一种基于云端信息保护的汉字计算模型

栗青生[1,3] 张 莉[2] 刘 泉[1] 熊 晶[3] 杨新新[3]
(武汉理工大学信息工程学院 武汉 430070)[1] (湖北经济学院工商管理学院 武汉 430205)[2]
(安阳师范学院计算机与信息工程学院 安阳 455000)[3]

摘 要 提出了一种基于信息内容保护的信息安全模型。该模型利用将汉字笔画抽象为有向图的方法,设计了汉字笔画图抽象的具体方案,实现了对汉字字形结构的动态描述;建立了动态汉字字形描述库,设计了汉字字形的生成算法,实现了汉字字形的Web存储和特征字形的客户端输出。所提模型为汉字信息的云端存储和云端数据安全性保护提供了一种解决方案,不仅有助于汉字信息的安全保护,而且有助于汉字认知计算、语义计算等深度汉字信息计算。

关键词 汉字,笔画,字形描述,有向图,信息安全
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.015

Chinese Character Computing Model Based on Cloud Information Protection

LI Qing-sheng[1,3] ZHANG Li[2] LIU Quan[1] XIONG Jing[3] YANG Xin-xin[3]
(School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China)[1]
(School of Business Administration, Hubei University of Economics, Wuhan 430205, China)[2]
(School of Computer and Information Engineering, Anyang Normal University, Anyang 455000, China)[3]

Abstract A new cloud information protection model was proposed based on information protection. The method of Chinese strokes abstract was designed by abstracting the digraph from the strokes of Chinese characters. Using this method, a dynamic description library for the glyphs of Chinese characters was constructed, and an algorithm of characters generation was shown based on the library. The dream of storing the glyphs of Chinese characters on the Web and outputting the characteristic glyphs on the client was realized. In addition, the model provides a solution to the storage of Chinese characters and the protection of information in the clouds. The model of Chinese computing based on information security is not only helpful for machines to understand linguistics but also helpful to protect information, semantic computation and information security computation.

Keywords Chinese character, Strokes, Character description, Digraph, Information security

1 引言

30多年前,为了让汉字进入计算机,国内外的汉字研究学者做出了很大的努力,最终在英文单字节编码基础上找到了汉字的双字节编码体系。30年后,汉字进入了互联网时代,就在我们尽情享受信息共享和信息交换给我们带来的巨大成功和幸福的同时,我们也碰到了汉字整字编码给中文信息处理技术带来的困难和问题。

汉字是象形文字,它的笔画或部件往往具有一定的语义信息。但是,目前汉字的运算是以整字编码参与计算,从汉字中分解出笔画和部件是一件非常困难的事情,这使得本来应该在汉字的语义计算中发挥作用的汉字部件失去了其独特的意义,给中文信息技术的研究和发展带来了困难。在互联网和大数据时代,在多元化的数据表示迅速发展的时代,这一问题已经引起了国内外很多研究学者的关注和重视[1,2]。

不可否认,汉字信息处理技术的发展已经取得了巨大的成就。汉字信息描述技术的研究经过几十年的发展,已经开始从宏观走向微观、从通用走向个性化、从规模处理走向精细处理阶段,在这个变化过程中,越来越多的用户对安全、无泄露、无风险的信息的需求日益增强。例如,平时读完的电子邮件或其它文本文档,如何能彻底销毁而不会被再次恢复?商业化的信息服务商提供给用户有限次阅读和访问的信息如何确保在过期之后文档能彻底"粉碎"等等。如果是日常的办公事务,办公人员可以将过期或不用的纸质文件用碎纸机对印刷文件进行彻底的粉碎以达到保密和保护的目的。但如果是电子文档,由于其存储媒体的特殊性,即使是删除后的文档,也可以通过数据恢复技术使其恢复原貌。

解铃还须系铃人,汉语表示的信息安全问题再一次让我

到稿日期:2014-10-29 返修日期:2014-12-14 本文受国家自然科学基金(60973051),河南省科技攻关项目(132102210115),河南省基础前沿研究项目(152300410089)资助。

栗青生(1966—),男,博士,教授,主要研究领域为自然语言处理、汉字字形计算;张 莉(1974—),女,博士,副教授,主要研究领域为服务管理、信息安全管理;刘 泉(1963—),女,教授,博士生导师,主要研究领域为信号处理和嵌入式技术;熊 晶(1979—),男,博士,副教授,主要研究领域为自然语言处理。

们考虑到汉字的编码和表示问题。如果从信息表示的源头切断了信息的安全隐患，或许可以给我们一种最更安全的网络环境。

文字是语言的细胞，对语言的理解、存储和判断不仅与字词句有关，而且与文字的描述和表示方式有关。通常所说的汉字内码就是汉字表示方式的一种，它与英文的 ASCII 码一样，每一个汉字都有唯一的内码，是我们进行中文信息处理的基础。汉字的外码也叫汉字输入码，它是为了人的理解和输入的方便，确切的说是为了人机交互的方便而进行的编码，汉字的外码可以有很多种编码，如汉语拼音编码、五笔字型编码等等。汉字的内码和外码两种描述方式解决了机器的信息交换问题，但解决不了机器的理解和认知问题，因此出现了对汉字的各种各样的描述方案。

