网络用户兴趣的智能挖掘方法研究

李 培,马 力
(西安邮电大学,陕西 西安 710061)

摘 要:目前网络上的重要应用都是围绕对用户兴趣的研究和发现而展开和完善的,主要的方式是借助于对用户的Web访问数据进行相关挖掘。该研究主要是通过建立一个从底层数据获取到上层数据处理的原型系统,对真实捕获的网络数据利用小世界网络模型提取中文文档关键字后处理为用户兴趣,再将用户的访问兴趣通过隐马尔可夫模型抽象成一种时间序列,依次反映用户兴趣的序列性,从而利用GSP算法得到用户的兴趣并供后续处理。实验证明,该原型系统从数据获取到最终处理,可以得到比较满意的结果。

关键词:兴趣挖掘;文本聚类;智能算法
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2014)02-0076-03
doi:10.3969/j.issn.1673-629X.2014.02.018

Research on Intelligent Mining Method for Web Users Interests

LI Pei, MA Li
(Xi'an University of Posts and Telecommunications, Xi'an 710061, China)

Abstract: At present important applications on the network is expanded and improved around research and discovery of user interest, the main way is through Web access to data mining. The research is mainly through the establishment of a prototype from underlying data acquisition to the upper data processing, for an actual network data captured apply the small world network model to extract keyword as user interests, which are extracted a kind of time series by hidden Markov model, which was used to reflect the sequential features of the user interest. Use GSP algorithm to mine the user's interests for processing. Experiments show that the prototype system from data acquisition to final disposal can be more satisfied with the results.

Key words: mining interests; text clustering; intelligent algorithm

0 引 言

网络的飞速发展及广泛普及,使得如何高效地从海量数据中获取有用知识,如何及时地获取最新信息,如何提高信息检索与推送的智能水平,以及如何满足各种用户不同的个性化需求等,成为了新的信息服务应用系统面临的挑战性课题。而Web数据挖掘成为了解决这些问题的有利工具,借助Web数据挖据,可以将网络应用从被动接受转化为主动感知,实现主动和有针对性的信息服务即个性化主动信息服务,而个性化主动信息服务的核心技术是获取网络用户并进行相关处理。

目前获取网络用户兴趣的途径主要有两个方面,一方面是用户浏览的网页内容,另一方面是用户浏览的页面路径。笔者认为其中的页面内容能最直观地反映用户兴趣,因此该研究没有采用基于页面路径的研究方式,而是从用户浏览的页面内容来挖掘用户的兴趣,并为了能够在真实环境下验证研究的正确性,搭建了获取和处理网络用户兴趣的原型系统。在底层捕获用户所使用的机器发送的数据包,然后提取出其中有效的URL,接着依据这些URL使用网页爬虫程序取得用户浏览的网页,通过页面的预处理,利用智能算法进行页面关键词提取,用户兴趣的获取,最终存储用户兴趣供后续其他应用的使用和处理[1-3]。

1 原型系统的设计

根据研究需要,该原型系统需要满足能够准确提供研究数据,能够根据研究需要进行扩展等原则,各部分的功能如下:

收稿日期:2013-05-02 **修回日期**:2013-08-04 **网络出版时间**:2013-11-29
基金项目:国家自然科学基金资助项目(61105064,61203311);陕西省自然科学基础研究计划项目(2011JM8007);西安邮电大学校青年教师科研基金(ZL2013-24)
作者简介:李 培(1980-),女,陕西西安人,硕士,讲师,研究方向为网络安全、智能信息处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20131129.0857.017.html

(1)数据包捕获模块:尽可能全部获取用户访问网络页面时所产生的网络数据包,并进行保存。

(2)网页采集模块:从数据包文件中进行数据包分析,提取其中的有用 URL,根据 URL 利用网页爬虫获取对应的网页并保存。

(3)网页预处理模块:从获取的网页中清除非内容文字,包括:HTML 代码、JavaScript 代码、CSS 代码、超链接内包含的文字、广告信息等,从而得到"干净"的纯文本文件。

(4)关键词提取模块:利用小世界网络理论(SWN)[4]提取经过预处理得到的页面文本文件中的关键词。

(5)兴趣提取模块:在提取文档关键词的基础上,利用 Newman 算法[5]进行文本聚类,提取出反映用户兴趣的词语。

(6)兴趣存储和处理模块:建立用户兴趣存储库,对用户兴趣进行存储和归类。

2　原型系统的实现

2.1　网页文档获取模块

网络用户浏览网页是通过使用 www 浏览的客户端程序浏览器,需要在浏览器的地址栏中输入想要访问页面的 URL,接下来按回车键后等待服务器返回给浏览器的网页信息。整个过程从实质上分析就是网络数据包的传输,而数据包中包含的就有所需要的信息,因此利用数据包获取原理来进行网页文档的获取。

