基于聚类分析的事件语义模式获取*

季陶美,刘茂福,张 璐,杨 晓
(武汉科技大学 计算机科学与技术学院,湖北 武汉 430065)

摘 要:为非结构化的 Web 页面标注事件语义信息,可以丰富 Web 页面结构化信息,加深对 Web 页面内容的理解。选取新闻类型的 Web 页面,遵照事件语义标注规范对选取的未标注 Web 页面进行事件语义标注。对标注了事件语义的语料实例进行抽象得到事件语义结构模式;利用层次聚类算法,将所得的事件语义结构模式进行聚类分析,得到不同类别的事件语义模式。实验结果表明,在已标注事件语义的语料实例的基础上,利用聚类算法进行分析,获取各种类别的事件语义模式,对 Web 页面内容分析与理解是非常必要的。

关键词:事件语义角色;事件语义结构模式;聚类分析
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1674-7720(2013)02-0063-04

Event semantics scheme acquisitions based on clustering analysis

Ji Taomei, Liu Maofu, Zhang Lu, Yang Xiao
(College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract: In order to enrich the structured information and further understand Web page contents, it is very important to firstly label the event semantic roles for the unstructured Web pages. After labeling Web pages using the event semantic rules, the event semantic scheme can be obtained by generalizing the event semantic corpus instances, and then the different clusters of the event semantic scheme can be gained based on hierarchical clustering algorithm. The experiment results show that the clustering analysis on the corpus instances labeled event semantic roles is necessary to understand the web page contents.

Key words: event semantic roles; event semantic; scheme clustering

随着互联网的发展,网络文本数据迅猛增加。要实现人机间相互理解,就意味着首先要让计算机理解自然语言语义。而自然语言语义一般又是由底层的事件语义组成的,因而基于已标注事件语义的语言语料,进行事件语义结构模式获取是非常必要的。近年来,事件研究在自然语言处理领域成为了热点,事件在很多语义计算理论和自动文摘、问答系统等应用领域中都很重要,因此,使用聚类分析获取事件语义结构模式是值得探索的。

语料实例指为语言研究收集的、用电子形式保存的语言材料,由自然出现的书面语或口语的样本汇集而成,用来代表特定的语言或语言变体。经过科学选材和标注,具有适当规模的语料库能够反映和记录语言的实际使用情况。语料实例已经成为语言学理论研究、应用研究和语言工程不可缺少的基础资源。

事件语义结构是语法和语义界面的结合。它充分考虑了事件的时间结构特性和内部构成关系对谓词句法表现的影响,有效地克服了以动词为核心的投射在句法解释方面的理论缺陷。

聚类分析是数据挖掘的核心部分。所谓聚类,就是将物理或抽象对象的集合组成由类似的对象组成的多个类或簇的过程。聚类生成的簇是一组数据对象的集合,同一簇中的对象应尽可能相似,而不同簇中的对象尽可能相异。聚类是在预先不知道目标数据到底有多少类的情况下,希望将所有的记录组成不同的类或者说“聚类”。

目前国内外对这方面的研究还在不断深入。JAMES 提出了事件结构的配价理论,并从词汇语义学的角度分析了事件结构中的语义角色[1];CHANG 基于事件谓词对事件结构内部的论元连接原则进行了讨论[2];JOOST 在通过情景语义分析事件路径的基础上,提出了事件轮廓与

*基金项目:国家自然科学基金资助项目(61100133);武汉科技大学大学生创新基金资助项目(11ZRA104)

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

轨迹的概念[3]；ELENA 从事件分类、语义角色、事体以及因果角度对事件结构进行了分析[4-5]。这些研究工作都是以事件谓语为中心，采用句法分析方法得到的。袁毓林等从认知角度研究了汉语的论元结构和描述框架，并进行了真实文本语义标注的实践[6-7]；吴平对特殊句式的事件语义结构进行了分析与研究[8-10]；李世奇等提出了一种基于特征组合和支持向量机的中文语义角色标注方法[11]；郝秀兰等提出了事件类定义角色语义表方法，将 HowNet 的事件类与语义解释联接起来[12]。

