文本分析与信息检索

彭甫阳、何新贵
北京系统工程研究所
北京9702信箱19号，100101

摘要：本文对信息检索的范型进行了分类，并将信息检索过程分解为文本分析和信息检索两个阶段。重点分析了信息检索环境下自然语言文本分析技术的主要特征、各种语义驱动的文本分析理论与技术，简要介绍了应用这些理论与技术的典型系统。

一. 信息检索的三种范型

范型(paradigm)一词最初由T.S.Kuhn用来描述科学革命的结构。该词在科学哲学中常用来表示相关的许多概念，很难给出一个准确的定义。R. Floyd在其1979年的Turing演说'程序设计的范型'中使用该术语来称呼程序设计风格、相关的典型数据类型、操作和控制结构。我们在此使用该词的含义与R. Floyd基本相同，即表示不同的计算模型及围绕它们所形成的流派或文化。不同的范型具有不同的方法学，使用不同的工具，基于不同的传统，拥有不同的用户群，适于不同的典型应用。

已提出的信息检索范型主要有三种。最经典的范型是布尔检索模型和统计检索模型。这两种范型基于的方法相对简单但稳健可靠。第三种范型是基于人工智能的检索模型，采用自然语言处理方法和基于知识的方法。

信息检索研究为满足对信息的需求而对自然语言文本进行处理的理论与技术。下面的五类功能都应归入信息检索的范围。一、分类（将文本归入某一固定的分类集中）；二、检索（确定满足任一检索表达式的所有文本的位置）；三、抽取（识别文本中的有用信息，并用一种具有确定语义的形式表示之，如用关系数据库的域表示）；四、综合（产生一内容与原文相关联的新的文本表示）；五、问题回答（基于对文章内容的理解回答用户提出的问题）。

自然语言要想成为科学、办公室及其它公共领域的一种合适的交流媒体须克服自然语言的固有复杂性。人们已经进行各种努力以攻克这种固有复杂性。信息检索系统的开发与这些努力是紧密交织在一起的。虽然不一定非得对文本充分理解也能从文本中抽取一些有用信息，这已被某些统计方法和文本浏览方法的部分成功所证实，但联机文本的急剧增长要求有更好的方法来获得文本形式的信息。在六十年代和七十年代，文献分析的统计学方法占统治地位，而进入八十年代后，基于知识的文献分析方法开始居统治地位。

范型的转变大大扩展了基于知识的信息检索系统的应用范围。许多功能是以布尔模型和统计模型为基础的文本分析所实现不了的。基于知识的信息检索系统功能的明显增强主要是由于对自然语言处理的统一观点。统计文本分析方法强调词汇语义而忽视文法现象；语言学描述过分强调文法结构的作用，相对来说，语义规定受到忽视。与这些方法不同，基于知识的文本处理方法提供一种合适的平衡，综合语言方面的描述和领域有关知识的适当建模手段。

本文集中讨论基于人工智能的信息检索范型，特别是基于知识的信息检索环境中语义驱动的文本分析的讨论。

二、信息检索的两个阶段

信息检索过程可以分为两个阶段：文本分析阶段和信息检索阶段。文本分析阶段的目的是从无结构的输入文本获得结构化的文本代表（text representative）。文本代表可以是输入文本的文法结构、语义结构、语境信息，或者是这三者的组合。获取文法结构的技术可以是简单的功能词剔除法、文法模版法，或复杂的分析器。语义结构可以借助主题词典（用文本中的词的组合表示意义）或知识库（用某种知识表示语言构造的数据结构表示语义）来构造。语境信息使得领域知识和实用知识对检索系统来说显性化。

信息检索阶段信息的检索可以看作是一个推理过程，它识别不等价概念之间的联系，推出未提及的隐概念。概念的识别通常需通过大量的规则进行推理，每条规则基于文本代表（如文本中某一词汇模式的出现）推出某一概念的存在。

三、信息检索环境的文本分析

3.1 信息检索环境下文本分析技术的主要特征

下面讨论信息检索环境下任何文本分析程序都须考虑的一些主要要求。

1．文法应覆盖所讨论语言的主要语言现象。与理论语言学和计算语言学长期形成的注重研究选择性语料的传统不同，信息检索中的自然语言分析必须面对不同作者日常使用自然语言产生的大量真实文本。这些真实文本向我们展示了所有的语言现象。遗憾的是，三十多年的形式文法研究并未导致任何完整的自然语言文法的产生，即使对研究较多的西方语言，如英、法、德语，情形也大致亦如此。由于自然语言的固有复杂性，在短期内提供完整的自然语言描述看来是不太可能的。因此，信息检索应用应把目标定为提供具有足够的语言和概念覆盖范围的语言描述。达到此目标的一条比较成功的途径是把精力和重点放在所讨论语言的主要结构的描述上，撇开次要的有时甚至是可能引入歧途的语言构造。