2 汉字的描述和信息保护

2.1 汉字信息的描述

心理学家曾经对文字和图画做过分类的认知实验，结果表明，人对图画的分类快于对英文词、汉字词的分类[3]。而对计算机而言，结果似乎正好相反，即互联网上对文字的检索远远要比对图画的检索方便得多。其原因与文字和图画的计算机表示有关，相对图画而言，文字检索是基于编码匹配的方法，表示文字的编码数据比较规范，适合计算机处理；而对于表示图画的非编码数据，计算机处理起来比较困难。因此让计算机检索一幅图画远比检索文字困难得多。

正是因为文字的编码表示在信息交换和信息检索方面的优越性，才使得文字信息表示技术的发展非常缓慢。从计算机出现到今天的互联网时代，数字图像表示、理解和感知技术始终是信息技术研究的热点领域，而有关文字信息表示的研究往往被忽视。以汉字为例，由于使用整字编码机制，汉字笔画和部件结构中蕴含的大量语义信息被忽略，使得本来应该很简单的汉字笔画分解变得异常复杂。汉语的认知和汉字的理解一样，需要有更多的知识存储、更规范的知识表达和更科学的计算方法，这就意味着要有更大的存储空间、更多的知识积累和更大的表示范围，才能满足基本的认知需求。

在这方面，我们可以从计算机处理复杂图像的过程中得到启示。面对复杂的图像，计算处理图像的过程基本上都要对复杂的图像进行抽象，略去一些次要因素，最后将对象抽象为可以自由存储和计算的骨架。与图像处理一样，计算语言学家也要对自然语言进行抽象、简化，以便利用现在的计算技术对它进行处理，将其从一种表现形式变换成另一种表现形式。

如果要对表示信息基本元素的汉字继续在字和词一级以下再抽象，目前的汉字编码机制是没办法实现的，因为汉字是整字编码，汉字在笔画和部件一级的编码机制是不可以计算的。为了解决这个问题，以对更细一级的笔画和部件进行计算，我们自然想到对现在的汉字编码进行一次再加工处理，这就是在汉字编码之外，根据汉字的语义信息、部件或笔画信息分别进行描述。

2.2 汉语信息的保护

传统的汉语信息是以结构化文档的方式（如数据库）进行存储，因此，通常所说的信息保护策略是基于结构化数据（数据库）的认证或准入机制的信息保护策略，这一策略的最大缺点是一旦认证机制失效，数据库中的所有信息都会出现极大的安全隐患。

结构化文档是云计算环境中实现服务组合与信息交互的核心载体，与其他非云计算环境不同，它不再是静态内容和单一版本，而具有动态性、时空立体性、多用户性、多安全等级、多媒体性与多版本性等“活”文档特征[4]。如何保护这种“活”结构化文档的机密性和完整性，实现对其安全访问，是云计算服务质量控制的关键。因此，迫切需要一种具有多级安全、高效且灵活的访问控制机制以实现基于内容的结构化文档的多级安全保障机制。

本文从信息描述的角度，提出了一种基于内容的文档信息保护方法。该方法通过定义比汉字笔画更小的描述单位——笔元，将编码汉字库抽象转换为非编码汉字信息描述库，克服了目前汉字字库只能在客户端文件存储的不足，实现了基于笔元描述的多级别汉字存储和云端存储功能。用户只需要将现有的编码文档进行一次编码转换，转换后的文档就具有了安全保护功能，既可以进行在线阅读，也可以进行离线存储，如果用户的许可超过了阅读期限和使用次数，文档中的汉字信息内容便会像纸片粉碎机一样将汉字信息进行粉碎，从而起到信息保护的作用。

3 相关工作

自然语言研究过程中，为了更多地发掘出语料中蕴藏的语言知识，通常要对语料进行标注、分析和处理。这里讲到的“标注”，从机器理解的角度来说就是用机器可以理解的描述表示机器不理解的描述。

3.1 以语料库和语言知识库建设为目的的字、词、句的标注和描述

语料库的建设是基于统计的自然语言研究方法不可缺少的基础。目前我国已有多个百万字以上容量的汉语语料库，用于语言信息处理的各种研究和应用，其中比较有代表性的是《人民日报》标注语料库[5]。语言知识库是用结构化的形式通过记录语言使用原貌来呈现语言知识，它是使语言信息处理系统得以运行的直接资源。语言知识库的典型代表是《知网》[6]，它是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。除此之外，还有其他一些语言知识库，例如，现代汉语语义词典、现代汉语短语结构知识库、中文概念词典、汉语句法树库等。近几年来，台湾中央研究院在语言知识资源方面也做了相当扎实的工作，例如“中文句结构树资料库”[7]和“中文词汇网络（Word Net）”[8]就是两个比较完整的具有 Web 服务功能的中文语义知识库。

语料库和语言知识库作为汉语智能化语义分析的工具，为汉语的认知和计算提供了极大的方便，但其单一的描述方式不能满足机器多方位和多功能的感知和理解。例如，对于人来讲，给定一个词或者一个概念，人能马上联想到这个词所表达事物的形状、大小、颜色，甚至图形或图像等等。而对于机器，其尽管不能和人一样去感知更复杂的图像，但可以通过分析和对比来实现机器的认知和计算。另一方面，这种基于句、词和字标签的标注方法，虽然能在安全信息的检测和敏感信息的过滤方面起到积极的作用，但不能保证存储文档的信息不会被盗取，基于内容的文档保护用这一方法实现起来比较困难。因此，我们需要更多的、更具体的或者说更确切的描述，这就是本文要介绍的几种在字词级及字词以下级进行的深度描述方案。

