标签时态特征分析及其在标签预测中的应用

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 标签作为用户生成的对资源的描述,反映了资源的语义和用户的兴趣。由于Web资源的动态性,标签数据相应地表现出较为明显的时态特征,已有相关研究中标签的时态特征却很少受到关注。针对这方面的不足,对标签数据的时态特征以及基于时态特征的标签间语义关联进行分析,并提出发现标签时态特征的时间段划分准则;为了评价标签时态特征的价值,以经典的统计主题模型为基础,提出新的模型用于分析数据时态特征对所生成主题的影响,并将其用于标签预测。在多个数据集上的测试验证了标签数据的时态特性及其对提高标签预测性能的影响。

关键词 标签,语义关联,时态,统计主题模型

中图法分类号 TP311.1 **文献标识码** A

Applying Temporal Features of Social Tags to Tag Predication

YUAN Liu[1] ZHANG Long-bo[2]

(College of Computer Science, Shaanxi Normal University, Xi'an 710062, China)[1]

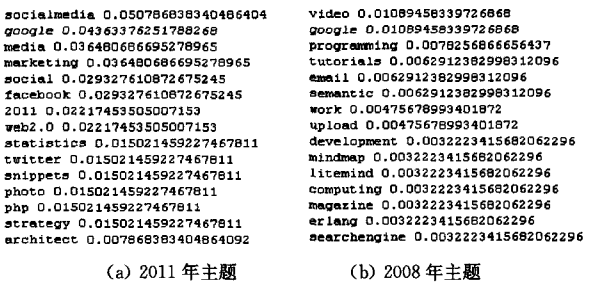
(School of Computer Science and Technology, Shandong University of Technology, Zibo 255049, China)[2]

Abstract Tag is a kind of description of Web resources generated by users, and it represents the semantics of resources and interests of users. Because the Web resources are dynamic, tags show some temporal features. However, few researches are concentrated on temporal features of tags. The temporal features represented by tags dataset were analyzed in this paper, and the semantic relations between tags based temporal features were discussed. The principle of time segmentation for discovering temporal features was proposed, and the effect of tags temporal on topics was analyzed by statistical topic model. The discovered temporal features were used in tags predication. The experiments based on different datasets shows that applying tags temporal feature to tags predication can improve the predication performance.

Keywords Tags, Semantic relation, Temporal, Statistical topic model

1 引言

在"以用户为中心"的Web2.0环境中,标签(Tags)已经成为用户标记、管理并分享自己所感兴趣资源的重要途径。作为一种用户生成的关于资源的元数据,标签对于研究Web资源的管理及用户的兴趣都有重要的价值。越来越多的研究关注于对标签数据的分析和利用[1]。由于标签数据反映了用户的兴趣和资源的内容,用户兴趣的转移和资源内容的变化将直接影响标签的生成,因此标签数据表现出了较为明显的时态特征。这种时态特征体现在两方面:一是标签所反映的资源主题的变化,新事物、新概念的出现都会影响用户标签数据的生成。图1显示了利用统计主题模型LDA得到的2008年和2011年的用户标签数据中与"Google"有关的主题对比。从中可以看出,即使对同一主题,其中所出现的标签数据也有明显变化。二是标签频率的变化趋势和标签之间在时态上的关联。标签之间在一定时间段内的相关程度可以表现出一定的变化趋势。图2显示了标签"gift"和"valentine"在一时间段内的关联程度变化。



(a) 2011年主题 (b) 2008年主题

图1 与"Google"相关的主题变化

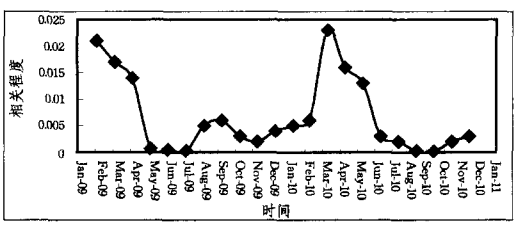


图2 标签"gift"与"valentine"从2009/03—2010/12相关程度的变化

关注标签的时态特征,无疑会提高标签数据分析与应用

到稿日期:2011-09-27 返修日期:2012-02-18 本文受国家自然科学基金项目(61003061)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:yuanliu@snnu.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。

