中文博客标签调查分析及标签推荐模型的研究*

宋洪鑫，李蕾，刘冬雪

北京邮电大学，计算机学院，智能与科学技术研究中心

shx-007@163.com, leili@bupt.edu.cn, beijing1112222@gmail.com

摘　要：随着标签作为 web2.0 时代一种重要的资源描述方式引起了人们极大的重视。本文主要分析了中文博客网站标签的标注情况，包括标签分类、热门标签、命名实体、网络流行语等，在此基础上总结了一些标签标注的规范，并提出了一种基于分类和主题词提取的标签推荐模型，目的在于帮助 blog 用户能够更准确的通过标签来描述自己的资源，以及在 blog 搜索中发挥更好的作用。本着大众分类的标准，验证了该模型有较好的性能。

关键词：标签推荐，博客，web2.0，主题词抽取

Research on Chinese Blog's Tags and Recommendation Model

Hongxin Song, Lei Li, Dongxue Liu

Center for Intelligence Science and Technology Research (CISTR), School of Computer Science and Technology,

Beijing University of Posts and Telecommunications, Beijing, 100876

shx-007@163.com, leili@bupt.edu.cn, beijing1112222@gmail.com

Abstract: As an important kind of Web 2.0 resource description tool, tag has draw great attention of researchers and users. This paper is fo cused on the usag e and application of Blog tags. We analyze the current situation and development of tagging on Chinese blog, including tag classification, hot tags, named entities in tags, new popular words in tags, etc. And we put forward a set of issues needing attention when tagging. Then we try to build a tag recommendation model according to these issues. The new proposed model is mainly based on text classification and key words extraction of Blog content in order to help blog writers adding more precise tags for his information and performing better in Blog retrieval. We evaluate the performance of this new model from the opinion of folksonomy.

Keywords: tag recommendation, blog, web2.0, key words extraction

1　引言

Web2.0 不再是个模糊的概念，而 web2.0 网站已经是现在网络很普及的了，其具有的信息交互性、提倡个人体验等新特点，为用户在网上进行信息的交流与共建提供了一个全新的平台。

blog 是 web2.0 最典型的代表技术之一，随着 blog 的发展，blog 页面的数量呈指数级别上升。人们通过 blog 发布自己的信息，也希望通过浏览别人的 blog，了解到自己感兴趣的知识和话题，blog 逐渐成为了一个人们共同发布信息、传播信息、发现信息的重要途径。而标签是 web2.0 的特色元素，一般来讲标签是用户在数字资源上标记的关键字，是一种常用的组织和发现资源的方法，反映了用户对资源的认识。对于博客而言，标签是一种更为灵活、更有趣的描述博文的方式，是博主自己写完博文后，自己填上的一串很具有代表性的关键词，通常平均在 3 到 5 个。因此标签是表示博文中心思想的重要部分。并且，对于搜索引擎来说也是一个非常值得利用的因素。根据钟义信教授的“全信息理论”[11]，全信息包含语法，语义，语境三个层次，标签主要体现

*本文承国家自然科学基金项目 60873001，教育部科学技术研究重点项目 108131，教育部科技发展中心网络时代的科技论文快速共享专项研究资助。

在语义这个层次上。但是往往由于博主的知识背景等各种原因在自己填写标签时，缺乏对语境因素的注意。例如，“苹果”，它可以表示一种水果，也可以表示一个计算机的品牌。因此在填写标签的时候往往需要一个合理的规范。

因此本文在引言部分首先介绍了标签的重要性，和标签标注可能出现的问题。在本文的第二部分，主要分析一下目前国际上大众标签标注系统（folksonomy, social tagging 和 collaborative tagging 等）和基于 blog 标签推荐系统的一些研究发展情况。第三部分主要立足于中文博客网站系统，对中文的标签标注情况，做一些详细的分析，分析出了现在中文博客网站的标签特点，并根据现有按“标签”搜索的博客搜索引擎的搜索结果分析出了标签发展的一些问题，并总结出了博客作者在添加标签时应注意的一些问题。在第四部分，根据对标签的分析，本文搭建了一个基于中文博客的标签推荐模型，并通过实验测试，验证了该算法的有效性，最后对本文的主要工作进行总结，并对未来工作改进作了展望。