3.2 以组字和组词为目的的汉字部件和笔画的描述

针对表示信息的汉字数量巨大以及汉字的机器组字、组词效率较低等问题而出现的基于汉字部件和笔画的描述方法，目前也被广泛应用于文档理解和语义识别。较有代表性的工作主要有香港浸会大学提出的 Han Glyph[9]、美国加州大学伯克利分校提出的基于笔画和汉字部件的字形描述语言 CDL(Character Description Language)[10] 等。如果将部件描述进一步细化，就是笔画描述，这方面 CDL 做得非常好，其兼顾了部件和笔画两种描述方法，特别受到了汉字学者的重视，并从多个方面得到了扩展，SCML(Structural Character Modeling Language)[11] 就是在 CDL 的基础上提出的一种将字形和结构融合起来进行综合描述的方法。

以组字为目的的汉字部件和笔画的描述，绝大部分使用数字标签对汉字的结构进行标记，其主要目的是实现汉字智能输入和文档的智能识别。其特点是将汉字的整体结构描述向前推进了一步，在笔画分解、识别和计算方面显然要优于用整字为单位的字词描述方案。但它同样存在缺少更多细节描述的缺陷，同样不能保证信息表示的安全，即使给汉字的部件再加上一个语义标签，也只是在字词一级的进一步细化，仍然不具备安全语义计算的条件。

3.3 以罕用字的表示为目的的汉字字形的笔段描述

笔段描述是将汉字部件描述继续细化的描述方法，这里的笔段是笔画的子集，它可以是笔画或笔画的一部分。在笔段描述方面，北京语言大学、内蒙古师范大学的宋柔和林民所领导的研究团队在基于笔段的汉字形式化描述方面做了很深入的研究工作，提出了基于笔段网格的生僻字、错字输入方案[12]。虽然在数量上有穷尽的可能性，但如果应用于基于内容的文档信息保护和汉字的语义计算，该描述方案还有很多工作要做。

3.4 以动态组字和生成字形为目的的笔元描述

笔元描述是我们在甲骨文数字化过程中提出来的一种有效的多字种描述方案[13-17]。简单的说，笔元就是有方向的笔段，它保留了笔画和笔段描述汉字的特点，加入了汉字书写过程中行笔方向性属性，并且将笔画或笔段之间的连接点以界点、驻点或势点进行分类描述，使描述过程更加简洁，更适合汉字(包括各类字体)的字形计算。

即将到来的下一代互联网，网络存储、云端存储、大数据服务等一系列新技术的应用，使得汉字的信息表示拥有了多元描述的条件，具备了多元描述的空间。因此，要实现让机器理解更容易的描述，应该面向互联网深入研究汉字的多元描述技术，结合汉字的特点，研究汉字信息深度认知计算的方法。

4 汉字信息的动态描述

设计甲骨文字形动态描述算法的最初目的是让计算机能自动地生成同一内容但不同形态的甲骨文字。甲骨文字形描述库以特征点数据存储甲骨文特征信息，没有编码，借助在互联网服务器端设计的各种服务，在互联网的客户端，各类甲骨文字形就可以动态生成。将这一思路应用于现代汉字的描述，就是建立现代汉字的字形描述库，在这个描述库中汉字的可控制性和可计算性将会得到极大的增强。

4.1 汉字笔画的特征点

根据汉字的书写规则，每一汉字的每一笔画的书写过程始终是一落(落笔)和一起(提笔)的过程，但不同的人有不同的书写习惯，有时会在一个落笔和一个提笔之间有数次的停顿，为此，除了落笔和提笔这两点之外，停顿点也是汉字的一种重要特征点。

定义 1(汉字的特征点) 汉字字形的特征点 T 是汉字书写或形成过程中的 3 类端点的集合，这 3 类端点是：书写笔画的开始点、停顿点和结束点。将笔画的开始点称为始结点 Ds，停顿点称为驻点 Dz，结束点称为尾结点 De。详细内容请参看参考文献[13,15]。

