●张玉峰[1]，蔡皎洁[1,2]
(1. 武汉大学 信息资源研究中心，湖北 武汉 430072；2. 孝感学院，湖北 孝感 432000)

基于数据挖掘的 Web 文本语义分析与标注研究*

摘 要：在领域本体已知和文本语义标注主要步骤的基础上，本文用数据挖掘技术实现文本语义信息的获取，提出了文本语义分析与标注的基本思想和处理流程，深入探讨了用聚类分析完成实例分析与标注过程，用关联挖掘和分类方法完成实例间关系的分析与标注过程。

关键词：语义分析；语义标注；数据挖掘；领域本体

Abstract: Based on the known domain ontology and the main semantic annotation steps of text, this paper realizes the acquisition of semantic information from Web text using data mining technology, proposes the basic idea and processing flow of semantic analysis and annotation of text, probes deeply into the analysis and annotation of the instances using cluster analysis, and fulfills the analysis and annotation of the relationships among instances using association mining and classification method.

Keywords: semantic analysis; semantic annotation; data mining; domain ontology

1998 年 12 月，Berners-Lee 首次提出语义网的设想，给出了它的总体框架，从而拉开了对语义网研究的序幕。语义网并非是全新的 Web，而是对现有 Web 的扩展，Web 上的信息更易被机器理解，使得机器和人类能够更好地协同工作。因此，需要在传统 Web 文本上添加语义信息，使其从机器可读提高到机器可理解状态，这种语义信息添加的过程称为语义标注，它是实现语义网的基础。但语义标注是一个复杂的工程问题，其具体的实施过程受多方面因素的影响，目前还没有达成统一的认识。如标注由谁来完成、标注过程中采用什么样的本体、语义标注的基本过程及具体方法等问题都引起国内外专家学者的广泛关注。

1 文本语义标注的基本内容

语义网的成功实现是依靠本体有效性及依据这些本体为 Web 网页标注元数据的过程[1]。本体在一个具体的应用中需要与实例联系起来，而语义标注技术正是能够丰富本体中实例的技术[2]。概括地讲，语义标注是一个在领域本体指导下为文档添加规范化知识表示的过程，该过程分为两大主要步骤[3]：①将文本中与领域本体中概念相对应的词标记出来，作为概念所对应的实例，通常以 RDF 资源形式描述。②找出这些实例间存在的与本体中属性相对应的关系，通常将关联的两个实例及实例间的关系表示为（R_1，P，R_2），即一个 RDF 陈述，P 为 2 个实例 R_1，R_2 间的关系，对应领域本体中的一个属性，即为实例标注过程和实例间关系标注的过程。本文也正是基于这关键的两大步骤，利用数据挖掘展开语义分析与标注过程。

一般来说，语义标注技术可分为 3 种类型：手工标注、半自动标注和自动标注。手工生成语义标注，如 Annotea + Amaya，Yawas，Edutella，SHOE，OntoAnnotea，HTML-A，WebKB，Karina，Mangrove，SMORE。在这些系统中，专家尽可能提供详尽的语义标注但过分依赖于他们的经验，该过程既耗时又昂贵，并且对日益增长的 Web 文档进行语义标注是非常困难的。因此手工标注对文档来说主要是静态的。半自动化语义标注，如 MnM，Melita，OntoAnnotate，Teknowledge，IMAT，这些系统预先假定一定量的手工标注的网页作为系统的训练集，但即使有机器的帮助，这样的标注过程仍然是费力的、耗时的以及易于出错的工作。自动化语义标注的生成，如 AeroDAML KIM，MnM，Magpie，虽然这些系统减轻了人工训练的负担，但是标注结果缺乏可靠性，急需利用文本语义分析的方法获取语义知识[4]。

2 文本语义分析与标注的方法研究

2.1 文本语义分析与标注的基本思想

应用数据挖掘技术研究文本语义分析与标注的基本思想是：首先将选好的 Web 文本集进行预处理，包括分词处理、文本特征标引和词频降维，其中用潜在语义标引方

* 本文为教育部人文社会科学重点研究基地重大项目的研究成果之一，项目批准号：08JJD870225。

法（LSI）的奇异值分解技术（SVD）进行词频降维，生成新的包含特征项间隐含语义关系的新矩阵是重要的一步；其次，在基于新矩阵转置的基础上，实施平面划分法进行特征项（实例）聚类，获取概念集簇，完成实例分析与标注过程；最后，在基于新矩阵的基础上，实施关联规则挖掘方法，发现具有强关联的特征项，运用概率统计方法发现具有等价关系、层次结构关系的特征项。先用人工方法产生上下位关系和整体部分关系的训练集，再用分类方法依据各自的分类区间及分类均值标准实施特征项间的语义关系分析，完成实例间关系的分析与标注过程。其流程如图1所示。

