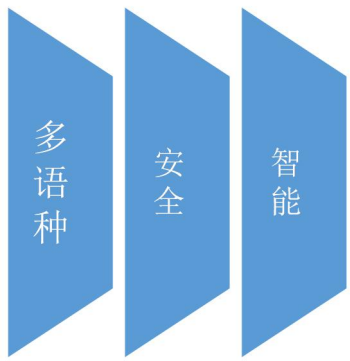
泥娃搜索技术白皮书



上海泥娃通信科技有限公司

2018-04

目录

泥娃搜索技术白皮书 2
1. 引言 2
2. 背景 2
3. 概述 5
4. 特点 7
4.1 产品特点 7
4.2 和关键词全文搜索对比 7
4.3 关键技术指标 8
4.4 功能 8
4.5 系统特色 9
5. 主要技术 10
5.1 关键技术 10
5.2 技术路线 10
5.3 语义树索引 11
5.4 密文搜索 13
6. 应用价值 14
7. 典型应用 15
7.1 文字信息的检索 16
7.2 联想语义 16
7.3 密文搜索 16
7.4 语义树应用 16
8. API 接口 18
8.1 内核管理 18
8.2 查询接口 22
8.3 文章导入接口 24
9. 技术支持 25
附录 25
1. utf8 编码分类 25
2. utf8 标点符号编码 30

泥娃搜索技术白皮书

1. 引言

信息技术的飞速发展，对信息的安全提出了更高要求，如何实现信息安全，从信息的安全存储，安全传输到信息的安全检索，是云计算时代必须面临的挑战，如何高效的检索这些加密的非结构化数据，还是一个亟待解决的难题。

泥娃多语种和密文全文搜索系统，构建一种基于语义树的全文搜索系统，在此基础上展开加密信息的全文搜索，在信息资源加密存储的前提下，通过对其构建密文全文索引，满足人们对于信息安全的需求。

人工智能可为许多行业带来巨大发展潜力，语义理解和智能搜索技术方兴未艾，借助独有的语义树索引技术实现一体化的多语种的自然语言理解服务，结合知识图谱实现智能化的搜索服务。

泥娃搜索旨在帮助企业、单位和个人，满足信息安全和信息搜索工作需求。统一的语义树索引处理技术，结合自主研发的密文索引算法，以满足人们对于加密信息安全检索的需求。结合 Newasoft® 分布式爬虫服务、文档信息索引服务，构成成套的技术解决方案，提供标准 API 接口和完善的技术服务，有助于开发、部署和集成智能全文搜索技术。

泥娃搜索基于语义树，不依赖于分词，致力于提供所有文字的全文检索服务。

泥娃搜索支持密文搜索服务。

泥娃搜索由上海泥娃通信科技有限公司提供。

关键词：

语义树索引 密文索引 zy6 消息摘要算法 sm3 sha256 utf8 全文搜索 密文搜索 可加密的搜索机制

约定：

分离码算法 flcode

前端 浏览器和客户端

2. 背景

搜索引擎是信息时代的基础服务之一，搜索引擎服务的核心为全文检索。常用的全文检索提供基于关键词的查找。研究一种基于句子的查找是研发的动机之一。

全文检索主要分为两个部分：文章索引和查找。传统的文章索引主要指的是关键词的索引。简单来说就是，索引程序通过扫描文章，为每一个词建立一个索引，记录该词在文章中出现的次数和位置，查询时，检索程序就根据事先建立的索引进行查找，并将查找的结果反馈给用户的检索方式。

全文检索系统是按照全文检索理论建立起来的，用于提供全文检索服务的软件系统。一般来说，全文检索需要具备建立索引和提供查询的基本功能。功能上，全文检索系统核心具有建立索引、处理查询返回结果集、增加索引、优化索引结构等等功能，外围则由各种不同应用具有的功能组成。结构上，全文检索系统核心具有索引引擎、查询引擎、文本分析引擎、对外接口等等，加上各种外围应用系统等等共同构成了全文检索系统。

全文信息搜索是信息时代的基本服务，文本信息检索是发展较快也较成熟的，其他的信息检索技术，往往也需要文本信息检索的支持。虽然搜索引擎已不仅仅局限于对文本进行检索，文本信息检索仍然是大部分搜索引擎的基础。常用的全文索引引擎有 Lucence 等,其主要是通过分词技术，结合文档关键词倒排序表实现全文信息的索引。

在信息检索系统的具体实现中，往往需要快速地找到文档中所包含的关键词。相比文档来说，关键词的个数是较少的，因此，以关键词为核心对文档进行索引是更加可行的方法。这就是信息检索领域常用的“倒排文档索引”技术。倒排文档索引可以被看成一个链表数组，每个链表的表头包含关键词，其后续单元则包括所有包括这个关键词的文档标号，以及一些其他信息。这些信息可以是文档中该词的频率，也可以是文档中该词的位置等信息。

倒排文档索引的优势不仅在于关键词个数少带来的检索效率提高，还在于其特别易于同信息检索技术结合。在实际应用中，查询中所包含的关键词往往是很少的，完全不包含查询中的所有关键词的文档，一般来说是不会被列入结果集的。因此，以关键词为主键进行索引，只需要用查询中包括的关键词，进行几次简单的查询就能够找出所有可能的文档。

全文索引主要技术是倒排文档索引技术，实质是词或者字的索引，结合特定的词典形成特定的关键词索引。

分词也是全文索引的关键技术之一。分词就是将连续的字序列，按照一定的规范重新组合成词序列的过程。中文分词(Chinese Word Segmentation) 指的是将一个汉字序列切分成一个一个单独的词。我们知道，在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在词这一层上，

中文比之英文要复杂的多、困难的多。分词技术的主要目的是减小倒排序表的存储，提高索引的效率。缺点是需要针对性的给出不同字典和分词的方法，缺乏统一的尺度。

对于常用的全文搜索来说，基本的功能就是分词加上倒排序文档。搜索引擎的服务随着信息量的增大，检索和存储量大，存在索引时间长，搜索速度慢等问题。

常规的全文检索对于分词技术和字典的依赖，使得全文搜索实施的难度加大。对于不同语种需要不同的字典和分词技术，对于同一语种不同专业的文档也需要不同的分词技术和字典，不同字典和分词技术也影响了系统的通用性。

传统的关键词搜索在语义理解方面支持存在极大的不足，这也是研发本产品的原因之一。研发支持所用语言的，实现智能和“更懂你”的搜索。

采用对文本信息进行特征序列的编码，形成相关的语义树，实质上提供提供一种基于语义树的索引方法和系统，不再依赖于分词、适合不同语种的全文搜索引擎，具有存储空间小，索引速度和查询速度快等特点。

3. 概述

构建语义树，通过语义树的构建提供一种快速匹配语义的方法，根据语句和文档的关系，查找到相关的文档信息。语义树的基本存储单元包括：最小语义单元，该单元的特征编码，前置单元的特征编码。泥娃搜索提供 windows 和 linux 下的 64 位版本，系统 c++编制，前端结合 jquery、d3.js 实现 web 页面。支持数据库 mongodb。系统采用 web 服务的架构提供，数据存储采用 mongodb，配置文件为 xml 文件。web 页面分为三大部分：查询、语义树查询和管理配置页面。

多语种全文搜索包括：数据导入，语义树编码、存储到数据库，查询是对查询的语句进行语义树特征编码，然后语义树查询，数据保存在数据库。内核文件存放在 xml 文件中，有关内核的配置也存放在数据库中。具体操作见下图。

