基于兴趣点标注的搜索用户分类研究*

张忠华[1] 刘 云[1] 左 肖[2]

(1. 北京理工大学管理与经济学院 北京 100081; 2. 百度营销研究院 北京 100085)

摘 要 查询背后的信息需求里蕴含着用户的兴趣信息,搜索引擎可以通过标注兴趣信息为用户提供多样化的服务。依据用户检索的关键词的语义标注用户的兴趣信息,将非结构化的关键词数据扩展为结构化的数据矩阵,利用聚类分析方法对搜索用户进行分类,并结合对应分析方法对不同类别的搜索用户的特征进行了解读。

关键词 用户兴趣 数据矩阵 聚类分析 对应分析

中图分类号 G203 **文献标识码** A **文章编号** 1002-1965(2013)06-0137-04

Classification of Search Users Based on Marking Interest

Zhang Zhonghua[1] Liu Yun[1] Zuo Xiao[2]

(1. School of Management and Economics of Beijing Institute of Technology, Beijing 100081;
2. Baidu Marketing Institute, Beijing 100085)

Abstract The information need behind the query contains the user's interest information, search engines can provide diversified service for users based on their interest information. In this paper, we design a data framework for keywords based on user's interest, and extend the unstructured keyword to structured data matrix, and then classify the web search users by using cluster analysis and correspondence analysis.

Key words users' interest data matrix clustering analysis corresponding analysis

0 引 言

随着互联网的发展,个性化信息服务和用户分类的研究逐渐成为热点。搜索引擎是连接网络信息资源与用户需求的纽带,它在满足用户信息需求的同时也反映出用户的搜索意图[1],搜索意图往往蕴含用户的兴趣信息,基于这些兴趣信息,可以实现对搜索用户的分类,搜索引擎可以据此为用户提供个性化的信息服务。

现有的搜索用户分类体系多是基于社会化标签构建的,社会化标签是指用户自发为某类信息进行的描述[2]。然而,社会化标签的模糊性、多样性、非结构行等缺陷[3],降低了分类体系的准确率。社会化标签应用于搜索用户的分类研究,还会产生数据稀疏问题[4]。很多研究[5-9]对基于社会化标签的用户分类方法进行了改进,效果并不好。

为了解决这些问题,本文利用标准化的标签标注用户关键词里蕴含的兴趣信息,将非结构化的关键词数据扩展为结构化的数据矩阵,进而利用聚类分析方法对搜索用户进行分类,并且结合对应分析方法对不同类别的搜索用户的特征进行解读。

1 数据和方法

1.1 数据获取 国外研究使用的搜索数据多来自于Google搜索引擎,对于中文搜索而言,百度搜索引擎市场份额高达70%以上,因此百度的搜索数据更具有普遍意义,本文使用的数据来源于百度搜索引擎。本文获得了2011年7月到2011年9月之间的987个cookie对应的43 722条关键词数据,也就是说,本文获得了987名搜索用户检索的关键词数据。由于本文研究的搜索用户分类并不针对特殊的用户群,所以,样本总体为全体搜索用户。

收稿日期:2013-03-07 修回日期:2013-05-03

基金项目:国家自然科学基金重点项目"国家自主创新体系国际化理论与政策研究"(编号:71033001);国家国际科技合作专项项目"应对气候变化南南科技合作监测系统构建"(编号:2012DFG11750)。

作者简介:张忠华(1989-),男,硕士研究生,研究方向:知识管理、技术创新管理;刘 云(1963-),男,教授、博士生导师,研究方向:科技评价、科技政策、技术创新管理;左 肖(1988-),男,硕士,营销分析师,研究方向:网络营销、搜索引擎营销。

从搜索引擎获取的原始数据很“脏”,存在一些与本研究不相关的信息,对数据进行预处理后,获取的数据形式的部分示例如表1所示:

表1中的编号代表搜索用户,时间是该搜索用户检索相应关键词的日期,关键词一列记录的是用户检索的关键词。

表1 数据格式

编号	时间	关键词	编号	时间	关键词
1	2011/7/21	山东大学排名	2	2011/7/11	支付宝登陆
1	2011/7/20	陕西人事考试网	2	2011/7/16	笔记本硬盘
1	2011/7/22	榆林人事人才网	2	2011/7/19	汕头硅谷学院
1	2011/7/22	NBA 频道	3	2011/8/28	僵尸先生
1	2011/7/22	QQ 游戏	3	2011/8/30	小魔仙第1部
2	2011/7/10	目前最好的显卡	3	2011/9/5..	起点中文网

1.2 数据处理 搜索引擎作为互联网上获取信息的最常用工具,是连接信息资源和用户需求的纽带。用户搜索的关键词从字意上可以被归为不同的类别,根据这些关键词的语义可以判断出对应的用户所关注的兴趣点。为了从关键词数据中“抽取”兴趣内容,获得更丰富的信息,实现基于兴趣点标注的用户分类,本文将这些蕴含在关键词中的兴趣信息标注在关键词之后,作为关键词数据的拓展部分。拓展数据的部分示例如表2所示:

表2 拓展数据

编号	时间	关键词	兴趣点
1	2011/7/20	陕西人事考试网	教育培训
1	2011/7/21	山东大学排名	教育培训
1	2011/7/22	榆林人事人才网	求职招聘
1	2011/7/22	NBA 频道	运动
1	2011/7/22	QQ 游戏	游戏
2	2011/7/10	目前最好的显卡	IT 数码产品
2	2011/7/11	支付宝登陆	网购
2	2011/7/16	笔记本硬盘	IT 数码产品
2	2011/7/19	汕头硅谷学院	教育培训
3	2011/8/28	僵尸先生	游戏
3	2011/8/30	小魔仙第1部	动漫
3	2011/9/5	起点中文网	文学爱好者

表2中的兴趣点一列记录的是根据对应关键词的语义手工标注的兴趣信息,以编号为1的用户(以下简称用户1)为例说明本文标注兴趣信息的过程。用户1检索了“陕西人事考试网”和“山东大学排名”,从这两个关键词的语义分析用户1可能是为了查询考试报考的相关信息,说明用户1比较关注“教育培训”这一兴趣点,故在两个关键词后标注“教育培训”。同理,在“榆林人事人才网”、“NBA 频道”、“QQ 游戏”后分别标注“求职招聘”、“运动”、“游戏”,这样就完成了对于用户1的兴趣信息的标注。

本文并没有选取量级巨大的数据进行研究,但是在实际应用时,必然会遇到海量数据的处理问题,人工标注兴趣点显然不能处理海量数据,所以,有必要对标注兴趣点的过程进行改进。本文提出一种方法以供参考,依据关键词的检索量,挑选可以归为某一兴趣点的核心关键词,制作相应的核心词表,当关键词与这一核心词表里的词匹配或包含这一核心词表里的词时,该关键词就被标注上相应的兴趣点信息。为每一个兴趣点整理出对应的核心词表,编写脚本通过导入这些词表给关键词自动标注兴趣点。

把搜索用户的兴趣按照个数汇总,得到搜索用户在每个兴趣点上的得分(个数),汇总结果的部分示例如表3所示:

表3 兴趣点汇总

编号	教育培训	求职招聘	运动	游戏	IT 数码	网购	动漫	文学
1	2	1	1	1	0	0	0	0
2	1	0	0	0	2	1	0	0
3	0	0	0	1	0	0	1	1

表3只简要列举了兴趣点汇总表的小部分,完整的数据处理结果中每个用户在各个兴趣点上的得分差异很大,为了减小得分的过大差异对分析结果的影响,本文对搜索用户在兴趣点上的得分进行了标准化处理。本文对数据进行标准化的原则是尽量保证标准化后不同的得分数量接近,从而有利于聚类分析得到好的结果。具体的标准化过程为:把得分为0的分值重定义为0,把得分大于等于1小于等于10的分值重定义为1,把大于10的分值重定义为2。

