基于概率主题模型的标签预测

袁 柳[1] 张龙波[2]
(陕西师范大学计算机科学学院 西安 710062)[1] (山东理工大学计算机学院 淄博 255049)[2]

摘 要 充分利用用户自定义标签信息,是理解Web资源语义,提高Web应用智能程度的重要途径。针对资源标签分派中大量存在的信息不完整、不一致的现象,建立基于用户标记行为特征的概率主题模型,利用概率主题模型实现对标记信息不完整资源的标签预测。根据每个资源所对应的标签的统计特征,可产生不同形式的标签文档,通过分析标签文档所生成主题的性能,确定适合于特定数据集的标签文档形式;利用同一主题内词汇间的高度相关性,设计合理的预测标签排序方法,从而实现对标记信息不完整资源的标签预测以及标签语义不一致现象的检测。在数据集DeliciousT 140和Wiki10+上的测试表明,所提方法能有效实现标签预测,并可提高信息检索的性能。

关键词 标签系统,标签预测,统计主题模型
中图法分类号 TP311.1 **文献标识码** A

Social Tag Predication Based on Probabilistic Topic Model

YUAN Liu[1] ZHANG Long-bo[2]
(College of Computer Science, Shaanxi Normal University, Xi'an 710062, China)[1]
(School of Computer Science and Technology, Shandong University of Technology, Zibo 255049, China)[2]

Abstract Tagging information created by users is important to understand the Web resource semantics and to improve the intelligence of Web applications. Probabilistic topic model was exploited to deal with the incompleteness and inconsistence of tagging systems. A probabilistic topic model generating technique based on tag statistical characteristics was proposed. According to tag statistical characteristics of each resource, tag documents with different format can be created. By analyzing the performance generated by different tag documents, document format that is appropriate for a certain dataset was confirmed. High relatedness between the vocabularies in the same topic was exploited to predicate the tag for resources with incomplete and inconsistence tags. Experiments on DeliciousT 140 and Wiki10+ show the effectiveness of the technique proposed.

Keywords Tagging system, Tag predication, Statistical topic model

1 引言

标签系统(Tagging Systems),例如Flickr,Del.icio.us,CiteULike等已成为最重要的Web应用之一[1-3]。用户通过标记(Tagging),可以为任意类型资源添加标签,即在任意资源(如文档、图像、视频等)和一组特定的词汇或短语之间建立关联。目前标签的作用主要体现在两方面:一是协助用户管理和组织自己生成的资源;二是从用户共享的资源中检索自己感兴趣的内容。因此标签可理解为对资源的语义标注,可用于Web搜索等。由于用户在标记过程中不受任何限制,并且建立标签的目的也各不相同,因此在标签系统中普遍存在着标记不完整、标签词汇过于通用或个性等现象。为了充分、有效地利用标签资源,需要处理标签系统中的标签不完整、语义不清晰等现象,通过分析较为完备的标签集合,发现标签间的内在联系,实现对资源标签的预测及对资源的有效检索。

本文将就资源标签数量过少、标注不完整等现象展开研究,将概率主题模型用于标签数据的分析。利用LDA(Latent Dirichlet Allocation)[4]模型从具有相对稳定、完整的标签集合的资源中获得隐含主题,根据隐含主题所包含的出现频率较高的词汇预测新资源的标签。与已有的相关研究不同,本研究从用户标签行为的特征分析入手,进而深入分析利用LDA进行处理的标签文档的表示形式,证明不同的标签描述形式会对预测结果产生重要影响。在此基础上提出一种基于标签行为特征的标签预测方法,并将所提方法用于提高Web资源检索性能。

2 相关工作

标签预测是Web 2.0环境下的研究热点之一。计算待预测资源与已标记资源间、用户之间以及标签之间的相似性,是解决该问题最为直观的方法。文献[5]根据同一用户所标记的Weblog间在内容、形式等方面的相似程度,为新的Weblog推荐标签;文献[6]通过检索内容相近的文档标签和

到稿日期:2010-08-03 返修日期:2010-12-06 本文受国家自然科学基金项目面向入侵检测的数据流挖掘研究(60873196)资助。
袁 柳(1979-),女,博士,讲师,主要研究方向为Web数据管理、语义信息检索,E-mail:yuanl@mail.nwpu.edu.cn;张龙波(1968-),男,博士,教授,主要研究方向为数据流与数据挖掘。

