西夏文字典《文海》的网络分析*

张光伟 / 陕西师范大学历史文化学院

摘　要：社会网络分析已经成为数字人文领域重要的数据可视化与分析的方法，文章将西夏文字典《文海》转换为字典网络，构建了《文海》中的西夏字解释关系模型。西夏学专家使用人工分析方法在西夏字研究方面已经做了大量的工作，但由于西夏字数量众多、互相之间的解释关系复杂，难以形成完整而清晰的概念。基于西夏人编写的字典《文海》构建的字典网络模型，能够帮助现代人深入了解西夏字型的构造特点，探讨西夏字意义形成的可能性。通过对《文海》字典网络可视化，直观地呈现了字典中解释关系的结构；基于网络中节点的可达性，提出了一种构建西夏基本字集的方法；通过检测《文海》字典网络中的循环和强连接部件，分析了它们在西夏字意义形成中可能的作用；根据《文海》字典网络中西夏字与基本字之间的定义距离，构建了《文海》西夏字解释的层次结构。此研究希望能够为现代人学习和研究西夏文提供一种优化方案，为还原《文海》的编撰思路提供一种技术手段，并为基于信息技术的中国古文字研究提供参考。

关键词：西夏字典　《文海》　网络分析　循环定义　定义距离　解释的层次结构

引　言

西夏文献对于研究西夏王朝存续时期的历史具有独特的重要意义，西夏文是研究这些文献的重要基础，但现在西夏文已经是一种近千年没有使用的“死文字”，而且字型构造复杂，学习和研究难度很大。存世西夏文献以黑水城出土

*本文是教育部人文社会科学研究青年基金项目“基于深度学习的西夏文自动识别系统”（17YJCZH239）的阶段性研究成果。

为主，包含大量的世俗文献和佛经文献，其中的辞书类文献如《文海》和《同音》为后人释读西夏文献提供了关键钥匙。《文海》是西夏人模仿《广韵》的体例编纂的一部西夏文详解字典，每个词条下都包括反切标音、字体构造分析和释义。[①]《文海》主要采用了八种类型的构造用语来解释西夏字：𗍫表示取某字的左边部分，𗏁表示取某字的右边部分，𗖵表示取某字的中间部分，𗰜表示取某字的上边部分，𗢳表示取某字的下边部分，𗀛表示取某字的左边和上边部分，𗦳表示取某字的整体，𗤓表示从某字中去掉一部分。西夏学者认为《文海》可能是西夏人对西夏文字的说明，并不是以西夏文字的构成原理为基准形成的书，但对于今天人们研究西夏文是很重要的线索。《文海》对西夏字的分析比较细致，但由于其缺失序跋等资料，因此难以确定《文海》作者所采用的词典构造方式。[②]如果我们能够深入分析《文海》的解释结构，就有可能解开西夏人编写《文海》时所遵循的规律，让现代人识读西夏文更加有效率。

西夏学专家龚煌城在重建西夏文字衍生过程中没有将西夏文字视为偏旁的排列组合，而是将西夏字看成是由基本字经过一些变换的规则（如“增添”“代换”“对调”等）依次产生出来的。他认为重建西夏文字衍生过程的第一个步骤是发现基本字，并提出了确定基本字的方法，在此基础上通过若干例子验证其提出的西夏字衍生过程理论，[③]但没有给出完整的西夏基本字集，这可能主要受限于人工分析的困难。西田龙雄认为研究西夏文字“首先必须发现、整理文字相互之间的联合关系。……西夏文字中，有一群基本的文字，以这些文字为中心，从这些基本文字开始，用某种步骤，联合若干派生出来的派生文字，这就是顺其自然的事了”，而且他认为这些基本文字一定为数不少，[④]这一点在本文第三部分提取的西夏文基本字集合可以得到验证。西田龙雄虽然根据其西夏基本字的方法对西夏字的构造进行了经验分析，但也提到思考西夏人当时如何考虑西夏文字的构成是有必要的。韩小忙也认为西夏字是由基本字衍生出其他西夏字，在其《西夏文的造字模式》中对西夏文字典《文海》中存在的字型解释的模式进行了深入研究，特别是对西夏文中字型的解释方面进行了全面的总结和梳理，提出从循环解说中提取基本字的方法。[⑤]上述学者基于西夏文基本字研究西夏字的构造以及衍生过程的理论是本研究的重要启发，本研究也希望为上述理论提供一种技术验证方法。