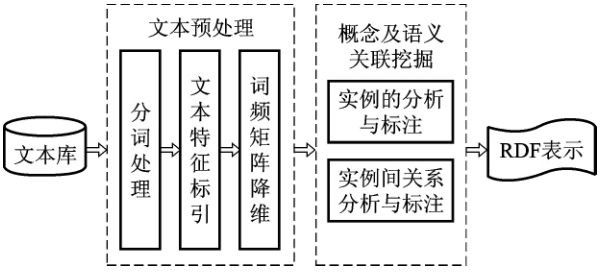

图1 基于数据挖掘的Web文本分析与标注流程

2.2 文本预处理过程

2.2.1 **中文分词处理** 计算机处理中文文本信息时，首先遇到的是词的切分问题。目前，汉语自动分词方法主要可分为三类：基于词典的方法、基于统计的方法和基于人工智能的方法。其中，基于统计的方法是中文分词技术方法的主流。基于词典的分词方法的三要素是分词词典、文本扫描顺序和匹配原则。文本扫描顺序有正向扫描、逆向扫描和双向扫描。匹配原则主要有最大匹配、最小匹配、逐次匹配和最佳匹配。

基于统计的分词方法所应用的主要统计量或统计模型有：词频、互信息、N元文法模型、隐Markov模型和最大熵模型等。这些统计模型主要利用词与词的联合共现概率作为分词信息。

基于人工智能的分词方法的主要思想是在分词的同时进行语法、语义分析，利用语法信息和语义信息来处理歧义现象，主要有专家系统分词法和神经网络分词法。

本文主要运用词频统计方法来进行中文分词处理。

2.2.2 **文本表示** 文本特征标引是用文本的特征信息集合来代表原来的文本。经典的文本表示模型是矢量空间模型（Vector Space Model，VSM）[4]，在该模型中，文档空间被看作由一组正交词条矢量组成的矢量空间，每个文档 d 可以表示为其中的一个范化特征矢量：$V(d) = (W_d(t_1), \cdots, W_d(t_i), \cdots, W_d(t_n))$，其中 t_i 为词条项，$W_d(t_i)$ 为 t_i 在 d 中的权值。$W_d(t_i)$ 一般被定义为 t_i 在 d 中出现频 $f_d(t_i)$ 的函数，即 $W_d(t_i) = \psi(f_d(t_i))$。$\psi$ 的计算通常有布尔函数、平方根函数、对数函数、TFIDF函数。

本文主要采用TFIDF函数[5]，因为它综合考虑了不同的词在所有文本中的出现频率以及这个词对不同文本的分辨能力，其公式如下：

$\psi = f_d(t_i) \times \log(N/n_i)$，其中，$N$ 为所有文档的数目，n_i 为含有词条 t_i 的文档数目。

2.2.3 **词频矩阵降维** 基于VSM向量模型开展数据挖掘分析存在以下缺陷：一是VSM主要依据词频信息来判断文本间的相似度，而忽略了关键词的同义词、多义词及上下文语境的限制，影响了结果处理的准确性和完整性。二是过高的文本向量空间维数而使词频矩阵非常稀疏，增加寻找词间关系的难度。潜在语义标引方法通过奇异值分解技术，将文档从高维向量空间投影到低维潜在语义空间中，有效地缩小了向量空间的规模。另外，LSI方法的出发点就是文本中词与词之间存在潜在的语义结构，与VSM相比，LSI将特征项映射到概念级，消除同义词、多义词的影响，提高文本表示的准确性。

设 A 为 $m \times n$ 特征项—权重矩阵，A 的奇异值分解定义为：$A = UWV^T$，其中 U 和 V 是正交矩阵，W 是奇异值对角矩阵。其中 U 为 $m \times m$ 阶，V 为 $n \times n$ 阶，W 为 $m \times n$ 阶，其对角线上元素由小到大的顺序排列，并保留 L 个最大的特征值，其余置为0，去掉置为0的多有行和列，得到对角阵 W_l，去掉 U 和 V 中相应的列，得到矩阵 U_l、V_l，可得到新的矩阵 $A_l = U_l W_l V_l^T$，此矩阵在最小平方意义下最接近原来的矩阵[6]。

对每个文档 d，设其初始向量的维数为 n，则 d 可映射成维数为 L 向量 d'，$d' = (dV_l W_l^{-1})$，在运用数据挖掘分析时，用该词组向量代替原有的文本特征向量。

