目 录

第一章 计算机情报检索概述 1—1

1.1 情报与社会的发展 1—1

1.2 情报检索与文献检索 1—2

1.3 文献情报检索系统的基本功能 1—3

1.4 文献情报检索系统的基本原理 1—4

1.4.1 文献与文献标识 1—4

1.4.2 文献——语词矩阵 1—6

1.4.3 三种基本的文献检索方式 1—7

1.5 联机情报检索 1—10

1.6 关于本课程的说明 1—11

第二章 基于倒排档的检索系统 2—1

2.1 倒排档检索技术发展简史 2—1

2.2 布尔逻辑 2—5

2.3 典型的文档结构 2—7

2.4 检索过程 2—11

2.5 检索式的逻辑运算 2—12

2.5.1 运算顺序的正确控制 2—13

2.5.2 集合的逻辑运算 2—17

2.6 倒排档检索机制的加强 2—19
2.6.1 邻接 2—19
2.6.2 截词 2—21
2.6.3 范围检索 2—21
2.6.4 加权 2—21
2.7 商业性检索系统介绍 2—22
2.7.1 DIALOG 系统 2—23
2.7.2 STAIRS 系统 2—24
2.7.3 MEDLARS 系统 2—29

第三章 文献情报检索的数据结构和检索技术 3—1
3.1 情报检索中的数据结构 3—1
3.1.1 逻辑结构与物理结构 3—1
3.1.2 线性表 3—5
3.1.3 树 3—8
3.1.4 图 3—13
3.2 查找技术 3—14
3.2.1 顺序查找 3—15
3.2.2 基于索引的方法 3—17
3.2.2.1 二分法查找 3—18
3.2.2.2 分块查找法 3—20
3.2.2.3 索引顺序法 3—23
3.2.2.4 B—树 3—28
3.2.3 基于Hash的查找方法 3—29
3.2.3.1 碰撞问题及其解决 3—30

3.2.3.2 截词检索 3—34
3.2.3.3 Hash法与情报检索 3—36

第四章 检索效果及其改善 4—1
4.1 检索效果及其测量指标 4—1
4.2 影响检索效果的主要因素 4—5
4.2.1 情报提问对情报需求的表达程度 4—6
4.2.2 数据库的选择和比较 4—8
4.2.3 检索途径的选择 4—9
4.2.4 检索词的选择与调节 4—9
4.2.5 检索式的结构 4—11
4.3 提高检索效果的反馈调整方法 4—12
4.3.1 反馈调整在检索过程中的作用 4—12
4.3.2 调节检索策略的若干方法 4—15

第五章 自动标引 5—1
5.1 自动标引和人工标引 5—1
5.2 西文自动标引方案简介 5—3
5.2.1 词频统计原理 5—3
5.2.2 逆文献频率法 5—6
5.2.3 信号——噪音法 5—7
5.2.4 词辨别值法 5—10
5.2.5 词短语的构造 5—16
5.3 自动标引中的词表 5—18

张庆国

第六章　聚类检索　6—1
6.1　问题的提出　6—1
6.2　SMART 系统　6—1
6.2.1　文献的向量表示和匹配度计算　6—2
6.2.2　聚类文件的生成和 SMART 系统的文档结构　6—4
6.2.3　提问式的反馈调整　6—10
6.2.4　动态文献空间　6—14
6.2.5　聚类检索和分类检索的区别　6—15
6.3　倒排检索和聚类检索的结合　6—16
6.3.1　SIRE 系统　6—16
6.3.2　加权的布尔检索　6—20

第七章　检索效果的改善（续）　7—1
7.1　文献—— 语词矩阵的若干推论　7—1
7.1.1　词联接矩阵　7—1
7.1.2　词结合矩阵和改良型文献—— 语词矩阵　7—2
7.2　与词结合矩阵相关的权和提问向量　7—4
7.3　通过结合反馈进行的提问自动修正　7—8
7.4　检索策略的最优化　7—11

