高技术通讯 2012 年 第22卷 第1期:48 ~ 54

doi:10.3772/j.issn.1002-0470.2012.01.008

基于社会标注的 Web 资源语义聚类研究①

杨 鲲②*,**,*** 马慧芳*,** 史忠植*

(*中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

(**中国科学院研究生院 北京 100049)

(***中国计量科学研究院 北京 100130)

摘 要 在深入分析社会标注系统中用户、标签及被标注 Web 资源之间的关联关系的基础上,提出了基于用户标签的 Web 资源语义描述获取算法,并基于所获取的 Web 资源语义描述及其与用户之间的关联关系,利用一种迭代的聚类算法对社会标注系统中的 Web 资源进行基于语义的聚类,该聚类算法通过迭代不断加强被聚类资源间的一致性信息,从而能够克服传统聚类算法所面临的数据稀疏以及性能问题。研究表明,对 Web 资源所处环境的各种关联关系的深入分析,能够帮助用户更好地理解和操作相关 Web 资源,尤其是对于本身特征不充分或难以获取的 Web 资源来说,关联关系的分析研究具有十分重要的意义。

关键词 社会标注,语义抽取,语义聚类算法,广义关联

0 引 言

社会标注服务系统是指目前广泛应用的允许用户对各种 Web 资源(如网页、图片、视频等)进行自由标注,从而方便用户管理并与他人共享这些 Web 资源的各种 Web 服务系统。当前用户从社会标注服务系统中获取 Web 资源主要通过基于关键词搜索和标签云视图两种方式。社会标注作为一种被实践证明的有效的 Web 资源组织方式,其用户群体迅速扩大,被标注的 Web 资源对象以惊人的速度增长。为帮助用户高效便捷地浏览日渐庞大的 Web 资源,近年来研究人员进行了卓有成效的研究。文献[1,2]探索了社会标注系统作为一个复杂系统的基本分布特性;文献[3-7]利用社会标注提供的语义信息改善了搜索结果,并提供了个性化的搜索方法;针对基于标签云视图的 Web 资源查找方式标签数量庞杂、浏览效率低下的不足,文献[8-10] 对社会标注的语义进行了探索和组织,为用户提供了一种高效的基于语义的层次化浏览方式;文献[11] 从社会标注中获取涌现语义,并基于获取的涌现语义来发现和搜索共享的 Web 页面标签;文献[12]提供了一种使得用户能观察系统的标注演化过程和进行交互的机制;文献[13]把图片的用户标签和视觉特征相结合,让用户非常直观地浏览图片。这些研究极大改善了社会标注系统的用户浏览和查找的体验,但是这些研究主要还是关注的社会标注本身的结构化展示,被标注的 Web 资源仍然是以平面无结构的方式呈现,当搜索结果中包含的 Web 资源数量庞大且包含多个主题时所遇到的困境仍然有待解决。

社会标注系统中包含网页、图片和视频等多种 Web 资源,目前对这些 Web 资源的分类聚类大都基于其自身的特性,但通常 Web 资源本身特征的表达能力有限,无法挖掘出蕴含在 Web 中的有用的知识,尤其是对图像和视频等其底层特征和其语义理解之间存在巨大鸿沟的 Web 资源。本文尝试通过对社会标注系统涉及到的多种对象实体集合之间的关联关系的深入分析,进一步挖掘蕴含在其中的知识从而获得对 Web 资源的深入理解,并基于这些关联关系对 Web 资源进行基于语义的聚类,从而对社会标注系统中的 Web 资源进行有效组织,使用户能够获得对 Web 资源集合更加全局和直观的认知,并高效和便捷地获取 Web 资源。

本研究利用社会标注系统所涉及的各种关联关

① 863 计划(2007AA01Z132),国家自然科学基金(60435010),973 计划(2007CB311004)和国家科技支撑计划(2006BAC08B06)资助项目。

② 女,1978 年生,博士生;研究方向:人工智能,语义网,服务计算;联系人,E-mail: yangkun@ics.ict.ac.cn

(收稿日期:2010-05-24)

