一种基于 PageRank 算法原理的会员人气度排序算法①

User's Renqi Ranking Algorithm based – on PageRank's Principle

李　强　（浙江大学 计算机学院 浙江杭州 310027）
（杭州电子科技大学 计算机学院 浙江杭州 310037）
王申康　（浙江大学 计算机学院 浙江杭州 310027）

摘　要： 社会化搜索提倡依靠众人的智慧改善搜索结果，其主要途径是利用搜索引擎的注册会员所收藏的网页，以及寻找具有相同兴趣主题的会员来帮助用户更加准确地找到所需信息。本文提出了一种依据 PageRank 算法原理来计算会员与兴趣主题的相关度的人气度排序算法，用于对会员的重要性进行排序，向用户推荐高质量的会员。通过对小数量会员进行会员搜索实验，取得了较好的效果。

关键词： 元搜索引擎 社会化搜索 会员搜索 人气度排序

1　引言

Google 引领的第二代搜索引擎采用关键字搜索技术，该技术出现后迅速风靡全球，并成为延续至今的主流搜索技术。但关键字搜索仅仅是进行关键词匹配搜索，它并不是基于对内容的理解，所以搜索结果往往会返回上万条或更多的查询结果，其中存在许多和查询关键词只是字形、词形的匹配，内容却毫不相关，其查准率还不能满足人们的需要[1]。

社会化搜索是随着 Web2.0 概念发展起来的搜索技术，是一种新的搜索模式。它将 Web2.0 这种人与人之间的交互共享的理念引入到搜索引擎当中，其目的是通过搜索引擎的众多用户的集体智慧获取和改善搜索结果[2]。它主要是通过搜索其他用户所保存的相同兴趣主题的收藏（收藏搜索），以及寻找具有相同兴趣主题的会员（会员搜索），形成自己的好友圈，来帮助用户更加准确地找到所需信息。

在国内首倡社会化搜索的 Bbmao 搜索引擎[3]就具有会员搜索的功能。它通过人气得分技术向用户推荐高质量的相关会员，帮助用户建立一个自己信任的好友网。本文提出了一种借助 PageRank 算法的原理来计算会员的收藏与查询主题相关度的会员搜索算法，该算法对会员的人气度进行计算并排序，以此向用户提供高质量的会员。

2　PageRank 算法简介

PageRank 算法[4]是在 1998 年由斯坦福大学的 Sergey Brin 和 Lawren Page 提出，它借鉴了传统情报检索理论中的引文分析方法：当网页 1 有一个链接指向网页 2，就认为网页 2 获得了一定的分数，该分值的多少取决于网页 1 的重要程度，即网页 1 的重要性越大，网页 2 获得的分数就越高。由于互联网的链接相互指向的复杂程度，该分值的计算过程是一个迭代过程，最终网页依据所得的分值进行排序并将检索结果送交用户。这个量化的分值就是 PageRank 值，其计算公式如下：

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \quad (1)$$

其中 PR(A) 是网页的页面级别，d 是界于 (0,1) 区

① 基金项目：浙江省自然科学基金项目（Y105567）资助。

间的衰减系数，一般取值 0.85，T_1、T_2、…T_n 为指向网页 A 的其它页面，$C(T_n)$ 是网页 T_n 向外指出的链接数目。

上述这种传统的 PageRank 算法还存在不少缺陷[5]，其中一点是页面级别只与页面的链接结构相关，而没有考虑链接页面的文本内容。Chakrabarti S. 等在文献[6]中提出结合链接分析和文本内容的 PageRank 算法，它计算文本内容与查询的相关性，再结合链接结构进行计算，其计算公式如下：

$$PR_q(A) = (1-d)PR'_q(A) + d\sum_{T_i \in T} \frac{PR(T_i) \times P_q(T_i)}{C(T_i)} \quad (2)$$

其中 $P_q(T_i) = \frac{R_q(T_i)}{\sum_{k \in F_i} R_q(k)}$，$PR'_q(A) = \frac{R_q(A)}{\sum_{k \in W} R_q(k)}$，$R_q(k)$ 是文档 k 与查询 q 的相关度，W 是整个网络的网页集合，F_i 是网页 T_i 的链接出网页集合。

