文章编号：1003-0077（2017）00-0000-00

基于搜索数据的生僻字知识库构建和应用

刘廷超[1] 王雨[1] 饶高琦[2] 杨兆勇[1] 荀恩东[3]

（1. 北京语言大学 信息科学学院，北京市 100083；
2. 北京语言大学 国际中文教育研究院，北京市 100083；
3. 北京语言大学 语言资源高精尖创新中心，北京市 100083）

摘要：本文探讨了生僻字的识别与应用问题。首先通过对搜索引擎日志数据的分析，将用户搜索生僻字的行为划分成直接搜索和拆字搜索两个类别。然后制定生僻字拆分规则，生成生僻字表，结合相关信息构建生僻字知识库。最后讨论了知识库的应用，包括构词领域分布、历时统计分析以及汉字拆分检索等方面。尽管面临样本数据有限和拆分歧义性等挑战，但本文的方法依然为生僻字的研究提供了新的视角和工具。这些成果为相关领域的进一步研究奠定了基础，并将为汉字文化的传承与发展起到了积极的推动作用。

关键词：生僻字，搜索关键字，汉字拆分

中图分类号：TP391 **文献标识码：**A

Construction and Application of Rare Character Knowledge Base Based on Search Queries

Tingchao Liu[1], Yu Wang[1], Gaoqi Rao[2], Zhaoyong Yang[1], Endong Xun[3]

(1. School of Information Science, Beijing Language and Culture University, Beijing, 100083, China;
2. Research Institution of Chinese International Education, Beijing Language and Culture University, Beijing, 100083, China;
3. Advanced Research Center of Language Resources, Beijing Language and Culture University, Beijing, 100083, China)

Abstract: This paper discusses the recognition and application of rare characters. Firstly, through the analysis of search engine log data, the user's search behavior for rare characters is divided into two categories: direct search and split search. Then, the rules for splitting rare characters are formulated, a list of rare characters is generated, and a knowledge base of rare characters is constructed based on relevant information. Finally, the application of knowledge base is discussed, including the distribution of word formation domains, diachronic statistical analysis, and Chinese character split retrieval. Despite the challenges of limited sample data and dismantling of different meanings, the method in this paper still provides new perspectives and tools for the study of rare characters. These achievements have laid a foundation for further research in related fields, and will play a positive role in promoting the inheritance and development of Chinese character culture.

Key words: Rare characters, Search queries, Split Chinese characters

0 引言

生僻字，又称冷僻字，指不常见的或不熟悉的文字[1]。生僻字是一个值得注意的字用现象，但这方面的研究一直关注度不足[2]。自有汉字以来，汉字也在不断发展中经历了“适者生存”的淘汰式选择和变化过程[3]。在历史上，生僻字也可能是常用字，生僻字的范围，是不断发展和变化的[4]。有些字虽然在一些时代、地区或领域比较常见，但

收稿收稿日期：20XX- XX-XX；**定稿日期**：20XX- XX-XX
基金项目：国家语委科研项目“基于大模型的国际中文教学多模态资源构建（YB145-88）”

不在相关环境中的用户并不能准确的掌握其读音和含义。

学界对生僻字还没有一个确切的定义，只有相对共识的理解，生僻是相对“常用”“通用”而言[5]。生僻字有两个典型特征：首先是这些字不常见，不为人们所熟悉，其次是生僻字是相对于常用字而言，具有相对性[2]。结合这些特征，本文对生僻字的定义为：人们在日常生活中偶尔会遇到，但不能够准确识读的汉字。

生僻字是我国汉字文学的组成部分，具有深厚的文化意蕴，是从最早的四千多个甲骨文到如今八万多字的发展过程中产生的[6]。尽管汉字总数庞大，但自古以来，常用的汉字数量保持相对稳定，大致在五千到六千字之间。由于地域和文化差异，某些汉字仅在特定地区或人群中广泛使用，特别是在人名和地名中更为常见。这种现象导致了汉字掌握和使用的不平衡性；例如，“临沂”和“曲阜”这样的地名，当地人耳熟能详，对外地人而言却可能构成挑战。同样，南方地区常用的地理名词如“滇池”和“盱眙”，对于北方人来说则相对较为陌生。如果不知道一个字的正确读音以及用法和含义，就会极大地阻碍文化交流与文化传播。

当用户遇到生僻字时，需要查找其读音及用法和含义，但如何输入和检索这些生僻字成为一个难题。此外，如何界定哪些字属于生僻字及用户遇到这个生僻字的频率也是挑战之一。

本文使用搜索引擎的日志数据，采取量化分析的方法，首先研究用户在搜索生僻字时所使用的搜索字符串的构成规律，提取出日志中用于搜索生僻字的搜索字符串；接着对这些搜索字符串进行分类，并依据不同类型的搜索字符串特点，构建起搜索字符串与目标生僻字之间的映射关系；随后，将搜索字符串转化为具体的生僻字，并统计这些生僻字的通用度。通过参考《通用规范汉字表》[6]，得到生僻字在字表中的分布情况，进而构建生僻字知识库；最后对知识库进行以下方面加以应用：统计生僻字组成的词语的领域分布，结合 BCC 语料库[7]生成生僻字以及其构成词语的历时分布，构造汉字拆分搜索服务。