第八章　数据检索系统　8—1
8.1　概论　8—1
8.2　数据库管理系统的结构　8—4
8.2.1　信息项的结构　8—4
8.2.2　关系数据库模式　8—8

8.2.3 层次数据库模式 8—13
8.2.4 网络数据库模式 8—18
8.3 查询和查询语言 8—19
8.3.1 分步法 8—21
8.3.2 “菜单”方法 8—22
8.3.3 表查询法 8—23
8.3.4 例举查询 8—24

第九章 事实检索 9—1
9.1 事实检索和自然语言处理 9—2
9.2 自然语言处理的句法分析系统 9—3
9.2.1 自然语言的处理层次 9—3
9.2.2 短语结构语法 9—4
9.2.3 转换语法 9—9
9.2.4 扩充转换网络语法 9—12
9.3 知识的表示 9—18
9.4 目前水平上的事实检索系统 9—23

第十章 情报信息的存贮和输入输出 10—1
10.1 数据标识的代码化 10—1
10.2 数据库的存贮载体 10—1
10.2.1 磁带数据库 10—5
10.2.2 磁盘数据库 10—7
10.2.3 其他存贮设备 10—8
10.3 情报资料的输入手段

10.3.1 键到纸介质方式 10—8
10.3.2 键到磁介质方式 10—9
10.3.3 联机终端输入方式 10—10
10.3.4 全自动字符识别方式 10—11
10.3.4.1 光学字符识别法 10—11
10.3.4.2 光学标记阅读装置 10—14
10.4 情报资料的输出手段 10—15

结 语 10—17

第九章　事实检索

9.1　事实检索和自然语言处理

在文献检索、数据检索和事实检索这三者中，事实检索是最高级也是难度最大的一种。

首先，事实检索涉及到自然语言的处理问题。在文献检索中，文献是以其标识化的形式——标引词来表示的。同时，提问式也是由若干个检索词以简单的逻辑关系组合而成的。我们在前面几章中看到，这种方法大大简化了文献检索工作，然而也带来了很大的检索效果问题。在数据检索中，我们依靠若干属性值来确定一个记录。例如，检索一个个人记录，我们必须指出这个人的职业是工程师，年龄33岁，工龄10年以上等等。这样，严格的数据格式和查询格式也简化了检索工作。事实检索则不同。在事实检索中，用户一般以自然语言提出问题，同时也要求自然语言的回答。只有自然语言能准确地表达用户的原意和回答用户的情报需求。

在知识的表示方法上同样有自然语言处理问题。在文献检索中，文献是通过单个的文献记录来表示的。这种单个的、独立的记录处理起来比较容易；在数据检索中，对于记录的格式甚至对于属性的表示格式都有严格的规定，因此能有效地对它们进行操纵。在事实检索中，知识以更接近其本来形式的形式存贮，不仅要存贮各知识的单元，而且要存贮知识单元之间的句法关系和语义关系，等等。一言以蔽之，我们构造的不是一个文献记录库或数据库，而是一个知识库。因此，从建立知识库，到接受用户提问并运用知识库中的知识对提问作出回答，都是一个自然语言处理问题。

其次，事实检索需要有推理能力。在文献检索系统或数据检索

系统中。我们只能检出系统中原来存在有的文献记录或数据记录；在事实检索中，不仅如此，往往还要检索系统中原来没有，但经过对系统中存贮的其他知识的推理运算而得出的新的知识内容，以回答用户的问题。例如，设系统中存贮了如下的两个知识：

John是一个雇员

所有的雇员都是工会会员

则系统应该能够回答用户的“John是工会会员吗”，这样的问题，虽然“John是工会会员这个知识原来并未存贮在系统中。更进一步地，如果上面的第二个知识换成。

80%的雇员是工会会员

则系统在回答用户的同一个问题时，还要告诉用户这个回答的可信程度，或者根据系统中的其他知识来改善对这个问题的回答。

在上述两种问题中，最重要也是难度最大的是自然语言处理问题。因此，本章重点在于介绍当前处理自然语言的一些方法。这些介绍将是很简略和很初步的，对事实检索原理的介绍也是如此。

