社交网络服务发展与现状研究

吴大愚

（西安政治学院军事系，陕西 西安 710068）

摘 要：介绍了社交网络服务（SNS）的定义、分类以及和其他类似应用的关系，对什么是社交网络服务，社交网络服务的发展历史与现状以及其流行的原因进行初步研究。通过对相关信息和文献的梳理评述，探讨了社交网络服务相关研究领域的内容。

关键词：社交网络服务；SNS；web2.0

社交网络服务（social network service,SNS）是近年来兴起的一类网络应用，其提供了一种新的方式供使用者联系、交流和学习。使用 SNS 的用户可以选择自己感兴趣的人、机构进行关注，并享受这些关注对象推送的各类信息。SNS 是以真实人际关系为基础的社会网络，将广大用户纳入到一个虚拟与现实相结合的平台。作为 Web2.0 的代表性应用，SNS 让一般用户自己发布信息而不再是由网站发布所有的信息。SNS 的发展引起了国内外学者的巨大兴趣，本文通过对当前研究 SNS 的代表性成果进行了梳理，以期对 SNS 研究和企业利用 SNS 的实践提供参考。

1 什么是 SNS

1.1 SNS 的定义

因为 SNS 发展非常迅速，因此有关 SNS 的定义有很多，常见的有两个。一是密歇根州立大学的 BOYD 和 ELLISON 给出的定义：SNS 是一种允许网民在一个受限制的系统上构建一个公开或半公开的个人空间的网络服务，在空间里面可以列出关注用户名单的链接，此外网民还可以查看自己的链接和相关用户的链接[1]。另一个是在维基百科上给出的 SNS 的定义，SNS 主要作用是为一群拥有相同兴趣与活动的人创建在线社区（或称为平台、网站）。SNS 为每个用户提供一个个人主页，包含用户的社会关系等信息。SNS 还为用户提供各种联系、交流的交互渠道，如电子邮件、实时消息服务等。此外，美国学者 Barn 认为 SNS 是能为人们提供在线个人空间并与他人分享的网站[2]。

1.2 SNS 的分类

SNS 根据功能可以分为综合性和垂直型的 SNS 网站，综合性的网站主要是指 Facebook、MySpace、人人网、开心网、朋友网这类网站，垂直型 SNS 网站主要是专注于某一类特定的应用，可以分为商务型、婚恋类、兴趣类、校友类等[3-4]，主要有 LinkedIn(商务)、世纪佳缘(婚恋类)、豆瓣(兴趣类)等。此外，随着时间发展，很多 Web2.0 的应用，例如微博、内容社区，论坛等也都逐渐加入了 SNS 的元素，各类应用之间的界限逐渐模糊，其分类也有待于进一步探究。

1.3 SNS 和其他类似应用的关系

Web2.0 是一种以 XML、RSS、AJAX 等技术为基础，并以 Blog、SNS、TAG、RSS、Wiki、微博、内容社区等应用为核心，满足不同用户社会化、人性化需求的互联网新一代模式[5]。SNS 应用是 Web2.0 时代的典型应用，微博、内容社区等应用也具有和 SNS 类似的功能，例如微博和以图片视频等内容为核心的内容社区（例如 YouTube、优酷、Flickr）也都具有用户发布内容、好友评论、分享等 SNS 所具有的功能。SNS、微博和内容社区的区别在于，SNS 是一种综合的平台，其强调的是用户的实名认证，是熟人网络。微博侧重于信息的传播，强调的信息分享的便捷性。内容社区强调的是内容，有了优质的内容才能吸引用户讨论和分享。

2 SNS 发展历史和现状

2.1 SNS 理论基础

SNS 其背后的理论基础源于哈佛大学的心理学教授 Milgram 提出的“六度分割理论”，是指“你和任何一个陌生人之间所间隔的人不会超过六个”[6]。影响 SNS 发展的还有一个“150 定律”，是指每个正常人日常密切联系的人际网络规模是 150 人左右[7]。