2.3 基于数据挖掘的文本语义分析与标注过程

2.3.1 **文本实例内容的分析与标注** 实例集合对应着领域本体中的概念。虽然LSI方法将特征项映射到概念级，消除了同义词、多义词的影响，但它缺乏清晰的表达能力。若基于SVD技术分解的新特征向量基础上，进行概念聚类，就能清晰地表达文本特征项（实例）集合所对应本体中的概念。

将经过SVD技术分解得到新矩阵 A_l 转置后得到“概念—文档”矩阵 N，矩阵的每一行表示一个概念、每一列表示一篇文档，则矩阵中的元素 N_{ij} 表示概念 t_i 在文档 d_j 中的权重。则该矩阵的行向量可表示为 $V(t) = (W_t(d_1), \cdots, W_t(d_i), \cdots, W_t(d_n))$。$d_i$ 表示第 i 个文本，$W_t(d_i)$ 表示第 i 个文本在概念 t 中的权重。本文试用平面划分法将概念集合分割为若干簇[7]，每一簇即为要标注的实例集合，实例聚类步骤如下。

1) 确定要生成的簇的数目 K（$K<n$）。

2) 依据领域本体所对应的概念，在矩阵 N 中找出 K 个上位概念作为聚类中心的种子 $S=\{s_1,\cdots,s_j,\cdots,s_k\}$，其中 $V(s_j)=(Ws_j(d_1),\cdots,Ws_j(d_i),\cdots,Ws_j(d_n))$。

3) 对每个特征项 t_i，依次计算它与各个种子 s_j 的相似度 sim（t_i，s_j）。其中特征项 t_i 和概念 s_j 之间的相似度可以用向量 $V(t_i)$ 和 $V(s_j)$ 的余弦来计算，公式如下:

$$\mathrm{sim}(t_i,s_j)=\frac{\sum_{i}^{n}W_{t_i}(d_i)\times W_{s_j}(d_i)}{\sqrt{\sum_{i=1}^{n}W_{t_i}(d_i)^2}\times\sqrt{\sum_{i=1}^{n}W_{s_j}(d_i)^2}} \quad (1)$$

4) 选取具有最大相似度的种子 $\arg\max \mathrm{sim}(t_i,s_j)$，将 t_i 归入以 s_j 为聚类中心的簇 C_j，从而得到特征项集的一个聚类 $C=\{c_1,\cdots,c_j,\cdots,c_k\}$。

5) 重新确定每个簇的中心点。

6) 重复步骤 2)、3)、4)、5)，直到中心点不再改变，文本中的特征项不再重新被分配为止。

2.3.2 文本实例间关系的分析与标注 本体中概念间的关系多种多样，主要有上下位关系、整体部分关系、属性关系、实例关系、同义关系或反义关系。一个关系通常包含定义域和值域两部分，这两部分限定了关系所适用的范围。在本体中，关系的定义域通常是一个概念，而值域既可以是概念，也可以是具体的取值域（如字符串和整数等），当值域为取值域的时候，关系便退化为属性，所以可以说属性是一种特殊的关系。如果只考虑关系的值域为概念的情况，关系集合 R 中的每个关系 r_i（c_p，c_q）便表示概念 c_p 和 c_q 间的二元关系。但此时这个关系只能表明概念 c_p 和 c_q 所对应的实例中可能存在关系 r_i，而并非任意取自这两个概念的实例都一定具有这样的关系。具体给定一组实例，它们之间是否存在某一种关系，需要借助标注指出。

关联规则分析是数据挖掘的一个重要方法，本文在 SVD 分解产生新矩阵的基础上运用关联规则挖掘方法发现特征项间的同义或反义关系、整体部分关系、上下位关系。将文本集 D 看作事务集，其中每个文本 d 视为一个事务；对于每一个文本 d，用 SVD 方法筛选得到的特征项组成新的向量替换原有的文本特征向量，新向量组成的特征项集视为项集。其应用过程如下:

1) 设置最小支持度阈值 $S_{\min}$ 和最小置信度阈值 $C_{\min}$，运用 Apriori 算法找出文本集 D 中所有的频繁特征项集 $W=\{t_1,\cdots,t_i,\cdots,t_n\}$，并由该频繁特征项集直接产生强关联规则集 $R=\{r_1,r_2,\cdots,r_i,\cdots\}$，$r_i=\{t_i\Rightarrow t_j\}$，其中 $t_i,t_j\in W$，且 $P(t_i\cup t_j)>S_{\min}$，$P(t_j|t_i)>C_{\min}$。

