

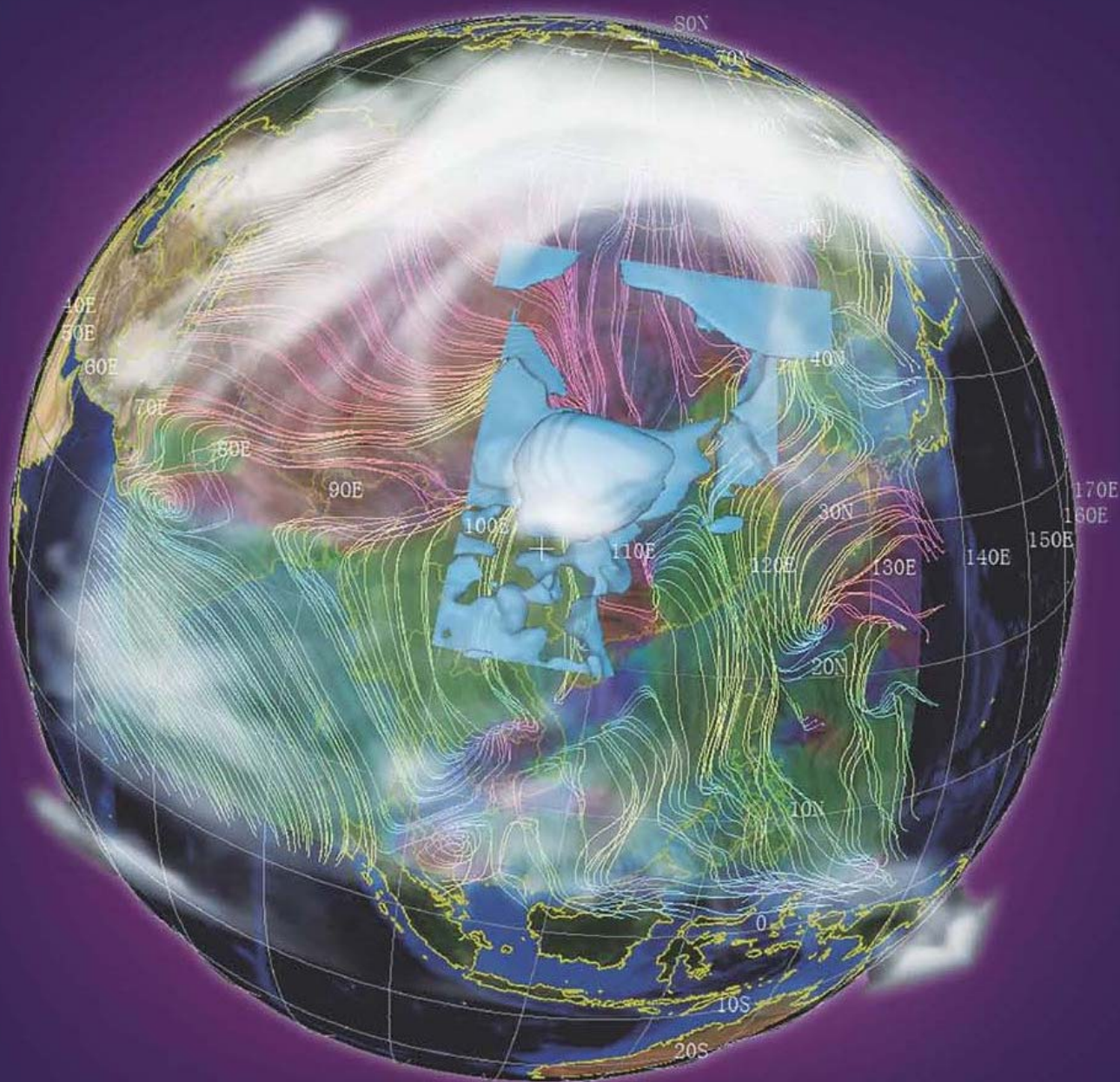


主办: 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2015
05
VOL.20

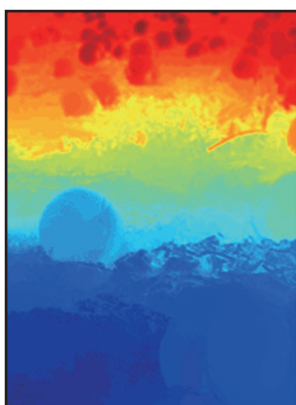
ISSN1006-8961
CN11-3758/TB



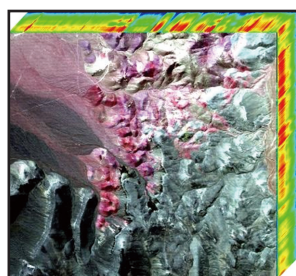
风场云场混合可视化



面向彩色增强图像的客观质量评价算法(第0643页)



使用柯西分布点扩散函数模型的单幅散焦图像深度恢复(第0708页)



基于蛙跳算法的离散粒子群优化端元提取(第0724页)

