第11卷 第28期 2011年10月
1671—1815(2011)28-6879-04

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 11 No. 28 Oct. 2011

基于社交网络的实时搜索引擎的排序算法研究

徐 婕 康慕宁 董谷音
(西北工业大学计算机学院,西安 710072)

摘 要 针对用户在社交网络中面对海量的信息和资源,如何实时地获取自己感兴趣的内容,给出一种基于社交网络的实时搜索模型,并根据社交网络的特点考虑对朋友、时间、相关度等因子对搜索结果进行排序。针对基于超链接网页排名的Pagerank算法,提出了一种基于用户朋友数的Pagerank排序算法。实测结果表明,该模型提高了搜索结果的实时性和相关度。

关键词 实时搜索 社交网络 API PageRank 概念
中图法分类号 TP391.3; **文献标志码** A

随着Web 2.0,特别是以MySpace、Facebook、Twitter为代表的社交网站的飞速发展,人们越来越喜欢在社交网络上发布和获取信息,社交网络将成为人们生活中的一部分,成为人们现实生活的延伸。社交网络已从单一的娱乐交友沟通平台演化至新闻传播平台[1]。人们开始从传统的新闻网站向这些社交网络转移,因为传统搜索引擎已经不能满足人们对实时信息的要求。根据OneRiot发布的一份报告,40%的搜索属于"浏览查询"。此类搜索并不是为了得到一个内心里特定的结果,更多的是希望发现关于主题的最新新闻。比如搜索某个名人时,极可能他们不是想找到关于这个名人的生平简介等入口点,而是想在网页上查看有关这个人的最新消息。人们对社交网络的信息越来越感兴趣,同时每分每秒都产生大量信息,这些信息过于庞大且具有杂乱、生命周期短等特点[2]。如何从这些庞大、杂乱的信息中获取最新的、高质量的信息是人们研究的热点之一。

社交网络,是指人和人之间通过朋友、血缘、交易、兴趣、链接等关系建立起来的社交网络结构,它强调人与人之间的关系及纽带[3]。以现实社会关系为基础,模拟和重建现实社会的人际关系网络,来提高社会交往的质量和效果。是一种关系化的,以用户为中心的,虚拟社交和真实社交的融合。其核心是"用户创造价值",以"用户为中心"来组织和传播内容。用户的意识和行为成为关注的焦点,满足用户多样化、个性化的追求是关键[4]。

目前随着Web2.0的发展,大多数社交网站实现了信息的实时化,但它仅对信息的时效性做了考虑,而没有对其他因素进行考虑。只有Twitter最近对实时功能做了新的改进,搜索结果除了时效性还考虑流行程度,将热门内容放在搜索的顶部。但这远远满足不了用户的需求,比如现在大多数用户想看到的与主题更相关的结果,想看到相关主题权威用户的说法,想看到相关主题内该用户有关朋友的消息。

因此本文在社交网络的基础上,构造了一种实时搜索模型,并针对现有社交网络实时搜索引擎的不足和社交网络的基本特性,主要对实时搜索结果进行了研究,提出了一种基于权威用户、朋友、时间、相关性等因子的搜索结果排序算法,用来更好的满足用户的搜索体验。

1 系统模型

为了提高搜索引擎的查全率、查准率,满足用户对搜索结果的需求,本文设计了一种基于社交网络的实时搜索引擎,它能够在某一特定的社交网络

2011年6月30日收到
第一作者简介:徐 婕(1985—),女,硕士研究生。研究方向:数据管理技术,信息检索。E-mail:xujie0230@126.com。

中进行实时搜索并进行相关性排序,回馈给用户更合理的数据结果。

1.1 模型结构

该模型是一个基于社交网络的实时搜索系统,目前大部分社交网站的将搜索实时化,动态的更新用户的最新消息,所以该模型以某一个社交网站为基点,并利用该社交网站提供的开放 API 实时的获取最新数据,当用户开始搜索时,采用周期式的制导性数据抓取来满足信息的实时化,并对搜索结果做相关性分析,并实时的更新知识库中的数据,回馈给用户最新的、相关的信息。模型结构网络如图 1 所示。

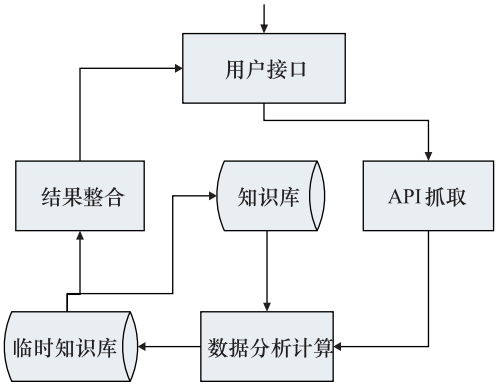

图 1 实时搜索模型结构图

1.2 数据结构

大多数成熟的社交网络为第三方开发提供了开放的 API,例如 Facebook 的 Status. get、Twitter 的 Streaming API、Search API 等,所以本模型利用第三方开放 API 从社交网站上抓取数据。社交网站一般以 JSON 格式或者以 XML 格式将信息结果返回,所以抓取下来的数据以 XML 格式进行存储。XML 数据格式如下:

```
<item>
<link> </link>
<content> </content>
<createtime> </createtime>
<username> </username>
…
</item>
```