2) 对于 t_i，$t_j\in W$，若同时满足 $t_i\Rightarrow t_j$ 和 $t_j\Rightarrow t_i$，那么特征项 t_i 和 t_j 的关系为等价关系[8]，笔者认为该等价关系包含同义或反义关系。

3) 整体部分关系和上下位关系都为层次结构。对于 t_i，$t_j\in W$，且 $\{t_i\Rightarrow t_j\}\in R$，$P(t_j|t_i)>P(t_i|t_j)$，则 t_j 出现的文档集合是 t_i 出现的文档集合的子集的概率，要大于 t_i 出现的文档集合是 t_j 出现文档集合的子集的概率，那么概念 t_i，t_j 之间存有层次结构，且 t_i 是 t_j 的上位概念。

4) 从 R 中筛选具有层次结构关系的强关联规则集 $R_1=\{r_1,\cdots,r_i,\cdots r_n\}$，其中 $r_i=\{t_i\Rightarrow t_j\}$，且 $P(t_j|t_i)>P(t_i|t_j)$。从 R_1 中人工选出具有继承关系的强关联规则训练集 $R_{\mathrm{is-a}}$，和整体部分关系规则训练集 $R_{\mathrm{part-whole}}$，计算它们各自关于 $P(t_j|t_i)$ 的最大值和最小值区间及平均值，作为具有层次结构概念间关系学习分类的标准。

5) 对任意 $r_i=\{t_i\Rightarrow t_j\}\in R_1$，若 $P(t_j|t_i)\in R_{\mathrm{is-a}}[P_{\min},P_{\max}]$，且 $P(t_j|t_i)\ :\ R_{\mathrm{is-a}}(\overline{P})$。则 r_i 应归属于上下位关系。同理可推断整体部分关系。

3 实验过程及结果

实验文本来源于中国农业网类的 20 篇文献，使用 ICTCLAS 汉语词法分析系统进行 Web 文本的预处理工作，每个文档取出权重最高的 5 个关键词，20 篇文档共取出 100 个关键词，形成 20×100 文档—概念矩阵。其部分截图如图 2 所示。

	桃花	苹果	鲜花	种植业	蔬菜	果品	种子	林业	花卉苗木	肥料	农业用药
	1	0	2	1	0	1	0	1	2	0	0
	0	0	0	2	0	0	0	0	0	1	1
▶	0	1	0	3	0	5	2	1	1	1	1
	1	1	1	2	1	1	1	1	1	1	1
	0	2	0	1	1	3	0	0	0	0	1
	0	0	4	1	0	1	1	4	3	1	2
	3	0	5	1	1	0	1	3	6	1	1
	0	1	1	0	3	0	4	0	0	3	2
	0	6	0	1	1	5	0	0	0	0	1
	1	1	1	2	1	1	1	3	0	1	1

图 2 文档—概念矩阵

对该矩阵进行奇异值分解和转置，得到具有潜在语义关系的概念—文档矩阵，并挑选聚类中心种子概念 $S=\{$种植业，林业，畜牧业，农业农药，农业机械……$\}$，其中如 V（种植业）=（1，2，…，2）。按照平面划分算法进行概念聚类，得到的部分概念集如表 1 所示。

基于上述概念—文档矩阵，再转置为文档—概念矩阵，实施关联规则挖掘。设最小支持度 $S_{\min}=0.5\%$，最小小置信度 $C_{\min}=30\%$，利用 Apriori 方法挖掘得到 123 个两项强关联规则，部分强关联规则集 $R=\{$种植业⇒蔬菜；种植业⇒果品；蔬菜⇒蔬菜种子；果品⇒果汁；畜牧业⇒蜜蜂；鲜花⇒花粉……$\}$，挑选出具有继承关系的训练样本集如｛种植业⇒蔬菜；种植业⇒果品；畜牧业⇒蜜蜂，｝和部分整体关系训练样本集如｛蔬菜⇒蔬菜种子；

表 1　部分结果概念集

种植业，蔬菜，脱水蔬菜，果品，种子，茶叶，棉花，食用菌，辣椒，牛椒，苹果，黑木耳
林业，花卉苗木，鲜花，桃花，树苗，乔木，灌木，竹类植物，球类植物，仙人掌，草坪，盆景
畜牧业，牛，羊，猪，刺猬，天山雪鸡，蜜蜂，大雁，山鸡，鸡苗，野猪，长毛兔，水貂
农业用药，肥料，饲料，农药，兽药，杀虫剂，复合肥，茶叶专用复合肥，保鲜剂，环保肥
农业机械，农田排灌机械，作物收获机械，农产品加工机械，蔬菜大棚，榨汁机，花架

