基于用户自描述标签的层次分类体系构建方法

刘苏祺[1] 白光伟[1,2] 沈 航[1]
(南京工业大学计算机科学与技术学院 南京 211816)[1]
(南京理工大学高维信息智能感知与系统教育部重点实验室 南京 210094)[2]

摘 要 模式层知识对于语义万维网的发展非常重要,然而当前开放链接数据(LOD)中模式层知识的数量十分有限,为突破这一局限,提出一种基于社交网络中用户自描述标签的层次分类体系构建方法。该方法首先设计基于搜索引擎的标签分块算法,将描述相同话题的标签划分到同一标签块中,然后采用基于半监督学习的标签传播算法挖掘相同标签块中标签间的上下位关系,最后运用基于启发式规则的贪心算法来构建层次分类体系,从而在社交站点中构建出大规模且高质量的层次分类体系。实验结果表明,该构建方法与现有相关工作相比在准确率、召回率以及F值上均有明显提高。

关键词 模式层知识,用户自描述标签,层次分类体系,标签传播
中图法分类号 TP182 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.7.040

Taxonomy Construction Based on User Self-describing Tags

LIU Su-qi[1] BAI Guang-wei[1,2] SHEN Hang[1]
(School of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)[1]
(Key Laboratory of Intelligent Perception and System for High-Dimensional Information of Ministry of Education of China, Nanjing University of Science and Technology, Nanjing 210094, China)[2]

Abstract Knowledge on schema level is vital for the development of semantic Web. However, the number of schema knowledge is limited in current linking open data (LOD). To optimize the issue, this paper proposed an approach for constructing a taxonomy using user self-describing tags in social network. This approach first designs a tag blocking algorithm based on search engine to partition tags into the same block, which describes the same topic. Then, it uses a label propagation algorithm based on the semi-supervised learning to detect hypernym relation between tags in the same block. Finally, it applies a greedy algorithm based on heuristic rules to construct a taxonomy. A large scale and high-quality taxonomy can be constructed after applying the proposed approach in social Web sites. The experimental results show that, compared with the existing related work, the proposed approach performs better in terms of precision, recall and F-score.

Keywords Knowledge on schema level, User self-describing tags, Taxonomy, Label propagation

1 引言

开放链接数据(Linking Open Data)[1]是目前最大的语义数据发布并互连的社区项目,该项目将仅包含网页与网页之间超链接的文档万维网(Web of Documents)逐渐转变成包含大量实体以及实体关系的数据万维网(Web of Data)。目前,共有超过200个知识库存在于开放链接数据中,位于这些知识库中心并与其它知识库进行互连的知识库包括DBpedia[2]、Yago[3]、Freebase[4]等。虽然它们拥有非常丰富的实例层知识,但模式层知识却非常有限。其中,Yago显式定义了模式层知识,包括概念间的包含关系以及属性的定义域与值域,但其准确率并不尽如人意;而Freebase则只是拥有一个非常浅层的包含领域与类别的层次分类体系;虽然DBpedia为了丰富模式层信息利用半自动的方法构建了DBpedia Ontology[5],但是其规模依旧有限。

近年来,随着社交网络的发展,涌现出大量的用户自定义内容,其中最为流行的便是以一组关键词形式出现的用户自定义标签。这类标签往往包括两种用途:第一,用户使用标签对特定的网页或文本进行标注,这类标注可被认为是被标注资源的关键词或简单总结,也称作分众分类法(Folksonomy);

到稿日期:2015-04-10 返修日期:2015-09-01 本文受国家自然科学基金(60673185,61073197),江苏省自然科学基金(BK2010548),江苏省科技支撑计划(工业)(BE2011186),南京邮电大学宽带无线通信与传感网技术教育部重点实验室开放研究基金资助课题项目(NYKL201304),江苏省六大高峰人才基金(第八批)资助。

刘苏祺(1990—),女,硕士生,主要研究方向为语义万维网、社交网络、网络协议以及无线网络编码,E-mail:sue900913@163.com;白光伟(1961—),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为移动互联网、无线传感器网络、社交网络、多媒体网络服务质量等,E-mail:bai@njtech.edu.cn(通信作者);沈 航(1984—),博士,讲师,主要研究方向为社交网络。

第二，用户使用标签进行自描述，此类标签并不是特定网页或文本的关键词，而是依据用户自身兴趣对用户本身特点的一种描述。因为第一种用途的标签可以获取到其本身所描述的文本信息，所以目前从分众分类法中挖掘知识的研究工作[6-8]较多。而对于第二种用途的标签，由于其本身缺乏上下文信息，因此从中进行知识挖掘的难度较大，导致目前鲜有关于从此类用途标签中进行知识挖掘的研究。

针对当前公开发布的数据集缺乏模式层知识的问题，本文尝试利用社交网络中的用户自描述标签构建层次分类体系，从而贡献大规模模式层知识。为此，首先设计基于搜索引擎的标签分块方法，将所有抽取到的标签划分到不同的且存在交集的标签块中的，这考虑了标签可能出现一词多义的情况。然后采用一种基于半监督学习的标签传播算法检测处于同一标签块中的两个标签间是否存在上下位关系，该学习算法综合利用了标签间的语言学相关度、社交相关度和语义相关度。最后运用基于启发式规则的贪心算法构建无环且无冗余边的层次分类体系。实验结果表明，通过本文方法可构建规模大且质量高的层次分类体系，在准确率、召回率以及 F 值上均有明显提高。

