基于 Web 信息使用改进的无监督关系抽取方法构建交通本体①

马 超

(复旦大学 计算机科学技术学院, 上海 201203)

摘 要: 领域本体是对领域概念及其关系的一种高效合理的展现形式. 在构建领域本体过程中, 常常遇到的问题就是尽管本体概念完备但概念间关系复杂多样导致人工标记关系代价过高. 使用无监督学习的关系抽取算法对包含丰富的领域概念的 web 信息进行抽取解决了这一问题. 然而, 传统的无监督学习的算法没有考虑到“单样例多概念对”的问题, 导致最终抽取的概念关系不完整. 本文利用交通领域的 Web 信息构建本体, 将样例概念关系对权重引入传统的无监督学习方法 Kmeans 中, 解决了此项问题并通过实验证明该算法取得了良好的效果.

关键词: 关系抽取; 本体; 无监督学习; 样例概念关系对权重

Using Improved Unsupervised Relation Extraction Method to Construct Traffic Ontology Based on Web

MA Chao

(School of Computer Science, Fudan University, Shanghai 201203, China)

Abstract: Domain ontology is an efficient and reasonable display form of domain concepts and their relationships. In the process of building domain ontology, the problem often encountered is that ontology concept is complete but concept relations are complex and diverse and artificial tag cost too much. Using unsupervised Relation Extraction algorithm on rich Web information related with domain Ontology concepts solve previous problem. But the traditional method based on unsupervised learning does not take into account the situation of a single sample with more concepts, leading to the final incomplete results. We used Web information in trafffic field to construct ontology, introduced the weight of sample concept relation pair to a traditional unsupervised learning approach-Kmeans to solve this problem and achieved good results through experiments.

Key words: relation extraction; ontology; unsupervised learning; sample concept relation pair weight

本体(Ontology)作为一个哲学概念在人工智能、信息系统领域被 Gruber 定义为“概念化的明确的规范说明”[1]. 本体的种类多种多样, 包含尝试本体、高层本体、领域本体、术语本体、形式本体等. 其中作为专业性的本体的领域本体(Domain Ontology)通过描述特定领域中的概念和概念之间的关系更为高效合理地展示领域知识. 现阶段很多领域本体构建过程中遇到的主要问题就是如何确定概念之间的关系. 通过领域专家手工梳理概念间的关系不仅成本过高而且可能还会有所遗漏, 所以研究通过大量领域文本信息自主学习或半自主学习概念间关系的方法变得尤为迫切.

国内外对于自主学习或半自主学习的关系抽取的研究已经有一定的积累. 可以分为有监督、半监督和无监督三个方向. 有监督的学习算法将关系抽取看做分类问题, 使用支持向量机(Support Vector Machine, SVM)、朴素贝叶斯(Naïve Bayes, NB)、最大熵(Max Entropy, ME)等算法训练分类模型. 半监督的学习方法使用模式学习方法迭代产生关系集合[2]. 但是它们都面临着一个问题, 如果关系的类型没有定义, 那么它们都将无法有效工作. 由此引入无监督的学习方法,

① 收稿时间:2015-04-14;收到修改稿时间:2015-06-08

其包含概念对聚类和关系标记两个过程. 最具代表的就是 Hasegawa[3]提出的无监督关系探索方法, 其使用全联通聚类对同一概念所有样例上下文合并归一后的数据进行聚类并用同一类中频率最高的词语描述该类关系. Rozenfeld B[4]在此基础上通过限制语料库领域类型和使用统计方法过滤掉具有多种关系的概念对解决了概念对具有多种关系的问题. Yan Y[5]基于维基百科文本信息并利用 web 网页的结构特征, 开发了基于模式的组合的无监督学习方法. Bonan Min[6]通过 Ensemble Semantics 算法解决了 web 实体关系抽取中因特征稀疏而产生的一词多义和同义问题, 并且可以高效的从 1470 万关系实例中抽取大量准确的关系. 但它们都存在一个相同的问题, 选择性忽略了一个样例中包含多个概念对的情景.

本文通过在无监督学习方法 Kmeans 中加入样例概念关系对权重, 解决了本体概念明确但概念间关系复杂多样、常规学习方法需要的手工标注训练数据成本高且无法全面覆盖概念间关系、传统无监督学习方法无法满足单样例多概念的情况等问题. 并最终使用此方法从大规模 web 交通信息中构建了交通本体.

1 算法流程

1.1 获取 web 语料库并预处理数据

从 web 中获取交通本体概念的相关信息. 如交通本体概念介绍, 包含交通本体概念的新闻、文章等.

提取语句中所有概念与概念关系. 这样一个样例语句可能包含多个概念关系对. 提取语句词频作为特征向量.

