基于凝聚式层次聚类算法的标签聚类研究*

曹高辉　焦玉英　成　全
(武汉大学信息资源研究中心　武汉 430070)

【摘要】对标签、标注、大众分类等概念进行界定,指出现有标签标注系统中存在着标签描述信息的精确度不高、标签检索结果相关度低、标签缺乏有效组织等问题,提出采用凝聚式聚类算法对标签聚类,从而实现对标签的重新组织,为用户提供更好的标签导航、浏览机制。最后通过实验对标签聚类方法进行验证。

【关键词】标签　标签聚类　凝聚式层次聚类
【分类号】G250.7

Research on Tag Cluster Based on Hierarchical Agglomerative Clustering Algorithm

Cao Gaohui　Jiao Yuying　Cheng Quan
(Center for Studies of Information Resources, Wuhan University, Wuhan 430070, China)

【Abstract】This paper firstly defines tag, tagging, folksonomy, then analyzes the limitation of collaborative tagging system. In order to achieve reorganization of the user tags and better tag navigation, browsing mechanism, the authors propose a method on using hierarchical agglomerative clustering algorithm to cluster the tags. Finally experiments certify the tag cluster method.

【Keywords】Tag　Tag cluster　Hierarchical agglomerative clustering algorithm

1　引　言

随着 Internet 技术的不断发展,互联网逐渐从以信息提供商为中心的 Web 1.0 向以用户为中心的 Web 2.0 转变。Web 2.0 作为一种架构在用户、微内容、应用基础上的高度网络化、自由化的互联网形态吸引了大量网络用户,衍生出诸如博客、播客、人际网络、网络文摘、维基百科等 Web 2.0 类应用。在 Web 2.0 时代,信息技术的发展使得网络用户广泛参与到信息资源的组织和描述活动中成为可能,用户不再仅仅是网络信息资源的消费者,同时也是信息资源的生产者、描述者、组织者。Flickr、del. icio. us、豆瓣网、和讯博客等网站都采用了协同标注[1]、大众分类法[2]让用户参与信息资源描述和组织,网站允许用户自由使用标签对文章、图片、视频、声音等对象进行描述,并利用这些用户标签完成信息资源的分类、组织、检索。然而,与传统的信息资源分类、组织方法相比,采用协同标注对信息资源进行描述、分类、组织、检索过程中存在着信息描述精确度不高、标签组织混乱等缺陷。本文将以豆瓣网的标签为研究数据来源,试图以凝聚式层次聚类方法将用户标签进行聚类,从而实现对用户标签的重新组织,为用户提供更好的标签导航、浏览机制。

收稿日期: 2007 - 12 - 03
收修改稿日期: 2007 - 12 - 08
* 本文系教育部人文社会科学重点研究基地重大项目"网络环境下数字化信息服务研究"(项目编号:06JJD870006)的研究成果之一。

2 研究背景

2.1 大众分类法与标签

大众分类法(Folksonomy)是美国信息架构专家Thomas Vander Wal和Gene Smith于2004年首先提出,由Folk和Taxonomy两个词合成而来,含义是"由大众的一致意见而产生的基于用户的分类体系"[3],中文翻译为"通俗分类"、"大众分类"、"自由分类"等。与传统结构严谨的登记体系分类法、复杂庞大的叙词法、分面分类法以及网站预设的信息分类组织体系不同,此种分类法并没有预先定义任何信息分类方法和词表,完全是用户根据个人的使用习惯,以自定义的自由词对数字资源对象进行标注和分类。

标签(Tag)是网络用户用于描述某个信息资源所使用的字、词或短语,是大众分类法的一个核心。大众分类法正是采用一种模糊化、智能化的"散秩分类"方式,鼓励大众本着自己的需要和个性化理解,对文字、图片、网页、视频等信息资源使用标签进行标注,通过互联网用户的大量交互以及相关的内容匹配,从而实现有效的信息检索和信息传播。

2.2 标签的局限性

标签作为新一代互联网形态Web 2.0的核心应用之一,实现了大众分类的思想,它的推广和应用将会引发评价标准和分类的多元化,促进人们的远程交流,有利于社会网络的形成。近年来,Flickr、del. icio. us、Technorati等国外的网站成功地将Tag运用于对图片、网络文摘等内容的描述、分类中。2005年,中文博客服务商BlogBus在国内最早引入了Tag理念,"博客中国"(Blogchina)则率先设计开发了Tag搜索引擎,搜狐、新浪等网站也广泛运用了Tag。然而,利用标签实现信息描述和检索还存在许多问题。