综述

中国图像工程: 2014

章毓晋 0585

图像分割中的超像素方法研究综述

宋熙煜, 周利莉, 李中国, 陈健, 曾磊, 闫钊 0599

图像处理和编码

结合通道共生和选择集成的彩色图像隐写分析

栗风永, 张新鹏 0609

复数基下的图像伪装算法

朱喜玲, 陈伟, 宋瑞霞, 齐东旭 0618

复杂自然环境下感兴趣区域检测

肖志涛, 王红, 张芳, 耿磊, 吴骏, 李月龙, 李峰 0625

结合调整差值变换的(K,N)有意义图像分存方案

欧阳显斌, 邵利平, 陈文鑫 0633

图像分析和识别

面向彩色增强图像的客观质量评价算法

李子印, 倪军 0643

基于局部模型匹配的几何活动轮廓跟踪

刘万军, 刘大千, 费博雯, 曲海成 0652

密集特征加权跟踪算法

罗会兰, 梅晶, 孔繁胜 0664

利用Kolmogorov-Smirnov统计的区域化图像分割

赵泉华, 张洪云, 李玉 0678

混合生成式和判别式模型的图像自动标注

李志欣, 施智平, 张灿龙, 王金艳 0687

基于协作表示残差融合的3维人脸识别

詹曙, 臧怀娟, 相桂芳 0700

图像理解和计算机视觉

使用柯西分布点扩散函数模型的单幅散焦图像深度恢复

明英, 蒋晶珏 0708

计算机图形学

用户驱动的微博可视化搜索

周霞娟, 汪飞, 金玲, 陈为, 王章野 0715

遥感图像处理

基于蛙跳算法的离散粒子群优化端元提取

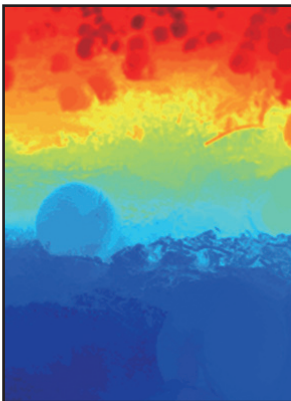
吴国伟, 赵艳玲, 王龙, 倪巍, 许立江, 冉艳艳 0724

CONTENTS

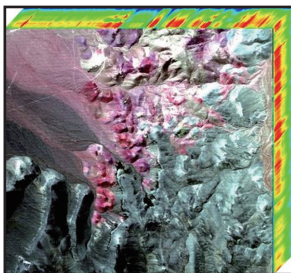
JOURNAL OF IMAGE AND GRAPHICS



Objective quality assessment
for enhanced chromatic
images(P0643)



Depth recovery from a single
defocused image using a
Cauchy-distribution-based
point spread function model
(P0708)



Discrete particle swarm
optimization endmember
extraction based on shuffled
frog leaping algorithm
(P0724)

Review

Image engineering in China: 2014 Zhang Yujin	0585
Review on superpixel methods in image segmentation Song Xiyu, Zhou Lili, Li Zhongguo, Chen Jian, Zeng Lei, Yan Bin	0599

Image Processing and Coding

Steganalysis for color images based on channel co-occurrence and selective ensemble Li Fengyong, Zhang Xinpeng	0609
Image disguise algorithm based on complex number base Zhu Xiling, Chen Wei, Song Ruixia, Qi Dongxu	0618
ROI detection under the complicated natural environment Xiao Zhitao, Wang Hong, Zhang Fang, Geng Lei, Wu Jun, Li Yuelong, Li Feng	0625
Meaningful (K, M) image sharing scheme combined with the adjusting difference transformation Ouyang Xianbin, Shao Liping, Chen Wenxin	0633

Image Analysis and Recognition

Objective quality assessment for enhanced chromatic images Li Ziyin, Ni Jun	0643
Geometric active contour tracking based on locally model matching Liu Wanjun, Liu Daqian, Fei Bowen, Qu Haicheng	0652
The dense feature-weighted object tracking Luo Huilan, Mei Jing, Kong Fansheng	0664
Regionalized image segmentation using Kolmogorov-Smirnov statistics Zhao Quanhua, Zhang Hongyun, Li Yu	0678
Hybrid generative/discriminative model for automatic image annotation Li Zhixin, Shi Zhiping, Zhang Canlong, Wang Jinyan	0687
Three dimensional face recognition by fused residual based on collaborative representation Zhan Shu, Zang Huaijuan, Xiang Guifang	0700

Image Understanding and Computer Vision

Depth recovery from a single defocused image using a Cauchy-distribution-based point spread function model Ming Ying, Jiang Jingjue	0708
--	------

Computer Graphics

User-driven visual micro-blog search Zhou Xiajuan, Wang Fei, Jin Ling, Chen Wei, Wang Zhangye	0715
--	------

Remote Sensing Image Processing

Discrete particle swarm optimization endmember extraction based on shuffled frog leaping algorithm Wu Guowei, Zhao Yanling, Wang Long, Ni Wei, Xu Lijiang, Ran Yanyan	0724
--	------

中图法分类号: TP39 文献标识码: A 文章编号: 1006-8961(2015)05-0715-09

论文引用格式: Zhou X J, Wang F, Jin L, Chen W, Wang Z Y. User-driven visual micro-blog search[J]. Journal of Image and Graphics, 2015, 20(5): 0715-0723. [周霞娟, 汪飞, 金玲, 陈为, 王章野. 用户驱动的微博可视化搜索[J]. 中国图象图形学报, 2015, 20(5): 0715-0723.] [DOI:10.11834/jig.20150514]

用户驱动的微博可视化搜索

周霞娟, 汪飞, 金玲, 陈为, 王章野

浙江大学 CAD&CG 国家重点实验室, 杭州 310058

摘要: 目的 微博作为一个社交与信息分享平台, 日信息量数以亿计, 如何高效地搜索用户感兴趣的信息成为亟待解决的问题。提出了一个新颖的用户驱动的可视化微博信息搜索方法。方法 采用特征词及其权重来建模用户的兴趣特征, 并基于此建立用户与特征词之间的相关关系。搜索微博信息时, 首先定位与检索词相关的微博用户, 在相关微博用户的微博中筛选与搜索相关的微博。另外, 采用关注度传递算法对搜索进行扩展, 将返回的特征词和微博用户进行可视化展示, 并提供交互供用户查看与选定特征词或用户相关的微博。结果 实验结果表明, 基于本文方法, 用户可以高效地定位感兴趣的微博信息。结论 以用户作为桥梁, 大大缩小了微博信息的搜索范围, 同时采用关注度传递算法对搜索进行扩展, 对结果进行可视化展示。实验表明本文方法能够使用户快速搜索出感兴趣的信息。

关键词: 可视化搜索; 扩展查询; 微博搜索; 关注度传递

User-driven visual micro-blog search

Zhou Xiajuan, Wang Fei, Jin Ling, Chen Wei, Wang Zhangye

State Key Lab. of CAD&CG, Zhejiang Univ., Hangzhou 310058, China

Abstract: **Objective** As a social platform for sharing information, a micro-blog generates millions of information daily. Thus, searching for wanted information is difficult. To solve this problem, this paper proposes a novel user-driven visual microblog search method. **Method** The method uses feature words to model the user's interest and to obtain the relationship between words and users based on the model. To search a relevant micro-blog, the method first locates micro-bloggers who are relevant to the keywords and then filters micro-blogs. The method extends the query via an attention-diverting algorithm. Finally, the relevant users and feature words are visualized on the basis of their degree of relevance. An interactive interface is provided to enable users to search micro-blogs that are relevant to the feature words or users that they are interested in. **Result** We show through a case study that users can efficiently find specific information by observing the visualization results and further interacting with the visualization interface. **Conclusion** We provide an efficient microblog search method that regards micro bloggers as a bridge to search for relevant microblogs and narrow down the search range. The proposed method also uses the attention-diverting algorithm to extend the query before the result is finally visualized. The proposed method is efficient and useful for users to find specific microblogs.