首先将网卡设置为混杂模式,就可以接收到物理媒体上的所有数据,利用 Winpcap 编程就可以获取网络中传输的各种数据包,并将其传输给调用它们的应用程序。Winpcap 能够设置过滤条件,因此可以根据需要捕获特定用户的特定协议的网络数据包。

网页用户浏览网页的时候,其发送的数据包中包含用户所要浏览的网页的 URL。因为想要获取用户访问的网页信息,一方面可以直接通过捕获接收到数据包,对数据包的内容进行分析还原得到服务器传送的网页内容,这样的好处是有实时性的优点,缺点是容易产生丢包影响结果;另一种方式可以从捕获用户发送的数据包,通过分析提取其中的 URL,根据这些 URL 利用网页爬虫程序获取其网页,这样的好处是相对准确,丢包少,缺点是不具有实时性。考虑到该研究中对用户兴趣获取的准确性需要,而用户兴趣的收集本身就需要一段时间的积累,因此实时性可以放在次要的地位,于是采用间接的方式获取网页信息。

2.2　关键词提取模块实现

根据 SWN 关键词提取算法,系统分成四个功能模块实现,自底层向上,依次包括:预处理模块、文本结构图构造模块、特征数据计算模块和关键词提取模块。

(1)预处理模块:实现对文本进行分词处理,并记录各个分词的相关信息,从而提供后续处理所需的数据。这里采用的中文分词系统是中科院计算所的开源的基于多层隐马模型的汉语词法分析系统(Institute of Computing Technology, Chinese Lexical Analysis System, ICTCLAS),其中词性标注方法选用的是北京大学设计的汉语文本词性标注标记集[6]。

(2)文本结构图构造模块:由于文本结构图中节点之间的边数较稀疏,所以采用邻接表作为结构图的物理结构。

(3)特征数据计算模块:实现计算平均路径、聚类系数、平均路径变化量、聚类系数变化量等功能。其中的平均路径,采用 Dijkstra 算法。

(4)关键词提取模块:先计算得到候选关键字集 Lmax 和 Cmax。然后合并候选关键字集即可得到关键字[7-8]。

2.3　提取模块实现

兴趣提取模块是该原型系统的核心功能模块。根据 Newman 算法,在实现时采用三个自下而上的模块实现,包括:构图子模块,聚类模块和兴趣生成模块。图 1 给出了兴趣提取模块实现的功能结构层次图,其中虚线框代表各个功能模块,带箭头的线代表模块间的调用关系。

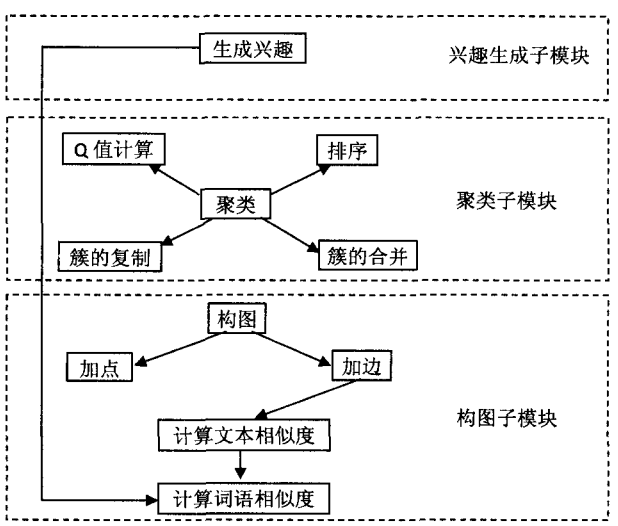

图 1　兴趣提取模块的功能结构层次图

该模块中的核心算法是 Newman 算法,该算法是一种基于图的聚类算法,构图中用节点表示独立的文本,连接节点的边为相应的文本相似度大于预先设定的文本相似度阈值而添加的,该边的权值为两个文本计算得到的相似度的量化值。

其中,文本相似度的计算是基于词语的相似度,采用了成熟的《知网》提供的词语语义相似度计算方法[9-10]。

3 原型系统的运行及分析

实验数据集是使用原型系统在局域网中针对某个用户捕获其一个月内所有的数据，通过分析提取了921个网页文本。该系统的实验包括关键字提取算法的验证、文本聚类算法的验证以及兴趣提取算法的验证。其中关键词提取算法的验证在之前的论文中已经阐述过，所以文中主要对文本聚类算法以及兴趣提取算法进行验证[11]。

文本聚类算法验证中，从提取的921个文本中分析和抽取了600个有用文本，并参照之前的实验语料库的分类标准[12]，即按照科技、体育、时事政治、娱乐、军事分为五类，先将其进行人工的手工分类，再使用提出的文本聚类算法来进行实验模拟验证。通过实验测试，得到词语相似度阈值设置范围是[0.7,0.95]，文本相似度阈值的范围是[0.3,0.6]，同时，如果词语相似度阈值或者文本相似度阈值增大，则簇的内部分裂程度增大，类之间分离程度越清晰。表1列举出了当词语相似度阈值设置为0.92时的聚类效果，可以看到五类文本都准确地进行了聚类。