果品⇒果汁；鲜花⇒花粉……}，其中各样本集数量占总样本集的 30%，并计算它们 $P(t_j \mid t_i)$ 的平均最小值和平均最大值，及平均值，按照算法，其中具有继承关系的训练集分类器区间为 [0.21，0.75]，均值为 0.43，在接近均值附近，不是绝对离群点的情况下，该实验将 97 个概念关系划分于继承关系中；同样其中具有部分整体关系的训练集分类器区间为 [0.42，0.68]，均值为 0.53，该实验将 19 个概念对划分于该关系中，其余 7 个概念对属于其他关系，如动宾关系，{蜜蜂⇒花粉……}。

经人工检查，该实验所选领域文本的实例可正确划分到所属领域本体概念簇中，实例间关系分析与标注的正确率可达到 80% 以上，经分析误差来自于领域本体的选取、领域文本数量选择等方面。因此，算法基本能够区分出实例间不同的语义关系，从而证明本文所提出的方法是有效的。

本研究表明，数据挖掘技术是实现文本语义分析与标注的有效方法，明显提高了文本分析与标注的自动化、智能化水平和质量。但也存在一定缺陷，如手工辅助仍有一定的比重，还需要更好的数据挖掘算法的支持以提高文本分析与标注的效率。□

参考文献

[1] HAN Lixin, et al. CMSA: a method for construction and maintenance of semantic annotations [J]. Lecture Notes in Computer Science. 2005 (1): 514-519.

[2] 陆建江，张亚非，等. 语义网原理与技术 [M]. 北京：科学出版社，2008：76-79.

[3] 荆涛，左万利，等. 中文网页语义标注：由句子到 RDF 表示 [J]. 计算机研究与发展，2008，45 (7)：1221-1231.

[4] 陈文伟. 数据仓库与数据挖掘教程 [M]. 北京：清华大学出版社，2006：132.

[5] 郭庆琳等. 基于 VSM 的文本相似度计算的研究 [J]. 计算机应用研究，2008，25 (11)：3256-3258.

[6] 戚涌，等. 基于潜在语义标引的 Web 文档自动分类 [J]. 计算机工程与应用，2004 (22)：28-31.

[7] 唐涛. 基于文本挖掘的领域本体学习研究 [D]. 武汉：武汉大学，2009：63.

[8] 赵心. 一种基于关联规则的中文概念集生成算法 [J]. 计算机科学，2004，31 (7)：175-177.

作者简介：张玉峰，教授，博士生导师。

蔡皎洁，博士生，讲师。

收稿日期：2009 - 09 - 23

(上接第 92 页)

5　结束语

面对网络信息资源激增的情况下，如果更加注重用户与用户之间的联系，对增加网络信息资源的可检索性有极大的帮助，也彰显了 Web2.0 网站最大的特性——社会性。因此，深层次挖掘用户之间的联系，搭建用户之间的社会语义的联系，放大信息用户深层次的社会化体验，是基于 Web2.0 的网络环境下情报学的核心内容之一。

再者，笔者在研究过程中发现，无论通过何种信息组织的方法，均殊途同归地让信息形成信息群，让用户形成用户群，而达到加大网络信息资源可检索性的目的。由此可以看出，解决“世界 2”与“世界 3”互动障碍问题的核心是让信息增加一些个性，让用户增加一些共性。“群”效应使两者之间达到了一定的平衡。□

参考文献

[1] BROOKES B C. 情报学的基础 (一) [J]. 王崇德，等译. 情报科学，1983，4 (4)：84-94.

[2] 刘春茂. 网络信息资源组织与服务的辩证分析 [J]. 情报科学，2003，52 (1)：42-44，52.

[3] 李玉梅. 面向“人—信息交互”的网络信息的社会性分析 [J]. 情报杂志，2008 (4)：83-85，89.

[4] 俞传正，李佳培. 社会导航研究 [J]. 图书与情报，2005 (2)：47-49，59.

[5] 杨达，毕强. 超媒体信息空间中导航用户认知地图的建构 [J]. 图书情报工作，2005，49 (8)：28-32.

[6] MORVILLE P. Ambient findability [M]. CA: O'Reilly Media, Inc., 2005.

[7] 刘春茂，王琳. 网络环境下情报学理论体系的创新 [J]. 图书情报工作，2001 (8)：15-19.

[8] 刘春茂，杨卫. 面向语义的网络信息资源整合的指示数据库案例研究 [J]. 情报学报，2006 (5)：620-628.

作者简介：刘春茂，男，1963 年生，教授。

米国伟，男，1982 年生，助理研究员。

收稿日期：2009 - 11 - 03