自然语言处理工作不仅对于事实检索，而且对于情报检索和其他许多领域都有重要意义。例如：

1．在文献检索中，我们前面讨论了一些自动标引方法，这些标引方法主要是基于词频统计的。实际上，利用自然语言处理中的句法或语义方法也可以实现自动标引工作，并且标引质量要更高，因为它已接近于人工标引的智力活动过程。利用自然语言处理方法还可以为文献自动作文摘。

2．在数据检索中允许用户用自然语言提问，将大大提高用户的方便性。

3．自然语言处理与人工智能有紧密联系。人工智能中的一个

重要内容就是研究人脑对于自然语言的理解和处理机制，然后用人工方法再现。

4。机器翻译工作实际上也是自然语言处理工作。这一问题一旦得决之后，人们的信息交流将在很大程度上克服了语种障碍。

等等

9.2 自然语言处理的句法分析系统

9.2.1 自然语言的处理层次

对自然语言的处理可以在多个层次上进行。这些层次主要有：

1。音位层次。处理语言或音位。研究语音理解系统或语音产生系统。语音学和我们目前的讨论关系不大。

2。词法层次或形态层次。处理个别词的形态和词的词素的识别。例如，词缀的处理，词干的产生等。

3。词汇层次。处理全词。在情报检索中，这包括功能词的删除，词典的操纵，用词类代替单个词等等工作。词汇层次我们在自动标引的有关内容中已有所涉及，本章不再讨论。

4。句法层次。研究如何将句子的成份组织成结构单位。例如名词短语，主——谓——宾结构等等。

5。语义层次。将上下文知识加进由句法分析得来的结构单位中，以使其具有实际的明确的意义。

6。语用层次。利用有关的社会背景知识，最后完成对语句的解释。

我们在本章中主要讨论句法、语义和语用这三个层次。在实际的自然语言处理过程中，这三个层次是紧密联系，互相影响的。

对于句法分析，人们已经提出了很多方法，其中比较有名的是

(下三种：短语结构语法、转换语法、转换网络语法。

2.2 短语结构语法

(Phrase Structure Grammar，短语结构语法对于句子生成和句子分析都是有效的。

考虑以下的重写规则：

$$S \rightarrow A + B \qquad (2.1)$$

这个重写规则表示“变量S可以被重写为A后跟B”。在重写规则中，大写字母被称为非终结符号，它可以再次被重写，只要它出现在重写规则的左边。

重写规则

$$S \rightarrow John + ran \qquad (2.2)$$

表示符号S可被重写成“John”和“ran”，由小写字母组成的符号“John”和“ran”称为终结符号，它们不再能被重写。因此，表达式(2.2)给出了一个句子：“John ran”。

在原理上，可以为每个待生成的句子使用一个重写规则，如

$$S \rightarrow John+ran \mid sally+jumped \mid \cdots\cdots \qquad (2.3)$$

其中竖线表示“OR”，这样便成为一种非常庞杂的语法。

考虑以下八个句子：

1. the man hit the ball
2. the man hit the man
3. the ball hit the ball
4. the ball hit the man
5. the man took the ball
6. the man took the man

7. the ball took the ball

8. the ball book the man

忽略其中的语义不合理现象。这八个句子可由八个“S→…”形式的重写规则产生。如果我们再增加一个词汇“rock”，则又可以产生十个句子：

9. the man hit the rock

10. the ball hit the rock

11. the rock hit the man

12. the rock hit the ball

13. the rock hit the rock

14. the man took the rock

15. the ball took the rock

16. the rock took the man

17. the rock took the ball

18. the rock took the rock

因此又需要十个重写规则。实际上，我们可以用以下语法来生成前述的1—8这八个句子：

$$
\begin{aligned}
&S \rightarrow NP + VP \\
&NP \rightarrow T + N \\
&T \rightarrow the \\
&N \rightarrow man \mid ball \qquad (2.4) \\
&VP \rightarrow V + NP \\
&V \rightarrow hit \mid took
\end{aligned}
$$

