第 42 卷 第 6A 期
2015 年 6 月

计 算 机 科 学
Computer Science

Vol. 42 No. 6A
June 2015

基于话题相关空间的微博用户兴趣识别及可视化方法

赵 华 纪晓文 曾庆田 郝春燕

(山东科技大学信息科学与工程学院 青岛 266590)

摘 要 微博已经成为获取用户兴趣的有效平台。在分析了用户发表微博的习惯及特点的基础上,提出了一种基于话题相关空间自动构建,同时融合位置信息的微博用户兴趣识别方法。该方法首先基于话题检测技术构建话题相关空间,提出了基于空间范围的 TFIDF 计算方法,然后融合位置信息计算微博词汇的兴趣表征值,最后采用 3D 标签云对兴趣识别结果进行了可视化。实验结果表明了所提方法的有效性。

关键词 微博,用户兴趣,话题相关空间,可视化

中图法分类号 TP391 文献标识码 A

Topic Related Space-based Microblog User Interests Mining and Visualization Method

ZHAO Hua JI Xiao-wen ZENG Qing-tian HAO Chun-yan

(College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

Abstract Microblog is an effective platform to get the user interests. Based on the analysis of the habit and the characteristic of publishing microblog, this paper proposed a topic related space-based microblog user interests mining method, with the help of the location feature. This method firstly creates the topic related space based on the topic detection, secondly calculates the interest index value of each word based on the combination of TFIDF and location feature, and finally visualizes the mining results based on the 3D tag clouds. The experimental results show that the proposed method is useful.

Keywords Microblog, User interests, Topic related space, Visualization

1 引言

作为 Web2.0 的典型代表,微博是一个基于用户关系的信息分享、传播以及获取平台,近来得到了快速的发展和广泛的应用。目前国外比较流行的微博平台是 Twitter,而在国内,新浪微博、腾讯微博等平台受到用户的喜爱。Twitter 最初的设计理念是可以"让用户更加简洁地与好友分享自己正在做什么",该理念是导致微博数据实时性的重要原因之一,正因为如此,微博平台成为目前获取用户兴趣的最新途径。有效地从用户发表的微博数据中挖掘用户的兴趣,对于提高个性化推荐的质量、提高用户体验具有重要意义。

用户兴趣挖掘近年来受到了研究者们的广泛关注,是个性化推荐研究中的基础性、关键性任务。为此,研究者们已经提出了一些较有效的挖掘方法,但由于微博是新近兴起的平台,同时具有鲜明的特点,因此传统的用户兴趣方法不能很好地适用于微博用户的兴趣发现。本文通过分析用户发表微博的特点,提出了一种基于话题相关空间自动构建的微博用户兴趣识别方法。该方法首先基于话题检测技术构建话题相关空间,然后在相关空间中基于 TFIDF 及位置信息计算微博词汇的兴趣表征值,最后通过 3D 标签云实现抽取结果可视化。

2 相关工作

微博的推广及应用吸引了大量的科研人员从各个方面对微博的各类资源进行分析研究,如文本资源、社交网络资源等,他们试图从中挖掘出有用的信息。基于微博挖掘用户的兴趣便是其中比较受关注的研究之一。

Wei Wu[1] 首次提出通过从 Twitter 中抽取关键字来表示用户的兴趣,并采用 TFIDF 和 TextRank 来抽取关键词,实验结果表明两种方法结合起来抽取微博关键词的准确率与从网页中抽取关键词的准确率相当。陈文涛等人比较了多种用于构建微博用户兴趣模型的主题模型,发现不同的主题模型在不同方面的性能是不同的[2]。Zhiyuan Liu[3] 等人基于机器翻译和词频从中文微博中抽取关键词,用以表示用户的兴趣。Yuto[4] 利用 Twitter 中特有的 Twitter 列表,从 Twitter 列表名称中提取标签,使用用户和标签的点互信息公式计算用户和标签之间的关系,最后通过分值为用户分配合适的标签。Ma 等[5] 提出可以从多个数据源如 Twitter、FaceBook、Linkedn 等挖掘用户兴趣,首先从每个数据源获取用户的兴趣,用户的兴趣用若干个关键词表示,然后通过信息融合策略,产生一个相对完整的用户兴趣集。Chen[6] 分别利用用户

