基于社会性标注的本体学习方法

刘凯鹏[1)] 方滨兴[1),2)]

[1)](哈尔滨工业大学计算机网络与信息安全技术研究中心 哈尔滨 150001)

[2)](中国科学院计算技术研究所网络重点实验室 北京 100190)

摘 要 由相互协作的用户在社会性标注系统中产生的大量的标注数据可以作为各种语义网应用的数据源.文中提出一种基于社会性标注的本体学习方法来挖掘蕴涵在社会性标注中的语义信息,提出一种隐含包容层次结构来刻画标签空间中潜在的结构,并基于此模型推导出本体学习算法.首先利用集合论的方法确定标签之间的包容关系,并将其表示为标签包容关系图.在将此图转化为层次关系时,为解决包容关系的不一致性,提出一种基于随机游走的标签普遍性排序方法.最后提出一种自顶向下的凝聚式层次聚类算法来生成概念层次结构.在实际社会性标注系统中采集的数据集上进行的实验表明,与目前的代表性方法相比,文中提出的方法在性能上有明显的提高.

关键词 社会性标注;本体学习;包容关系;随机游走;凝聚式层次聚类

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2010.01823

Ontology Induction Based on Social Annotations

LIU Kai-Peng[1)] FANG Bin-Xing[1),2)]

[1)](*Research Center of Computer Network and Information Security Technology*, *Harbin Institute of Technology*, *Harbin* 150001)

[2)](*Key Laboratory of Network Science and Technology*, *Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190)

Abstract This paper proposes an ontology induction approach to harvest the emergent semantics from the folksonomies which are built from the social annotations made by collaborative users. The authors leverage a latent subsumption hierarchy model to uncover the implicit structure of tag space. First, tag subsumptions are identified with a set-theoretical approach and model the tag space as a tag subsumption graph. Then, a tag generality ranking procedure is used to overcome the problem of inconsistent subsumptions. Finally, an agglomerative hierarchical clustering algorithm is utilized to generate the concept hierarchy. The authors conducted experiments on a dataset collected from a real-world system. Both qualitative and quantitative experimental results show a competitive performance of the proposed approach.

Keywords social annotations; ontology learning; subsumption; random walk; agglomerative hierarchical clustering

1 引 言

社会性标注是指相互协作的用户通过一个开放的平台,对共享的资源赋予简短而富于个性化的标签,实现对资源的管理和共享.社会性标注系统提供的简便易用的社会性协作机制吸引了大量用户的参与,并籍此形成了被称之为 Folksonomy 的社会性

收稿日期:2010-08-22.本课题得到国家自然科学基金(60703014,60933005)、国家"九七三"重点基础研究发展规划项目基金(2011CB302605)和国家"八六三"高技术研究发展计划项目基金(2006AA010105-02,2007AA01Z416,2007AA01Z442,2009AA01Z437)资助.**刘凯鹏**,男,1981 年生,博士研究生,主要研究领域为信息检索、数据挖掘和机器学习. E-mail: liukaipeng@gmail.com. **方滨兴**,1960 年生,男,博士,教授,中国工程院院士,主要研究领域为信息安全、信息检索和分布式系统.

标注数据.

定义 1(Folksonomy). Folksonomy 是一个四元组$\mathbb{F}:=(U,T,R,Y)$,其中U为用户的有限集合,T为标签的有限集合,R为资源的有限集合,$Y\subseteq U\times T\times R$为用户、资源和标签上的三元关系集合,称为标注集合.

本体是某个特定领域内的概念及其相互之间关系的形式化表示[1].本文采用下面的仅包含有词汇层和概念层次结构的简单的本体定义.

定义 2(本体). 本体是一个三元组$\mathcal{O}:=(C,root,\leq_C)$,其中$C$为概念标识符集合,$root$为根元素标识,偏序关系$\leq_C$为概念之间的上下位关系,上半格$\langle C\cup\{root\},\leq_C\rangle$称为本体的概念层次结构.

由于人工构建本体的复杂性,本体自动学习方法作为一种克服"知识获取瓶颈"的有效手段,一直受到研究者们的广泛关注.以往的研究大都基于非结构化数据[2-3],从大量的语料中挖掘概念以及概念之间的关系.虽然取得了一定的成果,但由于噪音数据的大量存在,这类方法的效果难以令人满意.

社会性标注系统的迅猛发展为人们提供了大量的蕴含着大众智慧的社会性标注数据,其协作性和动态性使其成为一种的重要的信息源.以往的研究[4-5]表明,大量活跃的社会性标注行为会形成一种能够代表其中语义信息的有限数量的标签的平稳分布.换言之,正如 Mika 指出的:"众多的、相互关联的个体在相互交互中最终会产生一种可以被看作语义的全局效应"[6].这表明,社会性标注数据中的确蕴含着丰富的语义信息.因此,基于社会性标注数据的语义挖掘已经成为当前的一个活跃的研究课题,而基于社会性标注的本体自动学习则是其中一个重要的研究领域[6-10].

与非结构化的文本语料相比,基于社会性标注的本体自动学习方法的优点包括:(1)结构化.社会性标注数据可以被抽象为用户、资源和标签之间的三元关系,无论是在表示上还是操作上,均比非结构化的文本语料更为容易,且噪音数据更少;(2)概念化.由于信息管理的需要,人们在标注资源时使用的短小、精炼的标签,明显具有更为"概念化"的特点,因此也更容易提炼出信息量较多的概念;(3)动态化.标签能够直接地反应人们的社会兴趣的改变,因此基于社会性标注学习得到的本体也具有较强的动态性.因此,有理由相信基于社会性标注进行本体自动学习能够获得更好的效果.

基于社会性标注学习得到的本体的典型应用包括:(1)检索机制.在社会性标注系统中单纯地基于标签的搜索往往召回率较低.可以利用自动学习到的本体进行查询拓展,以增加相关检索结果数量.当检索结果数量较大时,还可以利用本体进行检索结果聚类,以使检索结果界面更加友好;(2)导航界面.目前在社会性标注系统中广泛采用的导航界面包括最受欢迎的标签、最相关标签和标签云等.利用自动学习的本体可以提供更为直观方便的层次式的导航界面,支持按照人们自然的概念层次结构访问被标注过的资源;(3)标签推荐.标签推荐是指根据历史标注信息为当前用户对指定资源推荐能够令用户满意的标签.在自动学习的本体中,标签的祖先节点和后代节点以及对它们进行有限拓展所形成的标签集合可以用来进行不同概念级别的标签推荐.

本文提出一种基于社会性标注的自动本体学习算法,主要贡献包括:(1)利用隐含包容层次模型来揭示标签空间中隐含的概念层次结构;(2)利用基于随机游走的标签普遍性排序方法解决将标签包容图转化为概念层次结构时遇到包容关系不一致的问题;(3)提出一种基于凝聚式层次聚类算法的概念层次生成算法;(4)通过在实际系统中采集的数据集上进行的实验验证了本文提出的算法的有效性.

本文第 2 节介绍相关工作;第 3 节描述本体学习算法;第 4 节给出实验结果;第 5 节讨论算法中存在的问题;第 6 节总结本文工作并指出未来工作方向.

2 相关工作

目前已经有很多基于社会性标注的本体自动学习算法.这些算法基本上可以分为两类,即基于相似度的方法和基于集合论的方法.

