战略决策文本的语义分析研究

周 生[1,2] 胡晓峰[1] 罗 批[1] 李志强[1]
(国防大学信息作战与指挥训练教研部 北京 100091)[1] (解放军炮兵学院 合肥 230031)[2]

摘 要 针对采用从战略决策提取关键字检索虚拟新闻视频准确率不高的问题,分析了战略决策文本的特点、构成模式,给出了语义分析过程中构建字典库的原则、自动分词的方法、成分分析的方法和语义树构建的方法,以及与虚拟新闻视频匹配的方法,并进行了实验。结果表明,借助语义树的方法对战略决策文本进行语义分析后能够提高虚拟新闻视频检索的准确率。

关键词 战略决策文本,虚拟新闻,语义分析,战略对抗演习
中图法分类号 TP319.9 **文献标识码** A

Research on Semantics Analysis of Strategic Decision Text

ZHOU Sheng[1,2] HU Xiao-feng[1] LUO Pi[1] LI Zhi-qiang[1]
(Department of Information Operation & Command Training, NDU of PLA, Beijing 100091, China)[1]
(Artillery Academy of PLA, Hefei 230031, China)[2]

Abstract Aiming at the low precision rate of video retrieval for virtual television news caused by extracting keywords from strategic decision text, the features and construction modes of strategic decision text were analyzed. The principle of dictionary construction, automatic Chinese words separation, method of function analysis, semantic tree construction and matching approaches between virtual television news video were discussed in detail. Experiments were made on this research. The results show that the precision rate of virtual television news video retrieval can be increased by the employment of semantic tree method.

Keywords Strategic decision text, Virtual television news, Semantic analysis, Strategic seminar gaming

1 引言

战略对抗演习作为对国际社会焦点与热点问题的一种模拟实践活动[1],是通过模拟的方法和技术手段建立宏观决策的虚拟国家安全环境、虚拟信息环境以及相关的研讨与模拟环境进行的。战略决策文本是演习中参演人员针对演习想定和推演过程中出现的危机事件制定的各种措施文本,其内容涵盖政治、经济、军事、外交等各个方面,并且通过模拟环境尤其是虚拟新闻[2-4]的方式表现出来。

目前虚拟新闻系统在用模拟的电视新闻表现战略决策时是采用提取其中关键字并生成视频的方法[5]进行的。用关键字或关键词汇来代表、理解自然语言,在很多应用中是常见的做法,例如交通路网实时路况信息的研究[14]。但仅用几个关键字很难全面表达自然语言的全部语义,所以虚拟新闻从战略决策文本中提取关键字的方法影响了虚拟新闻表现的准确度。

本文针对以上问题提出一种语义树的方法对战略决策文本进行语义分析,以提高虚拟新闻表现战略决策的准确度。

2 战略决策的特点

战略决策是扮演各种不同角色的人员在演习中制定的措施,其表现形式就是文本。用虚拟新闻表现的战略决策与新闻报道非常类似。为了对其进行分析,下面从2010年3月16日的《参考消息》中选取几条与战略决策很接近的新闻报道进行说明。

(1)印度军方称,该国自主研发的拦截导弹在今天进行的导弹拦截试验中试射失败。(法新社印度巴内什瓦尔3月15日电)

(2)南美油田协议扩大了中国对全球的能源需求。(美国《纽约时报》网站3月14日报道)

(3)以色列总理内塔尼亚胡今天明确表示,东耶路撒冷的定居点建设将继续进行,而以色列和美国就东耶路撒冷定居点项目问题的外交争吵日益激烈。(法新社耶路撒冷3月15日电)

(4)泰国"红衫军"昨天在曼谷的大游行还是一片嘉年华。但15万"红衫军"今早却包围军方基地,逼阿披实中午前解散国会。(新加波《联合早报》网站3月15日报道)

(5)包括美国在内的8个国家今天开始商讨一项跨太平洋贸易协议。(路透社堪培拉3月15日电)