Key words: visual search; query expansion; microblog search; attention diverting

收稿日期: 2014-09-10; 修回日期: 2014-12-24

基金项目: 国家自然科学基金项目(61232012); 国家自然科学基金面上项目(61272302); 浙江省自然科学基金项目(LR13F020001); 教育部博士点基金项目(20120101110134)

第一作者简介: 周霞娟(1988—), 女, 工程师, 2014年于浙江大学获计算机应用专业硕士学位, 主要研究方向为可视分析及计算机图形学。E-mail: xiajuanzhou@pptv.com

通信作者: 王章野, 副教授, 硕导, E-mail: zywang@cad.zju.edu.cn

Supported by: National Natural Science Foundation of China (61232012); Project supported by the Natural Science Foundation of Zhejiang Province, China (LR13F020001)

0 引言

微博,如 twitter、新浪微博、腾讯微博等,作为一个新型的实时信息分享的平台近年来迅速发展壮大。用户在微博上分享生活中的所见所闻所感,发布或分享自己感兴趣的信息。微博已发展成为一个巨大的信息库。如何从这个巨大的信息库中更加迅速准确地搜索信息成为当前可视化领域研究的热点之一。与一般互联网信息相比,微博文本简短,如新浪微博文本字数限制在 140 字,信息量小;由于微博信息完全来源于微博的使用者,故其语言结构自由、组织松散^[1]。Teewan^[2]曾比较了 Twitter 和 Web 用户的搜索行为,发现 Twitter 上每个查询会话发出 1.45 次查询请求,而 Web 搜索则为 1.08 次。微博搜索过程中,用户对信息的需求比较模糊,往往关注于主题、概念、事件或地理位置,更乐于寻找特定问题的答案或发现更广泛的知识^[3]。因此,微博搜索更倾向于探索式搜索。智能搜索针对微博数据提出了多种度量标准进行评分排序。然而,由于博文简短,用户难以从搜索的结果中直接寻找答案,往往需要综合多方面信息方能得到有意义的结论。可视化搜索利用人眼的高带宽寻找图形化的数据特征,提高了搜索的准确性和效率。但是对于微博这样的复杂数据目前仍缺少有效的表示方法。

针对微博信息的特点,提出了用户驱动的微博可视化查询方法。微博用户,作为信息的发布者和传播者,其本身的特征将与信息密切相关^[4]。大多数情况下,用户发布或转发的内容都与自己的兴趣有关,而用户的兴趣、关注范围通常是有限的。在检索信息时,首先定位相关的微博用户,再基于这些用户搜索微博信息,搜索范围显著降低。同时基于微博用户的兴趣特征及兴趣相关性,用气泡图可视化出特征词和微博用户的搜索相关性,帮助用户更全面的明确搜索需求,做进一步查询,返回更满意的查询结果。

提出了一种新的搜索微博信息的方法,由微博用户驱动信息查找过程,基于微博用户特征计算特征词的搜索相关性,可视化扩展用户查询;提出了计算特征词和微博用户相关性的关注度传递算法,并对用户模型进行了改进,更加适用于描述微博搜索

中用户的兴趣特征。

1 相关工作

微博的可视化搜索的研究主要体现在两个方面:用户输入的交互界面和结果输出的可视化。

交互在可视化搜索过程中允许用户动态改变意图和兴趣,设计良好的交互方式可以避免复杂的输入。可视化信息检索工具通过关联可视化结果来浏览数据。信息查询人员无需使用特定查询语言编写冗长而易错的查询语句,就能够快速地浏览数据。HomeFinder^[5]是一个图形化的用户搜索系统,允许查询用户修改变量范围,通过鼠标拉动滑动条的方式过滤数据项。同时动态查询提供了可视化界面元素与底层数据库更紧密的耦合操作方式。当用户移动滑动条,界面立即更新相应的数据项。作为动态查询的扩展,属性浏览器^[6]提供了与交互元素一致的可视化变量,数据在每个维度上的分布都表示为图形元素,因此用户每改变一次查询就能直接看到结果是否满意。分面搜索(faceted search)是另一种交互技术。分面是指事物的多维度属性。例如一本书包含主题、作者、年代等分面。而分面搜索是指通过事物的这些属性不断筛选、过滤搜索结果的方法。可以将分面搜索看成搜索和浏览的结合。Abel^[7]将分面搜索应用于微博,从博文中抽取多个维度属性以及相关外部链接来增强博文的语义,让用户查看感兴趣的博文。FacetMap^[8]使用空间填充的方式将多属性可视化为平行排列的圆形或矩形区域,通过过滤属性可进一步浏览数据细节。

统计图可以使复杂的统计数字简单化、通俗化、形象化,使人一目了然,便于理解和比较。微软开发的 Tweet Archivist(<http://www.tweetarchivist.com/>)实时监测 Twitter 用户动态(最活跃的用户、出现最多的单词、出现最多的链接等)并以统计图表示。Twitonom(<http://www.twitonomy.com/>)具备更强大的分析功能,能够轻松发现单向关注的用户群。相似性布局是文本可视化的基本方法,在文本搜索中被广泛采用。Lvis^[9]、NIRVE(NIST information retrieval visualization engine)^[10]和 SPIRE(spatial paradigm for information retrieval and exploration)^[11]按主题在 3 维系统中进行可视化。WebSOM^[12]根据

主题相似性按照自组织图(SOM)布局,xFind^[13]将搜索结果表示为聚类的散点图。层次表示适用于结构清晰的文本。ResultMaps^[14]使用已有的层次元数据将搜索结果表示为树图结构,作为排名列表的补充。Hyperbolic Browser^[15]在2维双曲空间中表示层次树,节约了空间。InfoSky^[16]根据搜索结果的相似性进行分类,用Voronoi图对空间进行划分。

然而,现有的微博可视化搜索方面的工作,没有考虑微博源用户作为微博的发布者和传播者的主导作用,只单纯从微博本身进行搜索与分析,没有关注微博用户的兴趣相关性。为克服这种缺陷,提出了新的用户驱动的微博可视搜索方法,以微博源用户驱动微博搜索过程。