分级的方式实现对 Web 资源的标签预测。这类方法需要提取和分析被标记对象的特征。在标记对象类型复杂多样、数量巨大的 Web 环境下，设计合理的相似性计算算法并不容易，因此相似性计算无疑是一项艰巨的任务，并且计算结果的准确程度难以保证。一些研究考虑利用标签间的 co-occuring 关系进行标签预测[7]，但 co-occuring 需要满足一定的阈值才能保证预测结果的可靠性。高阈值尽管可以保证预测结果的质量，但实际的标签数据集中，大量存在的同义词、多义词、词形变化等现象对 co-occuring 的计算影响很大，标签间存在的隐含的语义关系往往被忽略。关联规则挖掘也被用于预测资源标签[8]，该方法所产生的预测结果十分直观，并具备可靠性，但是所挖掘出的关联规则有相当一部分描述的是词汇间“IS-A”类型的关系或因为词形变化而产生的词汇之间的关联[10]。这类词汇间的关联利用 WordNet 或概念分类树等途径可较为容易地获得，但不能反映用户自定义标签间隐含的内在联系，就标签预测任务而言，这类关联并无很大意义。一些研究考虑通过多种预测技术相结合的方法[11,12]来提高预测的准确性，这种策略的有效性已被验证，但同时增加了预测技术的复杂性。

通过分析用户标签集合中蕴含的语义来发现标签词汇间的关联，是解决标签预测问题的一条新思路。近期有研究开始关注将概率主题模型用于标签预测中[10]，将资源看作 LDA 模型中的文档，文档由与该资源相关的标签组成。在一个标签系统中，用户的标记行为特征，即用户、资源、标签之间关系的表现特征往往不尽相同，单纯利用资源及其所对应的原始标签会导致一些重要的用户行为特征信息的丢失，生成的文档并不能产生良好的隐含主题。目前还未见到研究用户行为特征在基于概率主题模型的标签预测过程中作用的成果，本文将就此展开研究。

3 标签预测

3.1 问题定义

设已给定资源集合 R、标签集合 T 以及用户集合 U，用户对特定资源的标签分派可描述为一个三元关系 X：$X\subseteq R\times T\times U$。对于资源 $r_i\in R$ 和用户 $u_j\in U$，书签 $b(r_i,u_j)$ 表示用户 u_j 分派给资源 r_i 的所有标签，且满足 $\bigcup_{\substack{u_j\in U\\ r_i\in R}} b(r_i,u_j)=T$，以关系代数运算符的形式可描述为 $b(r_i,u_j)=\pi_t\sigma_{r_i,u_j}X$[9]。标签预测的目标就是在分析标签信息相对完整的资源-标签分派集合 $Y=\sigma_{r\neq r_i}\pi_{r,t}X\subseteq R\times T$ 的基础上，为标签数量较少的资源预测新的标签，使其能够对资源进行较为全面的描述，以实现对资源的有效利用。本文研究设待预测资源的集合为 R_p，则 $R_p=\{r\mid r\in R,b(r,u_j),j\in\{1,\cdots,n\},n\leqslant 5\}$，即少于 5 个用户标记的资源都进行标签预测。

图 1 以 Delicious T140 数据集为例，描述了用户标签的主要特征，同时对标签预测任务进行具体说明。图 1(a)展示了将资源具有的标签数量按升序排列后资源-标签数量的散点图，x 轴为以标签数量升序排列的资源，y 轴表示对应资源的标签数量。可以看出，绝大多数资源标签数量在 5～24 之间。右上方的点代表标签数量较多的资源。就标签预测任务而言，这部分资源显然不是进行标签预测的对象。左下方的点则表示标签数量相对较少的资源，标签预测的任务就是为这些资源发现新的标签。图 1(b)为以 log-log 形式描述的标签-出现频率分布图，x 轴为以频率(本文以标签出现次数表示)升序排列的 66054 个不同的标签，y 轴为标签对应的出现频率。右上方的点代表了出现频率极高的标签，但就标签预测任务而言，这些标签的描述能力可能过于一般、过于通用而不具备在预测中的使用价值。左下方的点为出现频率极低的标签，一般为一些偶然出现的单词，如拼写错误、复杂短语等，这些对标记预测的意义也不大。在该数据集中，约 40%的标签出现频率少于 5 次。例如使用“*，+，2，3，a2，4－q1”等简单且意义不明确的标签对预测任务来讲是没有价值的。图 1(c)以 log-log 形式描述了资源-用户数量间的关系，x 轴以标记资源的用户数升序排列。图中左下方标记用户少于 5 个的资源是进行标签预测的对象。在该数据集包含的 144,574 个资源中，约 5%的资源少于 5 个用户进行标记。