系获取了被标注Web资源的语义信息,并对其中的Web资源集合进行了基于语义的聚类,而且通过实验验证了所获取的资源语义描述的有效性和语义聚类方法的有效性。

1　基于社会标注系统的Web资源语义获取

社会标注系统中的标签是由具有不同知识背景和不同认知能力的人们从各自不同的角度对Web资源进行标注而产生的,这为获取Web资源全面的语义信息以达到对其深入正确理解提供了可能性,也为理解图像和视频这些在语义理解和底层特征之间存在语义鸿沟的Web资源提供了新的途径。社会标注系统中没有集中控制,它的这个特点给人们提供充分自由的同时也带来了其他的应用问题。由于缺乏类似于本体这样严格的集中的词汇控制,在社会标注系统中不可避免地存在同义词和一词多义现象;同时由于社会标注系统中提供标注的人不再仅仅局限于有限的专家而是所有的用户,这些用户由于知识背景和认知能力的差异导致所提供的标注质量参差不齐。如何从用户提供的标注中获取高质量的Web资源语义描述是需要解决的问题。

1.1　Web资源语义获取的可能性

社会标注系统拥有复杂系统的相关特征,比如大量的用户,缺乏集中的协调控制,以及非线性的动态性。一般来说,这种类型的系统将会随时间延续产生一个被称为幂律分布的稳定分布。文献[1]的研究结果表明,当一个Web资源被标注100次左右,与该资源对应的标签集的分布趋于稳定,除非某个标签的语义随着时间的流逝发生了新的变化,这个稳定的分布会被打破,但会随着用户标注的次数增加再次趋于稳定。文献[2]的进一步研究表明,这个稳定的分布是幂律分布。标签的幂律分布特性表明,用户总是使用他们认为能够较好描述相应Web资源语义的标签来对其进行标注,高质量的标签被越来越多地重复使用。标签使用频度的高低能够在一定程度上反映该标签对相应Web资源语义描述的适合程度,这种现象为我们选取高质量的标签提供了可能性,同时也更好地反映了所选取概念的共享性。

1.2　Web资源语义获取算法

我们利用标签的分布符合幂律分布的特性对社会标注服务中的标签集进行了筛选。我们认为好的标签应该具有如下特征:

(1)使用频率高:类似于传统信息检索中的词频,如果一个标签集被很多用户用来标注一个特定Web资源,那么它是垃圾信息的可能性就比较小,就很有可能被新的用户用来标注该资源。

(2)涵盖范围广:每个用户都是从自己的角度和立场来对一个特定的Web资源进行自由标注,使得该Web资源的标签集是从各个不同的方面来对其进行描述,在对标签集进行筛选时应尽可能保留对该Web资源更多方面进行描述的标签。

(3)关联度高:如果某些词经常被很多用户放在一起对Web资源进行标注,表明人们认为这些词连在一起比其中任何一个词能更清晰地表达该Web资源的语义,因此在筛选时应该注意到这种情况,让关联度高的标签同时出现。

(4)一致性高:社会标注系统中没有一个全局的词汇控制,不同的人会使用不同的术语来表达同一个概念,这种差异性主要体现在两个方面,一个是句法方面,如blog, blogs, blogging,另一个是同义现象,如cell-phone和mobile-phone。筛选后的标签集应该具有较高的一致性。

(5)排除特定类型的标签:有些标签虽然出现频率很高,但是没有什么实际意义,比如用户基于个人目的标注的toread,toreference,以及系统自动产生的unfiled:system,这些标签应当被剔除。