(6)希腊财政部一位官员说,希腊将在明天和后天举行的欧盟财长会议上提交财政紧缩措施的最后结果。(美联社雅

到稿日期:2010-03-19 返修日期:2010-06-16 本文受863项目(2006AA01Z337),973项目(2006CB303106)资助。
周 生(1978-),男,博士生,主要研究方向为战争模拟、可视化表现,E-mail:unbend@126.com;胡晓峰(1957-),男,教授,博士生导师,主要研究方向为战争模拟系统与环境、军事运筹、军事信息系统工程等;罗 批(1974-),男,副教授,主要研究方向为战争复杂性、战争模拟、遗传算法等;李志强(1973-),男,副教授,主要研究方向为战争模拟、可视化表现、基础设施建模。

典3月14日电)

(7)中国已经取代美国成为世界最大的汽车市场。(日本《富士产经商报》3月15日报道)

以上7条与战略决策类似的新闻报道可以分为3类:第一类为由单个句子构成的简单句,例如(2),(5),(7);第二类为由几个分句构成的复句,例如(1),(3),(6);第三类为两个以上的简单句构成,这里称为复杂句(汉语语法里没有这种名称,为了表述方面,姑且称之),例如(4)。

从以上随机挑选的几条决策可以发现,战略决策文本不管属于哪一类,都具有以下特点:

① 语句的个数一般为1;

② 语句本身是自由的、非结构化的;

③ 语句中包含的时政领域词汇较多;

④ 语句比较严谨,没有任何歧义;

⑤ 语句都是陈述句;

⑥ 语句一般都是正常的一般语序,没有文学作品中倒装等语序;

⑦ 决策文本包含多个语义,且存在一定的逻辑关系。

3 战略决策文本构成模式

研究战略决策文本的构成模式是对其进行语义分析的前提和基础。战略决策文本除了自身的特点之外,由汉语中的一个句子或由几个句子组成。所以,研究战略决策文本的构成模式主要就是对其中包含的句子进行分析。因为句子是"表示相对完整意义的语言片段"[6]。

将战略决策文本划分为对句子的分析后,给定一个战略决策文本 ST,可以表示如下:

$$ST=\{S_1,S_2,\cdots,S_i,\cdots,S_n\} \tag{1}$$

式中,S_i 表示句子,n 是句子的个数。

从计算机处理的角度来看,S_i 是一个顺序排列的符号串,其中的符号可以是汉字字符、英文字符、标点符号和其他非汉字字符[7]。由于词是汉语中最小的、能独立运行的语言单位[8],因此本文不从符号的角度来研究和分析句子,而是将句子看成是一个顺序排列的词串,可以表示如下:

$$S=\{W_1,W_2,\cdots,W_i,\cdots,W_n\} \tag{2}$$

词串在形式上是没有关联性的,但是词串与它相邻的词串之间存在或强或弱的联系,即它们在语义上是相关的。它们之间的语义相关性就构成了汉语语句的结构。

一句完整的汉语语句结构主要由主语、谓语、宾语、定语、状语和补语构成。

句子=(定)主//[状]谓〈补〉+(定)宾

主语是谓语陈述的对象,指明说的是"什么人"或"什么事物"。谓语是陈述主语、说明主语的,即说明主语"是什么"或"怎么样"。宾语在动词后面,表示动作、行为涉及的人或事物,回答"谁"或"什么"一类问题。定语是名词前面的连带成分,用来修饰名词,表示人或事物性质、状态、数量、所属等。状语是动词或形容词前面的连带成分,用来修饰、限制动词或形容词,表示动作的状态、方式、时间、处所或程度等。补语是动词或形容词后面的连带成分,一般用来补充说明动作、行为的情况、结果、程度、趋向、时间、处所、数量、性状等。用结构成分对句子表示如下:

$$S=[A]\langle H\rangle[M]\langle P\rangle[C][A]\langle O\rangle \tag{3}$$

式中,[]表示可选项,〈〉表示必选项,H 是主语,P 是谓语,O 是宾语,M 是状语,A 是定语,C 是补语,它们又由词组成,表示如下:

$$H=\{W_1,W_2,\cdots,W_i,\cdots,W_n\}$$

$$P=\{W_1,W_2,\cdots,W_i,\cdots,W_n\}$$

$$M=\{W_1,W_2,\cdots,W_i,\cdots,W_n\}$$

$$O=\{W_1,W_2,\cdots,W_i,\cdots,W_n\}$$

$$C=\{W_1,W_2,\cdots,W_i,\cdots,W_n\}$$

$$A=\{W_1,W_2,\cdots,W_i,\cdots,W_n\}$$

而词 W 可以有各种不同的词性。词性与该词在句子中的结构有很大的关系。用词性对词 W 表示如下:

$$W=\langle n|a|v|p|c|ad|\cdots\rangle \tag{4}$$

式中,n 是名词、a 是形容词、v 是动词、p 是代词、c 是连词、ad 是副词。还有一些未列出的词性。

但是并不是所有的汉语语句都是这样的完整结构。我们对演习中产生的所有战略决策进行了统计,发现所有的这些决策语句都可以用式(3)表示,即战略决策文本的构成模式是具有主谓宾结构的语句。