初始化概念关系对相应的样例概念关系对权重. 假设“交通语句”拥有两个概念关系对 $(G_1, G_2, \mathrm{Re}lation_{G_1,G_2})$ 和 $(G_1, G_2, \mathrm{Re}lation_{G_1,G_2})$ 经过预处理便转化为: $(G_2, G_3, \mathrm{Re}lation_{G_2,G_3})$

$$\frac{1}{2}(TF_{w1}, TF_{w2}, TF_{w3}, ..., TF_{wn})(G_1, G_2, \mathrm{Re}lation_{G_1,G_2})$$

$$\frac{1}{2}(TF_{w1}, TF_{w2}, TF_{w3}, ..., TF_{wn})(G_2, G_3, \mathrm{Re}lation_{G_2,G_3})$$

其中 1/2 就是初始化该样例针对特定概念关系的权重.

整合相同概念关系对将具有相同概念关系对的特征累加得到该概念关系 $(G_i, G_j, \mathrm{Re}lation_{G_i,G_j})$ 的综合特征:

$$(\sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{w1,k}, \sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{w2,k}, ..., \sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{wn,k})(G_i, G_j, \mathrm{Re}lation_{G_i,G_j})$$

其中 $|D(G_i,G_j)|$ 表示 $(G_i, G_j, \mathrm{Re}lation_{G_i,G_j})$ 拥有特征向量数量. α_k 表示 k 样例的初始概念关系对权重. 此时所产生的概念关系对中还存在一些两个概念具有多个关系以及不存在的关系, 前者根据相同 (G_i,G_j) 直接取初始权值累加和最高的关系作为该概念关系对 $(G_i, G_j, \arg\max_{R\in \mathrm{Re}lation_{G_i,G_j}}\sum_{i\in R}\alpha_i)$, 后者通过在聚类每轮迭代中剔除权值累加和较少的概念关系对.

1.2 改进的聚类算法

聚类算法基于具有相同语义关系的概念对它们的上下文语境也应相似的假设. 这里使用传统的 Kmeans 算法结合样例概念关系对权重进行聚类. 该算法与 Kmeans 最大的不同之处是在 Kmeans 每次迭代后, 其会针对 $D(G_i,G_j)$ 重新计算所有样例概念对权重 α, 下一轮迭代会运行在在新的综合特征上.

$$(\sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{w1,k}, \sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{w2,k}, ..., \sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{wn,k})(G_i, G_j, \mathrm{Re}lation_{G_i,G_j})$$

下面将主要介绍如何更新样例概念对权重 α 设代价函数:

$$\cos t = \log\left(\frac{\sum_{C_i\in C}\sum_{f\in C_i}((\sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{w1,k}, \sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{w2,k}, ..., \sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{wn,k}) - center_{C_i})^2}{\sum_{C_i,C_j\in Candi\neq j}(center_{C_i} - center_{C_j})^2}\right)$$

该代价函数结合了类间距离和类内距离. 类间距离和类内距离的具体意义将在下一章讲解. 这样便可以通过梯度下降的方法更新 α:

$$\alpha_i = \alpha_i - \lambda\frac{\partial \cos t}{\partial \alpha_i}$$

改进的基于 Kmeans 的无监督聚类算法步骤:

输入:

经过预处理得到的 N 个概念关系对 $(TF_{w1}, TF_{w2}, TF_{w3}, ..., TF_{wn})(G_i, G_j, \mathrm{Re}lation_{G_i,G_j})$,0<*i*,*j*<*NG*(概念数)

聚类个数 K

输出:

K 个簇及其中心点

1 初始化

(1.1)所有概念关系对权重 α 设置为 *1/N*

(1.2)整合相同概念关系对将具有相同概念关系对的特征累加得到该概念关系 $(G_i, G_j, \mathrm{Re}lation_{G_i,G_j})$ 的综合特征:

$$(\sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{w1,k}, \sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{w2,k}, ..., \sum_{k\in|D(G_i,G_j)|}\alpha_k TF_{wn,k})(G_i, G_j, \mathrm{Re}lation_{G_i,G_j})$$

(1.3)从综合特征中, 随机选取 k 个中心点

2 迭代

(2.1)计算获得 K 个簇

针对综合特征, 根据 K 个中心点, 使用欧几里得距离结合概念关系对权重 α, 计算并获得 k 个簇.

$$distance_{i,j}=\sqrt{\sum_{w\in(1,n)}(\alpha_i TF_w-\alpha_j TF_w)^2}$$

(2.1)更新概念关系对权重α

针对单个概念对特征. 更新每个概念关系对权重α

$$\alpha_i=\alpha_i-\lambda\frac{\partial \cos t}{\partial \alpha_i}$$

(2.2)更新中心点

根据 2.1, 2.2 获得的簇及新概念关系对权重α计算新的 K 个中心点.