本文第 2 节深入分析利用不同资源自动化构建层次分类体系的相关工作；第 3 节提出基于用户自描述标签的层次分类体系的构建方法；第 4 节通过实验对上述方法进行性能评测与分析；最后总结全文。

2 相关工作

目前，关于自动化构建层次分类体系的资源主分为 3 种，分别是非结构化文本、半结构化数据以及社交网络中的分众分类。

在非结构化文本中，自动化挖掘上下位关系并构建大规模层次分类体系最为成功的工作是 Probase[9]。它利用 Hearst 模式[10]与概率模型迭代式地从超过 16 亿个网页文本中构建层次分类体系，其共包含约 260 万个概念与 2000 万个上下位关系，且 Probase 是目前已知的具有最多上下位关系的知识库，但是该知识库并未被公开。

基于半结构化数据构建层次分类体系的来源主要是百科数据中的分类系统与信息框，以及各站点的导航分类等。Wikitaxonomy[11]首次从维基百科的分类系统中，顺序使用了基于语法的方法、基于分类间连通性的方法、基于词法-句法的方法以及基于推理的方法，构建了一个约包含 12 万个分类及 10 万个上下位关系的层次分类体系。KOG[12]利用马尔科夫逻辑网络模型将信息框与 WordNet[13]进行关联并推断信息框间的上下位关系，从而构建出本体中的层次分类体系。Wang 等在 Zhishi. schema[14]中首次提出了开放链接模式的概念，并且利用机器学习算法从超过 50 个站点中的导航分类与标签中挖掘出超过 150 万个上下位关系。它是目前最大的中文层次分类体系。

从社交网络中的分众分类出发，构建标签间的层次分类结构是与本文最为相关的工作。Tang 等人提出一种基于分众分类的本体学习方法[6]，利用主题模型计算了标签在不同主题下的差异，包括上位差异度、合并差异度以及保持差异度。综合利用上述 3 种差异度，自底向上构建了标签的层次分类结构。文献[7]提出了关于社会性标注的本体学习方法，首先使用集合论的方法发现标签间的上下位关系，再利用一种基于随机游走的标签排序算法与自顶向下的层次聚类算法构建标签间的层次分类结构。上述工作均利用用户已标注文本资源辅助挖掘标签间的上下位关系，而在用户自描述标签本身无法获得相关文本资源的情况下，本文无法应用上述方法进行挖掘。

目前已知的与本文工作最为相关的是 Zhou 等人提出的一种用于构建标签层次关系的基于确定性退火算法的(Deterministic Annealing)无监督方法[8]。该工作也是从分众分类中挖掘标签间上下关系，却并未使用用户已标注文本资源，因此可以直接将其运用在本文的挖掘场景中。但是该工作并未处理标签的一词多义问题。在实验部分，将其与提出的方法作了详细的比较。

本文基于用户自描述标签，首先设计一种基于搜索引擎的标签分块算法对标签按话题分块，该算法可以将多义词划分至不同的标签块，基于以上标签块，采用基于半监督学习的标签传播算法检测同一标签块中标签的上下位关系，然后对已标注好上下位关系的标签块运用贪心算法去冗余、去环，构建层次分类体系。最后对本文所设计方法进行性能的分析与评价。

3 层次分类体系的构建方法

本节首先介绍所设计的层次分类体系构建方法的整个框架以及其中包含的组件。如图 1 所示，整个构建框架拥有 3 个组件，分别是标签分块器(Tag Blocker)、上下位关系检测器(Hypernym-hyponym Relation Detector)和层次构建器(Hierarchy Constructor)。

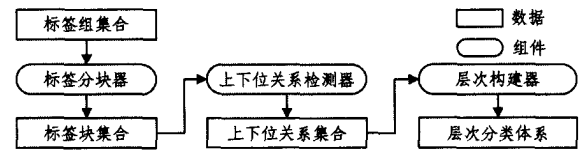

图 1 层次分类体系构建框架

标签分块器的输入为抽取自社交网络站点的用户自描述标签组集合。标签分块器旨在将描述相同话题的标签聚集在一起从而形成多个具有交集的标签块，而交集中的每个标签均具有多种含义。然后将上述输出的标签块集合作为上下位关系检测器的输入，从而检测同一标签块中的各标签之间是否存在上下位关系。最后，将已检测出的上下位关系的集合输入到层次构建器中，从而得到基于用户自描述标签的层次分类体系。下文将详细介绍构建框架中的 3 个组件。

3.1 标签分块

社交网络中用户自描述标签通常以标签组(Tag Group)的形式存在于具体网页中，比如某人微博主页的标签组为{“摇滚”，“hiphop”，“吉他”，“演出”}。由于标签的数量众多，且对于某个特定标签而言，与其真正存在上下位关系的标签并不多，因此直观地采取两两检测标签之间是否存在上下位关系的做法既费时又不合理。本文设计了搜索引擎的标签分块方法，对所有存在于标签组中的标签进行分块，并将描述相同话题的标签放置于同一标签块中。

标签分块方法不同于传统的聚类方法，传统的聚类方法是一种硬划分，即一个标签只可能存在于一个标签块中，从而

导致块与块之间不存在交集。但是具有相同字符串的标签往往表达的含义不尽相同，即存在一词多义的现象，如“苹果”既可以表示一种水果，也可以表示一个品牌，如果进行简单的硬划分，会导致多义词只拥有一种含义，故将传统的聚类方法用于此处并不合理。下面介绍如何利用搜索引擎来确定各标签组中的标签。