本文基于事件语义标注规范，使用事件语义标注工具，对 Web 上收集的未标注文本语料，进行尝试性标注和聚类分析，进而得到更加抽象的事件语义结构模式。

1 系统流程

文本选取新闻类型的 Web 页面，遵照事件语义标注规范对选取的未标注 Web 页面进行事件语义标注。对标注了事件语义的语料实例进行抽象得到事件语义结构模式；利用层次聚类算法，将所得到的事件语义结构模式进行聚类分析，得到不同类别的事件语义模式。整个系统的处理流程[9-10]如图 1 所示。

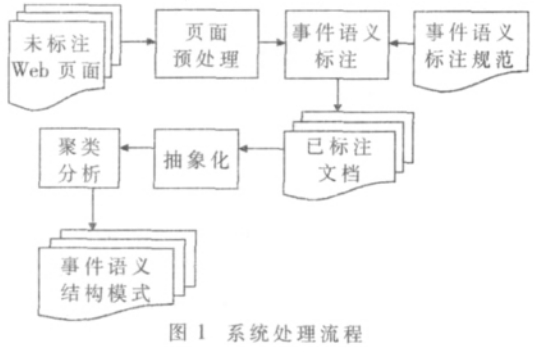

图 1 系统处理流程

其中，对于未处理的 Web 页面，页面预处理的主要功能是将未标注的 Web 页面中涉及到的事件进行拆分，如例 1 所示。

例 1 原句：2010 年朴文垚在日本首夺世界冠军，荣升中国第 30 位九段围棋手。

拆分后事件 E1：2010 年朴文垚在日本首夺世界冠军

拆分后事件 E2：荣升中国第 30 位九段围棋手

对选取的 Web 页面进行处理将获得事件集合，遵照事件语义标注规范对预处理后的 Web 页面进行事件语义标注。标注结果的语料实例如例 2 所示。

例 2 标注后事件 E1：<EVENT id="E1">[2010 年 T][朴文垚 A]在[日本 L][首 Ra][夺 EP][世界冠军 P] </EVENT>

标注后事件 E2：<EVENT id="E2">[荣升 EP][中国 Ra][第 30 位 Ra][九段围棋手 Re] </EVENT>

对此标注了事件语义的语料实例进行抽象得到事件语义结构模式，如事件 E1 抽象后的结果为“T，A，L，Ra，EP，P”。

其中 A 表示施事，P 表示受事，T 表示时间，EP 表示谓词，L 表示地点等。通过分析，对抽取的某个事件进行人工的事件语义标注，得到该事件的语义结构模式。最后，将大量的事件语义结构模式进行聚类即可得到不同类别的事件语义结构模式集合。

2 聚类算法

聚类[4]是将数据分类到不同的类或者簇的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。聚类是搜索簇的无监督学习过程。与分类具有类别标记不同，无监督学习不依赖预先定义的类或带类标记的训练实例，需要由聚类学习算法自动确定标记。聚类能够作为一个独立的工具获得数据的分布状况，观察每一簇数据的特征，集中对特定的聚簇集合作进一步的分析。

层次聚类方法通过将数据组织为若干组并形成树形结构来进行聚类，可以分为自上而下和自下而上两种。自上而下策略是将所有对象置于一个类中，然后渐渐分为越来越小的类，直到每个对象自成一类，或者达到了某个终结条件；自下而上策略是最初将每个对象（自身）作为一个基本类，然后将这些基本类进行聚合以构造越来越大的类，直到所有对象均聚合为一个类，或满足一定终止条件为止。自上而下和自下而上基本思想如图 2 所示。