2 本文方法

首先对微博源数据进行预处理,抽取出微博用户行为特征和兴趣特征,建立微博用户模型库。在给定检索词条件下,基于用户模型库构建用户兴趣特征网络(特征词-用户网络),再基于检索词执行关注度传递算法,计算兴趣特征词、微博用户的搜索相关性。基于兴趣特征词与检索词间的相关性进行扩展查询,基于微博用户与检索词间的相关性缩小微博搜索空间,提高搜索效率,并且用气泡图(bubble cloud)可视化出特征词、用户的搜索相关性,提供交互式的微博搜索可视化接口,其流程图如图1所示。

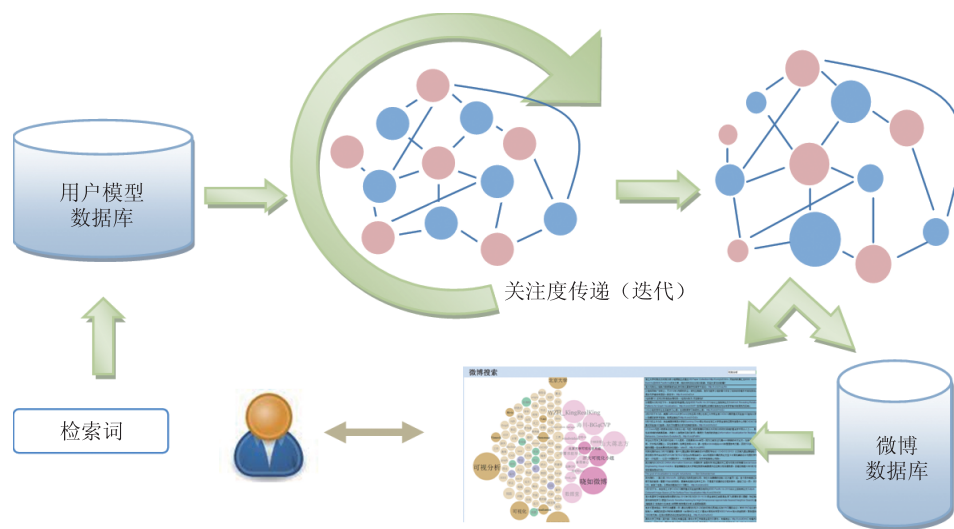


图1 流程图

Fig. 1 Flow chart

2.1 用户建模

用户建模分为用户兴趣特征建模和用户行为特征建模两部分。

2.1.1 用户兴趣特征建模

新浪微博作为一个信息分享平台,用户可以通过网页、外部程序,手机短信、彩信等发布140汉字(280字符)以内的信息,并可包含图片、视频链接、网页链接的资源,来分享他们的感兴趣的信息。本文将从微博文本内容中提取出用户的兴趣特征。首先要对微博条目进行过滤处理,删除掉一些与用户兴趣无关的内容,如图片、视频链接,表情,以及微博常用词语、符号,如“转发微博”,“//@ 用户名”,“#”等。

借助汉语词法分析系统 ICTCLAS (<http://ictclas.nlpir.org/>)以每个用户为单位对其微博文本进行分词和POS(Part of Speech)词性标注。POS标注主要包含名词、时间词、处所词、动词、形容词、外文单词等,其中名词(n)包含以下几个子类:人名(nr)、地名(ns)、机构团体名(nt)等。由于本文关注的是用户的兴趣特征,所以借助POS标注,过滤掉除名词和外文词外的其他词,并计算用户微博词频。这里使用向量空间模型(VSM)来描述用户微博,表示为 $D_k = (\langle t_1, tf_1 \rangle, \langle t_2, tf_2 \rangle, \dots, \langle t_n, tf_n \rangle)_k$,其中 t_i 表示第 i 个词, tf_i 表示词频, n 为词的数目。

期望用户向量能够最大程度地表征用户的兴趣

特征,显然词频并不是一个好的权重度量。Liu^[17]提出了一个描述微博用户兴趣特征的模型,其中引入一个词频权重系数,来计算词的权重。字词 t_i 的平均词频定义为

$$AF(t_i) = \sum_{k=1}^N (tf_i)_k / \sum_{k=1}^N in(t_i)_k \quad (1)$$

对于用户 k , 如果 $t_i \in D_k$, 则 $in(t_i)_k = 1$, 否则 $in(t_i)_k = 0$, N 为用户数。词频的权重系数定义为 $\Phi(t_i) = 1/(1 + e^{AF - tf_i})$ 。字词 t_i 的权重定义为 $w(t_i) = TF(t_i) \times \Phi(t_i)$, 其中, $TF(t_i)$ 为 t_i 的词频, $\Phi(t_i)$ 为词频权重系数。此模型虽然在分析个体兴趣特征时比较有效,但在分析用户兴趣相关性时,效果并不十分理想。比如,微博中的一些常用词如“微博”、“中国”、“粉丝”等与大部分词表现出强相关性。将 IDF (inverse document frequency)^[18] 引入上述模型,来避免常用词汇权重过大问题。同时由于微博自由开放的特点,用户可以发表任意形式的内容,可以无语法、无结构,甚至可以是任意字符的组合,其中可能不乏一些出现频率很高的“独有词汇”、“自创词语”,这些词汇对于搜索用户往往没有意义,成为“噪声”。为了避免“噪声”词因在文档集中出现极少而导致 IDF 值过大的问题,在 IDF 项中引入一个调节参数项 K 。IDF 项定义为

$$IDF(t_i) = \ln\left(\frac{N}{\min(N, n + K)}\right) \quad (2)$$

n 为包含词 t_i 的用户数, N 为用户总数。词 t_i 的权重定义为

$$w = TF(t_i) \times \Phi(t_i) \times IDF(t_i) \quad (3)$$

于是用户的兴趣特征表示为特征词向量

$$D_k = (\langle t_1, w_1 \rangle, \langle t_2, w_2 \rangle, \dots, \langle t_n, w_n \rangle)_k$$

2.1.2 行为特征建模

微博用户作为信息的来源,他们本身的行为特征将影响着信息的原创性、可靠性、准确性。这些特征包括用户的知名度、活跃度以及发布微博的原创性和习惯特性等。

用户知名度 知名度表现为微博使用者在社交平台上的影响力。粉丝数是用户知名度最为直接的表现。一般情况下,粉丝数越多,表明收听他们微博的用户也就越多。但也正是这个原因,导致了“僵尸粉”的出现。所谓僵尸粉,就是微博上的虚假粉丝,他们都是一些不活动的用户。一些微博用户为了提高知名度,会自己创建一些空账户并关注自己,