2.2 国外 SNS 历史与现状

最早的 SNS 网站创立于 1997 年，名字就叫 SixDegrees.com，是一个学术实验性质的 SNS 网站。Facebook2004 年在美国成立，是目前全球最大的 SNS 网站。截止 2011 年 6 月，Facebook 用户达到 7.5 亿，美国国内用户达到 1.5 亿以上，占到整个美国人口的一半[8]，在全球很多国家 Facebook 也都成为了当地最大的 SNS 网站，其发展前景被大多数人看好。Friendster 在美国于 2003 年成立，是商业 SNS 网站的鼻祖，一直被 SNS 业界称为全球首家社交网站，在全球范围内包括中国都掀起来了 SNS 网站热潮。但是之后在美国很快就失去了优势。MySpace 成立于 2004 年，2005 年被新闻集团收购后迅速发展，受到了美国大众的追捧，一度成为社交网络中最流行的网站。2008 年底达到其巅峰时期，随后被 Facebook 超过并逐步衰退。HI5 成立于 2003 年，一度于 2008 年全球排名第三，2010 年因效益不佳走向没落，2011 被收购。Orkut 是谷歌 2004 年创建的实验性 SNS 网站，未大力推广。Wallop 是微软 2004 年创建的实验性 SNS 网站，已经被关闭。LinkedIn 成立于 2003 年，是一家面向商业客户的 SNS 网站，其目的是让注册用户维护他们在商业交往中认识并信任的联系人。Google+是谷歌为了抗衡 Facebook 于 2011 创建的 SNS 网站，虽然得益于谷歌的用户群发展迅速，但因为失去先机，目前明显落后。Netlog（前身为 Facebox 和 Redbox）2004 年成立于比利时，以欧洲年轻人为目标用户。截至 2011 年 7 月，Netlog 仅在欧洲拥有超过 78 千万年轻用户，覆盖 39 个语种。Mixi2004 年成立于是日本，是日本最大的社交网站，目前是日本排名第三的网站。

2.3 国内 SNS 历史与现状

当前国内主流的综合性 SNS 平台主要有腾讯，人人网，开心网。腾讯借助其 QQ 巨大的用户群提供一系列的 SNS 服务，是目前国内最大的 SNS 平台，其 2011 年开放的朋友网是一个标准意义上的 SNS 网站。人人网前身是成立于 2005 年以大学校园为目标的校内网，功能上类似于 Facebook，2007 年开始面向整个社会开放。开心网成立于 2008 年，核心用户锁定为白领人群，偏重于娱乐性，早期发展势头非常迅速。相应的垂直型 SNS 还有，以建立在读书观影听音乐等兴趣上具有浓郁小资色彩的豆瓣网，以严肃婚恋为主题的世纪佳缘，淘宝建立的以分享购物体验为基础的淘江湖，6 到 14 岁儿童的 SNS 社区淘米网，走商务 SNS 路线的优仕网、经纬网，门户网站搜狐推出的面向白领的 SNS 社区白社会等。总体来说，国内各类 SNS 社区发展迅速，目前还没有形成一支独大垄断市场的局面。但相比 Facebook，国内各 SNS 社区普遍存在用户黏性不够容易流失，盈利方式不够明确等问题。