基于以上特征我们设计了一个算法来对每一个Web资源的标签集进行筛选。在介绍该算法之前,先介绍算法涉及的几个基本概念:$S(t,o)$是Web资源o的标签t的贡献度评估函数,其初始值是t被用来标注o的频率;$p(t_i \mid t_j;o)$表示某个用户在已经用t_j标注Web资源o的情况下再用t_i来对其进行标注的概率,这个概率表明了标签t_i和t_j之间的关联度,其值可以由同时用t_i和t_j对Web资源o进行标注的用户数和只用t_j对其进行标注的用户数之比值来衡量。$P(t_i \mid t_j)$表示任一个Web资源在被t_j标注的前提下被t_i标注的概率,这个概率表明了标签t_i和t_j所表达概念的重叠度,其值可以用同时使用了标签t_i和t_j的用户数和只使用标签t_j的用户数的比值来衡量。

算法的基本思想是反复从原始标签集中选出具有最高贡献度值$S(t,o)$的标签,每选出一个标签t_i就根据如下原则对剩余标签的贡献度评估函数$S(t,o)$进行更新:为了使被选中的标签集涵盖的范围更广,减小与t_i存在概念重叠的标签被选中的概

率,将每个剩余标签 t' 的贡献度评估函数值更新为 $s(t',o)-p(t'\mid t_i)\times s(t_i,o)$;为了提高那些与 t_i 关联度高的标签被选中的概率,将每个剩余标签 t' 的贡献度评估函数值更新为 $s(t',o)+p(t'\mid t_i;o)\times s(t_i,o)$。

具体的算法描述如下:

Web 资源语义描述获取算法:

输入:T(特定 Web 资源的原始标签集),X(需被剔除的标签集合),K(预先指定的需筛选出的标签个数),$R=\{\}$(结果标签集)

输出:R(结果标签集)

```
1.  T = T - X;
2.  为 T 中的每一个标签 t 计算其 s(t,o)值;
3.  While (T! = empty And |R| < K){
4.  t_i ∈ T 且 S(t_i,o) ≥ S(t_j,o) 对任何 t_j ∈ T 且 j ≠ i 成
    立 //找到具有最大 s(t,o) 值的标签 t_i;
5.  T = T - {t_i} //从 T 中移除选定标签;
6.  For each tag t' ∈ T{
7.  S(t',o) = S(t',o) - P(t'| t_i) × S(t_i,o) + P(t'|
    t_i;o) × S(t_i,o)
8.  }//更新剩余标签的 s(t,o) 值;
9.  R = R ∪ {t_i} //记录选定标签;
10. }
```

在该算法之前,我们用信息检索研究领域常用的 Porter 算法[14] 对 Web 资源原始标签集进行词干还原的预处理,以提高标签集的一致性。如果多个标签经过处理以后具有相同的词形,那么只留下用户使用频度最高的原型,其贡献度评估函数 $S(t,o)$ 的初始值更新为多个标签的使用频度之和。如:某个 Web 资源 o 被 blog 标注 10 次,被 blogging 标注 5 次,被 blogs 标注 3 次,那么经过预处理以后该资源被 blog 标注 18 次。

2 基于社会标注系统的资源聚类算法

社会标注系统中涉及到的实体对象有三类:用户、用户标注和被标注资源,三者之间的关系如图 1 所示。

其中所有的用户集合形成了用户空间 U,用户标注中所包含的所有标签词形成了标注空间 A,所有的 Web 资源形成了 Web 资源空间 R,且每一个 Web资源被一个统一资源标识符(URI)唯一标识。

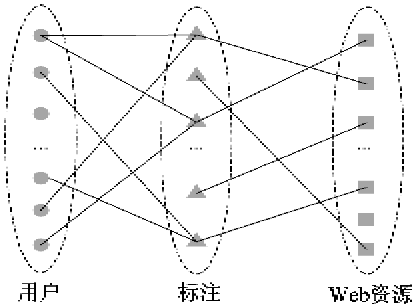

图 1 社会标注系统关系结构图

一个标注实例可以看作是连接用户和某个标签以及连接该标签和被标注 Web 资源的两条边。所有与某个 Web 资源有边相连的标签构成了对该 Web 资源进行描述的标签集合 T,对 T 利用上述算法进行处理,得到该 Web 资源内容的语义描述,然后利用传统的词频-逆文件频率(TF/IDF)技术将其表示为内容向量,并基于内容向量间的余弦相似度计算 Web 资源之间基于内容的相似度 $S_c(x,y)$。

