

基于义类搭配的汉语文本特征信息抽取研究^①

郭文宏 张永奎 龙一飞

山西大学计算机科学系

摘要: 本文通过统计与人工辅助分析相结合,建立了义类搭配信息集,并在此基础上提出了一个利用语义相关信息、以文本特定义类代码驱动的深度扫描算法来获取文本特征信息。该项研究对情报检索、自动文摘、自动分类等领域的研究有一定参考价值。

关键词: 信息获取 语义分类体系 深度扫描 义类搭配

Research of Extraction of Chinese Text Information based on Semantic Category Collocation

Guo Wenhong Zhang Yongkui Long Yifei

Department of Computer Science, Shanxi University, Taiyuan

Abstract A Semantic Collocation Information Table is built by combining statistics and human aided analysis methods, and then an algorithm which make use of semantic collocation information is provided. The Algorithm uses a depth scanning algorithm, which is specific semantic code driven, to extract the characteristic information of texts. The idea of the research is valuable to concerned domains such as information retrieval, automatic abstraction and automatic classification, etc.

keywords INFORMATION EXTRACTION, SEMANTIC CATEGORY SYSTEM, DEPTH SCANNING, SEMANTIC CATEGORY COLLOCATION

一、引言

文本信息的获取无论在信息检索和中文信息处理中都有极其重要的作用。当前,信息获取在词法、语法层面的研究较多。我们希望在基于规则和统计的基础上,从语义层面上做些探讨。尽管机器自动语义分析的研究工作十分艰辛,然而要想使信息处理系统具备相当水平的智能,就不能没有一定深度的语义分析,语义理解应该是自然语言理解的一种理想境界。

我们通过对一定规模标注语料义类组合的统计,辅以人工分析加工,建造了一个义类搭配信息表,提出了利用语义相关信息、以文本特定义类代码驱动的深度扫描获取文本特征信息的算法,在此基础上实现了一个试验系统。系统所处理的信息源为经过分词、义类标注等

^① 国家自然科学基金资助项目 69575011

加工处理后的文本，其存储形式为文本方式。我们研究获取信息的目的是提取文本的语义特征信息，用于基于文本的智能检索和信息过滤。

二、语义类别的分析及义类搭配关系的获取

传统的检索中文本分析大多停留在词一级，通过查主题词表或简单词频统计获取标引词或关键词，来确定文本的类别。为了提高特征信息获取的精度，系统对所处理的文本进行预处理，即对其进行分词和义类代码标注，以作为算法研究的基础。我们采用的语义分类体系以《同义词词林》为基础。《同义词词林》描述了一个由广泛概念到具体词义的语义分类体系，与此分类体系相对应的是一个词义的编码体系。机读资源的编码形式为：

<编码> ::= <大类> <中类> <小类> <四级类> <五级类>

由于汉语动词在句子中的特殊作用，我们主要研究了语义分类体系中的运动类、事物类在真实文本中结合的特性。系统进行句子分析以谓语句动词为核心，然后考虑动词前后关联的名词或名词短语。基于这一思想，系统在文本中所提取的对象限于义类体系中的事物类（第一至第四类）和运动类（第六至第十类），从词性角度看这两类分别对应名词和动词。

在总结《同义词词林》组织同义词规律的同时，经过对其义类体系的分析，我们得到了通常运动类和事物类间的相互搭配关系。详见表 1（其中“+”表示有搭配可能，“-”表示无搭配关系）

	F (动作)	G (心理活动)	H (活动)	I (自然现象与状态)	J (关联)
A (人)	-	+	+	+	+
B (物)	-	-	-	+	+
C (时间与空间)	-	-	-	+	+
D (抽象事物)	-	-	+	+	+

表 1 基于《同义词词林》义类体系的运动类和事物类间的相互组合能力关系表

我们对一定规模的经过义类代码的标注的真实语料和部分《现代汉语词典》的释义文本进行义类搭配统计。统计出的义类间的组合频率同上面的分析结果很接近。义类搭配表的初建正是基于上述统计的搭配规则。考虑到义类搭配库在算法中的重要性，规则库经过了长期的人工检测与加工。

