 算法前 10 页面排序

次序	网页标题	PR值	创建时间	k
1	英国赫瑞·瓦特大学项目	0.198122...	2008-10...	1.00351...
2	加拿大渥太华大学	0.154050...	2010-06...	0.94726...
3	意大利公立大学免学费留学2012年招生简章	0.151311...	2012-04...	0.69425...
4	留学德国—灵感与创新的国度	0.151133...	2009-10...	0.98246...
5	高中生直通美国项目	0.151078...	2010-06...	1.00910...
6	法国大学无正式文凭 国家文凭具同等价值	0.151078...	2009-09...	0.76316...
7	留学项目	0.150977...	2009-04...	1.23125...
8	意大利公立大学免学费留学2014年招生简章	0.150750...	2014-05...	1.16574...
9	德国名校--德累斯顿工业大学（免学费）	0.150750...	2013-11...	0.97647...
10	英国考文垂大学留学项目	0.150750...	2011-08...	1.12654...

图 3　WPR 算法前 10 页面排序

次序	网页标题	PR值	创建时间	k
1	英国赫瑞·瓦特大学项目	4.0655694	2008-10...	1.00351...
2	中外交流项目	2.3027806	2006-04...	1.00013...
3	英国考文垂留学项目	1.88907	2011-08...	1.26543...
4	留学项目	1.2227827	2009-04...	1.23125...
5	德国哈雷顿工业大学(名校，免学费)招生	0.92901...	2013-08...	0.99034...
6	加拿大渥太华大学	0.83339	2010-06...	0.94726...
7	报考须知	0.77787...	2010-06...	0.93742...
8	高中生直通美国项目	0.7109976	2010-06...	1.00910...
9	法国大学技术学院1.5+2.5	0.6997265	2010-06...	0.87936...
10	意大利公立大学免学费留学2012年招生简章	0.68224...	2012-04...	0.69425...

图 4　FPR 算法前 10 页面排序

次序	网页标题	PR值	创建时间	k
1	英国赫瑞·瓦特大学项目	4.0703464	2008-10...	1.00351...
2	留学项目	1.2992167	2009-04...	1.23125...
3	加拿大渥太华大学	0.839122...	2010-06...	0.94726...
4	高中生直通美国项目	0.7198789	2010-06...	1.00910...
5	意大利公立大学免学费留学2014年招生简章	0.6861755	2014-05...	1.16574...
6	德国名校--德累斯顿工业大学（免学费）	0.668758...	2013-11...	1.07647...
7	意大利公立大学免学费留学2012年招生简章	0.6562659	2012-04...	0.69425...
8	英国考文垂大学留学项目	0.6326603	2011-08...	1.12654...
9	留学德国—灵感与创新的国度	0.631007...	2009-10...	0.96546...
10	国际合作与交流简况	0.6300771	2010-06...	0.91321...

图 5　ITPR 算法前 10 页面排序

对比分析图 2—图 5，与查询关键词“留学项目”最为相关的网页“留学项目”在图 2 中排名第十，在图 3 中排名第七，在图 4 中排名第四，而在图 5 中排名进一步提升至第二，以网页“留学项目”为代表的主题相关网页排名的大幅提升一定程度上抑制了主题漂移现象；又例如，在图 2—图 4 中，由于网页“意大利公立大学免学费留学 2014 年招生简章”创建时间较短，链接结构不够完善，点击量较少，使得包含最新信息的网页“意大利公立大学免学费留学 2014 年招生简章”排名在同类旧网页“意大利公立大学免学费留学 2012 年招生简章”之后。由于旧网页后期点击量急剧下降，而新网页点击量有攀升之势，同时，在 ITPR 算法中，受时间因素影响的入链因子和出链因子在算法中所占比重降低，因此搜索结果如图 5 所示：新网页排名上升，旧网页排名下降，并降于新网页之后，偏重旧网页的现象得到一定的改善。

为评估 4 种算法的排序质量，采用人工评价查询结果集的方式，请 10 位测试人员分别对查询关键词为“留学项目”、“招生简章”、“留学德国”、“留学法国”、“留学英国”的查询结果进行评测，测试人员需将查询后的网页分为非常满意、满意、较满意、不满意 4 个不同级别。《Lucene in Action》中提到，用户一般只对查询结果集中前 20 的网页感兴趣，为了更好地体现改进算法的有效性，选取查询结果集的前 40 个网页交由测试人员评测。

查准率：查准率是一项用来评价搜索引擎的重要的指标，衡量的是系统检索出相关信息的能力。在评测结果中，凡级别为非常满意、满意或是较满意的都属于主题相关的网页，查准率计算公式为：

$$\text{查准率}=\frac{\text{排名前 40 且与主题相关的网页个数}}{\text{40(不足 40 的取搜索结果)}} \tag{13}$$