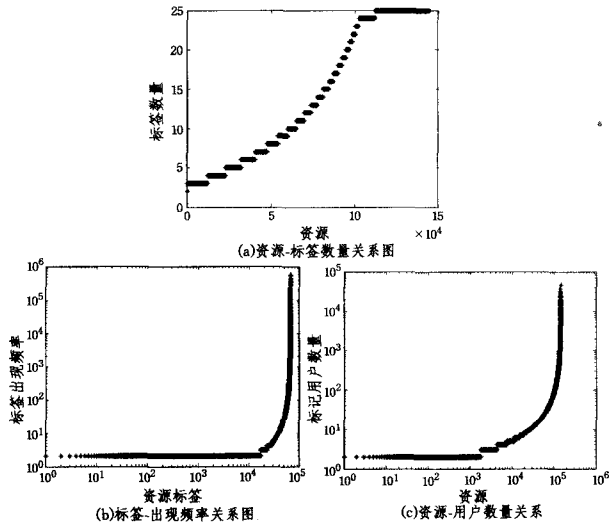

图 1 Delicious T140 中的标签行为描述

3.2 资源标签特征描述

如上所述，目前标签研究领域所广泛采用的三元关系 $X\subseteq R\times T\times U$ 从资源、标签、用户 3 者之间是否存在关联的角度对标签系统进行描述，但对关联的具体表现形式及特征没有更深的研究。在标签集合中存在着关于用户标记行为的重要的信息，如图 1 所示的资源-标签数量关系、标签-出现频率关系、资源-用户数量关系等，这些信息对于预测有着重要意义。如上分析，不是所有标签信息都适合预测任务。事实上，这些信息可以描述为用户标记行为的统计特征。表 1 为本文所考虑的标签系统中的统计特征列表。由于本文重点关注标签系统中资源-标签间的关联，因此忽略了对用户个人行为相关数据的统计，如单个用户所标注的资源数量、所使用的标签数量等，这也避免了一些过于个性化、不合常规的个人标记行为对整体分析的影响。

表 1 标签统计信息

符号	含义
NR	资源数量
NU	标记资源的用户总数
NT	标记资源的标签总数
N_t	标签 t 出现的总次数
N_a	资源集合中的标记总数
N(r,t)	使用标签 t 对资源 r 进行标记的次数
N(r,u)	标记资源 r 的用户数
α	每个标记平均使用的标签数量
β	每个资源平均分派的标签数量

在执行标签预测任务时，首先根据表1中所列出的条目分析数据集，再根据分析结果确定标签预测任务的对象、适合于预测任务的技术、算法中的参数等，从而做到依据数据集特征进行针对性的预测。已有一些研究也考虑到对标签统计特征进行分析[12]，但都忽略了这些特征与预测算法间的关系。

3.3 LDA模型与标签预测

3.3.1 LDA模型

统计主题模型的最大优势在于，在不需要计算机真正理解自然语言的情况下，可以从文档中提取被人理解的、相对稳定的隐含语义结构。LDA是一个多层的产生式概率模型，包含词、主题和文档3层结构。LDA将每个文档表示为一个主题混合，每个主题是固定词表上的一个多项式分布。LDA假设词由一个主题混合产生，同时每个主题是在固定词表上的一个多项式分布；这些主题被集合中的所有文档共享；每个文档有一个特定的主题比例，从Dirichlet分布中抽样产生。

一个LDA主题反映了某一概念范畴内的事物特征，即主题中所包含的词汇之间具有较强的相关性，而主题间的词汇相关性较弱。因此，如果标记资源的标签属于某个主题，则这个主题中的其他词汇可用于标记该资源的可能性就较大。据此就可以对标签信息贫乏的资源实现标签预测。

3.3.2 标签文档的生成

可以看出，产生LDA主题的表达能力直接决定了基于主题的标签预测的性能，因此使用LDA进行标签预测的首要任务是将资源集合R中所使用的标签以一种恰当的、LDA可处理的文档形式表示。在标签系统中，每个资源$r\in R$都可看作一篇文档，文档集合D由描述资源r的所有标签组成，即$D=[r_1,r_2,\cdots,r_n]^T$，r_i可描述为行向量$r_i=[t_{r_i,1},t_{r_i,2},\cdots,t_{r_i,l_i}]$，$t_{r_i,j}\in T$。其中，$n$表示资源数目，$t_{r_i,j}$表示标记资源$r_i$的第$j$个标签，$l_i$表示集合$D$中第$i$个文档的长度。根据表1所列的标签统计特征，文档$r_i$可有以下几种形式：

