茶学本体学习中的概念抽取①

程波波 张友华 李绍稳 辜丽川 朱利君 (安徽农业大学 信息与计算机学院 安徽 合肥 230036)

摘 要： 提出了一种基于茶学词典和统计算法相结合的茶学知识概念抽取方法。该方法以茶学词典为基础，首先对非结构化数据源进行中文分词处理，然后采用两种统计算法对分词结果进行概念抽取。通过使用丰富的茶学词典来降低统计算法时间复杂度，提高了中文分词和概念抽取的精度和效率。实验结果表明，词库的丰富程度决定了概念抽取的效果，可以通过不断丰富词库，进一步提高概念抽取精度。

关键词： 本体学习；概念抽取；茶学词典；统计算法

Concept Extraction in Tea Ontology Learning

CHENG Bo-Bo, ZHANG You-Hua, LI Shao-Wen, GU Li-Chuan, ZHU Li-Jun

(School of Information & Computer, Anhui Agricultural University, Hefei 230036, China)

Abstract: A concept extraction method is presented based on tea dictionary and statistics. The method takes tea dictionary as basis. Firstly, unstructured data source is in Chinese word segment processing, and then, two statistics algorithms are applied to extracted tea concept from Chinese segment results. The approach improves the precision and efficiency of Chinese segment and concept extraction by reducing the time complexity of statistical algorithms with the rich tea dictionary. The experimental results show that the degree of dictionary richness determines the efficiency of tea concept extraction, and can be improved by updating tea dictionary.

Keywords: ontology learning; concept abstract; tea dictionary; statistics

1 引言

目前领域本体的构建大多使用一些本体编辑工具手工构建本体，手工构建本体需要用户逐个的输入和编辑领域知识的概念、关系及属性等，然后才能基于这些知识进行推理或者知识发现，然而这是一项工作量大且容易出错的工作；另一方面，由于知识是一个不断更新和变化的过程，这就要求领域本体能够根据知识的变化实现动态更新，以保证领域本体的准确性和实效性。所以，手工构建本体具有工作效率低，且构建的本体动态更新、追踪新知识的能力弱等缺点。因此,如何利用知识获取技术来降低本体构建的开销是一个很有意义的研究方向.目前,国外在该方向的研究很活跃,把相关的技术称为本体学习(ontology learning)技术[1]。本体学习的目标是从数据源中自动或者半自动抽取本体的基本元素：概念、关系(分类关系和非分类关系)、公理来构建新本体或者扩充已有本体。概念是本体的最基本组成元素，是概念间关系的基础，所以概念获取的准确度和完备度是本体学习中一个重要参数指标。目前，对领域本体概念的自动抽取研究大多基于语言学和基于统计的方法，或者是两者相结合的方法。例如采用 WordNe[2]作为通用语言学本体，然后通过计算词频(Frequency)、术语和逆文档频率 TFIDF(Term Frequency Inverted Document Frequency)、互信息 MI(Mutual Information)等因子，来抽取领域概念。本文采用领域知识词典和统计相结合的方法实现概念抽取，并对实验结果进行了比较分析，得到了较好的效果。

① 基金项目:国家高技术研究发展计划(863)(2006AA10Z249);国家自然科学基金(30800663;30971691)
收稿时间: 2009-10-28;收到修改稿时间:2009-11-24

2　茶学知识概念提取框架

茶学知识的概念提取研究以茶学知识为背景语料集，数据源采用茶学茶虫知识非结构化数据文档集和已构建的茶学知识关系数据库茶学知识概念提取框架(Concept Extraction of Tea Knowledge Extra-ction of Tea Knowledge Framework-CETWF)如图 1 所示。整个流程从获取内容上可划分为三部分：

(1) 预处理部分

预处理部分主要是对非结构化知识源的中分词、词性标注、停用词的删除、标点及特殊符号的清除，为概念抽取提供格式化(归一化)的知识源。

(2) 概念的提取

概念的抽取是在数据源预处理的基础上，对格式化的知识源进行概念抽取。概念抽取采取两种方法：基于茶学词典的概念搜索算法和基于改进 TFIDF 算法的概念抽取。

(3) 结果评判

使用准确率和召回率作为评价概念抽取结果。

A:准确率＝抽取出正确的概念数目/所有的概念数目

B:召回率＝抽取出正确的概念数目/抽取出所有的概念数目(包含正确与不正确的)