在图 1 的基础上,我们可以获得用户和 Web 资源的直接关系:只要某个用户和某个 Web 资源之间有一条路径相通,则在该用户和该 Web 资源之间建立链接关系。用户空间 U 和 Web 资源空间 R 之间的关系如图 2 所示的二部图 Φ。

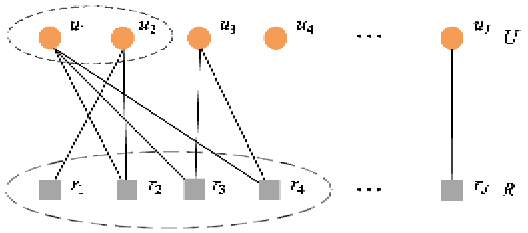

图 2 用户和 Web 资源间关系图

图 Φ 中的任意一个结点其链接属性都可以用一个向量来表示,该向量的每个元素对应于二部图 Φ 中与该结点所在边相对的另一边的所有结点。在这里我们同等地对待每一个用户,赋给每个用户的标注行为相同的权重,那么图 Φ 中的 u_1 和 u_2 这两个结点的链接属性可以分别表示为 $[0,1,1,1,0,\cdots,0]^T$ 和 $[1,1,0,\cdots,0]^T$。这样,相应的相似函数就可以定义为

$$S_l(x,y)=\cos(I_x,I_y) \tag{1}$$

其中 $S_l(x,y)$ 表示基于链接属性的相似度,而 I_x 和 I_y 分别表示结点 x 和结点 y 的链接属性。

基于以上获得的内容和链接的相似函数,我们定义总体相似函数如下:

$$S(x,y) = \alpha S_c(x,y) + (1-\alpha) S_l(x,y) \quad (2)$$

其中 $S(x,y)$,$S_c(x,y)$ 和 $S_l(x,y)$ 分别表示结点 x 和结点 y 的总体相似度、基于内容的相似度和基于链接的相似度,α 表示内容特征和链接特征的相对重要性。

基于链接属性对 Web 对象进行分类的方法被广泛应用于信息检索和协作过滤等领域[15,16],但是传统的方法面临着数据稀疏和性能方面的问题。同样,在社会标注系统中待处理的 Web 资源数量庞大,资源之间的链接信息非常稀疏,而且对 Web 资源的操作需要在合理的时间和空间限制下完成。为了应对这些挑战,我们采用了一种增量聚类算法,该算法在聚类的过程中能够不断加强内容属性和链接属性间的一致性信息,提高聚类效率和效果。算法过程描述如下:

迭代聚类算法过程:

输入:二部图 Φ(包括 Web 资源结点集合和用户结点集合以及他们之间的链接关系)和每个 Web 资源结点从标注中获取的内容属性 f_i。

输出:新的二部图 Φ',其中的两边分别表明用户集合和 Web 资源集合的聚类结果。

1. 基于 Web 资源的内容属性对其进行聚类;
2. do{
3. 基于上一步聚类结果,把每个类中的所有结点合并成一个结点看待,对结点之间的链接关系进行更新;
4. 转到用户结点集并根据更新后的链接属性对其进行聚类;
5. 还原 Web 资源结点集中初始结点及初始链接结构;
6. 将用户结点集的聚类结果的一类合并成一个结点看待,并对链接关系进行更新;
7. 转到 Web 资源结点集并根据更新后的链接属性对其进行聚类;
8. 使用评估函数 F 对 Web 资源结点集的聚类结果进行评估;
9. 还原用户结点集中的原始结点及其初始链接结构;
10. }until 评估函数的值可接受

该算法除第一次以外,每一次的聚类计算都是基于对应边聚类结果的基础上,这样被聚类对象的链接属性向量维数显著减少,从而能够在一定程度上解决传统聚类方法所面临的稀疏问题,同时使时间维和空间维上的性能得到显著提高。