汉字的笔画描述有以下两种特殊的情形：

(1)只有始结点和尾结点，但没有驻点的描述。

如“横”或“竖”，可以表示为：

$Z\{T[(Ds,De)]\}$

(2)两个笔画的连接描述。当一个笔画的始结点或尾结点与另一个笔画的始结点或尾结点重合时，我们称之为两个笔画的连接。

例如，一个笔画的尾结点和另一个笔画的始结点相连可以表示为：

$$Z\{\cdots T[(Ds_i,De_i)],T[(De_i,De_{i+1})],\cdots\}$$

或者表示为：

$$Z\{\cdots T[(Ds_i,Ds_{i+1})],T[(Ds_{i+1},De_{i+1})],\cdots\}$$

同样，两个笔画的始结点相连可以用类似的表达式表示。

为了和普通的特征点相区别。笔画连接处的特征点也称为连接点。连接点既可以由尾结点和始结点连接产生，也可以由两个始结点相连或两个尾结点相连来产生。

用定义 1 的特征点描述，可以实现结构相同或相近的不同汉字字体的统一描述，如图 1 所示。

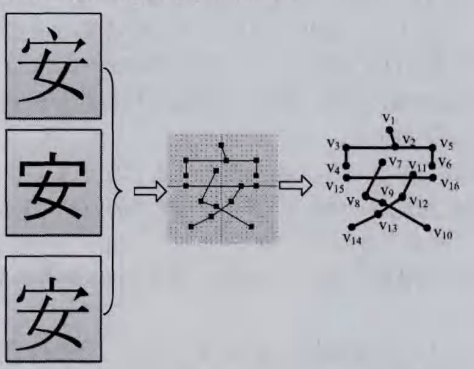

图 1 不同字体的“安”字对应的特征点

4.2 汉字笔画的笔元

从汉字特征点的定义 1 中可以看出，汉字每个笔画至少有两个端点，一个是开始点 Ds，称为始结点；一个为结束点 De，称为尾结点；始结点和尾结点之间的过程点称为驻点 Dz。

并不是汉字的每一笔画都有驻点，例如标准汉字中的“点”笔画，它只有一个始结点和一个尾结点。但是，如果是艺术或手写字体，在点的形成过程中有一个回笔过程，如果忽略这个回笔过程，汉字的字形计算将会失去其特有的书法特征。另外，同一汉字的描述随字体的不同而有所区别。如图 2 所示，“江”字三点水中点笔画的描述可以有驻点，也可以没有驻点，如果有驻点，驻点的数量和位置的不同会使得汉字的字形特征有所不同，因此，驻点是体现汉字不同特征的描述。为了实现对汉字更多特征的描述，引入汉字笔元的概念。

(a)黑体字形　　(b)行楷字形

图 2　同一汉字不同字形的汉字的描述

定义 2(笔元)　笔元是汉字笔画的特征点之间、特征点与驻点之间或驻点与驻点之间的一个有方向的线段。

设 $T(v_s,v_e)$是汉字笔画的两个特征点，$v_{z1},v_{z2},\cdots,v_{zn}$是其中 n 个驻点，则汉字的笔元可以表示为：

$Y(v_{zi},v_{zj})$

如果用笔元 Y 来表示汉字的笔画，则有：

$T(v_s,v_e)=Y(v_{z1},v_{z2})+Y(v_{z2},v_{z3})+\cdots+Y(v_{zi},v_{zj})$

其中，$i,j\in\{1,2,3,\cdots,n\}$。

由定义 2 可以看出，汉字的笔元集合是特征点集合的一个子集。这个集合在汉字字形的空间变换中起着重要的作用。

事实上，每一个汉字笔画中的驻点是随机的或不确定的，我们称之为动态的。因此，用笔元描述的汉字是动态汉字。

从字形描述的角度来分析汉字特征，汉字笔画特征可以分为字形特征和语义特征。字形特征是汉字外观表征。语义特征是汉字内容的表征。汉字特征点的描述中，在汉字笔画的特征点确定的情况下，增加或减少汉字的描述中的驻点数量或者增加或改变汉字的笔元数量，并不会改变汉字或笔画所包含的语义信息。

4.3　汉字笔画的图抽象

汉字书写过程中，多个笔画之间的书写关系还可以是相交或相切的关系。定义 1 和定义 2 已将落笔点设定为始结点，将提笔点设定为尾结点，行笔过程中的所有行笔方向的改变和停顿都视为驻点。始结点、尾结点和驻点是汉字笔画抽象中的重要特征点。

将汉字抽象为图，要解决两个最基本的问题：1)抽象时应该遵循的原则；2)对撇、捺、折等特殊笔画的处理。

4.3.1　汉字笔画抽象的原则

汉字抽象的原则是“由上到下、从左到右，笔画为先，兼顾直观”。

由于汉字的组成结构复杂，汉字的每一个笔画的特征点之间或笔画与笔画之间都有一些很重要的细节问题需要考虑，如特征点的位置关系问题和笔画之间的连笔问题(见图 3)等。如果完全不考虑这些笔画的细节，对不同字体的汉字特征的抽象一定会有影响，但过分地考虑笔画特征将会增加汉字抽象的难度。因此，我们以“笔画为先”，即要以笔画的书写规范为重要的参考依据；“兼顾直观”，即要充分考虑汉字形成的整体空间结构。例如：对“安”字的抽象描述中，重点考虑“安”字中点笔画和横折笔画之间的相切和相离关系，而对“安”字下半部分中“女”字的各个笔画之间的关系则不予考虑，如图 4 所示。

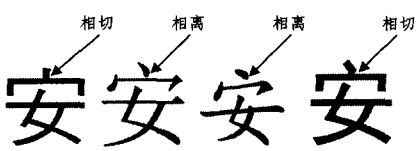

图 3　同一汉字不同字形的汉字的描述

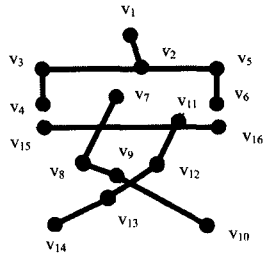

图 4　汉字的特征点描述

4.3.2　特殊笔画的处理

定义驻点是解决汉字中撇、捺、折等特殊笔画抽象的重要方法，在这些特殊笔画的抽象中，驻点不仅是连接始结点和尾结点的重要特征点，而且还起到表征书写方向的作用。例如，汉字中的撇，正是由于定义了驻点，使得笔元的描述更加符合汉字的字形特征，如图 5 所示。