3　会员搜索

3.1　用户模型

为获得用户的兴趣偏好，必须为用户建立用户兴趣模型。本文采用基于本体论的加权特征词形式表示用户的兴趣模型[7]。为了区分用户具有的不同兴趣类别，通常需要一个较为完整的兴趣分类参考模型。用户兴趣分类参考模型是基于本体论对世界知识体系的层次型主题分类结构的认知，从已有的开放分类目录和知识库获得主题的分类结构和相应的特征词。用户兴趣树是用户兴趣参考模型的一个子树，通过用户注册时手工建模，并在用户使用搜索引擎的过程中通过对查询关键词的学习和对查询结果的收藏反馈信息的强化学习进行更新。

定义 1：用户兴趣类节点 Node(c) 是一个二元组 (c,v)，其中 $c \in C$，v 为兴趣类别 c 的权值，C 是兴趣类别全集。

为实现系统的社会化搜索，还需要在用户模型中建立用户收藏库和用户好友库。系统依据用户兴趣树自动为用户建立用户收藏夹目录树和用户好友目录树。用户的收藏夹目录树、用户好友目录树和用户的兴趣树的类别结构是一一对应的，它们的每个目录对应到用户兴趣树上的每个兴趣类别，目录名为该兴趣类别名。以后随着用户兴趣模型中兴趣类别的更新，用户收藏夹目录结构和用户好友目录结构同步进行更新。系统的用户界面模块提供接口，使用户可以浏览、增删改自己的收藏树和好友信息树。

定义 2：好友信息节点 Fleaf(c,UserID) 它是个一元组(UserID)，用来保存用户在兴趣类别 c 下的好友信息，包含好友的用户 ID。

定义 3：用户收藏节点 CLeaf(c,Tag) 它是一个四元组(Tag，Title，Meta，Url)，用来保存用户在某个兴趣类别下的一个收藏品的信息，包含收藏文档的标签 Tag 、标题 Title、摘要 Meta 和 URL 地址 Url，其中标签又称为自由分类，用于对用户收藏文档的内容的概括，并用于对不同的收藏节点的分类。

定义 4 用户收藏兴趣类别节点 CNode(c) 它是一个二元组(c,v)，用来保存用户的收藏品的兴趣类别信息，包含兴趣类别名 c 和相应的权重 v，v 表示用户对该收藏类别的兴趣程度，用该兴趣类别下的收藏节点个数表示。

3.2　会员搜索的原理

本文认为某个兴趣主题的兴趣圈中的用户与该兴趣主题的相关度(在本文中称之为人气度)与以下三方面的因素有关：

(1) 用户兴趣树中该兴趣主题的权重。用户兴趣树的兴趣主题权重表示用户对该兴趣主题查询的次数。权值越大，表明用户对该主题查询的次数越多，对该主题的兴趣越大。

(2) 用户收藏树中该兴趣主题的权重。用户收藏树的兴趣主题权重表示用户收藏该兴趣主题的收藏品的数量。权值越大，表明用户对该兴趣主题的收藏越多，这是用户人气度的主要因数。

(3) 用户被加为好友的信息。一个用户如果被多个其他用户加为好友，表示该用户的收藏被众人认可，那么它就应该有更高的人气度。这好比一个网页，如果它被多个其它网页链接，那么该网页就越重要。

如果把一个用户被另一个用户加为好友的关系看成是一个网页被另一个网页链接的关系，那么就可以借助通过分析网页的链接结构来获得网页的等级值的原理，通过分析用户之间好友的结构关系来获得用户的人气度。本文借助 PageRank 算法的原理应用到用户人气度的计算，根据用户之间好友的结构关系，并结合其它两方面的因素，提出了一种计算用户人气度的算法 ScoreRank。

3.3　人气度排序算法 ScoreRank

在 ScoreRank 算法中，好友的结构关系看作是网

页的链接关系，用户对兴趣主题的权重看作是网页文档对查询的相关度。因为用户对兴趣主题的权重经常要发生变化，而 PageRank 值要通过迭代来计算，计算量较大，为提高检索的速度，ScoreRank 值应通过离线的方式计算，因而先按类似于传统的 PageRank 算法计算 ScoreRank，再以它为系数乘上用户对兴趣主题的权重来得到最后的 ScoreRank 值。