生僻字作为汉字文化的重要组成部分，具有丰富的历史文化内涵和独特的应用价值。通过构建生僻字表和建立生僻字知识库，可以更好地理解和应用生僻字，为汉字研究、教育推广以及信息检索等领域提供有力支持。

1 相关研究

关于生僻字有很多学者进行了研究，并将生僻字总共分为八大类，分别为：专有名词、历史词语、文言词语、联绵词语、成语词语、专业词语、方言以及网络用字[4]。由于各个信息系统使用的汉字编码及不同，导致很多自造字的出现，通用系统无法输入和显示这些自造字，导致了生僻字的使用困难重重，影响用户的日常生活[8]。

人名用字经常会使用到生僻字，在对人名生僻字的研究表明人名生僻字的问题十分普遍，即使 GB18030-2000 字符集已经收录了 27484 个汉字，仍然无法解决人名用字的覆盖问题[9]。

地名用字十分稳定，在对 1982 年的《地名大词典》的研究中，一共发现了地名生僻字 164 个，其中含地名生僻字的地名词条共有 349 条[10]。

互联网的普及导致一些罕用字开始流行，网络生僻字流行的一个主要原因是用户希望用单个汉字表达更复杂的含义，比如：“烎”“燚”等汉字在字型上会表达出额外的意义[11]。用户一开始接触网络生僻字时，并不知道它们的读音、意义，只能隐约地感受到它们的大体意思[12]。网络生僻字具有字形象似性、数量象似性、部件会意象似性、标记象似性和合音象似性这些特点，加上用户最求标新立异的心理，导致了一些生僻字的再度流行[13]。

网络生僻字主要来源于地名、古籍以及方言等领域，网络生僻字具有周期性和不稳定性的特点[14]。网络催生了新事物，网民根据象形、会意、形声等造字法创造性的使用生僻字，网络生僻字代表了新技术和新文化的一个发展方向[15]。网络生僻字是网络文化形态下的重要产物，至今不成熟、不完善，需要不断探索、归纳与整理。[16]生僻字的流行有利有弊，对社会存在着双面影响。如果能够把传统文化融入生活中，才是真正的文化承传[17]。

上述研究存在一个共同的问题：他们并未基于实际语料进行统计分析，因此无法直观反映用户在日常生活中遇到生僻字的真实情况。

随着互联网的普及，当用户遇到生僻字时，通常会使用搜索引擎来查找相关信息。对于那些难以输入的汉字，用户往往根据其字形特征构造搜索字符串，例如搜索“四个火念什么”或“日成

怎么读”。通过对这类搜索字符串的日志进行统计分析，可以获取生僻字的实际使用情况。

2 生僻字表构建

本文采用的数据来源于国内一家主流搜索引擎 2015 年 7 月至 2016 年 10 月期间经过脱敏处理的日志数据。数据按日期被划分为多个日志文件，每个文件包含用户输入的搜索字符串及其在当天的搜索次数。

通过分析搜索日志，可以了解到用户检索生僻字的方式。进而利用用户检索生僻字的搜索字符串，结合汉字的部件拆分数据，可以将搜索字符串转换为对应的生僻字。生僻字表的构建流程如图 1 所示。