评测结果显示，ITPR 算法的查准率较 FPR 算法、WPR 算法和传统的 PageRank 算法有所提升，检索主题相关网页的能力更强。查准率对比情况如图 6 所示。

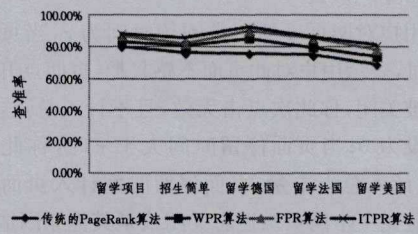

图 6　查准率对比情况

用户满意度评估：利用文献[14]中所采用的满意度评估公式 $S=\sum_{i=1}^{n}(n-i+1)*s_i$ 来评估排序后的前 n 个网页的满足度，其中 i 为 n 个网页中的第 i 个网页，s_i 为用户对第 i 个网页的满意系数。4 个级别即非常满意、满意、较满意、不满意对应的满意系数分别为 1.0、0.6、0.2、0。n 取 40，最终 PageRank 算法对应上述 5 次查询得到的满意度 S 的值为：445.6、415.8、508.6、418.6、372；Weighted PageRank 算法对应上述 5 次查询得到的满意度 S 的值为：466.4、426.2、577、495.8、402.4；FPR 算法对应上述 5 次查询得到的满意度 S 的值为：503.2、439、612.8、522、427.2；ITPR 算法对应上述 5 次查询得到的满意度 S 的值为：533.6、500.4、640.4、536.8、495.6。评测结果显示，ITPR 算法较前 3 种算法，用户感兴趣的网页排名靠前，明显提升了用户满意度，更好地满足了用户的信息搜索需求。用户满意度对比情况如图 7 所示。

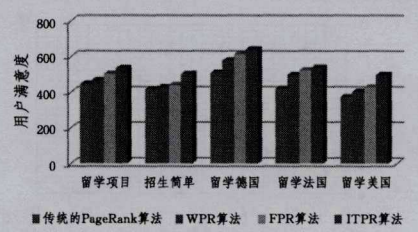

图 7　用户满意度对比情况

结束语　在分析传统 PageRank 算法不足的基础上，提出一种基于用户兴趣度与主题相关的 PageRank 改进算法即 ITPR，该算法融合了用户反馈（页面停留时间、链接点击量）和网页链接结构（链入结构、链出结构）等信息，对 PR 值进行了适当的修正与调整。利用 Lucene 打分机制，引入主题相关度因子，使与主题相关得分高的网页排名提前，有效地抑制了主题漂移的现象；用户兴趣度因子的添加，使 PR 值在传递时更偏向于用户感兴趣的网页。同时，由于用户兴趣度因子不受时间因素影响，使得因受时间因素影响而导致的算法偏重旧网页的入链因子和出链因子总权值降低，从而使得偏重旧

（下转第 312 页）

[8] Zhang Lei, Gao Shang. An image segmentation method based on elite theory-improved particle swarm optimization[J]. Computer applications and Software, 2009, 26(12): 89-92(in Chinese)
张磊，高尚. 基于精英粒子群优化算法的图像分割方法[J]. 计算机应用与软件，2009，26(12)：89-92

[9] Zhou Chang-ying. Research on image segmentation technology based on improved fuzzy BP neural network[J]. Computer Simulation, 2011, 28(4): 287-290(in Chinese)
周长英. 基于改进的模糊 BP 神经网络图像分割算法[J]. 计算机仿真，2011，28(4)：287-290

[10] Hsieh Sheng-ta, Sun Tsung-ying, Liu Chan-cheng, et al. Efficient population utilization strategy for particle swarm optimizer[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2009, 39(2): 444-456

[11] Zhan Zhi-Hui, Zhang Jun, Li Yun, et al. Adaptive particle swarm optimization[J]. IEEE Transactions on systems, Man, and Cybernetics, 2009, 39(6): 1362-1381

[12] Li Li, Xue Bing, Niu Ben, et al. The novel non-linear strategy of inertia weight in particle swarm optimization [C]// Proceedings of the Congress on Bio-Inspired Computing. 2009: 1-5

(上接第 278 页)

网页的不足得到改善；兴趣度预测因子能及时根据用户兴趣的变化调整网页排序，使用户感兴趣的网页快速上浮。仿真实验结果表明，ITPR 算法使得网页排序质量有一定程度的改善，进一步提升了网页查准率，提高了用户满意度。

参 考 文 献

[1] China Internet Network Information Center(CNNIC). The thirty-fourth statistical report of Chinese Internet development[R]. (2014-07). http://baike.baidu.com/view/14341540.htm(in Chinese)
中国互联网络信息中心(CNNIC). 第 34 次中国互联网络发展状况统计报告[R]. (2014-07). http://www.edu.cn/focus_1658/20140721/t20140721_1152815.shtml

