北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

属性和属性值组合的概念模板

程显毅[1,†] 沈学华[1] 施佺[1] 田宇贺[2]

1. 南通大学计算机科学与技术学院, 南通 226019; 2. 南通大学文学学院, 南通 226019; †E-mail: xycheng@ntu.edu.cn

摘要 基于本体抽取三元组〈概念, 属性, 属性值〉, 以词汇聚类为基础, 将概念表示为属性和属性值的组合向量, 对基于属性的概念模板和基于属性值的概念模板进行对比。研究发现, 基于属性和属性值组合的概念模板优于任何一个单独的模板。

关键词 概念; 属性; 属性值; 信息抽取

中图分类号 TP391

Concept Template of Combining Attributes with Attributive Values

CHENG Xianyi[1,†], SHEN Xuehua[1], SHI Quan[1], TIAN Yuhe[2]

1. School of Computer Science and technology, Nantong University, Nantong, 226019; 2. School of Literature, Nantong University, Nantong, 226019; † E-mail: xycheng@ntu.edu.cn

Abstract The authors extract triad 〈concept, attribute, property value〉 based on ontology. Concept is represented as a vector which is combined with attributes and attributive values based on vocabulary clustering. Comparasion between conceptual template based on attributes and based on attributive values is performed. The study shows that the concept template of combination attributes with attributive values is superior to any single template.

Key words concept; attribute; attributive values; information extraction

在信息时代存储大量数据是容易的, 但可用的信息在减少, 搜索引擎只能恶化“通过很少的关键词得到大量文本”这一问题。在这种背景下, 信息抽取(information extraction, IE)成为了研究热点, 其主要目的是将无结构的文本转化为结构化或半结构化的形式存储, 供用户查询以及进一步利用。为了抽取指定的信息, IE 系统需要完成以下具体任务。

1) 准确识别文本中各种命名实体, 如人名、地名、机构名、时间、货币以及各种数字等等。

2) 准确识别并标注指称相同的不同语言元素(共指消解)。

3) 利用领域知识进行推理, 在实体–实体之间及实体–事件之间建立关系。

对任务 1 和 2 的研究已经取得了一些实用成果[1], 而任务 3 的研究进展缓慢, 但却是文本结构化的基础和关键。

不同的研究者对关系抽取任务的表述不尽相同。Schutz 等[2]认为关系抽取是自动识别由一对概念和联系这对概念的关系构成的相关三元组。Katrenko 等[3]则从关系抽取的基本过程角度对关系抽取进行了界定。他们认为, 关系抽取可以看做是具有两个步骤的过程, 即: 识别存在关系的证据和检查是否存在关系。维基百科对关系抽取的解释是, 关系抽取是在自然语言处理过程中抽取文本中实体间命名关系的任务(http://en.wikipedia.org/wiki/Relationship_extraction,_February_8,2011)。抽取的实体间关系能够通过各种形式/语言来表达。而作为关系抽取权威评测会议的 ACE (automatic content extraction)将关系抽取任务表述为: 探测和识别文档中特定类型的关系, 并对这些抽取出的关系进行

国家重点基础研究发展计划(2012CB724108)资助
收稿日期: 2012–06–02; 修回日期: 2012–08–15; 网络出版时间: 2012-10-26 17:04
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1704.009.html

规范化表示[4]。这些关系中，有些对实体出现顺序敏感，有些对实体出现顺序不敏感。

关系抽取技术在很多领域具有应用价值。如在自动问答系统中，关系抽取自动关联相关问题和答案；在检索系统中，关系抽取使类似于“北京有哪些公司?”这样的语义检索功能的实现成为可能；在本体学习过程中，关系抽取能够发现新的实体间关系来丰富本体结构；在语义网标注任务中，关系抽取能够自动关联语义网知识单元[5]。

在从语料库进行概念识别的相关研究中，概念被认为是从语法结构中抽取出来的属性(值)向量[6]，或者看做属性(值)之间多维关系，例如：狗(白色，体积小，四条腿，咬人，嗅觉灵敏)。

在识别概念属性时，有两个问题需要解决。第一个问题是，属性值的抽取。我们发现，模式“A 有 B”可以用来发现“部分–整体”关系，这在发现属性值时特别有效；模式“B 是一种 A”可以用来发现“上下位”关系，这在发现属性时特别有效。第二个问题是，大型语料库中属性模式的实例不如属性值模式的实例多。