本文受国家自然科学基金(61170079,61202152),山东省自然科学基金(ZR2013FQ030),教育部网络时代科技论文快速共享课题(2013122),山东省优秀中青年科学家科研奖励基金(BS2012DX030),山东省高等学校科技计划项目(J12LN45),中国煤炭工业协会 2013 年度科学技术研究指导性计划项目(MTKJ2013-366),山东科技大学 2014—2015 年度研究生科技创新基金项目(YC140324,YC140326)资助。

赵 华(1980—),女,博士,副教授,CCF 会员,主要研究方向为自然语言处理;曾庆田(1976—),男,博士,教授,博士生导师,主要研究方向为个性化推荐、工作流等。

本身微博和用户的粉丝微博进行了用户兴趣发现的实验，结果显示基于用户本身微博得到用户兴趣的效果更好。Weng[7]收集用户所发的微博并将其整合成一个大文件，然后利用 LDA 模型发现用户潜在主题的兴趣。宋巍等人则提出了基于微博分类的用户兴趣识别[8]。Jilin Chen[9]分别采用用户自己发表的微博内容和用户粉丝发表的微博内容构建词袋来构建用户兴趣，实验结果表明采用用户自己发表的微博内容来构建词袋效果更好。

3 微博数据获取及预处理方法

3.1 基于新浪 API 接口的微博数据获取方法

新浪微博开发平台模仿 Twitter，也推出了自己的 API 接口。大部分 API 的访问如发表微博、获取私信、关注都需要用户身份，目前新浪微博开放平台的用户身份鉴权有 OAuth 2.0和 Basic Auth(仅用于应用所属开发者调试接口)。用户在申请应用获取授权之后可利用该接口实现很多功能，获取新浪微博上的数据。

本文在获取微博数据时选用新浪微博平台提供的开放 API，对微博进行数据采集工作。要使用微博的开放 API 进行大量的数据采集工作，首先要获得使用 API 的权限，获取 App Key 以及 App Secret，授权以后就可以获得需要的 access-token，提取用户发布的微博。

3.2 数据预处理方法

本文主要进行了两个数据预处理操作：分词与词性标注、数据清洗，下面分别进行介绍。

(1)分词与词性标注：由于本文处理的是中文微博，因此必须对其进行分词。实验中采用中科院分词工具 ICTCLAS2011(http://ictclas.org/ictclas-download.aspx)进行分词，同时采用计算所的二级标注级来标注词汇的词性。

(2)数据清洗(Data cleaning)：在微博中有很多的噪音数据，比如用户账号、表情符号、URL 等内容。为了准确抽取关键词，数据清洗部分首先将这些内容删除掉。

①用户账号：在新浪微博中，对于用户账号的引用有两种情况，对于不同的情况采取不同的操作。第一种情况是当前的微博 A 是对某一条微博 B 的回复，此时会用“//@用户账号：”引出 B，这种情况下，将“//@用户账号：”全部删除；第二种情况是用户在撰写微博时简单提及了某个用户，此时会用“@用户账号”来表示，这种情况下用户帐号其实是作为微博的内容的，所以此种情况下只是将@删除。

②表情符号：微博中通常含有较为丰富的表情符号，这些符号对于情感分析比较有用，但是对于关键词抽取意义不大，选择将其删除。下载的新浪微博中的表情符号通常用中括号(“[]”)相括，比如[阳光]、[疑问]等信息，本文根据这个特点制定规则来删除表情信息。