的准确率，但以往相关研究大多忽略了标签的时态特性，本文将对此展开研究。通过对标签时态特征的分析，建立基于时态特征的用户标签描述，并将其用于标签预测以提高预测的准确率。

2 相关工作

虽然针对标签主题及变化趋势的研究成果还较少，但研究文档主题时态特征以及标签语义的相关成果都对本文研究有重要参考价值。文献[2]提出了 Temporal LDA 模型来对文档主题建模，其思想是在经典的 LDA 模型上增加变量 K，对文档集合按照时态顺序进行划分。该模型可以抓住文档主题的时态特征，但其对文档集合时态划分的正确性难以保证，并且引入新的变量大大增加了计算的时间复杂度。文献[3]通过定义不同的标签间语义关系的计算方式，对标签的变化趋势进行检测，但该研究所提出的标签变化趋势仍属少量，不能反映标签时态变化的整体情况。文献[4]从用户兴趣的角度对标签的变化进行分析和建模，利用标签的变化对用户的兴趣行为进行分析，但也没有系统地对标签时态特征进行定义和分析。尽管目前缺少关于标签时态特征的有针对性的研究成果，但一些关于 Web 资源检索、资源内容演变等的研究已经考虑利用标签所提供的信息，如文献[5-7]等，这些成果对本文研究也具有一定的借鉴价值。

3 标签预测标签时态特征的定义

Folksonomy[8]作为用户标签数据集合的形式化描述，其基本形式为

Folksonomy：(*tag set*，*user group*，*source*，*occurrence*) (1)

式中，*tag set* 为所有出现的标签词汇的集合；*user group* 为参与标记活动的用户的集合；*source* 为 Folksonomy 被使用的场合(如社会网络站点、在线社区等)；*occurrence* 是标签出现频率的集合，反映了标签的受欢迎程度。Folksonomy 其他形式的描述几乎都以该形式为基础，但都很少考虑标签的时态特征。对此，本文首先对增加了标签时态特征的 Folksonomy 进行定义。

定义 1 增加时态特征的 Folksonomy(Temporal features enhanced Folksonomy)

Temp-Folksonomy：(*tag set*，*user group*，*source*，*occurrence*，*time_dimension*) (2)

式中，*tag set*，*user group*，*source*，*occurrence* 的含义同式(1)，*time_dimension* 则表示标签所产生的时间段集合。

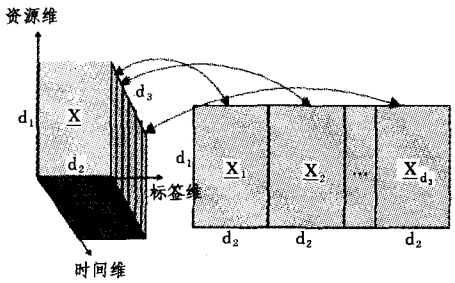


图 3 添加了时间约束的资源-标签关系

本文研究暂不考虑个体用户间标签行为的差异，故忽略 *user group* 集合，因此式(2)所定义的时态 Folksonomy 如图 3 所示。d_1，d_2，d_3 分别表示时态 Folksonomy $\underline{X}$的资源维、标签维和时间维。

可以看出，时间段的定义会直接影响在特定时间段内“标签-资源”的内容 X_{dt}。时间间隔越短，X_{dt} 间的差异就越小，从而能及时发现标签的产生、消失等现象，但相应的数据处理量会大大增加，且难以发现有价值的标签变化；时间间隔越大，可能越会丢失部分的标签变化，但数据处理的压力越小。因此选取合适的时间间隔观察标签的变化，以确保在发现有价值的标签变化信息的同时花费较小的数据处理代价，对研究标签时态特征具有非常重要的意义。本文研究中所采取的时间段划分准则如下。