更甚至购买大量僵尸粉。所以,粉丝数并不能完全体现用户知名度的高低。微博的转发率,评论率体现了微博本身的受欢迎程度,是用户知名度的一个重要衡量标准。用户的知名度定义为

$$Rep = (1 - \theta) \times \frac{n_{\text{follower}}}{n_{\text{follower}} + n_{\text{friend}}} + \theta \times (1 - e^{-Ave_r + Ave_c}) \quad (4)$$

式中, θ 为控制系数, $0 < \theta < 1$, n_{follower} 为粉丝数, n_{friend} 为关注数。微博转发率 Ave_r 定义为

$$Ave_r = \frac{\sum_{i=0}^{n_{\text{status}}} n_{\text{repost}}(i)}{n_{\text{status}}} \quad (5)$$

n_{status} 为微博数, n_{repost} 为微博的转发数。微博评论率 Ave_c 可类似定义。

账户活跃度 活跃度 (Act) 体现为用户使用微博的频繁程度,主要体现为发布或转发微博的频率。发微博的频率这里以日微博发布数来计算。另一方面,用户发布微博来源的多样化,也是体现用户微博活跃度的特征。

微博原创性 原创性 (Org) 可通过转发率来衡量。在用户发布的微博中,转发微博的比例越小,账户的原创性就越高。

账户规律性 账户规律性指用户发布微博时间分布上的规律性,这里用相邻两条微博发布的时间间隔的复杂程度来衡量用户发布微博的行为的规律性^[19]。熵率 (ER) 是过程复杂性的度量标准^[20]。熵率越低,表明过程的规律性越强,而高的熵率值表明了随机过程,相对中等的熵率对于一个相对复杂的过程。

微博用户的行为特征表示为 $U_{\text{beh}}(k) = \{Rep_k, Act_k, Org_k, ER_k\}$ 。

2.2 基于关注度传递的相关性计算方法

基于用户兴趣模型,构造一个用户-特征词关系网络,如图 2 所示。词集 $T = \{t_i\}$ 与用户集 $U = \{u_j\}$ 构成一个二部图 (bipartite graph) $G = \{V, E\}$ 。结点集为 $V = T \cup U$, 边集 E 表示用户与特征词间的关联关系,其中:边 $\langle t_i, u_j \rangle$ 属于 E , 即 $\langle t_i, u_j \rangle \in E$ 当且仅当用户 u_j 的兴趣特性向量中包含特征词 t_i , 边的权重为用户 u_j 的兴趣特征向量中对应于 t_i 的权值。

对于任何特征词结点,建立用户兴趣特征的反向索引^[21] (inverted index), 词结点表示为 $T_k =$

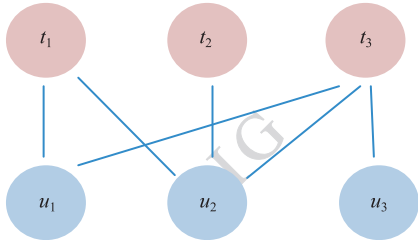


图2 用户-特征词关系图

Fig. 2 User- feature word relationship

$(\langle u_1, w_1 \rangle, \langle u_2, w_2 \rangle, \dots, \langle u_{n_k}, w_{n_k} \rangle)_k$, n_k 为与 T_k 关联的用户数。基于检索词的用户可以表示为 $U(k) = \{w_{q,k}, Beh_k, D_k\}$ 。

给定检索词 q , 可以从用户兴趣特征网络中提

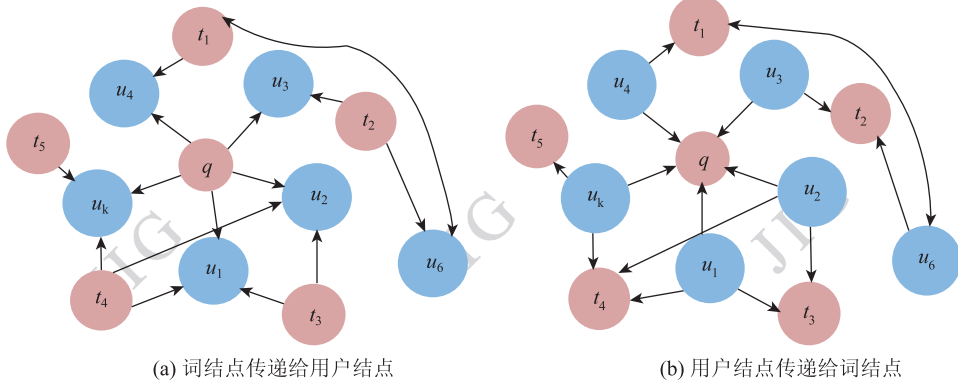


图3 关注度迭代过程

Fig. 3 Iteration process of attention degree

词结点 t_i 到用户结点 u_j 的关注度传递过程计算公式为

$$Att(u_j) = \frac{w_{t_i}(u_j)}{\sum_{k=1}^{N_u} w_{t_i}(u_k)} \cdot Att(t_i) \quad (6)$$

N_u 为与词结点 t_i 关联的用户数, $w_{t_i}(u_j)$ 为词结点 t_i 到用户结点 u_j 的关注度传递系数, 同时考虑用户的兴趣特征权重和用户行为特征, 定义为 $w_{t_i}(u_j) = \theta \cdot w_{t_i, u_j} + (1 - \theta) \cdot w_{beh}(u_j)$, 其中 w_{t_i, u_j} 为用户 u_j 兴趣特征向量中对应于特征词 t_i 的权重, $w_{beh}(u_j)$ 为用户 u_j 的行为特征值, θ 为调节参数。 θ 取值范围为 $[0, 1]$, 搜索者可以依据自己的搜索需求进行调整。如果搜索者偏向于了解信息内容本身, 可以设置较大的 θ 值; 如果偏向于了解相关微博用户的信息, 可以设置较小的 θ 值。用户结点 u_j 到词结点 t_i 传递关注度的计算公式可类似定义。

每次迭代结点的关注度累计变化差值定义为

取出搜索关注子网, 关注子网以检索词 q 为中心。如果搜索者关注特征词 q , 将有可能关注与 q 连接的微博用户 u , 同时如果用户关注 u , 那么也会关注与 u 相关的特征词, 以此类推, 搜索关注度将会在关注子网传递。下面, 将在关注子网执行关注度传递算法, 计算用户对特征词、微博用户的关注度。