2.1 基于相似关系的方法

基于相似关系的方法的主要特征为,使用概念之间的相似度来确定它们之间的关系[2]. Heymann 和 Garcia-Molina[7]采用一种简单高效的方法来发现标签空间中隐含的层次结构(参见第 4.2.1 节).将每一个标签表示为一个标注向量.标注向量中对应于每个资源的元素值为该标签被用来标注该资源的次数.标签之间的相似度定义为其对应标注向量之间的余弦相似度.然后,建立一个以标签为顶点的图.如果两个标签之间的相似度大于一个预先指定的阈值,则在图中增加一条连接它们的边,边的权值为它们之间的相似度.最后,标签按照其亲近中心性(closeness-centrality)递减的顺序逐步地加入到层

次结构中.对于一个新标签,如果其与层次结构中已经存在的、与其最为相似的叶节点的相似度大于预先指定的阈值,则其将被作为该节点的子节点加入到层次结构中,否则其将被作为根节点的子节点加入到层次结构中.与相似度相反,标签之间的差异度也可以被利用来进行本体学习.Tang 等人提出了一种自底向上的本体学习方法[11].使用主题模型对社会性标注数据建模,并使用标签在不同主题上的分布来计算标签之间的差异度.在构建层次结构时,通过最小化结合了分别称为上位差异度(hypernym-divergence)、合并差异度(merging-divergence)和保留差异度(keep-divergence)的 3 种差异度的目标函数来优化信息丢失(information loss).在初始时,所有的标签均为叶子节点.然后选择两个能够优化目标函数的节点,并根据其差异度特征执行从属、合并或保留操作.重复这一过程直至所有的节点最终合并到根节点.

2.2　基于包容关系的方法

基于包容关系的方法利用概念之间的包容关系来建立概念层次结构.通常根据概念的属性集合,利用集合论方法来发现概念之间的包容关系[2].Sanderson 和 Croft[3]最早描述了包容关系的统计模型.令 $D(x)$ 为包含 x 的文档集合,$P(x|y)=|D(x)\cap D(y)|/|D(y)|$,他们定义 x 包容 y 如果其满足 $P(x|y)=1$ 且 $P(y|x)<1$.他们同时注意到由于在一些文档中,y 没有 x 共现,导致 x 和 y 之间的包容关系没有被发现.因此,可以放松上式中的第 1 个条件为 $P(x|y)\geq 0.8$ 且 $P(y|x)<1$.Schmitz[8]在 Flickr 的标注数据集上拓展了 Sanderson 和 Croft 的工作,通过调整统计阈值来反映在该数据集上特殊的标签使用方式,并对一些很不常用的标签进行过滤.令 $D_{\min}$ 和 $U_{\min}$ 为预先指定的阈值,他定义 x 包容 y 如果其满足 $P(x|y)\geq t$,$P(y|x)<t$,且 $|D(x)|\geq D_{\min}$,$|D(y)|\geq D_{\min}$,$|U(x)|\geq U_{\min}$,$|U(y)|\geq U_{\min}$.Schmitz 等人将发现标签之间的包容关系的问题形式化为在形式背景(formal context)中进行关联规则挖掘的问题[9](参见第 4.2.1 节).他们使用了多种将社会性标注数据投影到二维空间以形成形式背景,从而进行关联规则挖掘的方法.对于挖掘得到的关联规则,只有其支持度和置信度大于预先指定的阈值时,才会被认定为有效.Mika[6]将传统的本体的两部图模型拓展为包含社会维度的三部图模型.他将在社会性标注数据中出现的不同种类的对象之间的共现关系建模为图,并提出可以利用这一图模型进行本体学习.在推导标签之间的包容关系时,他采用了与 Schmitz[9]等人相同的集合论方法.

3　本体学习算法

3.1　隐含包容层次模型

在本文中,为简化本体学习算法,假设用户使用单独的标签来表示概念,即 $C=T$.虽然这一假设具有一定的合理性,但是其过度地简化了社会性标注系统中用户复杂的标注行为.本文将在第 5 节中讨论这一问题.

Heymann 和 Garcia-Molina 曾经提出一种基于标签相似关系图的隐含层次结构模型[7]来阐述他们提出的本体学习算法的原理.他们提出了 3 个前提假设:(1)在导出本体中,概念之间的相互联系在标签相似关系图中也是存在的;(2)在标签相似关系图中存在噪音,即并非所有的关系都是有用的;(3)噪音关系较多地存在于本体概念层次结构的上层.本文采用一种更适合于揭示标签空间隐含结构的隐含包容层次模型.考虑将标签空间建模为以标签为顶点,以标签之间的包容关系为边的标签包容关系图(如图 1 的左半部分所示).本体学习算法试图将其转化为本体概念层次结构(如图 1 的左半部分所示).在隐含包容层次模型中包含如下 3 个前提假设:

(1)关系先在性.在导出本体概念层次结构中表示概念之间的上位关系和下位关系的边在标签包容关系图中也是存在的.这一前提假设是本体自动学习算法的基础.如果在导出本体概念层次结构中的关系并不预先存在于标签关系图中,那么将很难进行本体的自动学习.

(2)噪音存在性.在标签包容关系图中存在不相关关系和不一致关系.不相关关系是指那些不存在于导出本体概念层次结构中的关系.不一致关系是指那些虽然存在于导出本体概念层次中,但是方向相反的关系.根据对实际的社会性标注数据的分析,这些关系显然是存在的.它们的存在成为了基于社会性标注进行本体学习的最大障碍.

(3)噪音局部性.不相关的包容关系大多存在于本体层次结构的上层,而不一致关系则大多存在于本体层次结构的下层.一方面,在本体层次结构上层的更加普遍的标签倾向于包容更多的标签,造成了较多的不相关包容关系;而另一方面,在本体层次

结构下层的更加特殊的标签由于其出现次数较少，具有不稳定性，易于形成较多的不一致包容关系.这一前提假设为本体学习算法提供了清除噪音包容关系的有用的线索.

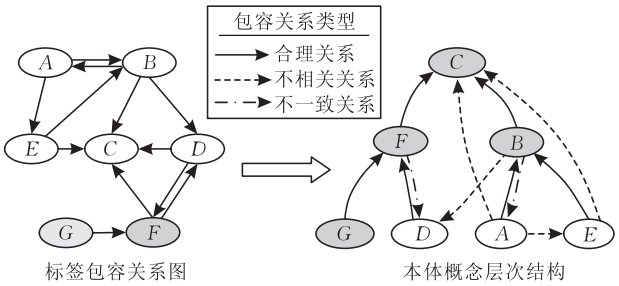

图 1 隐含概念层次模型

本文将在下一节中基于此隐含包容层次模型设计本体学习算法.首先将基于此模型中的关系先在性假设来发现标签空间中的包容关系，并构建标签包容关系图(见第 3.2 节).根据此模型中的噪音存在性假设，可以知道只有合理有效的清除此标签包容关系图中的不相关关系和不一致关系，才能将其最终转化为可用的本体概念层次结构.本文通过基于随机游走的标签普遍性计算，在标签空间中引入全序关系来解决标签包容关系图中存在不一致关系的问题(见第 3.3 节).最后，利用噪音局部性假设，使用自顶向下的凝聚式层次聚类算法生成本体概念层次结构(见第 3.4 节).