(1)计算同一标签组中任意两标签基于上下文的语义距离(Context-based Semantic Distance，CSD)。

众所周知，若要真正理解一个词的含义，需提供该词的使用语境，当“苹果”与“桔子”存在于同一标签组中，那么很有可能该“苹果”为一种水果，而若“苹果”与“平板”存在于同一标签组中，那么可以认为该“苹果”是一个品牌或一个公司。因此该步骤的目的在于在给定的标签组 tg 中，确定给定标签 t_1 的最相关标签 t_2 以辅助确定 t_1 的真正含义。当 t_1 和 t_2 同时存在于不同的标签组中时，其语义距离应该也是不同的，故本文采用搜索引擎的基于上下文的语义距离，它是对于规范化谷歌距离(Normalized Google Distance)[15]的扩展。

定义 1 给定一个标签 t 以及其所在的标签组 tg，那么标签组上下文 $Context(t,tg)=\{tags\ in\ tg\ except\ t\}$。

定义 2 给定两个标签 t_1 和 t_2，以及二者所在的标签组 tg，那么 t_1 和 t_2 基于上下文的语义距离 $CSD(t_1,t_2,tg)$ 的计算方式如式(1)所示：

$$CSD(t_1,t_2,tg)=\frac{\log A-\log B}{\log M-\log C} \tag{1}$$

其中，A,B,C 的计算方式如式(2)—式(4)所示：

$$A=\max\{f(t_1,CT(t_1,tg)-\{t_2\}),f(t_2,CT(t_2,tg)-\{t_1\})\} \tag{2}$$

$$B=f(t_1,t_2,CT(t_1,tg)\cap CT(t_2,tg)) \tag{3}$$

$$C=\min\{f(t_1,CT(t_1,tg)-\{t_2\}),f(t_2,CT(t_2,tg)-\{t_1\})\} \tag{4}$$

其中，CT 是 $Context$ 的缩写，$f(t_1,CT(t_1,tg)-\{t_2\})$ 表示将 tg 中除 t_2 外的所有标签共同作为查询词提交到搜索引擎后所得的结果数量，$f(t_2,CT(t_2,tg)-\{t_1\})$ 表示将 tg 中除 t_1 外的所有标签共同作为查询词提交到搜索引擎后所得的结果数量，$f(t_1,t_2,CT(t_1,tg)\cap CT(t_2,tg))$ 表示将 tg 中所有标签共同作为查询词提交到搜索引擎后所得的结果数量，而 M 表示搜索引擎所索引的总网页数。

在对所有标签组中任意两标签计算完基于上下文的语义距离后，在每个标签组中，对于任意 t，将与其语义距离最短的 t' 作为其最相关的标签。

(2)确定每个标签的话题

为确定给定标签的真正含义，将给定标签和与其最相关的标签共同作为查询词提交到百度知道[16]，即可获得相关问题，与此同时，亦可得到相关问题的所属分类，称为给定标签的相关分类(Related Category)。然后将获得的前 10 页的相关问题所属的相关分类映射到百度知道自身的层次分类体系中的顶层相关分类(Top Related Category)，并进行计数。最后，基于多数投票的思想，将出现次数最多的顶层相关分类作为给定标签的话题。

(3)构建标签块

经过步骤(1)、(2)，已确定所有标签组中所有标签所对应的话题，由于话题来源于百度知道的顶层相关分类，而百度知道共有 14 个顶层相关分类，分别是“电脑/网络”、“生活”、“医疗健康”、“体育运动”、“电子数码”、“商业理财”、“教育/科学”、“社会民生”、“文化艺术”、“游戏”、“娱乐休闲”、“烦恼”、“资源共享”和“地区”，因此可将描述相同话题的标签放置于同一标签块中，最终可得 14 个存在交集的标签块。

3.2 上下位关系检测

本文将标签之间的上下位关系检测问题当作一个二分类问题。对于每个标签块中每个待检测的标签对而言，首先计算标签对中两个标签的多种非对称相关度；然后将这些相关度作为上下位关系检测的特征；最后，运用基于半监督学习的标签传播(Label Propagation)算法[17]判断待检测的两个标签之间是否存在上下位关系。

为了从不同角度确定标签之间的相关度，共设计了 5 种表示标签的方法。定义一个标签 t，最为简单地表示 t 的方式是使用其字符串本身 $s(t)$，然而很多具有上下位关系的标签的字符串之间并不相似，如“苹果”hyponym of “公司”，所以仅利用字符串表示一个标签是远远不够的。根据用户自描述标签的特性易知，标签是从社交站点具体用户页面中抽取得到的，并且一个标签可能与其他标签共同出现在某些用户页面中，所以 t 可以被表示为 $OTS(t)=\{ot_1,ot_2,\cdots,ot_m\}$，而 ot_j 表示第 j 个与 t 共同出现在某一用户页面中的标签。又由于 ot_j 与 t 共现的次数可能不止一次，因此还可利用与 t 共现的标签进一步将 t 以向量的形式表示为 $OTV(t)=\langle ot_1(t),ot_2(t),\cdots,ot_m(t)\rangle$，其中 $ot_j(t)$ 表示标签 ot_j 与 t 共现的次数。