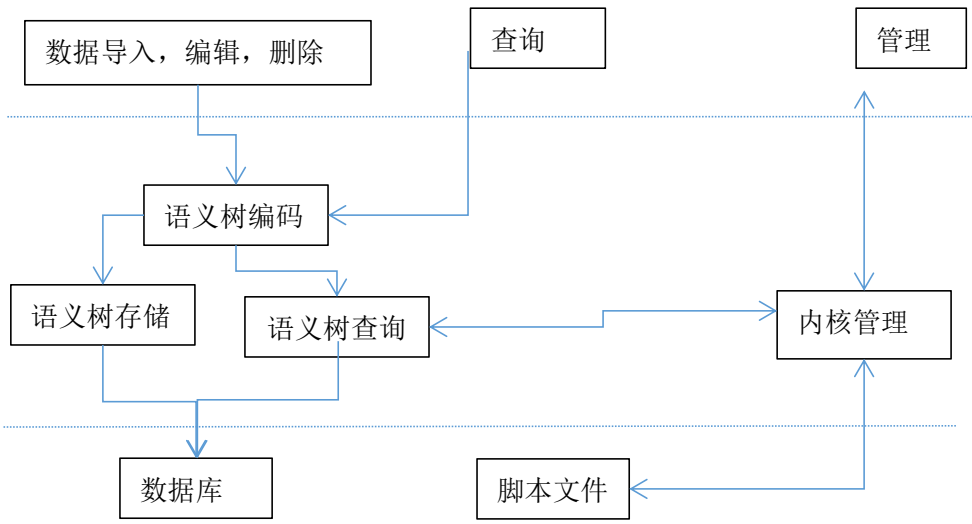


图 1 多语种全文搜索

密文全文搜索主要变化在数据导入和查询。导入数据要加密，查询的语句也要加密，服务器端注意内核的设置，要适合密文的文字环境，密钥保存在客户端。

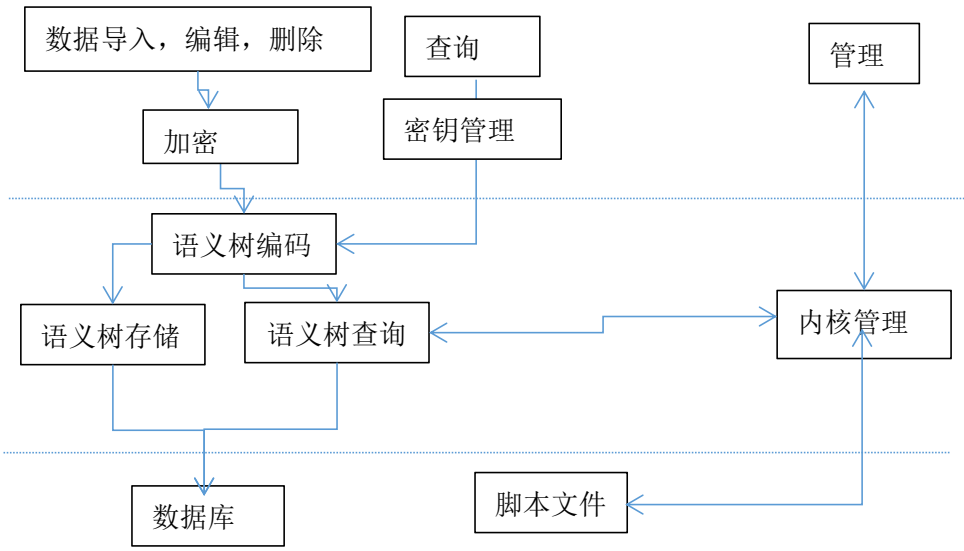


图 2 密文全文搜索

系统提供的是 B/S 的服务，所有的服务和 API 通过 http 协议提供。具体分为三层的架构：前端、全文搜索服务和数据层。

图 3 系统架构

4. 特点

泥娃搜索是一款基于语义树的全文检索服务系统，对语言的支持采用统一的标准，支持语言仅仅需要以下特点的：有最小的文字单元，语句可以切分。

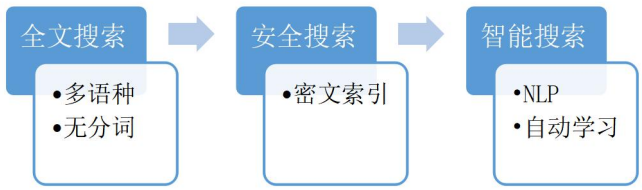


图 4 泥娃搜索进化趋势

4.1 产品特点

· 针对多语种的强大可扩展性。实现国际化界面设置，简单的设置配置文件即可实现不同语言的界面设置。统一的语义树索引技术，无需分词和字典，彻底实现和语言无关的全文搜索，便于多语种全文搜索服务的部署和实施。

· 快速、经济的全文搜索网络。语义树的索引技术，实现索引的体积仅为关键词索引的0.1%，先计算后搜索的技术保证查询时的计算仅为关键词搜索的1%，两者甚至于更低，并且随着数据量的增加，索引体积和文档的占比更低，计算效率更高。多语种全文搜索的无差异化特性，可以更经济部署多语种搜索。

· 真正的可搜索的加密机制。自主研发的分离码算法消息摘要算法实现了密文的索引，密文的索引无法还原原始的信息，保证安全的同时，还可以提供文的检索服务。

4.2 和关键词全文搜索对比

· **索引体积更小**：索引体积仅为0.1%，随着数据量的增大，索引和信息的比值将越来越小，语句不会重复索引，同样的语义树特征节点也不会重复索引。

· **计算占用资源更少**：语义树索引技术采用先计算后查找的方法，即先计算查询语句的语义特征码，后依据计算结果查找语句，检索效率更高。

· **多语种**：独有的算法实现语言统一处理标准，统一搜索方式，无论哪种语系，均同样处理。无需分词和设置简单，无需担心新词索引处理。

· **密文检索**：信息索引采用密文加散列的方式保存，通过密文索引不能还原原始信息，原始信息加密和密文索引在客户端完成，文档加密支持第三方，安全可靠。

· **智能语义理解**：具有语料自动分析功能，自动的提取语义单元，实现NLP和后续的关联查找；提供多语种的NLP服务，支持语义理解结合最小语义知识库，提供智能化的搜索服务。

4.3 关键技术指标

· 语义树索引支持2的96次方。

· 语义特征编码的散列算法支持：zy6、sm3、sha256；散列效果测试：语义树特征节点、语句特征节点和文档特征节点ID数目一致；同样计算条件下，循环测试，前一次散列结果作为下一次的输入：100万个384位信息，zy6，2540ms ；100万个256位信息，sm3:5760ms；100万个256位信息，sha256:5860ms。

· 密文索引关键算法flcode,密钥长度为K,安全解空间为K的阶乘；支持多次不同密钥的多次加密，加密后的结果支持语义树全文搜索。

· 支持所有utf8编码文字的全文索引。

4.4 功能

· 多语种搜索，支持所有utf8编码的语言的全文检索。

· 人机界面多语种支持，支持所有utf8编码语言，提供国际化部署的设置。

· 基于联想语义的全文检索。

· 支持密文信息的全文检索。

· 提供API接口，方便其他系统调用和数据的导入。

· 支持前端直接数据导入，js语句实现。

· 提供c语言和javascript语言的密文算法,方便密文文档的导入。

· 密文信息的查询实现，密文条件下查询的结果为密文，在用户侧进行信息的加解密计算。

4.5 系统特色

· 自然语句搜索，搜索的内容按自然语句的形式进行最大化匹配后展示搜索的结果。

· 支持多语句的查找，语句之间的关系为“和”。

· 可以绘制语义树，提供基于语义树的查询方式；

· 支持联想记忆查找，查找后给出该查找内容的后续文字记录，方便进一步的查找。

· 支持所有文字，对于文字处理仅仅需要分句和分字，即有划分语句的规则和区分文字单元的规则即可。

· 独有的语义特征编码技术，实现语句或者语句片段的快速查找。

· 适合进行语义理解的搜索，方便进行语义理解。

· 支持多核设置和分开存储。

· 支持mongodb数据库。

· 支持分布式的部署和负载均衡。

5. 主要技术

5.1 关键技术

1、语义树全文搜索引擎技术

多语种和密文全文搜索关键技术包括：信息安全算法，信息搜索算法等。语义特征编码技术是全文搜索的核心算法，数字指纹算法在其中起到关键性的作用；密文搜索的关键算法是分离码算法。