定义 2(有效时间间隔 d) 设 W^t 表示 t 时刻的标签集合，W^{t+d} 表示 $t+d$ 时刻的标签集合，若 $|W^{t+d}-W^t|>u$，或 $|W^t-W^{t+d}|>u$，则 d 称为标签数据集上的有效时间间隔。其中 u 表示标签集合间不同标签个数的阈值，且 *occurrence*$(W^t)>f_t$、*occurrence*$(W^{t+d})>f_t$，即 W^t、W^{t+d} 所对应的 *occurrence* 属性值大于标签频率阈值 f_t。*occurrence*(W^t) 的计算方式为

$$occurrance(W^t)=\frac{\sum_{w_i\in W^t} occurrance(w_i)}{|W^t|}$$

通过集合 W^t 与集合 W^{t+d} 之间的减法运算，可获得在$[t, t+d]$时间段内新生成或消失的标签集合，其反映了标签集合的变化，变化程度可用集合中元素个数 u 来描述。u 的选择直接影响到使用 LDA 模型进行标签预测的算法的性能：u 越小，可以捕捉到的标签变化越细致，但会产生更多的需要进行处理的标签文档集合，LDA 模型需要对文档集合进行多次迭代处理，以常规迭代次数 2000 为参考，文档数量的增加对算法的效率有很大的影响；u 越大，可能越会丢失一些标签时态信息，但减少了预测算法的计算代价。设定 LDA 模型所生成的每个隐含主题中的标签个数为 30，综合考虑算法效率、同一隐含主题中标签的变化以及标签集合的规模。借鉴先前的研究基础[9]，本文研究中确定以 W^t 中元素个数的 1/10 即 $\frac{|W^t|}{10}$ 为阈值来判定有效时间间隔。由于本文研究关注的是标签集合的时态特征是否会对标签预测产生影响，其前提条件是标签集合已具有明显变化的特征，因此忽略了关于标签集合的变化程度即 u 对预测结果的影响的研究。关于标签集合的时态变化特征等问题也是我们进一步研究关注的内容。

$\sum_{w_i\in W^t} occurrance(w_i)$ 表示集合 W^t 中所有标签出现频率(次数)的总和，$|W^t|$ 表示 W^t 中所包含标签的数量，*occurrence*(W^t) 即可表示 W^t 中标签的平均出现频率。平均出现频率可以简单直观地反映 W^t 中标签的受欢迎程度。在相关研究工作中，一般认为标签集合中出现频率较低(小于 5 次)的标签随机性太强，不能代表大众的观点，因此会将其清除后再对集合做进一步处理。如果 *occurrence*(W^t) 的值小于 5(即当 $f_t=5$ 时)，则表示标签集合 W^t 并不具备用于预测研究的价值。

如图 3 所示，在针对特定的 *source* 集合所产生的不同时间点上的标签数据集之间，有效时间间隔 d 可以是不同的。这种基于有效时间间隔的标签数据集合分割方式，不需要对不同时间点的标签数据集进行严格的排序，避免了排序结果对分析结果的影响。

4 标签时态特征的检测

4.1 利用统计主题模型实现标签主题变化检测

在特定时刻的标签集合 W^t，可以按照标签所属的资源表示为

$tag_1^1, tag_2^1, \cdots, tag_k^1$；

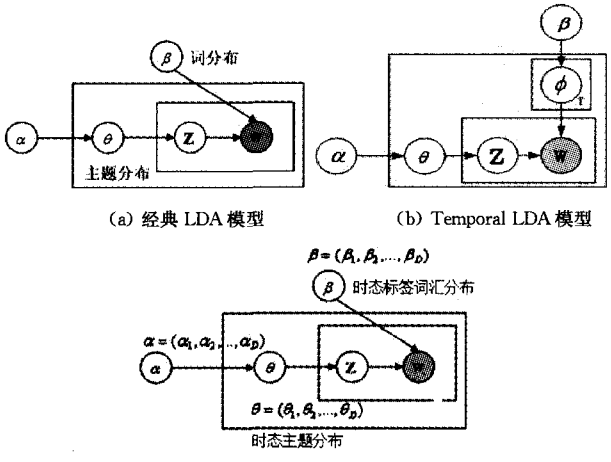
$tag_1^2, tag_2^2, \cdots, tag_l^2$；

……

$tag_1^n, tag_2^n, \cdots, tag_m^n$；

其中，tag_m^n 表示第 n 个资源 $source_n$ 所对应的第 m 个标签，因此标签集合可以看作由标签词汇组成的 n 个文档。研究文档语义及文档所包含的主题的方法可用于分析标签集合的主题。在文献[9]中，已经验证了 LDA 模型在分析标签语义方面的价值，本文研究将在文献[9]的基础之上，利用 LDA 模型对标签主题的变化进行分析。