上述 3 种表示方式均只利用了标签本身的特性，而不涉及语义层面。受显式语义分析(Explicit Semantic Analysis)[18]启发，本文还可将给定标签映射到某个知识库的多个概念上，这样即可使用这些概念来表示给定标签。在 3.1 节中，为确定每个标签的真正含义，将给定标签与其最相关的标签共同作为查询词提交到百度知道，从而获得前 10 页相关问题的相关分类，以及每个相关分类的出现次数。此处以百度知道为知识库，以百度知道中的相关分类为概念来表示标签是在语义层面对标签进行表示。因此 t 可以被表示为 $RCS(t)=\{rc_1,rc_2,\cdots,rc_n\}$，其中 rc_k 代表 t 的第 k 个相关分类。除了集合的表示形式外，还可以将 t 以向量的形式表示为 $RCV(t)=\langle rc_1(t),rc_2(t),\cdots,rc_n(t)\rangle$，其中 $rc_k(t)$ 表示 t 的第 k 个相关分类的出现次数。

对应上述 5 种不同的标签表示方式，共计算了标签间的 5 种非对称的相关度。

定义 3 给定一个标签对 (t_1,t_2)，该标签对中两个标签 t_1 与 t_2 的基于字符串的非对称的相关度的计算方式如式(5)所示：

$$Srel(t_1,t_2)=\frac{LCS(s(t_1),s(t_2))}{|s(t_1)|} \tag{5}$$

其中，$|s(t_1)|$ 表示 t_1 的字符串长度，$LCS(s(t_1),s(t_2))$ 表示 t_1 与 t_2 的最长公共子串。

定义 4 给定一个标签对 (t_1,t_2)，该标签对中两个标签 t_1 与 t_2 的基于共现标签集合的非对称的相关度的计算方式如式(6)所示：

$$OTSrel(t_1,t_2)=\frac{|OTS(t_1)\cap OTS(t_2)|}{|OTS(t_1)|} \tag{6}$$

其中，$OTS(t_1)$ 表示与 t_1 共现的标签的个数，$|OTS(t_1)\cap OTS(t_2)|$ 表示与 t_1 共现且与 t_2 也共现的标签的个数。

定义 5 给定一个标签对(t_1,t_2)，该标签对中两个标签t_1与t_2的基于共现标签向量的非对称的相关度的计算方式如式(7)所示：

$$OTVrel(t_1,t_2)=\frac{\sum_{ot\in OTS(t_1)\cap OTS(t_2)} ot(t_1)\cdot ot(t_2)}{\sum_{ot\in OTS(t_1)} ot(t_1)^2} \tag{7}$$

定义 6 给定一个标签对(t_1,t_2)，该标签对中两个标签t_1与t_2的基于相关分类集合的非对称的相关度的计算方式如式(8)所示：

$$RCSrel(t_1,t_2)=\frac{|RCS(t_1)\cap RCS(t_2)|}{|RCS(t_1)|} \tag{8}$$

其中，$RCS(t_1)$表示t_1的相关分类的个数，$|RCS(t_1)\cap RCS(t_2)|$表示t_1与t_2的共同相关分类。

定义 7 给定一个标签对(t_1,t_2)，该标签对中两个标签t_1与t_2的基于相关分类向量的非对称的相关度的计算方式如式(9)所示：

$$RCVrel(t_1,t_2)=\frac{\sum_{rc\in RCS(t_1)\cap RCS(t_2)} rc(t_1)\cdot rc(t_2)}{\sum_{rc\in RCS(t_1)} rc(t_1)^2} \tag{9}$$

基于字符串的相关度实质上是获取了两个待检测标签之间的语言学相关度，基于共现标签的两种相关度是获取了标签间的社交相关度，而基于相关分类的两种相关度则是获取了标签间的语义相关度。为了综合利用上述所有相关度，本文将 5 种相关度作为机器学习的特征进行二分类，从而检测两个标签之间是否存在上下位关系。

在训练分类模型之前，需要对部分数据进行标注，为了降低标注的工作量，本文采用了一种基于半监督学习的标签传播算法。标签传播算法的基本思想是首先构建一个图，图中每个节点代表一个数据点，每条边被赋予一个权值，权值实质上是数据点之间的相似度。之后将已标注数据的类标签(Class Label)传播给其在已构建图中的邻居，从而判定未标注数据的类标签。

本文将每个标签块中待检测上下位关系的标签对作为图中的节点来构建一个完全图，因为上下位关系是一种非对称的关系，所以(t_1,t_2)与(t_2,t_1)是不同的标签对，各自所对应的非对称的相关度也不相同。此处，令$L=\{(x_1,y_1),\cdots,(x_l,y_l)\}$为已标注数据，$U=\{(x_{l+1},y_{l+1}),\cdots,(x_{l+u},y_{l+u})\}$为未标注数据，其中$x_i$为图中节点，即一个标签对，$i\in\{1,2,\cdots,u\}$；已标注数据类标签均已知，为$Y_L=\{y_1,y_2,\cdots,y_l\}$；未标注数据类标签为$Y_U=\{y_{l+1},y_{l+2},\cdots,y_{l+u}\}$，但$Y_U$中类标签均未知。其中：

$$Y_{i\in(1,l)}=\begin{cases}1, & \text{不存在上下位关系}\\ 0, & \text{其他}\end{cases} \tag{10}$$

然后以式(11)计算已构建的图中任意两点间边的权重：

$$\omega_{ij}=\exp\left(-\frac{\|f(x_i)-f(x_j)\|^2}{\sigma^2}\right) \tag{11}$$

其中，$f(x_i)$表示节点x_i的特征向量，假设x_i节点对应的标签对为(t_i^1,t_i^2)，$f(x_i)=\langle Srel(t_i^1,t_i^2),OTSrel(t_i^1,t_i^2),OTVrel(t_i^1,t_i^2),RCSrel(t_i^1,t_i^2),RCVrel(t_i^1,t_i^2)\rangle$，$\sigma$为一常数。