(1)标签描述信息的精确度低

首先,标签作为一种原生态的自然语言,其固有语义模糊性、同义词、多义词等特性,这使得难以让大众按照统一规范的语词对信息资源进行描述。再则,对于同一个信息资源,不同的用户由于关注点、兴趣爱好不同,会采用不同的标签对其进行标注,实际上,同一个用户在不同的时间也有可能对同一个信息资源采用不同标签进行标注。换句话说,单一用户难以选择一个合适的标签准确地对某一个信息资源进行分类表述。

(2)标签检索结果相关度低

标签难以按照统一规范的语词对信息资源进行描述,导致使用标签检索到的信息,查全率和查准率都比较低。实践表明,在RSS、ATOM等信息聚合工具运用中都能根据用户自定义标签实现信息聚合,检索到信息的相关度很低。

(3)标签无序、缺乏有效组织

现在对标签的组织主要有利用标签搜索引擎和利用可视化的标签空间工具(如标签云)两种方式[4],这两种方式在一定程度上实现了标签和网络信息的有序组织,但标签仍然太凌乱和缺乏有效组织。

2.3 标签有序化研究

现阶段对于标签系统进行优化的研究主要集中于标签云[5,6]、标签的有序化组织[6,7]、垃圾标签识别与清除等方面,本文主要关注标签有序化研究。Begelman等人论述了现有标签系统存在的缺陷,提出采用聚类技术对大量标签进行聚类,从而提高系统的导航和检索性能。研究发现标签同时出现的频率会在某一个临界点显著变化,将这个临界点作为阈值,来确定两个标签是否相关。利用这种相关关系构建成一个加权无向图,图的顶点表示标签,边线的权重表示同时出现的次数。采用"Spectral bisection Clustering"算法对标签进行聚类分析,从RawSugar数据库中采集了200 000个页面和30 000个标签进行实验,实验表明通过对标签进行聚类,可以提高用户的标注经验和标签的组织。如在实验系统中检索"health",将能检索到shopping、nutrition、fitness、science等相关标签,但这些标签是杂乱的,通过对相关标签进行聚类,可以得到如图1所示的结果[4]:

Query tag: health
—shopping, research
—nutrition, food, diet
—fitness, workout, running
—article, science
—life, lifehack, product, howto, get

图1 "health"相关标签聚类结果

Heymann和Garcia – Molina试图构建标签的层次分类,他们从del. icio. us上搜集了60 000个标签,根据标签的向量相似度确定相关标签,将相关标签连接成无权重的无向图,采用相关算法将无向图转换为层次

结构的分类树[7]。

已有研究成果表明,对标签进行聚类,有助于实现标签的有序化组织。本文在综合上述学者研究的基础上,采用凝聚式层次聚类(Hierarchical Agglomerative Clustering,HAC)算法来探讨标签聚类的有效方法。

3　基于凝聚式层次聚类的标签聚类

3.1　聚类分析

聚类(Clustering)是分类的一种,是分类的逆向方法,它采用的分类规则是统计学的聚类分析方法[8]。聚类分析作为统计学的一个分支,已被广泛地研究了多年,主要集中在基于距离的聚类分析。基于 K 均值、K – medoids 和其他一些方法的聚类分析工具已经被加入到许多统计分析软件包或系统中。在机器学习领域,聚类是无指导学习的一个例子,与分类不同,聚类和无指导学习不依赖预先定义的类和带类标号的训练实例。在概念聚类中,一组对象只有当它们可以被一个概念描述时才形成一个簇,这不同于基于几何距离度量相似度的传统聚类。常见的聚类分析的方法有 4 类:划分方法、层次方法和基于密度的方法、基于网格的方法[9]。聚类分析常用于对大量文本聚类进行数据挖掘,本文试图将聚类分析技术应用于标签聚类中。

3.2　标签聚类的基本思想

标签聚类与传统的文本聚类存在很大差别,传统的文本聚类中,由于文本包含的信息量大,可以采用文档频度、信息增益、X^2 统计、互信息、给定词浓度等方法提取文本特征[10,11],构建文本矩阵,实现文本聚类。而标签一般都是单个的字、词或短语,包含的信息量很小,计算机难以自动抽取其特征。本文中将采用如图 2 所示的方法,借助标签系统中已有的标签,根据其他用户的标注,进行特征抽取,然后采用 HAC 算法实现标签的聚类,生成树状层次结构视图。如在豆瓣网中标签“小说”、“文学”、“数学”可以分别用“外国文学、文学、爱情、外国小说、美国、中国文学、中国、名著”、“小说、外国文学、中国、中国文学、散文、欧美、诗歌、名著”、“数学分析、Math、数学相关、科普、好玩的数学、概率论、经济数理基础、数学史”这样一些相关标签进行描述,可以看出“小说”和“文学”都有“外国文学”、“中国文学”相关标签,而“数学”与“小说”、“文学”没有相同的相关标签,根据相关标签计算它们之间的相关度,再进行聚类生成树状层次结构视图。