3　实验与结果分析

为了验证从社会标注系统中所抽取的 Web 资源语义描述的有效性,我们将基于该语义描述与基于传统的特征描述方法分别进行聚类,并对结果进行比较。由于 Web 页面的分类聚类是一个相对成熟的研究问题,人们在信息检索和协作过滤领域已经做了大量的工作,有许多的技术手段和标准可以参照,而且在人们已经手工分类的大量 Web 网页中,存在大量被社会标注系统用户标注的页面,为我们验证算法效果提供了很好的条件。而对于图片和视频目前还没类似于开放式分类目录检索系统(open directory project,ODP)[17]那样经过人工分类,且在社会标注系统中拥有大量标注的图片或视频集合,而用人工方式对数量庞大的图片或视频集进行处理和评价也比较困难。因此我们的实验定量评价过程主要针对 Web 页面集合,而对多媒体对象集合采用直观的定性的评价方式。Web 页面的聚类过程中所获得的一些特性能够指导我们更好地实现对图片和视频资源进行基于语义的聚类。

3.1　实验数据集

我们从 ODP 中选择内容差异大且相互交集小的 167 个子类别,并从中收集了大约 137000 个 Web 页面,每个类别中包含的 Web 页面数从 148 到 4702 不等。对每一个 Web 页面,我们在 Delicious(http://del.icio.us)中查找其标注记录。ODP 中的很多 Web 页面在 Delicious 中没有任何标注记录,剔除掉这些无标注记录以及标注用户少于 100 个的 Web 页面以后,剩下 12376 个 Web 页面。我们将这些 Web 页面的首页下载下来用 MSHTML 解析器进行解析,抽取出其中的文本内容,然后利用 Porter 算法[14]对其进行剪枝处理,最后用 TF/IDF 技术将每个 Web 页面的内容表示成关键词向量的形式,我们称之为页面内容特征。对于这 12376 个 Web 页面,根据其在 Delicious 上的标注历史,我们获得每个 Web 页面的基于标注的内容特征以及 Web 页面集合与用户集合之间的链接关系。

3.2　实验评估方法

ODP 中的 Web 页面都是人工分类好的,具有非常明确的类标签,这个人工的分类结果可以作为标准来对我们的聚类结果进行评估。基于信息论[18]中提出的用于对类的一致性或者纯度进行评估的熵的概念,我们采用一种量化的方法来对我们的聚类

结果进行评估。假定一个拥有 N 个类别的 Web 页面集合,其中的每一个页面对象都被赋予了它所在类别的标签,该集合经过聚类算法聚类以后分成了 S 类,A 表示其中的任意一个类,对于每个 $x_i \in A$(其中 $i = 1,\cdots,|A|$),其类别标签 $label(x_i)$ 的取值为 c_j(其中 $j = 1,\cdots,N$)。类 A 的熵定义为

$$Entropy(A) = -\sum_{j=1}^{N} p_j \cdot \log_2 p_j \qquad (3)$$

其中 $p_j = \frac{|\{x_i | label(x_i) = c_j\}|}{|A|}$,$N$ 表示类别数。聚类结果的熵可以表示为所有类的熵的加权和:

$$Total\_Entropy = \sum_{k=1}^{S} \frac{|A|}{n} Entropy(A_k) \qquad (4)$$

其中 S 表示聚类后类的数目,n 表示所有被聚类 Web 页面对象的数目。聚类结果的熵值反映了聚类结果的纯度,根据其定义容易发现,其值越小表明聚类效果越好。考虑极端情况,当每个 Web 页面对象自成一类的时候 $Total\_Entropy$ 的值为 0,因此在下面的实验中我们基于固定的聚类数目来比较聚类效果。

3.3 实验结果

在具体的聚类过程中,由于我们发现在很多时候聚类结果的熵在进行 5 到 6 次迭代时已收敛于一个比较稳定的状态,故每次实验迭代算法进行 10 次迭代足够。每一次迭代过程中对二部图的一边进行聚类时采用的是常用的 K-Means 算法,其迭代次数为 3 次。聚类结果的熵值在迭代过程中可能会呈现一定的扰动,因为此算法不能够确保单调。