3.2 标签包容关系发现

本文采用形式概念分析[12](formal concept analysis)中的术语来描述标签包容模型，因其非常适合于在定义 1 中给出的社会性标注的数据模型定义.

定义 3(形式背景). 形式背景是一个三元组 $K:=(G,M,I)$，其中 G 为对象集合，M 为属性集合，$I\subseteq G\times M$ 为定义在对象和属性集合上的二元关联关系.

对于 $A\subseteq M$，定义

$$G_A:=\{g\in G\mid \forall m\in A:(g,m)\in I\}.$$

定义 A 的支持度为

$$Supp(A)=\frac{|G_A|}{|G|} \tag{1}$$

对于两个属性集合 $A_1\subseteq M$ 和 $A_2\subseteq M$，定义它们之间关系的支持度和置信度为

$$Supp(A_1,A_2)=Supp(A_1\cup A_2) \tag{2}$$

$$Conf(A_1,A_2)=\frac{Supp(A_1\cup A_2)}{Supp(A_1)} \tag{3}$$

对于给定的最小支持度 $\theta_s\in[0,1]$和最小置信度 $\theta_c\in[0,1]$，认为 A_1 被 A_2 包含，表示为 $A_1\preceq A_2$，如果满足

$$Supp(A_1,A_2)\geq\theta_s\wedge Conf(A_1,A_2)\geq\theta_c.$$

利用上文定义的包容模型，通过将社会性标注数据投影到形式背景上可以发现标签之间的包容关系.注意到本文将每一个标签直接映射为一个概念，故在发现包容关系的过程中，只考虑形式背景中属性集合的单元素子集.可以采用不同的投影方式，例如

(1) $K^U:=(G=U,M=T,I)$，$(u,t)\in I\Leftrightarrow(u,t,r)\in Y$；

(2) $K^R:=(G=R,M=T,I)$，$(r,t)\in I\Leftrightarrow(u,t,r)\in Y$；

(3) $K^{U,R}:=(G=U\times R,M=T,I)$，$((u,r),t)\in I\Leftrightarrow(u,t,r)\in Y$.

例 1(标签包容关系发现). 假设在社会性标注数据中有两个用户 u_1 和 u_2，两个资源 r_1 和 r_2，两个标签 t_1 和 t_2 以及 5 个标注(u_1,t_1,r_1)、(u_1,t_1,r_2)、(u_1,t_2,r_1)、(u_1,t_2,r_2)和(u_2,t_1,r_2).

令 $A_1=\{t_1\}$，$A_2=\{t_2\}$，$A_1\cup A_2=\{t_1,t_2\}$.假定使用投影 $K^{U,R}$来发现包容关系，有

$$G_{A_1}=\{(u_1,r_1),(u_2,r_1),(u_2,r_2)\},$$

$$G_{A_2}=\{(u_1,r_1),(u_1,r_2)\},$$

$$G_{A_1\cup A_2}=\{(u_1,r_1)\},$$

$$G=\{(u_1,r_1),(u_1,r_2),(u_2,r_1),(u_2,r_2)\},$$

故，可以计算

$$Supp(A_1)=\frac{|G_{A_1}|}{|G|}=0.75,$$

$$Supp(A_2)=\frac{|G_{A_2}|}{|G|}=0.50,$$

$$Supp(A_1,A_2)=Supp(A_1\cup A_2)=\frac{|G_{A_1\cup A_2}|}{|G|}=0.25,$$

$$Conf(A_1,A_2)=\frac{Supp(A_1\cup A_2)}{Supp(A_1)}=0.33,$$

$$Conf(A_2,A_1)=\frac{Supp(A_1\cup A_2)}{Supp(A_2)}=0.50.$$

也就是说，A_1 被 A_2 包容的置信度为 0.33，而 A_2 被 A_1 包容的置信度则为 0.50.

本文在实验中采用 $K^{U,R}$，因其能够在最大程度上保留社会性标注数据中蕴涵的信息，从而获得最好的效果.

3.3 标签普遍性排序

通过发现标签之间的包容关系，可以构建以标签为顶点，以标签之间的包容关系为边的标签包容关系图.然而，正如在第 3.1 节中指出的，必须以合

理有效的方式清除图中的不相关关系和不一致关系，才能将此图转化为概念层次结构. 对于任意给定的置信度阈值 $\theta_c<1$，均有可能形成循环的包容关系链，造成了标签包容关系的不一致. 这种不一致的标签包容关系使得难以确定哪一个标签应该在概念层次结构较上层.

本文采用将标签按照普遍性进行排序，并将排序靠前的标签放置在概念层次结构的较上层的方法来解决这一问题. 标签普遍性排序的基本思想是，如果一个标签包容其它具有较高普遍性的标签，那么它本身也应该具有较高的普遍性. 换言之，通过利用标签之间的普遍性互增强关系，可以计算标签的普遍性. 每一个包容关系将被视为一个互增强关系，而这一互增强关系将被赋予正比于该包容关系置信度的权重. 所有的包容关系都将被用来传播标签普遍性. 基于这一思想，采用在机器学习和信息检索领域被广泛应用的随机游走方法来计算标签普遍性是非常适合的.

形式化定义标签包容关系图为 $G:=(V,E,\boldsymbol{W})$，其中 $V=T$ 为顶点集合，即标签集合，$E\subseteq T\times T$ 为边集合，即标签之间的包容关系集合，$\boldsymbol{W}$ 为邻接矩阵. 对于图中的两个顶点 $t_i\in T$ 和 $t_j\in T$，连接它们的边的权重 $w_{i,j}$ 定义为

$$w_{i,j}=Conf(\{t_i\},\{t_j\}) \tag{4}$$

对给定的标签包容关系图 G，使用其邻接矩阵 $\boldsymbol{W}$ 的行随机形式 $\boldsymbol{P}$ 作为随机游走过程的转移概率矩阵. $\boldsymbol{P}$ 中的元素 $p_{i,j}$ 代表从节点 t_i 到 t_j 的转移概率，定义为

$$p_{i,j}=\frac{w_{i,j}}{\sum_k w_{i,k}} \tag{5}$$

定义在标签包容关系图上的随机游走过程如下：

(1) 以概率 $\lambda\in[0,1]$，按照从当前节点出发的边的概率分布，游走到一个邻居节点；

(2) 以概率 $1-\lambda$，按照一个预先指定的在所有节点上的概率分布 $\boldsymbol{s}$，游走到任一节点.

使用上述的在标签包容关系图上的随机游走过程的稳态概率分布作为标签的普遍性分数. 可以采用迭代的方式计算稳态概率分布，即普遍性分数. 令 $g_i^{(n)}$ 为在第 n 次迭代时节点 t_i 的普遍性分数，则有

$$g_i^{(n)}=\lambda\sum_j p_{j,i}g_j^{(n-1)}+(1-\lambda)s_i \tag{6}$$

令 $\boldsymbol{g}^{(n)}\equiv[g_i^{(n)}]_{|T|\times 1}$ 为在第 n 次迭代时所有节点的普遍性分数形成的向量，上式也可以写成矩阵形式

$$\boldsymbol{g}^{(n)}=\lambda\boldsymbol{P}\boldsymbol{g}^{(n-1)}+(1-\lambda)\boldsymbol{s} \tag{7}$$

式(7)的收敛性在定理 1 中证明.