3 SNS 流行的原因

SNS 类应用能在短短不到十年的时间风靡整个世界，同时影响着互联网和真实世界，如此流行的背后有其深层次的原因。艾瑞咨询网 2011 年做了中国网站用户使用社交网络的原因调查[9]，其中排名前四的主要原因是休闲娱乐，广泛交友，维护朋友关系和寻找以前的好友。有学者总结 SNS 之所以流行，是因为其相对于其他网络应用更加真实，能够维系现实生活中的人际关系，并且具有管理边界，能很好的处理公开与隐私的问题[10]。此外，还有很多学者从不同的视角研究了网民使用 SNS 的影响因素，比较典型的有：从心理学的视角上看，影响网民使 SN 的因素是归属感、集体自尊感、社会规范、与他人交往的兴趣、接触专门知识的兴趣和自我效能等。有学者研究了社会认同感、孤独感和 SNS 使用之间的关系，得出孤独感越低、社会认同越低的人使用 SNS 越多的结论[11]。从网站设计来研究使用 SNS 的动机，主要包括趣味性、信任性、有用性和易用性[5-12]。对 SNS 影响因素的研究则多数聚焦在网民来源、年龄、性别、族群、教育程度等统计变量方面[13]，此外口碑传播也是 SNS 流行的重要影响因素之一[14]。

4 结束语

当前由于 SNS 实践的发展速度很快，特别是网络社区的 SNS 化运动的出现，学术界对 SNS 的研究还处于初级阶段，很多基础性研究的结论还需要不断更新。此外，学术界的研究和 SNS 实践领域存在着一定程度的脱节，学术界所热衷于研究的领域和实践环节目前热门的领域重合度有限。SNS 这一互联网革命对人类生活的影响将是长久并深刻的，还需要更多的学者去研究。

参考文献

[1]Boyd. D M,& Ellison N. B.(2007). Social network sites: Definition, history,and scholarship. Journal of Computer-Mediated Communication,13(1),article 11.

[2]Barnes S. A privacy paradox: Social networking in the United States [J].First Monday,2006,11(9)

[3]陈宁.360 圈社交网站发展战略研究[D].北京:北京邮电大学.2008.

[4]丁欣.决胜 SNS:产品设计运营开放平台社会化营销[M].北京:人民邮电出版社.2009.

[5]林颖.基于 web2.0 技术的 SNS 现状及发展趋势研究.[2011-12-21].http:

基于 DOM 的 Web 数据抽取研究

郭东峰

（新乡学院 计算机与信息工程学院,河南 新乡 453000）

摘　要:文章阐述了利用 XML 中的 DOM 树将 Web 数据结构分析,转化为结构化的 XML 数据,使用 Xpath 实现数据匹配查找数据,通过正则表达式实现数据抽取。同时,对目前数据抽取技术做一些简单探讨研究。

关键词:数据抽取;XML 数据;DOM 树

引言

随着 Internet 的快速发展,Web 上的数据信息急剧增加,成为了世界上规模最大的公共数据资源。目前虽然搜索引擎为用户查找信息提供了简便的方法,但它只是提高了 Web 文档的检索效率,只能根据用户提交的关键词返回一组网址,用户必须逐一浏览网址对应的 Web 页,采用人工的方式定位最终信息,现有的搜索引擎本身不能直接定位到所需的数据,更谈不上为数据增加语义。XML 技术出现之后,因为其定义严格,语法明确,结构良好,已经迅速成为互联网信息表示的事实标准,通过把 HTML 文档转换成 XHTML,借助于 DOM 分析技术,可以方便从中提取有用信息。

1 WEB 数据抽取

Web 信息抽取是一种从 Web 文档中抽取出有用信息的技术,可以大大的缩短了对资料的整理时间,为信息检索提供方便,有利于现实文档的存档管理。我们可以利用行业信息模型和领域特征做主题搜索,在收集信息时去除领域无关的信息,在信息检索时实现更优秀的查询扩展,从而提高搜索结果的查全率和查准率,有效解决通用搜索系统给出的检索结果往往过于繁杂,用户甄别信息价值的时间长问题。主题搜索利用逐渐成熟的文本分类技术,去除用户不关心数据,具有更多的针对性,减少搜索、浏览时间中的比重,使其满足人们对信息的精准化需求,提高工作效率。

2 信息抽取方法发展情况

2.1 手工方法:通过观察网页及其源代码,由编程人员找出一些模式,再根据这些模式编写程序抽取目标数据。然而这种方式无法抽取站点数量巨大的形式。手工方法由于设计难度大,只能针对少量网页抽取,目前基本不再使用。