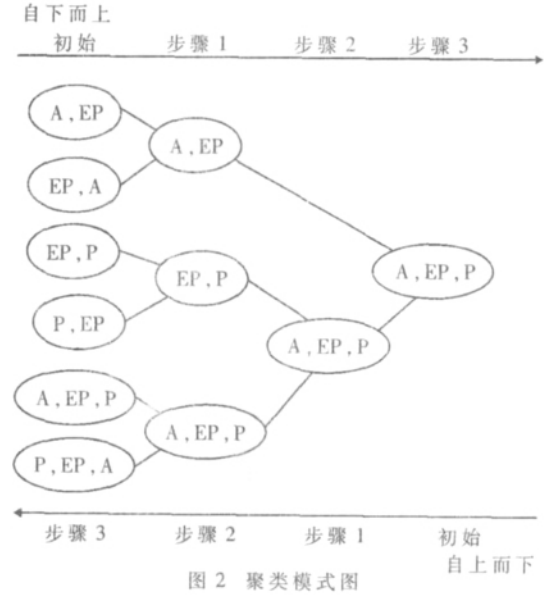

图 2 聚类模式图

本文采用自下而上的层次聚类算法对得到的事件模式集合进行处理。层次聚类算法的基本思想是：初始状态下属于数据集的每个数据对象自成一类，它们的合并代价初始值为 0；然后，假定任意两个簇合并，利用离差平方和的增量来度量两个簇合并后所需要付出的代价，在计算完所有的两个簇合并的代价后，选择合并代价最小的两个簇进行合并；算法反复迭代，直到所有的簇合并成一个簇或者达到预先设定的簇的数目 k 为止。Ward 层次聚类算法通常采用离差平方和函数做为目标

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

函数，如式(1)和式(2)所示。

$$S_t=\sum_{i=1}^{N_t}||x_{it}-x_t||^2 \quad (1)$$

$$S=\sum_{t=1}^{k}\sum_{i=1}^{N_t}||x_{it}-x_t||^2 \quad (2)$$

其中，S_t 为合并的两个事件语义结构模式中所有语义角色成分的离差平方和，S 为各个事件语义结构模式中所有语义角色成分的离差平方和的总和，k 为预先设定的需要最终凝聚成的事件语义结构模式的数目。假设两个事件语义结构模式要合并成一个事件语义结构模式 At，N_t 为合并后的事件语义结构模式的语义角色成分的个数，x_{it} 为 At 中的第 i 个语义角色成分，x_t 为 At 中所有语义角色成分的平均值。

算法描述如下：

(1)设定最终要凝聚的事件语义结构模式的数目 k；

(2)根据式(1)计算两个事件语义结构模式之间的距离，建立邻近度矩阵；

(3)根据之前的计算结果，合并两个距离最近的事件语义结构模式，生成新的事件语义结构模式 At；

(4)更新邻近度矩阵，反映出新的事件语义结构模式 At 与原来的事件语义结构模式之间的邻近性；

(5)直到事件语义结构模式的数目等于或者小于预先设定的数目 k 为止，否则转向步骤(2)。

3 实验结果与分析

在网络上选取新闻类型的 Web 页面，通过对 30 多篇 Web 页面语料的标注和分析，得出 5 000 个事件语义结构基本模式。将不同的基本事件模式进行初步整理之后，得出如图 3 所示的基本事件模式分布柱状图。

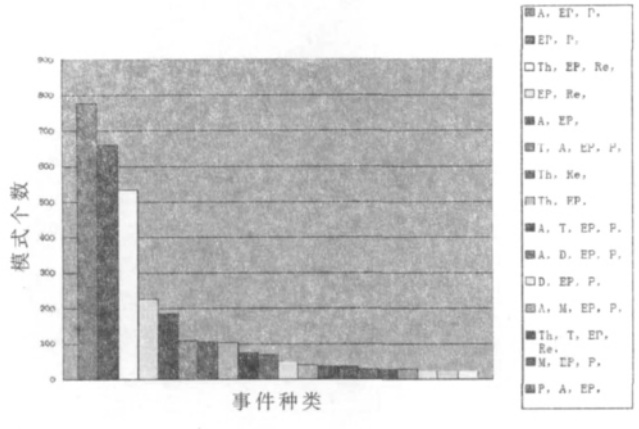

图 3 初始事件模式的分布直方图

从图 3 中得出：最多的两个事件语义结构模式是“A，EP，P”和“EP，P”，即“施事，谓词，受事”和“谓词，受事”，这主要是因为在现实生活中描写主体成分动作的情况非常普遍。而这两个事件语义结构模式的差别就在于后者缺少施事，也就是通常所谓的主体语义角色成分。在交流双方都明确知道的前提下，通常会省略掉“施事”。因此，缺少施事这一语义角色成分和补全这个语义角色成分的区别不大。