2、基于路径散列的消息摘要技术

通过信息分组、路径散列计算、结果序列调和散列，结合输出字符串的设定，从而输出消息摘要。本技术可以扩展和衍生不同的摘要算法，不同的分组，不同的变换序列，不同的路径选择算法，不同的散列算法均可以产生不同的消息摘要。

3、语义联想记忆技术

通过语义标识 ID 的链式存储，构建语义上下的关系，实现对语句的上下文搜索，从而实现一定程度的语义会话功能。系统主要用于人工智能领域的语义理解、智能机器人的人机对话、自然语言的语句搜索。

4、分离编解码技术

利用数字不同进制的转换结合码表，形成信息变换序列和位数序列分离，实现信息的编码；结合码表、变换序列和位数序列来解码实现信息还原。

5.2 技术路线

泥娃搜索由全文索引系统、索引安全算法、信息安全系统组成。全文索引系统采用基于语义树的索引系统组成，索引安全算法和信息安全系统采用分离码算法实现。

1、全文检索路线

（1）文档的导入，以文档中的句子为单位，通过对句子中文字信息的增量 hash 编码，构建句中文字的序列信息特征编码。

（2）语义树的构建，基于文字的表示习惯，以语句为单位构建文字和文字之间的前后关系。

（3）基于文字的编码规定，结合文字特有的分隔符先对语句进行切分，后对语句进行特征序列的编码处理。

（4）对组建语义树的编码范围给定，构建单一语种、多语种组合甚至不分语种的语义树。

（5）针对语句或者语句片段的查找，在语义树中查找记录，主要分以下步骤：1)特征序列的最大化查找；2)特征序列的递减查找；3）语

句特征序列的关联文档或者处理方法查找。

2、分离编解码路线

（1）制定码表：确定处理信息的单元位数，确定转换的进制，定义码表。

（2）编码：根据分离编解码分组单元要求读取 64 位（或者 128 位，或者其他）赋值给整数，然后根据要求转换成相应的进制（），转换结果记录到变换序列，转换后的位数记录到位数序列，一直持续到转换完毕，最后形成两个部分；变换序列的字符一定是码表的字符，位数序列主要记载转换结果对应在变换序列记录中的长度。

（3）解码：从位数序列读取位数信息，按位数信息从变换序列中读取相关的字符，依据分离码表变换成相应的数字，结合分离码表的进制定义，转换为整数，依次存入到解码结果，直到转换完毕，实现信息的解码。

3、密文索引路线

密文索引安全算法采用分离编解码算法，文档的加解密可以采用第三方算法。密文索引采用先对文档进行语句的分割，以语句为单位进行加密，加密后语句之间用标点符号分割，然后提供给多语种全文搜索服务系统，分离编解码算法保证密文和原始文档在基于语义树的搜索上的一致性。

5.3 语义树索引

语言的基本单位为语句，语句的基本元素为文字，由文字构成不同的语句，语句是文章或者人际交流最基本的语言单位。如果一句话为树的一个分支，那么同一文字起点的语句结合在一起就构成一棵语义树，树上的节点分为根节点，分支节点，果子节点（语句最后的节点，一般对应一篇文章，如果文章为果子的话，该节点为果子节点），这样所有的语句组成不同的语义树，整个语义树表示现代语言的语句的集合。

语义树的索引技术，通过该技术找到最大匹配的语句，从而得到果子，找到匹配的文章。该技术可以用于全文索引、密文全文索引和 NLP 语义理解等领域。

关键词：

语义树 增量 hash 链式存储 全文搜索 NLP 语义理解

语言是信息的高度浓缩，给人以记忆、联想，人们利用语言交流，写作，从事科研，人的活动可以通过文字的形式来表达。文字的形成是一件伟大的事，文字、语句和文章的组成需要满足该语言的规则，文字

的上下文联系现在看来既浑然天成又不可思议。

信息时代是伟大的时代，信息时代的记忆对比人来说，对比传统的文章而言，不可同日而语。

用信息时代的技术来描述和记载文字，形象的描述文字的上下文联系，就成了一件有意义的事，于是开始语义树的研究。

意义不仅于此，通过语义树可以描述语言上下文的同时，也为语言文字全文的检索提供了可行的支撑。

对文字的搜索提供一种独有的方式，按系统的算法对文字进行特定的编码，组建索引时存到关系表 word 中，为语句位的提供语句标识；存储语句和文档的关系到 docseg 表中；文档存储到 text 表中。

5.3.1 语义树索引技术简介

语义树索引技术主要是根据语句中文字的排列顺序，计算文字对应语句的特征编码，利用链式存储技术，实现对应语义树。形象的描述是：文字为语义树上的基本节点，语句为语义树的枝条，所有的枝条结合在一起构建语义树；语义树上的同一分枝具有同样的根节点文字。

语义树的索引技术指的是在语义树上查找语句，通过语句找到对应的文档。一般来说语义树的索引包括：

1、语义树。利用语句中文字的特征码，结合该文字前面的特征码组建。主要技术为特征码和链式存储。

2、语句和文档的关系。主要存储具有句尾标识的语句的特征码和文档 ID 的关系。

通过最大匹配语句找到文档的过程，称为语义树的索引。

语义树索引技术主要采用的技术：

1、特征编码技术。主要由增量 hash 算法构建，假定语句含有文字序列为$\{w_0,w_1,...,w_n\}$，则特征码计算如下：

$T_i=hash（T_{i-1}+w_{i-1}）$,当 i=0 时，$T_0=hash(w_0)$。

2、链式存储技术。语义树的存储单元为：$\{T_i，w_i，T_{i-1},f\}$，其中 f 表示是否为句尾。

3、语句和文档的关系。主要存储具有句尾标识的语句的特征码和文档 ID 的关系。

4、语句匹配查询技术。分为两种方式：计算语句的特征码，在语义树中查找最大匹配深度的特征码，然后依据特征码在语义树中匹配语句，根据匹配语句找到文档。在语义树中找到第一个文字对应的记录集合，从集合中取出特征码，结合查询的语句，去掉第一个字符，和后续的文字计算特征码，在语义树中继续查找，直到找到最大匹配深度的特征码，然后依据特征码在语义树中匹配语句，根据匹配语句找到文档。

5..3.2 语义树索引技术特点

语义树记载文字的上下文关系，特征码的计算和存储保障数据的最小化存储。语句中文字的特征码仅仅和该文字前的文字（包含该文字）以及文字排列的顺序相关。

语义树索引技术特点：

1、可视化的表示语言的组成。

2、真实的反应语言文字之间的关系。

3、有利于全文检索的精准查找。

4、语句的检索速度快，系统生成的索引体积小。

语义树索引采用增量 hash 技术的特点，和基于语句 hash 的不同之处在于：

1、增量 hash 保存的单元大小和格式固定。每个单元有：96 整形数表示的特征码，文字单元，前 96 整形数表示的特征码，句尾标识组成。

2、语句 hash 保存单元可以同样设置。每个单元里有：96 整形数表示的特征码，语句片段，前 96 整形数表示的特征码，句尾标识组成。

3、增量 hash 后续查找便捷，直接用该特征码结合后续文字计算即可，计算的特征码再查询。

4、增量 hash 有利于语义树的构建。

5、结合算法实现密文的语义树搜索。

5.3.3 语义树索引的应用

用于全文搜索方面，对比的有关键词倒排序表的技术；用于语义理解方面对应于常见的分词算法。

语义树索引的应用主要分为：

1、基于语义树的全文搜索。

2、基于语义树的密文全文搜索。

3、语义分析。

5.4 密文搜索

密文搜索主要的信息安全算法在前端完成。文章需要实现密文搜索，主要做两件事，第一完成信息的加密处理后，把加密信息输入到全文索引；第二查询条件加密后，送到服务端查询，查询的结果前端解密显示。