4 语义分析

对战略决策文本进行语义分析,第一步需要对其中的语句进行分词。而分词基础是需要针对战略决策文本的特点建立一个专用的词典,在词典的基础上进行分词。分词后才能分析句子的成分,得到句子的结构,最后才能建立战略决策文本的语义树。

4.1 词典库构建

在建立词典库之前,我们对大量的战略决策文本中的词语进行了认真分析,发现了几个不同于其他词典库(语料库)构建时遇到的现象。一是,词语在不同句子中的词性变化不大;二是,组合词语比较常见,比如战略轰炸机、商务部门等;三是,时政领域表示立场、态度的动词出现概率较大,比如发表、宣布、表示、决定等;四是,具体的人名、地名出现得较少。

因此,针对这4个不同的现象,在建立词典库时进行以下4点不用于一般词典库构建时的特殊处理:

第一,收录词语的词性时,只考虑其在战略决策文本中的词性,不追求一般词典库中词语词性的全面性,所以在设计词典库(为了表述的方便,取名为 com_dict)时,一个词语最多考虑4个词性的变化。

第二,对于由组合词语构成的一些特定领域词汇,比如"外交部部长"、"战略轰炸机"等这些词语,建立一个专门的特定领域专有名词词典库(取名为 noun_dict),但对于"部长"、"战略"、"轰炸机"仍然收录进 com_dict 中。

第三,由于地名、人名出现的机会少,如果将它们放进 com_dict 中,要切分出它们可能要扫描整个 com_dict,效率较低,因此将地名、人名也放入一个单独字典库中(取名 np_dict)。

第四,时政领域中表明立场、态度的动词在战略决策文本的语句中非常重要,几乎全部都是整个语句的谓语动词,所以也为它们建立一个单独的字典库(取名 verb_dict)。

在以上4点处理的基础上,设计字典库的结构如表1所列。

说明:

1)多数词语的词性数量为1或2,少部分超过2;

2)长度指的是词语的字符个数,所有的字典库都按照长

度的降序排列；

3)noun_dict 和 np_dict 中的词语只有一个词性，分别记为专有和人、地名；

4)verb_dict 中虽然记录的都是动词，但是必须考虑到有些动词其他的词性，比如“决定”在“同意这个决定”和“决定执行这个计划”中分别为名词和动词。

表 1　字典库的结构

名称	字段名	类型	长度
词汇	word	Nvarchar	50
第一词性	First_func	Nvarchar	10
第二词性	Second_func	Nvarchar	10
第三词性	Third_func	Nvarchar	10
第四词性	Fourth_func	Nvarchar	10
长度	Length	int	4

4.2　自动分词

汉语自动分词是对汉语文本进行自动分析的第一个步骤。

自动分词的困难主要有分词规范、切分歧义等问题[9]。现有的分词算法分为 3 大类：基于字符串匹配的分词方法、基于理解的分词方法、基于统计的分词方法。其中最成熟的、应用范围最广的是基于字符串的分词方法[10]。基于字符串匹配的分词方法又叫做机械分词方法，它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功(识别出一个词)。按照扫描方向的不同，串匹配分词方法可以分为正向匹配和逆向匹配；按照不同长度优先匹配的情况，可以分为最大(最长)匹配和最小(最短)匹配。

一般说来，逆向匹配的切分精度略高于正向匹配，遇到的歧义现象也较少。统计结果表明，单纯使用正向最大匹配的错误率为 1/169，单纯使用逆向最大匹配的错误率为 1/245。

经过对战略决策文本特点的分析和 4 个字典库 com_dict，noun_dict，np_dict，verb_dict 的构建，本文采取最长优先、逆向匹配的机械分词方法，具体的算法如下。

