TBIRS 中的信息抽取方法①

张永奎　　李红涛
（山西大学计算机科学系，太原　030006）

摘 要：TBIRS 是我们研制的一个基于文本的信息检索系统。该系统采用语义驱动的文本分析方法，将每一文本转换为一种标准形式，用于构造各种结构化的包括相关知识的文本数据库。本文以从百科全书《哺乳动物》中自动提取信息为例，介绍了 TBIRS 中的知识获取方法。

Information Extracting Method In Text—Based Information Retrieval Systems

Zhang Yongkui and Li Hongtao
(Dept. of Computer Science, ShanXi University, Tai Yuan　030006)

ABSTRACT: A TBIRS (stands for Text—Based Information Retrieval System has been developed. The system which uses semantic—driven text analysis techniques will transform each of texts into a standard canonical form. This can then be used to construct various forms of structured textual data—base with incorporated knowledge of their contents. In this paper an example, autoextracting information from a mammals encyclopedia, is presented to show the method of knowledge acquisition in the TBIRS.

1 引　　言

自由文本检索系统(Free Text Retrieval System)依据一个文档中出现过的一个或多个词来查找文挡。当使用的词是非常用词时，这种技术可以满意地工作。如果被检索的文档集合中所使用的词几乎在所有文档中都出现，对于这些文本，采用基于简单关键词的检索技术是不适用的，基于文本的智能检索可以较好地解决这个问题。智能信息检索系统是人工智能与信息检索技术相结合的产物，许多研究者把自然语言处理方法应用于智能信息检索中，并取得了一定的成效[1-10]。

几年来我们致力于文本信息抽取与信息检索的研究[11-17]，试图利用自然语言处理技术从无结构的自由文本中提取出相关信息并构造结构化的包括相关知识的非文本数据库(知识库)，从而实现对数据库的智能信息检索。应用自然语言处理方法于信息检索的成功与

① 本项研究得到国家自然科学基金(69375016 号)资助

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方法。作为实验，我们限定了一个研究领域即田园指南(Field Guids)类的文本[16]。这些文本都是描述生物标本的，其作者的目的是帮助读者准确地辨认标本。我们研制了一个基于文本的信息检索系统TBIRS，统采用语义驱动的文本分析方法[14]，可把每一文本转换成为一个标准形式，然后就可以用于构造各种结构化的包括相关知识的文本数据库。数据库的检索及自然语言查询运用差别表分析法和DCG方法，以灵活的自然语言提问方式进行查询[13]。

本文以从百科全书《哺乳动物》中自动提取信息为例，介绍TBIRS中的知识获取方法。

2 源文本文件样例

按照自动化领域的概念，信息自动抽取可以定义为：

$$Ma=(K,\Sigma,\Gamma,\sigma,q_0,F)$$

式中K为一个有限状态集合；Σ代表输入符号集合，即原始文献；Γ代表能输入和输出的有限符号集合，如原始文献或其它任何数据；σ是另一运动函数，由选择规则和转换所定义，这个函数还定义自动的输出；q_0代表初始状态，与原始的文献输入相对应；F代表结束状态，与输出的完成相对应。

源文本文件——百科全书《哺乳动物》[18]——描述了遍布于世界各地的426种哺乳动物，从最小的鼠类到最大的鲸。全书共426篇文章，每一篇包括一种动物的名称、分类、形态描述、分布地区、栖息地及行为特征、饮食习惯和繁殖情况等信息。本文以百科全书《哺乳动物》描述文本中较复杂的形态特征描述部分DESCRIPTION为例，说明形态描述的知识的表示与知识获取方法。

59

DESCRIPTION Of medium size, a robust fruit bat with large, rounded head; large eyes; prominent, long, tubular nostrils extending sideways; short tail. Fur soft and long, gray—brown above, darker along mid—back and on spinal stripe; underparts yellowish white; neck and sides tinged with yellow — orange; wings and ears speckled with irregular yellow spots. Length of head and body about 8—9 cm (3.2—3.5 in), forearm 5.5—6 cm (2.2 — 2.4 in); weight up to 45 g (1.6 oz).

