基于专业实践主题的博客信息内容组织与聚类推送*

胡昌平，余婷婷
（武汉大学信息管理学院，武汉 430072）

摘要：针对高等学校专业实践中利用博客信息所存在的问题，进行基于专业实践主题的博客信息组织构架；在博客信息采集、筛选和关联分析的基础上，提出面向专业实践选题、进程以及成果应用的聚类推送模式，以更好地实现博客信息的集中获取和交互共享。

关键词：专业实践；博客信息；内容组织；聚类推送
中图分类号：TP391
DOI：10.3772/j.issn.1673—2286.2014.09.02

高等学校的人才培养模式正在发生深刻变革，除深化专业理论教学和强调知识应用外，课程实验、实习和基于专业核心课程链的社会实践活动等逐渐得到重视。从面向社会需求的人才培养上看，专业实践是提高学生综合素质和保证学生培养质量的关键环节，因而基于数字网络的有关专业学习、知识应用与实践开展的博客交流，引起高校师生的广泛关注[1]。

从来源上看，与专业实践相关的博客信息，分布在不同的平台上。围绕专业实践进行博客信息查询，需要同时对多个网站的相关资源进行搜索[2]。另外，通过主题词检索获得的博客信息不仅数量巨大，而且涉及内容宽泛。事实上，用户的需求是对博客信息的集中获取，而不是全面浏览。鉴于此，本研究拟对专业实践主题的博客信息进行组织，以便在此基础上筛选、过滤和内容聚合，从而实现基于专业实践主题的博客信息的集中组织和交互共享。

1 基于专业实践主题的博客信息采集

1.1 专业实践主题博客的选取依据

专业实践主题博客是指学生从事专业实践活动中所发表的博客，以及与专业实践内容相关的学科知识、社会发展等方面的博客。从主题词结构和分布上看，专业实践博客的主题词与相关学科、社会发展、相关行业等领域的主题词之间具有关联关系。专业实践的主题词不仅包含在学科专业主题词之中，而且包含在本学科专业对应的行业以及对社会发展的影响情况的主题词之中。这说明，在内容选择上，应从学科和社会两方面进行架构；在来源选择上，应注重基于主题的多源博客信息渠道。

以信息管理与信息系统专业学生所进行的“城市智能交通”方面的专业实践主题为例，其专业实践在数字化环境和智慧城市建设背景下展开，面向现实问题的主题包括“智慧城市”、“公交管理信息化”、“城市智能交通”、“智能交通系统”、“智慧交通服务”、“城市道路安全”等；同时，进行“城市智能交通”方面专业主题实践，还需要利用本专业知识，所涉及的主题包括“城市信息管理”、“数字信息技术”、“智能信息系统”、“信息可视化”、“信息系统安全”、“数字地图服务”等方面的主题词。围绕这两个方面主题的博客信息发布者，不仅包括从事专业实践活动的学生、指导教师，还包括与智能交通有关的管理、行业部门人员、广大市民，以及智能交通方面的专家学者。从博客信息分布上看，博主可以在不同平台上发布信息，进行广泛的博客交流。

*本研究得到湖北省教学研究项目“信息管理类专业实践教学体系构建与改革研究”（编号：JG2012003）资助。

在博客信息组织的探索中，本研究根据博客平台的影响度，选取新浪博客、网易博客和搜狐博客三个具有代表性的博客，按主题词进行博客信息采集。在2014年8月所采集的数据中，依据搜索结果的相关度排序，选取每一个主题词在每一个博客平台上的前3页的博客内容作为数据源，在此基础上进行博客信息处理。

1.2 专业实践主题博客的信息内容抽取

由于通过网络爬虫获取的信息都是关于博客内容的HTML格式文本，由HTML标记和文字组成，因此需要对博客的HTML网页进行去噪，清除无用的标签，从而获取所需要的内容数据。通过定向选择，集中展示相关博客的标题、作者、发表日期、链接以及博客正文。相对于一般网页，同一平台上的博客网页比较特殊，其书写格式相对固定，而且比较规范，不同之处只是博客网页中的内容存在差异[3]。因此需要在分析博客HTML格式的基础上，按照笔者的自建模板（包括博客链接、标题、发表时间、作者以及文本内容），通过与模板库中的模板进行匹配来进行信息提取（见图1）。