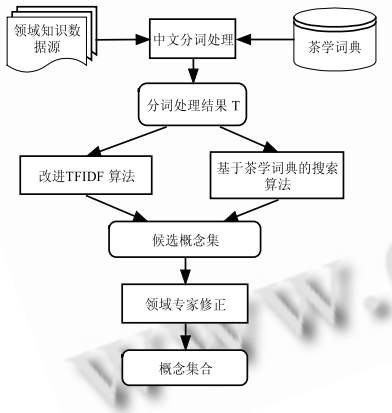

图 1　领域知识概念提取框架（CETWF）

3　茶学知识概念提取框架

3.1 分词及词性标注

由对于非结构化数据而言，词、短语的切分、词性标注是一个基础问题，因为概念是以词或短语形式出现。分词、词性标注采用开源 ICTCLAS [3]代码，并辅以茶学词典库进行中文分词，提高分词及词性标注效果[4]，经过调整后的分词系统，其正确率目前可以达到 90%以上，可以达到试验要求。

Example：

分词前：茶尺蠖是我国主要茶树害虫之一。分布遍及各主产茶区，主要在长江中下游，尤以苏、浙、皖接壤地区危害严重。国外分布于日本。幼虫食叶常致茶丛光秃。还危害大豆、豇豆、芝麻、向日葵、辣蓼等。

分词后：茶尺蠖/n 是/v 我国/n 主要/b 茶树/n 害虫/n 之一/r 。/w 分布/v 遍及/v 各/r 主/a 产/v 茶/n 区/n ，/w 主要/d 在/p 长江/n 中下游/n ，/w 尤/n 以苏/n 、/w 浙/j 、/w 皖/j 接壤/v 地区/n 危害/v 严重/a 。/w 国外/s 分布/v 于/p 日本/n 。/w 幼虫/n 食/v 叶/n 常/d 致/v 茶/n 丛/n 光/d 秃/a 。/w 还/d 危害/v 大豆/n 、/w 豇豆/n 、/w 芝麻/n 、/w 向日葵/n 、/w 辣蓼/n 等/u 。/w

其中汉语词性标记符号如表 1 所示。经过分词处理之后，像茶尺蠖、大豆、豇豆等名词被标注出来，并注有词性标注，方便术语的提取。

表 1　部分汉语词性标记符号

a	b	d	j	n
形容词	区别词	副词	地名	名词
p	r	s	v	w
介词	代词	处所词	动词	标点符号

3.2 概念抽取方法

在完成分词之后，采用两种方法实现概念的抽取，方法一：采用基于茶学字典的遍历查询，直接获取概念集；方法二：采用改进的 TF-IDF[5,6]方法抽取概念集。

Method 1：以先期构建的八大类、包含一万余条词和短语的电子词典为基础，采用正向最大[7]匹配算法直接从茶学知识非结构化文档中提取概念。这种方法实际上是把词典中的词或者短语作为概念候选项，通过和知识文档中字符的比较，若知识文档中存在词典中的词或者短语，则直接将其提取出来。例如对字符串 S = $C_1C_2C_3C_4$…进行匹配查询，其算法描述如下：

Step1.取一字 C_1 ,在字典中查找 C_1 并保存是否成词标记;

Step2.再取一字 C_2 ,判断字典中是否有以 C_1C_2 为前缀的词;

Step3.若 C_1C_2 不存在,则 C_1 为单字,一次查询结束;

Step4.若 C_1C_2 存在,判断 C_1C_2 是否为词,并取以 C_1C_2 为首的多字词的个数 n;

Step5.若 n 为 0,一次查询结束;

Step6.若 n 不为 0，则再取一字 C_i ,判断词表中是否有以 $C_1C_2\cdots C_i$ 为前缀的词;

Step7.若 $C_1C_2\cdots C_i$ 不存在,则返回最近一次能够成词的 $C_1C_2\cdots C_{i-1}$;

Step8.否则转(6);

Step9.从字 C_i 开始下一次查询。

这种方法较好的利用了茶学字典功能,通过把分词结果的特定短语和词典中词或短语进比较，直接的进行概念抽取，避免了大规模的文本计算。不过此方法也存在缺点,即直接把字典中的词和短语当作概念,这也是有待改进的地方。

