浙江科技学院学报,第 24 卷第 1 期,2012 年 2 月
Journal of Zhejiang University of Science and Technology
Vol. 24 No. 1, Feb. 2012

DOI: 10.3969/j.issn.1671-8798.2012.01.007

中文信息处理在动态几何软件领域的应用研究

陈晓霞[1,2]

(1.浙江科技学院 理学院,杭州 310023;2.国家数字化学习工程技术研究中心,武汉 430079)

摘　要:阐述了中文信息处理技术在几何作图等动态几何领域内的应用现状。鉴于几何语言其本身相较于一般自然语言的特点,介绍了利用中文分词技术及其他自然语言处理技术来实现基于自然语言输入的动态几何作图的方法,主要包括 GMMM 算法和基于"分词词典"的分词方法及建立同义词库、使用语模匹配和语模词典等方法来形式化和规则化几何命题等内容,并提出在该领域的研究展望。

关键词:中文信息处理;几何作图;动态几何;分词

中图分类号:TP301.6　　**文献标志码**:A　　**文章编号**:1671-8798(2012)01-0030-05

Application research of Chinese information processing in dynamic geometry software

CHEN Xiao-xia[1,2]

(1. School of Sciences, Zhejiang University of Science and Technology, Hangzhou 310023, China;
2. National Engineering Research Centre For E-learning, Wuhan 430079, China)

Abstract: The author represents the application situation of Chinese information processing in the field of dynamic geometry. In view of the distinguishing feature of geometry language, the author also introduces the dynamic geometry construction method based on the Chinese word segmentation technology and other natural language processing technology which includes GMMM algorithm, segmentation technology based on word segmentation dictionary, synonym database establishment as well as using language methods of pattern match and language pattern dictionary to formalize and regularize geometry proposition. Furthermore, some research prospects in this field are given.

Key words: Chinese information processing; geometry construction; dynamic geometry; participle

收稿日期:2011-05-03

作者简介:陈晓霞(1978—),女,山西省长治人,讲师,博士研究生,主要从事算法设计、自动推理理论等研究。

自 20 世纪 90 年代美国 Key Curriculum Press 公司制作出版了几何画板以来,世界上已经有近 50 种动态几何软件,目前国内被广泛使用的动态几何软件主要有几何画板和超级画板。但是,无论是几何画板还是超级画板,目前都没有实现基于自然语言输入的动态几何作图[1]。要让计算机像人一样能看会想、能听会讲,这是智能计算的最高境界,也是人类的梦想。如何将自然语言表述的初等几何命题自动转化为计算机可理解的作图语言是自然语言处理中比较新的内容,也是实现教育软件人机交互的难点。

目前国内广泛使用的计算机辅助教学软件如几何画板、超级画板等都需要使用者人工理解几何命题,再利用软件的绘图功能进行绘图,软件得到用户所提供的"作图序列"后,将"作图序列"转化为计算机能够理解的形式化参量,然后将具备几何关系的图形绘制显示给用户,用户在此基础上可以利用软件对所绘制的几何图形进行教学研究、数学实验等,但是,如果用户对几何命题的人工理解有偏差或者对软件的绘图功能无法熟练操作,则将导致后续的数学实验、教学研究无法进行。如果这些计算机辅助教学软件能够让使用者用他们所熟悉的自然语言提出几何问题,并且计算机也能正确理解,那么这样的教育软件将使人机交互更加方便、更加高效。

要使教育软件直接理解自然语言,首要的问题是输入问题,即软件能够根据用户所输入的自然语言进行作图,不需要用户进行翻译,计算机就像人脑一样直接读入这些自然语言,然后直接转化为形式化的参量和规则,并进行相应的动态几何作图。要实现基于自然语言的计算机作图,如同任何中文信息处理系统一样,首要的工作即对几何命题进行分词处理[2]。

1　分词技术在动态几何软件领域的应用现状

由于中文文本没有类似英文空格之类的显示词边界的标志,因此,相较于其他自然语言,对中文文本所要作的第一件事情是将词确定下来,即分词[1]。任何中文信息处理系统都需要建立在分词系统基础之上。从 20 世纪 80 年代初中文信息处理领域提出自动分词以来,有关方面的众多专家学者提出了许多分词方法,可以归结为:基于词典的方法、基于统计的方法和混合方法 3 类,同时也产生了一些实用性的分词系统[2]。

几何命题语言相较于普通语言有其自身的显著特点:初等平面几何范围内的命题相较一般的中文信息处理系统来说使用的词语量要少很多,仅 400 个词左右;所涉及词义主要是几何及部分代数术语,语义范围较窄;歧义较少,语义比较明确;句子之间逻辑关系比较强,句型变化较少;表达方式多为陈述句[1]。基于一般的分词方法,结合动态几何的特点,文献[1]和[3]构造了不同于一般分词系统的应用于专业几何作图领域的中文分词系统。

1.1　GMMM 算法

文献[3]提出了一种称为 GMMM 的算法实现几何语言的分词,以数据库的形式建立基本的词库字典,在此基础上首先使用了正向减词最大匹配法来实现基本的分词,算法思想为:

第一步,初始化两个字符串,s1 为待切分的字符串,s2 为已切分的字符串(初始化为空);

第二步,如果 s1 为空串,结束,输出 s2;

第三步,从 s1 左边复制一个子串 w 作为候选词,w 要尽可能长,但不超过预先设定的最大字符串长度;

第四步,如果此表中能找到 w 或者 w 的长度为 2 个字节,那么将 w 和一个分词标志一起加入到 s2 的右边,并且从 s1 的左边去掉 w,转到第二步;

第五步,去掉 w 中最后一个汉字(2 个字节),转到第四步;

第六步,输出 s2,结束。

由于汉字占 2 个字节,其他符号占 1 个字节,如果直接使用上述算法思想对混合了字母及汉字的文本

进行处理时，会导致分词结果出现乱码，因此还需要对这样的混合串进行切分，将其切分为若干个最大的只含字母的字母串和纯粹的汉字串的组合，然后对汉字串进行正向减词最大匹配算法，再进行初步分词。但对于含有几何意义的特殊符号的命题，存在命题结构不规范等问题，算法还有待进一步完善。

1.2 基于分词词典的分词方法

文献[1]构造了另外一种更加完善的分词方法，分词思想与 GMMM 算法基本相同，采用分词词典方法进行分词，基于字符串匹配原理进行分词和标注。初步实现了初等几何命题范围内的中文信息处理。要实现该方法需分以下几步完成。

第一步，建立分词词典，并对词素进行标注。文献[1]以近年来人民教育出版社出版的几何教材及习题集、试题集为基本素材，利用一般的自然语言分词技术对其中几何命题进行综合考虑，建立分词词典，并对词素进行词性标注：名词标注、动词标注和助词标注。其中名词标注是指将几何元素的基本元素名进行标注，例如“等腰三角形”标注为“三角形”；动词标注主要指几何元素之间的谓词关系的标注，例如“平行”“垂直”等；助词标注指表示范围和关系的词语如将“直线外一点”的“外”标注为助词。

第二步，初始化几何命题。用户输入命题往往是随心所欲的，例如线段 AB 平行于线段 CD，用户的输入可能是“AB//CD”“BA//CD”“线段 AB//线段 CD”“AB 平行于线段 CD”等多种形式，需要初始化用户随心所欲地输入的命题，包括：

1）对于混杂了中英文标点符号及字母大小写的输入进行统一化处理；

2）有特殊含义的符号的处理，如“//”“⊥”等转换成相应的中文处理；

3）对有可能产生歧义的符号，根据上下文情况判断它的含义，作初步的消歧，如“：”，在“已知：”和“AB∶BC=1∶2”中有不同含义；

4）去掉命题中与格式有关的符号，如空格、换行符等。

第三步，对命题进行初步划分。将命题中的所有字符依据几何元素名称、几何关系谓词、助词、标示字符、数字、断句符号、代数表达式等类别划分为 7 类。

第四步，建立分词树。对初步划分的几何命题根据分词词典的信息动态建立分词树，构造规则为：

1）以二叉树结构构造分词树（规则为每个节点的左子树节点词素小于当前节点词素，右子树节点词素大于当前节点词素，词素大小关系按拼音方式排列）；

2）树的每个节点为词素。

第五步，把命题拆分为一组有序的词素集合。对进行过初步划分的几何命题，使用向右搜索并且尽量套用最长语块的方式，通过不断搜索分词树，将一个完整命题拆分成一组有序的词素集合。

第六步，处理未登录词。为保证系统的纯洁性，忽略分词词典中未出现的词，可以在一定程度上消除输入性错误。

第七步，消歧。在分词结束后对于第二步未能解决的二义词再作进一步处理。

文献[1]中所使用的基于“分词词典”的分词方法相较文献[3]中的 GMMM 算法主要优点在于：建立的分词词典更具实用性；在分词词典中对词进行了预先标注处理，为后续建立完整的几何作图系统打下基础，并可以提高算法效率；对分词的几何命题进行预先规范化处理，处理了特殊符号，提高了分词的正确性和全面性。

2 利用中文信息处理技术建立几何作图系统

实现基于自然语言的几何作图，还需要在分词的基础上做更进一步的工作：句子拆分为词素及同义词集合代换、语模匹配、语义理解等几个部分。文献[4]实现了包括文献[1]中的分词技术在内的初等平面几何语言理解的语言系统。

2.1　利用同义词库对自然语言规范化处理

由于个人表达习惯的不同,以及语言的灵活性,虽然表达的是相同的含义,却会使用不同的表述方式,因此类似于一般的中文信息处理系统,还需要建立一个同义词库,在词库中建立一种对应关系,将表达相同含义词的词素都对应到一个词素上。利用这样的映射,可以将千变万化的自然语言在一定程度上规范化。

2.2　定义几何元素并建立几何元素库

将几何系统中的所有几何元素划分为基础几何元素和派生几何元素 2 类。其中基础几何元素只有 1 个,就是点。派生几何元素又分为简单几何元素和复杂几何元素。简单几何元素,各组成元素之间没有其他几何关系,仅有几个无序点派生而来,如直线、三角形等;复杂几何元素,各组成元素之间有约束条件,如平行线和梯形。平行线由 2 条平行直线构成,梯形有 2 条任意直线和 1 组平行线构成。其组成形式可描述为:

平行线(直线 1,直线 2)
　　直线 1(点 1,点 2)
　　直线 2(点 3,点 4)
梯形(平行线 1,直线 3,直线 4)
　　平行线 1(直线 2,直线 1)
　　直线 1(点 1,点 2)
　　直线 2(点 3,点 4)
　　直线 3(点 1,点 3)
　　直线 4(点 2,点 4)

文献[4]在几何元素的基础上建立了几何元素之间具有嵌套和继承关系的几何元素的集合,称为几何元素库,如图 1 所示。

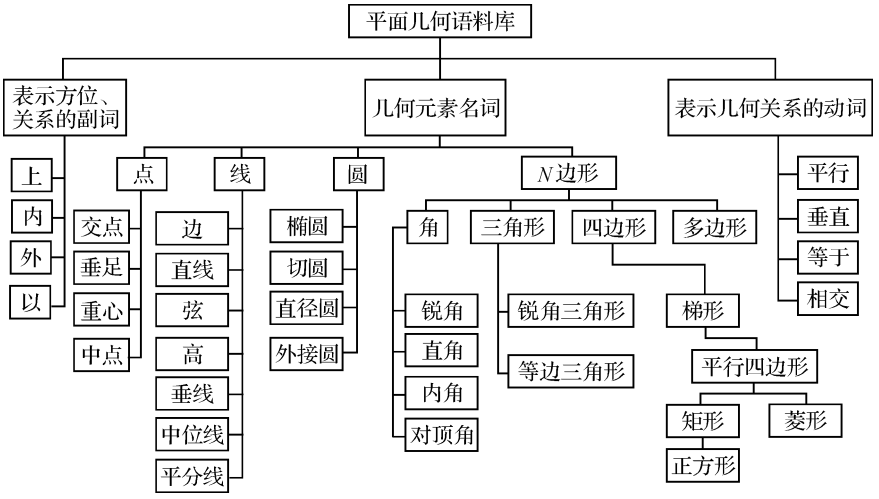

图 1　平面几何元素继承关系示意图

Fig. 1　Inheritance relationship diagram of plan geometry element

2.3　建立几何元素逻辑语义和作图语义

几何元素库建立了几何元素之间的结构关系,对象之间如面积、角度、边长等概念,还需要定义作图语义和几何对象之间的逻辑关系用于转换描述逻辑关系和作图含义的自然语言。如,在系统的正方形类建立如 CalculateArea 函数用于利用构成正方形的基础几何元素信息和其他几何约束信息来计算正方形的面积等。

2.4 利用语模匹配和语模词典形式化和规则化几何命题

利用同义词词库将几何命题转换为标准谓词、名词后，由这样的标准动词和几何元素所组成的语句称为语言模式，例如：自然语言所表达的“两圆的交点”，经标准化等处理后表述为“相交于(圆，圆)，返回值：点”，再形式化为计算机所能理解的形式化命题为“Intersection_point (circle(p1, p2), circle (p3, p4))”。把上述形式化的几何命题“Intersection_point (circle(p1, p2), circle (p3, p4))”组成一个线性表，通过查询这样的线性表，匹配成功后，即可产生相应的绘图命令。最终实现基于自然语言处理的计算机几何作图。

利用上述中文信息处理技术，结合动态几何特点所创建的基于自然语言处理的几何作图系统，可基本实现初等几何范围内基于自然语言输入的动态几何作图。

3 结 语

基于以上研究可以看出，在几何命题的机器理解中关于中文分词和标注的问题是一个庞大而复杂的系统，涉及未登陆词的处理、词典组织和选词、语构词法、汉语词语分类体系等各方面。在国内基于中文信息处理的动态几何系统还处于理论研究的阶段，还没有被广泛应用于当前的产业化软件中，要真正达到计算机能够理解汉语的程度，还要兼顾理解的正确性和语言的灵活性。一个完善的基于自然语言理解的动态几何作图系统可以为几何学专家系统、几何学信息检索系统等构建一个良好的平台，同时通过构建几何学专家系统等可以促进基于自然语言理解的动态几何系统的发展[5]。要构建一个基于自然语言的动态几何系统，并使之成功应用于教育、科研领域还有很长的路要走。

参考文献：

[1] 佘莉，符红光，方海光. 几何命题处理中的中文分词技术[J]. 计算机工程，2005，31(18)：180-182.

[2] 孙茂松，邹家彦. 汉语自动分词研究评述[J]. 当代语言学，2001，3(1)：22-32，77.

[3] 解烈军，候晓荣，周彩莲. 基于规则的几何语言自动分词算法[J]. 淮阴师范学院学报：自然科学版，2004，3(2)：152-155.

[4] 佘莉，符红光. 基于自然语言处理的计算机几何作图[J]. 计算机应用，2005，25(1)：7-10.

[5] 钟秀琴，符红光，佘莉，等. 基于本体的几何学知识获取及知识表示[J]. 计算机学报，2010，33(1)：167-174.