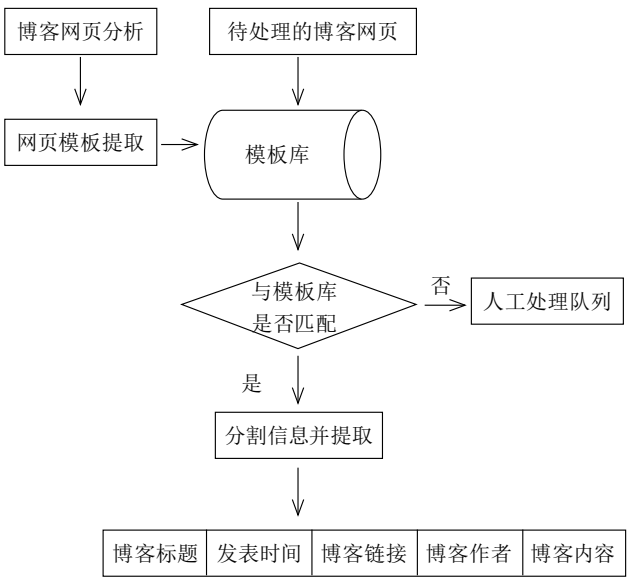

图1 博客信息抽取过程

如图1所示，对于给定的博客信息链接，分析链接，判定此博客来自于哪个博客平台；根据博客来源平台，从模板库中得到博客信息抽取的正则匹配式，进行信息的分割并提取博客标题、时间、链接、作者以及内容；对于无法匹配成功的特殊博客网页，放入人工处理队列，通过人工方式对其进行信息抽取；当所有的信息被成功提取出来后，顺序存入数据库中。

2 基于专业实践主题的博客信息内容组织

通过上述处理得到的面向专业实践的博客信息内容，由于文档表示的差异，需要进行统一的信息文本提取，按数字化形式进行文本表示。笔者采用向量空间模型方法，利用文本之间的夹角余弦作为相似度的度量[4]。对于“城市智能交通”方面的博客信息文档进行预处理后，转化为N维词条空间的一个向量，然后进行特征项提取和特征项赋权。

由于专业实践的主题比较明确，对于相关主题博客信息的文本表示可采用通行的方法进行。值得指出的是，博客信息内容深度，同时具有发文者的情感色彩和主观评论，文本中一般会包含助词、虚词等出现频率很高但是本身对于信息内容的重要程度很低的词，这些词语的存在不仅会增加存储空间，而且会影响文本表示的客观性，所以需要将其过滤掉。对此，可借鉴停用词表的处理模式，将缺乏实质意义的词语过滤掉。

同时，有关社会热点的博客信息，其内容表达往往不具规范性，这就需要建立不同博主用词之间的对应关系，设立专业实践方面的同义词处理规则，采用集合的方式进行同义词的描述。在内容处理上：通过遍历词表查找到该词；根据集合关系找到该词的序列；根据序列号将该词映射到序列的第一个词上面。通过以上算法，我们就可以将提取出来的特征项词条进行同义词消除，从而合并为一个特征项，这样不但降低了词条的数量，也降低了误差，提高了文本表示的准确度。

对于专业实践中经过分词得到的博客词条，其可用程度是不一样的，如一个在一篇文档中出现了20次的词，和出现了2次的词很明显重要程度不一样，这就需要采用赋值的方式来区分词条的重要程度。对于主题博客信息的权重设置，可以将所采集的博客集合视为一个整体，按博文中的词频对文本的重要性赋予一定的权重。这样，有用词条赋予较高权重，无用词条赋予较低权重。在“城市智能交通”方面的专业实践中，权重较高的词条往往是开展社会实践所需要解决的关键问题，如交通位置服务、交通智能安全、数字地图、智能交通查询等。