(1)$r_i=[t_{r_i,1},t_{r_i,2},\cdots,t_{r_i,l_i}]$，$t_{r_i,j}\in T$，$t_{r_i,j}\neq t_{r_i,k}$，且对于$\forall t\in\bigcup_{\substack{u_j\in U\\ r_i\in R}}b(r_i,u_j)$，$t$都出现在$r_i$的向量表示中。这意味着文档$r_i$中的标签互不相同，且只要存在用户使用标签$t$标记该资源，$t$就会出现在标签向量中；同一个标签被不同的用户多次使用来标记同一个资源，或者某标签被某个用户使用过一次，这种文档表示形式将不能反映。

(2)$r_i=[t_{r_i,1},t_{r_i,2},\cdots,t_{r_i,m_i}]$，$t_{r_i,j}\in T$，且对于$\forall t\in\bigcup_{\substack{u_j\in U\\ r_i\in R}}b(r_i,u_j)$，$t$都出现在$r_i$的向量表示中。这种表示允许文档$r_i$中有重复的标签，即若同一个标签$t$被不同用户多次使用，则$t$在文档中出现的次数为用户使用该标签进行标记的次数$N(r_i,t)$。

(3)$r_i=[t_{r_i,1},t_{r_i,2},\cdots,t_{r_i,l_i}]$，$t_{r_i,j}\in T$，$t_{r_i,j}\neq t_{r_i,k}$，$N(r_i,u)>5$，这种表示方法在(1)的基础上增加约束，即对于少于5个用户标记的资源，不会将其作为LDA的分析对象。

(4)$r_i=[t_{r_i,1},t_{r_i,2},\cdots,t_{r_i,m_i}]$，$t_{r_i,j}\in T$，$N(r_i,u)>5$。该表示方法在(2)的基础上增加约束，对于少于5个用户标记的资源，不会将其作为LDA的分析对象。

以上几种标签文档表示方式的差别根本上在于是否考虑了标签信息的统计特征：如标签出现频率、与资源相关的标签数量、标记资源的用户等信息。为了具体展示这些差别对生成隐含主题的影响，本文从Delicious T140中随机选取部分资源标签，表2(a)、(b)分别是在文档表示形式(1)、(2)的情况下所产生的与概念“finance”相关的LDA主题举例。

表2(a) 文档表示形式(1)所生成的主题

主题词汇	出现概率
finance	0.020238266
opensource	0.008755561
design	0.007320223
planning	0.007320223
free	0.005884885
webdesign	0.005884885
resources	0.005884885
guide	0.005884885
webcam	0.005884885
economics	0.005884885
sql	0.005884885
computers	0.004449547
blogs	0.004449547
xml	0.004449547
tv	0.004449547
rss	0.004449547
tips	0.004449547
iso	0.004449547
installation	0.004449547
freeware	0.003014209

表2(b) 文档表示形式(2)所生成的主题

主题词汇	出现概率
finance	0.185078066
money	0.172424701
economics	0.087240436
investing	0.049506292
banking	0.030978150
business	0.028266714
projectmanagement	0.027588856
personal	0.025103373
stocks	0.022391937
management	0.021036220
project	0.020358361
poverty	0.015161443
trading.	0.015161443
financial	0.013353819
investment	0.010642384
analysis	0.007930948
sociology	0.007704995
richest	0.007253089
budget	0.006575231
wealth	0.006349278

观察表2可明显看出，标签文档形式(2)包含了标签出现的频率信息，该信息反映了标签的权重信息(即对于同一个资源，使用同一个标签进行标记的用户越多，说明该标签描述该资源的能力越强)，从这种形式的文档中获得的主题内词汇间的内聚性、关联程度更强，这意味着主题所能描述的概念更准确、集中，如表2(b)所列。而表2(a)主题所描述的概念就较为分散，难以确定该概念所属领域。

表3是Delicious T140数据集标签文档中生成的一隐含主题。观察其中词汇可发现，在同一主题中存在一些词汇具有完全相同的含义，但由于用户定义标签时采用的词汇不受约束，表达相同含义的词被当做不同的词汇处理，例如. net，dotnet等。为了消除这种相同含义的标签由于描述方式的差异对LDA主题生成的影响，本文从数据集中选取了一组出现频率

较高、语义相同但词汇形式不同的标签，对其进行统一化处理。表 4 列出了对该数据集进行形式统一的部分标签。经过统一化的标签文档发现的隐含主题的概念描述能力更为准确。