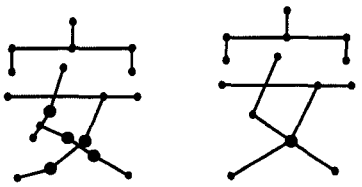

图 5　汉字的特征点(小)中的驻点(大)

另外，还有一种特殊情形，即汉字的连笔问题。在汉字书写过程中，汉字内部的连笔和汉字之间的连笔现象非常普遍。我们依然用始结点和尾结点来描述一个完整的笔画，且用第一个笔画的尾结点和第二个笔画的始结点的重合来描述这两个笔画之间的关系。

标准汉字偏旁三点水“氵”中，3 个点本来是相离的，但由于书写特征使其两个或三个笔画相连的现象就属于这种情形。如图 6 所示，标准状态下 v_6 和 v_7 应该是分离的，但对于书写特征的描述，这两个点可以重合，重合之后的点仍然是两个点，它既是上一个笔画的尾结点，又是下一个笔画的始结点。

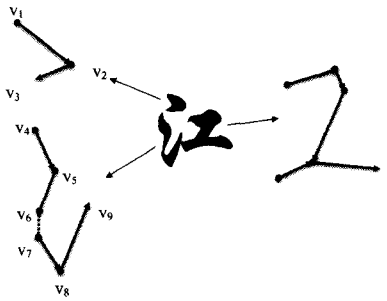

图 6　汉字抽象中的连笔书写问题

4.3.3　笔画图抽象步骤

给定的汉字笔画 T 转化为有向图 G 需要从汉字笔画中抽象出顶点和边并对这些顶点进行标号，具体的抽象和标号步骤如下：

步骤 1　每个孤立的笔画 T 中始点抽象为图 G 的一个顶点，并按从上到下、从左到右的方向和书写顺序，每次在原来方向上的改变视为图 G 的另一个顶点，直到孤立的笔画结束(尾点)为止，依次标号为 $v_1,v_2,\cdots,v_r$，其中 r 为此类顶点的数目。

步骤 2　每个不与其它笔画相交/相切的笔画的始点抽象为图 G 的一个顶点，并按照从上到下、从左到右的方向和书写顺序，每次在原来方向上的改变视为图 G 的另一个顶

点，直到孤立的笔画结束(尾点)为止，依次标号为 $v_{r+1}, v_{r+2}, \cdots, v_{r+s}$，其中 s 为此类顶点的数目。

步骤 3　在每条不与其它笔画相交/相切的封闭笔画上任取两个点作为图 G 的两个顶点，并按照从上到下、从左到右的方向依次标号为 $v_{r+s+1}, v_{r+s+2}, \cdots, v_{r+s+t}$，其中 t 为此类顶点的数目。

步骤 4　每个与其它笔画相交/相切的笔画的始结点抽象为图 G 的一个顶点，并按照“从上到下、从左到右，笔画为先，兼顾直观”的书写方向和顺序，每次在原来方向上的改变视为图 G 的另一个顶点，直到孤立的笔画结束(尾点)为止，以相交/相切的端点为顶点，依次标号为 $v_{r+s+1}, v_{r+s+2}, \cdots, v_{r+s+p}$，其中 p 为此类顶点的数目。

步骤 5　将每个与其它笔画相交/相切的笔画的端点(也称连接点)抽象为图 G 的顶点，并按照从上到下、从左到右的方向的书写方向和顺序，每次在原来方向上的改变视为图 G 的另一个顶点，直到孤立的笔画结束(尾点)为止，依次标号为 $v_{r+s+t+p+1}, v_{r+s+t+p+2}, \cdots, v_{r+s+t+p+q}$，其中 q 为此类顶点的数目。

步骤 6　将经过上述步骤得到的每两个相邻顶点连接为一条边。

例如，行楷汉字“江”字的第一笔画——“点”笔画，是一个孤立笔画，经过步骤 1 时将始结点抽象为图 G 的一个顶点，中间方向改变了一次，因此增加一个顶点(该顶点其实是定义 1 中的驻点)，此时，$r=3$。在这个笔画的抽象过程中，由于不满足步骤 2 到步骤 5 的条件，因此，不会抽象出其它顶点，执行步骤 6 后，这个笔画的抽象结果如图 7(a)所示。同理，第二笔画是上一个笔画和下一个笔画相连的笔画，不满足步骤 1 到步骤 2，执行步骤 3－步骤 6 后的抽象图如图 7(b)所示。图 7(c)是整个汉字的抽象图。

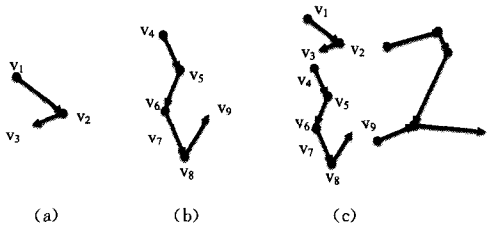

图 7　汉字笔画的抽象图

5　动态汉字字形的生成算法

实现在笔画一级汉字的计算，需要根据第 4 节所提供的汉字抽象的技术和方法对汉字的每一笔画进行抽象描述，将抽象描述的结果根据字体类型和汉字笔画的类型进行分级存储，形成动态汉字字形描述库