这对分析器提出了一定的要求，要求分析器至少应具有区分重要词和非重要词的能力。所谓重要词包括概念上关联的词（它们表示领域知识库中的概念，如名词和形容词）和语言上关联的词（如否定词，某些连接词和修饰词等）。非重要词指与检索目标无关的词，是一个相对的概念。重要词应作为词汇文法描述的起点，而非重要词应在进一步分析之前剔除掉。

2．分析机构应具有较好的鲁棒性，不至于因文法描述的不完整性和语言文本错误而导致失败。因此，在较长时期内，自然语言分析器设计要求之一仍然是它能够从无足够词法和语法描述的情况下恢复。

自顶向下方法，或称期望驱动方法，是一种浏览文本中的特定信息而忽略不熟悉词或结构，从而避免语言复杂性困绕的一种较好方法。自顶向下处理和自底向上处理的结合使得文本分析系统能够利用分析过程中要用到的所有信息源。如果有比较完整的语法和词汇知识时，它能够进行深入的分析；当碰到生词或不熟悉的结构时也能进行浅层的分析。

另一种有用的方法是根据文本中词模式的出现进行概念推理。这些模式用某些简单的文

法模版或描述将相关的词结合起来，如规定某些词必须出现在同一句、 同一段落或其它指定上下文中。这些规则当然没有自然语言分析器那么精确， 但它不象许多分析器那样在碰到困难的语言结构时容易导致失败。

3．语法描述手段和分析系统必须为语言工程提供良好的方法和计算工具，即能进行语言知识的增加和修改。分析器需要拓展其覆盖面以不断逼近完整文法， 这意味着语言描述会不断变化。因此，要求语言描述工具和与分析器配套的工具，如文法编辑器、编译器、 调试器及集成化文法/分析器工作台，支持易扩充性、易维护性和增量式开发。

4．文法描述应提供篇章文法以便进行篇章级的语言描述。当今文献数据库中已越来越多地存放全文而非二次文献，但传统的信息检索方法（串模式匹配、布尔/邻接检索算符、以主题词表/分类为基础的领域模型）面对全文环境其性能急剧下降。我们必须寻求处理全文中知识的新的自然语言处理方法，特别是涉及超出句子级描述的语言现象。

3.2 信息检索中应用自然语言处理方法的可能行

自然语言处理所使用的技术很多，从简单的为抽取索引短语而设置的语法模版， 到利用自然语言分析器产生详细的文法和语义表示。非NLP技术包括各种由经过训练的或未经训练的人员进行的内容分析。 这些技术的共同点是它们都产生一种新的表示来显化原文本结构中的一些信息，以便于后续的强有力的推理。

如果允许一定的误差和非确定性，那么在受限领域进行语义分析， 在任何领域进行文法分析都是可能的。Sparck Jones和Tait的工作表明，即使非完整的、 有二义性的语义表示对检索也是有帮助的[9]。 所谓现在的自然语言处理技术面对真实文本必然失败的说法恐怕仅适用于象问题回答这样的复杂任务，并不适用于文本检索任务。

3.3 语义驱动的文本分析

我们这里所说的语义驱动的文本分析包括所有的以知识表示结构驱动自然语言分析的各种方法。 文法驱动的自然语言分析和语义驱动的自然语言分析的主要区别不是在是否有语义信息的存在，而是在使用这些知识进行决断的控制方式， 即是使用文法知识还是使用语义知识作为构造性决策的主要知识源。在语义分析器中， 从表层字串到表示结构的映射主要由表征特定论域的语义知识控制，其它知识源（如文法）只在需要时才访问。 在文法分析器中，这一映射主要由表征自然语言可接受的成分结构的文法规则控制， 领域知识只在文法分析完成后再起作用，也就是说， 文法驱动的分析器中分析过程的控制可以明显地分成顺序进行的几个阶段，其中文法成分分析优先于语义和语用解释。

概念分析、词专家分析和优先语义学是三类最常用的语义分析方法。

概念分析程序根据基于记忆的概念分析观进行语义分析， 近来这一方法已发展为标记传递方法(marker passing paradigm)[1]。一般采用某种受限的目标表示结构，如概念依存图，脚本，记忆组织包MOP，题元抽象单元TAU等。早期的概念分析程序采用某种全局的控制机构，后来逐步被更灵巧的守护神控制交换(control exchange via demons)模型所取代[7]。

词专家分析程序极力强调由词汇程序执行的相互之间的立即控制的作用。严格词汇化的分析过程的控制问题是词专家语义分析模型十分关注的问题[2, 3]。Hahn[8]给出了一个基于行为者消息传递方法的词汇分布文法的形式描述。

基于优先语义学的分析模型由Wilks提出[4, 5]。它与概念分析方法的主要区别在于应用的语义原语(semantic primitives)具有不同的理论基础，同时按优先顺序对非标准解释排队的能力亦不相同。Wilks的语义原语来自自然语言本身，而Schank则声言语义原语是完全独立于语言的实体。关于这两种模型方法上的差别的详细讨论，请参见[5]。