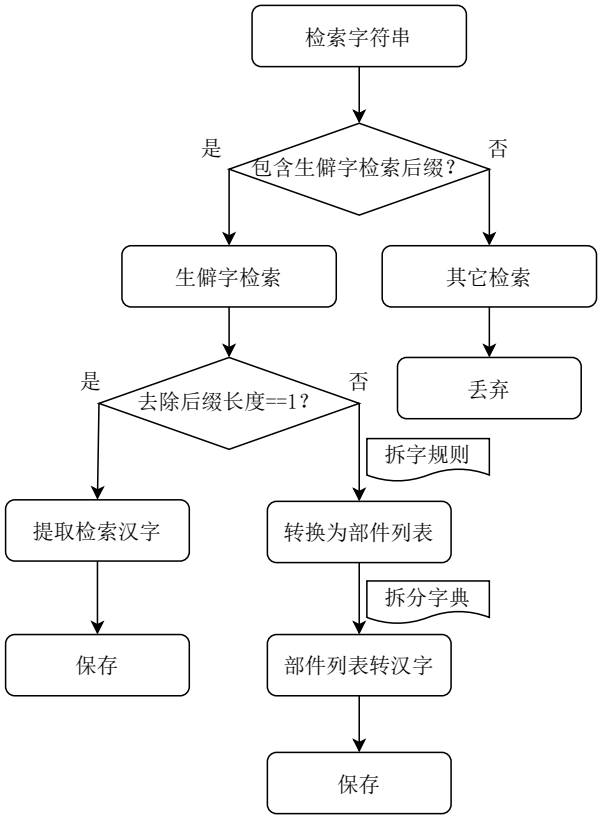


图 1 生僻字表构建流程

2.1 生僻字检索方法

当用户遇到生僻字时，通常会使用搜索引擎来查找其含义和用法。用户的检索行为可以分为两种情况：一种情况是用户可以通过直接输入或者复制粘贴方式查询生僻字，称为直接生僻字搜索；另一种情况是用户无法通过直接输入或者复制粘贴方式查询生僻字，在这种情况下，用户会尝试将汉字拆解成部件列表，并通过描述这些部件来搜索生僻字，称为拆字生僻字搜索。

用户在能够得到生僻字的情况下，经常以待查询的生僻字加上后缀的方式进行查询。以用户检索“堃”字为例，常见的搜索查询包括“堃字怎么读”“堃读什么”“堃念什么”“堃怎么读”等。通过对搜索日志中一天的用户搜索数据按照后缀进行过滤，结果如表 1 所示：

表 1：直接生僻字搜索

搜索字符串	次数
灏读什么	53
闰字读什么	6
斐读什么	6
…	…

然后继续搜索“灏”“闰”等字在日志中出现的其他相关搜索字符串，整理出生僻字搜索常用的后缀列表。其中，“字读什么”“字念什么”“是什么字”“念什么字”“念什么”“读什么”覆盖了绝大部分情况，本文采用这六种后缀作为生僻字搜索字符串的后缀。把搜索字符串中的后缀去掉后，如果只剩下一个汉字，那么这个汉字就是用户搜索的生僻字。

对于拆字生僻字搜索，由于用户无法直接输入生僻字，他们会通过拆字的方式来搜索。以用户搜索“三个火”为例，可以看到搜索后缀与直接生僻字搜索的后缀相同，但拆字部分存在着多种描述方式，如表 2 所示。

表 2：拆字生僻字搜索

搜索字符串	次数
三个火字读什么	160
三个火念什么	156
三个火加一个木是什么字	53
…	…

拆字生僻字搜索的搜索字符串的拆字部分形式多样，需要分析拆字部分的构成方式和特点，整理出生僻字的拆分规则，将搜索字符串还原为汉字部件列表，然后通过汉字部件数据反查出用户搜索的生僻字。

2.2 生僻字拆分规则

拆字生僻字搜索的转换规则比较复杂，用户会根据汉字的拆分结构信息，描述生僻字的组成

部件。

在用户描述待检索生僻字时，会使用数量词表示部件的个数，使用部首名称指代部件，使用简体字代替繁体字，以及使用方位词表示汉字的结构信息等。因此需要制定搜索查询到部件的转换规则，实现搜索字符串拆字部分到汉字构成部件之间的映射。经过对搜索字符串拆字部分的分析整理，总结出以下六条转换规则：

1）**数量词还原规则**。在搜索字符串中经常出现“三个土”“2 个马”等信息，需要对这些数量词进行还原，得到实际的汉字部件。

2）**偏旁部首还原规则**。搜索字符串中包含汉字部件的信息时，用户使用“三点水”“单人旁”等来描述，需要将这些部首描述还原为实际的汉字部件。

3）**结构提取规则**。方位词在搜索字符串中起到描述汉字结构的目的，有些汉字部件相同但是结构不同，如“岭”“岑”，都是由“山”“令”两个部件构成的，因此需要通过方位词给出汉字的结构信息。

4）**对“加”的转换规则**。通常“加”用来连接两个部件，本身并没有实际意义，但在个别情况下“加”也是组成部件。在数量词还原、方位词提取后，如果“加”处于部件列表的首位或者尾部则保留，否则就进行删除处理。

5）**修饰语转换规则**。在搜索字符串中经常包含一些修饰性的词语，如：“合起来”“组成的”等。这些修饰语在检索时并不起作用，遇到的时候直接删除。

6）**对“去除”的转换规则**。除了使用部件组合成字以外，还有一种情况是将一个已知汉字去掉某些部件或笔画变成待搜索的生僻字。例如：“游去掉三点水念什么”，可以转换为：“游”字由“氵”以及什么字组成，因此，可以确定搜索的生僻字为“斿”字。

在搜索日志里，单个搜索条目或许会涵盖众多转换规则。以“三个火加一个木是什么字”为例，需要结合规则 1 与规则 5，以便将其准确还原为部件列表。灵活运用这些转换规则，能够把搜索字符串转变为汉字的部件列表。具体的搜索字符串转换示例，详见表 3。