---

①桑明裘：《我国西夏文辞书〈文海〉及其他》，《辞书研究》1980年第1期。
②史金波：《新见西夏文偏旁部首和草书刻本文献考释》，《民族语文》2017年第2期。
③龚煌城：《西夏文字衍生过程的重建》，《“国立”政治大学边政研究所年报》1984年第15期。
④西田龙雄、鲁忠慧：《西夏文字的分析》，《西夏研究》2012年第2期。
⑤韩小忙：《西夏文的造字模式》，北京：中国社会科学出版社，2016年，第13—99页。

《文海》中主要采用的四字解释模式：各取两个字的一部分组合起来解释另一个字，如“靴”被解释为“腿左蔽右”，即“腿”的左边部首加上“蔽”的右边部首构成了“靴”。这种字型结合字义的解释模式对于当时的人识读西夏字应该提供了重要的帮助，但是其中存在的循环解释在学者寻找字型根源方面造成了很大的困难。韩小忙在《西夏文的造字模式》一书中对此进行了细致梳理，列举出《文海》中大量的循环解释。西夏字数量庞大，共有约5,900字，《文海》等西夏辞书中包含约4,090个字；[①]《文海》中多采用每个西夏字由另外两个西夏字来解释，这样字与字之间的解释关系就有近一万条，而且它们之间互相交织构成了一个巨大的关系网络，人工梳理难度可想而知。

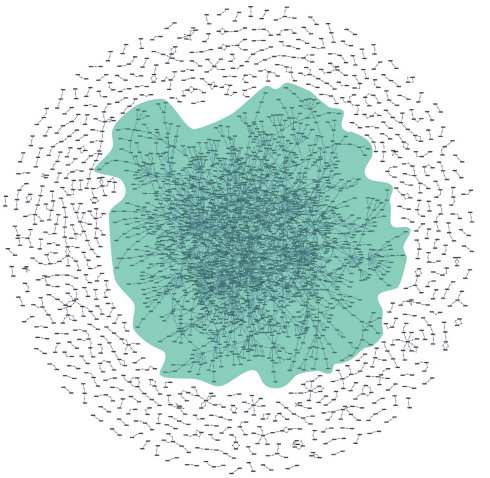

图 1a　《文海》西夏字定义的整体网络结构

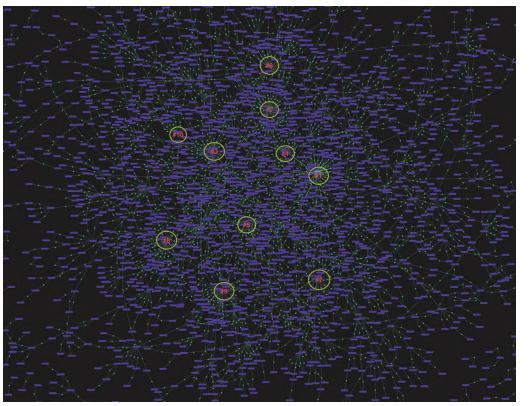

图 1b　重要节点示意

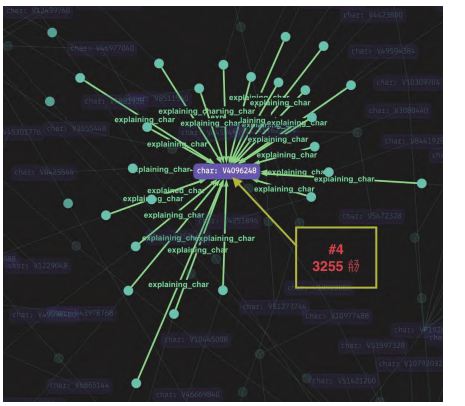

图 1c　重要节点细部

①韩小忙：《西夏文的造字模式》，第7页。

我们构建的《文海》字型解释网络如图1所示，可以直观地看出《文海》中西夏字的互相解释关系的复杂，特别是图的绿色区域的节点构成了一个非常复杂的网络。该图中的节点为《文海》中的西夏字，每个节点关联另外两个节点形成图中的两条边；边表示《文海》中西夏字的解释关系，即边的一个节点代表的西夏字解释另一个节点所代表的西夏字。这种规模的信息量再加上西夏字型的复杂，即便是资深的西夏学专家，人工分析其结构和规律的难度也是非常大的。