1.3 研究方法

1.3.1 K-means 聚类分析方法。用户细分的关键是要找出能够据以分类的用户特征,通过扩展关键词数据,本文已经描述了搜索用户的兴趣特征,通过一些统计分析方法可以依据搜索用户的兴趣特征对其进行分类。用户细分的实质就是对用户分类,聚类分析方法是常用的分类方法。

聚类分析中的 K-Means 算法简便实用,是较为常用的聚类分析算法,在 K-means 算法中事先并不知道目标数据应该被分成多少个类别,在 K-means 算法中需要根据初始聚类中心来确定一个初始划分,然后对初始划分进行优化,初始聚类中心的选择对聚类结果有较大的影响[6]。本文在数据分析处理时,并不能事先确定分类个数,所以使用该算法可以取得较好的效果。

1.3.2 对应分析。K-Means 聚类分析结果的可读性较差,仅仅观察聚类结果表中的信息,并不能很好的描述每一类搜索用户的特征,本文总结现有研究文

献，尝试使用对应分析的方法解读聚类分析的结果。

对应分析是近年新发展起来的一种统计分析技术，通过分析由定性变量构成的交互汇总表来揭示变量间的联系。它最大特点是能把众多的样本和众多的变量同时作到同一张图解上，将样本的大类及其属性在图上直观而又明了地表示出来。对应分析适用于数据是频次或频率的资料[7-8]。

2　结果及分析

2.1　聚类分析结果　结合 K - Means 聚类的相关特征，本文根据所要处理数据的量级将 K 分别设置为 5、6、7，通过对比分析每个 K 值对应的聚类结果，得出的结论是，将样本搜索用户分成 5 类的效果最好，具体的聚类结果如表 4 所示:

表 4　搜索用户分类结果

	1	2	3	4	5		1	2	3	4	5
IT 数码产品	1	1	1	1	1	腾讯 QQ 消费品	1	0	0	0	0
动漫	0	0	0	0	0	投资理财	0	0	0	2	0
房产家居	0	0	0	0	0	网购	0	1	0	1	0
服装日用品	0	0	0	0	0	文学	0	0	0	0	0
婚恋	0	0	0	0	0	新闻媒体	0	0	0	0	0
家电	0	0	0	0	0	星座	0	0	0	0	0
教育培训	0	1	1	1	1	医疗保健	0	0	0	0	1
旅游票务	0	1	0	1	0	音乐	0	0	0	0	0
汽车	0	0	0	0	0	饮食	0	0	0	0	0
求职招聘	0	0	0	0	0	游戏	2	0	0	0	0
社交	0	0	0	0	0	娱乐八卦	0	0	0	0	0
视频	1	0	2	0	0	运动	0	0	0	0	0
书刊	0	0	0	0	0						

表 4 中第一行的数字代表不同类别的用户，第一列是兴趣点信息，表中的数值表示的是每类用户在对应的兴趣点上的得分。每个类别的搜索用户数如表 5 所示:

从表 5 中可以看出，第 5 类用户的数量稍多，第 4 类用户的数量略少，每个类别的搜索用户数比较接近，说明本文对搜索用户的分类结果是有效的。

表 5　每个类别搜索用户数

类别	人数	类别	人数
1	205	4	135
2	196	5	243
3	208		

2.2　对应分析结果　根据对应分析算法的特点，本文依据现有数据，构造用以度量用户类别和用户兴趣点之间联系强弱程度的数据。根据完整的拓展数据结果，计算出每一类搜索用户在每个兴趣点上的平均得分，得到的每类用户的平均得分数据的部分示例如表 6 所示。