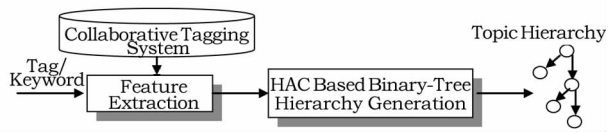

图 2　标签聚类系统示意图

3.3　标签特征抽取

在传统的信息检索模型中,特征项是指文档中能够代表文档性质的基本语言单位(如字、词),也就是通常所指的检索词。而标签本身就是单独的字或词,难以对其再进行抽取。现有的标签系统往往根据用户使用标签的频率提供相关标签,如在豆瓣网中,与“小说”相关的标签有“外国文学”、“文学”、“爱情”、“外国小说”、“美国”、“中国文学”、“中国”、“名著”,这些相关标签在一定程度上对“小说”这一标签的概念进行了描述,但系统中相关标签采用平行视图,难以精确展示标签之间的关系,不能为用户提供良好的导航视图。笔者将在已有的标签系统中,检索与某一个标签相关的标签,抽取与检索词同时出现的标签为描述项,采用传统空间向量模型表示,计算标签之间的相关程度,从而实现标签聚类。具体实现步骤如下:

①对于标签 p,检索相关标签,并对相关标签进行预处理,得到 n 个相关标签,以 $D_p=(t_1,t_2,\cdots,t_n)$ 表示;

②采用空间向量模型来表示特征选项,$V(d)=(t_1,w_1(d);\cdots;t_i,w_i(d);\cdots;t_n,w_n(d))$,其中 t_i 表示第 i 个相关标签,$w_i(d)$ 表示 t_i 在向量 V(d) 中的权重,采用传统的 TF – IDF 权重计算方法来计算每个相关标签在相应向量中的权重,计算公式见公式(1):

$$w_{t,i}(d)=(1+\log f_{t,i})\times\log(N/n_i+0.01) \quad (1)$$

其中 $f_{t,i}$ 表示 t_i 在 V(d) 中的频率,N 表示标签总数量,n_i 表示 t_i 在相关标签向量中出现的次数;

③计算标签向量相关度,以 $sim(v_a,v_b)$ 表示向量 v_a 和 v_b 的相关度,计算公式见公式(2):

$$sim(v_a,v_b)=\cos(v_a,v_b) \quad (2)$$

④计算向量集合的平均相关度,$sim_A(C_i,C_j)$ 表示标签集 C_i 和 C_j 的平均相关度,计算公式见公式(3):

$$sim_A(C_i,C_j)=\frac{1}{|C_i||C_j|}\sum_{v_a\in C_i}\sum_{v_b\in C_j}\cos(v_a,v_b) \quad (3)$$

3.4　凝聚式层次聚类算法

凝聚式层次聚类是一种常用的文本层次聚类方法,它自底向上分析目标文档,每个目标文档最初被看

成一个最小的聚类,两个相似度最大的聚类被合并为一个最大的聚类,这个过程一直持续到所有文档合并为一个聚类或者达到一个终止条件。其实现的基本思想如下[12,13]:

①计算 n 个文档两两之间的相似性 S_{ij},记为初始相似性矩阵;

②初始构造 n 个簇类,每个文档自成一簇类;

③合并距离最近的两个类为一个新类;

④计算新类与当前簇类的相似性,更新相似性矩阵,若簇类数等于 1 或达到一个停止准则,则跳到步骤⑤,否则跳到步骤③;

⑤选定分类阈值,决定簇类的个数和簇类。

所谓停止准则是指用户提前定义聚类数目或者聚类程度,当达到这个数目或程度时,聚类自动停止。具体过程如下:

①将文档集 $D=\{d_1,\cdots,d_i,\cdots,d_n\}$ 中的每一个文档 d_i 看作是一个具有单个成员的簇类 $C_i=\{d_i\}$,这些簇类构成了 D 的一个聚类 $C=\{C_1,\cdots,C_i,\cdots,C_n\}$;