本研究将《文海》中每个字的定义提取出来构建了一个字典网络对《文海》的西夏字解释关系进行建模。在此基础上我们主要做了以下工作：（1）利用图算法找出《文海》的字型解释中循环的方法，供西夏语言文字学者进行下一步的深入研究；（2）根据网络中节点的可达关系提取出《文海》的基本字集合，根据这个集合可以推导出所有剩余的西夏字；（3）根据《文海》字典网络中所有字与基本字集之间的“定义距离”构建了《文海》中解释西夏字的层次结构。我们构建的西夏基本字集和解释的层次结构有可能为现代人学习西夏文提供一种较为高效的策略，也有可能为西夏学专家推断西夏人创造西夏字或者西夏人学习西夏字的思维特点提供一种工具。

一、《文海》字典网络

本部分介绍《文海》字典网络的构造方法，基于网络可视化方法的《文海》字典网络结构的视觉呈现，以及基于《文海》字典网络的西夏基本字集的提取方法。

（一）字典网络的构造

西夏文字典《文海》中的定义多采用“四字解释”模式，即分别取两个字（构字）的一部分用于解释一个字（被解释字）（这里的被解释字和构字采用韩小忙的命名方法），例如A是由B的左边部分和C的右边部分组成，表示为A：B左C右。表1列出了《文海》中字型解释的几个例子，可以看出这些解释不仅仅是从字型方面来进行文字的解读，从汉译词条来看，构字的选择和被解释字在意思方面也相近，因此我们可以假设：如果能够识读一个字的解释中的所有构字，那么被解释字的意思是可以推导出来的。[①] 这样如果能找到一个西夏字的基本字集合，根据字典定义就可以推出其他的所有西夏字，这样现代人在学习西夏文时可以优先掌握这些基本的西夏字，然后借助《文海》高效地掌握所有其他西夏字。

①Philippe Vincent-Lamarre et al., "The Latent Structure of Dictionaries," *Topics in Cognitive Science*, vol. 8, no. 3, 2016, pp. 625-659.

表1 《文海》字型解释举例

西夏字	《文海》解释	汉译
[illegible]	[illegible]	愚：昏围笨右
[illegible]	[illegible]	垂：挂下下右
[illegible]	[illegible]	地狱：鬼右地左
[illegible]	[illegible]	靴：腿左蔽右

如前文所述《文海》中主要包含八种构字解释关系，但我们在构建《文海》字典网络时对于这些关系并不作区分，将其合并为一种关系，即解释关系：被解释的字和用于解释的字（构字）之间的关系。我们构建的《文海》字典网络是一个有向图，每一条边的方向都是从构字指向被解释字，这样上述解释关系的例子中的ABC三个字在字典网络中就形成了两条有向边B→A和C→A，表示A的解释中有B和C；当然ABC都可能出现在其他字的解释中，即它们都可能指向其他的节点。根据这种解释关系,《文海》中所有的西夏字就构成了一个网络，即字典网络，网络中的节点是西夏字，边为解释关系。通过分析节点的可达关系以及网络中的循环关系我们可以找出西夏字中的基本字；通过计算所有西夏字到基本字集合的距离，我们能够分析《文海》中西夏字解释关系的层次结构，从而有可能为现代人学习西夏文提供一条优化的学习路径。

下面我们简要介绍一下《文海》字典网络的构建（由于网络分析广泛使用图理论和图算法，因此本文中图和网络的含义是等价的）。

字典网络是一个有向图$G=(V,E)$，其中G表示字典网络或图，V表示图中的节点集合，一个节点对应字典中的一个字，E表示节点之间构成的边的集合。(v_1,v_2)表示图G中的一条有向边$(v_1,v_2)\in E$，v_1和v_2为该边的两个节点，表示字典中的两个字，边的方向为从v_1指向v_2，表示字典定义中v_1出现在v_2的定义中；图G中可能有多条有向边指向v_2，因为字典中一个字往往由多个字来解释；v_1除了有指向v_2的边还可能有多条指向其他节点的边，因为一个字可能会出现在多个字的解释中。所以，一个简单的解释关系就让几千个西夏字组成图1所示的庞大而且错综复杂的网络，这也是复杂网络的共同特点。