定理 1.　式(7)的迭代过程收敛.

证明.　迭代式(7)，有

$$\boldsymbol{g}^{(\infty)}=\lim_{n\to\infty}(\lambda\boldsymbol{P})^n\boldsymbol{g}^{(0)}+\lim_{n\to\infty}(1-\lambda)\Big(\sum_{i=1}^{n}(\lambda\boldsymbol{P})^{i-1}\Big)\boldsymbol{s} \tag{8}$$

对 $\boldsymbol{g}^{(\infty)}$ 的第一个分量，考虑

$$\begin{aligned}\sum_j(\lambda\boldsymbol{P})_{i,j}^n&=\sum_j\sum_k(\lambda\boldsymbol{P})_{i,k}^{n-1}(\lambda\boldsymbol{P})_{k,j}\\&=\sum_k(\lambda\boldsymbol{P})_{i,k}^{n-1}\Big(\lambda\sum_j\boldsymbol{P}_{k,j}\Big)\end{aligned} \tag{9}$$

由于 $\sum_j\boldsymbol{P}_{k,j}=1$，有

$$\sum_j(\lambda\boldsymbol{P})_{i,j}^n=\lambda\sum_k(\lambda\boldsymbol{P})_{i,k}^{n-1}\equiv\lambda\sum_j(\lambda\boldsymbol{P})_{i,j}^{n-1} \tag{10}$$

对于 $\lambda\in(0,1)$，存在 $\gamma\in(0,1)$ 满足 $\lambda\leq\gamma$，故有

$$\sum_j(\lambda\boldsymbol{P})_{i,j}^n\leq\gamma\sum_j(\lambda\boldsymbol{P})_{i,j}^{n-1} \tag{11}$$

迭代式(11)，有

$$\sum_j(\lambda\boldsymbol{P})_{i,j}^n\leq\gamma^n \tag{12}$$

故有

$$\lim_{n\to\infty}\sum_j(\lambda\boldsymbol{P})_{i,j}^n=0 \tag{13}$$

这表明，由于 $(\lambda\boldsymbol{P})^n$ 的每行元素之和收敛于 0，故 $\boldsymbol{g}^{(\infty)}$ 的第一个分量将收敛于一个零向量. 注意到 $\boldsymbol{P}$ 是行随机的，根据式(8)，有

$$\boldsymbol{g}^{(\infty)}=(1-\lambda)(\boldsymbol{I}-\lambda\boldsymbol{P})^{-1}\boldsymbol{s} \tag{14}$$

式中 $\boldsymbol{I}$ 为单位矩阵. 式(14)即为式(7)的收敛解.

证毕.

在算法实现中，式(7)的迭代过程终止于：(1) 评分向量在两次迭代中的值 $\boldsymbol{g}^{n+1}$ 和 $\boldsymbol{g}^n$ 满足 $\|\boldsymbol{g}^{n+1}-\boldsymbol{g}^n\|_2/\|\boldsymbol{g}^n\|\leq\theta$，其中 θ 为预先指定的阈值(在实验中取 $\theta=0.001$)；或(2)迭代次数大于预先制定的阈值 $n_{\max}$(在实验中取 $n_{\max}=100$). 如果有可供参考的关于标签普遍性的知识，则可以利用概率分布 $\boldsymbol{s}$ 对排序结果进行定制. 在实验中设定 $\boldsymbol{s}$ 为均匀分布，即 $s_i=1/|T|$.

3.4　概念层次生成

将标签按照其普遍性排序后，最后使用一个自顶向下的凝聚式聚类算法来生成本体概念层次结构. 算法的详细过程如算法 1 所示. 算法中使用到了两个辅助函数分别为：(1) $GeneralitySort(T)$. 将 T 中的所有标签按照其普遍性递减的顺序排序；(2) $Ancestors(t,\leq_C)$. 返回标签 t 的除了根节点 $root$ 之外所有祖先节点.

算法 1. 本体概念层次生成算法.

输入：$\mathbb{F}=(U,T,R,Y)$为社会性标注数据

输入：θ_s为最小支持度

输入：θ_c为最小置信度

输出：$\mathbb{O}=(C,root,\leq_C)$为导出本体

1. $C\leftarrow\varnothing$;
2. $\leq_C\leftarrow\varnothing$;
3. $T_{\text{generality}}\leftarrow GeneralitySort(T)$;
4. foreach $t_i\in T_{\text{generality}}$ do
5. $t_{\text{parent}}\leftarrow root$;
6. $c_{\max}\leftarrow 0$;
7. foreach $t_j\in C$ do
8. if $Supp(t_i,t_j)\geq\theta_s$ and
9. $Conf(t_i,t_j)\geq\theta_c$ then
10. if $Conf(t_i,t_j)>c_{\max}$ then
11. $context\leftarrow$ true;
12. foreach $t_k\in Ancestors(t_j,\leq_C)$ do
13. if $Supp(t_i,t_k)<\theta_s$ or
14. $Conf(t_i,t_k)<\theta_c$ then
15. $context\leftarrow$ false;
16. break;
17. end
18. end
19. end
20. if $context$ then
21. $t_{\text{parent}}\leftarrow t_j$;
22. $c_{\max}\leftarrow Conf(t_i,t_j)$;
23. end
24. end
25. end
26. $C\leftarrow C\cup\{t_i\}$;
27. $\leq_C\leftarrow\leq_C\cup\{t_i\leq t_{\text{parent}}\}$;
28. end
29. return $\mathbb{O}=(C,root,\leq_C)$;

算法首先将 T 中的所有标签按照其普遍性递减的顺序排序，得到 $T_{\text{generality}}$（第 3 行）. 然后，对 $T_{\text{generality}}$中每一个标签，选择一个已在概念层次结构中的，包容该标签且该包容关系置信度最大的标签作为该标签的父节点（第 5～25 行）. 当选择父节点时，只有置信度和支持度不小于预先指定的阈值 θ_s 和 θ_c的包容关系才被考虑. 另外，除了这一节点本身之外，其所有的祖先节点和新加入的标签之间的包容关系的支持度和置信度也要不小于预先指定的阈值 θ_s和 θ_c（第 11～23 行）. 如果这里不对整个路径上的包容关系都进行检查，则可能会生成诸如 macintosh≤apple≤fruit≤plant 的包容关系链. 在这种包容关系链中，每一个包容关系都是合理的，但整个包容关系链却是不合理的. 具有多种词义的标签（如 apple）可能会造成这类问题. 本文将在第 5 节中讨论这一问题. 如果可以找到一个合适的父节点，则将标签作为该节点的子节点加入到概念层次结构中；否则将标签作为根节点的子节点加入到概念层次结构中（第 26～27 行）. 重复此过程直至所有的标签都被加入到概念层次结构中为止.