表 3：拆字转换规则示例

规则	示例	结果
数量词	一个立一个羽 两个方一个土	立羽 方方土
	三个火加一个木	火火火木
部首	三点水一个女 一个单人旁一个吉 草字头下面一个长	氵女 亻吉 艹长
结构	斌下面一个贝 左边一个革右边一个斤	上下结构 左右结构
加	三个火加一个木 上面一个加下面一个可	删除 保留
修饰语	男女男合起来 更生组成的	删除 删除
去除	埠去掉土 演去掉三点水	埠|土 演|氵

通过对上述规则的协同运用，可以对搜索日志中的搜索字符串进行有效转换，生成部件列表及汉字结构信息。以“昱”作为范例，包含在搜索日志中的搜索字符串都能被转化为如“日立”这样的部件列表，同时还可能附带结构信息。详细数据展示于表 4 之中。

表 4：“昱”的转换结果

序号	搜索字符串	部件	结构
1	日下面加个立	日立	上下结构
2	上边日下边立	日立	上下结构
3	日下面立	日立	上下结构
4	日字下面立	日立	上下结构
…	…	…	…
26	一个日字一个立	日立	
27	日下面加立	日立	上下结构
28	一日一立	日立	
29	上面一个日字下面一个立	日立	上下结构

2.3 部件还原生僻字

将部件列表还原为汉字的关键步骤在于找到部件列表与汉字之间的对应关系。本文采用了开放词典[1]项目提供的汉语拆字字典数据，以实现从部件列表到汉字的转换。

汉语拆字字典的拆分规则是尽可能把汉字拆成最大的组成部件。以“灏”字为例，其构成的汉字部件拆分数据呈现为“氵颢”，而“颢”进一步拆分为“景页”，“景”字则拆分为“日京”。因此，在执行部件列表匹配之前，需要通过逐级拆解的方法，将汉字分解成多种可能的部件列表，以便于对同一汉字不同拆分方式的匹配。详细的汉字部件拆分数据，请参见表 5。

[1] https://github.com/kfcd/chaizi

表 5：汉字部件拆分数据

汉字	拆分 1	拆分 2	拆分 3
奮	大明	大日月	
奮	大周		
奮	大朋	大月月	
燚	炎炎	火火火火	
蕊	艹惢	艹心心心	
灪	氵灥	氵景页	氵日京页

鉴于用户通常以简体字输入，而部分生僻字并无对应的简体字，因此设计了一套容错机制，以支持简繁汉字之间的无缝转换。这样，在使用简体字进行搜索时，也能自动将其映射至相应的繁体字，确保检索的全面性和准确性。

此外，考虑到用户在输入某些部件时的表述多样性，例如，当用户键入“四点底”或“四个点”时，实际上意图匹配的是“灬”这一汉字部件。为了响应这类用户输入，增加相应的映射数据，以优化部件识别的精确度。

汉字部件拆分容错数据设计，增强了系统的鲁棒性。详细的汉字部件拆分容错数据，请参见表 6。

表 6：部件拆分容错数据

汉字	拆分	容错	说明
鱻	魚魚魚	鱼鱼鱼	简繁映射
龘	龍龍龍	龙龙龙	简繁映射
驫	馬馬	马马	简繁映射
灬	丶丶丶丶	点点点点	部件映射

经过上述处理，最终建立了直接生僻字搜索以及拆字生僻字搜索的生僻字字表。该字表为生僻字的识别、检索及研究提供了强有力的数据支持。

3 生僻字知识库构建

首先依据《通用规范汉字表》的数据，获取生僻字表中的汉字的等级，揭示了这些生僻字在不同级别中的分布特征。随后，通过对搜索日志中生僻字出现频率的统计，获取了生僻字的通用度信息，这为理解生僻字在现代语言环境中的应用提供了重要依据。

进一步地，对搜索日志中包含生僻字的搜索字符串进行了统计，以此定义了生僻字的常用词汇。为了更全面地构建生僻字知识库，还整合了汉字的拼音数据、构成部件数据，以及搜索日志中的搜索字符串等相关信息。

这一综合性的构造流程不仅确保了生僻字知识库的丰富性和准确性，而且为汉字研究、教育以及信息处理等领域提供了宝贵的数据资源。整个构建流程如图 2 所示，清晰地展示了从生僻字表到知识库构建的过程。

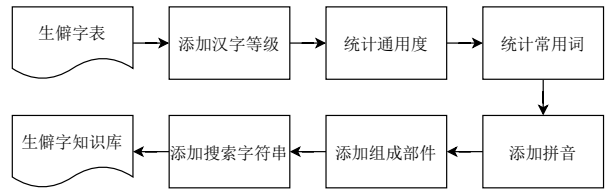


图 2：生僻字知识库构建流程

3.1 获取分布信息

所谓词语通用度，是指词语在语言应用的各个领域里常用性的综合指标。通用度就是词语在语言应用的各个领域里通用的程度。通用度已经兼顾到词语的分布率和频率两个方面，把两者有机地结合起来[18]。