本文进行了两个实验，目标是测试基于属性值和属性组合的模板是否对概念产生一个更好的描述。使用属性值或属性作为概念向量的元素，通过聚类实现概念识别。在第 2 节讨论如何用 Web 数据来建立基于属性和属性值的概念向量。第 3 节讨论了两个实验，一个只有 10 个概念的小型概念集聚类实验；一个来自 1998 年《人民日报》的 214 个概念的大型概念集的聚类实验[7]。

1 属性和属性值抽取

图 1 给出属性和属性值抽取流程。给定一段文本，首先用 GATE (general architecture for text engineering)实现分句、分词、消歧、命名实体识别、词性标注和句法树分析，在此基础上分 4 个步骤进行属性和属性值抽取：1) 根据语言学知识和本体抽取概念；2) 依据数据库多值依赖理论抽取属性值，特别是识别混合内容的属性值；3) 根据依存关系过滤掉不大可能担当属性的短语，根据意见挖掘机制抽取语句中的属性，根据半监督学习联想不在语句中出现的属性；4) 基于本体生成三元组<概念，属性，值>。具体细节见文献[8]。

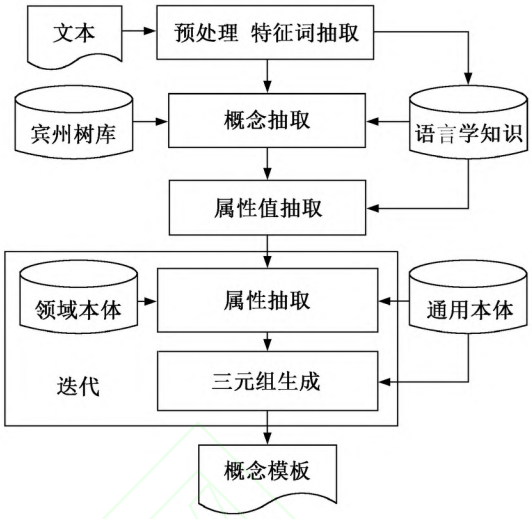

图 1 属性和属性值抽取过程

Fig. 1 Precess of extracting attribute; attributes and attribute values

2 基于属性和属性值组合的概念模板

本文基于属性/属性值的概念模板的构建比在其他语义分析中使用的概念模板，至少在两个方面要简单得多。首先，我们只抽取作为修饰语的属性值和概念，而忽略了语言结构的内容，在语言结构中概念是论元[9]。提出这种简化模板是为了比较基于属性和属性值的概念聚类和基于复杂语言关系的概念聚类，结果发现前者比后者在概念聚类方面具有更好的表现性能。其次，文本模板是基于单词的，不需要进行复杂的关系分析。但当建立完整的概念描述，并且允许更加一般的文本模板描述概念时，进行全面的关系分析是必要的，但计算昂贵得多，尤其是当使用 Web 数据时。我们发现，简单的文本模板不需要复杂的关系分析，就足以提取大量概念实例，其精度和性能良好。

2.1 基于属性值的概念模板

除了上面提到的两个简化，抽取属性值的方法类似文献[10]中使用的方法，即：仅仅考虑表达潜在的修饰语。我们使用的抽取属性值的模板如下。

模板 1 *A+N*。其中，名词 N 为概念，属性词 A 为属性值，(*)代表通配符。该模板匹配的文本例子如“一条白色的狗趴在门口左侧”。

属性词，也称“非谓形容词”或“区别词”，是从传统语法中名词、动词和形容词中脱离出来的一类新兴词类。全面统计考察《现代汉语词典》(第 5 版)标注的属性词来源、意义和功能，发现属性词是定

位、黏附性较强的黏着饰词, 其内部结构对分布功能会产生重要影响。属性词与所饰词语存在着较密切的语义结构关系。属性词表示事物的属性或等级, 少数属性词(修饰动词时)表示动作的方式或性质。属性词绝大多数只做定语, 小部分只能做定语和状语。属性词可以组合的实词, 绝大多数是名词, 少部分可以是谓词。《现代汉语词典》(第 5 版)第一次标注了属性词, 标注为属性词的共有 550 个(属性词义项的有 615 个), 并在词条中释义为: “属性词只表示人、事物的属性或特征, 具有区别或分类的作用。属性词一般只能做定语, 不能做谓语。

模板 2 *Ad+N*。其中, 名词 N 是概念, 副词 Ad 是属性值, 匹配该模板的文本如“小狗”、“老李”。

模板 3 *N+N*。其中, 第一个名词 N 是概念, 第二个名词 N 是属性值, 匹配该模板的文本如“张经理”、“李教授”。

模板 4 * N+是+[A|N]*。其中第一个名词 N 为概念, 属性词 A 或第二个名词 N 为属性值, 匹配该模板的文本如: “桌子是木质的” (N+是+Adj); “李莉是女性”(N+是+N)。