Step1　根据标点符号，将战略决策文本分成 n 个句子；

Step2　按照顺序取其中的第 i 个句子；

Step3　根据句子中的逗号，将第 i 个句子划分成 m 个部分；

Step4　按照顺序取第 j 个部分，计算其长度 len；

Step5　根据 len 的大小，分别利用 noun_dict，np_dict，verb_dict 对其进行分词；

Step6　计算未被切分的余下部分最长的长度 k，根据 k 的大小，利用 com_dict 对余下部分进行切分；

Step7　若有未被识别的词语(未登录词语)，将其存入数组 a 中；

Step8　取第 $j+1$ 个部分，重复 Step4－Step7；

Step9　取第 $i+1$ 个句子，重复 Step3－Step8；

Step10 对 a 中未被识别的词语，进行人工干预，确定其词性，并存入相应的字典库中；

Step11 算法结束。

说明：

1)长度 len 是切分时相应的字典库的入口点，这样避免了从头到尾扫描整个字典库，这也是字典库为什么按照长度降序排列的原因；

2)自动分词时，对未登录词进行人工干预并存入相应的字典库中。对于字典库采取的是边使用边增加的原则，因为没有任何一个字典库能够收录所有的词汇；

3)自动分词时要按照表 len 中的第一到第四记录每个词语的可能词性，这是后续词语分析的基础。

4.3　词语成分分析

4.3.1　词性标注

分词只是语义分析中的第一步。分词后的结果是每个词语可能有多个词性，也可能有多个语义。但是在具体的战略决策文本中，每个词的词性和语义都是确定的。怎样确定一个词在语句中的词性，称为词性标注。词性标注是对战略决策文本进行语义分析的前提。一般来说，词性标注有两个基本的原则：

第一，观察词语上下文的词性。在汉语中很多词语与词语的搭配是固定的，而有些搭配基本上是不可能的。针对这条原则，我们参考现代汉语语法并对大量的战略决策文本进行分析，列举并建立了出现在其中的 20 种词语的搭配模式，比如 $v+v$，$v+a$，$a+n$，$ad+v$，$ad+a$，$v+n$ 等。

第二，词语本身的信息，虽然一个词可能有多个词性，但并不是每个词性的使用频率都相同。针对这条原则，我们在字典库构建时统计了每个词的词性使用情况，将使用频率最高的作为该词的第一词性，次之的作为第二词性，依此类推。

依据以上两条原则以及建立的词语搭配模式和词典库构建的先期工作，在自动分词的基础上，对战略决策文本进行词性标注。步骤如下：

第一步，扫描战略决策文本 ST，找出其中未确定词性的词语；

第二步，按照 20 种词语的搭配模式，对未进行词性标注的词语进行词性标注；

第三步，若第二步中仍然有少数词语未能确定词性，则按第一词性对这些词语进行词性标注。

通过以上 3 个步骤，就可以完成对战略决策文本的词性标注。

4.3.2　成分确定的方法

怎样对汉语语句进行成分分析，即确定语句的结构，是自然语言理解领域的一个难题。因为汉语的语句结构非常灵活，难以用统一的方法来分析。经过仔细分析大量的战略决策文本，本文提出一种新的确定战略决策文本中汉语语句结构的方法，称为以动词为中心的最近邻居确定法。

以动词为中心的最近邻居确定法指的是在战略决策文本中，动词是整个句子的中心，尤其是从左至右的第一个动词，就是整个语句的谓语动词，与谓语动词左侧最近的名词短语邻居即为主语，与名词左侧最近的形容词邻居即为定语，与动词左侧最近的副词邻居即为状语，与动词右侧最近的副词短语、介宾短语、动词短语、孤立形容词邻居(不修饰名词)为补语，除此之外的名词即为宾语。对该方法的补充说明如下。