利用 LDA 对标签集合进行分析的结果是，对每个标签词汇产生形如〈*word*：*topic*〉的词汇-主题对；对每一个标签词汇，都分派了一个主题，即使从语义上讲该词汇几乎不可能属于该主题，这种情况可视为主题的强制分配。强制分配的情况在具有时态特征的标签数据集合上表现尤为突出。由于标签数据的时态特性，不同时间段内标签所涉及的主题往往不同。为了避免不同时间段之间的主题强制分配，本文将对传统 LDA 模型进行扩展，以支持具有时态特征的标签数据集合上的主题检测，如图 4 所示。



(a) 经典 LDA 模型　(b) Temporal LDA 模型

(c) 包含了时态特征的 LDA 模型

图 4　LDA 模型对比

与传统的 LDA 模型相比，该模型的各个参数和随机变量都有所变化：在文档集合级别，模型参数 α 和 β 按照时间段被分为 D 组，即 $\alpha=\{\alpha_1,\alpha_2,\cdots,\alpha_D\}$，$\beta=\{\beta_1,\beta_2,\cdots,\beta_D\}$，参数对 (α_d,β_d) 确定 d 时间段内对应的一组隐含主题。α_d、β_d 分别与传统 LDA 模型中的 α、β 具有相同的形式：α_d 是一个列向量，每个分量 α_{d_i} 表示第 d 个时间段内的第 i 个隐含主题的先验概率；β_d 是一个矩阵，元素 $\beta_{d_{ij}}$ 表示第 d 个时间段内的第 i 个隐含主题生成词 j 的概率。模型中的隐含随机变量包括文档级的 θ 以及词一级的 z，其在形式上也有变化。目标文档 m 的隐含概念分配向量 θ 按照时间段分为 D 组 $\theta=\{\theta_1,\theta_2,\cdots,\theta_D\}$，其子向量 θ_d 表示文档 m 在时间段 d 的一组隐含主题上的分配额度。z 和 θ 类似，也是按照时间段分为 D 组，$z=\{z_1,z_2,\cdots,z_D\}$，z_d 为一个矩阵，$z_{d_{mn}}$ 表示时间段 d 内文档 m 中第 n 个词的主题索引。

在传统 LDA 模型引入时间段的约束之后，对于每个标签词汇，形如〈*word*：*topic*〉的直观结果将以〈*word*：*temporal_topic*〉的形式呈现，其中 *temporal_topic* 表示该标签所存在的时间段内的主题。值得注意的是，与图 4(b)相比，由于没有引入表示领域的随机变量，因此该模型并没有增加模型的计算复杂度。

4.2 标签间的时态关联检测

对于属于同一主题的标签，可以认为它们之间在语义上是有关联的，但并不能直接表达出标签间的关联程度。对此，本文研究提出了描述两个标签之间的关联程度的量化指标。首先需要对标签个体及资源个体关于标记行为的一些统计特征进行分析。具体包括如下量化指标：

$RS(tag_i,t)$：在时间点 t，使用标签 tag_i 所标记的资源集合；

$CS(resource_j,t)$：在时间点 t，用于标记资源 $resource_j$ 的标签集合；

$freq(tag_i,resource_j,t)$：在时间点 t，资源 $resource_j$ 使用标签 tag_i 进行标记的频率。

在以上基本统计量的基础上，本文将采用两类不同的度量方式对标签间的时态关联进行量化。

方法 1　基于被标记资源重叠程度的关联计算

$$corr_{M1}(tag_i,tag_j,t)=\frac{RS(tag_i,t)\bigcap RS(tag_j,t)}{RS(tag_i,t)\bigcup RS(tag_j,t)}$$

方法 2　基于标签向量余弦相似性的关联计算

$$corr_{M2}(tag_i,tag_j,t)=\frac{V_{tag_i,t}\cdot V_{tag_j,t}}{|V_{tag_i,t}|\cdot|V_{tag_j,t}|}$$