2.2 基于属性的概念模板

模板 5 *N+的+N*。其中 N 为名词, 第一个 N 是一个概念, 第二个 N 是一个属性, 模板中[的]的限制是为了确保第一个 N 实际上代表了一个概修饰语。匹配该模板的文本如“房子的价格又涨了”。

2.3 基于属性和属性值组合的概念模板

模板 6 *N+的+N+是+[N|Q|A]*。其中 N 为名词, 第一个 N 是一个概念, 第二个 N 是一个属性, 第三个 N 为属性值, Q 为数词, A 为属性词。匹配该模板的文本如: “张三的手机号码是 13338888666” (N+的+N+Q); “雪的颜色是白的”(N+的+N+A); “手机的生产厂家是惠普公司”(N+的+N+N)。

模板 7 *P+的+N+是+[N|Q|A]*。其中 N 为名词, 是一个属性, P 为代词, 是一个概念, 第二个 N 为属性值, Q 为数词, A 为属性词。匹配该模板的文本如“我的手机号码是 13338888666”。

模板 8 *N+是由+N+N+的*。其中 N 为名词, 第一个 N 是一个概念, 第二个 N 是属性值, 第三个 N 是属性。匹配该模板的文本如“手机是由惠普公司生产的”。

上面的 8 个模板都满足 Hearst[11]关于好模板的必要条件: 频繁的, 精确的, 容易识别的。

3 数据采集

最近几年, 越来越多的证据表明, 使用 Web 作为语料库能大大减少数据稀疏问题, 而且其大小也弥补了平衡不足问题[12]。使用 Web 作为语料库, 特别是当使用简单的文本模板抽取语义关系时, 好处甚至超过了使用大规模语料库(如 1998 年《人民日报》)。表 1 是关于使用 Web 和 1998 年《人民日报》的一些模板实例数字对比。

我们使用 Google 搜索引擎从网上收集数据, 通过 Google Web API (http://www.google.com/apis/)免费访问。对每个搜索请求, 该 API 只允许返回前 1000 个结果, 为了克服这个限制, 我们使用 Google 搜索请求的 daterage 功能。此功能允许用户对搜索空间分成若干时期, 因此可以检索已在指定时期更新的网页。这里提出两个实验, 目标是使用 daterage 功能对每个检索收集 10000 个相匹配的结果。我们的搜索时间为 100 天, 从 2000 年 1 月开始, 至 2011 年 5 月(我们使用的程序不能保证能收集到期限内的所有实例, 因为如果在一个时期内有超过 1000 的实例, 则只有前 1000 的实例被收集)。

我们对 Google 的搜索请求采取的一般形式是 “S1* S2”, 其中 S1 和 S2 是两个字符串, 通配符(*)表示一个未指定的词语。例如, 与搜索请求“一辆 * 车”相匹配的实例有: [一辆红色的车], [一辆大卡车], [一辆跑车], …。值得一提的是, Google 不重视标点符号, 这有助于我们的分析。

当接收 Google 检索的结果时, 并没有实际访问 Web 页面, 相反我们处理的是由 Google 返回的片段(片段是通过在 Web 页面中嵌入 HTML 标签所摘取

表 1 1998 年《人民日报》和 Web 中一些模板的使用频率比较

Table 1 Comparasion of the frequency of the modelin "1998 People's Daily" with the model in some Web

	模板	使用频率	
		Web	1998 年《人民日报》
1	A+N	161560	8768
2	Ad+N	193340	78610
3	N+N	184243	28905
4	N+是+[A|N]	26452	56
5	N+的+N	43414	128
6	N+的+N+是+[N|Q|A]	67545	210
7	P+的+N+是+[N|Q|A]	87691	76
8	N+是由+N+N+的	28750	8

的部分文字, 我们通过移除 HTML 标签, 抽取检索请求里指定的目标文本来处理这些返回的片段)。

4 实验分析

使用 Google 的一个缺点是, 即使把日限制增加 20000, 也不会真正实现对 1998 年《人民日报》100000 个名词性概念的灵活聚类。出于这个原因, 我们在两个实验中用到的概念集要小得多。第一个集合允许对 10 个概念聚类结果进行比较, 在第二个集合中包含 1998 年《人民日报》中的大量概念。