我们构建《文海》字典网络时，主要选择了《文海》中的完整的四字解释词条，即每个西夏字都是由另外两个西夏字解释，因此网络中每个节点的入度都为2，如表2所示。其中，target_id和source_id列所对应的数字是西夏字的字典序号，target和source列对应的是具体的西夏字；每两行表示《文海》中的一个解释条目，每一行表示字典网络中的一条有向边，方向都是从source节点指向

target节点。这种格式的数据能够直接使用网络分析软件Gephi进行可视化，如图1所示。我们构建的《文海》字典网络总共包括3,781个节点（西夏字），其中的绿色区域包含2,970个节点，是本文主要研究的对象，其余区域包含811个节点，后续除了可视化之外网络分析部分不涉及这811个节点。这两部分的分离主要是因为（1）现存的《文海》字典是不完整的，有些词条的解释缺失;（2）本文主要关注的解释模式为占字典主要部分的“四字解释”词条，还有一小部分不符合这种模式。《文海》字典网络中基本字集所代表的2,970个节点部分，虽然节点数少了一些，但网络结构完整，能够集中反映该字典文字定义的模式，研究结论和方法能够推广应用于分析其他语言的字典定义模式。

表2 《文海》字典网络的边举例

target_id	target	source_id	source
3909	[illegible]	3911	[illegible]
3909	[illegible]	2653	[illegible]
3911	[illegible]	3909	[illegible]
3911	[illegible]	2262	[illegible]
1059	[illegible]	3287	[illegible]
1059	[illegible]	0377	[illegible]
2678	[illegible]	3255	[illegible]
2678	[illegible]	3911	[illegible]

（二）字典网络的可视化

复杂网络分析的数据规模往往比较大，数据之间的关系也比较复杂，虽然使用网络对数据进行了建模，但是如果不能以一种直观的方式呈现出来，我们也很难快速把握数据的本质。复杂网络的可视化将网络呈现在二维或三维空间中，为我们快速理解一个复杂网络的结构提供了高效的方法，与专业经验相结合有可能引出与研究问题相对更为明确的探索方向。网络可视化可以根据节点的出度值或入度值设置节点的大小，从而从一个侧面反映节点的重要性。我们需要根据研究目的以及网络结构的不同，选择合适的网络可视化布局，从而直观地呈现网络结构。有很多专门进行网络可视化布局的算法，其中将节点模拟为物理粒子之间引力关系的力导向图（Force-Directed Graph），如Kamada Kawai、Fruchterman Reingold等算法，在检测和呈现网络中的社团（Community）比较有效；Circle Pack Layout[①]是一种能够较直观地反映较大规模数据集合层次关系的网络布局形

①Weixin Wang et al., “Visualization of Large Hierarchical Data by Circle Packing,” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2006, pp. 517-520.

式，它能够将网络中的节点聚类。我们主要使用这两种网络可视化布局，除此之外还有很多不同的可视化布局算法，这里不做更多介绍。

网络可视化可以借助专门的网络可视化分析软件，如Gephi[①]、Pajek[②]等实现，也可以使用Python、R等通用编程语言的网络分析包实现，图1和图2是在Gephi中分别使用Fruchterman Reingold和Circle Pack Layout布局呈现的结果。专门的网络可视化分析软件的使用相对比较容易，但功能相对固定，不够灵活；编程实现网络分析和可视化，相对比较复杂，但灵活性高。我们在进行《文海》字典的网络构建与分析过程中，这两种方法都有所应用。

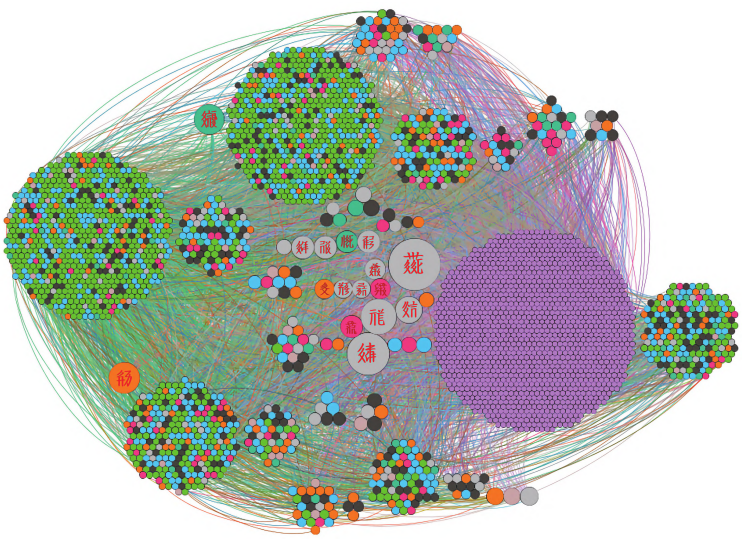

图2 《文海》字典网络层次结构可视化

我们首先使用网络可视化方法将《文海》字典网络的结构展示出来，如图1所示。该图呈现的《文海》字典网络的结构让我们对其大致的结构有所了解，而且中间的绿色区域是我们重点研究的部分。在大致了解了《文海》字典网络的整体结构之后，我们使用Circle Pack Layout布局方式按照节点的出度对《文海》字典网络中的节点层次结构进行了可视化，如图2所示，该可视化方法对比较频繁出现在其他字的解释中的字所代表的节点进行了标记，即图中较大的节点，共16个，这些字都是西夏学专家所认可的基本字。这也初步表明根据网络中节点

①Gephi-The Open Graph Viz Platform, https://gephi.org/.

