Hans Journal of Data Mining 数据挖掘, 2013, 3, 27-31
http://dx.doi.org/10.12677/hjdm.2013.32006 Published Online April 2013 (http://www.hanspub.org/journal/hjdm.html)

Study on Real-Time Search Engine Based on Social Network

Zhengwei Huang, Huijuan Huang, Yanni Yang

The Department of Economy and Management, China Three Gorges University, Yichang
Email: 339372775@qq.com

Received: Dec. 4th, 2012; revised: Jan. 17th, 2013; accepted: Feb. 2nd, 2013

Abstract: With the rise of social network, Internet users have put forward higher request on the speed and real-time for information. Through the time-consuming web crawler, traditional search engine grab the web page, and establish the index, which can not meet the users' needs. In contrast, the real-time search can capture the data flow directly from the network, thus it meets the needs of the users for real-time information more quickly. Some real-time search engines have already appeared at home and abroad, such as the United States OneRiot, Google, Collecta and Scoopler, etc., which provide people with the latest search content through grabbing the mass information released by the users on Twitter, Facebook, Digg, etc. at any time.

Keywords: Real-Time Search; Traditional Search; Social Network

基于社交网络的实时搜索引擎探究

黄正伟，黄会娟，杨艳妮

三峡大学经济与管理学院，宜昌
Email: 339372775@qq.com

收稿日期：2012 年 12 月 4 日；修回日期：2013 年 1 月 17 日；录用日期：2013 年 2 月 2 日

摘　要：随着社交网络的兴起，网民对信息资讯的速度和实时性提出了更高的要求。传统搜索引擎通过费时的网络爬虫抓取网页，建立索引，这已远不能满足用户的需求。相比之下，实时搜索能直接从网络捕获数据流，从而更迅速地满足用户对实时信息的需求。目前国内外已经出现了一些实时搜索引擎，如美国的 OneRiot、Google、Collecta 和 Scoopler 等，它们通过抓取 Twitter、Facebook、Digg 等社交网络上用户随时发布的海量信息，给人们提供最新的搜索内容。

关键词：实时搜索；传统搜索；社交网络

1. 基于社交网络的实时搜索引擎的产生

随着 Web2.0，特别是以 Facebook、Twitter 为代表的社交网站的飞速发展，人们越来越热衷于在社交网络上发布和获取信息，社交网络将成为人们生活的一部分，成为人们现实生活的延伸。同时，以 Twitter 为代表的实时网络成为了互联网最热门的应用领域之一，社交网络不再是过去单一的娱乐交友沟通平台，而更作为一种新闻传播平台存在和发展。人们获取新闻的渠道也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。社交网络信息的爆炸性增长，以及其更新快，生命周期短的特性，使得人们越来越关注于如何从这些庞大、杂乱无章的信息中获取高质量、实时性的信息，这也成为 IT 行业的技术研究重点。正是基于网民对新兴资讯速度和实时性的更高要求，搜索引擎领域应需而动，基于社交网络的“实时搜索”

概念应运而生。

所谓“实时搜索”是指用户能够查阅实时发布的信息资料。也就是说，信息发布和信息查询之间几乎没有时间延迟。例如，你拍张照片并张贴，几秒钟后，可能全世界都能看到这张照片[1]。实时搜索的出现，引起了社会的广泛关注，并对搜索引擎领域的发展产生了巨大的影响，被列为 2010 年全球十大新兴科技之一。

2. 实时搜索引擎的实现原理

2.1. 传统搜索引擎的原理

传统搜索引擎通过网络爬虫，从互联网上自动搜集网页，自动访问互联网，并沿着网页中所有的 URL 爬到其他网页，跟踪网络链接，通过不断地重复该动作来搜集所有爬过的网页，并将之存储到数据库中。

抓取网页是整个搜索引擎工作的开始。最简单的抓取网页的路径是按照超链接的拓扑顺序进行的。首先爬虫会拥有一个初始的 URL 列表，访问到对应的网页中，分析该网页中的＜a href=…＞结构，获取新的 URL，并将之插入到原有队列中[2]。同时根据需要也可以获取＜img src=…＞等结构，抓取相应的其它类型的文件。重复地进行这个过程，直到抓取到指定数量的网页为止。

将抓取到的页面文件进行分析、分解，按照一定的算法，通常根据网页中关键词的匹配程度，出现的位置/频次，链接质量等——计算出各网页的相关度及排名等级，然后根据关联度高低，建立索引，按顺序将这些网页链接返回给用户。