三、初始特征义类信息的选取策略

通过对一部分有标题的文本测试集的统计分析发现，多数文本的标题在很大程度上体现

文本主题，因此，在处理时，可考虑从文本的标题入手，首先找出标题中的关键词，然后，在文本中搜索标题中出现的关键词类，一旦搜索到便根据其出现在句子中的位置确定它在句子中可能充当的成分，然后，以义类搭配关系为主要依据，找到其上下文相关信息，进而提取出和主题密切相关的语义代码集。

在语义范畴内，词是能够表达一定意义的最小语义单位，因此在对语料进行语义分析时，词是基本的分析单位，然而词所表示的概念的外延不够广，而用词的义类代码来表示文本的特征要比单纯用词好的多，因为类要比单个的对象能更好地在语义空间涵盖所表示的信息特征。准确地找出初始特征词的特征义类信息(初始特征义类信息)，才能以其为驱动信号挖掘出语义上相关联的其他文本义类信息，初始特征义类信息找得越准确(即从量上看，找出的初始特征义类信息尽可能多；从质上看，要真正能够体现文本的主题)，最后提取出的信息就越接近文本主题信息。为了尽可能准确地找到初始特征义类信息，我们试验了两种方法，一是首先对所处理文本中表示运动类和事物类的词所对应的义类代码进行频次统计，根据统计结果按特定比例确定阈值，从而选出一定数量的高频义类代码，在以后检索中就把这些义类代码作为初始特征义类信息；第二种方法是以文本标题中表示运动类和事物类的词的义类代码作为初始特征义类信息。从试验结果看，在少量测试文本的情况下，有标题的文本采用后一种方法要比前一种方法准确率高。但是，随着测试文本数量的增加，两种方法提取信息的准确率相差不大，本文提出的算法是上述二种方法——基于标题义类代码驱动和高频义类代码驱动结合起来进行文本信息获取的算法。

四、深度扫描算法

该算法根据用户对目标信息的要求将被处理的文本扫描有限次(N次)，并把上次扫描过程中提取出语义中的新义类代码作为下次扫描时的中心代码，而第一次扫描所用到的中心词对应的义类代码便是文本初始特征义类信息。它们是表示初始运动类和事物类的义类代码，这样如果把每次扫描过程中搜索到的符合一定语义规则的义类代码都当作一个元素，那么，系统对某文本T扫描N次后最终构成一个义类代码集，该义类代码集即是文本T的特征信息。算法如下：

- (1) 设 S1, S2 为空集
- (2) TitleFlag=true, 判断文本是否有标题, 如没有则 TitleFlag=false, 转(4)
- (3) 标题义类代码驱动——从标题中提取属于事物类与动作类的义类代码加入集合 S1
- (4) 高频义类代码驱动——统计文本中属于事物类与动作类的不同义类代码出现的频次, 设定阈值, 提取高频次义类代码加入集合 S2
- (5) 确定文本扫描深度 N (N>0)
- (6) For I=1 to N
 - 扫描待处理文本 T 中的第一个词对应的义类代码 CodeNext.
 - While NOT Eof(T)
 - If codeNext∈S1, 分析 codeNext 所在句子并查义类搭配表, 如有

```

    关联代码 codej , Then S1= S1+{ codej }
    If codeNext∈S2, 分析 codeNext 所在句子并查义类搭配表, 如有
    关联代码 codej , Then S2= S2+{ codej }
    扫描待处理文本 T 中的下一个词对应的义类代码 CodeNext,
Wend
Endfor
(7) If TitleFlag=false Then 文本特征信息为 S2 中的义类代码
    If TitleFlag=True Then 文本特征信息为 S2∩S1 中的义类代码

```

五、算法试验结果及分析

该算法的一个特点之一是通过用户对系统扫描文本的深度加以控制。试验表明, 当 $N < 4$ 时, 系统对文本的扫描深度不同, 文本特征信息获取的程度也不同; 当 $N \geq 4$ 时, 文本的扫描深度的增加对系统获取的文本特征信息程度的影响不明显。为了防止 N 过大时获取的信息失真, 算法控制义类代码集的元素保持在某个值内。

下面是该系统对文章《在新的历史条件下继承和发扬爱国主义传统》的分析结果, 文本的预处理格式如下:

```

TITLE: 在|Kb010301| 新|Eb280101| 的|Kd010101| 历史|Da070103| 条件|Da210401| 下
|Ca040101| 继承|Hj100301| 和|Kc010101| 发扬|Hj160101| 爱国主义|Di080113| 传统|