5.4.1 加解密算法

密文索引算法采用分离码算法实现。

密文全文搜索系统，主要包括：信息的加密，加密信息索引，全文索引系统。主要利用密文和原文前缀一致性，通过查询前缀匹配实现密文的检索，保证密文的查询和原始语句的查询结果一致。

索引加密算法的特征：语句前缀相同，加密的结果具有相同的一致性，这样保证采用密文检索和原语句检索的结果是一致的。加密信息全文检索算法可以采用分离码算法（指的是《一种分离编解码的方法和系统》实现的算法），文档按标点符号进行语句分割，形成信息加密单元，通过分离码算法加密，结合标点符号构成全文加密索引的相关字段，导入全文索引。

文档内容可以采用其他的加密方式进行。

5.4.2 密文索引的建立

信息全文需要建立密文索引信息和信息密文，密文索引信息建立过程如下：

首先对原文进行断句，然后对每一句进行分离码算法加密，组合所有加密语句，句与句之间增加标点符号或者自定义的语句分隔符，组成的信息为密文索引信息；信息原文的密文直接采用分离码算法加密即可。

支持多重分离码算法加密。

5.4.3 密文搜索

密文搜索主要的信息安全算法在前端完成。文章需要实现密文搜索，主要做两件事，第一完成信息的加密处理后，把加密信息输入到全文索引；第二查询条件加密后，送到服务端查询，查询的结果前端解密显示。

6. 应用价值

信息技术的飞速发展，对信息的安全提出了更高要求，如何实现信息安全，从信息的安全存储，安全传输到信息的安全检索，是云计算时代必须面临的挑战，如何高效的检索这些加密的非结构化数据，还是一个亟待解决的难题。

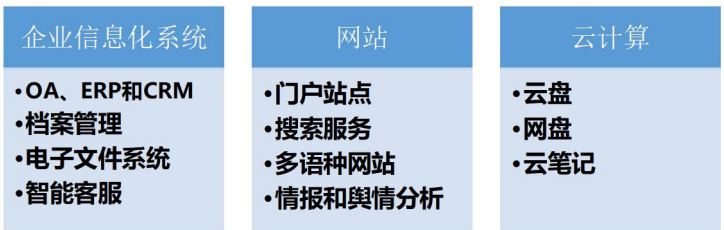
泥娃多语种和密文全文搜索系统，构建一种基于语义树的全文搜索系统，在此基础上展开加密信息的全文搜索，在信息资源加密存储的前提下，通过对其构建密文全文索引，满足人们对于信息安全的需求。

人工智能可为许多行业带来巨大发展潜力，语义理解和智能搜索技术方兴未艾，借助独有的语义树索引技术实现一体化的多语种的自然语言理解服务，结合知识图谱实现智能化的搜索服务。

泥娃搜索旨在帮助企业、单位和个人，满足信息安全和信息搜索工作需求。统一的语义树索引处理技术，结合自主研发的密文索引算法，以满足人们对于加密信息安全检索的需求。结合 Newasoft® 分布式爬虫服务、文档信息索引服务，构成成套的技术解决方案，提供标准 API 接口和完善的技术服务，有助于开发、部署和集成智能全文搜索技术。

泥娃提供更简单、更灵活的全文搜索平台，以满足人们对于全文搜索的工作需求。通过开放的 API 接口实现和其他系统的互通，容易集成到企业的 OA、ERP、档案管理和云存储服务，提高相关文档的利用率。

泥娃搜索可帮助企业以更低的风险、更轻松地利用颠覆性的新技术。重要的是，它还可灵活地支持各种文档的导入工作，因此可以更轻松地集成其他业务和技术应用，节省成本，同时减少专门系统之间的大型文档集迁移需求。



7. 典型应用

泥娃搜索除了用于全文检索，还提供联想语义查询和语义树的查询功能。

典型应用场景：

1、企业级市场技术服务，服务的商家包括：办公软件、和办公系统厂商，office 办公和 OA 系统，ERP 系统，电子文件系统，档案系统厂商提供全文搜索加密文全文搜索服务。

2、互联网技术服务，提供网站、电商平台等互联网平台的内容搜

索服务，包括搜索引擎提供商，云计算厂商等，提供多语种和密文全文搜索服务。

3、人工智能市场技术服务，提供智能语义理解解析服务。

4、软件市场，主要提供多语种和密文全文搜索产品，作为独立的软件提供服务。

5、情报收集、专业文献和论文的检索服务。

7.1 文字信息的检索

作为独立的全文搜索引擎使用，满足文字信息的索引和检索工作。适合：网站、网站群的站内搜索；档案、图书馆、专利平台的数据检索；金融、通信、公安等行业及企业级的全文检索服务。

适合不同的文字的全文检索，可以作为特定语言的网页信息检索的工具。

7.2 联想语义

提供文字的上下文关联检索，甚至于提供语句的上下文检索。适用于各行各业及个人的服务需求，尤其是人工智能领域，如：人机对话、智能客服等等。

7.3 密文搜索

提供密文索引的建立，密文全文搜索服务，适用于各行各业及个人基于信息安全服务的需求，如：科研机构、金融行业、通信行业、机关、学校和企事业单位、企业、个人等等，个人对于信息安全下文章全文检索服务的要求。