为衡量一个已标注节点的类标签传播给其他节点的概率，此处定义一个$(l+u)\times(l+u)$的概率转移矩阵$\boldsymbol{P}$，其中

$$P_{ij}=P(i\to j)=\frac{\omega_{ij}}{\sum_{k=1}^{l+u}\omega_{ik}} \tag{12}$$

此处，P_{ij}表示节点x_i的类标签传播给节点x_j的概率。与此同时，还定义了一个$(l+u)\times 2$的类标签矩阵$\boldsymbol{Y}$，$\boldsymbol{Y}$中第 1 列表示图中各节点的类标签为 0 的概率，而第 2 列表示各节点的类标签为 1 的概率。对已标注节点x_t而言，$t\in\{1,2,\cdots,l\}$，若其类标签为 1，则$Y_{t1}=0,Y_{t2}=1$；若其类标签为 0，则$Y_{t1}=1,Y_{t2}=0$。此外，根据文献[17]，$\boldsymbol{Y}$对应的未标注节点的初始值并不重要。最后，可按如下步骤执行基于半监督学习的标签传播算法。

1. 计算 $\mathbf{Y}\leftarrow\mathbf{PY}$；
2. 将 $\mathbf{Y}$ 中每一行的概率按比例进行归一处理；
3. 将已标注节点的概率分布恢复到初始值；
4. 从步骤 1 开始重复，直至 $\mathbf{Y}$ 收敛。

当标签传播算法执行完毕后，当未标注节点类标签为 1 的概率大于类标签为 0 的概率时，判定该节点所对应的标签对中的两个标签存在上下位关系；否则，判定不存在上下位关系。

3.3 层次构建

所有已发现的上下位关系可将标签互相连接从而构成一个有向图，有向图中边与边的权值为 3.2 节中判定未标注节点类标签为 1 的概率，即标签间上下位关系成立的概率。但是该有向图存在两个问题：1)如图 2(a)所示，图中有环，出现此问题的原因在于将上下位关系的逆关系判定为正确的上下位关系；2)如图 2(b)所示，图中存在一些冗余边，究其原因是上下位关系存在传递性，就一个合理的层次分类体系而言，应该将冗余的边去掉。因此本文提出了一个基于启发式规则的贪心算法来构建一个层次分类体系，其本质是一个无冗余边的有向无环图，具体规则如下。

规则 1 在构建有向无环图的过程中，当出现环时，须去除环中权值最小的边，以保证无环。

规则 2 在构建有向无环图的过程中，当一个节点到另外一个节点的路径不止一条时，仅保留最长路径(即路径中包含的节点数最多)。

(a)图中有环　　(b)存在冗余边

图 2 初始有向图问题实例

实际存在多种不同的有向图符合无环且无冗余边的要求。为了使所构建的有向无环图的所有边的权值和在所有可构建的有向无环图中最大，使用如算法 1 所示的贪心算法进行构建。

算法 1 无冗余边的有向无环图构建

输入：$R=\{t_1\mathbf{h}\ t_2\mid t_1\in T,t_2\in T\}$
输出：$G=\{V,E\}$

```
1. V←∅,E←∅,C←∅;
2. while R≠∅
3.   if C=∅
4.     choose t1 h t2 from R with the highest weight;
5.     E←t1 h t2;
6.     V←t1,V←t2;
7.     remove t1 h t2 from R;
8.     C←{t1 h t2 | t1 h t2∈R,{t1,t2}∩V≠∅};
```

9. else
10. choose t_1 **h** t_2 from C with the highest weight;
11. add t_1 and t_2 to V, t_1 **h** t_2 to E according to rule 1 and 2 when adding t_1 h t_2 to G;
12. remove t_1 **h** t_2 from C;
13. remove t_1 **h** t_2 from R;
14. C←{t_1 **h** t_2 | t_1 **h** t_2∈R, {t_1, t_2}∩V ≠∅};
15. end if
16. end while

算法 1 中 T 表示标签集，R 表示加权的上下位关系集，t_1 **h** t_2 意为 t_1 是 t_2 的下位词，该算法首先从集合 R 中选取权值最高的一条边以及连接此边的两个标签，再将 R 中各条连接上述两个标签的所有边置入集合 C 中(步骤 3－8)。然后选取 C 中权值最高的一条边，并检测将其加入有向无环图 G 中时是否满足规则 1 与规则 2。若皆满足，则将此边加入 G；反之，则舍弃此边。最后将 R 中各连接 V 中所有节点的边加入 C 中(步骤 10－15)。重复上述所有步骤(步骤 3－15)，直至集合 R 中所有边都在构建过程中被使用过。

4 性能分析与评价

本文的实验数据抽取自网易微博[19]，包括 50000 个不同用户页面中的标签组共 60901 个不重复的标签，平均每个用户页面标签组中的标签数目约为 4.5，所有标签的抽取时间为 2014 年 10 月。