2 研究背景

标签作为对资源描述的一种有力工具，在国际上，形成了一些著名协作标签系统，即允许用户自由使用标签对资源做标注，从而对资源进行准确和快速的说明，如社会标注网站 Del.icio.us[1]，图片共享管理网站 Flickr，视频共享网站 YouTube 等。由于易于使用并具有较高的交互共享性，这些 web2.0 网站都取得了巨大的成功，吸引了大量的用户的参与和使用。大量的标签，被用户自由公开地标注到各种电子资源上，如网页、视频、图片等等。随着时间的推进，这些社会标准系统已经累积了数以百万计的标签数据。Del.icio.us[1]允许用户对同一网络资源自由的填写标签，在许多用户的共同标注下，每种标签按照被标注的多少排序，这样就形成了一些比较有权威的标签，这样的标签往往更加带有语境信息。

基于这些大众分类的标签系统，许多学者都专研于一种社会标签推荐模型的研究，这样可以使用户更加准确的填写标签，描述文章的内容，还可以使用户用一些比较流行的标签来标注文章，进而可使用户与其他用户进行更好的交流。其中在 blog 标签推荐上，例如：Shigeru Fujimura[2]基于 k-NN 算法，提出了一种多层次的标签推荐系统，这种方法可将具有相同意义标签的聚类，从而识别更有信息量的标签。可是这种算法依赖于已被标注 blog 标签的好坏程度。Brooks[14]开发了一个根据原文信息内容的标签推荐模型，抽出 TF-IDF 得分较高的三个关键词做为推荐标签，但是他将其主要应用在对 blog 的聚类上，并且没有考虑文章作者的自身信息。

因为社会标注系统是从每个浏览网络资源的用户的角度出发，融汇了大众对同一资源的普遍看法，所以社会标注系统往往会更有效率，更有权威性，可以减少用户在添加标签时出现的语义模糊等问题，所以社会标注系统在标签标注方面有很大的优势，但是对于 blog 这一网络资源来说，往往只有一少部分的博客可以被大众发现并且进行标注[4]。这是由于不同的 blog 在“长尾效应”中的体现。“长尾”就是统计学中幂率分布的一个口语化表达方式，如果把 blog 按受到的关注度与长尾理论结合起来，blog 的分布也是服从幂率分的，分布曲线有一个长长的“尾巴”[13]。人们只关注那些处在曲线头部的 blog，这些 blog 往往是一些名人的，或者最先受到关注大众标注的 blog，被大众标注的越多，这样的 blog 就越受关注，这种现象就会导致处在曲线头部的 blog 越来越受到人们的关注，而处在“尾巴”上的 blog，就很少受到人们的关注。这样大众标注系统对这些 blog 而言，起到的作用非常小，而这部分 blog 往往占总数的一大部分。而前

面的理论研究都没考虑到这个问题带来的影响。Subramanya[4]研究了一个为 blog 自动推荐社会标签的系统，目的是使社会标注系统能够对这些处在“尾巴”上的 blog 起到更好的作用，但是没有从根本解决这个问题。现在许多学者都在研究一些方法，可以使这个问题得到很好的解决。

而本文提出的标签推荐系统是针对 blog 的作者，即网络资源的贡献者角度出发，因为一个 blog 的作者在为自己的 blog 填写标签时往往比别人更具有权威性，更能清楚的表达自己的意图，但是由于每个 blog 作者的专业知识背景不同，在填写标签时缺乏慎重的考虑，往往会错用一些没有信息的标签或者忽略一些有价值的标签。因此，本文是根据分析中文 blog 标签的标注现状和特点，主要关注于文章内容而提出一种基于分类和主题词提取的推荐模型。

3 中文标签标注情况的具体分析

3.1 中文各个博客网站标签的“贴标率”分析

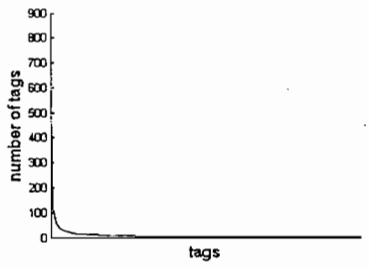
贴标率，即博文被标注标签的频率。这里我们给出一个贴标率的计算公式如下：

贴标率= 抽查中贴有标签的博文数/抽查的博文总数

我们对目前比较成熟的中文的博客网站首先进行了一个粗略的统计，选取了 25 个网站，对其中每个博客网站随机抽取了 100 篇博客，统计该博客网站的贴标率。我们发现，其中博客标签标注情况，较好的为新浪博客，搜狐博客，和讯博客等网站，贴标率均在 80%以上，而平均贴标率仅为 24.61%。在抽取中发现目前有许多博客网站，不但没有标签，而且基本的分类标注也没有，这些网站都忽略了或者没有意识到标签在博客当中的重大作用（将在下文进行详细讨论）。可见目前，中文网站的博客标签发展还不完善。