5.1　汉字字形的动态生成算法

对 4.3 节中汉字笔画的图抽象过程进行逆化处理即可得到汉字笔画的生成算法。

为了节约篇幅，在此只给出简单笔画文字的生成算法，有关复杂笔画生成的详细内容请参考文献[13,15]。

Step 1　定义并初始化

笔画的始点 $V_1(x_1, y_1)$初始化。

笔画的尾点 $V_2(x_2, y_2)$初始化。

端点 D(x,y) 初始化。

驻点数组初始化 Z[x,y]。

Bool M;//笔画相交/相切或笔画不相交/相切。

int L;// 相交/相切笔画数，L=1 是孤立笔画。

输出类型：

N;// 1:正常输出；

2:粉碎输出；

Step 2　笔画类型判断

Step 2.1　If L=1

Then {

$V_1(x_1, y_1)$←始结点；

$V_2(x_2, y_2)$←尾结点；

Z[x,y]←驻点；

goto Step 3；

goto Step 4；}

Else

goto Step 2.2；

//孤立的笔画，取得笔画中的特征点，调用绘制程序，结束该笔画的绘制。

Step 2.2　$V_1(x_1, y_1)$←始结点；

If $V_1(x_1, y_1)$!=D(x,y)

Then {

$V_2(x_2, y_2)$←尾结点；

Z[x,y]←驻点；

goto Step 3；

L=L−1；}

Else

goto Step 2.3；

//非孤立的笔画，不相交/相切，开始为始结点的笔画绘制。

Step 2.3　$V_1(x_1, y_1)$←端点；

$V_2(x_2, y_2)$←尾结点；

Z[x,y]←驻点；

goto Step 3；

If L1 !=0

Then

goto Step 2.2；

Else

goto Step 4；

//非孤立的笔画，不相交/相切，开始为端点的笔画绘制。

Step 3　绘制笔元

Step 3.1　调用相应的 GDI 绘图指令或选定特定的数字墨水绘制技术；

Step 3.2　设定起始点、结束点的状态，初始化输出笔画大小Strokes Size、变化空间值 Deform Value、X-Y 方向位移 X-Y Move 及在 X 和 Y 方向的膨胀值 Expand X 和 Expand Y 等；

Step 3.3　判断输出类型

Case1:标准或正常输出

Case2:变形或粉碎输出

Step 3.4　返回

Step 4　当前笔画绘制结束。

5.2　算法验证实验

在 5.1 节设计的算法的基础上进行字形输出验证实验，实验分为标准输出实验和变形输出实验两部分。标准输出实验主要验证本算法汉字输出的有效性，如图 8 所示；变形输出实验主要验证汉字输出的变化空间大小，如图 9 所示。

从图 8 可以看出，汉字字形描述库的标准汉字输出完全正确，并且依据 Strokes Size 的大小具有一定的变化规模。

从图 9 可以看出，多达几百甚至几千种的字形变化可以通过控制 Deform Value 的值由字形描述库来动态生成。

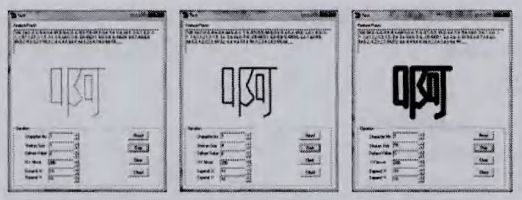

图 8　标准输出实验

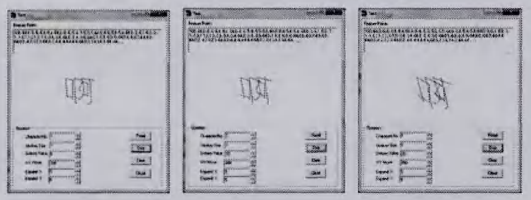

图 9　变形输出实验

上述实验中，如果再考虑生成汉字在 X-Y 方向的位移值 X-Y Move、在 X 和 Y 方向的膨胀值 Expand X 和 Expand Y 等变量值及其组合，完全有可能实现我国人均一套汉字字形系统，使汉字信息像指纹一样受到保护。

6　基于内容的网络信息保护

动态汉字字形描述库提供了一种基于内容的网络信息保护方式，这一保护方式可以根据用户的需求，将需要保护的文档通过文档转换，将通用编码机制下的文档转变为基于动态汉字字形描述库的受保护文档，通过服务器端的有效控制，实现基于内容的网络信息的保护。例如，用户可以设定信息输出的次数和变化范围，根据使用次数来设定每一次的输出信息中文字的变化状况，或者根据文字的变化状况来识别用户的合法性等等。

另外，由于字形描述库是存储在服务器端的汉字特征描述信息，受保护的文档必须完全具备这些特征才能正确使用，因此可以根据受保护信息的需要，将描述库信息分别存储于不同的服务器，实现信息的多级保护，如图 10 所示。