设用户 A 在兴趣类别 c 下被用户 $U_1, U_2, \cdots, U_n$ 加为好友，则用户 A 对兴趣类别 c 的人气度 $SR_c(A)$ 为：

$$SR_c(A) = (1-d) + d\left(\frac{SR_c(U_1)}{C_c(U_1)} + \frac{SR_c(U_2)}{C_c(U_2)} + \cdots + \frac{SR_c(U_n)}{C_c(U_n)}\right) \quad (3)$$

其中 $C_c(U_i)$ 表示用户 Ui 在兴趣类别 c 下的好友个数，d 是界于(0,1)区间的衰减系数，一般取值 0.85。通过迭代计算，得到兴趣主题 c 的兴趣圈中所有用户的人气度 SR_c 值。

用户对兴趣主题 c 的兴趣权重与用户在用户兴趣树中对兴趣主题 c 的兴趣权重和用户收藏树中对兴趣主题 c 的兴趣权重相关，而后者占主要因数。当用户对兴趣主题 c 的查询次数达到一定的次数后，表明用户对该兴趣主题比较稳定，因此可以认为用户对查询主题的兴趣度与该主题相关的查询次数存在对数函数的相关关系，由此得出用户 A 对兴趣主题 c 的兴趣权重 $W_c(A)$ 为：

$$W_c(A) = \lg(Node(c).v) \times CNode(c).w \quad (4)$$

最终用户 A 对兴趣类别 c 的人气度为：

$$RqScore_c(A) = SR_c(A) \times W_c(A) \quad (5)$$

根据公式 5 计算出兴趣类别 c 的兴趣圈内所有用户的人气度值，再按他们的人气度从高到低排序显示，向用户推荐高质量的会员。

4　实验结果与分析

为了对本文提出的会员排序算法进行实验，我们在 Windows 平台上用 ASP. NET 建立了一个结合个性化搜索与社会化搜索的元搜索引擎，并对实验结果进行了分析。

4.1　实验准备

(1) 建立用户兴趣模型。用户兴趣分类参考模型采用网易的中文 ODP 开放目录的前两级目录，共 14 个一级类别和 152 个二级类别，并通过中文维基百科进行二级类别下的特征词扩充。邀请 100 个大学生作为注册会员，每个会员注册一个用户帐号，选定计算机网络、程序设计、电脑游戏、音乐和体育报道这 5 个二级类别作为会员的共同兴趣类别，个人还可以选择其它的兴趣类别，通过这种手工的方式建立用户个体的初始兴趣模型。在用户的 30 天的使用过程中，系统对用户的查询关键词和对查询结果的反馈信息自动进行机器学习，更新各个用户的兴趣树。

(2) 建立用户收藏库和用户好友信息库。在用户注册会员手工建立用户个体兴趣模型的同时，系统自动为每个用户建立相应的用户收藏树和用户好友信息数的兴趣类别目录结构。用户在使用系统进行搜索的过程中，对搜索结果中自己感兴趣的搜索结果设定标签后保存到自己的收藏树中的兴趣类别下，并且不定期地输入查询词对会员进行搜索，将自己感兴趣的会员添加为相应兴趣类别下的好友，建立用户的好友信息树。经过 30 天的使用，用户收藏库中保存了 3723 条收藏信息，用户好友信息库中保存了 375 条加为好友的记录。

表 1　对不同兴趣类别的查询词进行会员排序的运行时间比较

查询词	查询意图兴趣类别	运行时间（秒）
元搜索引擎	计算机网络	2.1
对象	程序设计	2.3
征途	电脑游戏	1.9
花腔	音乐	0.8
大师杯	体育报道	1.3
流感	医学	0.4
J-10	军事	0.5
股票	经济金融	1.0
围城	文学	0.5
考研	校园学习	1.3
共同兴趣类别内的平均运行时间		1.68
一般兴趣类别内的平均运行时间		0.74
平均运行时间		1.21

4.2　实验方法和结果分析

为了验证上述会员搜索算法的运算效率和有效

性，用一台 CPU 为 PIII800、内存为 512MB、操作系统为 Windows XP 的计算机作为测试平台。以一个用户身份登录进入系统，分别输入 5 个属于共同兴趣类别和普通兴趣类别的查询词进行会员搜索，检测系统进行会员搜索的运行时间，来对会员搜索算法的运行效率进行分析；同时对会员人气度排序结果进行人工检测分析，来判断会员搜索算法的有效性。表 1 列出了对 10 个查询词进行会员搜索时系统运行的时间，单位为秒。