表 3 ".net development"相关的主题

LDA 主题词汇(.net development)：
;ruby,rails,asp.net,subyonrails,.net,development, programming,ecommerce,excel,c#,facebook,web,gmail,dotnet,validation, actionmailer,iis,paras,deployment,payment

表 4 标签统一化说明

同义标签	统一化标签
.net,dotnet	.net
code,codes,coding	code
advertising,ads,ad	*advertising*
lists,list	list
networking,network,networks	network
apps,application,applications	application
templates,template	template
tools,tool	tool
algorithms,algorithm	algorithm
camera,cameras	*camera*
computers,computer	computer
Mac,Macintosh	Macintosh
extensions,extension	extension
bookmarks,bookmark	bookmark
authors,author	author
books,book	book
downloads,download	download
reviews,review	review
photos,photo	photo
images,image	image
articles,article	article
movies,movie	movie
graphics,graphic	graphic
webdevelopment,webdev	webdev
resources,resource	resource
plugins,plugin	*plugin*
bloggind,blog	blog
humour,humor	humor

3.3.3 标签的聚合与排序

从标签文档所生成的隐含主题中选择与已有资源标签最相关的标签词汇，是实现标签预测的最后步骤。设 C_t 为与用户自定义标签 t 对应的候选标签集合，其中包含了与 t 最相关的标签。根据 LDA 模型从标签文档中获取的主题，对于 $\forall t\in T$，可能存在多个主题都包含标签 t。设 $Z^t=\{z_1^t,z_2^t,\cdots,z_m^t\}$ 表示 t 所属的主题集合，$p(t|z_i^t)$ 表示标签 t 在主题 Z_i^t 中出现的概率，$tag(Z_i^t)$ 表示包含在主题 Z_i^t 中的标签，则

$$C_t=\{t_c|t_c\in tag(z_i^t),z_i^t\in Z^t,p(t_c|z_i^t)>\sigma\}$$

σ 为主题中标签出现概率的阈值，即标签 t 的候选集合为以较大概率值出现在主题中的若干个标签。

设 P_r 为与特定资源 r 相关的标签预测集合，对于任意资源，可能存在多个自定义标签，则根据每个标签所对应的候选集合 C_t，可确定该资源的标签预测集合 P_r。本文在计算标签预测集合时同时考虑标签的频率特征。如上所述，对于出现频率过高或过低的标签，其具备的描述资源特征的意义并不大，是不可靠的，本文处理时予以忽略。

综上所述，对于资源 r_i，有标签集合 $T_{r_i}=\{t_1,t_2,\cdots,t_k\}$，$C_{t_i}$ 为标签 t_i 对应的候选集合。最终的预测标签的生成通过对标签候选集合的聚合实现。本文中，设定 $P_{r_i}=C_1\cup C_2$，其中 $C_1=\bigcup_{\substack{t_i\in T_{r_i}\\ t_j\in T_{r_j}}}(C_{t_i}\cap C_{t_j})$，即同时属于两个以上标签候选集合的标签；$C_2=top_k(\bigcup_{t_i\in T_{r_i}}C_{t_i})$，即所有标签候选集合的并集中概率值最高的前 k 个标签。

值得注意的是 C_1，C_2 可能会包含相同的标签。事实上，C_1 是一项类似于 co-occurring 关系但强于 co-occurring 的标签预测指标，从标签语义即标签所属隐含主题的层面选择所预测的标签。若 C_1 所能提供的标签数量不能满足标签预测任务的需求，则可选择 C_2 中的标签作为预测结果。