Method 2： 对于某一领域的知识文集而言，文档中总会有一些词或者短语可以反映文章的主题。就像茶虫知识而言，生活习性、病虫防治、生态特征等词汇出现频率较高,因为每一种茶虫都具有这些特征。基于这一点，本文采用改进的特征值提取算法—TFIDF 方法，其数学表达式如式(1)所示：

$$\mathrm{TFIDF}=\sum_{j=1}^{n} tf_{ij} * \log\left(\frac{(df_i)_j}{(df_i)_j+(df_i)_k}\right) * n \qquad (1)$$

该算法基本思想是：假设目标文集定义为 $D=\{d_1,d_2,\cdots,d_j,\cdots,d_n\}$,其中 d_j 为 D 的一个文档，每个的文档 d_j 中被计算的字符串称为 $term_i$，通过利用改进的TFIDF方法计算每个文档中 $term_i$ 的权重抽取术语。公式中 tf_{ij} 表示 $term_i$ 在文档 d_j 中的出现次数,次数越高就意味着 $term_i$ 对于文档 d_j 就越重要，也就表明 $term_i$ 分辨领域主题的作用越大；$(df_i)_j$ 表示在当前类中含有 $term_i$ 的文档数量，$(df_i)_k$ 表示在其他类中含有 $term_i$ 的文档数。$(df_i)_j$ 越大就意味着 $term_i$ 为术语的可能性方面越大，$(df_i)_k$ 越大就意味着 $term_i$ 为术语的可能性越小。其算法描述如下：

定义 1.设文档集 $D=\{d_1,d_2,\cdots,d_j,\cdots,d_n\}$，$d_j$ 表示每一个独立的数据源文档；

输入：预处理文档集合 $D=\{d_1,d_2,\cdots,d_j,\cdots,d_n\}$；

定义 2.每个文档的分词处理结果为 $W=\{w_1,w_2,\cdots,w_j,\cdots,w_n\}$,W 中包含 n 个词序列，每个词序列定义为 wordi；

定义 3.设术语抽取结果为 $T=\{term_1,term_2,term_3,\cdots,term_i,\cdots,term_n\}$,T 为一个文档，文档中包含 n 个术语 term；

输出：$T=\{term_1,term_2,term_3,\cdots,term_i,\cdots,term_n\}$。

处理步骤：

Step1.输入预处理文档集合 $D=\{d_1,d_2,\cdots,d_j,\cdots,d_n\}$,进行中文分词处理得到结果 $W_j=\{w_1,w_2,\cdots,w_j,\cdots,w_n\}$;

Step2.计算 $word_i$ 在当前文档中的词频 $w\text{-}f_i$，计算 $word_i$ 在其他文档中的逆文档词频 $w\text{-}idf_i$；

Step3.计算 $tf\text{-}idf = (w\text{-}f_i) * (w\text{-}idf_i)$;

Step4.对于 tf-idf 进行阀值判定，当 tf-idf>w 时，输出 $word_i$ 到 T 中，形成 $term_i$，返回 Step2。

对于文集中的词或者短语，在得到其 TFIDF 值之后，可以设定一个阀值 w，低于此阀值的将被过滤，不作为术语。阀值的获取要通过训练文本得到一系列的 TFIDF,由人工确定 W 值的大小，目的是为了获取术语的准确程度；另外对于一些经常出现的常用词，像“我们”、“的”、“不”、“他”等常用词语，虽然 TFIDF 较高，但并不能反映文章主题，所以要设计停用词表，自动屏蔽此类词和短语。

4 实验与结果分析

数据来源主要有两个方面，一是利用中国知网期刊数据库检索茶学茶虫文献，共 107 篇，包括生活习性、防治方法等方面；二是采用书籍《中国茶树害虫及其无公害治理》(张汉鹊 谭济才主编，安徽科学技术出版社出版社)中第 155 页到第 325 页的数据，首先对书本数据源进行电子化，然后将文本中的图、表等非字符数据删除，只保留字符数据，调整后的数据源文档共 171 页,其中每个茶虫都有独自的生活习性、不同生命阶段的形态特征、茶园危害、防治方法等方面，不同方面文档的数量如表 2 所示。实验内容主要包括两部分：一是对概念的抽取；二是利用准确率及召回率对实验结果进行计算，并和本体学习工具 Text2Onto 作比较分析。