7.4 语义树应用

基于语义树的信息分析和展示及搜索结果的存储和交流，它既可以直观的呈现搜索的结果和结果之间的关系，也可以提供基于语义树的信息分享。

基于语义树的机器自动学习和智能语义理解，智能情报分析服务。

通过语义树的比较，分析关注点，例如不同地区的不同事物，或者不同语言的同样意思的语句，生成语义树的比较。

搜索“武汉小吃”和“重庆小吃”给出的语义树如下：

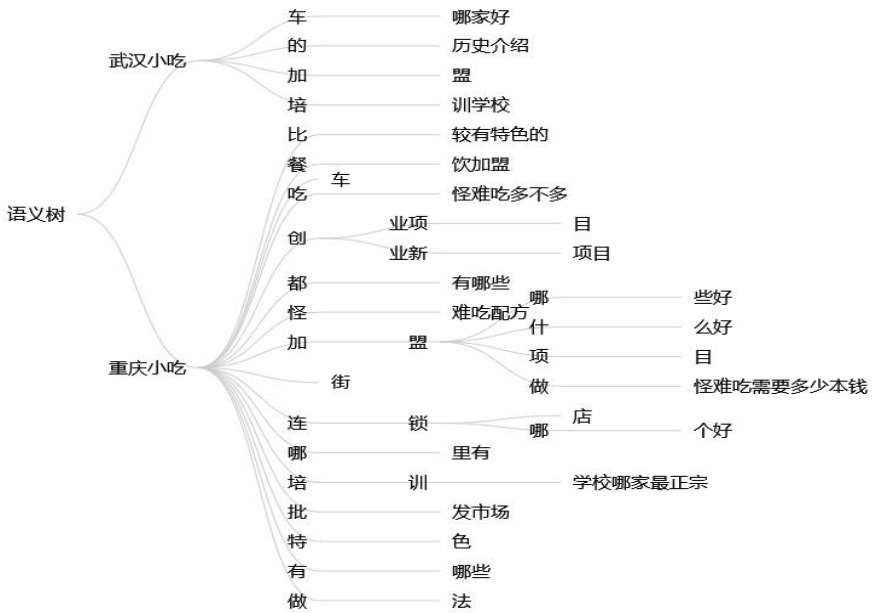


图 4 语义树

搜索“上海东方明珠”的语义树：

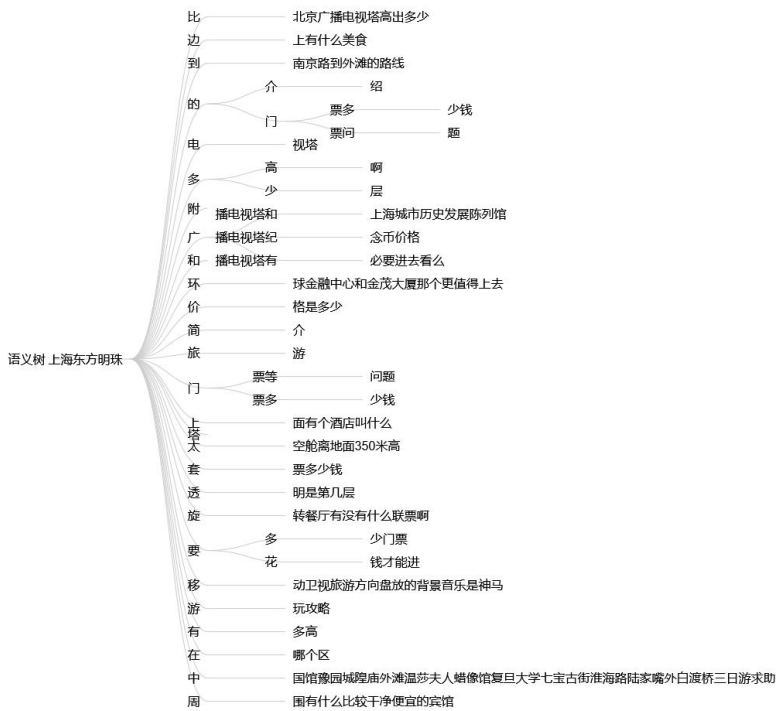


图 5 语义树

8. API 接口

系统接口基于 web 服务，接口均采用 http 服务的方式提供。

8.1 内核管理

8.1.1 内核列表接口

模式：get
接口：core_list.action
访问形式：
例如：http://www.01wa.net/core_list.action
返回结果：
{"num":5,"res":[{"name":"aq","db":"tq","style":"mongodb","engn":""},{"name":"cx","db":"cx","style":"mysql","engn":"InnoDB"},{"name":"m1","db":"QA","style":"mysql","engn":"InnoDB"},{"name":"m8","db":"arvin","style":"mysql","engn":"InnoDB"},{"name":"m_1","db":"test","style":"mysql","engn":"InnoDB"}]}

8.1.2 增加内核

模式：post
接口：core.action
参数：
opt：add
value：加入内核的内容
访问形式：

Request：
{"opt":"add","value":{"name":"test","db":"test","style":"mongodb","engn":""}}:
Response：
{"num":5,"res":[{"name":"aq","db":"zy6q","style":"mongodb","engn":""},{"name":"keyword","db":"keyword","style":"mongodb","

engn":""},{"name":"miwen","db":"miwen","style":"mongodb","eng
n":""},{"name":"sybigdata","db":"sy","style":"mongodb","engn":
""},{"name":"test","db":"test","style":"mongodb","engn":""}]}

8.1.3 编辑内核

- 设置空格分词的语种

参数

style：3，

{"opt":"write","core":"aq","style":3,"value":"{\"split_word_b
y_blank\":[{\"min\":\"0020\",\"max\":\"007F\"},{\"min\":\"00A
0\",\"max\":\"00FF\"},{\"min\":\"0100\",\"max\":\"017F\"},{\"
min\":\"0370\",\"max\":\"03FF\"},{\"min\":\"0400\",\"max\":\"
04FF\"},{\"min\":\"0500\",\"max\":\"052F\"},{\"min\":\"0530\",
\"max\":\"058F\"},{\"min\":\"0590\",\"max\":\"05FF\"},{\"min\
":\"0600\",\"max\":\"06FF\"},{\"min\":\"0700\",\"max\":\"074F
\"},{\"min\":\"0750\",\"max\":\"077F\"},{\"min\":\"0780\",\"m
ax\":\"07BF\"},{\"min\":\"0900\",\"max\":\"097F\"},{\"min\":\
"0980\",\"max\":\"09FF\"},{\"min\":\"0A00\",\"max\":\"0A7F\"},
{\"min\":\"0A80\",\"max\":\"0AFF\"},{\"min\":\"0B00\",\"max\":
\"0B7F\"},{\"min\":\"0B80\",\"max\":\"0BFF\"},{\"min\":\"0C00
\",\"max\":\"0C7F\"},{\"min\":\"0C80\",\"max\":\"0CFF\"},{\"m
in\":\"0D00\",\"max\":\"0D7F\"},{\"min\":\"0D80\",\"max\":\"0
DFF\"},{\"min\":\"0F00\",\"max\":\"0FFF\"},{\"min\":\"10A0\",
\"max\":\"10FF\"},{\"min\":\"1200\",\"max\":\"137F\"},{\"min\
":\"1380\",\"max\":\"139F\"},{\"min\":\"13A0\",\"max\":\"13FF
\"},{\"min\":\"1400\",\"max\":\"167F\"},{\"min\":\"1680\",\"m
ax\":\"169F\"},{\"min\":\"16A0\",\"max\":\"16FF\"},{\"min\":\
"1700\",\"max\":\"171F\"},{\"min\":\"1720\",\"max\":\"173F\"},
{\"min\":\"1740\",\"max\":\"175F\"},{\"min\":\"1760\",\"max\":
\"177F\"},{\"min\":\"1800\",\"max\":\"18AF\"},{\"min\":\"1900
\",\"max\":\"194F\"},{\"min\":\"1950\",\"max\":\"197F\"},{\"m
in\":\"1980\",\"max\":\"19DF\"},{\"min\":\"1A00\",\"max\":\"1
A1F\"},{\"min\":\"1B00\",\"max\":\"1B7F\"},{\"min\":\"1F00\",
\"max\":\"1FFF\"},{\"min\":\"2C00\",\"max\":\"2C5F\"},{\"min\
":\"2C80\",\"max\":\"2CFF\"},{\"min\":\"2D00\",\"max\":\"2D2F
\"},{\"min\":\"2D30\",\"max\":\"2D7F\"},{\"min\":\"2D80\",\"m
ax\":\"2DDF\"},{\"min\":\"A000\",\"max\":\"A48F\"},{\"min\":\

"A490\",\"max\":\"A4CF\"},{\"min\":\"A700\",\"max\":\"A71F\"},{\"min\":\"A720\",\"max\":\"A7FF\"},{\"min\":\"A840\",\"max\":\"A87F\"},{\"min\":\"AC00\",\"max\":\"D7AF\"},{\"min\":\"FF00\",\"max\":\"FFEF\"},{\"min\":\"10300\",\"max\":\"103\"},{\"min\":\"10330\",\"max\":\"103\"},{\"min\":\"10380\",\"max\":\"103\"},{\"min\":\"103A0\",\"max\":\"103\"},{\"min\":\"10400\",\"max\":\"104\"},{\"min\":\"10450\",\"max\":\"104\"},{\"min\":\"10800\",\"max\":\"108\"},{\"min\":\"10900\",\"max\":\"109\"},{\"min\":\"10A00\",\"max\":\"10A\"},{\"min\":\"12000\",\"max\":\"123\"},{\"min\":\"12400\",\"max\":\"124\"},{\"min\":\"1D000\",\"max\":\"1D0\"},{\"min\":\"1D100\",\"max\":\"1D1\"},{\"min\":\"1D200\",\"max\":\"1D2\"}]}"}:
Response：为 true 表示成功；
{"result":true}

- 设置索引的语种

参数
style：1
value：word 表示特定的文字；multi_word 表示 utf8 区块

{"opt":"write","core":"aq","style":1,"value":"{\"word\":\"[]\",\"multi_word\":[]}"}:
{"result":true}