3 基于专业实践主题的博客信息聚类推送

互联网上，关于专业实践主题的博客信息数以万计，而且每天又有大量的博客更新和加入。面对如此巨量的信息，对信息内容进行重新聚类是解决信息快速获取的有效途径。以“城市智能交通”方面的实践为例，在此主题下，有的博文是关于国家政府的项目招标，有的博文是关于智能交系统的构建，有的则是关于智能交通的科技成果展示，因此我们需要将博文信息进行重新组织，进行不同主题下的聚类，以便快速定位相应的信息内容。

3.1 专业实践主题博客的聚类方法

一般说来，聚类通过一定的规则将抽象的对象集合按照相似性划分为若干个类或簇，由于同一个簇中的相似度较大，不同簇中的相似度较小，其聚类比较容易实现[5]。然而面向具体问题的博客信息，由于发文者思考问题的角度和认识上的差异，对同一主题的博文论述很难用简单的相似度计算方法来解决，因此需要进行深层聚类处理。在细分聚类中，基于层次划分的聚类方法、基于密度的方法以及基于网格的方法具有实用性。相比较而言，层次聚类算法使用简单，它不同于基于划分的K-means算法对初始数值的依赖性，其对数据类型要求较低，因此可应用于面向专业实践的博客信息聚类组织。

层次聚类算法是根据计算每两个对象之间、对象和分组之间以及分组和分组之间的距离，然后按照距离的大小构建一个层次聚类层次。其依据是计算两个对象之间的相似度，然后将距离最近的两个簇合并成为一个新的簇，这样循环下去，直到满足某个终结条件[6]。例如，簇C1（公交地图服务）中的一个“数字地图”对象和簇C2（公交线路查询）中的一个“位置显示”对象之间的距离相对较小，其数字地图显示和线路位置显示都需要通过数字地图的利用来实现，那么C1和C2就可以合并，这样可以进行专业实践中相关技术应用博文内容的整合。具体过程如下：在初始阶段，对于给定的文档集合，将有待处理的新文档作为一个独立的类别，即每一个文档为一个簇，然后根据相似度计算公式，计算每两个簇之间的相似度，得到相似度最大的两个簇，并且将其合并为一个簇；否则，独立列为一个新的簇类。如此循环下去，计算有关“城市智能交通”社会实践方面的簇之间相似度，合并相似度最大的簇，直到将所有的文档被划分为恰当的簇类位置。以此出发，保证专业实践主题博客信息聚类组合的动态性和完整性。

3.2 专业实践主题博客的聚类推送模式

面向社会实践的博客信息内容聚类推送的出发点在于，按专业实践选题需求和实践活动的开展需要进行[7]。从博客信息推送内容上看，着重于以下几方面的聚类：

（1）面向专业实践选题的博客信息聚类推送

专业实践选题具有面向现实问题的针对性，“城市智能交通”方面的实践主题选择，往往是社会共同关注的问题，这些问题在相关博客里面可以得到全面体现；同时，专业实践的开展与专业学习内容息息相关，其选题应是本专业所能解决的问题，专业知识应用于社会问题解决的博客信息内容具有关键作用。这两方面的博客信息聚类推送在于为专业实践选题论证提供依据，其中前者关系到实践选题的必要性和社会意义，后者关系到实践选题的可行性和专业依据。在专业实践主题热点聚类中，拟按照热点的不同内容进行不同主题归类；在专业知识应用的博客信息聚类中，其细分类按博客信息内容的专业逻辑关系组织。在“城市智能交通”的专业实践选题聚类推送中，可以将如国内城市智能交通现状与发展方向、中国城市道路智能交通行业调研报告、我国智能交通发展现状与城市应用探讨等方面的博文，进行内容细分和聚类，以此进行内容推送。