当然存在一些事件语义结构模式出现的频率很低，如“A，T，Rn，EP，P”。这一类的事件语义结构模式，即“施事，时间，原因，谓词，受事”，除了包括事件语义结构模式中最重要的主体、谓词、客体成分，还涵盖了凭借成分、环境成分这些附加的事件成分，使得这一类的事件语义结构模式的语义角色成分比较多。事件语义角色成分越多，事件语义结构模式的限定也就越多，所表达的含义就越明确，而通常在使用时会省去时间语义角色成分，所以这一类的事件语义结构模式就很少见了。

对一些看似是两个不同的事件语义结构模式，而实际上表达了相同含义，模式相似度达到 50%的两个事件语义结构模式进行合并，合并之后事件模式的分布直方图如图 4 所示。

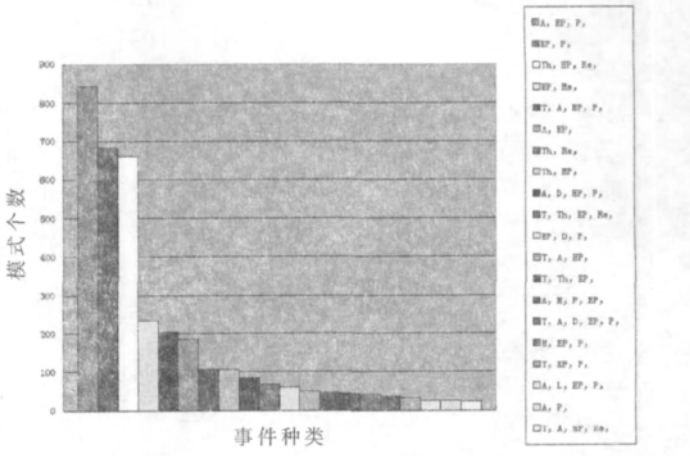

图 4 合并之后事件模式分布直方图

例 3 事件语义结构模式 M：“A，EP，P”。事件语义结构模式 N：“P，EP，A”。

例 3 中 M 包含的 3 个语义角色成分与 N 中包含的语义角色成分是完全相同的，唯一不同点在于语义角色的排列顺序。在汉语中，由于对句子进行了倒装处理或者是将某些语义角色成分前置改变事件语义角色成分的顺序，但是这种情况并没有增加或减少事件语义结构模式中语义角色成分的数目，更没有改变原有事件的含义。如例 4 所示。

例 4 (1)我被老师夸奖了。(2)老师夸奖了我。

在例 4 中，句(1)得到的事件语义结构模式是“P，A，EP”，而句(2)得到的事件语义结构是“A，EP，P”，但句(1)和句(2)的句子成分和句子所表达的客观含义是一致的，因此可以认为这两个句子是相同的。类似的情况还有很多，如“EP，P”与“P，EP”、“Th，EP，P”与“P，EP，Th”等。因此，这样的两个事件语义结构模式是可以合并的，也就是说，这样的两个事件语义结构模式可以视为同一个事件语义结构模式。