- 断句设置

参数
style：2
value：sen 表示特定的文字；multi_sen 表示 utf8 区块;noinc 排除字符

{"opt":"write","core":"aq","style":2,"value":"{\"sen\":[],\"noinc\":[],\"multi_sen\":[{\"min\":\"2000\",\"max\":\"206F\"},{\"min\":\"3000\",\"max\":\"303F\"},{\"min\":\"0020\",\"max\":\"002F\"},{\"min\":\"003A\",\"max\":\"0040\"},{\"min\":\"005B\",\"max\":\"0060\"},{\"min\":\"007B\",\"max\":\"007E\"},{\"min\":\"00A0\",\"max\":\"00BF\"}]}"}:
{"result":true}

- 初始化，主要初始化文档索引和存储的字段

参数：

style：0；save 表示存储的字段；index 表示索引的字段

{"opt":"write","core":"aq","style":0,"value":"{\"save\":[\"_id\",\"title\",\"title_s\",\"content\",\"answer\"],\"index\":[\"title\",\"title_s\",\"answer\"]}","create_table_str":""}:
{"result":true}
{"opt":"read","core":"aq","style":0}:
{"result":true,"core":"aq","idx_str":"{\"save\":[\"_id\",\"title\",\"title_s\",\"content\",\"answer\"],\"index\":[\"title\",\"title_s\",\"answer\"]}","create_table_str":""}
{"opt":"read","core":"aq","style":3}:
{"result":true,"core":"aq","split_word_by_blank":"{\"split_word_by_blank\":[{\"min\":\"0020\",\"max\":\"007F\"},{\"min\":\"00A0\",\"max\":\"00FF\"},{\"min\":\"0100\",\"max\":\"017F\"},{\"min\":\"0370\",\"max\":\"03FF\"},{\"min\":\"0400\",\"max\":\"04FF\"},{\"min\":\"0500\",\"max\":\"052F\"},{\"min\":\"0530\",\"max\":\"058F\"},{\"min\":\"0590\",\"max\":\"05FF\"},{\"min\":\"0600\",\"max\":\"06FF\"},{\"min\":\"0700\",\"max\":\"074F\"},{\"min\":\"0750\",\"max\":\"077F\"},{\"min\":\"0780\",\"max\":\"07BF\"},{\"min\":\"0900\",\"max\":\"097F\"},{\"min\":\"0980\",\"max\":\"09FF\"},{\"min\":\"0A00\",\"max\":\"0A7F\"},{\"min\":\"0A80\",\"max\":\"0AFF\"},{\"min\":\"0B00\",\"max\":\"0B7F\"},{\"min\":\"0B80\",\"max\":\"0BFF\"},{\"min\":\"0C00\",\"max\":\"0C7F\"},{\"min\":\"0C80\",\"max\":\"0CFF\"},{\"min\":\"0D00\",\"max\":\"0D7F\"},{\"min\":\"0D80\",\"max\":\"0DFF\"},{\"min\":\"0F00\",\"max\":\"0FFF\"},{\"min\":\"10A0\",\"max\":\"10FF\"},{\"min\":\"1200\",\"max\":\"137F\"},{\"min\":\"1380\",\"max\":\"139F\"},{\"min\":\"13A0\",\"max\":\"13FF\"},{\"min\":\"1400\",\"max\":\"167F\"},{\"min\":\"1680\",\"max\":\"169F\"},{\"min\":\"16A0\",\"max\":\"16FF\"},{\"min\":\"1700\",\"max\":\"171F\"},{\"min\":\"1720\",\"max\":\"173F\"},{\"min\":\"1740\",\"max\":\"175F\"},{\"min\":\"1760\",\"max\":\"177F\"},{\"min\":\"1800\",\"max\":\"18AF\"},{\"min\":\"1900\",\"max\":\"194F\"},{\"min\":\"1950\",\"max\":\"197F\"},{\"min\":\"1980\",\"max\":\"19DF\"},{\"min\":\"1A00\",\"max\":\"1A1F\"},{\"min\":\"1B00\",\"max\":\"1B7F\"},{\"min\":\"1F00

\",\"max\":\"1FFF\"},{\"min\":\"2C00\",\"max\":\"2C5F\"},{\"min\":\"2C80\",\"max\":\"2CFF\"},{\"min\":\"2D00\",\"max\":\"2D2F\"},{\"min\":\"2D30\",\"max\":\"2D7F\"},{\"min\":\"2D80\",\"max\":\"2DDF\"},{\"min\":\"A000\",\"max\":\"A48F\"},{\"min\":\"A490\",\"max\":\"A4CF\"},{\"min\":\"A700\",\"max\":\"A71F\"},{\"min\":\"A720\",\"max\":\"A7FF\"},{\"min\":\"A840\",\"max\":\"A87F\"},{\"min\":\"AC00\",\"max\":\"D7AF\"},{\"min\":\"FF00\",\"max\":\"FFEF\"},{\"min\":\"10300\",\"max\":\"103\"},{\"min\":\"10330\",\"max\":\"103\"},{\"min\":\"10380\",\"max\":\"103\"},{\"min\":\"103A0\",\"max\":\"103\"},{\"min\":\"10400\",\"max\":\"104\"},{\"min\":\"10450\",\"max\":\"104\"},{\"min\":\"10800\",\"max\":\"108\"},{\"min\":\"10900\",\"max\":\"109\"},{\"min\":\"10A00\",\"max\":\"10A\"},{\"min\":\"12000\",\"max\":\"123\"},{\"min\":\"12400\",\"max\":\"124\"},{\"min\":\"1D000\",\"max\":\"1D0\"},{\"min\":\"1D100\",\"max\":\"1D1\"},{\"min\":\"1D200\",\"max\":\"1D2\"}]}"}

8.1.4 删除内核

参数：
opt:del
value：name，删除内核的名称
{"opt":"del","value":{"name":"test"}}:
{"num":4,"res":[{"name":"aq","db":"zy6q","style":"mongodb","engn":""},{"name":"keyword","db":"keyword","style":"mongodb","engn":""},{"name":"miwen","db":"miwen","style":"mongodb","engn":""},{"name":"sybigdata","db":"sy","style":"mongodb","engn":""}]}

8.2 查询接口

8.2.1 一般查询

模式：post
接口：search.action

参数：json 格式
core：查询内核名称
query：查询内容
skip: 从 skip 开始查询
limit: 返回记录数
byid: 为 0 标识根据文字内容查询；为 1 表示按编码的 id 查询
{core:”core”,query:”query”,skip:”0”,limit:”10”,byid:1, flag: 0 }

8.2.2 联想查询

模式：post
接口：find_children.action
参数：json 格式
core：查询内核名称
query：查询内容
style: 1 根据文字查；0 根据 id 查
w:
{"core":"aq","query":"李白","style":1,"w":""}:
{"w":"","core":"aq","query":"李白","style":1,"children":[{"w":"李白","QTime":1.269,"children":[{"w":"李白","_id":"cc0f3b41c58e008632a3e265","a":1,"children":[{ "a" : true, "_id" : "915633ff93ae8fbbc4d840ab", "w" : "诗" },{ "_id" : "28924c82aa1f889040d841c2", "w" : "只" }]}] }]}
{"core":"aq","query":"7adb09ef0c422ac21ad84399","style":0,"w":""}:
{"w":"","core":"aq","query":"7adb09ef0c422ac21ad84399","style":0,"children":[]}

8.2.3 NLP 查询接口

模式：post
接口：aisearch.action
参数：json 格式
core：查询内核名称
query：查询内容

{"core":"aq","query":"李白"}:
{"result":[{ "w":"李白","QTime":0.024,"children": [{"w":"李白","_id":"cc0f3b41c58e008632a3e265","a":1,"v":[1]},{"w":"
","_id":"0","a":1}]}],"core":"aq","tm":0.024}

8.3 文章导入接口

模式：post
接口：/core/update/json
参数：json 数组，数组内为提交的文档 json。
[{"_id":"","title":"泥娃搜索提供多语种和密文全文搜索服务","title_s":"泥娃搜索由上海泥娃通信科技有限公司开发，由苏州泥娃软件科技有限公司提供技术支持","content":"","answer":"产品特点：\n1、多语种支持。采用语义树索引技术，支持不同语言文字的全文检索服务，具有占用少，搜索效率高的特点，搜索精准度高，具有上下文的联想提示功能。\n2、密文搜索支持。支持加密文档的加密搜索服务。\n3、智能语义理解。"}]:
{"time":14.148,"result":"suc"}