为了检测所提出或采用的 3 种算法的性能，本节按照构建层次分类体系顺序设计了 3 个实验分别对以上算法进行检测评估，并且每一部分分别与现有方法进行对比。

4.1 标签分块算法

为了对标签分块算法进行评测，首先需建立一个已标注好的标准集，由于对所有数据进行手工标注是不现实的，因此从抽取获得的所有标签中随机选取 500 个标签对，且每个标签对中的两个标签源自同一标签组。然后邀请 5 位硕士研究生对该 500 个标签对进行标注，每个标签对可被标注为“相关”和“不相关”。根据多数投票的思想，超过半数研究生将给定的标签对标注为“相关”时，即可认为该标签对中的两个标签相关。根据标注结果，共有 362 个标签对被标注为“相关”，138 个标签对被标注为“不相关”，本文将这 500 个已标注的标签对作为评测标签分块算法的标准集。

针对上述 500 个已标注标签对中的 982 个不重复标签，利用提出的标签分块算法与文献[20]中表现最好的 $K=19$ 的基于 IBV 的 k 均值算法与基于 TABV 的 K 均值算法分别进行划分。然后将划分结果与标准集进行对比评测，评测结果如图 3 所示。

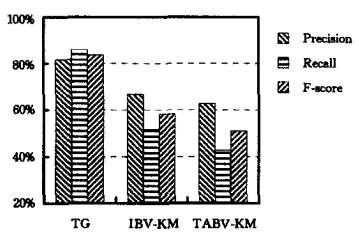

图 3 标签分块算法评测结果

图 3 中 TG、IBV-KM、TABV-KM 分别表示所提出的标签分块(Tag Blocking)算法、基于 IBV 的 k 均值算法与基于 TABV 的 K 均值算法，其中 $K=19$。纵坐标表示的是评测指标，包括准确率(Precision)、召回率(Recall)和 F 值(F-score)。此处认为，若一个标签对中的两个标签均属于同一标签块或标签簇时，则可判定这两个标签相关，反之则判定为不相关。再用上述标签分块或聚类的结果与标准集进行比对，即可得出准确率、召回率与 F 值。根据评测结果，所提出的标签分块算法不论用准确率、召回率还是 F 值评测，其结果均超过现有方法至少 15%，足以证明其有效性。

除了对标签本身的划分进行评测外，由于该标签分块算法的另一个重要特性是可以对多义标签进行合理的划分，因此对多义标签的划分情况进行评测也是必要的。在对所有 60901 个标签利用标签分块算法划分后，取出所有处于标签块间交集的共 3008 个标签，并将每个处于交集的标签与其所在的标签块进行组合，共构成 6910 个形如(t_i, b_j)的数据对，其中 t_i 表示标签，而 b_j 表示标签块。之后，同样让 5 位硕士研究生标注上述数据对，若认为一个数据对(t_i, b_j)中的 t_i 属于 b_j 是正确的，则将该数据对标注为“正确”，反之则标注为“错误”。比如，若标签“苹果”出现在标签块“电子数码”中，则可判定为“正确”，若标签“Web”出现在标签块“医疗健康中”，则可判定为“错误”。此处的标注策略也采用多数投票的思想，经过对所有数据对的标注，该标签分块算法对多义标签的划分正确率达到 81.65%。

4.2 上下位关系检测算法

与 4.1 节相同，为了评测上下位关系检测算法，需先建立一个标准集。此处标准集的抽样机制为随机从每个标签块中选取 200 标签对，共得 2800 个标签对。为保证随机抽取的标签对中的两个标签有一定的相关性，在随机抽取的过程中需保证每个标签对中的两个标签间的 5 种相关度至少有一个高于 0.5。然后同样由 5 名硕士研究生根据多数投票的思想进行标注，可将标准集中每个标签对标注为“上下位关系”与“非上下位关系”。根据标注结果，该标准集中共有 87 个上下位关系与 2713 个非上下位关系。

若要确定标签传播算法中所构建完全图中边的权值，则需确定常数 σ 的值。此处 σ 的确定方法与文献[15]中的基于启发式的方法相同，即利用克鲁斯卡尔算法并且依据图中节点间的欧氏距离构建一棵最小生成树，在构建最小生成树的过程中，当首次将两个类标签不同的已标注节点相连时，记录这两个节点的欧氏距离 d^0，且令 $\sigma=d^0/3$。

为了评测本文使用的标签传播算法，在标准集中的 87 个上下位关系中仅随机保留其中 30 个，2713 个非上下位关系中仅随机保留其中 100 个，然后将其余标签对作为未标注数据。为了更好地体现标签传播算法的效果，本文将以在绝大多数情况下效果最好的监督学习算法[21]支持向量机(SVM)和随机森林(Random Forest)作为对比基准，并采用 5 折交叉验证的方式进行评测。此外，还将标签传播算法与对现有最相关工作中的确定性退火算法[8]进行了对比评测。上述评测的结果如图 4 所示。

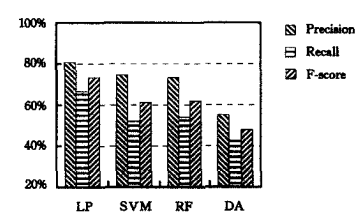

图 4 上下位关系检测算法的评测结果

图4中LP、SVM、RF和DA分别表示标签传播算法、支持向量机、随机森林以及确定性退火算法。不难看出，使用本文提出的5种非对称相关度作为特征的机器学习模型不论在准确率、召回率还是F值上均高于确定性退火算法。此外，本文所利用的基于半监督学习的标签传播算法也比传统监督学习算法中的支持向量机以及随机森林的效果要好。

为了考察在标签传播算法中使用的5种特征各自的重要性，将标签传播算法在标准集中重新运行5次，每次去除一个特征，从而可以得到去除每个特征后F值的下降程度，结果如表1所列。