① 主语在许多情况下不只是一个孤立的名词，还可是复合名词短语，比如“D 方国防部”、“F 方外交部长”等；

② 宾语有可能是动词的宾语、介词的宾语，这需要根据其前置词来确定；

③ 确定成分时以逗号作为标志将战略决策中的语句划分为几个部分，对每个部分按照本文提出的方法进行成分确定，但是主语和谓语只出现在第一部分中；

④ 若战略决策文本由多个句子组成，则对每个句子按照

以上的方法来确定其成分；

⑤ 对由多个句子组成的战略决策文本确定成分时，若当前句子无主语，则取前一个句子的主语作为该句的主语。

根据以上的方法和补充说明，对战略决策文本的成分确定的算法如下。

Step1 获取分词后的战略决策文本；

$ST=\{W_1W_2\cdots,W_i\cdots,\cdots W_n,\cdots\}$；

Step2 对 ST 中词性相同的连续部分进行合并，即将它们合并为一个复合词语；

Step3 将 ST 按照句号进行分割；

$ST=\{ST_1ST_2\cdots ST_i\cdots ST_n\}$；

Step4 对 $ST_i(i=1\cdots n)$按照逗号进行分割；

$ST_i=\{S_1S_2\cdots S_j\cdots S_m\}$；

Step5 对 $S_j(j=1\cdots m)$按照以动词为中心的最近邻居确定法确定其成分；

Step6 对确定成分后的 ST_i 进行扫描，若缺少主语，则取前一个句子的主语放置句首作为其主语；

Step7 取下一个 ST_i 按照同样的方法来确定其成分，直到 ST 中所有的句子成分都确定完毕；

Step8 算法结束。

4.4 语义树构建

虽然确定了战略决策文本句子的各种成分，但是这些成分是按照从左至右的顺序出现的，是一种线性描述。为了便于对战略决策进行进一步的分析以及为其检索到更准确的视频来生成虚拟新闻，本文提出一种三叉语义树的方法来表示战略决策文本的句子结构。

定义 1(三叉语义树) 表示战略决策文本中的一个句子，树中每个节点最多拥有 3 个子节点，因此称为三叉语义树。

定义 2(语义森林) 当战略决策文本中包含的句子不止一个时，每个句子都用一棵三叉语义树来表示。由这些树构成的森林，称为战略决策文本的语义森林。

定义 3(节点) 表示战略决策文本语句中的一个词语，这个词语是已经合并基础上的复合词语。每个节点最多有 3 个子节点，分别称为左子节点、中子节点和右子节点。

定义 4(左子节点) 修饰某个节点词语的部分，比如形容词性词语、副词性词语等称为该节点的左子节点。

定义 5(中子节点) 对某个词语进行补充说明的部分，主要指补语，称为该节点的中子节点。

定义 6(右子节点) 表示动作施加的对象，即宾语，称为该节点的右子节点。

定义 7(根节点) 整个句子的主语是三叉语义树的根节点。根节点的右子节点是整个句子的谓语。

根据以上的定义，一个句子就可以用图 1 所示的树形结构表示。

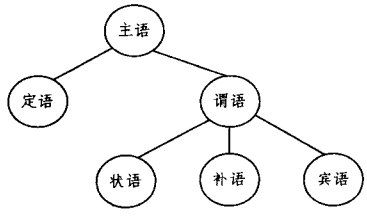

图 1 三叉树句子结构

对以上语义树的补充说明如下：

① 定语、状语、补语、宾语也有可能本身就是一棵子树；

② 定语只能作为主语和宾语的左子节点；

③ 若补语位于谓语动词之后，则补语作为该动词的中子节点；若补语位于宾语名词之后，则作为该名词的中子节点。

根据以上的方法和补充说明，对战略决策文本的语义树构建算法如下。

Step1 获取确定成分后的战略决策文本；

$ST=\{W_1W_2\cdots,W_i\cdots,\cdots W_n,\cdots\}$；

Step2 将 ST 按照句号进行分割；

$ST=\{ST_1ST_2\cdots ST_i\cdots ST_n\}$；

Step3 取 $ST_i(i=1\cdots n)$构建语义树；

Step4 ST_i 的主语作为第 i 棵树的根节点 Root，谓语作为 Root 的右子节点 RightNode，主语的定语作为 Root 的左子节点 LeftNode，宾语作为谓语 RightNode 的右子节点，状语作为谓语 RightNode 的左子节点，补语作为谓语 RightNode 的中子节点；

Step5 对下一个 ST_i 按照同样的方法来构建语义树；

Step6 得到战略决策文本的语义森林；

Step7 算法结束。

5 与虚拟新闻视频匹配方法研究

5.1 视频语义描述

对战略决策文本进行语义分析的根本目的是检索到更匹配、更准确的视频在虚拟新闻系统中进行表现。因为分析了战略决策文本的语义，所以对于视频不能再利用传统的几个关键字标注的方法进行描述，需要进行语义描述。