通过利用标签普遍性排序的结果，此算法充分地应用了在第 3.1 节中描述的隐含包容层次结构模型. 通过选择具有最高置信度的包容关系来确定父节点，有效地清除了不相关的包容关系. 通过将标签按照普遍性递减的顺序加入到概念层次结构中，有效地清除了不一致的包容关系. 由于不相关的包容关系较多地存在于概念层次结构的上层，采用自顶向下的凝聚式聚类算法保证了在生成概念层次结构上层时，仅有较少的可供选择的父节点，且这些节点的差异性较大，从而降低了将不相关的包容关系引入到概念层次结构中的可能性. 而当生成概念层次结构的下层时，需要注意到不一致的包容关系增多的问题. 由于采用自顶向下的方法，当生成概念层次结构下层时可供选择的父节点增多，普遍性较低的标签此时被加入到概念层次结构中，不易将不一致的包容关系引入到概念层次结构中.

参数 θ_s和 θ_c在算法中起到相对重要的作用. 偏高或偏低的参数值将导致覆盖率较高或精确度较高的本体概念层次结构. 本文将在第 4 节中通过实验的方式选择合理的参数设置.

3.5 算法复杂性分析

对于标签包容关系发现，令 m 为用户的平均标注数量，对每一对标签，计算它们之间的包容关系的支持度和置信度需要 $O(m)$步，故此步骤的算法复杂性为 $O(m|T|^2)$. 对于标签普遍性排序，令 n 为式(7)的迭代过程的收敛所需的平均迭代次数，则计算每一个标签的普遍性需要 $O(n|T|)$步，故此步骤的算法复杂性为 $O(n|T|^2)$. 对于概念层次结构生成，为每一个标签选择父节点时最多需要检查 $|T|$ 个节点，故此步骤的算法复杂性为 $O(|T|^2)$. 尽管整个算法的理论时间复杂性较高，但在实际应用时，该算法能够应用于较大规模的数据集，原因是：在第 1 步中，用户的平均标注数量 m 值相对较小（在实验数据集中 $m<20$）；在第 2 步中，迭代次数 n 值也相对较小（在 20～50 之间）；另外第 1 步和第 2 步均可以并发执行，可部署在诸如 MapReduce[13] 之类的平台上；在第 3 步中，通过预先标记包容当前节点的所

有节点，并在选择父节点时只检查这些被标记过的节点，则实际需要检查的节点数目将远远小于$|T|$.

4 实　验

在本节中，首先介绍实验中使用的数据集(见第4.1节)，然后简要地介绍评价方案(见第4.2节)，最后给出详细的实验结果(见第4.3节和第4.4节).

4.1 实验数据集

本文中使用的实验数据集是从实际的社会性标注系统Delicious(http://www.delicious.com/)中采集的.该数据集能够基本上客观地反映在Delicious系统中大约半年时间内所发生的标注活动.从2008年11月开始，利用定制的爬虫在Delicious网站上采集网页，并从网页中抽取用户、资源、标签和标注时间等信息，形成包含825401个用户、59623937个资源、6420685个标签和459686018个标注的原始数据集.

对原始数据集的预处理分两步进行.在第1步中，使用Porter算法对所有标签取词根以清除原始数据集中因不同词形变化而产生的噪音数据.在第2步中，从原始数据集中移除出现次数少于k次的用户、资源和标签.在这一步中，不活跃的用户、不常见的资源和不常用的标签被清理出原始数据集以去除更多的噪音数据.实验表明，不同的k值对相对实验结果影响很小.在下面的实验中，取$k=20$.经过预处理的实验数据集包括282016个用户、90790个资源、32615个标签和30902845个标注.

4.2 评价方案

可以将本体学习算法看作是一个以领域相关的语料集为输入，以能够代表该语料集中涉及到的概念及其相互之间的关系的本体为输出的自动化过程.因此，可以通过评价导出本体的质量来评价本体学习算法的性能.本文通过比较本文提出的算法和目前的代表性算法来评价算法性能.

4.2.1 代表性算法

选择两个代表性算法进行比较分析，其分别为Heymann等人在文献[7]中提出的算法和Schmitz在文献[9]中提出的算法(参见第2节).Heymann方法[7]将标签表示为标注向量，并基于标注向量之间的余弦相似度来计算标签之间的相似度.对应于标签t的标注向量为一个$|R|$维向量，其中对应于资源r的元素值为资源r被标签t标注的次数.在本文中，为了进行公平的比较，将对应于标签t的标注向量扩展为$|U|\times|R|$维矩阵$\boldsymbol{A}^t$.如果用户u对资源r使用过标签t来标注，则$\boldsymbol{A}^t_{u,r}=1$，否则$\boldsymbol{A}^t_{u,r}=0$，并使用$\sum_{u\in U}\sum_{r\in R}(\boldsymbol{A}^t_{u,r}\cdot\boldsymbol{A}^{t'}_{u,r})\Big/\sqrt{\sum_{u\in U}\sum_{r\in R}\boldsymbol{A}^t_{u,r}\cdot\sum_{u\in U}\sum_{r\in R}\boldsymbol{A}^{t'}_{u,r}}$来计算$t$和$t'$之间的相似度.在初步实验的结果表明，使用修改后的标注向量可以小幅度地提高Heymann的算法的性能.Schmitz方法[9]中没有具体描述如何根据挖掘出的关联规则来最终生成本体.为了评价Schmitz方法的性能，首先使用Schmitz方法挖掘出关联规则，然后根据这些关联规则构建一个以标签为顶点、以关联规则为边、以关联规则的置信度为边的权重的图.对图中的每个顶点，只保留其权重最大的一条入边.对图中剩余的边所形成的环，删除环中权值最小的一条边.连结概念层次结构的根节点和图中剩余的所有的树的根节点，生成最终的概念层次结构.

对Heymann的算法，采用与其相同的参数设置[7]，即令最小相似度阈值为0.099.对Schmitz的算法，采用经验的最小支持度阈值0.00001.对本文提出的算法，在计算标签普遍性时，在式(7)中采用经验值$\lambda=0.95$；在生成概念层次结构时，采用经验的最小支持度阈值$\theta_s=0.00001$，并在初步实验中通过尝试不同的最小置信度阈值$\theta_c=(0.05,0.10,\cdots,0.95)$来选择使得词汇层F-值(参见第4.4.1节)最优的参数设置，即$\theta_c=0.15$.

4.2.2 评价方法

对本体质量进行合理的评价是一项非常困难的工作.尽管已经有一些关于本体评价的研究工作[14~16]，但它们大都仅对本体的某方面特性进行评价.在本文中，为了对本体进行全面深入的评价，将结合以下的两种方法来进行评价：

(1)人工评价.人工评价方法在以往的研究中经常被采用[3,8,17].通过对从导出本体中提取的上下位关系进行人工评价来比较不同本体学习算法的性能(见第4.3节).

(2)基于标准本体的评价.基于标准本体的评价是指将导出本体与预先选择的标准本体进行对比以评价本体学习算法性能的方法.在这种评价方法中，本体的词汇层和结构层都将被比较，以获得更加全面的评价结果(见第4.4节).

4.3 人工评价

尽管定性分析可以提供很多有用的信息，但是其对算法性能的评价较为笼统.为了更加客观地评价不同的本体学习算法，下面将采用定量的评价方法.定量的评价方法有两种，即人工评价方法和基于标准本体的评价方法.本小节和下一小节将分别描

述这两种评价方法的详细过程并给出评价结果.