格框架分析程序介于语义驱动的分析器和文法驱动的分析器之间，其描述性知识是语义的，但多数处理工作是以文法驱动的方式进行的。另外一种折中方法是文法和语义集成分析模型，该模型按照每一分析步的实际需要平衡全语义成分和文法成分之间的控制，而没有任何固有的规则。

其它领域，如知识工程、计算语言学和认知科学等，提出的一些理论和技术经适当修改可以用于信息检索环境下的文本分析。框架(特别是格框架)、CD图、语义网、规则等在文本分析的各种知识表示中都得到了应用，如表示用户模型、世界知识、领域知识、词汇知识和篇章知识等；概念分析、标记传递机制和词专家分析已成为语义驱动的文本分析的主要认知模型。表格分析器(Chart parser)，如果和自底向上的控制策略相结合，特别适用于处理真实文本，因为其部分成分表示能力使得即使完整的分析失败，它也能输出一些有用信息。

四、典型系统

下面列表为几个采用基于知识的语义驱动方法的典型信息检索系统。

ADRENAL

主要特点是:

- 视信息检索为推理任务
- 用NLP技术产生文本表示，以便对文本内容进行更精确的推理
- 与传统的统计方法相结合的混合文本分析

ARGON

主要特点是:

- 基于框架的语言KANDOR
- 基于知识的文本检索系统

CANSEARCH

主要特点是:

- 分布式专家信息系统
- 以Prolog语言为基础
- 用于MEDLINE数据库中与癌症治疗有关的文献检索系统的前端

CODER

主要特点是:

。基于框架的方法表示文本、用户和查询知识
。处理框架定义的类型管理器和处理实例的对象管理器分离
。基于知识的文本检索
。框架语言既用于分析，又用于检索
。使用框架表示文献知识和领域知识
。系统按专家系统的面向对象方法构造
。使用范围为[0,1]的关联值
。在线帮助功能
。不同模块之间通过向黑板结构传递消息进行通讯
。系统实现语言为Prolog

I(3)R

主要特点是:

。使用关系数据库表示文献
。主要的知识处理工作限于检索，概念框架由系统与用户的交互过程创建，它含有对领域知识进行推理的识别规则
。系统由以黑板/调度器为中心的相互合作的一系列专家组成

RUBRIC

主要特点是:

。基于产生式规则的商品化系统
。系统实现语言为Common LISP
。系统按专家系统的面向对象的方法构造
。使用范围为[0,1]的关联值而不是取集合{0,1}
。检索过程中任一点的在线帮助

TOPIC

主要特点是:

。基于框架的语言
。基于知识的文本检索
。工作集中在文献的分析，使用语义分析将文本映射到框架表示结构

五、结语

可以胜任的信息检索系统必须能稳健地处理大量的真实文本。这意味着对自然语言处理我们应当持一种工程的、而非理论的观点。应用自然语言处理方法于信息检索的成功与否很大程度上取决于NLP技术的选择、应用领域的选择及知识获取的方法。本文的结论是语义驱动的文本分析和基于知识的信息检索可能是攻克信息检索问题的最好方法。

参考文献

[1] C.K.Riesbeck and C.E.Martin(1986)
Direct memory access parsing
in Experience,Memory,and Reasoning, J.L.Kolodner and C.K.RiesbeckZ(Eds)

[2] C. Rieger and S. Small (1979)
Word expert parsing
IJCAI-79, vol. 2, 723-728
[3] S. Small and C. Rieger (1982)_
Parsing and comprehending with word experts (A theory and its realization)
in Strategies for Natural Language Processing, W. G. Lehnert and M. H. Ringle (eds)
[4] Y. Wilks (1975)
An intelligent analyzer and understander of English
CACM, 18 (5), 264-274
[5] Y. Wilks (1983)
Deep and superficial parsing
in Parsing Natural Language. Proc. 2nd LuganoTutorial, 1981, M. King (Ed.)
[6] S. L. Lytinen (1987)
Integrating syntax and semantics
in Machine Translation: Theoretical and Methodological Issues
[7] M. G. Dyer (1981)
Integration, unification, reconstruction, modification: an eternal parsing braid
IJCAI-81 (7th), vol. 1.
[8] U. Hahn and U. Reimer (1986)
Topic essentials
Technical Report TOPIC-19/86
[9] K. Sparck Jones and J. I. Tait (1984)
Automatic search term variant generation
J. of Documentation, 40 (1), 50-66.

Text Analysis and Information Retrieval: An Overview

Fu-Yang Peng and Xin-Gui He
Beijing Institute of Systems Engineering
P. O. Box 9702#19, Beijing 100101

Abstracts--A classification of information retrieval paradigms and segregation of information retrieval process into text analysis phase and information retrieval phase have been proposed in this paper. Then emphasis is shifted to the issue of text analysis in information retrieval environment. criteria characterizing text analysis techniques in IR environment have been given, semantically driven text analysis theories and techniques as well as several typical IR systems using semantically driven approach have been discussed.