式中，$V_{tag_i,t}$ 是关于标签 tag_i 的向量，向量中第 j 个元素为 $w(tag_i,resource_j)$，表示资源 $resource_j$ 对标签 tag_i 的权重。其计算依据是，首先对 $freq(tag_i,resource_j,t)$ 值进行标准化，再根据不同资源所对应的标签数量进行加权，标签数量较多的资源将获得更高的权重。

$$w(tag_i,resource_j)=\frac{freq(tag_i,resource_j,t)}{\sum_{tag_k\in CS(resource_j,t)}freq(tag_k,resource_j,t)}\cdot\log_{\sum_{tag_k\in CS(resource_j,t)}freq(tag_k,resource_j,t)}$$

4.3 基于标签时态特征的标签预测

由于用户标记行为不受约束，因此资源的标签存在不完整、不规范、语义不清晰等情况。标签预测的目的是在分析标签信息相对完整的资源-标签分派集合的基础上，为标签数量较少的资源预测新的标签，以便能够对资源进行较为全面的描述，实现对资源的有效利用。如上文分析，标签时态特征的主要表现形式为在不同的时间段用户标签所涵盖的主题不尽相同，即使就同一主题来看，其中所包含的词汇也会产生变化；以词汇出现频率为统计特征的标签词汇的变化可以是周期性的，也可以是逐渐增强或逐渐减弱的趋势。因此，在标签预测的过程中考虑标签的时态特征，将有助于获得更为准确的预测结果。结合 4.1 节和 4.2 节的方法，本文提出了基于标签时态特征的标签预测方法。具体步骤如下。

步骤 1　划分训练集和测试集，即确定用于预测的训练集中的标签数据所在的时间段以及待预测的资源集合；

步骤 2　使用 4.1 节所述的 LDA 模型，获得训练集中标

签包含的主题；

步骤3　对于属于同一主题的标签词汇，使用4.2节的方法计算其间的相关程度；

步骤4　对于待预测的资源，找到与其已有的标签词汇相关程度较高的标签。

标签预测过程中将根据数据集的特点确定LDA模型及语义相关性计算中的重要参数。

5　实验

5.1　数据集

本文实验的数据来自标签服务提供网站Del.icio.us。为了研究标签的时态特征，实验中标签的生成时间覆盖了2003年－2011年的各年份。由于2006年之前用户对资源进行标记的普及程度远不如近年，在Del.icio.us上，无论是用户数量还是标签数量都远远不及2006年之后的数据，因此在分析数据时，将2003年－2006年这一阶段所生成的数据整合在一起进行分析。具体的标签数据来源如表1所列。使用JGibbsLDA[10]作为LDA模型的实现算法，算法中的参数(如迭代次数、主题数目、每个主题中所包含的标签个数等)根据数据集的大小进行调整。

在分析之前，首先对标签数据进行了预处理，以提高利用LDA模型的结果的准确率，具体包括：大小写转换，将标签词汇全部转换为小写形式；去除标签中出现的标点等非字符符合；删除出现频率过低的标签等。

表1　实验所用的标签数据集

数据集时间	标签数量	资源数量	数据来源
2003－2006	1,314,979	433,495	http://tagora.ecs.soton.ac.uk
2007.12	51,000	173,000	
2008.6	67,000	144,574	数据集DeliciousT 140
2009.9	82,000	155,000	http://arvindn.livejournal.com/116137.html
2010.4	61,000	136,000	通过Delicious RSS Feed直接获取
2011.6	61,000	136,000	通过Delicious RSS Feed直接获取