③URL：在微博中常常包含着一些 URL，本文采用正规表达式将其识别并删除。

(3)去除停用词。借助于停用词表，将停用词去掉。

4 微博用户兴趣识别方法

4.1 基于话题相关空间的词汇兴趣表征值计算方法

通过浏览微博用户发表的微博数据，我们发现：人们从事某些特定活动时，比如到某地旅游或者参加某次学术会议，通常会连续发表若干条与该活动相关的微博。本文将主题相关的微博集合称为微博话题相关空间。也就是说，每个微博用户都会有若干个微博话题相关空间，而每个微博话题相关空间都包含若干条微博数据。本文基于话题检测方法实现话题相关空间的自动构建[10]。

有了话题相关空间以后，本文提出基于话题相关空间来计算词汇的 TFIDF 值，计算公式如下：

$$RS\text{-}TFIDF(word_i)=TF(word_i)\times\log\left(\frac{|RS(W)|}{RSF(word_i)}+1\right) \tag{1}$$

$$RSF(word_i)=\sum_{W_k\in RS(W)}Appear(word_i,W_k) \tag{2}$$

其中，$word_i(1\leqslant i\leqslant|W|)$是微博 W 中的词汇，$RS(W)$表示 W 的话题相关空间，$TF(word_i)$是 $word_i$ 在 W 中的词频，$RSF(word_i)$是 $word_i$ 在 $RS(W)$ 中的文档频次，由式(2)计算得到。如果 $word_i$ 在 W_k 中出现，$Appear(word_i,W_k)=1$；否则 $Appear(word_i,W_k)=0$。

本文将不建立话题相关空间时进行的 TFIDF 方法称为传统方法。基于话题相关空间的 TFIDF 方法和传统方法都是依据 TF 和 IDF 值进行计算，但是 TF 和 IDF 统计的范围不一样。

4.2 融合位置特征的词汇兴趣表征值计算方法

微博自身的短文本特性决定了许多微博用户在发表微博时通常采用“开门见山”的叙述方式，即越靠近微博开始的词汇越是重要的词汇。位置特征已经被广泛应用于多种自然语言处理研究中，并被证明是一种较为有效的特征。为此，本文将词汇的位置信息也作为用户兴趣关键词的特征之一，并最终采用式(3)计算词汇的兴趣值：

$$Interest(word_i)=\frac{RS\text{-}TFIDF(word_i)}{LOC(word_i)} \tag{3}$$

其中，$Interest(word_i)$是本文中用以度量词汇兴趣值的，$LOC(word_i)$是 $word_i$ 在微博中的序号。式(3)表示，序号越小即越靠近微博开头的词汇越重要。

经过上述步骤，得到了词汇的兴趣表征值。然后，本文采用 TOP-N 的方法来抽取代表用户兴趣的关键词，即将词汇按照兴趣表征值从高到低的顺序进行排序，取前 N 个兴趣表征值较高的词汇作为用户的兴趣。

5 实验结果与分析

5.1 语料及评价指标

由于缺乏公开的标记语料，为了评测提出的方法的有效性，本文手动建立了语料库。建立过程为：首先随机选取 10 个微博用户，其次通过新浪微博提供的 API 函数(指定用户 ID)下载特定用户最新的 200 条微博(注：微博 API 的限制使得只能下载 200 条)。最后，根据该 200 条微博内容，手动标记代表该用户兴趣的关键词。

评测过程中，本文采用传统的正确率、召回率和 $F1$ 作为评价指标，各自的计算公式如下所示：

$$precision=\frac{c}{m} \tag{4}$$

$$recall=\frac{c}{n} \tag{5}$$

$$F1=\frac{2\times precision\times recall}{precision+recall} \tag{6}$$

$precision$、$recall$、$F1$ 别是抽取用户的兴趣关键词时的准确率、召回率和 $F1$ 值，其中 c 是系统正确抽取的兴趣关键词的个数，m 是系统返回的兴趣关键词的个数，n 是语料中人工标记的兴趣关键词的个数。

5.2 用户兴趣识别方法实验结果与分析

为了验证本文提出方法的有效性，进行以下 3 组实验。

实验 1：本组实验在未建立话题相关空间的基础上，采用传统的 TFIDF 方法计算词汇的兴趣表征值，也可以认为某用户的 200 条微博信息属于同一个话题相关空间。实验结果如表 1 所列（限于篇幅，本文中的实验结果只列出 $F1$ 值）。