9. 技术支持

泥娃搜索由上海泥娃通信科技有限公司开发，苏州泥娃通信科技有限公司提供技术支持。

企业网站：http://www.newasoft.net

附录

1. utf8 编码分类

1. 【0020-007F】 Basic Latin 基本拉丁字母
2. 【00A0-00FF】 Latin-1 Supplement 拉丁字母补充-1
3. 【0100-017F】 Latin Extended-A 拉丁字母扩充-A
4. 【0180-023F】 Latin Extended-B 拉丁字母扩充-B
5. 【0250-02AF】 IPA Extensions 国际音标扩充
6. 【02B0-02EF】 Spacing Modifier Letters 进格修饰字符
7. 【0300-036F】 Combining Diacritical Marks 组合音标附加符号
8. 【0370-03FF】 Greek and Coptic 希腊字母
9. 【0400-04FF】 Cyrillic 西里尔字母
10. 【0500-052F】 Cyrillic Supplement 西里尔字母补充
11. 【0530-058F】 Armenian 亚美尼亚文
12. 【0590-05FF】 Hebrew 希伯来文
13. 【0600-06FF】 Arabic 基本阿拉伯文
14. 【0700-074F】 Syriac 叙利亚文
15. 【0750-077F】 Arabic Supplement 阿拉伯文补充
16. 【0780-07BF】 Thaana 塔纳文
17. 【07C0-07FF】 N’Ko
18. 【0900-097F】 Devanagari 天城体梵文字母
19. 【0980-09FF】 Bengali 孟加拉国文
20. 【0A00-0A7F】 Gurmukhi 古尔穆基文
21. 【0A80-0AFF】 Gujarati 古吉拉特文
22. 【0B00-0B7F】 Oriya 奥里亚文
23. 【0B80-0BFF】 Tamil 泰米尔文
24. 【0C00-0C7F】 Telugu 泰卢固文
25. 【0C80-0CFF】 Kannada 卡纳达文
26. 【0D00-0D7F】 Malayalam 马拉亚拉姆文

27. 【0D80-0DFF】 Sinhala 僧伽罗文
28. 【0E00-0E7F】 Thai 泰文
29. 【0E80-0EFF】 Lao 老挝文；寮国文
30. 【0F00-0FFF】 Tibetan 藏文
31. 【1000-109F】 Myanmar 缅甸文
32. 【10A0-10FF】 Georgian 格鲁吉亚文
33. 【1100-11FF】 Hangul Jamo 谚文字母
34. 【1200-137F】 Ethiopic 埃塞俄比亚文
35. 【1380-139F】 Ethiopic Supplement 埃塞俄比亚文补充
36. 【13A0-13FF】 Cherokee 切罗基文
37. 【1400-167F】 Unified Canadian Aboriginal Syllabics 加拿大土著统一音节文字
38. 【1680-169F】 Ogham 欧甘文
39. 【16A0-16FF】 Runic 北欧古文
40. 【1700-171F】 Tagalog 他加禄文
41. 【1720-173F】 Hanunoo 哈努诺文
42. 【1740-175F】 Buhid 布什德文
43. 【1760-177F】 Tagbanwa 塔格巴努亚文
44. 【1780-17FF】 Khmer 高棉文
45. 【1800-18AF】 Mongolian 蒙古文
46. 【1900-194F】 Limbu 林布文
47. 【1950-197F】 Tai Le 傣哪文；德宏傣文
48. 【1980-19DF】 New Tai Lue 新傣仂文
49. 【19E0-19FF】 Khmer Symbols 高棉符号
50. 【1A00-1A1F】 Buginese 布吉文
51. 【1B00-1B7F】 Balinese 巴利文
52. 【1D00-1D7F】 Phonetic Extensions 音标扩充
53. 【1D80-1DBF】 Phonetic Extensions Supplement 音标扩充补充
54. 【1DC0-1DFF】 Combining Diacritical Marks Supplement 组合音标附加符号
55. 【1E00-1EFF】 Latin Extended Additional 拉丁字母扩充附加
56. 【1F00-1FFF】 Greek Extended 希腊文扩充
57. 【2000-206F】 General Punctuation 一般标点符号
58. 【2070-209F】 Superscripts and Subscripts 下标及上标
59. 【20A0-20CF】 Currency Symbols 货币符号
60. 【20D0-20FF】 Combining Diacritical Marks for Symbols 符号用组合附加符号
61. 【2100-214F】 Letterlike Symbols 似字母符号

62. 【2150-218F】 Number Forms 数字形式
63. 【2190-21FF】 Arrows 箭头符号
64. 【2200-22FF】 Mathematical Operators 数学运算符号
65. 【2300-23FF】 Miscellaneous Technical 混合专门符号
66. 【2400-243F】 Control Pictures 控制图像
67. 【2440-245F】 Optical Character Recognition 光学字符识别
68. 【2460-24FF】 Enclosed Alphanumerics 括号字母数字
69. 【2500-257F】 Box Drawing 制表符
70. 【2580-259F】 Block Elements 区块组件
71. 【25A0-25FF】 Geometric Shapes 几何形状
72. 【2600-26FF】 Miscellaneous Symbols 混合什锦符号
73. 【2700-27BF】 Dingbats 什锦符号
74. 【27C0-27EF】 Miscellaneous Mathematical Symbols-A 混合数学符号-A
75. 【27F0-27FF】 Supplemental Arrows-A 补充性箭头符号-A
76. 【2800-28FF】 Braille Patterns 盲文；盲人点字
77. 【2900-297F】 Supplemental Arrows-B 补充性箭头符号-B
78. 【2980-29FF】 Miscellaneous Mathematical Symbols-B 混合数学符号-B
79. 【2A00-2AFF】 Supplemental Mathematical Operators 补充性数学运算符号
80. 【2B00-2BFF】 Miscellaneous Symbols and Arrows 混合什锦符号和箭头符号
81. 【2C00-2C5F】 Glagolitic 格拉戈尔字母
82. 【2C60-2C7F】 Latin Extended-C 拉丁字母扩充-C
83. 【2C80-2CFF】 Coptic 科普特文
84. 【2D00-2D2F】 Georgian Supplement 格鲁吉亚文补充
85. 【2D30-2D7F】 Tifinagh 提非纳格字母
86. 【2D80-2DDF】 Ethiopic Extended 埃塞俄比亚文扩充
87. 【2E00-2E7F】 Supplemental Punctuation 补充性标点符号
88. 【2E80-2EFF】 CJK Radicals Supplement 中日韩部首补充
89. 【2F00-2FDF】 Kangxi Radicals 康熙部首
90. 【2FF0-2FFF】 Ideographic Description Characters 汉字结构描述字符
91. 【3000-303F】 CJK Symbols and Punctuation 中日韩符号和标点
92. 【3040-309F】 Hiragana 平假名
93. 【30A0-30FF】 Katakana 片假名
94. 【3100-312F】 Bopomofo 注音符号