表1 标签传播算法中各特征重要性评测结果

	Srel	OTSrel	OTVrel	RCSrel	RCVrel
F值下降比例(%)	7.08	2.63	4.46	1.24	4.51

从表1中可以看出，基于字符串的非对称相关度Srel最为重要，这并不难理解。当两个标签的字符串存在交集时，二者很有可能存在上下位关系，比如："水果"与"苹果"、"星座"与"双子座"等。而基于相关分类集合的非对称相关度RCSrel的重要性最低，其原因可能是虽然将给定标签映射到百度知道的相关分类带来了一定的语义信息，但是同时也带来了一些噪音。而基于集合的标签表示方式又将所有相关分类等价看待，使得噪音影响了RCSrel的效果。

在将标签传播算法作用于所有标签之前，首先将每个标签块中的标签两两组合生成所有待检测的标签对。考虑到直接将所有待检测的标签对作为数据点而构建成的完全图的规模太大，以及绝大多数待检测的标签对中的两个标签都是没有上下位关系的，此处预先计算出所有标签对中的两个标签的5种相关度，若给定标签对中两个标签的5种相关度均低于0.5，则认定这两个标签没有关系，并将其从待检测的标签对中删除。经过上述清理后，随机从剩余的99410个待检测的标签对中选取10000个标签对进行标注。经过5位硕士研究生的标注后，其中共有1013个标签对被标注为"上下位关系"，8987个标签对被标注为"非上下位关系"。执行完标签传播算法后，对所获得的11133个不同的上下位关系进行人工标注，正确率达到76.21%。

4.3 层次构建算法

为考察所获得的11133个不同标签间的上下位关系中冗余关系的比例，首先以标签为点、上下位关系为有向边，构建一个初始图，如果两个标签的上下位关系可由初始图中的中间边推断得到，则判定该上下位关系为冗余关系，经过统计，冗余关系的比例如图5所示。

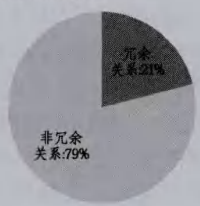

图5 冗余上下位关系比例

另外，在对初始图进行深度优先遍历的过程中，共发现672个环。所以经过上述统计，在构建基于标签的层次分类体系的过程中删除冗余边并去环的工作是非常必要的。

在利用本文提出的基于启发式规则的贪心算法构建层次分类体系之后，所有的环与冗余边均被去除，而该层次分类体系中共有9126个分类(即标签)，8090个上下位关系，上下位关系的正确率较未构建层次分类体系前提高了4.18%，达到80.39%。

结束语 本文提出了一种基于用户自描述标签的层次分类体系构建方法。该方法首先设计了基于搜索引擎的标签分块算法，将描述相同话题的标签划分至同一标签块中，其中可将具有多种含义的标签置于不同标签块的交集中，从而解决了在标签划分过程中由于标签具有歧义可能导致的误划分或漏划分问题。然后采用基于半监督学习的标签传播算法检测同一标签块中任意两标签间是否存在上下位关系。最后运用基于启发式规则的贪心算法将已挖掘得到的标签间的上下位关系的进一步整合为无环且无冗余边的层次分类体系。实验评测结果表明，利用本文提出的方法所构建的层次分类体系中的上下位关系的正确率超过80%，该层次分类体系共包含9126个标签与8090个上下位关系。另外，经过对比，本文提出的标签分块算法与上下位关系检测算法在正确率、召回率以及F值上均优于现有的代表性方法，如IBV-KM、TABV-KM、SVM、RF、DA等。

未来的工作主要包括：1)将基于用户自描述标签的层次分类体系构建方法运用到其他社交站点中，以进一步挖掘大规模模式层知识。2)将构建好的各社交站点的层次分类体系与当前最大的中文开放链接模式Zhishi.schema进行链接，从而将此类模式层知识作为中文开放链接数据的重要补充并公开发布在互联网中。

参考文献

[1] Linking Open Data[OL]. [2014-10-11]. http://linkeddata.org

[2] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A nucleus for a Web of open data[C]// Proceedings of the 6th International Semantic Web Conference. 2007: 722-735

[3] Suchanek F M, Kasneci G, Weikum G. Yago: A large ontology from wikipedia and wordnet[J]. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(3): 203-217

[4] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge [C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008: 1247-1250

[5] Bizer C, Lehmann J, Kobilarov G, et al. Dbpedia-a crystallization point for the Web of data[J]. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(3): 154-165

[6] Tang J, Leung H, Luo Q, et al. Towards ontology learning form folksonomies[C]// Proceedings of the 21st International Joint Conference on Artificial Intelligence. 2009: 2089-2094

[7] Liu Kai-peng, Fang Bin-xing. Ontology Induction Based on Social Annotations[J]. Chinese Journal of Computers, 2010, 33(10): 1823-1834(in Chinese)
刘凯鹏，方滨兴. 基于社会性标注的本体学习方法[J]. 计算机学报，2010，33(10)：1823-1834

(下转第239页)

tute of Computer Science, 2009

[5] Chen Xiao. Chinese Organization Names Recognition Based on Support Vector Machine[D]. Shanghai: Shanghai Jiao Tong University, 2007(in Chinese)
陈霄. 基于支持向量机的中文组织机构名识别[D]. 上海: 上海交通大学, 2007