初始状态下, 给定检索词 q , 其关注度为 $Att(q) = 1.0$, 其他特征词与用户的被关注度初始值为 0。每一次迭代分为两个过程, 如图 3 所示, 关注度由词结点传递给用户结点, 再由用户结点传递给词结点。如此不断迭代, 直至用户和词结点的关注度值趋于稳定状态。

$$\Delta Att = \sum_{t \in |T_i|} \sum_{k=1}^{N_t} |Att_{new}(t_k) - Att(t_k)| + \sum_{u \in |U_j|} \sum_{k=1}^{N_u} |Att_{new}(u_k) - Att(u_k)| \quad (7)$$

当 $\Delta Att < \varphi$ (φ 为接近 0 的常数) 时停止迭代。迭代过程如下:

- 1) 初始化检索词 $\{Att(q) = 1.0\}$, $\Delta Att = 0$;
- 2) 更新用户结点关注度, 并累积变化差值 ΔAtt ;
- 3) 更新词结点关注度, 并累积变化差值 ΔAtt ;
- 4) 如果 $\Delta Att < \varphi$, 转步骤 5); 否则转步骤 2);
- 5) 结束。

实验结果表明, 关注度的收敛速度与数据规模无关。从图 4 看出, 关注度的变化差值成递减趋势, 若干次迭代后趋向于 0。特征词和用户的关注度经过若干次传递, 最终趋于稳定状态。此时, 特征词和用户的关注度值从一定程度上反映了搜索者对微博用户及特征词的感兴趣程度。

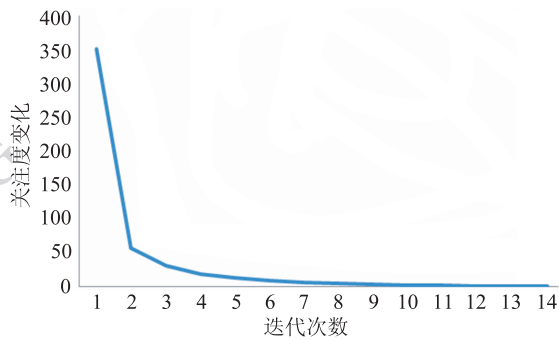


图4 关注度传递算法收敛速度

Fig. 4 Convergence rate of attention diverting algorithm

2.3 可视化搜索

在输入检索词后,从用户模型库中提取出特征词-用户子网,执行关注度传递算法,得到带有权重的相关特征词和用户列表。特征词的权重越大,表明其与检索词的相关性越大。用户的权重越大,表明该用户成为搜索微博来源的可能性越大。基于特征词与用户的权重计算结果提供一个可视化的交互界面。可以选取若干自己感兴趣的特征词或微博用户,将依据用户的反馈返回查询结果。

本文可视化界面如图5所示。左边部分采用气泡图对检索词相关的微博用户和特征词进行可视化。右边部分对相关的微博内容进行展示。

1) 气泡图。气泡图是一种基于力引导力的布局,其中的每个结点受向中心聚集的力和排斥力两种力。聚集力使得结点向中心聚集,从而使得结点在页面中排放紧凑,在有限空间内能够展示更多的信息;排斥力使得结点之间互相排斥,保证了结点之间互不遮挡。

本文的气泡图结点包括特征词与用户两种类型。气泡的大小编码结点相关度,气泡越大,其对应特征词或用户的搜索相关度就越大。对于特征词气泡,其颜色编码了特征词的不同类别,主要来自于POS标注结果,包括人名(nr)、地名(ns)、专用名(nz)及其他名称(n)。本文只显示相关度不低于一定阈值的特征词和用户。

采用辅助高亮的方式表示特征词与微博用户之间的关系:当搜索用户鼠标停留在某个特征词上时,与其关联的微博用户会高亮显示,反之亦然。

2) 微博列表。提供一定的交互供用户搜索相关微博内容。用户可以直接选择某个或若干个特征词气泡,检索与这些特征词相关微博,或者通过选择某些用户来搜索这些用户的微博信息。

依据用户的交互和检索词提取数据库中的微博内容,然后按照下面的准则对微博条目进行评分,并按评分结果从上至下排列微博条目,显示在右边的微博列表中:

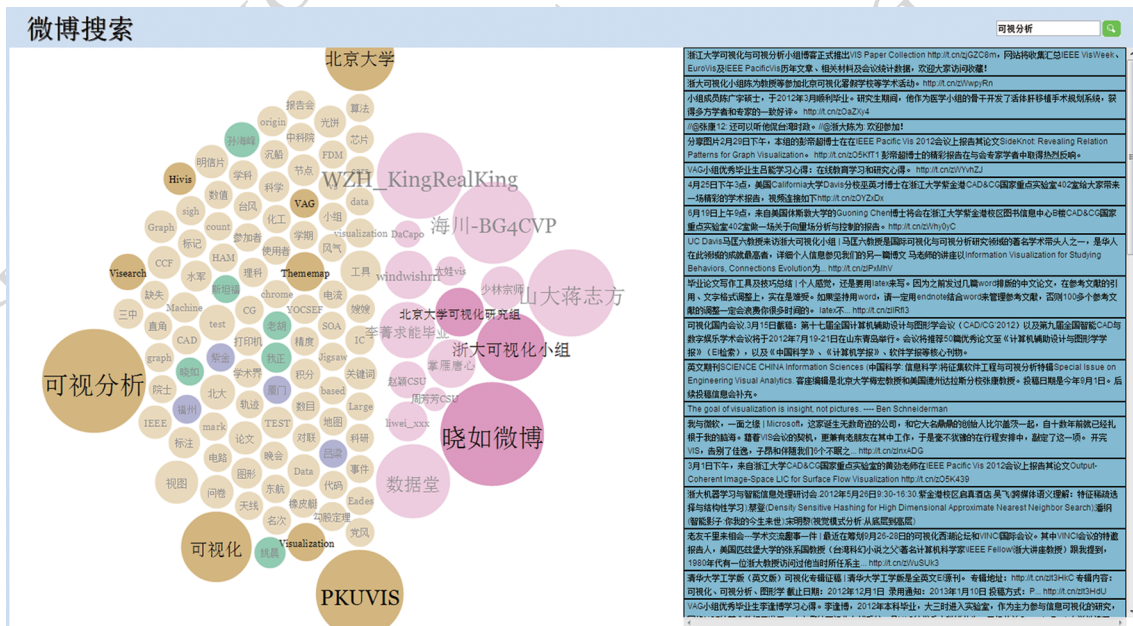


图5 可视化界面

Fig. 5 Visualization interface

(1)匹配度。指微博与反馈结果之间的匹配程度。

(2)文本长度。微博文本的长度从一定程度上决定了信息量的大小。一些文本长度短的微博将被直接过滤掉。

(3)发布时间。微博发布的时间决定了信息的时效性。特别是对于微博这样一个实时的信息分享平台,微博的发布时间在信息的价值评判上起着十分重要的作用。

(4)受欢迎程度。微博条目的受欢迎程度以转发此条微博的相关微博用户(搜索子网中的用户)数来衡量。

(5)用户评分。用户的评分指发布此条微博的用户的评分,主要包括用户的知名度、账户活跃度、微博原创性等等。

2.4 实验评估

本文使用的数据来自新浪微博,是通过新浪微博开放平台(<http://open.weibo.com/>)提供的API自动抓取的,抓取了2013年6月之前近10万名用户的最近五千条微博数据及个人信息。表1为用户的属性信息,主要包括用户ID、用户名、性别、所在城市、认证类型、注册时间等。

表1 用户属性

Table 1 User attributes

属性	备注
uid	用户id
screen_name	用户名
gender	性别(男、女)
location	位置信息(省、市)
verified	认证类型(未认证:-1)
created_at	账户注册时间
nBifriends	朋友数(互粉)
nFriends	关注数
nFollowers	粉丝数
nStatus	微博数
description	个人描述

表2为微博的属性,包括微博ID、发布时间、是否转发(转发微博ID、用户ID)、微博来源、微博内容等。

表2 微博属性

Table 2 Microblog attributes

属性	备注
TID	微博id
UID	用户名
create_at	发布时间
RID	转发微博(原创:-1)
RUID	转发用户ID(原创:-1)

通过对这两个数据集的预处理,建立了用户的行为模型和兴趣模型。共提取了约20万的特征词来描述用户的兴趣特征,这样,用户的兴趣模型就是一个约30万结点的特征词-用户网络。

对于每一次搜索任务,首先从用户兴趣模型中提取出以检索词为中心的特征词-用户子网,执行关注度传递算法,计算特征词、用户的搜索相关性。同时以气泡图可视化特征词与用户的相关性,搜索者可以交互的搜索自己感兴趣的微博。

将从时间效率和用户体验两个方面来评估本文的模型:

1)时间效率。在预处理部分已对用户的行为特征和兴趣特征进行计算,并将结果保存于MongoDb中。搜索时间主要消耗在执行关注度传递算法和微博的查询和排序上。

实验结果表明,关注度传递算法基本可以在常数迭代后收敛,并且由于在预处理部分已对特征词和用户之间的关系建立了正反向索引,每次迭代查询数据库的时间非常快,故时间可忽略不计。

查询相关微博时,首先基于特征词和用户之间的关系,定位相关度大于一定阈值的微博用户,将这些用户的微博读入内存中,再从中选取与特征词和检索词相关的微博,最后进行评分和排序,返回评分大于一定阈值的微博内容进行可视化展示。实验表明,相关微博用户的数目不会超过一个常数,故从数据库中提取微博的时间可忽略不计;对微博内容的筛选在内存中进行,其时间亦可忽略不计,最终筛选出的微博数目也不会超过一个常数,排序时间亦可忽略。实验证明,每次的搜索和交互响应时间不会超过1s,在用户可接受的范围之内。

2)实例分析。搜索“可视分析”,从图5可以看到,出现了“可视化”、“北京大学”、“PKUVIS”等比较明显的特征词,“晓如微博”、“浙大可视化小组”、

“北京大学可视化研究小组”等用户。浙大可视化小组和北大可视化小组是国内领先的两个做可视化、可视分析方向的团队。PKUVIS 是北大可视化小组的简称,晓如微博是北大可视化小组带头人袁晓如的微博账号。进一步探索发现,Visearch、Hivis、Thememap 为北大可视化小组开发的 3 个微博可视分析工具。

从微博列表中,可以了解到了“可视分析”这个方向其他学术领头人的信息,如美国 UC Davis 的马匡六教授等。同时发现一些对可视化、可视分析这一领域的分析和评价信息,如“可视化的评价是可以量化的。比如对特定分析目标的正确率,成功率。当然这是和使用者有关,人在其中的因素也要考虑。//@ 山大蒋志方:可视化效果评价,是个很依赖于个人主观感觉的问题,确实是个问题,和认知科学,交互技术有关 //@ 毕毕毕熙然:有木有人关注可视化效果评价?”;“可视化不是只是为了更好地表达数据。好的可视化,是可以能够发掘数据内在

的规律,找到核心的内容。在大数据时,abstract/filtering 是关键”。

如搜索“杭州”、“美食”这两个特征词,结果如图 6 所示。从系统界面中观察到了“甜品”、“海鲜”、“火锅”、“包子”等一些与美食有关特征词,“好客美食-杭州”、“大众点评网”、“19 楼私房菜”等介绍美食的源用户。从一些微博用户的名称中,可以了解到杭州地区受欢迎的美食品牌,如“仙踪林”、“甘其食”包子、“外婆家”等。进一步地,通过基于用户和特征词的查找,获取到了“西塘河有台湾美食街”等信息。通过查询,了解到“好客美食网”是一个本地美食搜索和美食资讯分享的 SNS 网站,支持消费者手机移动点评模式,网站覆盖江苏、浙江、上海、北京、广州等全国 30 多个主要城市。“大众点评网”是中国领先的本地生活信息及交易平台,同时也是全球最早建立的独立第三方消费点评网站,时至今日分支机构遍布中国各大主要城市。

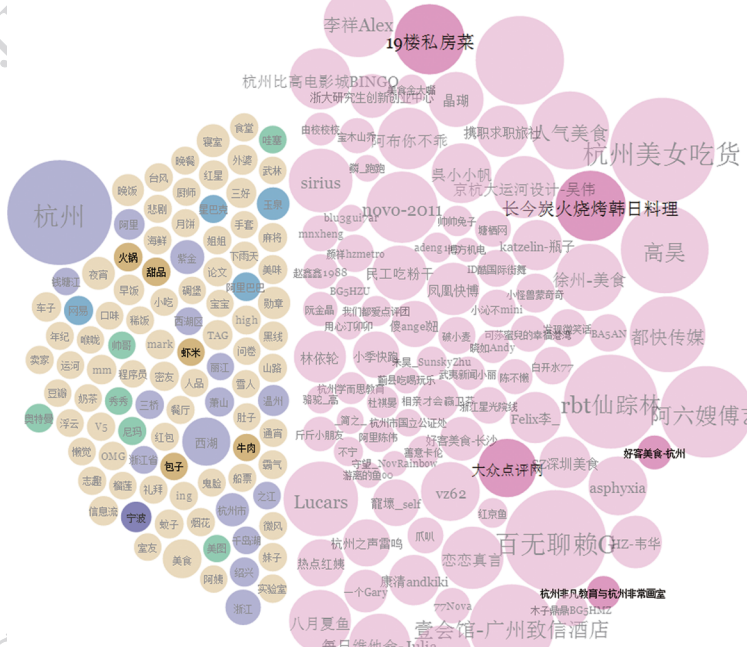


图 6 “杭州美食”气泡图

Fig. 6 Bubble chart with keywords “Hangzhou meishi”