1.3 模型实现流程算法

当用户开始搜索时,根据用户输入的关键字触发后台服务程序对数据进行抓取,然后对抓取到的数据进行数据分析,同时将抓取到的数据存入临时知识库,并对临时知识库的内容进行排序,存入并更新知识库,将知识库中的内容反馈给用户。为了能实时的给用提供最新的数据,本系统采用周期式的制导性抓取,将每一批数据实时的显示给用户,同时也动态的更新旧的知识内容,共用户浏览和参考。

数据分析模块是对下载下来的每一批数据进行分析、过滤和相关性计算等一系列的操作,是实现算法的关键部分,也是本系统的核心。流程图见图 2。

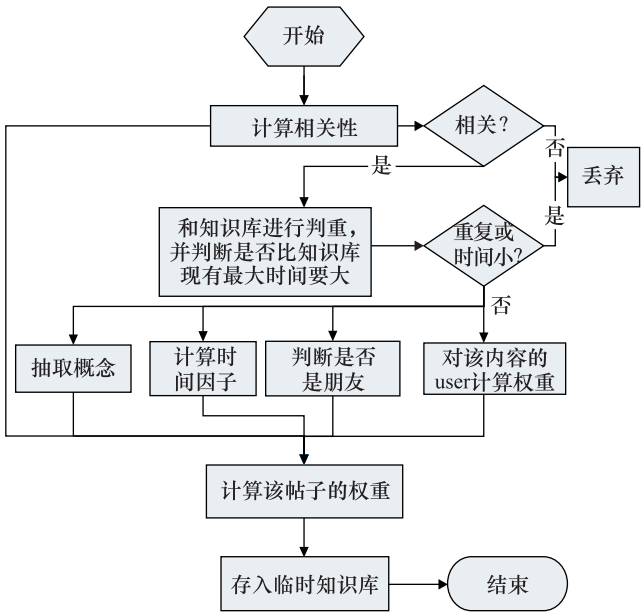

图 2 数据分析流程图

2 实时搜索排序算法

社交网络的主要特征是以用户为中心,用各种关系组成了一个小的社交关系网。同时,社交网络产生的信息对实时性的要求很高。所以对用户搜索的信息排序主要考虑的因子有以下几点。

2.1 时间因素

因为用户想看到的是最新的动态消息,所以对时间的要求很高。

$$T_v = \frac{a}{now - createtime} \quad (1)$$

在式(1)中,*now* 表示用户开始搜索的时间,*createtime* 为内容的创建时间,*a* 表示相关性系数。

2.2　内容相关性

对搜索内容进行关键词拆分,根据 TF - IDF 算法[5]来评估某一关键词在某一搜索到得 XML 文本中的重要程度。其相关值为:

$$R_v = \sum_{k=1}^{q} TF_{vk} IDF_k \quad (2)$$

在式(2)中,q 为搜索内容关键词的个数,TF_{vk} 为关键词 k 在某一 XML 文档 v 中出现的频率;IDF_k 为关键词 k 的逆文本词频。

$$IDF_k = \lg\left[\frac{N}{DF(w_k)}\right] \quad (3)$$

在式(3)中,N 为本次搜索到得所有 XML 文本数,$DF(w_k)$ 为这些文本中至少包含一次关键词 k 的文本数。

2.3　概念相关性

对搜索到的每一个信息内容确定若干个概念 $U(C_{u1}, C_{u2}, \cdots, C_{um})$,这里考虑用户正在搜索的关键词为一个概念。如果搜索到的信息中的概念集合中包含用户正在搜索的关键词的话,则认为具有更好的相关性,用 C_v 表示。

2.4　发表微内容用户的朋友数

PageRank 算法的基本思想是借鉴传统的学术文献的引文分析方法,并把这一思想应用到了 Web 页面中,即一篇文献的重要性可以通过其他文献对其引用的数量来衡量[6]。如果页面 A 通过超级链接指向了页面 B,相当于页面 A 给页面 B 投了一票,页面 A 需要把自己的一部分 PageRank 值分给页面 B,重要的页面会在搜索引擎的搜索结果中位于前列,如果一个网页有许多网页都指向它,那么它可能获得很高的 PageRank 值;如果一个网页被一个本身 PageRank 值很高的页面所指向,那么它同样可能具有很高的 PageRank 值[7]。同样,在社交网络中,如果一个用户有一个朋友意思是该朋友给他投了一票,如果该用户有很多的朋友,说明他在该社交网站中具有很高的威望,同样,一个人如果有一个朋友的威望很高,说明他的威望也很高。所以根据这个特点利用 Pagerank 算法进行排序并对该用户的微内容计算一定的权值。

$$PR(v) = (1 - d) + d\sum_{u \subset f_V} PR(u)/N_u \quad (4)$$

在式(4)中,$PR(v)$ 表示发表该帖子的用户 v 的权重,d 表示衰减因子,其取值介于 0 ~ 1 之间,通常设定为 0.85,fv 指用户 v 的朋友,$PR(u)$ 表示朋友 u 的权重,N_u 表示朋友 u 的朋友个数。