②Andrej Mrvar, Vladimir Batagelj, "Analysis and Visualization of Large Networks with Program Package Pajek," *Complex Adaptive Systems Modeling*, vol. 4, no. 1, 2016, pp. 1-8.

的出度来衡量所表示的西夏字的重要性与西夏学专家传统的经验式研究是大致相符的。图中节点越大表示该节点对应的字出现在其他字的解释中的次数越多；节点的颜色表示该节点在网络中的离心率（Eccentricity）：一个节点作为起点到网络中其他节点最长的距离。离心率作为衡量网络中心度的一个指标，在字典网络中可以用来衡量一个字在整个网络的解释链条中传播的距离。节点颜色表示的距离关系为紫色＜绿色＜蓝色＜黑色＜桔黄色＜红色＜灰色，紫色表示距离为0，灰色表示大于8的距离。这个指标与第三部分所提出的字典网络的解释层次结构并不矛盾，从某些节点出发可能会有很长的解释传播路径，但是这条路径大部分节点可能处在解释层次结构的同一个层次内部，参考图4。

这两种不同的网络可视化布局方式对同一个网络的呈现，是对不同方面的侧重，由此可见，网络可视化能够将复杂的数据进行直观呈现，特别是有些网络可视化布局为我们理解数据产生启发。但仅仅可视化还不足以把握网络所反映的数据的内在特点，因此网络分析方法一定要与具体的研究深度融合：我们需要将网络分析的一些关键指标直接或者将不同的指标组合起来与具体的研究问题相关联，如西夏基本字集用《文海》字典网络的核心节点集描述，《文海》的解释结构层次用西夏字到基本字集的定义距离来表示等。这样基于网络分析的研究方法能够真正落实到人文研究的问题上，从而能够通过网络分析为人文问题的解决提供证据。

（三）西夏文基本字集的构建

学习任何一门外语时，该语种的词典都是非常关键的，但是词典中收录的词条数量大，利用网络分析或许可以找到一个优化的单词学习策略。1936年，英国语言学家奥格登（C. K. Ogden）经过统计分析列出了850个单词作为英语的基础词汇，使用它们几乎可以表达所有其他复杂的词和概念。[①]西夏文作为一种死文字，对于现代人来说相当于一门外语，因此在学习西夏文的过程中如果能够从一些基本字开始，学习的效率也能够得到优化。西夏学者如龚煌城、西田龙雄、韩小忙等从西夏字的构造角度提出了一些确定西夏基本字的方法，本文借助网络分析技术以西夏人编写的字典《文海》为依据，构建一个西夏基本字集，以期为西夏文的学习和研究提供一种优化方案建议，也有可能为西夏文字专家提供一种分析西夏人创造或学习西夏字时所遵循原则的一种工具。

①Camilo Garrido, Claudio Gutierrez, "Dictionaries as Networks: Identifying the Graph Structure of Ogden's Basic English," Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 3565-3576; COGNITIVE 2018: The Tenth International Conference on Advanced Cognitive Technologies and Applications, 2018, Barcelona, Spain, pp. 18-23.

集合U是字典网络G的节点集合V的一个子集，$R^l(U)$定义为根据字典定义从集合U能够直接推导出的集合，$R^l(U)$表示经过i次推导获得的集合。如果有一个整数k，对于所有l，如果$l>k$，都有$R^l(U)=R^k(U)$，那么$R^k(U)$就是集合的可达集，即从集合U的元素出发能够推导出的所有元素的集合，记为$R^*(U)$。如果$R^*(U)=V$，则表示U是V的一个核心集合，即根据U和字典定义我们可以推导出字典包含的所有字，这是我们根据《文海》定义寻找西夏基本字集合所遵循的原则。

代码1　删除出度为0的节点（Python）

```
1 out_degrees = G.out_degree()
2 od0 = [x for x in out_degrees if x[1]==0]
3 for n in od0:
4 G.remove_node(n[0])
```