3.2 新浪网站博客的标签分析

通过上面的分析，我们选择贴标率比较高的新浪博客进行进一步的详细分析。我们利用网络爬虫分三期对新浪的 blog 进行抓取，共抓取了 12279 篇博客，从中共提取出 49444 个标签。经过分类统计，得出 17458 种不同的标签，按照频数从高到低排序，分布形式如下图：



其中，x 轴代表标签，y 轴代表每种标签出现频数。

从图中可以看出中文标签也基本符合幂率分布。这里我们假设每种标签个数多于 10 个的，认为为“热门标签”，热门标签率为：2.56%，假设每种标签在两个以下的为“尾标签”，尾标签率为 70.72%。并且从这三期数据的贴标率来看，新浪博客的用户在填写标签的意识上有了进一步的提高。

热门标签的进一步分析，这三期数据的热门标签基本一样，取出按标签频数排名的前 10 名标签，如表 3-1 所示：

杂谈	娱乐	文化	旅游	健康	教育	情感	体育	房产	财经
2736	828	686	664	618	555	540	490	465	420

表 3-1

从表 3-1 中可以看出，热门标签中，一般都是类别标签，这说明有许多用户在给博文添加标签的时候，都喜欢给自己的博文，定义一个大方向的类。这些数据也直接反应了这些篇博文大概的分类情况。其中“杂谈”这种标签，在新浪的博客中最为常见，这反映出也有很多用户不愿意去刻意的把他们的博文规定类别。

标签的词性分析，本文又人工统计了这些标签中的命名实体，一共统计出了 3634 个命名实体，占标签总数比例为 7.35%，其中人名占 63.46%，地名占 25.65%，机构名占 10.90%，在统计过程中，我们发现人名多在娱乐，文学，体育类中出现，地名多在旅游，文化类中出现。人名出现的最多，可以看出用户在标注的时候也通常将自己的名字或者比较关注的名字添加到标签中。

另外在统计过程中，我们还发现一些新的词语，和网络流行的词汇，这对信息检索、数据挖掘，热点跟踪等研究有很大的帮助。

3.3 新浪博客搜索标签分析

我们又对新浪博客按标签搜索的结果排序方案，进行了分析。我们搜索了一百个热门标签，在搜索结果其中发现，排序方案首先是时间，时间对结果起到了很重要的，其次，标签在文中的权重，还有阅读次数，和评论次数分别对搜索结果的排序有着重要的影响。因此，用户在填写标签的时候要注意自己的标签所在文中的权重，这样可以提高自己文章的排名，其次在填写标签的时候不要把自己的标签变成一个孤立的“岛屿”，这样别的人通过标签找不到你的博文，这样也没办法进行信息之间的交流。

3.4 标签标注规范总结

本文通过对新浪博客标签标注情况的详细分析，根据标签对于博客本身的影响，试着给出博客作者在填写一组标签时应注意的问题：

[1] 首先不要有歧义的标签，

[2] 其次标签在文章中占有一定的权重，

[3] 还要根据文章类别加上一些关键的命名实体，

[4] 最后一组标签不应该全是生僻的，处在“长尾巴”[4]上的标签。

4 基于分类和主题词提取的标签推荐系统

根据以上对标签的分析和得出的结论，从用户喜好和标签在文中占的权重出发，我们试着提出了一种基于文章分类和主题词提取的标签推荐模型。

4.1 系统功能模块

系统大致分为四个模块

1，博客文章的预处理模块，包括中文的分词，去除停用词，以及词性标注三部分。这些程序由北京邮电大学智能与科学技术研究中心提供。

2，主题词提取模块，这个模块的功能，是基于[10]提出的单篇文档的核心词抽取算法,它的提取基本原理，是把文中的关键词按照词性和词频的权重进行排序，进行单个词提取，在此基础上，我们又对经常连续出现的多个单词进行串分析，提取的结果是两组词的向量。其中一组是单个词向量(tagKey1)，例如“经济”，“规律”；另一组是组合词向量(tagKey2)，即按照多个单个词的共现频率、位置和在文章中出现的次数进行组合，例如“经济规律”。