[6] Che Wan-xiang, Zhang Mei-shan, Liu Ting. Active Learning for Chinese Dependency Parsing[J]. Journal of Chinese Information Processing, 2012, 26(2): 18-22(in Chinese)
车万翔, 张梅山, 刘挺. 基于主动学习的中文依存句法分析[J]. 中文信息学报, 2012, 26(2): 18-22

[7] Tong S, Koller D. Support Vector Machine Active Learning with Applications to Text Classification[J]. The Journal of Machine Learning Research, 2002, 2(1): 45-66

[8] Li S, Xue Y, Wang Z, et al. Active Learning for Cross-Domain Sentiment Classification[C]// Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2013: 2127-2133

[9] Blitzer J, Dredze M, Pereira F. Biographies, Bollywood, Boomboxes and Blenders: Domain Adaptation for Sentiment Classification[J]// ACL, 2012, 31(2): 187-205

[10] Liu K, Zhao J. Cross-Domain Sentiment Classification Using a Two-Stage Method[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 1717-1720

[11] Zhang Hong-yu, Zhou Quan, Hu Xue-gang. Feature Selection for Cross-Domain Sentiment Classification[J]. Pattern Recognition and Aitificial Intelligence, 2013, 26(11): 1068-1072(in Chinese)
张玉红, 周全, 胡学钢. 面向跨领域情感分类的特征选择方法[J]. 模式识别与人工智能, 2013, 26(11): 1068-1072

[12] Wei Xian-hui, Zhang Shao-wu, Yang Liang, et al. Cross-Domain Sentiment Analysis Based on Weighted SimRank[J]. Pattern Recognition and Aitificial Intelligence, 2013, 26(11): 1004-1009 (in Chinese)
魏现辉, 张绍武, 杨亮, 等. 基于加权 SimRank 的跨领域文本情感倾向性分析[J]. 模式识别与人工智能, 2013, 26(11): 1004-1009

[13] Tan S, Wu G, Tang H, et al. A Novel Scheme for Domain-transfer Problem in the context of Sentiment Analysis[C]// Proceedings of the 16th ACM Conference on Information and Knowledge Management. New York: ACM, 2007: 979-982

[14] Jiang J, Zhai C X. Instance Weighting for Domain Adaptation in NLP[C]// Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Stroudsburg, PA: ACL, 2007: 264-271

[15] Dai W, Yang Q, Xue G R, et al. Boosting for Transfer Learning [C]// Proceedings of the 24th International Conference on Machine Learning. Corvallis, Oregon, USA, 2007: 193-200

[16] Zhao Chuan-jun, Wang Su-ge, Li De-yu, et al. Cross-Domain Text Sentiment Classification Based on Grouping-AdaBoost Ensemble[J]. Journal of Computer Research and Development, 2015, 52(3): 629-638(in Chinese)
赵传君, 王素格, 李德玉, 等. 基于分组提升集成的跨领域文本情感分类[J]. 计算机研究与发展, 2015, 52(3): 629-638

[17] Liao X, Xue Y, Carin L. Logistic Regression with an Auxiliary Data Source[C]// Proceedings of the 22nd International Conference on Machine Learning. New York: ACM, 2005: 505-512

[18] Xu Lin-hong, Lin Hong-fei, Pang Yu, et al. Constructing the Affective Lexicon Ontology[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 180-185(in Chinese)
徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185

[19] Chen S, Wang Y. Mining the Emotional Words from Chinese Reviews Based on Part of Speech and Syntax[C]// 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet). IEEE, 2012: 1904-1907

(上接第 229 页)

[8] Zhou M, Bao S, Wu X, et al. An unsupervised model for exploring hierarchical semantics from social annotations[C]// Proceedings of the 6th International Semantic Web Conference. 2007: 680-693

[9] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding[C]// Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. 2012: 481-492

[10] Hearst M A. Automatic acquisition of hyponyms from large text corpora[C]// Proceedings of the 14th Conference on Computational Linguistics. 1992: 539-545

[11] Ponzetto S P, Strube M. WikiTaxonomy: A Large Scale Knowledge Resource[C]// Proceedings of the 18th European Conference on Artificial Intelligence. 2008, 178: 751-752

[12] Wu F, Weld D S. Automatically refining the wikipedia infobox ontology[C]// Proceedings of the 17th International Conference on World Wide Web. 2008: 635-644

[13] Fellbaum C, et al. WordNet: An electronic lexical database[M]. MIT Press, 1998

[14] Wang H, Wu T, Qi G, et al. On publishing Chinese linked open schema[C]// Proceedings of the 13th International Semantic Web Conference. 2014: 293-308

[15] Cilibrasi R L, Vitanyi P M B. The google similarity distance[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383

[16] 百度知道[OL]. [2014-10-11]. http://zhidao.baidu.com

[17] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation[R]. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002

[18] Gabrilovich E, Markovitch S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis[C]// Proceedings of the 20th International Joint Conference on Artificial Intelligence. 2010: 1606-1611

[19] 网易微博[OL]. [2014-10-11]. http://t.163.com

[20] Zhou Jin, Chen Chao, Yu Neng-hai. Tag Clustering Algorithm Using Object-based Feature Vector[J]. Journal of Chinese Computer Systems, 2012, 33(3): 525-530(in Chinese)
周津, 陈超, 俞能海. 采用对象特征向量表示法的标签聚类算法[J]. 小型微型计算机系统, 2012, 33(3): 525-530

[21] Fernández-Delgado M, Cernadas E, Barro S, et al. Do we need hundreds of classifiers to solve real world classification problems? [J]. The Journal of Machine Learning Research, 2014, 15(1): 3133-3181