表1 不同词语相似度阈值情况下的聚类效果

P =0.92	召回率	准确率
科技	0.857	1
体育	0.556	1
时事政治	0.889	1
娱乐	0.833	1
军事	0.4	1
平均值	0.707	1

通过图2和图3可以看到簇的数目越小，即簇中元素个数越多，召回率越高，由于这时簇的数目与实际的数目相差很多，因此模块度也较小。同时由于采用的是无监督聚类算法，聚类个数不定，会出现没有形成簇的单元素的噪声簇。

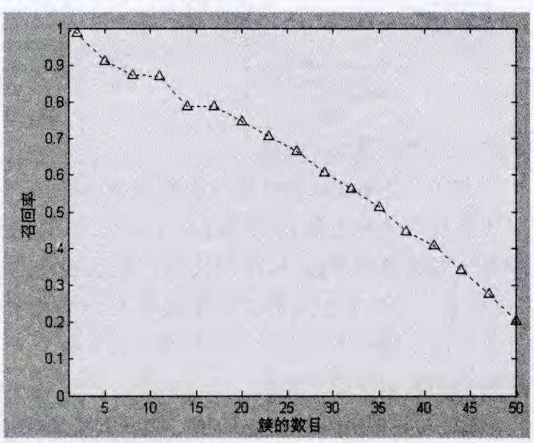

图2 簇的数目与召回率的关系

对于兴趣提取算法的验证，由实验得出：兴趣度阈值一般设置范围是[0.4,0.8]。兴趣度阈值与准确率的关系如图4所示。

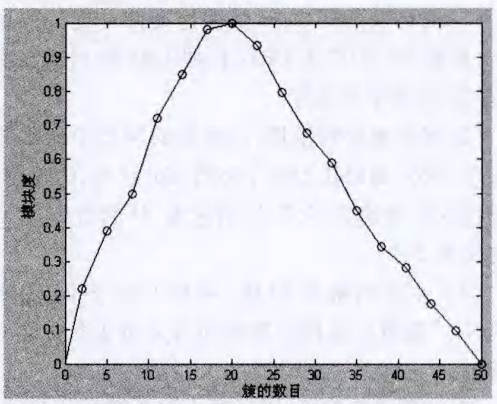

图3 簇的数目与模块度的关系

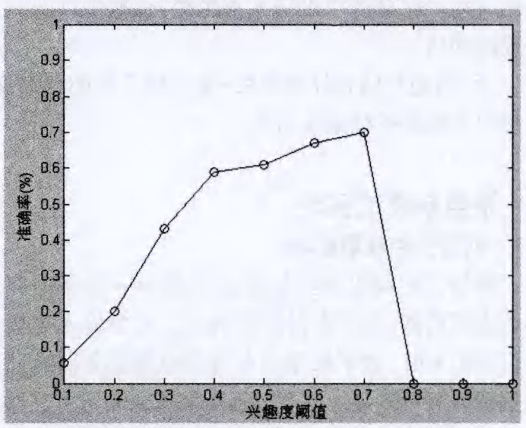

图4 兴趣度阈值与准确率的关系

根据图4的分析，为了能够得到准确的兴趣，将兴趣度阈值设置为0.78，在此基础上得到的兴趣如表2所示。

表2 兴趣度阈值设置为0.78的情况下系统提取的兴趣

网页类型	兴趣
科技	计算机 手机 通信
体育	篮球 足球
时事政治	中国 钓鱼岛 大选
娱乐	电影 音乐
军事	武器

4 结束语

文中所涉及的网络用户兴趣挖掘的智能方法研究，采用了研究算法以及原型实现两个方面进行，文中强调了实现方面的技术和方法，并利用原型系统对真实网络数据进行了测试与分析，对使用的算法进行验证。如果想要进一步地为大量的用户提供准确有效的个性化主动信息服务，还需要对获取的用户兴趣进行

（下转第83页）

参考文献:

[1] Gijsenij A, Gevers T, van de Weijer J. Computational color constancy: Survey and experiments[J]. IEEE transactions on image processing, 2011, 20(9): 2475-2489.

[2] Kibria A F M G, Monirul I M. Reduction of over segmentation in JSEG using canny edge detector[C]//Proc of 2012 international conference on informatics, electronics and vision. [s. l.]: [s. n.], 2012: 65-69.

[3] Yeek T S, Wei B C, Dickson L. Building detection with loosely-coupled hybrid feature descriptors[J]. Lecture notes in computer science, 2012, 7458: 552-563.