其中NP表示名词短语，VP表示动词短语，T表示冠词，V表示动词，N表示名词。这样，需要加入一个新的词“rock”时，我

们只要增加一个重写规则

N→rock

便产生了前述的后十个句子。

重写规则

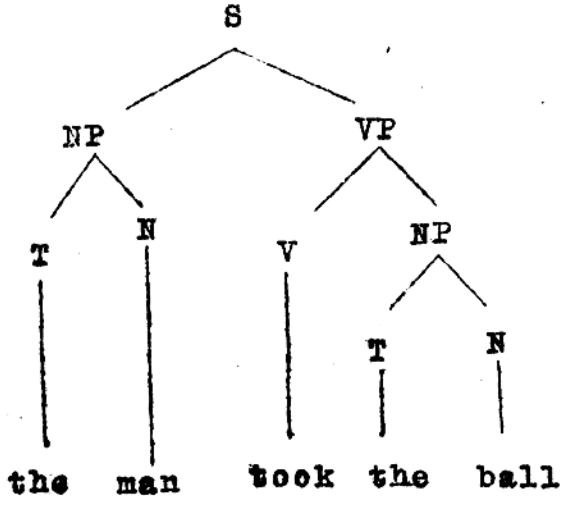
S→A＋B

可表示成

表达式(2. 4)中的每一个重写规则也都可以类似地表示。于是句子

the man took the ball

可以表示成



重写规则(2. 4)解决了一种最基本的句型。它可以产生相当多的句子。然而，在丰富的自然语言中，有很多句型不能为这少量的重写规则所解决。对于这些句型，要增加重写规则。例如，考虑

句子

19 John phoned Mary

20 John phoned up Mary

我们用以下重写规则可以产生这两个句子

S→NP + VP

NP→John|Mary

VP→V + NP (9.5)

V→phoned| phoned up

但是，要产生句子

21 John phoned Mary up

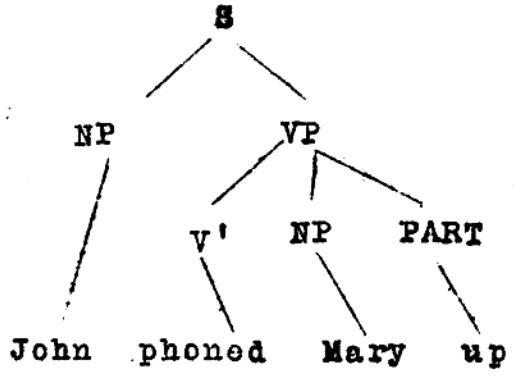
则需要在上述重写规则的基础上再增加重写规则

VP→V'+NP + PART

V'→phoned (9.6)

PART→up

即



再例如，我们必须考虑主语名词的单复数对其所用的动词形状的影响，即主谓一致问题。考虑下列语法

$$
\begin{aligned}
&S \to NP + VP \\
&NP \to T + N \\
&T \to the \\
&N \to man \mid men \mid ball \mid balls \\
&VP \to V + NP \\
&V \to have \mid has
\end{aligned}
\qquad (2.7)
$$

这组重写规则虽然能产生如“the man has the ball”这样的正确句子，但是也可能产生象“the man have the ball”这样的错误句子。为解决这一问题，我们以NP_s表示单数名词主语，NP_p表示复数名词主语，类似地还给定符号VP_s、VP_p、V_s和V_p，于是产生新的语法

$$
\begin{aligned}
&S \to NP_s + VP_s \mid NP_p + VP_p \\
&NP_s \to T + N_s \\
&T \to the \\
&N_s \to man \mid ball \\
&NP_p \to T + N_p \\
&N_p \to men \mid balls \\
&VP_s \to V_s + NP_s \mid V_s + NP_p \\
&V_s \to has \\
&VP_p \to V_p + NP_s \mid V_p + NP_p \\
&V_p \to have
\end{aligned}
$$