4.1 小型概念集实验

在第一个实验中, 聚类的概念有: 毛泽东、邓小平、战争、环保、汽车、中共、中国、走私、计算机以及操作系统, 用这 10 个概念来对基于属性的概念聚类、基于属性值的概念聚类和基于属性–属性值的概念聚类进行对比, 该实验中使用第 2 节描述的模板收集概念描述。

聚类输入的是一个频率表, 行表示概念, 列表示属性值、属性或者属性和属性值。表中的每一个单元包含了概念与属性或属性值共同发生的频率, 聚类之前, 频率转换成 t 检验的加权值。t 检验公式如下:

$$t_{i,j}=\left(\frac{\mathrm{Count}(\mathrm{Concept}_i,\mathrm{Attribute}_j)}{N}-\left(\frac{\mathrm{Count}(\mathrm{Concept}_i)\times\mathrm{Count}(\mathrm{Attribute}_j)}{N^2}\right)\right)\Bigg/\sqrt{\frac{\mathrm{Count}(\mathrm{Concept}_i,\mathrm{Attribute}_j)}{N^2}},\quad(1)$$

其中 N 是关系总数, Count(*)是一个计数函数, 使用 CLUTO[13]的 vcluster 命令来进行聚类, 相关参数: similarity function =extended jaccard coefficient, 聚类方法=图形分割, 聚类数=10。

表 2 是使用不同大小的向量时, 基于属性值、属性及其组合的聚类精度。结果表明, 概念描述的向量值为 20 时, 属性(94%)远比属性值(61%)准确; 大于 20 时, 属性的精度变化可以忽略不计, 而属性值的精确度有所提高, 但比属性(94%)的精度低得多。这表明, 属性比属性值有更多的识别力: 一个大小为 20 的属性向量几乎和使用 5 倍大小的向量的结果一样准确。最有趣的结果是, 组合属性和属性值的向量, 随着向量值的增加, 可以得到完美的精度。这表明, 虽然属性是一个通用内容的好方法, 但并不是所有的概念内容, 概念的内容应该包括属性/属性值对。

4.2 大型概念集实验

为了得到一个更现实的评测方法, 实验中使用一个更大的概念集合, 我们选择 1998 年《人民日报》预语库中 214 个来自 13 个不同类别的概念。

属性和属性值的频率如在第一个实验中一样被再次收集。然而, 这些数据的使用方式不同, 为了定义权重, 我们在 t 检验中使用布尔值, 来代替原来的频率, 这只会影响 $\mathrm{Count}(\mathrm{Concept}_i, \mathrm{Attribute}_j)$, 其他计数将不受影响。把所有正频率视为 1, 其他值为 0, 这消除了在原始数据频率变化的影响。直观看来, 频率不会被添加到概念的语义中。我们感兴趣的是, 概念有一个给定的属性/属性值的事实, 而不管遇到多少次。转换表是一个二进制表, 其中只含有 0 和 1。表 3 是基于属性和属性值相结合的布尔和频率的列联表, 表明布尔数据比表 1 中更准确。

为了聚类, 我们使用扩展的 Jaccard 相似度函数来计算概念间的相似度, 公式如下:

$$\mathrm{sim}(\mathrm{Concept}_m,\mathrm{Concept}_n)=\frac{\sum_i(t_{m,i}\times t_{n,i})}{\sum_i(|t_{m,i}|+|t_{n,i}|-(t_{m,i}\times t_{n,i}))},\quad(2)$$

其中 $t_{m,i}$ 和 $t_{n,i}$ 是概念 m 和概念 n 的共有属性/属性值 i 的加权值, 计算公式如式(1)。

计算每对概念的相似度, 组成相似矩阵, 并传送给 CLUTO 以便聚类。然后, 用以下参数调用 CLUTO 的 scluster 命令: clustering method = Graph Partitioning, 聚类数=214。评测的结果如表 4 所示。

表 2 使用不同大小的矢量时, 属性值, 属性及其组合的聚类精度

Table 2 Cluster precision using different sizes of vector, property values, the properties and their combination

模板	聚类精度/%				
	向量值=20	向量值=30	向量值=40	向量值=50	向量值=100
只使用属性值 1, 2, 3, 4	61.0	62.0	62.3	62.8	63.0
只使用属性 5	94.0	94.3	94.5	94.5	95.0
同时使用属性值和属性 6, 7, 8	70.0	82.0	88.0	92.0	96.0

表 3　基于属性和属性值结合的布尔和频率的列联表

Table 3　Column league table of the Boolean and frequency based on the properties and property values of the combination