2.2 包装器归纳:即有监督学习方法,是半自动的。从手工标注的网页或数据记录集中利用机器学习方法序列覆盖学习一组抽取规则。随后这些归则即被用于从具有类似格式的网页中抽取目标数据项。由于需要手工标注的工作,不适合对大量站点抽取,并且维护开销大。

2.3 自动抽取:即无监督学习方法,给定一张或数张网页,这种方法自动从中寻找模式或语法,以便进行数据抽取。自动化抽取的主要优点是它能处理大量站点的情况,并且维护开销小,主要缺点是因为系统不知道用户对什么感兴趣,它可能抽取了大量不需要的数据。

3 DOM 树的解析、扩展和 Xpath 使用

文件对象模型(Document Object Model,简称 DOM),是 W3C 组织推荐的处理可扩展置标语言的标准编程接口。DOM 可以先将 XML 文档解析成结点对象以元素、属性、实体和注释等节点形式存放信息的树形分级结构,然后以节点树的形式在内存中,由于树形数据结构应用较为广泛,有很多成熟的算法可以用来遍历、搜索、编辑 XML 文档树,同时借助于 JDOM、DOM4J、SAX 等技术类库可以更加方便的访问分档中的数据。

XPath 是一种用于查询 XML 文档中的信息的语言，是定位 XML 文档节点的声明式语言，是 W3CXSLT 标准的主要组成部分。Xpath 规范定义了允许到 XML 文档各个部分的路径说明的表达式语法和支持这些表达式的核心库基本函数。主要用于识别、选择和匹配 XML 文档中的各个组成部分，包括元素、属性和文本内容等。XPath 可以使用路径表达式方便地定位 XML 节点,所以很适合于数据抽取。

4 Web 信息抽取的概念及实现流程

Web 信息抽取就是从 Web 页面中抽取目标信息的问题，从网页中所包含的无结构或半结构的信息中识别用户感兴趣的数据,并将其转化为结构和语义更为清晰的格式(XML、关系数据、面向对象的数据等)。基于 XML 技术抽取的流程为:首先,从网络中获取 HTML 文档；然后，经 Tidy 等工具处理后转换为符合 XML 格式的 XHTML 文档,再使用 XSL 保存的数据抽取规则,经 XSLT 处理抽取出 XML,中对原始的 HTML 文件加工清洗,经过使用工具 Tity 对网页语法检查及纠错，将 HTML 文档转换为结构完整的 XHTML;第三，使用 HTMLParser 等工具解析 XML 文档生成 DOM 树模式;最后,利用 Xpath 和正则表达式信息抽取规则提取有价值的信息存储到数据库中以便使用。

5 DOM 子树最大匹配求方法

设有两棵树 T1=RA:<A1,…,Ak>和 T2=RB:<B1,…Bn>,RA,RB 分别为两棵树的根,Ai 和 Bj 分别是 T1 的第 i 个和 T2 的第 j 个第一层子树。设 M(T1,T2)为求 T1,T2 最大匹配的节点个数。当 RA 和 RB 相同时，即两棵树的根部相同,T1 和 T2 的最大匹配就是 M(T1,T2)=M(<A1,…Ak>,<B1,…Bn>)+1，否则 M(T1,T2)=0。其中有递推公式:M(<A1,…Ak>,<B1,…Bn>)=max(m(<A1,…Ak-1>,<B1,…Bn-1>)+M(Ak,Bn)，m(<A1,…Ak-1>,<B1,…Bn>)，m(<A1,…Ak>,<B1,…Bn-1>)),M(<>,<>)=0,M(s,<>)=M(<>,S)=0;计算出 DOM 结点的最大匹配值,就可以通过选择合适的阀值,找出具有相同结构模式的 DOM 子树，这些子树一般为网页表格中的行<tr>…</tr>或列表项
…
就是需要集中抽取的数据区域。

6 结束语

Web 数据抽取技术目前还处在不断发展之中，是 Web 数据挖掘研究领域中的难题和热点。本文论述了基于 DOM 技术查找网页中的数据区域方法,维护开销小,具有很强的实用价值。值得注意的是还存在着改进的地方，比如抽取了一部分用户不感兴趣的数据,这可以尝试使用领域分词过滤掉不需要的信息加以完善。

参考文献

[1]蔚晓娟.基于 DOM 的 XML 解析与应用[J].计算机技术与发展,2007.17(4).