图1 一个源文本文件样例(摘自文献[18])

3 知识的表示

3.1 对结构的标准形式的需求

在许多基于知识的系统中，知识表示结构中的链用于表示结构中部件的意义或真实世

界关系。这些允许程序使用结构来处理省缺值、部件间的关系、继承和具体情况。但是结构的多变形式并不理想地适合于从文本中有效地进行变换，并且它的固有的复杂性将显著地增加使这个提取信息系统适用于其它领域的艰难性。为使这个系统尽可能地是通用的和实用的，我们决定不试图强加任何与对象相关的结构。在分析过程中，除了使用一个四级层次的统一形式外，并不把一个对象各种部件间的内部联系作为部件的性质被存储。

3.2 描述结构的定义

一个描述结构被定义为一个四层结构，它是由一个描述文本变换的(见图 2)。

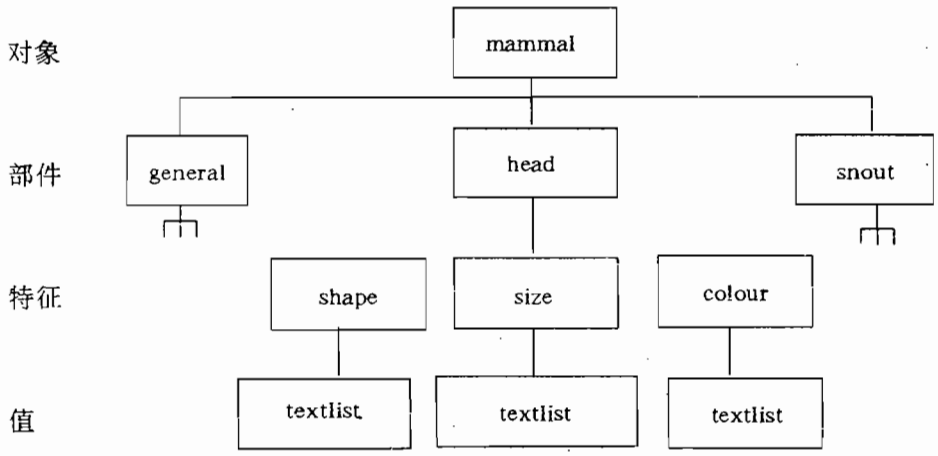

图2 一个描述结构的例子

结构中的各层是：

对象：顶层，表示一个对象的完整描述，如一个动物。

部件：支持对象的可能被描述的每一部分，如头、眼睛、牙齿等。

特征：一个部件应具有的每一个有意义的特征，如大小、形状、颜色等。

值：抽取的每一特征的值。例如，头的大小[large]，头的形状[round]。

3.3 Prolog 表示

从正文中提取的知识是用 Prolog 谓词来表示的。正文中 DESCRIPTION 部分内容较复杂，拟采用结构模板来表示结构化的知识。整个模板用一个 Prolog 事实 obj—structure 表示。与图 2 所示例子对应的一种动物形态特征描述的结构模板如下所示：

```
obj__structure(
animal(
        general(shape(Gen1),colour(Gen2),size(Gen3)),
```

```
head(shape(Hea1),colour(Hea2),size(Hea3)),
snout(shape(Sno1),colour(Sno2),size(Sno3)),
...
)).
```

结构模板的第二层是被描述的实体，即由部位名词汇表中挑选并加工合成的 26 个关键部位。我们采用了基于统计、词典的计算机辅助编辑方法获取部位名词汇[16]，包括真实文本的预处理、词的统计分析、专门词汇的识别、使用核心词汇表和电子词典中的语义码、附加部位名信息源等。

3.4 修改结构模板

模板作为一个 Prolog 事实随同所有其他处理规则被读入机器。可以使用一个文本编辑器修改模板，然后重新读入 Prolog 数据库中。如果被处理的文本不包括某一特征，那么可以通过省略结构中相关的部分而忽略它。类似地，可通过向结构附加的办法使新的谓词和/或部件能被检测。

4 信息的自动抽取

4.1 语义驱动的文本分析

语义驱动的文本分析包括了以知识表示结构驱动自然语言分析的各种方法。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。文本分析过程通过词汇查询、文本修正、文本切分三个阶段来完成[17]。最后，原始文本被分割为小的描述单元片断(segment)，每个片断同相应的描述实体相联系，其描述焦点是通过关键词的识别来实现的。单元片断表示形式如下：

```
seg(<Seg_number>,<Keyword>,<Text_list>,<Original_form>).
```

其中，seg 事实的四个项依次表示文本片断的编号、描述对象的实体、与该实体相关联的文本片断及该片断中出现的真实部位名。

4.2 特征信息提取

特征抽取是以文本分解后的 segment 片断为依据的，通过特征规则识别动物形态描述的各个部位的特征。对于每种特征（颜色、形状、大小）都有相应的语义代码，例如：

```
colour:   'CO'——colour in LDOCE
          'co'——colour in core_word vocabulary
          'Ap'——Appearence
```

```
shape:    'Sh'——primary code : shape
size:     'Si'——words like large and small in the core word vocabulary
          'si'——words like length and size in the core word vocabulary
          'NB'——numeric is a LDOCE code
          'Qu'——quantity is attached to qualifying words like much
```