表 1 实验 1 的实验结果

Users \ N	3	4	5	6	7	8	9
User1	0.4528	0.4541	0.4621	0.4425	0.4587	0.4952	0.4625
User2	0.4872	0.4756	0.4905	0.5021	0.5213	0.4879	0.4923
User3	0.5027	0.5110	0.5101	0.4989	0.5245	0.4785	0.5025
User4	0.4129	0.4214	0.4407	0.4369	0.4512	0.4476	0.4352
User5	0.3514	0.3321	0.3625	0.3712	0.4009	0.4112	0.4025
User6	0.5121	0.5201	0.5197	0.4987	0.5021	0.5078	0.5211
User7	0.3711	0.3803	0.3815	0.4021	0.4321	0.4206	0.4512
User8	0.4516	0.4418	0.4525	0.4478	0.4621	0.4425	0.4716
User9	0.6001	0.5978	0.5901	0.5849	0.5810	0.5608	0.5528
User10	0.4723	0.4802	0.4889	0.4658	0.4587	0.4952	0.4715

实验 2：本组实验在自动建立话题相关空间的基础上识别用户兴趣，即采用式(1)计算词汇的兴趣表征值。实验结果如表 2 所列。

表 2 实验 2 的实验结果

Users \ N	3	4	5	6	7	8	9
User1	0.4678	0.4589	0.4578	0.4570	0.4425	0.4895	0.4728
User2	0.4729	0.4978	0.5214	0.4989	0.5410	0.4978	0.5031
User3	0.5145	0.5058	0.5245	0.5312	0.5478	0.5012	0.5325
User4	0.4278	0.4412	0.4345	0.4578	0.4689	0.4876	0.4578
User5	0.3578	0.3412	0.3590	0.3711	0.39845	0.4178	0.4102
User6	0.5215	0.5325	0.5025	0.5046	0.5105	0.5325	0.5412
User7	0.3811	0.3789	0.3841	0.4109	0.4295	0.4308	0.4575
User8	0.4486	0.4512	0.4858	0.4627	0.4695	0.4620	0.4878
User9	0.6212	0.6032	0.6001	0.5978	0.5912	0.5412	0.5501
User10	0.4881	0.4936	0.4725	0.4585	0.4802	0.5052	0.4915

从实验 2 的实验结果可以看出，建立了话题相关空间以后，用户兴趣识别的效果得到了一定的提升，证明建立话题相关空间对于用户兴趣的识别是有一定的帮助的。

实验 3：该组实验在计算词汇兴趣表征值时融入了位置信息，即采用式(3)计算词汇的兴趣表征值。实验结果如表 3 所列。

表 3 实验 3 的实验结果

Users \ N	3	4	5	6	7	8	9
User1	0.4889	0.4785	0.4645	0.4787	0.4325	0.4945	0.4689
User2	0.4956	0.5012	0.5345	0.5023	0.5397	0.5206	0.5104
User3	0.5212	0.5256	0.5302	0.5412	0.5541	0.5120	0.5415
User4	0.4478	0.4523	0.4457	0.4545	0.4789	0.4980	0.4658
User5	0.3658	0.3745	0.3878	0.3698	0.4102	0.4258	0.4304
User6	0.5302	0.5102	0.5251	0.5321	0.5102	0.5452	0.5308
User7	0.4012	0.4102	0.3987	0.4201	0.4159	0.4258	0.4345
User8	0.4523	0.4412	0.4987	0.4564	0.4758	0.4859	0.4936
User9	0.6312	0.6032	0.6078	0.5898	0.6028	0.5365	0.5658
User10	0.4881	0.49878	0.4978	0.4720	0.4987	0.5325	0.5478