3 结 论

提出了一种新的交互式微博可视化搜索方法,该方法将微博用户作为“桥梁”来查找信息,通过改进用户模型,使之更能适合于描述微博搜索中用户

的兴趣特征;通过在用户兴趣特征子网上进行关注度传递来计算特征词与用户的搜索相关度,提供给搜索者一个直观的可视化查询接口来进行探索式信息发掘。用户可以通过可视化结果更加全面地明确搜索需求。

实验结果表明,本文方法可以帮助搜索者能够

更迅速地定位信息,更好地理解 and 发掘所包含的信息,这在目前的大数据时代具有重要的应用前景。

参考文献 (References)

- [1] Naveed N, Gottron T, Kunegis J, et al. Searching microblogs: coping with sparsity and document quality [C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2011: 183-188. [DOI: 10.1145/2063576.2063607]
- [2] Teevan J, Ramage D, Morris M R. TwitterSearch: a comparison of microblog search and web search [C]//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2011: 35-44. [DOI: 10.1145/1935826.1935842]
- [3] Kules B, Capra R, Banta M, et al. What do exploratory searchers look at in a faceted search interface [C]//Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital libraries. New York: ACM Press, 2009: 313-322. [DOI: 10.1145/1555400.1555452]
- [4] Wang J, Zhu K, Wang B Q. Survey on microblog research based on information data analysis [J]. Computer Application, 2012, 32(7): 2027-2029. [王晶, 朱珂, 汪斌强. 基于信息数据分析的微博研究综述 [J]. 计算机应用, 2012, 32(7): 2027-2029.] [DOI: 10.3724/SP.J.1087.2012.02027]
- [5] Ahlberg C, Shneiderman B. Visual information seeking: tight coupling of dynamic query filters with starfield displays [C]//Proceedings of the SIGCHI Conference on Human factors in Computing Systems. New York: ACM Press, 1994: 313-317. [DOI: 10.1145/191666.191775]
- [6] Spence R, Tweedie L. The attribute explorer: information synthesis via exploration [J]. Interacting with Computers, 1998, 11(2): 137-146. [DOI: 10.1016/S0953-5438(98)00022-8]
- [7] Abel F, Celik I, Houben G J, et al. Leveraging the semantics of tweets for adaptive faceted search on twitter [M]//The Semantic Web-ISWC. Berlin Heidelberg: Springer, 2011: 1-17. [DOI: 10.1007/978-3-642-25073-6_1]
- [8] Smith G, Czerwinski M, Meyers B R, et al. FacetMap: a scalable search and browse visualization [J]. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(5): 797-804. [DOI: 10.1109/TVCG.2006.142]
- [9] Börner K, Dillon A, Dolinsky M. LVis-digital library visualizer [C]//Proceedings of IEEE International Conference on Information Visualization '2000, Piscataway: IEEE Press, 2000: 77-81.
- [10] Sebrecths M M, Cugini J V, Laskowski S J, et al. Visualization of search results: a comparative evaluation of text, 2D, and 3D interfaces [C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1999: 3-10. [DOI: 10.1145/312624.312634]
- [11] Wise J A, Thomas J J, Pennoek K, et al. Visualizing the non-visual: spatial analysis and interaction with information from text documents [C]//Proceedings Information Visualization. Piscataway: IEEE Press, 1995: 51-58. [DOI: 10.1109/INFVIS.1995.528686]
- [12] Kaski S, Honkela T, Lagus K, et al. WEBSOM-self-organizing maps of document collections [J]. Neurocomputing, 1998, 21(1): 101-117. [DOI: 10.1016/S0925-2312(98)00039-3]
- [13] Andrews K, Gutl C, Moser J, et al. Search result visualisation with xfind [C]//Proceedings. Second International Workshop on User Interfaces to Data Intensive Systems. Piscataway: IEEE Press, 2001: 50-58. [DOI: 10.1109/UIDIS.2001.929925]
- [14] Clarkson E, Desai K, Foley J D. Resultmaps: visualization for search interfaces [J]. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 1057-1064. [DOI: 10.1109/TVCG.2009.176]
- [15] Lamping J, Rao R, Pirolli P. A focus + context technique based on hyperbolic geometry for visualizing large hierarchies [C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press/Addison-Wesley Publishing Co., 1995: 401-408. [DOI: 10.1145/223904.223956]
- [16] Andrews K, Kienreich W, Sabol V, et al. The infosky visual explorer: exploiting hierarchical structure and document similarities [J]. Information Visualization, 2002, 1(3-4): 166-181. [DOI: 10.1057/palgrave.ivs.9500023]
- [17] Liu D, Wu Q, Han W. Mining micro-blogging users' interest features via fingerprint generation [C]//Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering. Paris: Atlantis Press, 2013: 1469-1472. [DOI: 10.2991/icsee.2013.370]
- [18] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval [J]. Information Processing & Management, 1988, 24(5): 513-523. [DOI: 10.1016/0306-4573(88)90021-0]
- [19] Chu Z, Gianvecchio S, Wang H, et al. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? [J]. IEEE Transactions on Dependable and Secure Computing, 2012, 9(6): 811-824. [DOI: 10.1109/TDSC.2012.75]
- [20] Cover T M, Thomas J A. Elements of Information Theory [M]. State of New Jersey: John Wiley & Sons, 2012. [DOI: 10.1002/047174882X]
- [21] Catena M, Macdonald C, Ounis I. On Inverted Index Compression for Search Engine Efficiency [M]//Advances in Information Retrieval. Berlin: Springer, Switzerland, 2014: 359-371. [DOI: 10.1007/978-3-319-06028-6_30]