[2] Feng Hai-tao. An improved PageRank algorithm with web time weight[J]. Journal of Xi'an University of Posts and Telecommunications, 2013, 18(2): 121-124(in Chinese)
冯海涛. 基于网页时间权值的 PageRank 算法改进[J]. 西安邮电大学学报，2013，18(2)：121-124

[3] Shi Ming-ming. Research on Weighted PageRank algorithm[J]. Software Guide, 2013, 12(2): 30-32(in Chinese)
史铭茗. 加权 PageRank 算法研究综述[J]. 软件导刊，2013，12(2)：30-32

[4] Brin S. The anatomy of a large hypertextual Web search engine [J]. Computer Networks and ISDN System, 1998, 30(98): 107-117

[5] Shao Jing-jing, Li Bo, Liu Han-ping. An improved pagerank algorithm-adjusting the damping factor[J]. Mathematica Applicata, 2008, 21(S1): 57-61(in Chinese)
邵晶晶，李波，刘汉平. PageRank 的改进算法——调整阻尼因子[J]. 应用数学，2008，21(S1)：57-61

[6] Lovasz L, et al. Random Walks on Graphs: A Survey [J]. Combinatorics, 1993, 8(4): 1-46

[7] Xing W, Ghorbani A. Weighted PageRank algorithm[C]// Proceedings of Second Annual Conference. Piscataway: IEEE Press, 2004: 305-314

[8] Manning C D, Raghavan P, Schutze H, et al. Introduction to information Retrieval [M]. Beijing: Post & Telecom Press, 2010

[9] Tyagi N, Sharma S. Comparative study of various Page Ranking Algorithms in Web Structure Mining (WSM) [J]. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2012, 1(1): 14-19

[10] Taher H. Topic-sensitive PageRank [C]// Proceedings of the 1th International Conference on World Wide WEB. Honolulu. Hawaii: ACM Press, 2002: 784-796

[11] Li Wei-dong, Lu Ling. Research and application of pageRank algorithm combined with VSM technique[J]. Computer and Modernization, 2011(7): 96-98(in Chinese)
李卫东，陆玲. 融合 VSM 技术的 PageRank 算法研究与应用[J]. 计算机与现代化，2011(7)：96-98

[12] Wang Zhong-fei, Gong Biao. Improved pageRank algorithm based on anchor texts similarity[J]. Computer Engineering, 2010, 36(24): 258-260(in Chinese)
王钟斐，工彪. 基于锚文本相似度的 PageRank 改进算法[J]. 计算机工程，2010，36(24)：258-260

[13] Chakrabarti S, Dom B, Gibson D, et al. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text [C]// Proceedings of the 7 ACM-WWW International Conference. Brisbane: ACM Press. 1998: 65-74

[14] Duan H C, Hu P. Improved pagerank algorithm based on topic character and time factor[J]. Computer Engineering and Design, 2010, 31(4): 866-868(in Chinese)
段淮川，胡平. 基于主题特征和时间因子的改进 PageRank 算法[J]. 计算机工程与设计，2010，31(4)：866-868

[15] Kumar G, Duhan N, Sharma A K. Page Ranking Based on Number of Visits of Links of Web Page[C]// International Conference on Computer & Communication Technology (ICCCT). 2011: 11-14

[16] Peng Cong, Wu Qiang, Li Ren-fa. An Improved Algorithm of Web Page Ranking[J]. Microcomputer Information, 2010, 26(33): 72-74(in Chinese)
彭聪，吴强，李仁发. 一种改进型的网页排序算法[J]. 微计算机信息，2010，26(33)：72-74

[17] Fang S F. Based on User Feedback PageRank algorithm [J]. Computer Technology and Automation, 2012, 31(1): 89-92(in Chinese)
方树峰. 基于用户反馈的 PageRank 改进算法[J]. 计算技术与自动化，2012，31(1)：89-92

[18] Wang D G, Zhou Z G, Liang X. Analysis of pagerank algorithm and its improvement [J]. Computer Engineering, 2010, 36(22): 291-293(in Chinese)
王德广，周志刚，梁旭. PageRank 算法的分析及其改进[J]. 计算机工程，2010，36(22)：291-293

[19] Mccandless M, Hatcher E, Gospodnetic O, et al. Lucene in action [M]. Beijing: Post & Telecom Press, 2011

[20] Qiu Z, Fu T T. Lucene 2.0+ Heritrix[M]. Beijing: Post & Telecom Press, 2007(in Chinese)
邱哲，符滔滔. 开发自己的搜索引擎：Lucene 2.0+ Heritrix [M]. 北京：人民邮电出版社，2007