从实验 3 的几组实验可以看出，加入了位置特征以后，用户兴趣识别的效果得到了大幅度提升，再次证明位置特征在微博用户兴趣识别中也是一个非常有效的特征。

从上述几组实验结果可发现：不同用户的兴趣识别效果在不同的 N 值达到最好，原因在于不同用户个性不一样，有些兴趣广泛，对于这样的用户来讲 N 值比较大时识别效果更好，而有些用户在微博中只关注某一个特定方面，所以相对来说其兴趣识别效果在 N 取较小值时比较好。同时，为了更好地提取用户兴趣，应该再加入一些语义分析技术。

5.3 用户兴趣抽取结果可视化

实验中，采用 3D 标签云（3D Tag Cloud）[11] 来可视化抽取到的微博用户的兴趣，采用的标签云的具体版本号是 3.5.2。可视化结果如图 1 和图 2 所示。

图 1 用户兴趣抽取可视化结果 1

图 2 用户兴趣抽取可视化结果 2

结束语 本文对微博用户兴趣识别方法进行了研究，提出了基于话题相关空间自动构建，同时融合位置信息的微博用户兴趣识别方法，实验结果表明该方法能比较有效地识别微博用户兴趣。实验中我们发现不同类型的用户的兴趣分布情况差别比较大，下一步的工作将在对用户个性识别的基础上进行用户兴趣识别研究，同时考虑融入一些语义分析技术。在下一步工作中也将在更大规模语料上进行实验。

参 考 文 献

[1] Wu Wei, Zhang Bin, Ostendorf M. Automatic Generation of Personalized Annotation Tags for Twitter Users[C]// Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL. 2010: 689-692

[2] 陈文涛，张小明，李舟军. 构建微博用户兴趣模型的主题模型的分析[J]. 计算机科学，2013：40(4)：127-130

[3] Liu Zhi-yuan, Chen Xin-xiong, Sun Mao-song. Mining the interests of Chinese microbloggers via keyword extraction[J]. Frontiers of Computer Science in China, 2012, 6(1): 76-87