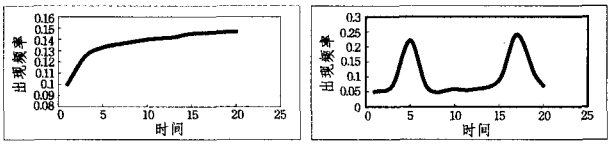
5.2　标签的时态特征的检测

首先是对标签主题变化的检测。分别对表1所列的标签数据集进行处理，得到不同时间段的数据集上的标签主题集合。根据标签的出现频率，选择了与“Google”、“Apple”、“Facebook”、“Network”这4个标签最相关且出现频率较高的标签所形成的主题，分别对其变化进行了具体说明，如表2所列。观察发现，标签主题呈现出较为明显的随时间变化的趋势。尽管一些“热门标签”在不同时间点的出现频率都较高，如主题“Apple”中的“mac”、“os”等，但仍有相当数量的标签词汇发生变化，引导着标签主题的变化。例如主题1中的词汇“rss”、“Android”等在2007年之前较少出现；在Web环境中不断会有新的主题产生，例如“Facebook”相关主题的产生与变化与这项应用的发展和普及是紧密相关的。一般认为属于同一主题的标签词汇之间具有较强的语义关联，随着主题中标签出现频率及标签词汇的变化，标签间的语义关联会随时间而变化，表2可反映这一特征。

表2　不同时间段标签主题的变化

	2003－2006	2007	2008	2009	2010	2011
主题1 (Google)	Google search Email map Web App tools Development Programming Code internet	Google search Pagerank Pandaro map Web App tools Development Programming Code	Google Web2.0 Map flickr Web app Tools programming opensource code	Goolge toread rss Web Programming Tools code seo Map gps mobile geo	Google seo Android search Email mobile App development opensouce	Google seo Android tools Email mobile App media development Media Software Gui html5
主题2 (Apple)	Apple mac Os jobs Music Ipod gui laptop software	Apple mac Os jobs Music Ipod gui Laptop software	Apple mac Osx ipod Utilities Jobs Mobile Iphone Chinese lvm webkit	Apple mac Software ios Google Ebook mac imac Jobs itouch Itunes support Css3	Apple mac iphone Design ipad app Ipod ios imported Registered blog Osx tips software google	Apple mac iphone Design ipad app Ipod ios imported Registered blog Osx tips software google
主题3 (Facebook)	NULL	Facebook Web gmail Validation gmail socialmedia Web2.0 Blogs Howto	Facebook socialmedia Web2.0 Blogs Social posts Google Howto online	Facebook socialmedia Web2.0 Blogs Social posts Google Howto online Twitter	Facebook socialmedia image Blogs Social posts movies Howto online Twitter Join	Facebook socialmedia Marketing twitter social Trends privacy google+ How to fanpage advertising
主题4 (network)	Network sysadmin security firewall solaris Monitoring Bind unix Ip opensource	Network sysadmin security firewall solaris Monitoring Unix Ip opensource	Network Social Wifi hardware Game programming Tutorial Security opensource	Network Wifi hardware socialmedia Web2.0 Bandwidth Myspace Security Howto Programming game	Network wifi wireless Linux android tools Utility security privacy Hardware 802.11n Free opensource Howto game	Network wifi wireless Linux android tools Utility security privacy Hardware 802.11n Free opensource Howto game

除了标签主题的变化，单个标签的出现频率在一定的时间区间内也呈现出了特定的趋势。以 2008 年 10 月至 2010 年 4 月为时间区间，图 5 分别展示了两种不同类型的标签频率变化。在计算标签出现频率时，本文对标签进行了词法上和语义上的处理，将语义上相同的标签作为同一标签处理。了解这些变化趋势对提高标签预测的准确率具有重要的参考意义。



(a) 标签"twitter"变化趋势　　(b) 标签"Valentine"变化趋势

图 5　单个标签的变化趋势

5.3　时态特征对标签预测结果的影响

直接验证标签间具有时态特征的语义关联的准确性较为困难，本文实验将以标签预测为途径，通过分析标签间的语义关联对标签预测结果的影响，检验了反映时态特征的标签间关联的应用价值。文献[9]在不考虑时态特征的情况下，讨论了 LDA 模型用于标签预测的效果，以该研究工作为基础，本文实验将进一步分析标签时态信息对预测结果的影响。具体方法是：首先，对于表 1 所列的标签数据集合，以训练集与测试集之比为 9:1 的比例分别对各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 JGibbsLDA 可处理的标签文档形式；其次，分别使用经典的 LDA 模型和 4.1 节所提出的 LDA 模型对测试集中的标签进行预测，将各个测试集预测结果的平均值作为预测的最终结果。同一主题内标签间语义相关程度的取值为 $\max(corr_{M1}(tag_i,tag_j,t),corr_{M2}(tag_i,tag_j,t))$，以遗漏语义相关的标签词汇。设定不同的"词汇：主题"的概率阈值 σ，分别使用经典的 LDA 模型和包含标签时态信息的 LDA 模型进行标签预测，结果如表 3 所列。