$$newCenter_k=\frac{(\sum_{i\in Cluster_k}\alpha_i TF_{w1,i},\sum_{i\in Cluster_k}\alpha_i TF_{w2,i},.....,\sum_{i\in Cluster_k}\alpha_i TF_{wn,i})}{|Cluster_k|}$$

(2.3)去除低权重和概念关系对

累加得到所有概念关系$(G_i,G_j,\mathrm{Re}lation_{G_i,G_j})$的$\alpha$权重和.

$$\sum_{i\in(G_1,G_2,\mathrm{Re}lation_{G_1,G_2})}\alpha_i$$

去除α权重和最低的概念关系对.

(2.4)若收敛或达到设置的迭代次数, 结束. 否则继续迭代.

最后, 针对算法输出的 K 个簇及其中心点. 从聚类中 Relation 频次最多的几个 Relation 名中抽象出该聚类概念间关系名称.

2 实验及结果

交通本体概念明确, 共拥有 3145 个概念, 但是无法确定概念间的关系. 本节试验将从 web 信息中抽取交通本体概念间关系并与其它算法对比聚类后果.

本文采用的实验数据是利用《交通大辞典》[7]中的交通术语从互联网爬取的百度百科交通领域相关的文本语料, 共计 1049 篇文档, 14M 数据. 按照第一节算法处理并人工抽象出关系名称后得到以下 25 个概念间关系.

为证明 kmeans+样例概念关系对权重具有更好的聚类效果. 使用实现 Hasegawa[3]方法的传统无监督学习 Kmeans, Bonan Min[6]通过 Ensemble Semantics 算法与 kmeans+样例概念关系对权重进行对比, 设定 k 为 25. 通过比较两种算法的归一化类间距离和归一化类内距离对比他们的聚类效果. 下面给出归一化类间距离、归一化类内距离的定义及试验对比结果.

归一化类间距离: 描述两个聚类的区分程度, 越大表示聚类效果越好. 其公式定义为:

$$interClassDistance=1-\exp(-\frac{2*\sum_{C_i,C_j\in Candi\neq j}(center_{C_i}-center_{C_j})^2}{|C|(|C|-1)})$$

归一化类内距离: 描述一个聚类的紧凑程度, 越小表示聚类效果越好. 其公式定义为:

$$innerClassDistance=1-\exp(-\frac{\sum_{C_i\in C}\sum_{f\in C_i}\frac{(f-center_{C_i})^2}{|C_i|}}{|C|})$$

其中 f 为聚类C_i中的一条综合特征.

表 1 关系列表

关系 ID	关系名称	关系描述
1	乘坐关系	乘坐关系是本体乘客和交通工具之间的关系, 在本体库中具体表示为: 乘客“乘坐”交通工具.
2	位于关系	位于关系可以是多个本体和交通地理本体的关系. 例如交通事故“位于”路口, 或者交通设施“位于”路段等.
3	停靠关系	停靠关系是本体交通工具和交通地理之间的关系, 在本体库中具体表示为: 交通工具“停靠”交通地理, 例如公交车“停靠”公交站点等.
4	去往关系	去往关系也是本体交通工具与交通地理之间的关系, 在本体库中可以表示为: 交通工具“去往”交通地理, 例如出租车“去往”单位等.
5	经过关系	表示本体交通工具与交通地理之间的关系.
6	处理关系	处理关系是本体交通警察和本体交通事故之间的关系, 在本体库中表示为: 交通警察“处理”交通事故.
7	实施关系	实施关系是本体养护工人和本体设施维护之间的关系, 在本体库中表示为: 养护工人“实施”设施维护.
8	属于关系	属于关系是本体公交车与公交车线路, 出租车与出租车线路, 地铁与轨交线路之间的关系, 在本体库中表示为: 公交车“属于”公交车线路, 地铁“属于”轨交线路等
9	影响关系	影响关系是本体交通事件与交通指标之间的关系, 在本体库中表示为: 交通事件“影响”交通指标.
10	搭载关系	搭载关系是本体交通工具与乘客之间的关系, 在本体库中表示为: 交通工具“搭载”乘客.
11	收集关系	收集关系是本体交通设施与交通相关信息之间的关系, 在本体库中表示为: 交通设施“收集”交通相关信息.
12	显示关系	显示关系是本体交通设施与交通指标以及交通信息的之间的关系, 在本体库中表示为: 交通设施“显示”交通指标和交通设施“显示”交通信息等. 例如在信息板上显示附近停车场的信息.
13	行驶于关系	行驶于关系表示本体交通工具与交通地理之间当前相对位置状态的关系. 例如: 公交车“行驶于”路段.
14	靠近关系	靠近关系可以用来表示本体交通地理子本体之间的关系. 例如: 公交站点“靠近”单位.