《文海》字典基本字集的构造原则：任何字典中都有一些字没有出现在任何其他字的定义中，将这些字从网络中去掉，不影响字典的基本字集合，这是因为出度为0的字不用于解释其他任何字，而且可以由基本字推导出来。所以，我们可以不断重复以下操作：删除当前字典网络中出度为0的节点，直到网络中没有节点能够继续被删除时，我们就找到了字典网络的基本字集。代码1中显示的是一次迭代所做的工作，第1行计算字典网络所有节点的出度。第2行找出所有出度为0的节点。第3—4行将出度为0的节点从网络中删除。经过一次迭代之后，原来出度为1的节点中：有一部分，因为由其导出的子节点被删掉，其出度变为0，因此需要继续上述过程，将当前状态下出度为0的节点继续删除；另一部分节点由于循环的存在其导出节点并不会被删除，因此这些节点的出度不变。经过数轮迭代，我们能够得到一个网络，其中的每一个节点的出度都不为0，最后这个网络的节点所代表的西夏字我们可以认为就是唯一的基本字集。

实验中，我们构建的《文海》字典网络经过六次上述的迭代之后剩余的网络中不存在出度为0的节点，迭代结束，节点数量从最初的2,970精简为1,389，即本文构建的《文海》字典网络基本字集包含1,389个西夏字，占总体的比例为46.8%（注意：不同的网络结构所需的迭代次数不同）。我们将《文海》基本字与英文词典网络提取的基本词进行比较：英文词典的核心词汇占总体的比例一般在10%左右，如*Longman*和*Cambridge*分别为8%和7%，《韦氏词典》12%；[①]《文海》基本字占总字数的比例要比英文词典明显高很多，这主要是由于《文海》中定义

①Olivier Picard et al., "Hierarchies in Dictionary Definition Space," arXiv:0911.5703 [cs], 2009.

一个西夏字多仅仅使用另外两个西夏字，因此表示整个西夏语集合所需的基本字相对较多，而且现存的《文海》不完整，本文研究的是其结构相对完整的部分。

二、《文海》字典网络分析

本部分主要基于《文海》字典网络介绍《文海》中循环解释的检测，《文海》字典网络中强连接部件的检测以及《文海》中西夏字解释层次结构的构建。

（一）《文海》中循环解释的检测

如果我们放大前面可视化呈现的《文海》字典网络图，可以看出其中有一些节点之间是存在循环的，但肉眼很难将所有的循环全部找出，但是寻找网络中的循环是一个经典问题，有一些高效的算法。

字典中的循环解释指的是字A的“构字”中有B，而且B的“构字”中有A，即在字典网络中可以表示为A←→B，这种循环解释是直接循环，相对比较容易找出；还有一些路径更长的循环解释，例如A→B→C→D→A，这样就构成了间接循环解释。单语种字典中的循环解释是非常普遍的，甚至是不可避免的，循环对于一个字典中字意的形成有其独特的意义，这在国外一些单语种字典的研究中已经得到证明，[①]本文对循环解释在字典中的功能暂不做深入讨论，而只是关注与使用网络分析方法快速高效地找出《文海》中所有的循环解释。《文海》中的循环解释很多，西夏文字专家在进行西夏字型研究中从《文海》找到了若干直接循环和间接循环，[②]但这些工作多数由西夏学专家手工进行处理，由于字典中包含西夏字数量众多，字之间的解释关系网络非常复杂，这远远超出人工可以完成的规模，因此手工的方法很难找到《文海》中所有的直接循环解释和间接循环解释。

网络分析算法在计算机领域已经算是比较成熟的，很多问题都有经典的解决算法，也有一些打包好的工具供人文学者使用。我们使用Python中的网络分析包NetworkX查找《文海》字典网络中包含的循环解释，从构建的《文海》字典中找到的循环解释进行了可视化，如图3所示，左边节点用西夏字表示能够看出循环解释中涉及的西夏字，右边不显示西夏字，着重呈现该字典网络中循环解释的结构特点。从这一点也可以看出信息技术能够为人文学者的研究打造基本的实验研究工具和方法，借助这些工具人文学者就有了自己的“望远镜”“显微镜”，甚至是“冷冻电镜”。

---

①David Levary et al., “Loops and Self-Reference in the Construction of Dictionaries,” *Physical Review X*, vol. 2, no. 3, 2012, DOI:10.1103/PhysRevX.2.031018.