（2）专业实践进行中的博客信息内容聚类推送

专业实践活动在社会调研和前期准备的基础上开展，其进行中需要按预定方案解决具体的专业问题。除相关的研究成果信息外，博客信息由于具有面向专题研究的交互性，因而有一定的参考价值。据此，专业实践进行中的博客信息内容聚类，拟按专业实践研究的问题进行主题推送。例如，“城市智能交通”方面的专业实践，对于城市数字交通地图、基于地图的数字服务技术、面向多用户的位置查询服务、城市交通诱导上端系统的设计与实现、城市轨道交通信息控制、城市路口闯红灯检系统、网络图像监控光端机在城市智能交通中的应用、基于GIS平台的城市智能交通管理系统构架研究等方面的博文，需要按内容进行解决方案的细分，进

行主题推送。同时专业实践进行还需要解决前所未有的新问题，这就需要进行关键问题解决技术上的创新，其博客信息推送主题还应包括本专业理论与方法应用方面的博文。

（3）面向专业实践成果应用的博客信息聚类推送

专业实践的目的，一是通过专业知识的应用和相应的社会现实问题的解决，培养学生的实践能力；二是在专业实践中围绕当前所要解决的理论与现实问题，利用本专业优势进行理论、方法与技术创新。因此，专业实践中的成果具有拓展应用前景。专业实践成果的应用，首先需要对成果进行客观评价，与此相关的博客信息有必要进行展示；同时，与专业实践活动相关的政府部门、研发机构、企事业单位和公众的博客信息，必然涉及具体成果的应用，此类博客信息也需要进行展示。有关“城市智能交通”的成果应用的城市智能交通防雷系统的应用、江苏南京市将构建城市综合智能交通系统、上海热力打造智能交通发展智慧城市等博客，拟按内容进行细分和主题聚类推送。

4 结语

与论坛信息和微博信息相比，博客信息具有内容详细、分析全面和专业性较强的特点，在高等学校专业实践的信息利用中，是其他信息源所无法取代的。同时，博客信息发布，往往围绕社会共同关注的问题和一些前沿性问题展开，往往是学生进行专业实践选题和开展专业实践活动的重要参考文献源。目前存在的问题是，博客信息的提供和服务往往局限于博文内容的采集和基于标题的内容发布，缺乏面向需求的内容细分和集中推送。针对这一现实问题，本文进行了基于主题的博客信息跨平台搜索、内容关联分析和深层聚类，以此为基础实现了博客信息的文本表示和面向专业实践的主题推送构架，并以“城市智能交通”专业实践为例进行了实证。博客信息的主题聚类推送不仅在专业实践中需要，而且在其他主题活动中也是重要的，因而本文的研究具有普遍意义。

参考文献

[1] 李朝红.博客在高职电子商务专业实践教学中的应用[J].创新与创业教育,2010,1(6):62-64.

[2] 胡昌平,李琛.基于内容聚合的博客辅助学习模式研究[J].情报科学,2008,26(6):801-804.

[3] 庞俊.基于确定话题和情感极性的博客文本聚类研究[D].武汉:武汉理工大学,2010.

[4] SALTON G. The SMART Retrieval System Experiments in Automatic Document Processing [J]. Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[5] 黄晓斌,赵超.文本挖掘在网络舆情信息分析中的应用[J].情报科学,2009,27(1):94-99.

[6] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: An efficient data clustering method for very large databases [C]// Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal: ACM Press, 1996: 73-84.

[7] 胡昌平,等.创新型国家的知识信息服务体系研究[M].北京:经济科学出版社,2011.

作者简介

胡昌平，男，1946年生，武汉大学信息管理学院教授、博士生导师，研究方向：信息服务与用户。
余婷婷，女，1990年生，武汉大学信息管理学院硕士生，研究方向：信息服务，E-mail：837857547@qq.com。

Blog Information Content Organization and Clustering Push Based on Professional Practice Topic

HU ChangPing, YU TingTing
(School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: This paper studies on how to construct the blog information cluster organization based on the professional practice topic, aiming at institutions of higher learning in professional practice problems in using blog information. On the base of blog information gathering, screening, correlation analysis, put forward the pattern for topic selection, process and results' application of clustering push, in order to better access and interaction of the blog information sharing.

Keywords: Professional practice; Blog information; Content organization; Clustering push

（收稿日期：2014-09-11）