3.4 基于 LDA 和标签特征的标签预测方法

综上所述，利用 LDA 主题中词汇的相关性预测资源标签的步骤如图 2 所示。

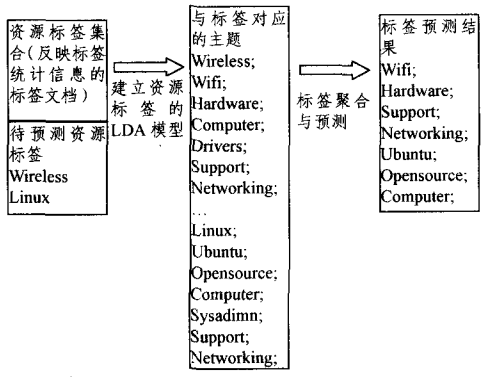

图 2 基于 LDA 和标签特征的标签预测步骤

首先，根据用户对资源所定义的标签，建立资源集合的标签文档表示；

然后，利用 LDA 模型从标签文档中发现隐含主题，并从每个主题所包含的标签中确定可用于标签预测的标签集合；

最后，对不同 LDA 主题所包含的标签进行聚合，生成最终的预测标签集合。

4 实验

4.1 数据集与评价标准

本文分别利用 DeliciousT 140[13] 和 Wiki10+[14] 两个数据集验证所提出的方法。DeliciousT 140 收集了 Delicious 站点 2008 年 6 月份的用户书签记录，包含 144,574 个不同的 URL，每个 URL 都有其所对应的标签组，共计约 67,000 个不同的标签。该数据集中包含了被描述文档 URL、文件类型、标记资源的用户数量、标记资源的标签及使用该标签的用户数量等信息，这些信息以 XML 文件的形式描述，约 197MB。Wiki10+ 收集了 2009 年 4 月在 Delicious 上对 Wikipedia 中的文章进行标记的数据，包含 20,764 个英文 Wikipedia 文章的 URL 及其对应的在 Delicious 上的标签，每篇文章至少有 10 个标签，以 XML 文件形式描述，共约 38MB。DeliciousT140 主要用于验证标签统计特征对基于 LDA 的标签预测方法的影响；Wiki10+ 主要用于验证基于 LDA 主题的标签分析在 Web 文档检索中的性能优势。

本文采用信息检索领域通用的性能评价指标 Precision/Recall 对所提出的标签预测方法进行评价，并以基于关联规则挖掘的标签预测方法作为比较基准。以训练集与测试集之比为 9:1 的比例分别对两个数据集进行随机分割，对测试集

中的标签仅保留随机选择的 3～5 个，将训练集中的资源-标签数据转化为 3.3.2 节所讨论的标签文档形式(4)。使用 JGibbsLDA[15]作为 LDA 模型的实现算法，发现文档中的隐含主题。实验中 JGibbsLDA 算法的主要参数以如下值设定：迭代次数 1000；主题数目 100；每个主题中所包含的标签词汇数目 50。这些参数的变化可能会对所发现的隐含主题有所影响，由于本文主要关注资源-标签信息的统计特征及 LDA 模型在标签预测中的实用价值，算法内部参数对结果的影响不是本文研究的重点，因此实验中对 LDA 算法参数对标签预测结果的影响不做深入的讨论。算法参数的设定以能够清晰地证明统计主题模型在标签预测中的作用为目的。

4.2 标签预测

表 5(a)是利用数据挖掘工具 Weka[16]提供的关联规则挖掘功能预测 Delicious T140 测试集标签的结果，设定最小支持度为 0.05，置信度分别在 0.1～0.9 之间取值，$N(r,u)$取值为 5。对标签文档词汇进行了统一化处理，尽管处理方式较为简单，但可明显减少语义上冗余的关联规则，如“humour→humor”，“photography,photos→photo”，“software,macintosh→mac”等包含同义标签的关联规则被有效地控制。表 5(b)是当支持度＝0.05，置信度＝0.9 时，利用关联规则对 $N(r,u)$值为 1～5，>5 的资源进行标签预测的结果。

表 5(a)　支持度＝0.05，$N(r,u)$＝5 时不同置信度条件下的关联规则预测结果

Conf.	Precision	Recall	F-Measure
0.9	0.618	0.082	0.145
0.7	0.533	0.162	0.249
0.5	0.401	0.201	0.268
0.3	0.322	0.303	0.372
0.1	0.288	0.412	0.339

表 5(b)　支持度＝0.05，置信度＝0.9 时不同 $N(r,u)$值条件下的关联规则预测结果

N(r,u)	Precision	Recall	F-Measure
1	0.726	0.048	0.090
2	0.698	0.057	0.105
3	0.674	0.064	0.117
4	0.632	0.077	0.137
5	0.618	0.082	0.145
>5	0.709	0.093	0.164

表 6(a)　$N(r,u)$＝5 时不同 σ 条件下的基于统计主题模型的预测结果

阈值 σ	Precision	Recall	F-Measure
0.01	0.699	0.183	0.290
0.005	0.603	0.277	0.379
0.001	0.412	0.355	0.381
0.0005	0.378	0.522	0.438
0.0001	0.384	0.563	0.456
0.00001	0.192	0.658	0.297

表 6(b)　σ＝0.01 时不同的 $N(r,u)$取值条件下基于统计主题模型的预测结果

N(r,u)	Precision	Recall	F-Measure
1	0.653	0.102	0.176
2	0.690	0.147	0.242
3	0.682	0.151	0.247
4	0.676	0.178	0.282
5	0.699	0.183	0.290
>5	0.727	0.201	0.292