为了比较从 Web 页面本身的内容特征和从用户标注获取的语义描述对聚类效果的影响,我们分别以页面内容特征向量和标注内容特征向量作为 Web 页面的内容特征,以用户集合和 Web 页面集合之间的链接关系作为 Web 页面的链接特征,用上述聚类算法对所获得的数据集进行聚类,实验结果如图 3 所示,从中可以看出不同的内容特征对聚类结果的影响经过多次迭代后趋于一致。

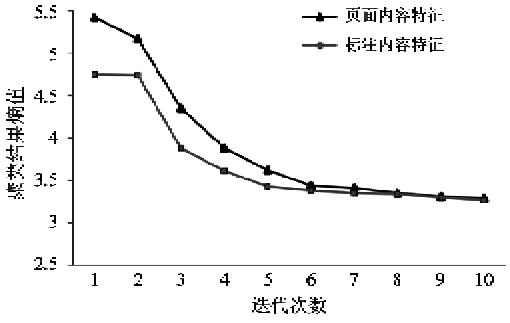

图 3 不同内容特征聚类结果

为了考察内容特征和链接特征的对聚类结果的影响的相对重要程度,在利用式(2)计算总体相似度时,令 α 取不同值使得内容特征和链接特征对总体相似度贡献度的比率为 r(即 $r = \frac{\alpha}{1-\alpha}$),当 r 取不同值时,聚类结果如图 4 所示,从中可以看出当内容特征所占的比重较大,如 r 分别为0.7,0.8,0.9 时,聚类的结果不是很理想,当完全不考虑内容特征($r = 0$)时聚类结果也不是很好,而当 r 取0.05到0.6之间的值时聚类结果较好且极为接近。

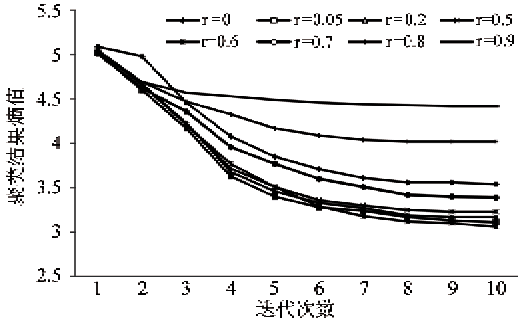

图 4 不同内容特征和链接特征比率聚类结果

从以上两个实验我们发现,在我们的聚类方法中内容特征对聚类结果产生的影响相对较小,这是一个非常重要的特征,特别是对于内容特征难以获取或者获取不精确的 Web 资源的研究非常有意义。当然这个结论是建立在用户集合和 Web 页面集合之间的链接具有一定密度的基础上的。那么这个链接密度到底对聚类效果有什么样的影响呢? 在我们获得的数据集中,每个 Web 页面对象都有上百个用户结点与之链接,我们从这些链接关系中随机选取 1 个,5 个,10 个,25 个,50 个,100 个链接分别作为 Web 页面对象的链接属性进行聚类,结果如图 5 所示。

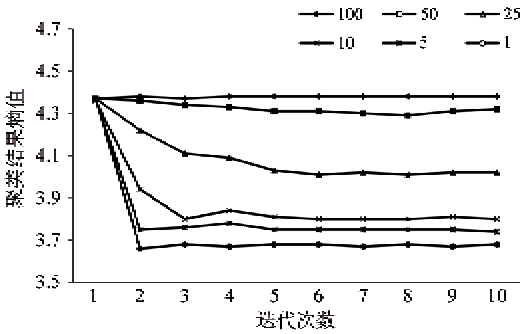

图 5 不同链接数目的聚类结果

从图 5 可以看出,聚类结果随着链接数目的增

多而改善,当链接数目达到 25 个时,聚类结果已经比较好了。

根据上面的实验结论,我们从 Flickr 网站针对查询"Apple"所返回的 2140479 个结果中筛选了涵盖水果和品牌两个主题,被标注次数在 25 次以上,且用户相对集中的 478 幅图片作为我们的测试数据集。