表 6　用户在兴趣点上的平均得分

分类	兴趣点	得分	分类	兴趣点	得分
1	IT 数码产品	0.698565	4	动漫	0.08209
2	IT 数码产品	0.883249	5	动漫	0.05
3	IT 数码产品	0.730769	1	房产家居	0.15311
4	IT 数码产品	1.044776	2	房产家居	0.345178
5	IT 数码产品	0.641667	3	房产家居	0.134615
1	动漫	0.368421	4	房产家居	0.395522
2	动漫	0.035533	5	房产家居	0.433333
3	动漫	0.105769			

表 6 只列举了搜索用户在兴趣点上的平均得分数据的一部分，实际运算结果中的得分数据要大得多，平均得分的数值度量的是人群与兴趣点之间的关联程度的强弱。

对兴趣点变量进行数值转换并进行对应分析的运算。运算后的结果包含对应分析结果摘要和对应分析图，其中对应分析结果摘要是整个对应分析结果的汇总表，是输出结果中最重要的一个，主要用于确定使用多少个维度来对结果进行解释。对应分析图是对变量间相互关系进行直观描述的图形，是对应分析的主要结果，主要用于在一个低维度空间描述各个变量之间的相互关系，具体结果见表 7 和图 1:

表 7　对应分析结果摘要

维数	奇异值	惯量	卡方	Sig.	解释	累积	标准差	相关
1	.393	.154			.540	.540	.167	.148
2	.252	.063			.222	.762	.201	
3	.240	.058			.202	.964		
4	.101	.010			.036	1.000		
总计		.285	9.150	1.000[a]	1.000	1.000		

如表 7 所示，第一维(0.54)、第二维(0.222)的惯量比例积累为 0.762，这表明第一维度和第二维度分别解释了总信息量的 54% 和 22.2%，共同解释了信息量的 76.2%，因此，采用二维图形可以反映两变量之间的绝大部分信息。

根据图 1 中所示的每类用户的兴趣特征，本文从兴趣特征角度描述 5 类用户的特点。

第一类用户与其他四类用户的特征差异非常明显。根据图 1 并结合点在两个维度上的得分可以看出，第一类用户(1.016，-0.491)更加关注游戏(1.114，-0.660)、动漫(1.327，-0.633)、QQ 消费品(0.785，-0.640)这三个兴趣点，而且这三个兴趣点和其他兴趣点的得分差异很大，与其他兴趣点的关联程度不高。这类用户在全体样本网民中具有鲜明的独特性，其兴趣特征相较于其他样本网民有很大的区别，从

他们关注的兴趣点可以推断他们是比较年轻的一类网民。

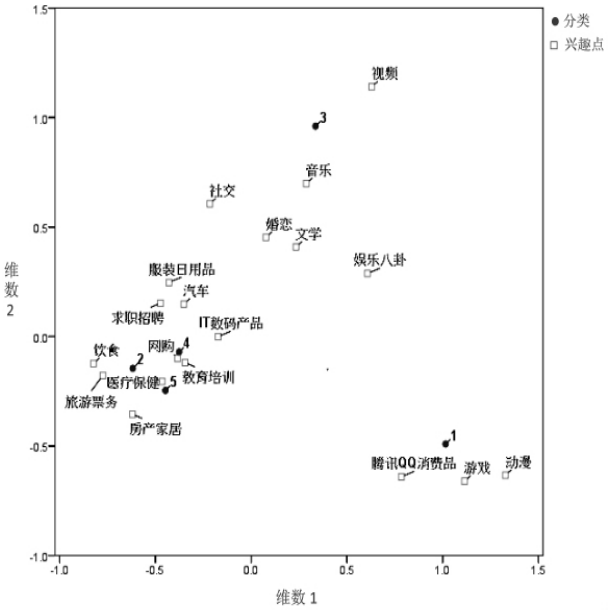

图1 每类用户兴趣特征

第三类用户与其他四类用户的特征差异也比较明显。从图1中可以看出,表示第三类用户的点和表示其他四类用户的点距离都很大。这类用户关注的兴趣点比较多,包括音乐、婚恋交友、视频、社交、文学、娱乐八卦,从图1中点的距离可以看出,这类用户对这些兴趣点的关注是一种"弱关注",而且这些兴趣点之间也没有很强的关联。第三类用户所关注的兴趣点多与娱乐有关,这类用户喜爱娱乐,但没有特别喜爱的娱乐项目,这类用户的兴趣特征比较契合大部分网民的上网动机。