4.3.1 评价方法

人工评价的方法在很多以往的研究工作中被采用[3,8,17].导出本体经过一定的方式被转化为易于被人工评价的形式呈现给参与评价的领域专家.在本文中,将本体分解成为以下两种类型的关系,以便于人工评价:

(1)父节点-子节点.考察一个节点与其子节点形成的关系是否合理;

(2)祖先节点-后代节点.考察一个节点与其所有的后代节点形成的关系路径是否合理.

由于从整个的本体中抽取出的这两种关系的数量过多,为使人工评价成为可能,只对一些具有代表性的子树进行评价.这些子树的根节点都是非常普遍和常用的概念,便于人工评价.对于每个子树,随机选择 100 个子节点-父节点关系和 50 个叶节点-祖先节点关系来进行评价.

评价人员为 3 名计算机专业的博士研究生,他们能够熟练地使用互联网查找不熟悉的标签的意义.对于每个关系,评价人员根据下面列出的准则来判断其是否合理.合理的关系包括以下两种类型:

(1)类型-实例,如 tennis 是 sport 作为类型的一个实例;

(2)整体-部分,如 wheel 是 car 作为整体的一个部分.

如果有不少于两个的评价人员认为一个关系是合理的,那么则认为其是合理的.使用学习正确率,即在所有被评价的关系中合理的关系所占的比例,来对学习算法的性能进行评价.

4.3.2 评价结果

人工评价的结果如表 1 所示.对于父节点-子节点关系,本文方法在共 15 个子树中的 11 个上获得最优的结果,并且在平均正确率上比 Heymann 方法和 Schmitz 方法分别提高了 20.9%和 5.1%.本文方法能够获得较好的性能的原因是,通过有机地结合标签普遍性排序和自顶向下的本体生成算法,有效地减少了本体中的不相关和不一致的包容关系.

表 1 对不同学习算法得到的本体的人工评价结果比较

根节点	父节点-子节点			祖先节点-后代节点		
	本文方法	Heymann 方法	Schmitz 方法	本文方法	Heymann 方法	Schmitz 方法
sport	**0.54**	0.45	0.50	**0.38**	0.24	0.26
scienc	**0.52**	0.33	0.44	**0.26**	0.18	0.20
news	**0.49**	0.19	0.45	0.28	0.12	**0.30**
program	**0.49**	0.34	0.39	**0.22**	0.14	0.18
histori	**0.64**	0.44	0.60	**0.32**	0.22	0.30
book	0.51	0.19	**0.52**	**0.28**	0.10	**0.28**
cultur	0.47	**0.61**	0.44	0.26	**0.34**	0.20
comput	**0.54**	0.21	0.43	**0.28**	0.12	0.22
game	**0.53**	0.29	0.49	0.24	0.12	**0.26**
educ	**0.51**	0.26	0.47	**0.40**	0.14	0.32
resourc	0.70	0.10	**0.72**	0.42	0.06	**0.44**
media	**0.46**	0.24	0.35	**0.30**	0.14	0.20
shop	**0.61**	0.30	0.55	0.28	0.16	**0.32**
graphic	0.49	**0.68**	0.49	0.30	**0.35**	0.28
health	**0.67**	0.41	0.57	**0.42**	0.18	0.28
平均	**0.545**	0.336	0.494	**0.309**	0.176	0.269

对于祖先节点-后代节点关系,本文方法能够在共 15 个子树中的 9 个上获得最优的结果,并且在平均正确率上比 Heymann 方法和 Schmitz 方法分别提高了 13.3%和 4.0%.Heymann 方法倾向于生成较长但质量较低的关系路径,故其在祖先节点-后代节点关系上的性能较差.与 Schmitz 方法相比,本文方法通过在向概念层次结构中加入新节点时同时检查其与所有祖先节点的包容关系的有效性以防止产生不合理的关系路径,有效地提高了性能.

4.4 基于标准本体的评价

人工评价在一定程度上受到评价人员的主观因素的影响,并且规模受到限制.因此,需要一种能够自动地对本体质量进行评价的方法.基于标准本体的评价方法通过将导出本体与预先指定的标准本体进行比较,因此能够客观全面地评价本体质量.

4.4.1 评价方法

本文采用 ODP(Open Directory Project, http://dmoz.org/)的概念层次结构作为标准本体.ODP 是一个免费、公开、由志愿者维护的 Web 目录.ODP 的概念层次结构中的每一个节点都代表一个标注有主题(如 Sports、Arts 等)的 URL 集合.由于 ODP 是由相互协作的志愿者使用相对自由的词汇创建并

维护的，因此与由专家使用受控的词汇创建的本体相比，其与从社会性标注数据中导出本体在概念范围和结构特征上有较大的相似性.

使用拓展精确率、召回率和 F-值等概念得到的指标[15]对本体的概念词汇层和层次结构层进行评价. 词汇精确率(LP)、词汇召回率(LR)和词汇 F-值(LF)表征了在不考虑层次结构因素的情况下，导出本体和标准本体在概念词汇上的相似程度[15]. 对给定的导出本体$\mathcal{O}_{\mathrm{I}}=\{C_{\mathrm{I}},root,\leq_{C_{\mathrm{I}}}\}$和标准本体$\mathcal{O}_{\mathrm{G}}=\{C_{\mathrm{G}},root,\leq_{C_{\mathrm{G}}}\}$，它们被定义为

$$LP(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})=\frac{|C_{\mathrm{I}}\cap C_{\mathrm{G}}|}{|C_{\mathrm{I}}|},$$

$$LR(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})=\frac{|C_{\mathrm{I}}\cap C_{\mathrm{G}}|}{|C_{\mathrm{G}}|},$$

$$LF(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})=\frac{2\cdot LP(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})\cdot LR(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})}{LP(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})+LR(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})}.$$

结构精确率(TP)、结构召回率(TR)和结构 F-值(TF)表征了导出本体和标准本体在概念层次结构上的相似程度[15]. 对于不同的评价目标，存在不同形式的 TP 的定义. 本文采用基于共同语义上下位集(common semantic cotopy，csc)的 TP[15]作为评价指标. 具体地，对给定的两个本体$\mathcal{O}_1=\{C_1,root,\leq_{C_1}\}$和$\mathcal{O}_2=\{C_2,root,\leq_{C_2}\}$，概念 c 的共同语义上下位集定义为其在两个本体中所有的共同的祖先节点和后代节点构成的集合，即

$$csc(c,\mathcal{O}_1,\mathcal{O}_2)=\{c'\mid c'\in C_1\cap C_2\wedge(c'\prec_{C_1}c\vee c\prec_{C_2}c')\}.$$

对给定的导出本体$\mathcal{O}_{\mathrm{I}}$和标准本体$\mathcal{O}_{\mathrm{G}}$，概念 c 的局部结构精确率(tp)定义为

$$tp(c,\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})=\frac{|csc(c,\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})\cap csc(c,\mathcal{O}_{\mathrm{G}},\mathcal{O}_{\mathrm{I}})|}{|csc(c,\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})|}.$$

最后使用$\mathcal{O}_{\mathrm{I}}$和$\mathcal{O}_{\mathrm{G}}$中共同的概念的 tp 来计算 TP、TR 和 TF，即

$$TP(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})=\frac{1}{|C_{\mathrm{I}}\cap C_{\mathrm{G}}|}\sum_{c\in C_{\mathrm{I}}\cap C_{\mathrm{G}}}tp(c,\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})$$

$$TR(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})=TP(\mathcal{O}_{\mathrm{G}},\mathcal{O}_{\mathrm{I}}),$$

$$TF(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})=\frac{2\cdot TP(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})\cdot TR(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})}{TP(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})+TR(\mathcal{O}_{\mathrm{I}},\mathcal{O}_{\mathrm{G}})}.$$

4.4.2　评价结果

导出本体中包含很多不常用的标签，且它们中的很多都位于本体的第一层. 如果直接将导出本体和标准本体进行比较，对所有的算法都将获得非常差的评价结果，使得算法之间的比较缺乏意义. 因此，将通过比较的部分限制在有限的子树内，可以使得评价结果更加合理. 这些子树与在第 4.3 节中选择的子树相同，能够代表大部分的热门概念. 另外，还对由这些子树构成的相对“完整”的本体和标准本体中相应的子树集合进行了对比，比较的结果在标识为“全部”的行中给出.