通用度的计算公式如下：

$$T = \frac{(\sqrt{n_1} + \sqrt{n_2} + \cdots + \sqrt{n_k})^2}{k}$$

将按照日期的生僻字的频次信息代入到通用度公式，得到生僻字的通用度信息。通用度大于等于 1 的汉字，直接生僻字搜索结果有 3652 个汉字。前 50 个字的通用度如表 7 所示。

表 7：直接生僻字搜索 TOP50 通用度

阜	3483	阚	3312	丰	2451	靳	2310
衢	1995	羽	1783	翟	1703	鄞	1658
赣	1597	邹	1558	晟	1496	祁	1436
华	1417	虞	1417	闫	1391	濮	1390
沂	1374	睢	1254	绥	1233	力	1194
青	1190	肇	1185	斤	1183	忻	1162
页	1065	郴	1006	堃	1004	煜	986
暨	980	黔	966	日	958	缪	934
毓	921	筱	910	婺	897	綦	890
岑	874	文	846	湛	842	兖	829
昱	821	于	820	茜	811	行	809
芮	806	樊	802	覃	789	龚	787
卜	777	荥	775				

通用度大于等于 1 的汉字，拆字生僻字搜索结果有 2929 个汉字，前 50 个字的通用度如表 8 所示。

表 8：拆字生僻字搜索 TOP50 通用度

垚	130038	燚	71547	堃	60613	焱	56376
昶	53550	昱	45034	犇	43931	赟	42286
翊	37987	燊	36919	靳	35119	晟	32497
鼛	28081	仝	28033	淼	26516	旻	25866
囡	23602	婧	22678	煜	22353	幢	21515
砼	21167	臻	20346	喆	20150	歆	19575
夯	18582	翀	17948	頓	17582	騭	17039
咎	16962	圩	16894	艮	16408	泵	15813
彧	14935	泠	14761	斛	14485	岑	14196
黔	14009	粿	13218	昕	13148	趸	13018
墾	12839	柘	12565	灪	12485	龘	11708
阚	11553	怼	11448	颢	11349	聿	10995
珩	10863	梓	10768				

通过对比直接搜索与拆字搜索的结果，可以发现两者在笔画数量、字形结构以及包含的汉字上存在显著差异。直接搜索得到的生僻字在笔画数和结构上相对简洁，更多的是人名、地名等专有名词；而拆字搜索的结果则显示了更为复杂的笔画组合和结构，其中包括了大量的叠字。

3.2 获取等级信息

自 2013 年发布以来，《通用规范汉字表》已成为汉字规范化的核心参考资料。该字表分为三个层级，共涵盖 8105 个汉字。一级字表奠定了日常交流的基础；二级字表满足了更为广泛的语言应用需求；而三级字表则聚焦于姓氏人名、地名、科学技术术语以及中小学语文教材中的文言文用字。

对通过两种不同搜索策略获取的生僻字进行了字表分布统计。分析结果表明，在直接搜索得到的生僻字中，常用汉字的比例显著高于拆字搜索所得。具体而言，直接搜索的生僻字中，属于一级和二级汉字的比例高达 85.95%，相比之下，拆字搜索的比例则为 69.14%。此外，直接搜索中超出规范字范围的生僻字占比为 10.16%，而拆字搜索则为 18.63%。这些汉字的等级信息对于深入了解生僻字的分布特征极为关键，并将被纳入最终构建的知识库中，作为生僻字属性的重要组成部分。

两种搜索方式中出现频率最高的前 100 个生僻字，有 22 个汉字在两种搜索结果中均出现，包括：“阜臻靳昱晟昇阚赟砼煜覃堃仝尹暨尧昕昶灪黔沂岑”。前 100 生僻字在规范字表中的等级分布情况，详见表 9。

表 9：TOP100 生僻字等级分布

级别	方式	数量	汉字列表
一级	直接	28	钦欠赣华卜辛斤行于矢文页石乐柏羽殷玉今童尧力青日沁黔中肇
	拆字	12	矗尧兢滇邑泵汝寅幢黔酉夯
二级	直接	66	鑫筱阜臻靳芮翟裴睢荥昱殇缪郸晟邵兖龚懋鄞詹岑忻瞿菁牟阚邹暹蔺谌赟倪睿砼煜衢钰虞茜郴覃廖毓樊蓟綦胥褚郝雒邳尹麓暨绥祁昕昶灪濮婺湛栾沂斐
	拆字	62	囡阜銮臻靳疃趸舸岫颉昱累珩柘辊晟厝訾岑忝泺怼阚赟罡珏梓歆彧砼煜毳翊晁姝乜旻喀骅廿覃劼熘淦弋尹轶婧咎槎暨烨颢昕圩昶灪斛艮沂焱聿
三级	直接	5	仝堃羋闫昇
	拆字	15	犇喆燚燊堃仝甦鼛淼狨墾崟昇晔垚
级外	直接	1	尛
	拆字	11	叕騭咲羴灬龘頔鱻爓粿忝