第四类用户与第二、五类用户的特征比较接近,但也存在一些差异。根据图1并结合点在两个维度上的得分可以看出,第四类用户(-0.376,-0.071)对网购(-0.383,-0.100)、教育培训(-0.345,-0.119)这两个兴趣点的关注程度非常强,而且这两个兴趣点之间也有很强的关联。这类网民热衷于网购,是各大电子商务网站的活跃用户,他们也很关注教育培训。

第二类用户关注的兴趣点比较多,包括饮食、旅游、网购、医疗保健、教育培训,从图1中可以看出,这类用户对这些兴趣点的关注也是一种"弱关注",而且他们关注的兴趣点之间既没有很强的关联又没有相似的特征。

第五类用户与第二、四类用户的特征比较接近,但也存在一些差异。根据图1并结合点在两个维度上的得分可以看出,第五类用户(-0.448,-0.246)对医疗保健(-0.465,-0.205)这一兴趣点的关注程度很强。

3 结 论

a. 搜索关键词蕴含网络用户的兴趣信息。搜索引擎作为互联网上获取信息的最常用工具,是连接信息资源和用户需求的纽带。搜索关键词是网络用户真实需求的体现,根据这些关键词的语义可以推断出用户所关注的兴趣点,通过标注这些兴趣点,描述出用户的兴趣特征。

b. 基于关键词搜索可以实现用户分类。本文依据标准化的兴趣标签,通过标注搜索关键词所对应的用户兴趣信息,扩展关键词数据,得到用户的兴趣数据,根据兴趣数据并结合聚类分析方法将搜索用户分类,通过对应分析方法完成对于不同类别搜索用户特征的解读。

c. 基于人群分类的精准营销理论适用于搜索引擎营销。网络营销的核心是精准营销,搜索引擎营销作为网络营销中发展最好的领域,可以为广告主提供精准的营销价值。网络营销的发展趋势是实现基于人群分类的精准营销。搜索引擎作为互联网的入口应用,收集了海量的用户数据,通过分析这些用户数据,完全可以实现搜索引擎领域基于人群分类的精准营销。

参 考 文 献

[1] Border, A. A Taxonomy of Web Search [M]. SIGIR Forum 36(2), 2002

[2] 史云飞,吴江宁,宣照国,等. 基于用户兴趣扩散模型的网络资源推荐方法[J]. 情报学报, 2012, 31(3):275-280

[3] 熊回香,王学东. 大众分类体系中标签概念空间的构建研究[J]. 情报学报,2012,31(9): 984-992

[4] 张富国. 基于标签的个性化项目推荐系统研究综述[J]. 情报学报,2012,31(9):963-972

[5] Schafer J, Frankowski D, Herlocker J, et al. Collaborative Filtering Recommender Systems [C]. Lecture Notes In Computer Science, 2007, 4321:291-324

[6] 杨天霞,王治和,王 华,王凌云. 聚类初始中心点选取研究[J]. 南京师大学报: 自然科学版, 2010,33(4):161-165

[7] 王 阶,何庆勇. 基于聚类分析和对应分析的稳定型心绞痛证候要素组合规律的研究[J]. 中西医结合学报,2008,6(7):690-694

[8] 邢雁伟,王 阶,衷敬柏,等. 采用聚类分析和对应相关方法研究1069例冠心病心绞痛证候应证组合规律[J]. 中华中医药志,2007,22(11):747-750

(责编: 白燕琼)