②韩小忙：《西夏文的造字模式》，第17页。

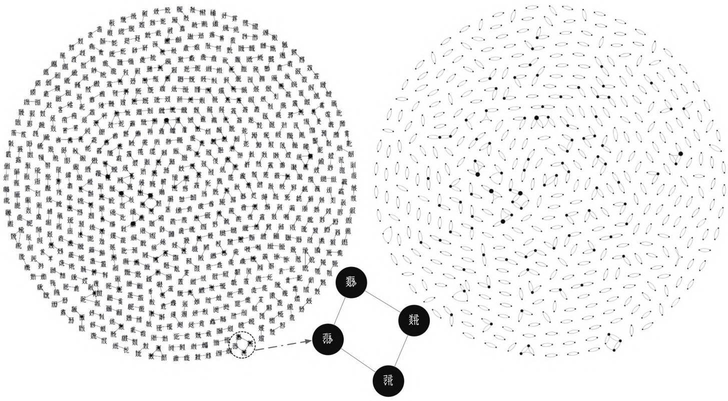

图3 《文海》中的循环解释
（左右两图中节点的大小反应节点的度数）

图3展示的《文海》中的循环解释数量比较多，总共472个，而且比较复杂，循环解释有可能共同参与某些字意义的形成；[①]字典网络中的强连接部件虽然也构成循环，但不完全是前面介绍的循环解释，而可能是多个单独的循环解释相互关联形成，因此强连接部件对于探讨《文海》所体现西夏字义形成原则的可能具有更大的作用。

（二）《文海》字典网络中的强连接部件的检测

强连接部件（Strongly Connected Components, SCCs）是网络中可以相互到达的全部节点的集合，一个网络中可能会有多个强连接部件，查找网络中的强连接部件有比较成熟的算法。我们在查找《文海》字典网络的强连接部件时只针对《文海》的基本字集合构成的网络。

表3 《文海》基本字网络强连接部件（SCCs）举例

六字（1组）		五字（1组）		四字（17组）	
西夏字	中文释义	西夏字	中文释义	西夏字	中文释义
[illegible]	子、幼子、孩童	[illegible]	抽、取	[illegible]	堪、胜、能、可
[illegible]	孩子、小儿、男	[illegible]	摘去、拔去	[illegible]	胜
[illegible]	男、子	[illegible]	扯、摘去、拔去	[illegible]	强、能、胜
[illegible]	生、产	[illegible]	抽、拔	[illegible]	胜、能
[illegible]	安静、徐徐	[illegible]	抽、拔、取		
[illegible]	安详、徐徐				

①David Levary et al., “Loops and Self-Reference in the Construction of Dictionaries,” 2012.

在《文海》1,389个基本字构成的网络中，存在371个强连接部件，其中最大的强连接部件只有一组，包括六个节点，即这六个西夏字在解释关系方面是能够互通的，如表3所示，从其中文释义来看，六字SCC构成了一个“男子、小孩”相关的语义（后面两个表示“安静、徐徐”的西夏字在语义构成中的作用我们留在后续的研究中解决）；五字SCC也只有一组，这些字构成了“抽、拔”动作相关的语义；四字SCC共有17组，表中给出的四字SCC构成了“胜、强”相关的语义。

（三）《文海》中西夏字解释层次结构的构建

《文海》基本字集能够定义所有其余的字，我们通过计算《文海》中每个字距离基本字集的定义路径长度，来确定《文海》中西夏字的解释层次结构。我们首先提出一个概念：定义距离，记为$dis(w)$，其中w表示《文海》中的任意一个西夏字，定义距离表示为:《文海》中的任何一个字以基本字集中的某个字为起点，通过字典中词条的定义，到达该西夏字需要经过的推导次数（即网络中，从基本字集合中的某字到该字的路径长度）。这样，如果该字包含在基本字集中，则$dis(w)=0$；如果是从基本字经过一步就可以将w推导出来，即该字的所有“构字”都在基本字集中，则$dis(w)=1$，否则$dis(w)=\max\{dis(v)\}+1$，其中v表示出现在w的解释中的所有字（《文海》中为两个“构字”）。由此可见该定义距离是一个递归定义。

由上述定义距离原则构建的《文海》西夏字解释层次结构如图4所示，图中表示了三个层次结构的示例:（1）中心黄色区域表示字典网络的基本字集，包含若干强连接部件，其中的西夏字（空白圆圈）的定义距离$dis(w)$皆为0;（2）浅蓝色区域中的西夏字（蓝色圆点）的定义距离$dis(w)$皆为1;（3）浅绿色区域中的西夏字（绿色圆点）的定义距离$dis(w)$皆为2。注意绿色圆点代表的处在第三层次的西夏字的“构字”中至少有一个是来自第二层次（蓝色圆点表示），也可能有来自基本字集的“构字”，如图中标出的1、2、3条解释边所示，由此可见：一个字所处的层次是根据其所有“构字”中所处的最大