系统答案		模板答案	
		Yes	No
布尔	Yes	1294	503
	No	387	20607
频率	Yes	1117	950
	No	564	20160

表 4 中，$P=\frac{a}{a+b}$，$R=\frac{a}{a+c}$，$\text{Falseout}=\frac{b}{b+d}$，$\text{Tureout}=\frac{c}{a+c}$，$F=\frac{2RP}{R+P}$。$a$, b, c 和 d 来自列联表(表 5)。

表 4　基于属性值，属性及其组合的聚类评测

Table 4　Evaluation based on the attribute values, properties, and their combination of clustering

模板	*P*/%	*R*/%	Falseout/%	Tureout/%	*F*/%
1, 2, 3, 4	58.48	52.23	1.98	61.10	55.18
5	55.20	77.83	0.87	20.19	64.59
6, 7, 8	73.00	81.11	0.10	10.98	76.84

表 5　列联表

Table 5　Column league table

系统答案	模型答案	
	Yes	No
Yes	*a*	*b*
No	*c*	*d*

当用精确度衡量时，基于属性值的概念模板比基于属性的产生的聚类结果要好(分别是 58.48%和 55.20%)，但其他概念模板都显示基于属性比基于属性值的更好：基于属性值 F=55.18%，基于属性 F=64.59%。造成这种差别的原因是，如果每个概念能够正确聚类，测量精度仅仅做简单的计算，而其余的方法关注概念之间的关系。

通过我们的算法，聚类了 30 个有关汽车概念的属性：雨刮器、方向盘、照明、发动机、底盘、座椅、电池、轮胎、速度、重量、颜色、舒适度、外观、内饰、历史、耗油量、图案、品牌、规格、拥有者、驾驶员、制造商、道路、导航仪、设计、产地、价格、销售、购买、安全。

5　结语

简单的文本模板可以用来自动抽取基本的基于属性和属性值的概念描述，以达到概念聚类的目的。初步的研究结果表明：首先，当对大量数据(如网络等)进行访问时，这些简单的模板可能足以计算进行识别的描述，至少对明显属于不同类的概念集是可行的。其次，我们发现，即使属性比属性值少，基于属性的描述也没有必要与基于属性值的一样长来达到一样好或更好的识别结果。最后，我们发现最佳概念描述应该包括属性和属性值的组合。我们下一步的工作是通过聚类得到的概念属性，自动生成商品数据库。

参考文献

[1] ACE. Automatic content extraction 2008 evaluation plan (ACE08) [EB/OL]. (2008–05–30) [2012–04–10]. http://wenku.baidu.com/view/71385200de80d4d8d15a4f6e.html

[2] Schutz A, Buitelaar P. RelExt: a tool for relation extraction from text in ontology extension // Proceedings of the 4th International Semantic Web Conference. Galway, 2005: 593–606

[3] Katrenko S, Adriaans P. Learning Relations from Biomedical Corpora Using Dependency Tree Levels // Proc Benelearn conference (2006). Cambridge: MIT Press, 2006: 867–877

[4] ACE. The nist ace evaluation website [EB/OL]. (2007–09–06)[2012–04–12]. http://www.nist.gov/speech/tests/ace/ace07/

[5] 程显毅，朱倩，王进. 中文信息抽取原理及应用. 北京：科学出版社, 2010

[6] Maite T, Julian B, Milan T, et al. Lexicon-based methods for sentiment analysis. Computational Linguistics, 2011, 37(2): 267–307

[7] 俞士汶，朱学锋，王惠，等. 现代汉语语法信息词典详解. 北京：清华大学出版社, 2008

[8] 程显毅，朱倩. 未定义类型的关系抽取的半监督学习框架研究. 南京大学学报：自然科学版，2012, 48(4): 878–892

[9] 朱聪慧，赵铁军，韩习武，等. 动词次范畴英汉论元对应关系获取. 中文信息学报, 2010, 24(2): 23–29

[10] Almuhareb A, Poesioa M. Attribute-based and value-based clustering: an evaluation. Computational Linguistics, 2006, 32(1): 1–8

[11] Hearst M A. Automated discovery of WordNet relations // Fellbaum C. WordNet: an electronic lexical database. Cambridge: MIT Press, 1998: 879–911

[12] Keller, Lapata M. Using the Web to obtain frequencies for unseen bigrams. Computational Linguistics, 2003, 29(3): 214–219

[13] Karypise Lab. CLUTO — Software for Clustering High-Dimensional Datasets [CP/OL]. (2006) [2012–03–20]. http://glaros.dtc.umn.edu/gkhome/cluto/ cluto/overview