特征提取过程的控制通过上文所述的结构模板来实现。采用的最基本的操作是 univ 谓词(=..)[19],它把一个模板结构分解成各种表和子表。分析谓词的最后结果返回一个带有模板中所有未例示的变量的完整的描述结构。下面给出自动生成的与图 1 所示文本样例对应的一种动物形态特征描述文本数据库内容:

```
description(
  m59(
      general(size(["medium-size","45-gram","1.6-oz"])),
      head_body(size(["length","8-9-centimetre"])),
      coat(size(["long"])),
      upperpart(colour(["brown"])),
      head(shape(["round"]),size(["large"])),
      ear(shape(["irregular","spot"]),colour(["yellow-spot"])),
      eye(size(["large"])),
      snout(size(["long"])),
      neck(colour(["tinge","with-yellow-orange"])),
      forearm(size(["5.5-6-centimetre"])),
      back(colour(["dark","stripe"])),
      wing(shape(["irregular","spot"]),colour(["yellow-spot"])),
      underpart(colour(["yellow-white"])),
      tail(size(["short"]))
      )).
```

5 结　论

本文对语义驱动的文本分析及数据库自动生成进行了初步研究和探讨,对于知识获取亦作了有益的尝试,并自动生成了百科全书 426 种哺乳动物描述文本数据库(知识库)[17]。从提取的结果来看,对于文本中较难处理的形态特征描述部分,能较好地提取出相关信息,结果令人满意。这里讨论的是生物学领域的知识获取问题,但这一整套方法同样适用于其它主题领域。实验结果表明,对于受限领域中的具有一定写作规范的文本,这种方法是可行和有效的。对于不同的主题领域,只须重新建立相应的规则集。然而,由于自然语言固有的复杂性,在信息提取过程中还存在一些不足之处有待完善,如:对于否定、比较以及各部位的用途等处理不太理想。这亦是我们进一步研究和解决的关键问题。

参 考 文 献

[1] P. Norvig; P. Jacobs ed., Text—Based Intelligent System, Artificial Intelligence, 65(1994)

[2] W. B. Croft, Knowledge—Based and Statistical Approaches to Text Retrieval, IEEE EXPERT, 1993

[3] B. Bernard and R. Daniel, Automated Knowledge Acquisition from Regulatory Texts, IEEE EXPERT, 1992

[4] G. Sabah, Knowledge Representation and Natural Language Understanding, AICOM, 6(1993)

[5] M. S. Palmer etc., The KERNEL Text Understanding System, Artificial Intelligence, 63(1993)

[6] 李东、董振东、黄昌宁,"中文信息处理平台(CIPP)"工程,《计算语言学研究与应用》北京语言学院出版社,1993

[7] 王永成等,《中文信息处理技术及其基础》,上海交通大学出版社,1992

[8] 王建波、王开铸,自然语言篇章理解及其基于理解的自动文摘研究,《中文信息学报》,1992年第6卷第2期

[9] 彭甫阳、何新贵,文本分析与信息检索,《计算语言学研究与应用》,北京语言学院出版社,1993

[10] 赖茂生、王延飞、赵丹群,《计算机情报检索》,北京大学出版社,1993

[11] Y. K. Zhang(张永奎)、J. R. Cowie, Using Cluster Analysis for Processing English Texts, In Proc. of Pacific Asia Conference on Formal and Computational Linguistics (Taipei, Aug. 30—31), 1993

[12] 张永奎,聚类分析在自然语言处理中的应用,《情报学报》,1993年第12卷第5期

[13] 张永奎等,动物数据库自然语言前端的设计与实现,《计算语言学研究与应用》,北京语言学院出版社,1993

[14] 张永奎,从文本中提取信息,《情报学报》,1994年第13卷第2期

[15] 张永奎、J. R. Cowie,机器可读词典的快速查找技术,《中文信息学报》,1994年第8卷第2期

[16] 张永奎、李红涛,从田园指南文本中获取部位名词汇,《人工智能新进展》,清华大学出版社,1994

[17] Y. K. Zhang(张永奎)、J. R. Cowie, Building a Mammalsbase from an Encyclopedia, In Proc. of International Conference on Information & Knowledge Engineering (Dalian, PRC), 1995

[18] L. Boitani and S. Bartoli, The Macdonald Encyclopedia of Mammals, Macdonald, London, 1983

[19] The SD—Prolog Programmer's Reference Manual, Pembroke House, Camberly, Surrey, 1987