视频数据结构复杂，内容丰富，但从内容的构成上，视频可以看作是在时间轴上进行动态演化的物体组成，物体是视频的基本组成部分[11]。

因此，本文在综合和比较其他研究者有关视频语义描述的基础上[12,13]，结合本文的研究背景，给出了一种新的视频语义模型。

定义 8(视频语义) 指的是其所包含的对象(语义物体)的行为属性及其时间关系、空间位置关系和逻辑关系的描述。可以用下面一个五元组表示：

VideoSemantics={SO,AP,TR,SR,LR}；

SO(Semantic Objcets)代表语义物体；

BP(Behavior Properties)代表行为属性；

TR(Temporal Relations)代表时间关系；

SR(Space Relations)代表空间关系；

LR(Logic Relations)代表逻辑关系。

SO,BP,TR,SR,LR 分别描述如下。

$SO=\{Obj_1,Obj_2,Obj_3,\cdots,Obj_n\}$；一段视频常常包含若干个语义物体。

BP={

$\{\langle Obj_1,Act_1\rangle,\langle Obj_1,Act_2\rangle,\cdots,\langle Obj_1,Act_k\rangle\}$,

……

$\{\langle Obj_n,Act_1\rangle,\langle Obj_n,Act_2\rangle,\cdots,\langle Obj_n,Act_k\rangle\}$

}；行为属性与具体的语义物体有关，可能有多个。

$TR=\{\langle Obj_i,Obj_j,Rt_k,Act_i,Act_j\rangle\cdots\}$；时间关系发生在两个不同的语义物体不同的动作之间。

$SR=\{\langle Obj_i,Obj_j,Rs_k,Act_i,Act_j\rangle\cdots\}$；与时间关系类似，空间时间关系也发生在两个不同的语义物体不同的动作之间。

$LR=\{\langle Obj_i,Obj_j,Rl_k\rangle\cdots\}$；逻辑关系是一种静态的关

系，与时间和动作无关。

根据以上的语义模型，分别对 SO，BP，TR，SR，LR 建立标准的字典库，字典库中的许多词语直接就来自于我们在战略决策文本语义分析时构建的字典库。根据以上语义模型，就可以得到视频的语义关系，如图 2 所示。

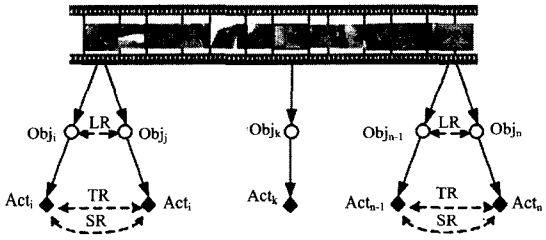

图 2　视频语义关系图

5.2　匹配方法

战略决策文本是演习中应对危机事件的措施，是一种静态描述。而虚拟新闻视频是对演习中危机事件的一种动态描述，具有时空结构。虽然虚拟新闻视频具有时空结构，但是其中最核心、最根本的信息是语义物体及其行为属性。战略决策文本中最核心、最关键的信息是由主语、谓语、宾语 3 部分组成的，而状语、定语、补语等主要起补充说明的作用。因此，基本思想是将战略决策中的主语、谓语和宾语与视频中的语义物体、行为属性进行匹配。在进行具体的匹配计算说明之前，先引入几个概念。