②计算 C 中每对簇类 (C_i,C_j) 之间的相似度 $sim(C_i,C_j)$,形成初始相似性矩阵 S;

③选取具有最大相似度的簇对 $max(sim(C_i,C_j))$,并将 C_i 和 C_j 合并为一个新的簇类 $C_k=C_i\cup C_j$,从而构成 D 的一个新的簇类 $C=\{C_1,\cdots,C_{n-1}\}$,更新相似矩阵;

④重复上述步骤,直到 C 中只剩下一个类或者满足停止规则为止。

4 标签聚类实验

4.1 数据采集

实验从豆瓣网(www. douban. com)采集数据,选择了 Java、Python、编程、软件 4 个标签,分别用 P1、P2、P3、P4 表示,从豆瓣网检索相关标签,对检索数据进行处理,选择出与查询标签最相关的前 20 个标签,如表 1 所示。

4.2 相关标签权重计算

采用的公式(1)计算相关标签的权重,计算结果如表 2 所示。

表 1 前 20 个最相关标签表

P1 相关标签	频率	P2 相关标签	频率	P3 相关标签	频率	P4 相关标签	频率
编程	155	编程	185	计算机	604	软件开发	312
J2EE	108	计算机	61	算法	443	编程	305
计算机	138	程序设计	37	C ++	234	计算机	285
重构	105	脚本	22	UNIX	212	设计模式	212
软件开发	97	技术	22	软件	209	项目管理	181
Spring	91	Oreilly	17	软件开发	200	Java	161
软件工程	78	程序语言	15	设计模式	175	交互设计	136
Hibernate	69	Django	12	C	158	重构	105
技术	65	网络	12	JavaScript	154	管理	95
设计模式	50	WxPython	12	程序设计	131	代码大全	83
Design	35	Cookbook	10	计算机科学	114	设计	74
虚拟机	41	计算机科学	10	经典	107	软件	73
JUnit	27	Web	9	重构	105	敏捷开发	67
Struts	21	Twisted	9	Python	99	Design	67
模式	21	软件开发	8	Java	96	UI	56
Web	18	Nutshell	8	代码大全	83	人月神话	49
Effective	17	教程	8	计算机系统	82	Code	49
Tomcat	15	GUI	7	Dom	82	程序设计	45
Framework	14	入门	7	Design	67	Agile	40
Eclipse	14	Design	6	算法导论	59	Pattern	38
JSP	13	开源	5	SICP	41	敏捷	37

表 2　相关标签权重表

Tag	P1	P2	P3	P4
Agile	0	0	0	0.1781
C	0	0	0.3633	0
C ++	0	0	0.538	0
Code	0	0	0	0.2181
Cookbook	0	0.0751	0	0
Design	0.0022	0.0003	0.0011	0.0021
Django	0	0.0901	0	0
Dom	0	0	0.1885	0
Eclipse	0.1254	0	0	0
Effective	0.1523	0	0	0
Framework	0.1254	0	0	0
GUI	0	0.0525	0	0
Hibernate	0.6182	0	0	0
J2EE	0.9677	0	0	0
Java	0	0	0.111	0.3603
JavaScript	0	0	0.3541	0
JSP	0.1165	0	0	0
JUnit	0.2419	0	0	0
Nutshell	0	0.0601	0	0
Oreilly	0	0.1276	0	0
Pattern	0	0	0	0.1691
Python	0	0	0.2276	0
SICP	0	0	0.0943	0
Spring	0.8154	0	0	0
Struts	0.1882	0	0	0
Tomcat	0.1344	0	0	0
Twisted	0	0.0676	0	0
UI	0	0	0	0.2493
项目管理	0	0	0	0.8057
虚拟机	0.3674	0	0	0
重构	0.1999	0	0.0513	0.0993
UNIX	0	0	0.4875	0
Web	0.0811	0.034	0	0
wxPython	0	0.0901	0	0
编程	0.2952	0.2952	0	0.2885
程序设计	0	0.059	0.064	0.0426
程序语言	0	0.1126	0	0
代码大全	0	0	0.0959	0.1857
管理	0	0	0	0.4229
计算机	0.0089	0.0033	0.01	0.0091
计算机科学	0	0.0377	0.1318	0
计算机系统	0	0	0.1885	0
技术	0.2928	0.083	0	0
交互设计	0	0	0	0.6054
脚本	0	0.1652	0	0
教程	0	0.0601	0	0
经典	0	0	0.246	0
开源	0	0.0375	0	0
敏捷	0	0	0	0.1647
敏捷开发	0	0	0	0.2982
模式	0.1882	0	0	0
人月神话	0	0	0	0.2181
入门	0	0.0525	0	0
软件	0	0	0.2416	0.1633
软件工程	0.6989	0	0	0
软件开发	0.0062	0.0004	0.0033	0.01
设计	0	0	0	0.3294
设计模式	0.0952	0	0.0855	0.2006
算法	0	0	0.9832	0
算法导论	0	0	0.1357	0
网络	0	0.0901	0	0