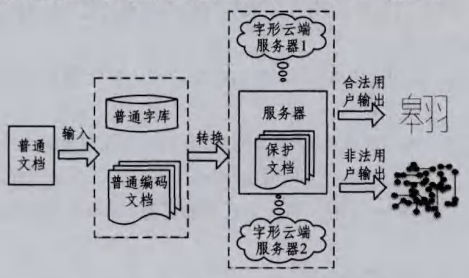

图 10　编码汉字转换和保护输出过程示意图

从图 10 可以看出，汉字字形描述库是通过客户端的动态生成来实现信息的安全服务。它与现代汉字字库的区别主要表现在以下 4 个方面：

(1)存储方式不同。动态汉字字形描述库中汉字是非确定编码汉字，可以使用分布式存储和计算，并且可以广泛存储在云端服务器中，为用户提供安全的个性化的字形服务。而汉字字库中的汉字是编码汉字，只能存储在本地客户端，仅适合字库文件计算。

(2)使用方式不同。非确定编码汉字可以在客户端根据用户的需求动态显示信息结果，因此具有信息安全的保护功能。而字库在客户端显示信息的前提是客户端必须安装相应的字库文件系统。

(3)信息的保护方式不同。目前通用的网络信息保护方式采用的是认证和准入机制，这种信息保护方式的缺点是一旦这一机制失效，信息保护手段就面临失败的危险。而使用动态汉字字形描述库可以实现在认证和准入机制失效的情况下，利用云服务端的多级汉字特征描述库，通过服务器端的动态控制，实现基于内容的信息保护。

(4)信息保护的级别不同。目前，通用的文档信息安全保障机制是在标准编码下的多级机制，级数由算法来确定，而基于内容的信息保障机制可以根据在描述层的分级来增加保护的级别。例如，可以将汉字信息的描述在汉字字形描述库的基础上再分别进行词或句层的描述，这样数据信息的安全保障则可由 n 级数据安全保障机制来完成，n 的取值由描述层的深度即被描述文字在笔画层、字层、词层或句子层的结构数量确定。假设目前使用的信息加密算法对数据进行了 k 级数据保护，那么，使用基于内容的数据信息保护就可以达到 $k\times n$ 级。

对标准字库来说，假设保护 1 个汉字的级别为 1，那么，两个汉字的词的保护级别就可以提高到两个字的组合即 2，3 个汉字的词的保护级别可以提高到 3 个字的组合即 6。而对汉字描述库来说，以每个汉字平均为 10 个特征笔元点来分析，以上情况的保护级别就可以得到 10 倍的保护规模，如图 11 所示。

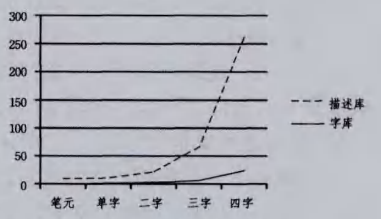

图 11　字库和描述库信息保护级的比较

结束语　快速发展的互联网技术为多特征描述的汉字信息存储创造了条件。与传统的文件系统发展一样，随着应用需求的变化和计算机技术的进步，以信息交换为主要服务内容的字库文件系统也会逐渐演变到具有丰富的文学字形计算服务功能的多种不同类型的网络文件系统、并行文件系统或分布式文件系统等。本文基于图形理论，使用动态汉字字形描述库实现了汉字笔元、笔画和汉字部件结构的多层次描述；扩展了汉字信息的存储方式和存储途径；建立了一种基于内容的汉字信息保护模型；提高了汉字信息的保护手段和保护级别，为互联网环境下中文信息安全性的保护找到了一种有效的解决方案。接下来将基于汉字字形描述库，研究更深层次的描述机制——基于内容的网络词语和语言识别技术。

参 考 文 献

[1]　俞士汶. 语言随计算齐飞[J]. 当代语言学，2009，11(2)：97-99
Yu Shi-wen. Languages with calculation and ebony[J]. Contemporary Linguistics，2009，11(2)：97-99

[2]　马希文. 可计算的语言[M]//逻辑-语言-计算. 北京：商务印书馆，2003：45-60
Ma Xi-wen. Computable language[M]//logic-languages-calculation. Beijing：The Commercial Press.，2003：45-60

[3]　张积家，王娟，刘鸣. 英文词、汉字词、早期文字和图画的认知加工比较[J]. 心理学报，2011，43(4)：347-363
Zhang Ji-jia，Wang Juan，Liu Ming. The Comparative Study on English Words，Chinese Words，Early Words and Pictures[J]. Acta Psychologica Sinica，2011，43(4)：347-363

[4] 熊金波,姚志强,等.基于行为的结构化文档多级访问控制[J].计算机研究与发展,2013,50(7):1399-1408
Xiong Jin-bo, Yao Zhi-qiang, et al. Action-Based Multilevel Access Control for Structured Document [J]. Journal of Computer Research and Development, 2013, 50(7): 1399-1408

[5] 俞士汶,段慧明,朱学锋,等.北京大学现代汉语语料库基本加工规范[J].中文信息学报,2002,16(5):49-56
Yu Shi-wen, Duan Hui-ming, Zhu Xue-feng, et al. The Basic Processing of Contemporary Chinese Corpus at Peking University SPECIFICATION[J]. Journal of Chinese Information Processing, 2002, 16(5): 49-56

[6] 《知网》[OL]. [2014-2-20]. http://www.keenage.com

[7] 台湾中央研究院.中文句结构树资料库[OL]. [2014-2-20]http://TreeBank.sinica.edu.tw

[8] 黄居仁.中文词汇网络[OL]. [2014-2-20]. http://lope.linguistics.ntu.edu.tw/cwn

[9] Yiu C L K, Wong Wai. Chinese character synthesis using METAPOST[OL]. [2014-2-22]. http://www.tug.org/TUGboat/tb24-1/yiu.pdf