定义 9(语义物体集)　指的是标注单个视频的语义物体集合，用 So 表示，$|So|$ 表示集合中元素的数目。

定义 10(行为属性集)　指的是标注单个视频的行为属性集合，用 Be 表示，$|Be|$ 表示集合中元素的数目。

定义 11(主语和宾语集)　指的是战略决策文本中主语和宾语集合，用 No 表示，$|No|$ 表示集合中元素的数目。

定义 12(谓语和动词集)　指的是战略决策文本中谓语和动词集合，用 Ve 表示，$|Ve|$ 表示集合中元素的数目。

$|So\cap No|$ 表示语义物体集与主语和宾语集的公共元素数目，$|So\cup No|$ 表示两个集合元素总数。

$|Be\cap Ve|$ 表示行为属性集与谓语和动词集的公共元素数目，$|Be\cup Ve|$ 表示两个集合元素总数。

匹配度 Sim 的计算公式为：

$$\mathrm{Sim}=\frac{|So\cap No|+|Be\cap Ve|}{|So\cup No|+|Be\cup Ve|} \tag{5}$$

根据以上方法，战略决策文本与虚拟新闻视频匹配的算法如下。

Step1　提取战略决策文本中的主语、宾语，消除其中重复的元素，加入到集合 No 中；

Step2　提取战略决策文本中的谓语、动词，消除其中重复的元素，加入到集合 Ve 中；

Step3　$i=1$，指向视频库中第 i 个视频；

Step4　取第 i 个视频中的语义物体，消除重复元素，加入到集合 So 中；

Step5　取第 i 个视频中的行为属性，消除重复元素，加入到集合 Be 中；

Step6　利用式(5)计算其匹配度 Sim；

Step7　i 指向下一个视频，重复 Step4－Step6；

Step8　取 Sim 值最大的作为候选视频在虚拟新闻系统中播出；

Step9　算法结束。

6　实验结果分析

采用语义树的方法对战略决策文本进行分析，然后检索虚拟新闻视频在实际的战略对抗演习中进行实验，结果如表 2 所列。

表 2　实验结果

类　型	比例	视频检索准确率
简单句	9.4%	98.9%
复句	64.2%	82%
复杂句	26.4%	65%

从表 2 的实验结果来看，简单句语义分析得比较好，视频检索率也较高。复句在整个战略决策文本中所占的比例很高，视频检索的准确率为 82%。汉语中由于有些复句的结构比较复杂，宾语和补语非常容易混淆。复杂句的检索准确率更低，这是因为复杂句由多个句子组成，对其进行语义分析更为困难。

结束语　战略决策文本是汉语中一种普通的文本，对其进行语义分析也属于自然语言理解的范畴。由于其特殊的应用背景，决定着战略决策文本有着自身的特点。本文针对这些特点，分析了战略决策文本的构成模式，提出了用三叉树的方法对其进行语义表示。根本目的是为了在虚拟新闻系统中为其检索到更准确的视频，为构建一个逼真的虚拟信息环境而服务。从实验结果来看，平均检索准确率为 81.97%，比起以前的效果有所增强。怎样在此基础上更加准确地分析战略决策文本的语义，提高检索准确率，需要进一步研究。

参 考 文 献

[1]　胡晓峰，司光亚，吴琳，等. 战争复杂系统仿真分析与实验[M]. 北京：国防大学出版社，2008

[2]　陈芳莉，胡晓峰，吴琳，等. 虚拟新闻模拟系统的研究与设计[J]. 计算机仿真，2007，24(8)：5-7

[3]　周生，胡晓峰，罗批. 战略对抗演习中的态势表现方法研究[J]. 装备指挥技术学院学报，2009，20(3)：96-99

[4]　董献洲，胡晓峰，吴琳，等. 虚拟新闻的表达与生成及其系统设计与实现[J]. 系统仿真学报，2006，18(12)：3634-3636

[5]　周生，胡晓峰，罗批. 虚拟新闻系统的自动化设计研究[J]. 计算机工程与应用，2008，44(12)：20-23

[6]　房玉清. 实用汉语语法[M]. 北京：北京语言出版社，1992

[7]　韦向峰，张全，熊亮. 汉语语句形式结构到语义结构的理解自明度[J]. 计算机科学，2006，33(12)：142-144

[8]　苗多谦，卫志华. 中文文本信息处理的原理与应用[M]. 北京：清华大学出版社，2007

[9]　卢亮，张博文. 搜索引擎原理、实践和应用[M]. 北京：电子工业出版社，2005

[10]　孙宾. 现代汉语文本的词语切分技术[R]. 北京：北京大学计算语言学研究所，2003

[11]　杨士强，孙立峰，崔鹏，等. 视频的语义挖掘[J]. 中国计算机学会通讯，2009，5(7)：30-35

[12]　余卫宇，余英林. 视频语义信息的研究[J]. 计算机工程与应用，2004，40(6)：27-29

[13]　王煜，周立柱，邢春晓. 视频语义模型及评价准则[J]. 计算机学报，2007，30(3)：337-351

[14]　陈传彬，陆峰，励惠国，等. 自然语言表达实时路况信息的路网匹配融合技术[J]. 中国图像图形学报，2009，14(8)：1669-1676