15	驾驶关系	驾驶关系表示本体驾驶员与交通工具之间的关系. 在本体库中表示为: 驾驶员“驾驶”交通工具.
16	维护关系	维护关系表示本体养护工人与交通设施和交通地理之间的关系. 例如: 养护工人“维护”桥梁, 养护工人“维护”交通设施等.
17	等价于关系	等价于关系表示两个名称不同, 意义完全相同的两个概念本体之间的关系. 例如: 出租车“等价于”计程车, 准点“等价于”正点等.
18	相关关系	相关关系是量标本体之间可能具有的关系, 需要通过分析具体数据来获取, 又可以分为正相关关系和负相关关系. 例如, 出租车运行速度与天气“相关”, 客运收入与客运量“正相关”等.
19	指导关系	指导关系表示需求与规划之间的关系. 在本体库中表示为: 需求“指导”规划.
20	导致关系	导致关系表示违法违章行为与交通事件/事故之间的关系. 在本体库中表示为: 交通违章“导致”交通事故.
21	保障关系	保障关系表示交通法规与交通秩序之间的关系. 在本体库中表示为: 交通法规“保障”交通秩序.
22	规范关系	规范关系表示交通管理与交通运行之间的关系. 在本体库中表示为: 交通管理“规范”交通运行.
23	可换乘关系	可换乘关系表示站点与线路之间的关系. 例如公共汽电车站点“可换乘”轨道交通线路等.
24	基于关系	基于关系表示生活指数与天气之间的关系. 在本体库中表示为: 生活指数“基于”天气.
25	妨害关系	妨害关系表示交通违章与交通秩序之间的关系. 在本体库中表示为: 交通违章“妨害”交通秩序.

表 2 聚类效果对比

	实现 Hasegawa 方法的 kmeans	Ensemble Semantics 算法	kmeans+样例概念关系对权重
归一化类间距离	0.327	0.441	0.495
归一化类内距离	0.108	0.089	0.064

由表 2.2 可知加入样例概念关系对权重的 kmeans 聚类效果的归一化类间距离、归一化类内距离分别为 0.495、0.014. 实现 Hasegawa 方法的 kmeans 聚类效果归一化类间距离比其低了 0.168, 归一化类内距离高了 0.044. Ensemble Semantics 算法类效果归一化类间距离比其低了 0.054, 归一化类内距离高了 0.025. 显而易见, kmeans+样例概念关系对权重较实现 Hasegawa 方法的 kmeans 和 Ensemble Semantics 算法的聚类效果好, 聚的类类内紧凑度高、类间区分度高.

3 总结

本文通过在传统无监督学习方法中加入样例概念权重, 使得可以从单样例多概念对中正确的抽取概念关系. 在实验室中与传统的无监督学习方法对比发现, 该方法聚类效果具有较高的类间距离和较低的类内距离. 并以此方法为基础构建了交通本体.

在从原始语句抽取概念关系对时, 一些不存在的概念关系对也被抽取出来. 现阶段只是在每轮迭代中通过简单的统计剔除累加权重较少的概念关系对, 后续工作将研究使用语法分析方法剔除不存在的概念关系对以提高算法的整体性能.

参考文献

1 Gruber TR. A translation approach to portable ontology specifications. Technical Report, KSL 92-71, Knowledge System Laboratory, 1993: 12–16.

2 Zhang M, Zhou G, Aw A. Exploring syntactic structured features over parse trees for relation extraction using kernel methods. Information Processing & Management, 2008, 44 (2): 687–701.

3 T.H. Discovering Relations Among Named Entities From Large Corpora. Acl Proc. of Annual Meeting on Association for Computational Linguistics. 2004.

4 Rozenfeld B, Feldman R. High-performance unsupervised relation extraction from large corpora. ICDM-06 2006: 1032 – 1037.

5 Yan Y, Okazaki N, Matsuo Y, et al. Unsupervised relation extraction by mining Wikipedia texts using information from the web. Acl Proc. of the Joint Conference of Annual Meeting of the Acl & Intern. 2009, 2:1021–1029.

6 Min B, Shi SM, Grishman R, Lin CY. Ensemble semantics for large-scale unsupervised relation extraction. Proc. of EMNLP -CoNLL 2012.

7 《交通大辞典》编辑委员会.交通大辞典.上海:上海交通大学出版社,2005.