2.5　朋友因子

根据社交网站的特点,用户更多的想关注和他搜索主题相关的朋友的动态消息。

综上所述,所以发表该微内容的权重值为:

$$W_V = T_v + C_v + R_v + PR(V) + bIsfriend + others \quad (5)$$

在式(5)中,*Isfriend* 为是否是朋友,如果是则为 1,否则为 0。b 为相关因子,如果增大朋友因子对排序的影响,可增大 b 的系数的设置。*others* 为其他相关因素。

以上算法既满足了用户对实时信息的要求又对搜索的信息进行相关性排序且符合社交网络的特征。

3　实验与结果

本实验以 Twitter 社区网络作为实验平台,利用 Twitter 提供的第三方接口如:Search API 来获取和主题相关的信息。GET followers 接口来获取和用户相关的朋友信息。建立一个专门的服务程序来抓取数据并将数据存放在文件队列里。建立一个专门的服务程序来进行相关性计算并对内容进行过滤,然后分别调用相应的线程来计算帖子的权威度、时间因子、朋友权值,最后计算该帖子的权值并将内容存入临时知识库中,建立一个专门的线程来对临时知识库中的帖子进行排序。这样能保证用户能看到实时更新的信息。由于 Twitter 提供了 Streaming API,不停的将最新的数据回馈给用户,和 Search API 相比实时性比较好,搜索范围比较全,但是不能得到几天前的数据,主题相关性没有 Search

API 好。所以根据这两个接口的优缺点,将二者合理利用,当第一次搜索时利用 Search API 来搜索最近的 100 条数据,以后通过调用 Streaming API 来获取相关的实时的数据。以 20 个用户为中心输入关键字,并和 Google、Bing、Twitter 实时搜索引擎的搜索结果进行相关性比较,比较结果如图 3 所示。

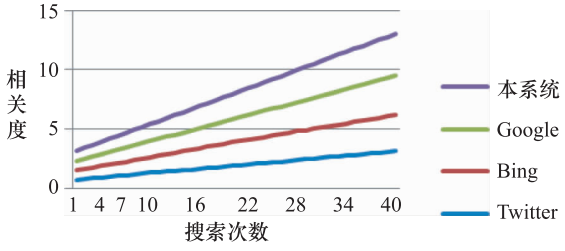

图 3 关键词与搜索结果相关度曲线图

经过多次实验表明,该系统的查询准确率比其他实时搜索引擎高很多,并且符合社交网络的特征,在实时更新方面也要高很多,查询响应时间在用户接受的范围内,性能良好,验证了本文算法的可行性,也验证了本文算法的合理性。

4 结论

实测表明,本文提出的基于社交网络的实时搜索引擎模型不仅能够提供实时的搜索结果,并且能够提高搜索的效率。尤其是在搜索结果排序算法中,对搜索内容进行相关性计算和概念抽取提高了信息的相关度和准确性;考虑时间因素提高了信息的实时性;考虑朋友和权威用户能更好的符合社交网络的特征,更好的为用户提供基于社交网络的实时搜索服务。但是在考虑权威用户的时候没有考虑权威用户的朋友与搜索主题相关性,容易差生主题漂移的现象,这是今后进行研究和改进的地方。

参 考 文 献

1 Young A L, Quan-haase A. Information revelation and Internet privacy concerns on social network sites: a case study of facebook. Porc of the 4th International Conference on Communities and Technologies. New York: ACM Press,2009; 265—274

2 Li Yungming, Hsiao Hanwen. Recommender service for social network based application. Proc of the 11th International Conference on Electronic Commerce. New York: ACM Press,2009; 378—381

3 李勇军,代亚非. 社交网络. 中国计算机学会通讯,2010;6(3):47—51

4 林 容. 社交网络的特性及其发展趋势. 新闻界,2010;(5):32—34

5 徐宝文,张卫丰. 搜索引擎与信息获取技术. 北京:清华大学出版社,2003

6 原福永,张园园. 基于链接分析的相关排序方法的研究与改进. 计算机工程与设计,2007;28(7):1603—1662

7 杨 彬,康慕宁. 基于概念的权重 PageRank 改进算法. 情报杂志,2006;(11):70—72

Sequencing Algorithm Based on the Social Network Real-time Search Engine

XU Jie, KANG Mu-ning, DONG Gu-yin

(Department of Computer Science, Northwestern Polytechnical University, Xi'an 710129, P. R. China)

[**Abstract**] For users of social networks in the face of a flood of information and resources to real-time access to content of interest, a social network is give based real-time search model, and in accordance with social networking features to consider friends, time, relevance and other factors to sort the search results. For the hyperlink-based page ranking Pagerank algorithm, a number of user-based friends Pagerank sorting algorithm is presented. Experimental results show that the model improves the real-time and relevance of the search results.

[**Key words**] real-time search social network API PageRank concept