[2]李雪竹.一种基于 XML 的 Web 数据抽取的实现[J].科学技术与工程,2008(9).

[3]尹津其.基于 WEB 的数据抽取及应用实例[J].中国新技术新产品,2009(19).

//wenku.baidu.com/view/0481db4c852458fb770b5652.html

[6]百度百科.六度空间理论.[2011-12-25].http://baike.baidu.com/view/200573.htm

[7]百度百科.150 定律.[2011-12-25].http://baike.baidu.com/view/1560858.htm

[8]网易科技.扎克伯格证实 Facebook 全球注册用户 7.5 亿.[2011-12-12].http://news.xinhuanet.com/games/2011-07/07/c_121635946.htm

[9] 艾瑞咨询集团.2010-2011 年中国社交网络用户行为研究报告简版.[2011-12-20].http://report.iresearch.cn/Reports/Free/1606.html

[10]陈卉.社会性网络服务(SNS)流行原因分析.新闻世界.2009(05):114-115.

[11]阴良.孤独感、社会认同与 SNS 使用之研究-以人人网为例.新闻大学.2010(4):8-18.

[12]Hossain L,Anjalide S. Exploring User Acceptance of Technology Using Social Networks [J]. Journal of H igh Technology Management Research,2009(20):1-18

[13]Baker V. Older Adolescent´s Motivations for Social Network iSite use: The Influence of Gender,Group Identity and Collective Self-esteem[J].Cyber Psychology & Behavior,2009,12(2):209-213

[14]Trusov M,Bucklin R E,Pauwels K. Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site[J]. Journal of Marketing,2009,73(September):90-102.

作者简介:吴大愚,西安政治学院军事系讲师,少校。研究方向:兵棋,信息安全,网络舆情,软件工程。

社交网络服务发展与现状研究

作者：吴大愚
作者单位：西安政治学院军事系,陕西 西安 710068
刊名：科技创新与应用
英文刊名：Technology Innovation and Application
年，卷(期)：2013(18)

参考文献(14条)

1. Boyd. D M;Ellison N. B Social network sites:Definition, history,and scholarship 2007(01)

2. Barnes S A privacy paradox:Social networking in the United States 2006(09)

3. 陈宁 360圈社交网站发展战略研究[学位论文] 2008

4. 丁欣 决胜SNS:产品设计运营开放平台社会化营销 2009

5. 林颖 基于web2.0技术的SNS现状及发展趋势研究 2011

6. 六度空间理论 2011

7. 150定律 2011

8. 扎克伯格证实Facebook全球注册用户7.5亿 2011

9. 艾瑞咨询集团 2010-2011年中国社交网络用户行为研究报告简版 2011

10. 陈卉 社会性网络服务(SNS)流行-因分析 2009(05)

11. 阴良 孤独感、社会认同与SNS使用之研究-以人人网为例[期刊论文]-{H}新闻大学 2010(04)

12. Hossain L;Anjalide S Exploring User Acceptance of Technology Using Social Networks 2009(20)

13. Baker V Older Adolescent's Motivations for Social Network iSite use:The Influence of Gender,Group Identity and Collective Self-es-teem 2009(02)

14. Trusov M;Bucklin R E;Pauwels K Effects of Word-of-Mouth Ver-sus Traditional Marketing:Findings from an Internet Social Networking Site 2009(September)