3.3 构建知识库

基于直接生僻字搜索和拆字生僻字搜索两种不同的日志数据，构建了生僻字知识库[2]。直接生僻字知识库涵盖了拼音、常用词语、通用度以及等级等多项内容，而拆字生僻字知识库则在此基础上，增加了组成部件和搜索字符串等内容。

在汉字拼音方面，本文采用了 mozillazg 的 pinyin-data[3]项目所提供的拼音数据。通用度指标则是根据汉字在搜索日志中的出现频率统计得出，反映了生僻字字的普遍程度和使用频率。汉字等级信息则参照《通用规范汉字表》，分为 1 至 3 级，对于不在字表中的生僻字，则统一归为 4 级。

组成部件数据源自汉字拆分数据库，记录了生僻字的拆分部件列表。常用词语则是通过对搜索日志中的相关搜索字符串的统计分析得出，为每个生僻字列出了高频使用的词语。搜索字符串则保存了用户在搜索日志中拆字搜索时的搜索字符串的拆字部分。

[2] https://github.com/paineliu/RCKB

[3] https://github.com/mozillazg/pinyin-data

所有这些信息以 JSON 格式保存为生僻字知识库文件，便于通过编程方式访问和操作，极大地提高了生僻字知识库的实用性和可扩展性。

4 生僻字知识库分析与应用

生僻字知识库是一项宝贵的资源，它不仅提供了生僻字的读音，还包括了反映其通用度的信息、常见的词语、详细的部件拆分数据，以及用户在搜索过程中实际输入的检索字符串等多维度信息。本文将从三个不同的应用场景出发，探讨生僻字知识库的实际应用价值。首先是生僻字常用词语在不同领域的分布情况；其次是生僻字及其构词的历时分布特征；最后是汉字拆分检索的实际应用效果。通过这些应用，可以更全面地理解生僻字知识库在语言研究、教育推广以及信息检索等多个领域的潜在价值。

4.1 知识库特点

生僻字知识库不仅涵盖了汉字的读音、等级等基础信息，还包括常用词、组成部件、搜索字符串在内的多维度数据，使用户能够从多个角度深入理解生僻字的用法和含义。以“昱”字为例，知识库中信息如表 10 所示。

表 10：生僻字知识库数据示例

汉字	昱
拼音	yù
通用度	45034
汉字等级	2
组成部件	日立
常用词语，词语通用度	王昱珩，52143 朱相昱，47549 瑞昱，14756 …，…
搜索字符串	日立 上日下立 …

4.2 生僻字领域分布

生僻字在其应用领域中展现出独特特性，通过对生僻字常用词语的分析，可以揭示其构词能力及其适用的领域。本文根据生僻字知识库中常用词语的类别分布，将生僻字的应用领域细分为姓名、组织机构、地名、网络、方言以及其他类别。

在姓名用字方面，这一类别涵盖了姓氏和人名两个子类。例如，“阚”作为姓氏用字，而“堃”则常见于人名之中。

组织机构用字的范围较为广泛，一些生僻字主要应用于组织机构的命名，如“鬞”“犇”这类字眼虽不常见于个人姓名，却在“鑫鬞大厦”“犇腾牛排”等机构名称中频繁出现。同时，姓名用字和地名用字也常被用于组织机构的命名，例如“垚”字既可用于构成“陆垚”“刘垚昕”等人名，也可用于“安徽垚森”“继垚生态园”等机构名称。

地名用字相对稳定，但由于地域差异导致的用户认知不同，仍有些地名用字被部分用户认为是生僻字，如“兖”“黔”等。地名用字在组织机构命名中也十分常见，如“临沂大学”“五邑大学”等。

网络用字则是随着互联网文化的兴起而重新进入公众视野的罕用字，例如“槑”字在网络用语中的使用，构成了“槑痘痘”、“泵槑孖”、“丄線發槑”等表达方式。方言中的某些汉字也因文化传播而逐渐流行，如“乜”字在方言中的使用，形成了“乜嘢”、“冇乜”等词汇。

此外，还有些汉字无法归属到上述分类中，例如：“斛”，它构成了“石斛”“铁皮石斛”等词语。

4.3 生僻字以及构词的历时分布

为扩展生僻字知识库的应用场景，本文通过利用 BCC 历时语料库[19]系统地整理出生僻字的历时信息。包括生僻字的历时字频统计、成词能力、例词和词频等多维度信息，如图 3 所示。