3，分类模块，这里我们设计了两种分类模式，一是系统自动分类，我们采用了比较简单的朴素贝叶斯分类方法，对文章进行分类，我们结合了百度百科和新浪博客的分类方式，确定了 12 类，其中收集了包括人名日报，百度百科，和新浪一些类别较明显的文章，作为训练语料集，每类的训练语料文章数，平均大概在 400 篇左右。二是，根据我们提供的可选类别，由用户手动添加,这里主要考虑到“杂谈”的影响。此模块，在确定分好类之后，把分类作为一个确定标签加入到推荐标签组(tagRe)中，而每个类别下，有一组参照标签系（tagRefVector)，例如，“技术类”中包含一组参照标签，“互联网”、“ 数码”、 “软件”、“硬件”、“工程”、“通信”、“计算机”和“信息”。这些标签是我们从百度百科分类和新浪的热点标签中采集的，目的是用来区分抽取出主题词与该类别的相关度。

4，语义计算模块，本文参考 Dai[12]等人提出 Hownet 语义相似度算法来计算主题词对参照标签系的得分，进而来根据得分来选择推荐的标签。

其中计算单个词的相似度得分计算公式如下：

$$V = Sim(key, tag\,\mathrm{Re}\,fVector)$$
$$= \sum_{i=1}^{n} Sim(key, tag\,\mathrm{Re}\,f_i)$$

其中 n 为 tagRefVector 中的参照标签个数，key 为单个主题词，$tag\mathrm{Re}f_i$ 为第 i 个参照标签;

组合词的相似度得分计算公式如下：

$$keyGR = (key_1, \cdots, key_i, \cdots, key_n),$$
$$V = Sim(keyGR, tag\,\mathrm{Re}\,fVector)$$
$$= 1/n \sum_{i=1}^{n} Sim(key_i, tag\,\mathrm{Re}\,fVector)$$

其中 keyGR 为组合词，n 为组合词中单个词的个数，key_i 为第 i 个单个词。

4.2 系统实现

算法具体步骤：

[1]:对博文的预处理。

[2]: 对博客进行分类，如果分类结果是“杂谈“，那么直接转到[5]，否则得到类别标签 tag1，和相关标签参照系 tagRefVector。

[3]: 对博客进行主题词提取，从单个词组中选取 10 个词 tagKey1，再从组合词组中选取 6

个词 tagKey2，作为侯选词。

[4]:分别计算出 tagKey1 中的词与 tagRefVector 中每个参照标签的相似度求和的分数，按分数由高到低取出 top5，同理在 tagKey2 中取出 top3，连同 tag1 得到推荐标签组 tagRe，结束。

[5]：对博客进行主题词提取，从单个词组中选取 top5 个词 tagKey1，再从组合词组中选取 top3 个组合词 tagKey2，连同“杂谈”作为推荐的标签组 tagRe，结束。

4.3 实验结果与分析

由于标签质量缺乏统一的评价标准，我们的实验数据，是从 del.icio.us[1],上收集了 200 篇已经大众标注过的中文博客文章，并以此作为标准来检测算法的有效性。我们实验采用[14]TFIDF 算法来做为 baseline 系统，其中 TFIDF 算法选取的标签数和我们实验室选取的标签数相同。并通过计算准确率（P）和召回率(R)，来检测我们算法（简称为 TR）的有效性，准确率（P）和召回率(R)定义如下：

p_i = 博客b_i的大众标签与推荐标签的匹配数/推荐给b_i的标签总数

R_i =博客b_i的大众标签与推荐标签的匹配数/博客b_i的大众标签总数

$$P=\sum_{i=1}^{n}P_i/n \ , R=\sum_{i=1}^{n}R_i/n$$

作者标注	我们，美国人，管理，分歧
大众标注	管理，分歧，领导，文化，中国
TR 推荐标注	文化，美国，管理，分歧，领导，下属，中国人，中国人管理，中国人当官

表 4-1

表 4-1 以其中一篇博文为例，可以看出我们推荐的标签，和大众标注的标签有很大的相似性，而用户自己填写标签的比较少，这是由于用户没有注意到本文 3.4 节提到的问题，还有这篇文章引起许多人的关注，因此在大众填写标签的时候，标注的人越多，最后积累的标签也就规范和权威。而在一些不好的情况下，我们推荐的标签和用户自己的标签，还有大众添加的标签相差较远，这是因为我们推荐的标签比较注重原文，而用户在添加标签的时候，更加灵活，还有在大众标注中，这篇文章被标注的人比较少，这样的标签可能具有比较低的参考价值。