我们利用文献[19]介绍的 CTM 算法对上述数据集作基于视觉特征的聚类,部分聚类结果如图 6 所示,而利用本文算法,令 $r = 0.1$,对上述数据集进行聚类,部分聚类结果如图 7 所示(每行代表聚类结果中的一个类)。与基于视觉特征的聚类结果相比,基于本文算法的聚类结果从直观上印证了算法对于语义聚类的支持。

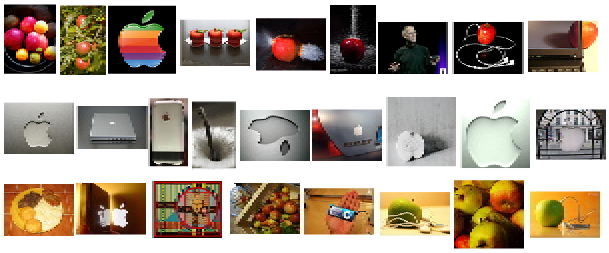

图 6　基于视觉特征的聚类结果

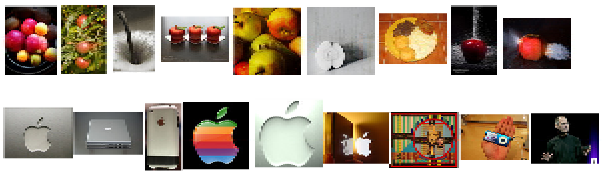

图 7　基于本文算法的聚类结果

3.4　讨论

聚类过程中,如何确定聚类数目一直是个开放性的问题。很多关于聚类的研究都是预先指定聚类数目的,我们的实验评估也是在对实验数据集特性进行分析后预先指定聚类数目的基础上进行的。在实际的搜索过程中,用户不可能了解搜索结果集的特性从而在聚类之前指定聚类数目,所以必须要对如何确定聚类数目进行研究。文献[20]采用光谱聚类技术,利用连续特征值之间的差异来确定聚类数目,但是该方法的性能对聚类对象间的拓扑结构依赖度较高,文献[20]的研究同时表明通过引入其它的特征表示方法能够对此有较大改善,我们将遵循这个思路,通过深入分析社会标注系统中的资源特征表示,力图自主地确定聚类数目。

本文所介绍的算法在上面列举的数据集上初步展示了较好的基于语义聚类的潜力,我们今后的工作将致力于对算法进行更全面的评估和改善,从而使算法的处理能力和鲁棒性更强,为下一步开发实用的基于社会标注的搜索系统奠定基础。

4　结　论

将搜索结果基于不同的语义分主题进行聚类呈现给用户,将会极大改善用户体验,帮助用户迅速便捷地定位所需资源。

本文的初步研究表明,Web 资源所处的环境中的各种关联关系为进一步获取相关知识和加深对 Web 资源的深入理解提供了可能性和有效途径。这些关联关系包括网页中实体与实体的关联,多媒体(图片、视频等)与网页的关联,用户访问日志中的用户与网页、关键字与网页等各种关联,社会标注系统中标签与网页的关联,等等。我们把所有的各种关联关系统一称为广义关联,对这些广义关联关系进行深入挖掘研究将具有重要的应用和研究价值,将会在提高 Web 搜索的质量,拓展数据挖掘的研究以及为图像和视频的分析和理解走出一条新的道路等方面产生重要的推动作用。

参考文献

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

[2] Halpin H, Robu V, Shepherd H. The complex dynamics of collaborative tagging, In: Proceedings of the 2007 International Conference on World Wide Web, Banff, Canada, 2007. 211-220

[3] Zhou D, Bian J, Zheng S, et al. Exploring social annotations for information retrieval. In: Proceedings of the 2008 International Conference on World Wide Web, Beijing, China, 2008. 715-724