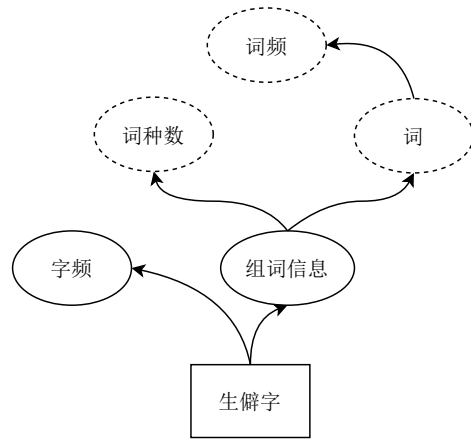


图 3：生僻字历时数据

具体而言，通过设定时间范围（1872 年至 2015 年），在语料库中执行检索算法，从而获得生僻字

在各个时间点的详细信息。这些数据以结构化的形式呈现，便于存储和查询，为后续的数据分析提供了坚实的基础。

获取到的生僻字历时信息在多个方面发挥着重要作用。首先，通过对比不同时间点的数据,揭示了生僻字在历史长河中的演变轨迹；其次，生僻字的历时信息可以构建和完善生僻字知识库，为语言学研究、汉字教学以及文字保护等领域提供支持。此外，生僻字的历时信息还具有一定的应用价值。在信息检索、自然语言处理等领域，了解生僻字的使用情况有助于提高系统的准确性和效率。例如，在搜索引擎中引入生僻字历时信息，可以帮助用户更准确地找到与特定历史时期相关的文献和资料；在自然语言处理任务中，考虑生僻字的使用习惯可以提高文本处理的准确性和流畅度。

4.4 生僻字拆分检索服务

面对用户在使用生僻字时无法直接输入的问题，本文提供的解决方案是直接使用部件信息输入和搜索相关信息。

本文基于知识库提供的生僻字拆分数据，对一个大语言模型进行了微调，专门优化了其支持生僻字拆分检索的能力，旨在简化用户输入和搜索生僻字。

在模型微调时，大语言模型的输入为用户检索生僻字时的搜索字符串，而输出结果则为部件序列，而非直接生成生僻字。这样的设计优势在于，模型可以根据搜索字符串的语义内容进行部件的组合，甚至能够自动识别并处理新的汉字部件，从而显著提升模型的泛化能力和适应性。

具体实施步骤包括：选用 Qwen-7B 模型[20]进行 LoRA（Low-Rank Adaptation）微调[21]。将搜索字符串与相应的部件列表按照 Qwen 微调的格式要求，整理成 JSON 格式的微调数据。为了明确任务的导向，在每条搜索字符串前添加了指令前缀：“请将下文解析成汉字部件列表：”以指导模型准确理解意图。

采用知识库中的 9846 条非重复搜索字符串作为数据集,这些记录可解析为 3023 个不同的汉字构成部件。数据集被划分为训练集（80%，共 7878 条）、验证集（10%，共 984 条）和测试集（10%，共 984 条）。同一汉字构成部件的样本会被随机分配至不同集合，确保各集合间样本的独立性。测试集包含 810 个汉字构成部件样本，其中 108 个在训练集中未出现，这样的设计有助于全面评估模型的训练效果与性能。

在微调过程中，设定最大迭代轮次为 20，并采用 Qwen 官方推荐的 LoRA 微调单 GPU 脚本进行操作。经过微调，选出最优模型并将其集成至原始模型中，成功实现了生僻字的拆分检索功能。

微调模型在测试集上的表现优异，仅有 3 条样本数据转换出现错误。深入分析后发现，这些错误主要源于训练集中数据的缺失。例如，“熟去掉四点底”被错误识别为“熟|丶”，未能准确解析出“去掉四点底”;“更生组成的”被误判为“更生组成”，未能剔除修饰词“组合”。由于“去掉四点底”和“组成”的描述方式在训练集中未出现，导致了模型转换失误。这一问题可通过扩充训练集来改善，从而提升模型的泛化能力。

微调模型在测试集上的准确率高达 99.69%，成功实现了从拆分字符串到生僻字构成部件的高效转换。如图 4 展示，该功能已在实际产品中得到应用，用户可通过拆字方式便捷地查询生僻字信息。

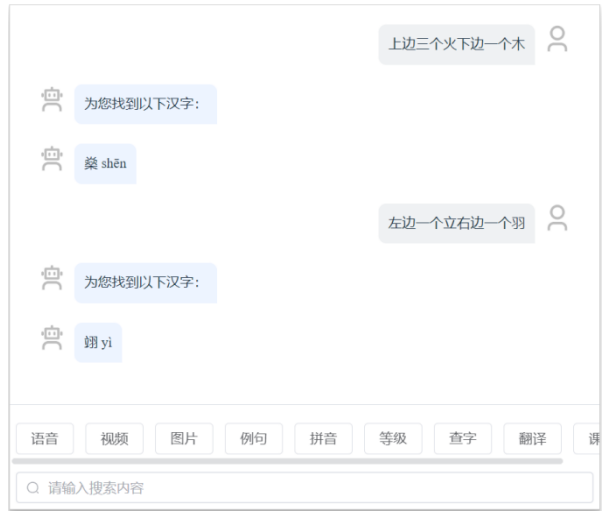


图 4：汉字拆分检索效果

5 结论

本文致力于解决生僻字识别与应用的挑战，通过构建生僻字表以及建立生僻字知识库，实现了对生僻字的高效检索和应用。

本文的主要贡献，首先是总结生僻字检索方法，包括直接查询和拆字查询，并通过实验验证了其有效性;其次是制定了一套生僻字拆分规则，并通过部件还原和容错处理提高了识别准确性；最后是构建了一个全面的生僻字知识库，为生僻字的领域分布、历时统计和拆分检索等应用提供了有力支持。