表 6 是在不考虑标签资源统计特征的情况下构造标签文档，利用统计主题模型进行标签预测的结果。其中表 6(a)反映了当 $N(r,u)$取值为 5 时概率阈值 σ 的变化对预测结果的影响。表 6(b)是当 σ＝0.01 时不同的 $N(r,u)$取值条件下预测结果的变化。与关联规则预测结果进行对比可发现，使用统计主题模型对 F-Measure 值的提高是明显的。

表 7 在蕴含标签统计特征信息的标签文档基础上构造隐含主题，实现标签预测的结果。其中表 7(a)反映了当 $N(r,u)$取值为 5 时概率阈值 σ 的变化对预测结果的影响。表 7(b)是当 σ＝0.01 时不同的 $N(r,u)$取值条件下预测结果的变化。与表 6 的结果相比可以看出，由于合理的标签文档结构有助于生成语义上更明确的主题，因此基于隐含主题的标签预测可以获得更好的性能。

表 7(a)　$N(r,u)$＝5 时不同 σ 条件下的考虑标记行为特征的统计主题模型的预测结果

阈值 σ	Precision	Recall	F-Measure
0.01	0.807	0.212	0.336
0.005	0.721	0.293	0.423
0.001	0.544	0.401	0.461
0.0005	0.423	0.533	0.472
0.0001	0.434	0.558	0.488
0.00001	0.288	0.702	0.408

表 7(b)　表 6(b) σ＝0.01 时不同的 $N(r,u)$取值条件下考虑标记行为特征的统计主题模型的预测结果

N(r,u)	Precision	Recall	F-Measure
1	0.732	0.153	0.253
2	0.794	0.188	0.304
3	0.788	0.191	0.307
4	0.801	0.200	0.320
5	0.807	0.212	0.336
>5	0.819	0.223	0.351

表 8 反映了 σ＝0.01 时在不同的 $N(r,u)$取值条件下 LDA 主题数目的变化对标签预测结果的影响。可以发现，选取合适的主题数目，可提高预测结果的性能。在本实验数据集中，当主题数目为 300 时，结果相对较优，但计算所花费的时间也更多。本实验在计算时间耗费与结果质量间权衡后，对主题数目为 100 时的情况进行深入分析。

表 8　主题数目变化对预测结果影响

N(r,u)	LDA 主题数目			
	50	100	300	500
1	0.224	0.253	0.262	0.229
2	0.259	0.304	0.315	0.267
3	0.272	0.307	0.337	0.276
4	0.288	0.320	0.349	0.301
5	0.300	0.336	0.369	0.313
>5	0.309	0.351	0.378	0.317

4.3 基于标签的检索

本文使用 Wiki10+数据集验证标签预测在资源检索中的作用，表 9 展示了仅使用原始标签和利用标签预测技术进行资源检索的性能。以简单的关键词匹配作为检索成功的评判标准，即只要资源包含与用户检索条件词汇 t_a 相同的标签，则认为该资源可满足用户要求。对于仅使用原始标签的检索，将符合检索要求 t_a 的资源按 $freq(t_a,r)$值排序：

$$freq(t_a,r)=\frac{N(r,t)}{\sum_{t_i\in r}N(r,t_i)}$$

$freq(t_a,r)$反映了标签 t_a 标记资源 r 的频率。对于利用标签

预测结果的检索，以 t_a 在 LDA 主题中出现的概率值对符合要求的资源排序。

由于返回的检索结果已按与检索要求的相关性进行排序，故本文选择能够反映排序结果准确性的评价指标 MAP (Mean Average Precision)对两种不同方式的检索结果进行度量。

$$\mathrm{MAP}(Q)=\frac{\sum_{r=1}^{N}(P(r)\times rel(r))}{\text{数据集中相关文档的数量}}$$

式中，r 表示文档的排序，N 是检索到的文档数量，$rel()$表示给定排序相关性的函数，$P(r)=\frac{|\{d|rank(d)\leqslant r\}|}{r}$，$rank(d)$表示文档 d 在检索结果中的排序。分析满足检索要求的 top 20 个资源，判断检索结果。实验设计了 3 个查询问题，分别为“music AND China”，“travel AND Asia”以及“programming and Linux”。表 9 为对不同 $N(r,u)$值的资源使用不同方式进行检索的 MAP 值比较，表中数值为 3 个查询结果的 MAP 平均值。可以看出，利用标签预测可以使检索结果的 MAP 提高 3 倍以上。