对于传统搜索引擎而言，其自动搜集网页功能分为两种。一种是定期搜索，即每隔一段时间(比如 Google 一般是 28 天)，搜索引擎主动派出“蜘蛛”程序，对一定 IP 地址范围内的互联网站进行检索，一旦发现新的网站，它会自动提取网站的信息和网址加入自己的数据库。另一种是提交网站搜索，即网站拥有者主动向搜索引擎提交网址，它在一定时间内(2 天到数月不等)定向向你的网站派出“蜘蛛”程序，扫描你的网站并将有关信息存入数据库，以备用户查询[3]。但主动提交网址并不保证你的网站能进入搜索引擎数据库，因此目前最好的办法是多获得一些外部链接，让搜索引擎有更多机会找到你并自动将你的网站收录。

由此，传统搜索引擎之下，用户面临的问题是：首先，返回的信息是在几小时或几天其至几个月以前被预先存放在搜索引擎的数据库中的，实际页面中的内容可能已经改变。第二，搜索引擎服务器中存放的主要是静态页面的内容，而无法记录动态发布的信息。但动态信息却是互联网上的重要部分，尤其是在互联网上发布的数据库的内容越来越多，也越来越重要。因此，返回的 URL 虽然多，但是符合用户需求的甚少，用户经常只能以猜测的方式选择其中的一些进行访问，或者侥幸找到需要的信息，或者不得不放弃搜索。

2.2. 实时搜索引擎的原理

以社交网络网站、博客、新闻媒体、出版商内容管理系统等作为外部资源库，实时搜索引擎直接从这些实时网络中捕获数据流，接收数据。搜索服务订阅社交网站的新内容通知，使用 HTTP、FTP 或该网站的其他协议并通过网站 API 检索。由于社交网站一般以 JSON 格式或者以 XML 格式将信息结果返回，所以实时搜索引擎获得的数据将以 XML 格式进行存储[4]。这些资源直接将内容“推入”实时搜索的过滤引擎中，通过主题词建立索引并根据提交的时间、是否包含垃圾信息和与查询之间的相关性进行快速过滤和组织数据。内容最后通过一个称之为流水口(firehose)的机制到达用户界面(UI)来显示搜索结果。实时搜索对内容进行排序和索引的算法是根据内容的提交时间、兴趣的直接程度、查询相关性、跟随者数量代表的作者信誉、读者转发次数代表的链接信誉等。立足于社交网络，实时搜索对搜索结果排序的算法能够给用户带来更新鲜的搜索体验，满足用户的个性化搜索需求。

这种搜索结构每小时能够索引数百万页面。实时搜索引擎检索和索引数据如此之快是因为它不像传统搜索引擎一样需要靠网络爬虫抓取页面来收集和索引信息，它是直接从 Twitter、Facebook 等社交网络的种子获得大多数数据，因此能够立即索引和过滤网络素材。

此外，实时搜索引擎不需要借助预先保存的网页内容，而是实时地进行网上搜索，因而返回给用户的信息更快更新。同时这个搜索引擎还能按用户的个性

化查询要求对搜索结果进行综合，免去了用户大量繁重的人工操作。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会网络的特征，更好的为用户提供基于社交网络的实时搜索服务[5]。

图 1 给出了传统搜索引擎与实时搜索引擎的各自的工作原理和二者的流程比较。首先，从搜索效率上而言，相比于传统搜索引擎费时的网络爬虫抓取页面，实时搜索能够更高效直接地从各类外部数据源获取数据，并迅速将查询结果返还给用户。其次，从搜索质量上，传统搜索引擎的数据库更新慢，主要记录静态内容，而实时搜索引擎立足于社交网络，满足了用户对实时信息的要求。此外，对搜索的信息进行相关性排序和索引的算法符合社交网络的特征。

3. 目前出现的实时搜索引擎

目前已经出现了不少实时搜索引擎。Twitter 在 2009 年集成了 tweets 的实时搜索到社交网络服务中，是目前最大的实时数据来源。

3.1. OneRiot

OneRiot 实时搜索引擎从 Delicious、Digg、Friend-Feed、Twitter 等网站和 OneRiot 自己的搜索工具栏(从 Facebook 和 Myspace 检索数据)来更新实时内容，

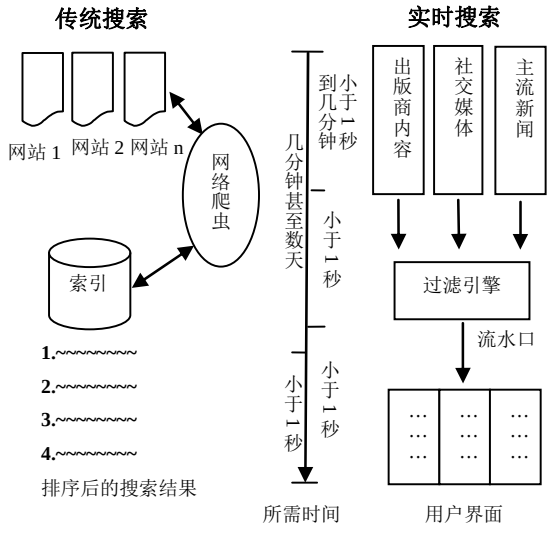

Figure 1. Comparison between traditional search process and real-time search process
图 1. 传统搜索与实时搜索流程比较