```

```

AUTHOR: 江泽民|Aa010201|

```

```

BEGIN: 在|Kb010301| 我国|| 历史|Da070103| 上|Cb030101| , |end| 爱国主义
|Di080113| 从来|Ka100201| 就是|Ja010102| 动员|Hc120201| 和|Kc010101| 鼓舞
|Je060201| 人民|Aa010201| 团结|Ie080101| 奋斗|Hj080101| 的|Kd010101| 一面
|Dd050205| 旗帜|Bp200201|

```

```

.....
爱国主义|Di080113| 是|Ja010101| 一个|| 历史|Da070103| 范畴|Dd050101| , |end|
在|Kb010301| 社会|Di010102| 发展|Ie120101| 的|Kd010101| 不同|Jb020101| 阶段|| ,
|end| 不同|Jb020101| 时期|Ca030101| 有|Jd010101| 不同|Jb020101| 的|Kd010101| 具体
|Ed100101| 内容|Dk100101| 。|end| 我们|Aa020201| 所|Kd010301| 讲|Hi120101| 的
|Kd010101| 爱国主义|Di080113| .....
1 9 9 8 || 年|Ca180101| 4 || 月|Ca210101| 2 4 || 日|Ca230101|

```

试验结果:

```

当N=1时 S1={ Da070103 Da210401 Hj100301 Hj160101 Di080113}
当N=2时 S1={ Da070103 Da210401 Hj100301 Hj160101 Di080113 Di080112 Df010203}
当N=3时S1={ Da070103 Da210401 Hj100301 Hj160101 Di080113 Ie080201
Df010203 Jd060501 }
当N=7时 S1={ Da070103 Da210401 Hj100301 Hj160101 Di080113 Ie080201
Df010203 Ha010101 Di080106 Di210204 }
当N=8时 S1={ Da070103 Da210401 Hj100301 Hj160101 Di080113 Ie080201

```

该算法使用义类代码集来表示文本的语义特征信息，在研究过程中，我们发现许多文本的特征用线性结构难以表示清楚。对于多主题的文本，能寻求更好的特征信息表示方法，将会给大大提高信息获取的准确度。

六、结束语

我们的试验系统要成为一个实用系统，还有许多工作需做，这是由中文信息的研究现状和语义本身的难度决定的。我们的算法是建立在许多标注和统计工作的基础上的，标注和统计的准确率对算法的结果会有直接的影响；义类搭配表规模小，收集的词搭配信息甚少，这也是影响结果的关键问题之一；在对测试结果进行评价时，由于采用人工判断的方法，因此结果难免有一定主观性。本算法的继续完善将是我们下一步所要做的工作。

总之，自然语言的语义研究任重道远。有些问题需要我们今后进行深入研究，

参考文献

- [1]. Tadashi Nomoto & Matsumoto, "Data reliability and its effects on automatic abstracting", Processing of fifth workshop on very large corpora. Aug. 1997, 113-126
- [2]. Gerard Salton, "Automatic text structuring and summarizing", Information Processing & Management, 1997, 33(2): 193-207
- [3]. Gerard Salton, "Automatic text decomposition and structuring", Information Processing & Management, 1996, 32(2): 127-138
- [4]. Cremmins E.T., "The art of abstracting", Information Resource Press, Arlington, 1996
- [5]. M. Taketa, F. Matsuo & J. Suda, "Identification of nouns of scientific and technical literature", Trans. Information Processing Society of Japan, 8, Aug. 1995, 1828-1837
- [6]. Ellen Riloff & Wendy Lehnert, "Information Extraction as a Basis for High-Precision Text Classification", ACM Transaction on Information Systems, Vol. 12, No.3, July 1994, 296-333
- [7]. 郭文宏、张永奎、余明山, "基于语义信息的分词后处理研究", 第五届中国人工智能联合学术会议论文集, 1998年, 227-231
- [8]. 王永成、许慧敏, "OA中文文献自动摘要系统", 情报学报, 1997, 16(2), 128-132
- [9]. 吴立德等, "大规模中文文本处理", 复旦大学出版社, 1997
- [10]. 陈海虹, "文献标引深度的控制", 情报学报, 1991, 10(3), 224-229