图 5 是对不同的事件语义结构模式进行聚类分析之后得到的分析柱状图。根据某个语义角色成分在规定的语料范围内出现的频率决定其加权值。利用聚类算法

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

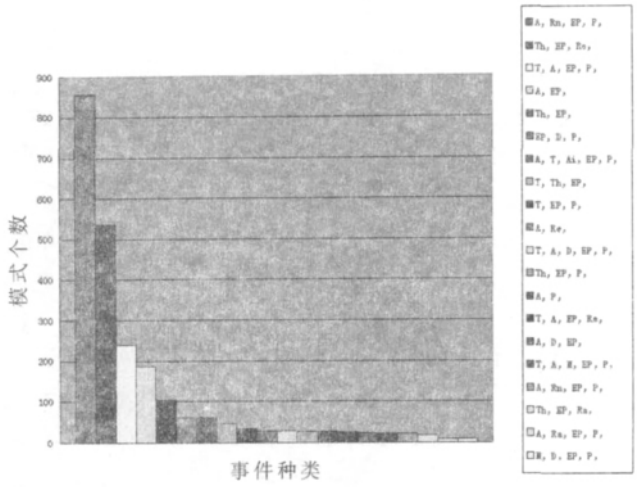

图 5 聚类分析之后的事件模式的分布直方图

对事件语义结构模式相似度高的两个事件语义结构模式进行合并，得到一个事件语义结构模式，经过多次聚类将得出事件中最普遍的事件语义结构模式集合。

例如，事件语义结构模式“A，EP，P”和事件语义结构模式“A，Rn，EP，P”，其中“A，Rn，EP，P”中事件语义角色成分“原因(Rn)”相对于事件语义结构模式“A，EP，P”这个整体所造成的影响是可忽略的。因此这两个事件语义结构模式在某种程度上达到了一致。事件语义结构模式中往往还含有一些对整体模式的影响可以被忽略的语义角色成分，如“使用工具”、“环境成分”等。这些语义角色成分对事件语义结构模式中那些主要的成分进行修饰或者补充说明。例如事件语义结构模式“A，Rn”中的“(原因)Rn”语义角色成分，可以适当忽略该成分对整体事件语义结构模式的影响，将其与事件语义结构模式“A，EP，P”进行合并操作。

本文基于事件语义标注规范，使用事件语义标注工具，对从 Web 上收集的未标注文本语料，进行尝试性标注和聚类分析，进而得到更抽象的事件语义结构模式。实验结果表明，在已标注事件语义的语料实例基础上，利用聚类算法进行分析，获取各种类别的事件语义模式，对 Web 页面内容分析与理解是非常必要的。本文利用上述的聚类算法，对获得的事件语义结构模式进行分析，虽然实验结果还存在一定的问题，如聚类算法不够完善等，但是实验结果说明对事件语义结构模式进行研究还是很有意义的。

参考文献

[1] JAMES P.The syntax of event structure[J].Journal of Cognition，1991，41：47-81.

[2] CHANG Jung-hsing.Event structure and argument linking in Chinese[J].Language And Linguistics，2003，4(2)：317-351.

[3] JOOST Z.Event shape：paths in the semantics of verbs[EB/OL].Ms.Radboud University Nijmegen & Utrecht University. http://www.let.uu.nl/users/Joost.Zwarts/personal/EventShape.pdf，2006.

[4] ELENA P.Event structure in russian：semantic roles，aspect，causation[J].Journal of The Prague Bulletin of Mathematical Linguistics，2009(92)：5-20.

[5] ELENA P.Event structure：taxonomy，semantic roles，aspect，causation[J].Journal of Automatic Documentation and Mathematical Linguistics，2009，43(3)：196-202.

[6] 袁毓林.基于认知的汉语计算语言学研究[M].北京：北京大学出版社，2008.

[7] 袁毓林.用动词的论元结构跟事件模板相匹配——一种由动词驱动的信息抽取方法[J].中文信息学报，2005，19(5)：37-43.

[8] 吴平.汉语特殊句式的事件语义分析与计算(第 1 版)[M].北京：中国社会科学出版社，2009：67-85.

[9] 吴平.论元控制谓词与非论元控制谓词的逻辑语义分析与计算[J].外语与外语教学，2006(3)：5-10.

[10] 吴平.“使”字句事件结构的语义分析[J].浙江大学学报(人文社会科学版)，2009，39(3)：157-164.

[11] 李世奇，赵铁军，李晗静，等.基于特征组合的中文语义角色标注[J].软件学报，2011，22(2)：222-232.

[12] 郝秀兰，杨尔弘，舒鑫柱.基于 HowNet 的事件角色语义特征提取[J].中文信息学报，2001，15(5)：26-32.

(收稿日期：2012-09-25)

作者简介：

季陶美，女，1991 年生，学士，主要研究方向：自然语言处理。

刘茂福，男，1977 年生，副教授，博士，主要研究方向：自然语言处理。

(上接第 62 页)

科技资讯，2009(31)：98.

[8] 倪建立.电力地理信息系统及其发展[J].电力设备，2004，5(7)：85-88.

[9] 黄志龙，邱家驹.配网 SCADA 和 GIS 功能的集成[J].电力系统及其自动化学报，2000，12(4)：36-41.

(收稿日期：2012-09-26)

作者简介：

苑振宇，男，1987 年生，硕士研究生，主要研究方向：GIS 研究、设计、开发与应用。

张明明，女，1981 年生，博士研究生，讲师，主要研究方向：GIS 应用及开发，三维 GIS，三维地质体建模及应用。

王结臣，男，1973 年生，博士，教授，主要研究方向：GIS 理论与应用。

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net