显然，为处理丰富的自然语言现象，我们需要不断地增加重写规则。但是，一条新的重写规则可以解决相当数量的自然语言问题，因此，重写规则数量的增长速度会越来越慢，最终达到以有限的重写规则解决无限的自然语言句法处理问题的程度。

9.2.3 转换语法（Transformational Grammars）

转换语法的基本特征是引进如下形式的所谓“上下文敏感”重写规则：

$$wAx \to wyx \qquad (9.9)$$

其中A是语法中的非终结符号，y是一终结符号或非终结符号的字符串。对w和x没有限制。重写规则（9.9）表示，当变量A出现在上下文的w和x之间时，它可以由字符串y代替。

将上下文敏感重写规则

$$\text{phoned} + \text{up} + \text{NP} \to \text{phoned} + \text{NP} + \text{up} \qquad (9.10)$$

加进重写规则（9.5），可以产生句子21，过程为：

$$\begin{aligned} S &\to NP + VP \\ &\to John + VP \\ &\to John + V + NP \\ &\to John + phoned + up + NP \qquad (9.11) \\ &\to John + phoned + NP + up \\ &\to John + phoned + Mary + up \end{aligned}$$

对于主谓一致的情况也可以处理。引进符号sing和pl，令其分别代表单数和复数。给出重写规则

$$\begin{aligned} &N \to man \mid ball \\ &V \to have \\ &sing + have \to has \qquad (9.12) \\ &pl + have \to have \\ &man + sing \to man \\ &man + pl \to men \\ &ball + sing \to ball \end{aligned}$$

ball+pl→balls

其中的下划线表示是非终结符号。即该词在词形上还可以（并需要）变化（被重写）。

我们再通过一些例子来看转换语法的作用：

22 Chomsky proved the theorem

23 the theorem was prved by Chomsky

24 Chomsky did not prove the theorem

25 did Chomsky prove the theorem?

26 Was the theorem proved by Chomsky?

27 the theorem was not proved by Chamsky

28 did not Chomsky prove the therem?

29 was not the theorem proved by Chomsky?

显然，这些句子中的后七个句子都可以由第一个句子（句子22）转换而来。这些转换包括：主动→被动，肯定→否定，陈述→疑问。下表列出这八个句子的特性。

主动	肯定	陈述	句子号
√	√	√	22
×	√	√	23
√	×	√	24
√	√	×	25
×	√	×	26
×	×	√	27
√	×	×	28
×	×	×	29

表9.1 八个句子的特性

设句子22可以由以下规则产生

$$S \to NP_1 + V + ed + NP_2$$

其中NP_1和NP_2表示特定名词短语的实际值。显然，我们很容易给出上下文敏感语法的句子转换规则：

主动→被动：

$$NP_1+V+ed+NP_2 \to NP_2+was+V+ed+by+NP_1$$

肯定→否定：

$$NP_1+V+ed+NP_2 \to NP_1+did\ not+V+NP_2$$

陈述→疑问

$$NP_1+V+ed+NP_2 \to did+NP_1+V+NP_2$$

其他的转换规则如被动→疑问、否定→疑问等等，也都可以很容易地产生。

在短语结构语法中，语言的分析或识别是直接的，因为它由一个最初的句子符号S开始，反复运用重写规则，直到所有的终结符号都已产生，同时也不再有非终结符号。在转换语法中，识别过程要复杂些。转换语法实际上可分为两部分：首先是语法的基本成份，产生句子的所谓“深层结构”。深层结构表示这个输入语句的本质句法和语义；其次是语法的“转换成份”，它在基本成份的输出基础上工作，并产生句子的所谓“表层结构”。表层结构反映句子实际的语音表示。转换语法的句子生成过程如下图所示：