表 2 给出了基于标准本体的概念词汇层评价结果. 从表 2 中可以看出，本文方法在共 15 个子树中的 11 个上获得最优的词汇精确率，而其它的方法则在词汇召回率上性能较好. 利用标签普遍性排序和自顶向下的层次聚类算法，本文方法倾向于抽象出较少但信息量较大的概念，因此获得了较高的词汇精确率和较低的词汇召回率. 当使用词汇 F-值来平衡词汇精确率和词汇召回率时，本文方法在 8 个子树上获得了最优的结果，且在整体比较上，本文方法要优于其它两种算法. 这表明本文方法能够很好地在词汇的精确率和召回率上做出折衷.

表 2　基于标准本体的概念词汇层评价结果

根节点	词汇精确率(LP)			词汇召回率(LR)			词汇 F-值(LF)		
	本文方法	Heymann	Schmitz	本文方法	Heymann	Schmitz	本文方法	Heymann	Schmitz
sport	0.7838	**0.9375**	0.7500	**0.0504**	0.0261	0.0470	**0.0948**	0.0508	0.0884
scienc	**0.5517**	0.2588	0.4583	0.0170	0.0626	**0.0934**	0.0330	0.1009	**0.1552**
news	**0.2647**	0.0615	0.1067	0.0677	0.0602	**0.3233**	0.1078	0.0608	**0.1604**
program	**0.4162**	0.2687	0.3333	**0.3404**	0.0851	0.0260	**0.3745**	0.1293	0.0482
histori	0.3333	**0.5455**	0.3919	0.0455	0.0606	**0.1465**	0.0800	0.1091	**0.2132**
book	**0.1429**	0.0545	0.0694	0.0108	**0.0323**	0.0269	0.0200	**0.0405**	0.0388
cultur	**0.2500**	0.0065	0.1429	0.1333	0.0667	**0.2000**	**0.1739**	0.0118	0.1667
comput	0.5532	0.2058	**0.5753**	0.0173	**0.0334**	0.0280	0.0336	**0.0574**	0.0534
game	**0.5621**	0.2252	0.5183	**0.1466**	0.0525	0.1312	**0.2326**	0.0851	0.2094
educ	**0.0667**	0.0062	0.0130	0.0204	**0.0612**	0.0204	**0.0312**	0.0112	0.0159
resourc	**1.0000**	0.0030	0.3333	0.0526	**0.5789**	0.0526	**0.1000**	0.0061	0.0909
media	**0.5000**	0.0147	0.2857	0.0938	**0.2188**	0.0625	**0.1579**	0.0275	0.1026
shop	**0.4828**	0.2627	0.4485	**0.0938**	0.0297	0.0584	**0.1571**	0.0533	0.1033
graphic	**1.0000**	0.0500	0.0526	0.0208	0.0208	**0.0417**	0.0408	0.0294	**0.0465**
health	0.5000	0.4583	**0.5542**	0.0459	0.0531	**0.1111**	0.0841	0.0952	**0.1851**
全部	**0.4864**	0.0647	0.2945	0.0517	**0.0618**	0.0526	**0.0935**	0.0632	0.0893

表 3 给出了基于标准本体的层次结构层评价结果. 从表 3 中可以看出,本文方法获得了很高的结构精确率和相对较低的结构召回率. 在第 3.1 节中的分析指出,绝大部分的层次结构错误是由包容关系中的噪音引起的. 在本文方法中,根据 3.1 节中的标签包容层次模型,通过结合标签普遍性排序和自顶向下的凝聚式聚类方法,解决了不一致和不相关的包容关系的问题,从而获得了很高的结构精确率. 然而,这些过程也使得更多的标签由于无法找到合适的父节点而最终成为根节点的子节点,从而降低了本文方法的结构召回率. Heymann 方法生成了最深的层次结构,从而使更多的标签出现在本体中较为靠下的层中,使其结构召回率较高,但结构精确率与其它方法差距较大. 当使用结构 F-值来平衡结构精确率和结构召回率时,本文方法在 7 个子树上获得了最优的结果. 在整体比较上,本文方法要优于其它两种算法.

表 3 基于标准本体的层次结构层评价结果

根节点	结构精确率(*TP*)			结构召回率(*TR*)			结构 F-值(*TF*)		
	本文方法	Heymann	Schmitz	本文方法	Heymann	Schmitz	本文方法	Heymann	Schmitz
sport	**0.8284**	0.4110	0.7915	0.3604	**0.4334**	0.3682	0.5023	0.4219	**0.5026**
scienc	**0.7024**	0.2819	0.5792	0.3374	**0.3798**	0.2793	**0.4558**	0.3236	0.3769
news	**0.6869**	0.1342	0.6673	0.2891	0.3064	**0.3432**	0.4069	0.1866	**0.4532**
program	**0.7126**	0.2716	0.4230	0.3065	**0.3189**	0.2875	**0.4287**	0.2933	0.3423
histori	**0.9205**	0.3834	0.8994	0.4049	**0.5545**	0.4046	**0.5624**	0.4534	0.5581
book	0.3590	0.0996	**0.5890**	0.3140	**0.4398**	0.4393	0.3350	0.1624	**0.5033**
cultur	0.5714	0.0000	**0.7000**	**0.2857**	0.0000	0.2810	0.3810	—	**0.4010**
comput	**0.6265**	0.1322	0.5120	0.3327	0.3220	**0.3466**	**0.4346**	0.1875	0.4133
game	**0.9198**	0.2449	0.9192	0.2850	0.2865	**0.2901**	0.4351	0.2641	**0.4410**
educ	0.0000	**0.2514**	0.0000	0.0000	**0.2861**	0.0000	—	**0.2676**	—
resourc	0.0000	**0.0510**	0.0000	0.0000	**0.6467**	0.0000	—	**0.0945**	—
media	0.5000	0.1544	**0.5833**	0.2357	**0.6429**	0.2639	0.3204	0.2489	**0.3634**
shop	0.7731	0.2274	**0.8379**	**0.3649**	0.3556	0.3591	0.4958	0.2774	**0.5027**
graphic	**0.5135**	0.0000	0.0000	**0.2589**	0.0000	0.0000	**0.3443**	—	—
health	**0.8546**	0.2958	0.4624	0.4928	0.5200	**0.5346**	**0.6251**	0.3770	0.4959
全部	**0.8091**	0.2458	0.7173	0.3505	**0.3719**	0.3519	**0.4891**	0.2960	0.4722

5 结束语

本文提出了一种基于社会性标注的本体学习算法. 根据标签包容关系层次模型,首先使用一种集合论方法来发现标签包容关系,并使用标签普遍性排序和自顶向下的凝聚式聚类算法来解决不一致和不相关的包容关系的问题. 通过在 Delicious 数据集上进行实验,结合定性分析、人工评价和基于标准本体的评价,验证了本体提出的方法的有效性. 本体学习算法中的标签包容关系发现、标签普遍性排序和概念层次结构生成等环节均与本文提出的标签包容关系层次模型非常适合,这是本文提出的算法能够获得很好性能的主要原因.