4.3　聚　类

根据相关标签权重表，分别计算标签之间的相关度，计算结果为 Sim(P1,P2) = 0.0153，Sim(P1,P3) = 0.0296，Sim(P1,P4) = 0.0997，Sim(P2,P3) = 0.0279，Sim(P2,P4) = 0.2714，Sim(P3,P4) = 0.1191，聚类结果示意图如图 3 所示。

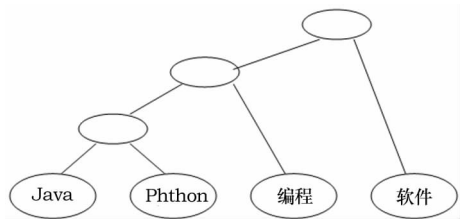

图 3　P1、P2、P3、P4 聚类结果示意图

5　结　语

本文提出采用层次凝聚算法对用户标签进行聚类，以达到为用户提供更好的标签导航机制。通过实验表明，文中所提出的方法能够完成标签聚类，对聚类结果进行可视化后，能够在一定程度上表现标签之间的语义关系。接下来，笔者将对下面的内容作进一步研究。

（1）完善标签权重算法，将用户数引入到权重计算中；

（2）对聚类结果进行优化，使其能更好的表达标签概念之间的语义关系；

（3）研究其他聚类方式在标签聚类中的应用，并比较不同聚类算法在标签聚类中的效率。

参考文献：

[1] Golder S, Huberman B. Usage Patterns of Collaborative Tagging Systems[J]. *Journal of Information Science*, 2006(2): 198 - 208.

[2] Hammond T, Hannay T, Lund B, et al. Social Bookmaking Tools (I): a General Review[EB/OL]. [2007 - 12 - 05]. http://www.dlib.org/dlib/april05/hammond/04hammond.html.

[3] Mathes A. Folksonomies - Cooperative Classification and Communication Through Shared Metadata[EB/OL]. [2007 - 11 - 10]. www.adammathes.com/academic/computer - mediated - communication/folksonomies.html.

[4] Begelman G, Keller P, Smadjia F. Automated Tag Clustering: Improving Search and Exploration in the Tag Space[C]. In: *Collaborative Web Tagging Workshop*, 15*th International World Wide Web*

Conference, Edinburgh, UK, May 22 – 26, 2006.

[5] Speroni. Tagclouds and Cultural Changes[EB/OL]. [2007 – 11 – 10]. http://blog. pietrosperoni. it.

[6] Owen K, Daniel L. TagCloud Drawing: Algorithms for Cloud Visualization[C]. In: *proceedings of Tagging and Metadata for Social Information Organization* (*WWW*2007), 2007.

[7] Heymann P, Garcia – Molina H. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems[EB/OL]. [2007 – 11 – 10]. http://dbpubs. stanford. edu:8090/pub/2006 – 10.

[8] 孙建军, 成颖. 信息检索技术[M]. 北京:科学出版社,2004: 201 – 202.

[9] 张建辉. K_means 聚类算法研究及应用[D]. 武汉:武汉理工大学,2007.

[10] Yang Y, Pedersen J P. Feature Selection in Statistical Learning of Text Categorization[C]. *In the* 14*th Int. Conf. On Machine Learning*, San Francisco, 1997.

[11] Chuang S L, Chien L F. Taxonomy Generation for Text Segments: a Practical Web – based Approach[J]. *ACM Transactions on Information Systems*, 2005,23(4): 363 – 369.

[12] Chuang S L, Chien L F. Towards Automatic Generation of Query Taxonomy: A Hierarchical Query Clustering Approach[C]. In: *Proceedings of the* 2002 *IEEE International Conference on Data Mining*, Maebashi City. Japan: IEEE Computer Society Press, 75 – 82.

[13] Brandes U, Gaertler M, Wagner D. Experiments on Graph Clustering[C]. In: *Proceedings of the* 11*th Annual European Symposium on Algorithms* (ESA'03), volume 2832 of Lecture Notes in Computer Science, 2003:568 – 579.

(作者 E – mail: ghcao@ mail. ccnu. edu. cn)