表 3　标签预测结果比较

阈值 σ	precision		recall		f-Measure	
	LDA	时态特征+LDA	LDA	时态特征+LDA	LDA	时态特征+LDA
0.01	0.681	0.693	0.194	0.189	0.302	0.297
0.005	0.617	0.624	0.269	0.272	0.375	0.379
0.001	0.455	0.471	0.302	0.298	0.363	0.365
0.0005	0.312	0.346	0.544	0.553	0.397	0.426
0.0001	0.354	0.403	0.573	0.577	0.437	0.475
0.00001	0.198	0.266	0.692	0.686	0.308	0.383

从预测结果可以看出，具有时态特征的语义关联对于改善标签预测的准确率 *precision* 有较明显的作用，在阈值 σ 取值较小的时候其作用更为明显。先前将 LDA 模型用于标签语义分析的基础研究表明[9,11]，即使一组词汇表现出了很明显的主题语义，大多数描述"词汇：主题"的概率值也都处于较低的水平。这说明本文方法对于所属主题相对分散的词汇集合更加有效。由于时态特征的引入将标签间语义关联的计算限定在一定的范围内，因此对查全率 *recall* 略有影响。从整体上分析，实验结果验证了标签时态变化会对标签预测结果产生积极的影响。为了进一步提高标签预测的各方面性能，还需要对标签时态特征进行深入的研究并对预测算法进行优化。

结束语　标签作为用户所生成的资源描述信息，其所具有的时态特征对于研究标签管理及 Web 数据管理是不可忽视的。本文从标签时态特征的描述、获取和应用几方面，展开对标签数据集合中时态特征的研究。研究结果表明，同一主题所包含的标签会随着时间的推移而发生演变，监测标签主题变化对于提高标签预测的性能有积极的意义。

本文关于标签时态特征的研究还仅限于较短时期内的标签数据，标签数据在更长时间段内的时态特征还需要继续跟踪研究。更深入的研究主要还包括对于标签时态特征的形式化定义和描述、将标签看作资源的语义元数据、如何对其存储和实施有效的管理，以及标签的更深层次、更有价值的应用。

参 考 文 献

[1] Knoll G. Folksonomy: Tagging-Brief Summary of the Ideas and Solutions to Tagging[EB/OL]. http://www.cs.colorado.edu/~kena/classes/7818/f06/lectures/folksonomy.pdf

[2] Reese K W, Shafto P. Towards Temporal Latent Dirichlet Allocation Model

[3] Hsu M-H, Chang Yu-Hui, et al. Temporal Correlation between Social Tags and Emerging Long-Term Trend Detection[C]// Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. 2010:255-258

[4] Yin Da-wei, Hong Liang-jie, et al. Temporal Dynamics of User Interests in Tagging Systems[C]// Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. 2011: 1279-1285

[5] Sato N, Uehara M. Temporal Ranking for Fresh Information Retrieval[C]// AsianIR'03 Proceedings of the sixth international werkslep on Information retrieval with Asian Languages. 2003

[6] Zubiaga A, Martinez R, et al. Getting the Most Out of Social Annotations for Web Page Classification[C]// Proceedings of the 9th ACM Symposium on Document Engineering. 2009

[7] Shen Zhi-yong, Luo Ping, et al. Topic Modeling for Sequences of Temporal Activities[C]// Proceedings of ICDM. 2009:980-985

[8] Kim H L, Scerri S, et al. The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies[C]// Proceedings of Int'l Conf. on Dublin Core and Metadata Application. 2008:128-137

[9] 袁柳，张龙波. 基于概率主题模型的标签预测[J]. 计算机科学，2011,38(7):175-180

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

[11] 袁柳，张龙波. 基于概率主题模型的多粒度 Web 文档标注[J]. 计算机应用，2010,30(12):3401-3406