表 2　文档数据分布

	生活习性(篇)	形态特征(篇)	茶园危害(篇)	防治方法(篇)
网页	18	16	31	42
书籍	125	108	136	167

图 2 为表 3 实验结果的曲线图，图 3 为与本体学习工具 Text2onto[8,9]在概念抽取准确率方面的比较。

表 3　准确率与召回率

	生活习性	形态特征	茶园危害	防治方法
准确率	62.30%	64.40%	65.20%	66.70%
召回率	58.60%	60.24%	60.80%	60.20%

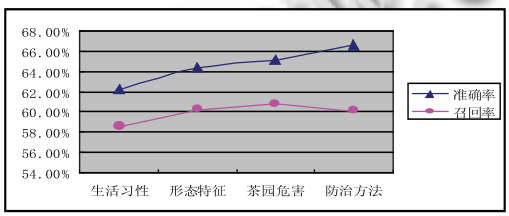

图 2　实验结果曲线图

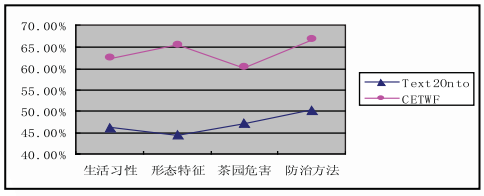

图 3　准确率比较

从实验结果来看，该方法在概念提取精度上有了一定的提高，这主要归于茶学字典的应用，因为大多统计方法的应用都是以一定的语料库为基础，语料库越健全，获取的结果越好；另外中文分词及词性标注的结果对概念的抽取也有很大影响，例如，分词前，一个句子的长度为 n 个字符，对其处理之后变成 m 个 n/m 词或短语，利用 Method1 进行搜索查询时其时间复杂度就提高了 O(n)/O(n/m)，从中可以看出 m 越大效率越高。

5　结语

当前本体学习中概念的提取，大多基于统计学的方法。如果单纯依靠统计学的方法对非结构化数据源进行概念抽取，计算量会随着数据源的增加呈现指数级的增长，而且中文语句句法复杂，这就导致了概念抽取效率不高、抽取精度很低等问题。本文采用统计和茶学词典相结合的方法对概念进行抽取，并在预处理部分对非结构化文档进行分词、及词性标注，将复杂的中文语句切换成简单的词或短语，节省了计算量并提高了效率。通过分析算法可以看出，实验结果的好坏，与所构建的领域知识词典具有很大关系，词典的丰富程度直接决定了分词的效果，而分词效果又直接影响两种方法的效率。所以，为了提高概念的抽取精度，可以通过丰富领域知识词典的容量来避免复杂的统计算法，达到提高概念抽取效率的目的。

参考文献

1. Gomez PA, MACHO MD. An over view of methods and tools for ontology learning from texts. The Knowledge Engineering Review, New York: Cam bridge University Press, 2004:187－212.
2. Zhang J, Li CP. WordNet-based concept vector space model for text classification. Computer Engineering and Applications, 2006,4:174－178.
3. Zhang HP, Yu HK, Xiong DY. HHMM-based Chinese Lexical Analyzer ICTCLAS. 2nd SIGHAN Workshop Affiliated with 41th ACL, Sapporo Japan, 2003:184－187.
4. 张友华,熊范纶.基于句子相关度的文本自动分类.中国科学技术大学学报, 2006,36(5):540－545.
5. 张玉芳,彭时名,吕佳.基于文本分类 TFIDF 方法的改进与应用. 计算机工程, 2006,32(19):76－78.
6. 徐文海,温有奎.一种基于 TFIDF 方法的中文关键词抽取算法.信息系统, 2008,31(2):298－302.
7. 吴栋,滕育平.中文信息检索引擎中的分词与检索技术.计算机应用, 2004,24 (7):128-131.
8. Volker J, Langa SF, Sure Y. Supporting the cons- truction of Spanish legal ontologies with Text2Onto. Computable Models of the Law, Heidelberg: Springer Press, 2008.105－112.
9. Philipp C, Johanna V. Natural language processing and information systems. Heidelberg: Springer Press, 2005. 227－238.