其界面如图 2 所示。该系统根据一种考虑 26 个因素的算法来过滤垃圾信息，获得查询的结果或答案。其中一种过滤是热度，即某个链接前一分钟在社交网络上被分享的程度，可以表明该内容在某段时间内流行程度的增加或降低。还有一种是用户的在线信誉。信誉建立在跟随者数量以及他们的帖子被转发的频率上。OneRiot 确定链接的流行程度是根据发送者的跟随者数量、链接被分享的速度和次数。随着搜索流量的不断增长，OneRiot 开放了自己的应用程序编程接口(API)，允许其他网站和应用程序嵌入它的实时搜索功能，并通过查询和搜索结果相关的实时广告获得收入。

3.2. Google

Google 公司于 2010 年 8 月推出实时搜索新功能，在搜索结果的显示页面，通过点击“显示选项”，允许按照时间来过滤显示结果，分类包括：“最新”、“过去 24 小时”、“过去 1 周”、“过去 1 年”、“用户指定时间范围”，其搜索界面见图 3。“最新”检索结果中实际是 Flickr、Friend-Feed、Twitter 和博客帖子的实时搜索结果。此外，Google 已经与 Facebook 和 MySpace 签订协议，能够更有效索引他们的公开内容。当然，如果用户愿意，他们可以通过 Facebook 和 MySpace 的隐私控制阻止 Google 索引他们的内容。Google 的实时搜索能够在新的相关信息出现几秒后自动滚动显示。

Google 实时算法实现了搜索的实时更新和质量评价。采用的过滤器包括结果与查询之间的相关性、读者转发次数、作者的跟随者数量等属性。Google

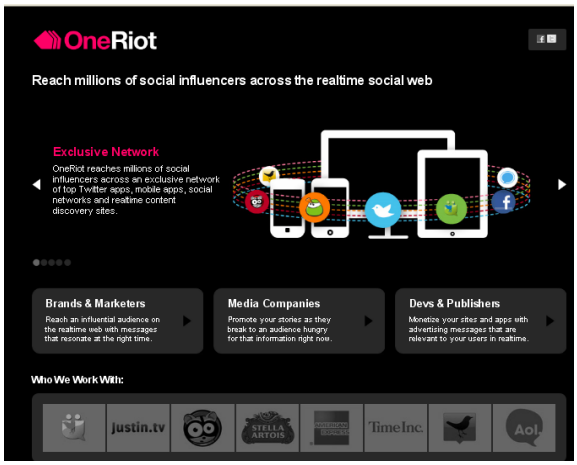

Figure 2. OneRiot real-time search site
图 2. OneRiot 实时搜索网站

爬虫现在几乎能够检索和显示出现的所有网页。公司不断开发新技术来检测每天超过 1 亿的更新文档。其中一种语言算法模型是，句子更新时，比较单词序列在其它文档中的顺序，以确定是否包含更新信息。其它算法还包括解决语义识别来明确内容的意思，缓存历史查询和响应来提高性能、测量搜索结果与查询主题的相关性、查看搜索结果集合形式来确定近期信息的重要性等。

3.3. Collecta

该搜索引擎监控 Flickr、Twitter 和 WordPress 等实时博客和网站的更新数据流，并能够在张贴后的查询中尽快显示结果。该引擎采用 XML 可扩展消息处理现场协议，使得接近实时的信息快速反馈给搜索者。服务使用长轮询方式。如果 Collecta 检测的网站服务器没有新数据，请求将持续保持到有信息产生。如果有新数据，Collecta 立即返回结果并发送另外一个请求，从不间断的数据流中提供信息给用户。Collecta 对检索结果并不排序，而是通过相关性和垃圾过滤的顺序流来显示，其搜索界面见图 4 所示。

3.4. 其他搜索引擎

微软实时搜索：Bing(必应)是微软的新版实时搜索引擎，拥有访问 Twitter 的实时数据种子的权限，并获得授权使用 Facebook 的 API 服务。Twitter 是 Bing 搜索的主要数据来源。通过嵌入 Delicious、Digg、Flickr 和 Twitter 等多个社交网络服务，雅虎与微软的搜索联盟(新搜索引擎名称 Scoopler)能够提供现场的、自动更新的实时搜索结果。Scoopler 用一列实时提供了查询相关的最流行的链接、视频和图片，根据内容多新和社交网络共享的程度来排序。其他相关网络内容出现在另一列。管道数据延迟最大为 30 秒。友好的种子允许用户提交查询和对网上朋友帖子做出实时反应。该服务还引人 Flickr、Twitter 和 YouTube 的即时更新。