表 9　标签预测对检索结果 MAP 值的影响

N(r,u)	MAP	
	仅使用原始标签的检索	利用标签预测的检索
1	0.032	0.121
2	0.045	0.143
3	0.052	0.163
4	0.059	0.177
5	0.068	0.183
>5	0.077	0.213

4.4　实验结果分析

通过实验发现，基于统计主题模型的标签预测较已有方法，其 Precision，Recall 值有明显提高，但也存在一些问题。首先，隐含主题的生成计算相对复杂，时间消耗较大，尤其是对标签文档较长、大数量的数据集的处理。而大数据量是 Web 环境下标签系统的基本特性，因此如果要将本方法用于实时环境，还需要改进。其次，LDA 算法对参数较为敏感，不同的数据集所适合的参数也不尽相同，为了获得满意的预测结果，需要多次试验以确定合适的参数。这无疑提升了本文方法的难度，同时也增加了计算的时间消耗。

根据表 1 对数据集进行分析的结果是确定基于 LDA 的标签预测算法各项参数的重要依据，这是正确理解数据集的基础。本文实验没有对所提出方法的时间消耗进行具体的量化分析。但值得注意的是，本文方法中分析数据集特征所耗费的时间并不会对标签预测任务的完成产生明显的影响。与其对提高标签预测结果性能的贡献相比，少量的时间消耗是十分有价值的。

结束语　本文对统计主题模型在标签预测中的应用进行了深入的研究，着重分析了不同形式的标签文档对标签预测的影响。实验结果不但验证了统计主题模型在标签预测中的效用，更重要的是表明了标签文档形式对预测结果的影响，蕴含标签资源相关统计信息的文档可以产生语义更为清晰、集中的隐含主题，因此可以获得更好的预测结果。

进一步的研究工作将主要围绕提高算法的效率展开，主要分两方面：一是根据标签预测任务的特征和需求，有针对性地改进 LDA 算法本身，降低算法的时间消耗；二是考虑将多种标签预测技术相结合，如基于关联规则的预测、基于标签频率的预测等，根据预测任务的特征选择合适的预测策略，在保证预测结果可靠的同时使计算效率得到最大地提高。

参 考 文 献

[1] del. icio. us[EB/OL]. http://www. del. icio. us

[2] flickr[EB/OL]. http://www. flickr. com

[3] CiteULike[EB/OL]. http://www. citeulike. org

[4] Blei D M, Jordan Y N A, Michael I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003(3): 993-1022

[5] Mishne G. Autotag: a collaborative approach to automated tag assignment for weblog posts[C]// Proceedings of the 15th international conference on World Wide Web (WWW '06). New York: ACM, 2006: 953-954

[6] Chirita P A, Costache S, Nejdl W, et al. P-tag: large scale automatic generation of personalized annotation tags for the Web[C]// Proceedings of the 16th International Conference on World Wide Web(WWW '07). New York: ACM, 2007: 845-854

[7] Ornsson B S, van Zwol R. Flickr tag recommendation based on collective knowledge[C]// Proceeding of the 17th International Conference on World Wide Web (WWW '08). New York: ACM, 2008: 327-336

[8] Schmitz C, Hotho A, Jäschke R, et al. Mining Association Rules in Folksonomies[J]. Data Science and Classification, 2006: 261-270

[9] Heymann P, Ramage D, Garcia-Molina H. Social tag prediction [C]// Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Rretrieval(SIGIR '08). New York: ACM, 2008: 531-538

[10] Krestel R, Fankhauser P, Nejdl W. Latent Dirichlet Allocation for Tag Recommendation[C]// Proceedings of RecSys'09. New York: ACM, 2009: 61-68

[11] Limpens F, Gandon F, Buffa M. Linking Folksonomies and Ontologies for Supporting Knowledge Sharing: a State of the Art [R]. ANR, 2009

[12] Zhang Ning, Zhang Yuan, Tang Jie. A Tag Recommendation System for Folksonomy[C]// Proceedings of SWSM'09. New York: ACM, 2009: 9-16

[13] Zubiaga A, García-Plaza A P, Fresno V, et al. Content-based Clustering for Tag Cloud Visualization [C] // Proceedings of ASONAM 2009, International Conference on Advances in Social Networks Analysis and Mining. 2009: 316-319

[14] Zubiaga A. Enhancing Navigation on Wikipedia with Social Tags [C]// Proceedings of Wikimania. 2009

[15] JGibbsLDA[EB/OL]. http://gibbslda. sourceforge. net

[16] Weka[EB/OL]. http:// www. cs. waikato. ac. nz/~ml/weka