在本文提出的本体学习算法中存在一些值得讨论的问题. 第 3.1 节中假定标签空间中的标签与概念空间中的概念是一一对应的. 这个假定简化了本体学习算法,但忽略了实际存在于社会性标注数据中的一些问题. 首先,不同的标签可以被用来表示相同的概念. 对于这一问题,最直接的解决方法是使用词典等知识源(如 WordNet[18])将同义标签映射到单一的概念上. 另一种更为复杂的方法是,使用恰当的相似度指标对标签聚类以发现同义标签,并将标签聚类映射到概念空间中. 其次,相同的标签可以用来表示不同的概念. 解决这一问题可能更加困难,且目前在本体学习领域的相关工作很少. 然而,如果将这一问题看作为词义消歧问题,则可以利用已有的相关技术[19-21],结合社会性标注数据的特点来解决它.

未来的工作主要包括:(1) 研究更为恰当的模型来揭示标签空间的结构特征. 本文提出的标签包容层次关系模型中的假设虽然合理有效,但并未经过严格的实验或理论验证. 如果能对这些假设进行检验,则在此过程中可能获得更为有用的启示;(2) 研究在标签空间中挖掘更多语义信息的方法. 具体来说,研究在社会性标注数据中的标签语义消歧问题和对标注数据的概率建模问题等,以期加深对社会性标注数据的理解.

参 考 文 献

[1] Gruber T R. Toward principles for the design of ontologies used for knowledge sharing. International Journal Human-Computer Studies, 1995, 43(5-6): 907-928

[2] Cimiano P. Ontology Learning and Population From Text: Algorithms, Evaluation and Applications. Heidelberg, Germany: Springer, 2006

[3] Sanderson M, Croft B. Deriving concept hierarchies from text//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'99). Berkeley, CA, USA, 1999: 206-213

[4] Golder S A, Huberman B A. Usage patterns of collaborative tagging systems. Journal of Information Science, 2006, 32(2): 198-208

[5] Halpin H, Robu V, Shepherd H. The complex dynamics of collaborative tagging//Proceedings of the 16th International Conference on World Wide Web(WWW'07). Banff, Alberta, Canada, 2007: 211-220

[6] Mika P. Ontologies are us: A unified model of social networks and semantics. Web Semantics: Science, Services and Agents on the World Wide Web, 2007, 5(1): 5-15

[7] Heymann P, Garcia-Molina H. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Stanford University: Technical Report: 2006-10, 2006

[8] Schmitz P. Inducing ontology from flickr tags//Proceedings of the Collaborative Web Tagging Workshop (WWW'06). Edinburgh, Scotland, UK, 2006

[9] Schmitz C et al. Mining association rules in folksonomies//Proceedings of the 10th Conference of the International Federation of Classification Societies(IFCS'06). Ljubljana, Slovenia, 2006: 261-270

[10] Plangprasopchok A, Lerman K. Constructing folksonomies from user-specified relations on flickr//Proceedings of the 18th International Conference on World Wide Web (WWW'09). Madrid, Spain, 2009: 781-790

[11] Tang J et al. Towards ontology learning from folksonomies//Proceedings of the 2009 International Joint Conference on Artificial Intelligence(IJCAI'09). Pasadena, CA, USA, 2009: 2089-2094

[12] Wille R. Restructuring lattice theory: An approach based on hierarchies of concepts//Proceedings of the 7th International Conference on Formal Concept Analysis (ICFCA'09). Darmstadt, Germany, 1982: 314-339

[13] Dean J, Ghemawat S. Mapreduce: Simplified data processing on large clusters. Communications of the ACM, 2008, 51(1): 107-113

[14] Brank J, Grobelnik M, Mladenic D. A survey of ontology evaluation techniques//Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD'05). Ljubljana, Slovenia, 2005: 166-169

[15] Dellschaft K, Staab S. Strategies for the evaluation of ontology learning//Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. Amsterdam, The Netherlands, 2008: 253-272

[16] Brewster C et al. Data driven ontology evaluation//Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, 2004: 164-168

[17] Snow R, Jurafsky D, Ng A Y. Semantic taxonomy induction from heterogenous evidence//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics(ACL'06). Sydney, Australia, 2006: 801-808

[18] Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge, MA, USA: The MIT Press, 1998

[19] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods//Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95). Cambridge, MA, USA, 1995: 189-196

[20] Mihalcea R. Using wikipedia for automatic word sense disambiguation//Proceedings of the 9th American Chapter of the Association for Computational Linguistics (NAACL-HLT'07). Rochester, NY, USA, 2007: 196-203

[21] Navigli R. Word sense disambiguation: A survey. ACM Computing Surveys, 2009, 41(2): 1-69

LIU Kai-Peng, born in 1981, Ph. D. candidate. His current research interests include information retrieval, data mining and machine learning.

FANG Bin-Xing, born in 1960, Ph. D., professor, member of Chinese Academy of Engineering. His current research interests include information security, information retrieval and distributed systems.

Background

The folksonomies built from the large-scale social annotations made by collaborating users are perfect data sources for bootstrapping Semantic Web applications. Extracting the emergent semantics from folksonomies becomes an attractive research topic. Specifically, many efforts have been conducted into learning lightweight ontologies from folksonomies. Unlike the ontologies manually created by domain experts with ontology editing tools, those automatically induced from the collaborative knowledge of the folks have the advantages of inexpensive to generate, dynamically evolving in time, and easy to be deployed in real-world systems.

In this paper, the authors develop an ontology induction approach to harvest the emergent semantics from folksonomies. They propose a latent subsumption hierarchy model to uncover the hierarchical structures of tag space. They address the problem of noisy subsumptions while turning a subsumption graph into a concept hierarchy and propose a random walk based generality ranking mechanism to settle it. They propose a simple yet effective agglomerative hierarchical clustering algorithm to generate the concept hierarchy. They conduct experiments on a dataset collected from a real-world system to evaluate the proposed approach. The experimental results show a competitive performance of the proposed approach.

This work was partially supported by the National Natural Science Foundation of China under grant Nos. 60703014, 60933005, the National Basic Research Program (973 Program) of China under grant No. 2011CB302605 and the National High Technology Research and Development Program (863 Program) of China under grant Nos. 2006AA010105-02, 2007AA01Z416, 2007AA01Z442 and 2009AA01Z437. These projects aim to study the mechanisms within a virtual computing environment and build a Web information retrieval and data mining system with an emphasis on network monitoring.