从表 1 我们可以看出系统进行会员搜索的平均时间只有 1 秒多，该排序算法在运行效率上是可行的，当然这样少的运行时间与会员数量少有很大的关系。还可以看出对属于共同兴趣类别内的查询词进行会员搜索系统所耗费时间要比一般兴趣类别的要多，这主要是因为共同兴趣类别的兴趣圈会员数量相对较多，会员之间加为好友的关系更复杂，排序算法在进行迭代计算时要耗费更多的时间。

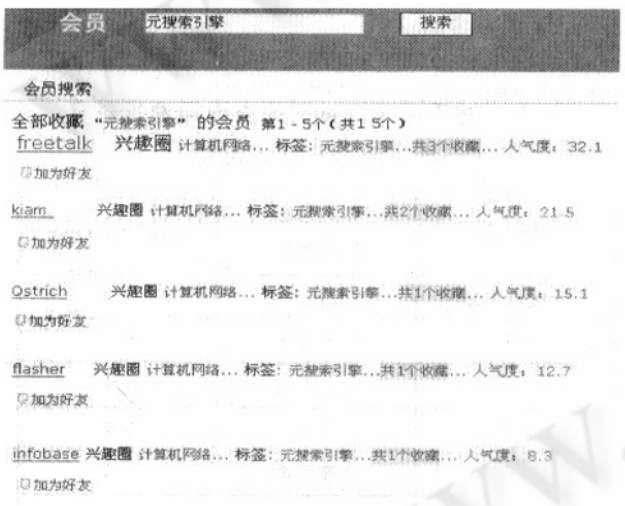

图 1　对"元搜索引擎"查询词进行会员搜索得到的会员人气度排序结果

图 1 是对"元搜索引擎"感兴趣的会员的搜索结果的运行截图。从图 1 中看到系统将"freetalk"会员推荐排在首位，它有 3 个以"元搜索引擎"为标签的收藏。通过好友库人工查看该会员被 4 个用户加为好友，这 4 个用户对"计算机网络"类别的兴趣度都较高。排在第 2 为的 kiam 会员被 2 个用户加为好友。根据 ScoreRank 算法来计算它们的重要性，得到它们的人气度分别为 32.1、21.5，能较合理地根据用户被加为好友的情况来判断用户对兴趣类别收藏的重要性。通过点击这些会员名可以链接到它在"计算机网络"兴趣类别下的所有收藏。

5　结束语

本文在元搜索引擎上引入社会化搜索，提出了一种借鉴结合链接分析和文本内容的 PageRank 算法的 ScoreRank 人气度算法。。由于受系统本身条件的限制，在进行系统实验时用户数量和用户使用时间都相对较少，用户的收藏库内容不多，用户之间的好友关系也比较简单。从以上初步的实验分析可以看出该会员排序算法在运行时间和有效性上都有较好的结果，但还需经过大量的用户长期使用来检验其效果。

参考文献

1　彭洪汇、林作铨，Internet 上的搜索引擎和元搜索引擎[J]，计算机科学，2002，29(9)：1 - 12.

2　Yutaka M.，Junichiro M.，Masahiro H. An Advanced Social Search Engine System from the Web[C]. Proceedings of the 15th International Conference on World Wide Web WWW'06，May，2006，Edinburgh，Scotland：397 - 406.

3　BBMao 搜索引擎[EB/OL]：http://www.bbmao.com.cn

4　Page L.，Brin S. The PageRank Citation Ranking：Bringing Order to the Web[EB/OL]. http://www.db.stanford.edu/~backub/PageRanksub.ps，1998 ~ 2001.

5　Haveliwala，T. H. Topic - sensitive PageRank[C]. Proceedings of the 11th International World Web Conference，Hoho Lulu Hawaii，2002.

6　Chakrabarti S.，Dom B.，Gibson D.，et al. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text[C]. Proceeding of the 11th ACM - WWW International Conference. Brisbane：ACM Press，2002：65 - 74.

7　黄国景、崔志明，基于 Ontology 的个性化元搜索引擎[J]，微电子学与计算，2004，21(11)：100 - 103.