标注方法	准确率(P)	召回率(R)
TR	31．43%	40．86%
TFIDF	25．74%	32．33%

表 4-2

从表 4-2 分析的结果中可以看出，我们的推荐算法与 TFIDF 算法相比，无论是在准确率和召回率方面，都有了一定的提高。而对于算法本身而言我们得到的准确率和召回率并不是很高。因为本文提出的推荐算法是从标签的分析出发，考虑到用户的心理等因素提出的。而相对于大众标注来说，因为每篇的大众标签的权威度高低不等可能会影响我们的实验效果，还有大众添加标签

可能更加灵活，概括的比较抽象，这样我们标签推荐系统来说有一定差别，但仅出结果来看，我们的标签推荐系统也有一定的作用。因此本文提出的标签推荐算法在总体上来说是有效的。

5 结论与展望

本文对中文博客网站的标签标注情况进行了统计分析，说明了中文标签还处在发展阶段，并以新浪博客为例，分析了博客标签的分布情况，和用户标注标签的特点，和反映出用户的一些心理偏好，并且通过对标签的分析可以发现网络的热点词汇。我们还发现了标签中存在的一些问题，并分析了标签在博客搜索中的应用情况，以及影响结果排名的因素。最后给出了一些规范填写标签的建议。最后在分析的基础上，提出了一个基于分类和主题词提取的标签推荐系统，并通过实验验证了方法的有效性。在今后的工作中，我们将会收集更多的数据，更加细化分类，收集热门标签，完善标签推荐系统，使推荐结果更符合用户的心理需要。

参 考 文 献

[1] http://del.icio.us/

[2] Shigeru Fujimu ra, Ko Fujimura. Blogosonomy : Autotagging an y text using bloggers' kno wledge. International Conference on Web Intelligence, 2007. 205~212.

[3] Lara Marcellin, Roberto Polit i .Tag Vision: Soci al Knowledg e for Collabor ative Search. ACM, 2009. 325~326.

[4] Shankara B. Subramanya, Huan Liu. SocialTagger - Collaborative Tagging for Blogs in the Long Tail. ACM SIGIR, 2008. 19~26.

[5] Ching-man Au Yeung, Nicholas Gibbins, Nigel Shadbolt. Contextualising Tags in Collaborative Tagging Systems. Proceedings of the 20th ACM conference on Hypertext and hypermedia, 2009. 251~260.

[6] Jengun Hwang , Ming Zhang. Personalized Recommendation Based on T ag Clustering. http://www.paper.edu.cn, 2008.10.

[7] Huang Yanjing, Zhang Ming, Feng T ianxiao. A T ag-based Approach of Or ganizing and Ex ploring Literature Resources. http://www.paper.edu.cn, 2007.

[8] Xinghua Hu, Bin W u. Auto matic Key word Extr action Us ing Linguistic Features. Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), 2006.19~23.

[9] Y.MATSUO, M.Ishizuka. Key word Extr action f rom a S ingle Docum ent us ing Word Co-occurr ence Statistical Information. International Journal on Artificial Intelligence Tools, 2004, 13(1): 157-169.

[10] Jun WANG, Lei LI , Lixia LONG. Research of the method for obtaining seman tic knowledg e in the system of CI-NLU based text processing after speech recognition. NLP-KE, 2009. 1~6.

[11] Yixin ZHONG. Principle of Information Science 3nd edition. BUPT Press, 2002. 189.

[12] Liuling Dai , B in Liu . Measuring Semantic Si milarity b etween W ords Using HowNet. Computer Science and Information Technology, 2008. 601~605.

[13] http://www.shirky.com/writings/powerlaw_weblog.html.

[14] Christopher H. Brooks. Improv ed Annotation of th e Blogosph ere via Autotag ging and Hier archical Clustering. Proceedings of the 15th international conference on World Wide Web, 2006. 625~632.