[4] Ajay M, Sanjeev S, Edwin H. Detection of edges in color images: A review and evaluative comparison of state-of-the-art techniques[J]. Lecture notes in computer science, 2012, 7326: 250-259.

[5] Zhu Yahui, Peng Guohua. Method for image edge detection based on UGM model[J]. Lecture notes in electrical engineering, 2012, 154: 1502-1507.

[6] 雒　涛,郑喜凤,丁铁夫. 改进的自适应阈值 Canny 边缘检测[J]. 光电工程, 2009, 36(11): 106-111.

[7] 林生佑,石教英. 基于 HVS 的彩色图像边缘检测算子[J]. 中国图象图形学报: A 辑, 2005, 10(1): 43-47.

[8] 袁春兰,熊宗龙,周雪花,等. 基于 Sobel 算子的图像边缘检测研究[J]. 激光与红外, 2009, 39(1): 85-87.

[9] 黄　伟,周鸣争,李小牛. 一种基于四元数的彩色图像边缘检测改进算法[J]. 计算机技术与发展, 2008, 18(3): 121-124.

[10] Ivan D, Carlos J, Xiao Jizhong. Incremental registration of RGB-D images[C]//Proceedings of IEEE international conference on robotics and automation. [s. l.]: [s. n.], 2012: 1685-1690.

[11] Wu Xiaohua, Jean-Pierre H. Visualization of continuous stream of grid turbulence past the langston turbine cascade [J]. AIAA Journal, 2012, 50(1): 215-225.

[12] Dos S D F D, Da S I R, Denise G, et al. Combining color and topology for partial matching[C]//Proceedings of international conference on tools with artificial intelligence. [s. l.]: [s. n.], 2012: 770-777.

[13] Hong Soon-Won, Lynn C. Automatic recognition of flowers through color and edge based contour detection[C]//Proc of 2012 3rd international conference on image processing theory, tools and applications. [s. l.]: [s. n.], 2012: 141-146.

[14] Gonzalez R C, Woods R E. Digital image processing[M]. 2nd ed. Beijing: Publishing House of Electronics Industry, 2009.

[15] Ibaa J, Akram M U, Anam T. Retinal image preprocessing: Background and noise segmentation[J]. Telkomnika, 2012, 10(3): 537-544.

[16] Koschan A. A comparative study on color edge detection [C]//Proceedings of 2nd Asian conference on computer vision. [s. l.]: [s. n.], 1995: 574-578.

[17] Asmare M H, Asirvadam V S, Iznita L. Color space selection for color image enhancement applications[C]//Proc of 2009 international conference on signal acquisition and processing. [s. l.]: [s. n.], 2009: 208-212.

[18] Kong N S P, Ibrahim H. Color image enhancement using brightness preserving dynamic histogram equalization [J]. IEEE transactions on consumer electronics, 2008, 54(4): 1962-1968.

(上接第 78 页)

进一步的分析研究,对该原型系统的性能和容量进行改进和完善,即在兴趣后期处理算法和系统实现上进一步深入。

参考文献:

[1] 宋丽哲,牛振东,余正涛,等. 一种基于混合模型的用户兴趣漂移方法[J]. 计算机工程, 2006, 32(1): 4-6.

[2] Koychev I, Schwab I. Adaptation to drifting user's interests [C]//Proceedings of the ECML2OOO/MLNet workshop on machine learning in the new information age. [s. l.]: [s. n.], 2000: 39-45.

[3] Grabtree I, Soltysiak S. Identifying and tracking changing interests[J]. International journal of digital libraries, 1998, 2(1): 38-53.

[4] 司徒俊峰. Internet 的小世界网络研究[J]. 情报杂志, 2004(12): 59-63.

[5] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Phys Rev E, 2004, 69: 026113.

[6] RabinerL R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proc of IEEE, 1989, 77(2): 257-286.

[7] 周雅夫,马　力,董洛兵. 基于 SWN 理论提取复合关键字系统的设计与实现[J]. 西安邮电学院学报, 2007, 12(5): 82-86.

[8] Watts D J, Strogatz S H. Collective dynamics of small world' networks[J]. Nature, 1998, 393(6684): 440-442.

[9] 唐歆瑜,乐文忠,李志成,等. 基于知网语义相似度计算的特征降维方法研究[J]. 科学技术与工程, 2006, 6(21): 3442-3446.

[10] 周　粉,夏幼明. 一种改进的基于知网的语义相似度计算方法[J]. 云南大学学报(自然科学版), 2008, 30(S2): 215-218.

[11] 马　力,谭　薇,李　培. 基于 Web 访问信息的用户兴趣迁移模式的研究[J]. 计算机科学, 2011, 38(5): 175-179.

[12] 李素建. 基于语义计算的语句相关度研究[J]. 计算机工程与应用, 2002(7): 75-77.