（下转第 509 页）

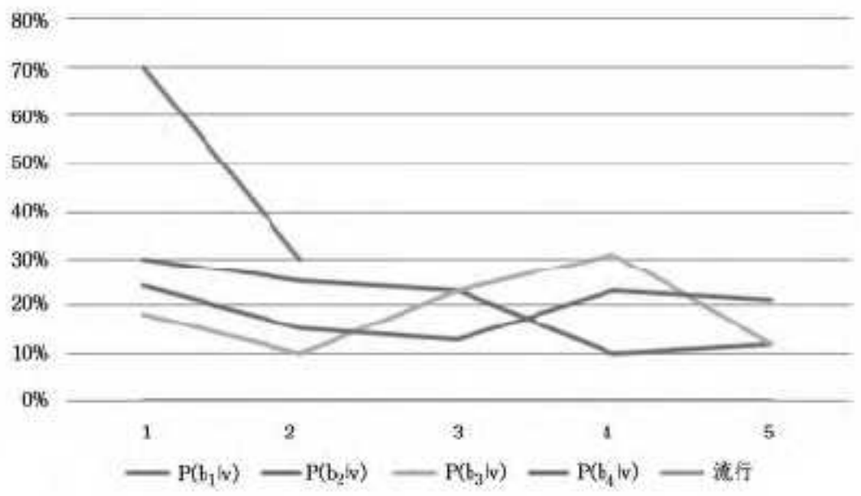


图 7　各类别偏好测试者中带有不同应用特征的概率分布

(6)实验结果

按以上方法处理数据，根据贝叶斯数据集计算修正系数(以测试者 1 为例)，数据的散点图分布如图 8 所示。

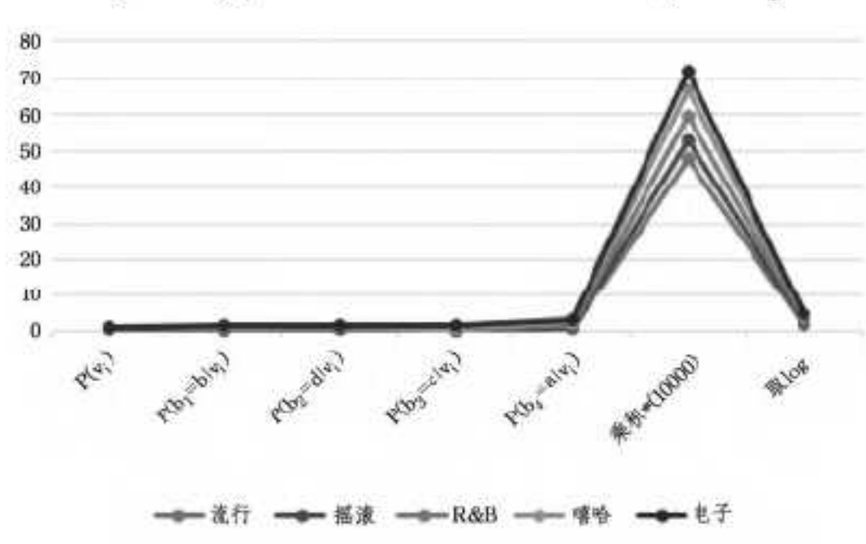


图 8　修正系数的散点图分布

根据协同过滤算法得原始推荐分数及音乐(以测试者 1 为例)，如图 9 所示。

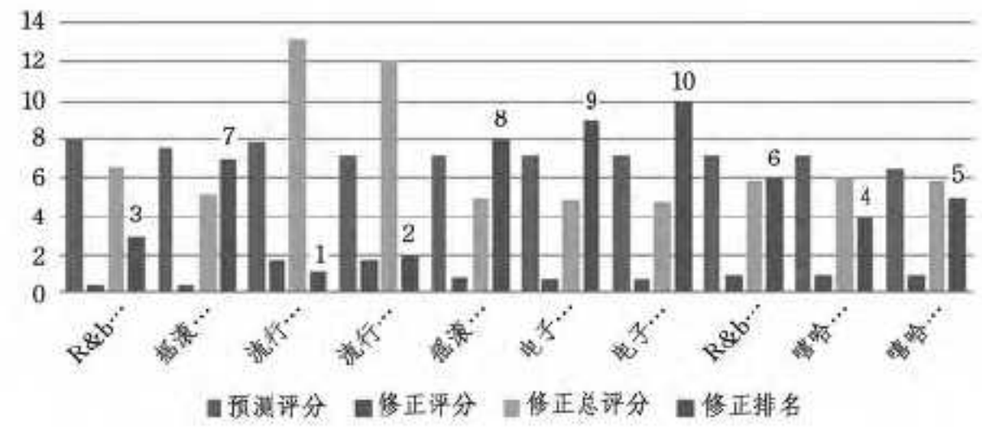
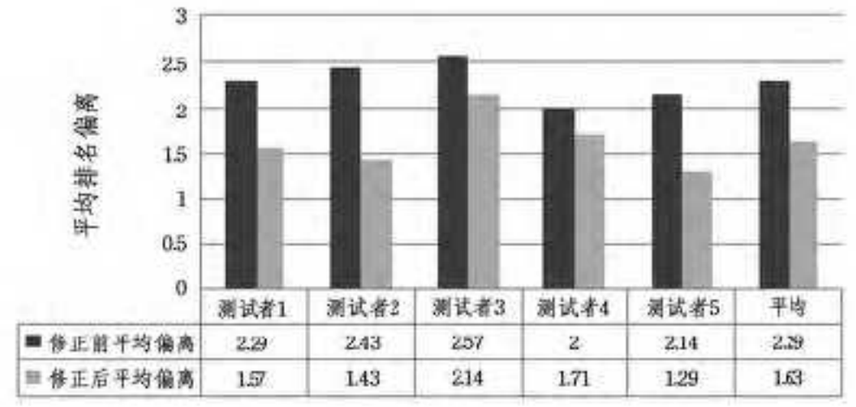