[4] Hotho A, Jaschke R, Schmitz C, et al. Information retrieval in Folksonomies: search and ranking. In: Proceedings of the 2006 European Semantic Web Conference, Berlin, Germany, 2006. 411-426

[5] Bao S, Xue G, Wu X, et al. Optimizing web search using social annotations, In: Proceedings of the 2007 International Conference on World Wide Web, Banff, Canada, 2007. 501-510

[6] Noll M G, Meinel C. Web search personalization via social bookmarking and tagging. In: Proceedings of the

2007 International Semantic Web Conference and Asia Semantic Web Conference, Busan, South Korea, 2007. 367-380

[7] Xu S L, Bao S H, Fei B, et al. Exploring folksonomy for personalized search. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, 2008. 155-162

[8] Zhou M W, Bao S H, Wu X, et al. An unsupervised model for exploring hierarchical semantics from social annotations. In: Proceedings of the 2007 International Semantic Web Conference and Asia Semantic Web Conference, Busan, South Korea, 2007. 680-693

[9] Li R, Bao S, Yu Y, et al. Toward effective browsing of large scale social annotations, In: Proceedings of the 2007 International Conference on World Wide Web, Banff, Canada, 2007. 943-952

[10] Brooks C H, Montanez N. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proceedings of the 2006 International Conference on World Wide Web, Edinburgh, UK, 2006. 625-632

[11] Wu X, Zhang L, Yu Y. Exploring social annotations for the semantic web. In: Proceedings of the 2006 International Conference on World Wide Web, Edinburgh, UK, 2006. 417-426

[12] Dubinko M, Kumar R, Magnani J, et al. Visualizing tags over time. *ACM Transaction on The Web*, 2007

[13] Aurnhammer M, Hanappe P, Steels L. Augmenting navigation for collaborative tagging with emergent semantics. In: Proceedings of the 2006 International Semantic Web Conference, 2006. 58-71

[14] Porter M. An algorithm for suffix stripping. *Program*, 1980, 14(3): 130-137

[15] Yongjian Fu K S, Shih M Y. Clustering of web users based on access patterns. In: Proceedings of the 1999 KDD Workshop on Web Mining, San Diego, CA, 1999

[16] Ungar L, Foster D. Clustering methods for collaborative filtering. In: Proceedings of the Workshop on Recommendation Systems at the 15th National Conference on Artificial Intelligence, Madison, USA, 1998

[17] ODP: Open Directory Project, http://dmoz.org/

[18] Cover T M, Thomas J A. Elements of Information Theory. Hoboken: Wiley & Sons, Inc, 1991

[19] Yu H, Li M, Zhang H J, et al. Color texture moments for content-based image retrieval. In: Proceedings of the2002 International Conference on Image Processing, Rochester, USA, 2002. 929-932

[20] Cai D, He X, Li Z, et al. Hierarchical clustering of WWW image search results using visual, textual and link information. In: Proceedings of the 12nd Annual ACM International Conference on Multimedia, New York, USA, 2004. 952-959

Semantic clustering of web resources based on social annotation

Yang Kun[*,**,***], Ma Huifang[*,**], Shi Zhongzhi[*]

([*]The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([**]Graduate University of Chinese Academy of Sciences, Beijing 100049)

([***]National Institute of Metrology, Beijing 1001301)

Abstract

By analyzing the correlations between users, tags and Web resources in social annotation systems, this paper proposes an algorithm to acquire the semantic descriptions of Web resources based on users' tags. And based on the acquired semantic descriptions and the correlations between the descriptions and users, an iterative algorithm is proposed for semantic clustering of the Web resources in social annotation systems. By mutually reinforcing the agreed information between Web resources during the clustering process, the clustering algorithm can tackle, to some extent, the challenges faced by traditional clustering algorithms such as the data sparseness and the performance constraints. The research illustrates the importance of the analysis of the correlations in the environment of Web resources, especially to those whose features are not sufficient or difficult to acquire.

Key words: social annotation, semantic extraction, semantic clustering algorithm, general correlation