[10] CDL: Character description language [OL]. [2014-2-22]. http://www.wenlin.com/cdl

[11] Peebles D G. A Structural Representation for Chinese Characters[OL]. [2014-1-24]. http://www.cs.dartmouth.edu/reports/TR2007-592.pdf

[12] 林民,宋柔.汉字字形形式化方法研究[J].计算机科学,2007,34(11):185-189
Lin Min, Song Rou. Formal Description of Chinese character Glyph[J]. Computer Science, 2007, 34(11): 185-189

[13] 栗青生,熊晶,吴琴霞,等.基于特征加权的汉字点笔画生成研究[J].北京大学学报(自然科学版),2014,50(1):153-160
Li Qing-sheng, Xiong Jing, Wu Qin-xia, et al. Study of Feature Weighted-based Generation Method for Dian Strokes of Chinese Character[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 153-160

[14] 吴琴霞,栗青生,高峰.基于语义构件的甲骨文字库自动生成技术研究[J].北京大学学报(自然科学版),2014,50(1):161-166
Wu Qin-xia, Li Qing-sheng, Gao Feng. Study on the Technique of Automatic Generation of Oracle Characters Based on Semantic Component [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 161-166

[15] 吴琴霞,栗青生.基于动态描述库的汉字字形自动生成技术研[J].科学技术与工程,2013,13(15):4425-4431
Wu Qin-xia, Li Qing-sheng. Research of Chinese Character Auto-generation Technology Based on Dynamic Description Library [J]. Science Technology and Engineering, 2013, 13(15): 4425-4431

[16] 栗青生,吴琴霞,杨玉星.甲骨文字形动态描述库及其字形生成技术研究[J].北京大学学报(自然科学版),2013,49(1):61-67
Li Qing-sheng, Wu Qin-xia, Yang Yu-xing. Dynamic Description Library for Jiaguwen Characters and the Research of the Characters Processing [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(1): 61-67

(上接第 64 页)

表 1 稀疏矩阵特征

稀疏矩阵	维度	非零个数	块数
apache1	80800	542184	28
ted_B	10605	144579	4
bcsstk36	23052	1143140	9
2cubes_sphere	101492	1647264	169
nasasrb	54870	2677324	19
boneS01	127224	5516602	46
ASIC_100ks	99190	578890	169
Zd_Jac6	22835	1711557	8
rajat26	51032	247528	44
cage12	130228	2032536	162
cage11	39082	559722	19
viscoplastic2	32769	381326	13

表 2 填零数目比较

稀疏矩阵	k=2		k=4		k=8	
	优化前	优化后	优化前	优化后	优化前	优化后
apache1	66000	4	160840	56	220568	72
ted_B	4593	1	13737	1	41733	5
bcsstk36	10100	2	47088	16	89652	20
2cubes_sphere	93686	70	245140	248	591312	568
nasasrb	23632	2	108756	24	204460	60
boneS01	57472	10	220026	58	589734	158
ASIC_100ks	73642	84	222414	250	522758	614
Zd_Jac6	13435	3	53711	11	147291	35
rajat26	39386	24	95000	80	298152	152
cage12	286062	30	1013220	88	2629448	192
cage11	41148	0	129326	6	358694	38
viscoplastic2	36720	4	112170	14	324794	42

结束语 本文针对目前面向定制结构的稀疏矩阵分块方法和表示方法的不足,提出了一种稀疏矩阵二维均匀分块方法,并给出了相应的表示方法。实验结果表明,本文提出的稀疏矩阵分块方法和表示方法能够有效减少填零个数,将有效开发定制结构的并行性。

参 考 文 献

[1] Dorrance R, Ren F, Marković D. A Scalable Sparse Matrix-Vector Multiplication Kernel for Energy-Efficient Sparse-BLAS on FPGAs[C]// Proceedings of the 2014 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA). ACM, 2014: 161-170

[2] Fowers J, Ovtcharov K, Strauss K, et al. A High Memory Bandwidth FPGA Accelerator for Sparse Matrix-Vector Multiplication[C]// Proceedings of the 2014 IEEE 22nd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). IEEE, 2014: 36-43

[3] Zhuo L, Prasanna V K. Sparse Matrix-Vector Multiplication on FPGAs[C]// Proceedings of the 13th ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA). ACM, 2005: 63-74

[4] Sun J, Peterson G D, Storaasli O O. Sparse Matrix-Vector Multiplication Design on FPGAs[C]// Proceedings of the 15th Annual IEEE Symposium Field-Programmable Custom Computing Machines (FCCM). IEEE, 2007: 349-352

[5] Stathis P T. Sparse Matrix Vector Processing Formats [D]. Delft, Netherlands: Delft University of Technology, 2004

[6] Williams S, Vuduc R, Oliker L, et al. Optimizing Sparse Matrix-Vector Multiply on Emerging Multicore Platforms[J]. Journal of Parallel Computing, 2009, 35(3): 178-194

[7] The University of Florida Sparse Matrix Collection[OL]. http://www.cise.ufl.edu/research/sparse/matrices