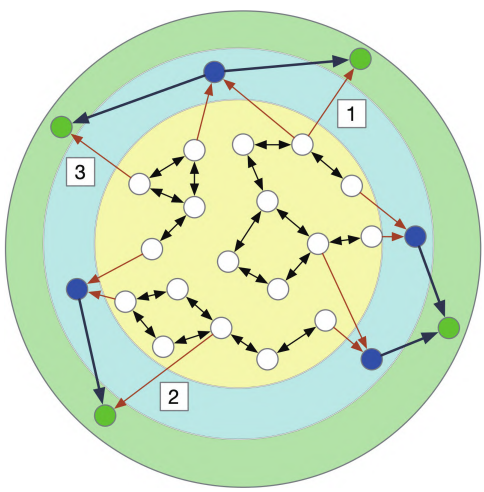

图4 《文海》字典解释层次结构示意图

表4　不同定义层次的西夏字数量

定义距离	字数
1	1,141
2	346
3	70
4	20
5	4

的层次+1得到。我们计算《文海》中每个字的定义距离，在基本字集以外，每一层次的西夏字的数量如表4所示。

通过表4的数据我们可以看出，整个《文海》的西夏字根据定义距离可以分为6个层次：第0层为基本字共1,389个字，第1—5层共1,581个字，合计2,970个字，对应图1中绿色区域覆盖的节点总数。从得到的西夏字解释网络的层次结构来看，多数西夏字集中在基本字集和第一、二层次，距离基本字集较远的层次上的字数相对比较少。虽然本文分析的数据是《文海》中西夏字的一部分，但这部分字解释关系完整，能够体现《文海》中西夏字解释的特点。

结　语

网络分析，特别是社会网络分析，在数字人文领域已经开始发挥重要的作用，人文学者对于网络分析方法的应用也越来越重视。本文以西夏文《文海》字典网络的构建与应用为例讨论了基于网络分析技术的循环解释结构以及强连接部件的检测方法，以《文海》字典的字型解释关系构建西夏基本字集的方法，以及根据《文海》解释中西夏字与基本字之间的定义距离构建《文海》西夏字解释的层次结构的方法。本研究希望能够为现代人学习和研究西夏文提供一种优化方案，能够为还原《文海》的编撰思路提供一种技术手段的支持。

《文海》中的字型解释关系在一定程度上能够反映当时的人们在学习西夏文时所遵循的大致顺序，但是这与当时实际生活中的应用可能存在区别，而这些可能从西夏文历史文献的用字方面有所体现，因此，如果我们借助现代人工智能方法将西夏文世俗文献进行分析，与《文海》字典网络的研究结论互相印证和补充，将能够更加准确地反映西夏字语义的形成与衍生过程，也有可能为基于信息技术的中国其他古文字的研究提供参考。

Network Analysis of the Tangut Dictionary *Wenhai*

Zhang Guangwei

Abstract: Social network analysis has become an important data visualization and analysis method in digital humanities. This paper constructed a definition model of Tangut characters in the Tangut dictionary *Wenhai* by converting it into a dictionary network. Tangut scholars have done a lot of research on Tangut character patterns using manual analysis methods. However, it is not easy to form a widely accepted interpretation due to the massive number of Tangut characters and the complex definition relationships. The dictionary network constructed based on *Wenhai* written by the Tangut people can help us gain insight into the glyph characteristics and the origin of the meaning of the Tangut characters. We showed the definition structure of *Wenhai* with the visualization of the dictionary network. We proposed a method for constructing the core set of the Tangut characters based on the reachability of nodes in the dictionary network. We detected the loops and strong connected components in the *Wenhai* dictionary network and analyzed their roles in the meaning formation of Tangut characters. We constructed the definition hierarchies in *Wenhai* based on the definition distance from the core set to the Tangut characters. We hope to provide an optimized solution for modern people to study Tangut characters and a technical means to support the restoration of *Wenhai*. This method might be helpful for the study of ancient Chinese characters based on information technology.

Keywords: Tangut Dictionary; *Wenhai*; Network Analysis; Definition Loop; Definition Distance; Definition Hierarchies

（编辑：许可）