尽管本文已经取得了显著的进展，但仍然存在一些局限性。首先，生僻字表和知识库的构建仅基于有限的样本数据，这可能未能全面覆盖所有生僻字。其次，拆分规则和部件还原方法仍需进一步优化和完善。部分生僻字很难用汉字部件描述，导致用户在构造有效的拆字检索时面临困难，例如，“羋”字在直接检索中的频率较高，但在拆字检索中却未出现。此外，某些汉字的拆分结果存在歧义，如“邑”和“吧”均可拆分为“口巴”。若用户未提供结构信息，仅依靠部件列表无法准确判断用户意图，此时需要人工干预以辅助判断。最后，关于生僻字知识库的应用场景讨论相对有限，未来研究可以进一步挖掘其在更广泛领域的应用潜力。

参考文献

[1] 夏征农，陈至立. 辞海（第六版彩图本） 第六版[M]. 上海：上海辞书出版社，2009:2021.

[2] 魏傲. 字詞系統視域下《說文・馬部》生僻字研究[D]. 兰州大学,2023.

[3] 范亚茹. 文化传承视角下“生僻字”流行的再认识[J]. 汉字文化,2019,(18):38-39.

[4] 陈立民,徐路. 汉语生僻字的叙事性设计研究[J]. 设计,2023,36(01):46-49.

[5] 商伟凡. 试论我国政区名称生僻汉字的治理[J]. 语文建设,1999,(06):13-16.

[6] 慈舒. 通用规范汉字表[M]. 北京：语文出版社，2013.

[7] 荀恩东，饶高琦，肖晓悦，臧娇娇. 大数据背景下BCC语料库的研制[J]，语料库语言学，2016(1).

[8] 艾卓码. 2021. 姓名生僻字应用的困境与对策[J]. 信息技术与标准化，(10):77-82.

[9] 丁芸. 2009. 我国公民姓名用字中的生僻字分析[J]. 现代商贸工业,21(18):240-241.

[10] 黄小英. 2021. 《中华人民共和国地名大词典》地名生僻字研究[D]. 四川外国语大学.

[11] 石真一. 2018. 网络语言中单个生僻字的流行现状及成因分析[J]. 职大学报，(05):57-60.

[12] 刘冬青,施建平. 2010. 试说网络生僻词——以“烎”为例[J]. 语文建设，(06):34-35.

[13] 蔡文慧. 2021. 网络流行生僻字的象似性分析[J]. 汉字文化，(S2):24-26.

[14] 范慧敏. 2021. 网络语言流行生僻字研究[D]. 辽宁大学.

[15] 杜永青. 2010. 网络生僻字探源[J]. 新乡学院学报(社会科学版),24(06):123-125.

[16] 唐璐璐. 2022. 网络古生僻词研究[J]. 文化创新比较研究,6(12):57-61.

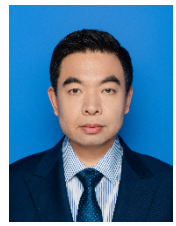
[17] 范亚茹. 2019. 文化传承视角下“生僻字”流行的再认识[J]. 汉字文化，(18):38-39.

[18] 尹斌庸,方世增. 1994. 词频统计的新概念和新方法[J]. 语言文字应用，(02):69-75+113

[19] 荀恩东，饶高琦，谢佳莉,等. 现代汉语词汇历时检索系统的建设与应用[J]. 中文信息学报，2015,29(3):169-176.

[20] Jinze Bai, Shuai Bai, Yunfei Chu ,et al. Qwen Technical Report. arXiv, 2023, arXiv:2309.16609.

[21] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-Rank Adaptation of Large Language Models[J]. 2021. DOI:10.48550/arXiv.2106.09685.



刘廷超（1978—），博士研究生，主要研究领域为信息检索、自然语言处理。
E-mail：liutingchao@hotmail.com

王雨（1993—），博士研究生，主要研究领域为语言资源建设、国际中文智慧教育、语义分析。
E-mail：1213546256@qq.com

饶高琦（1987—），通信作者,博士，副研究员，主要研究领域为语言信息处理、教育技术、语言资源、语言规划。
E-mail：raogaoqi@blcu.edu.cn