Topsy：显示 Twitter 上发布的帖子的实时搜索结果，并且根据链接出现在 Tweets 上的次数来排序。越多的人转发一个帖子，该帖子会被给予更多的权重。

CrowdEye：是个 beta 版本。采用自己的实时算法对来自 Twitter 的搜索结果进行排序。主要依据发帖者的跟随者数量和被转发的次数来确定。

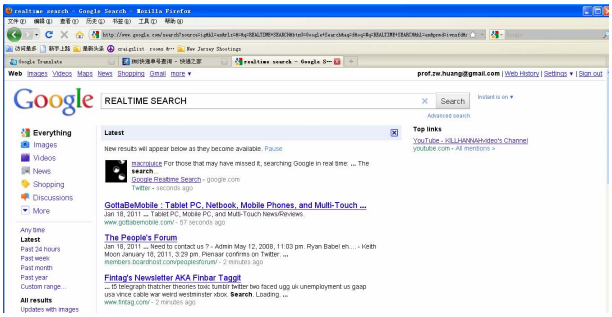

Figure 3. Google real-time search site
图 3. Google 实时搜索网站

Figure 4. Collecta real-time search engine interface
图 4. Collecta 实时搜索引擎界面

有道搜索：率先在国内推出的实时搜索服务，搜索结果有的来自新华网、网易等主流新闻媒体，但更多是来自主流微博和百度贴吧的内容。对检索内容进行时间相关性排序后显示，可以按照“最新”、“一天内”、“一周内”、“一月内”、“一年内”分类显示信息。

4. 实时搜索引擎展望

实时搜索尚处在起步阶段。2008 年以前，实时搜索技术还不存在，了解和使用实时搜索的用户还不多。虽然经过了多年的开发，但目前的实时搜索和过滤功能还算不上完善，搜索结果可能带来很多无关的或者重复的数据，更可能会导致严重的垃圾邮件和广告。由于实时搜索需要考虑很多因素，并在此基础上进行快速、复杂的计算，因此需要大量的计算开销。目前 Twitter 占据了实时信息来源的绝大部分，这也确定了它在实时搜索中的统治地位。实时搜索现在主要用于与 Twitter 相关的信息发布类网站或者微博应用上。

不远的将来，更多的实时搜索应用将会产生。1) 精确定位广告促销。实时搜索能够马上提供与用户搜索主题相关的产品广告。例如，用户搜索某一个影星时可能会出现该影星所代言的 DVD 促销广告。2) 用于定制用户感兴趣的话题。用户能够定制获取他们感兴趣的主题或者参与事件的信息来源。将来，实时搜索可能集中应用于与位置相关的问题。例如，移动设备用户可能会通过社交网络发布交通阻塞情况，并不断地更新状况，人们可以利用实时搜索发现这个信息，这便构成了一个让其他人受益匪浅的路况报道。3) 社交网络中正面宣传公司形象。社交网络中可以以正面方式讨论关于公司品牌的对话，他们可能成为响应客户需求和投诉的对话的一部分，并且指导正面的对话。4) 辅助决策。用户甚至可以采用实时搜索来进行现场财务和其他决策。

总而言之，实时搜索将会变得更加普遍，消费者将进入一个永远在线的网络。而且，随着进一步的应用推广，以及与物联网、实时通讯等技术的进一步融合，实时搜索将会出现在电子商务、物流服务等多种场合。实时搜索与传统搜索的一个很大区别是过滤效果不佳，在搜索时间敏感性不强的主题方面的搜索精度不如传统搜索引擎。但是，随着人们对搜索引擎的关注和搜索市场的增长，实时搜索技术也将不断进步，越来越多的实时搜索应用将会诞生。

参考文献 (References)

[1] D. Sullivan. What is real time search, 2009. http://searchengineland.com/what-is-real-time-search-definitions-players-22172

[2] B. J. Jansen. Real time search on the web: Queries, topics and economic value, 2011. http://collecta.com/#q=real%20time%20search

[3] 侯震宇. 基于 Fish 算法的实时搜索系统的实现[J]. 现代图书情报技术, 2002, 6: 33-35.

[4] 徐婕, 康慕宁, 董谷音. 基于社交网络的实时搜索引擎的排序算法研究[J]. 科学技术与工程, 2011, 11(28): 6879-6882.

[5] 邓志宏. 实时信息搜索技术的研究[J]. 信息技术, 2011, 11: 27-30.