图 9　协同过滤推荐后的分布

(7)精确性对比

分别将两份名单中排名靠前的 7 首音乐推荐给测试者，测试者根据偏好情况，给出 1 到 7 的偏好排名。按平均排名偏离水平方法比较两种名单的排名准确度，5 位测试者的数据如图 10 所示，其中第六列为前 5 列的平均值。



	测试者1	测试者2	测试者3	测试者4	测试者5	平均
修正前平均偏离	2.29	2.43	2.57	2	2.14	2.29
修正后平均偏离	1.57	1.43	2.14	1.71	1.29	1.63

图 10　综合推荐排名与原推荐排名准确度的比较

(8)实验总结

通过实验可看出，修正后的排名提高了推荐音乐排名的精准性，说明在综合了实验者的移动应用情境后，排名更接近测试者的真实偏好。综合考虑移动应用特征后，对原推荐排名进行调整，将更符合用户的真实偏好，接近用户情境。

结束语　本文定位于根据移动端用户音乐行为特征，完善其在音乐市场得到的音乐推荐准确性。在传统的协同过滤算法中，两维用户_项目模型(U-I 模型)为普遍认可的算法模型。通过寻找相似邻居的项目评分，确定目标用户在 U-I 中的位置，进行应用推荐。在该模型中加入 *Situation*(情境)维度，在用户进入模型时，首先寻找其相似情境，确定 *Situation* 维度的值，再在该情境下寻找相似邻居。第三维情境模式，包括所处位置、网络状况、所处时间段、是否是某些重大节日。将这 4 种元素通过贝叶斯算法聚合为用户对某种应用的类别偏好，由此将 *Situation* 更加详细地描述为类别情境。添加类别情境的优势在于，更加精准地判断某个相似邻居是否为情境相同的相似邻居。

参考文献

[1] 莫同，李伟平. 一种情境感知服务系统框架[J]. 计算机学报，2010，33(11)：2084-2092

[2] 周玲元，段隆振. 移动情境感知服务系统研究[J]. 图书馆学研究，2014，36(8)：36-44

[3] 徐步刊，周兴社，等. 一种场景驱动的情境感知计算框架[J]. 计算机科学，2012，39(3)：. 216-222

[4] 张静. 基于情境感知的自适应个性化知识服务研究[J]. 情报科学，2011，29(11)：1658-1661

[5] 叶剑，李锦涛，等. 分布式情境感知分层模型设计与角色分析[J]. 电子学报，2012，8(8)：1572-1575

[6] 於志勇，周兴社，王海鹏，等. 动态上下文知识的获取与共享[J]. 计算机科学，2009，36(9)：218-223

(上接第 502 页)

[4] Yamaguchi Y, Amagasa T, Kitagawa H. Tag-based User Topic Discovery Using Twitter Lists[C]// Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining. 2011：13-20

[5] Ma Y, Zeng Y, Ren X, et al. User Interests Modeling based on multi-source personal information fusion and semantic reasoning [C]// Proceedings of the 7th international conference on active media technology. 2011：195-204

[6] Chen J, Nairn R, Nelson L, et al. Short and tweet：experiments on recommending content from information streams[C]// Proceedings of the 28th International Conference on Human Factors in Computing Systems. 2010：1185-1194

[7] Wen J, Lim E P, Jiang J, et al. TwitterRank：finding topic-sensitive influential twitterers[C]// Proceedings of the third ACM International Conference on Web Search and Data Mining. 2010：261-270

[8] 宋巍，张宇，谢毓彬，等. 基于微博分类的用户兴趣识别[J]. 智能计算机与应用，2013，3(4)：80-83

[9] Chen J, Nairn R, Nelson L. Short and tweet：experiments on recommending content from information streams[C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2010：1185-1194

[10] Zhao Hua, Zeng Qing-tian. Micro-blog Hot Event Detection Based on Dynamic Event Model[J]. Lecture Notes in Artificial Intelligence，2013，8041：161-172

[11] http://wordpress.org/plugins/cardoza-3d-tag-cloud/