95. 【3130-318F】 Hangul Compatibility Jamo 谚文兼容字母
96. 【3190-319F】 Kanbun 汉文标注号
97. 【31A0-31BF】 Bopomofo Extended 注音符号扩充
98. 【31C0-31EF】 CJK Strokes 中日韩笔画部件
99. 【31F0-31FF】 Katakana Phonetic Extensions 片假名音标扩充
100. 【3200-32FF】 Enclosed CJK Letters and Months 中日韩括号字母及月份
101. 【3300-33FF】 CJK Compatibility 中日韩兼容字符
102. 【3400-4DBF】 CJK Unified Ideographs Extension A 中日韩统一表意文字扩充 A
103. 【4DC0-4DFF】 Yijing Hexagram Symbols 易经六十四卦象
104. 【4E00-9FFF】 CJK Unified Ideographs 中日韩统一表意文字
105. 【A000-A48F】 Yi Syllables 彝文音节
106. 【A490-A4CF】 Yi Radicals 彝文字母
107. 【A700-A71F】 Modifier Tone Letters 声调符号
108. 【A720-A7FF】 Latin Extended-D 拉丁字母扩充-D
109. 【A800-A82F】 Syloti Nagri
110. 【A840-A87F】 Phags-pa 八思巴字母
111. 【AC00-D7AF】 Hangul Syllables 谚文音节
112. 【D800-DB7F】 High Surrogates 高半代用区
113. 【DB80-DBFF】 High Private Use Surrogates 高半专用代用区
114. 【DC00-DFFF】 Low Surrogates 低半代用区
115. 【E000-F8FF】 Private Use Area 专用区
116. 【F900-FAFF】 CJK Compatibility Ideographs 中日韩兼容表意文字
117. 【FB00-FB4F】 Alphabetic Presentation Forms 字母变体显现形式
118. 【FB50-FDFF】 Arabic Presentation Forms-A 阿拉伯文变体显现形式-A
119. 【FE00-FE0F】 Variation Selectors 字型变换选取器
120. 【FE10-FE1F】 Vertical Forms 竖式标点
121. 【FE20-FE2F】 Combining HalF Marks 组合半角标示
122. 【FE30-FE4F】 CJK Compatibility Forms 中日韩相容形式
123. 【FE50-FE6F】 Small Form Variants 小写变体
124. 【FE70-FEFF】 Arabic Presentation Forms-B 阿拉伯文变体显现形式-B
125. 【FF00-FFEF】 Halfwidth and Fullwidth Forms 半角及全角字符
126. 【FFF0-FFFF】 Specials 特殊区域

127.【10000-1007F】 Linear B Syllabary 线形文字 B 音节文字
128.【10080-100FF】 Linear B Ideograms 线形文字 B 表意文字
129.【10100-1013F】 Aegean Numbers 爱琴数字
130.【10140-1018F】 Ancient Greek Numbers 古希腊数字
131.【10300-1032F】 Old Italic 古意大利文
132.【10330-1034F】 Gothic 哥特文
133.【10380-1039F】 Ugaritic 乌加里特楔形文字
134.【103A0-103DF】 Old Persian 古波斯文
135.【10400-1044F】 Deseret 犹他大学音标
136.【10450-1047F】 Shavian 肃伯纳字母
137.【10480-104AF】 Osmanya
138.【10800-1083F】 Cypriot Syllabary 塞浦路斯音节文字
139.【10900-1091F】 Phoenician 腓尼基字母
140.【10A00-10A5F】 Kharoshthi 佉卢字母
141.【12000-123FF】 Cuneiform 楔形文字
142.【12400-1247F】 Cuneiform Numbers and Punctuation 楔形文字数字及标点
143.【1D000-1D0FF】 Byzantine Musical Symbols 东正教音乐符号
144.【1D100-1D1FF】 Musical Symbols 音乐符号
145.【1D200-1D24F】 Ancient Greek Musical Notation 古希腊音乐谱记号
146.【1D300-1D35F】 Tai Xuan Jing Symbols 太玄经符号
147.【1D360-1D37F】 Counting Rod Numerals 算筹记数式
148.【1D400-1D7FF】 Mathematical Alphanumeric Symbols 数学用字母数字符号
149.【20000-2A6DF】 CJK Unified Ideographs Extension B 中日韩统一表意文字扩充 B
150.【2F800-2FA1F】 CJK Compatibility Ideographs Supplement 中日韩兼容表意文字补充
151.【E0000-E007F】 Tags 语言编码卷标
152.【E0100-E01EF】 Variation Selectors Supplement 字型变换选取器补充
153.【FFF80-FFFFF】 Supplementary Private Use Area-A 补充专用区-A
154.【10FF80-10FFFF】 Supplementary Private Use Area-B 补充专用区-B

2. utf8 标点符号编码

1. 【2000-206F】 General Punctuation 一般标点符号
2. 【3000-303F】 CJK Symbols and Punctuation 中日韩符号和标点
3. 【0020-002F】 ASCII 标点符号
4. 【003A-0040】 ASCII 标点符号
5. 【005B-0060】 ASCII 标点符号
6. 【007B-007E】 ASCII 标点符号
7. 【00A0-00BF】 拉丁文第一增补集标点符号
8. 【2E00-2E7F 增补标点符号
9. 【FF01-FF0F】 全角 ASCII 标点符号
10. 【FF1A-FF20】全角 ASCII 标点符号
11. 【FF3B-FF40】全角 ASCII 标点符号
12. 【FF5B-FF5E】全角 ASCII 标点符号
13. 【FE10-FE1F】 竖排标点符号

以空格为分词语言：

1. 【0020-007F】 Basic Latin 基本拉丁字母
2. 【0370-03FF】 Greek and Coptic 希腊字母
3. 【0400-04FF】 Cyrillic 西里尔字母
4. 【0500-052F】 Cyrillic Supplement 西里尔字母补充
5. 【0530-058F】 Armenian 亚美尼亚文
6. 【0590-05FF】 Hebrew 希伯来文
7. 【0600-06FF】 Arabic 基本阿拉伯文
8. 【0700-074F】 Syriac 叙利亚文
9. 【0750-077F】 Arabic Supplement 阿拉伯文补充
10. 【0780-07BF】 Thaana 塔纳文
11. 【0900-097F】 Devanagari 天城体梵文字母
12. 【0980-09FF】 Bengali 孟加拉国文
13. 【0A00-0A7F】 Gurmukhi 古尔穆基文
14. 【0A80-0AFF】 Gujarati 古吉拉特文
15. 【0B00-0B7F】 Oriya 奥里亚文
16. 【0B80-0BFF】 Tamil 泰米尔文
17. 【0C00-0C7F】 Telugu 泰卢固文
18. 【0C80-0CFF】 Kannada 卡纳达文
19. 【0D00-0D7F】 Malayalam 马拉亚拉姆文
20. 【0D80-0DFF】 Sinhala 僧伽罗文
21. 【0F00-0FFF】 Tibetan 藏文

22. 【10A0-10FF】 Georgian 格鲁吉亚文
23. 【1200-137F】 Ethiopic 埃塞俄比亚文
24. 【1380-139F】 Ethiopic Supplement 埃塞俄比亚文补充
25. 【13A0-13FF】 Cherokee 切罗基文
26. 【1400-167F】 Unified Canadian Aboriginal Syllabics 加拿大土著统一音节文字
27. 【1680-169F】 Ogham 欧甘文
28. 【16A0-16FF】 Runic 北欧古文
29. 【1700-171F】 Tagalog 他加禄文
30. 【1720-173F】 Hanunoo 哈努诺文
31. 【1740-175F】 Buhid 布什德文
32. 【1760-177F】 Tagbanwa 塔格巴努亚文
33. 【1800-18AF】 Mongolian 蒙古文
34. 【1900-194F】 Limbu 林布文
35. 【1950-197F】 Tai Le 傣哪文；德宏傣文
36. 【1980-19DF】 New Tai Lue 新傣仂文
37. 【1A00-1A1F】 Buginese 布吉文
38. 【1B00-1B7F】 Balinese 巴利文
39. 【1F00-1FFF】 Greek Extended 希腊文扩充
40. 【2C00-2C5F】 Glagolitic 格拉戈尔字母
41. 【2C80-2CFF】 Coptic 科普特文
42. 【2D00-2D2F】 Georgian Supplement 格鲁吉亚文补充
43. 【2D30-2D7F】 Tifinagh 提非纳格字母
44. 【2D80-2DDF】 Ethiopic Extended 埃塞俄比亚文扩充
45.【A000-A48F】 Yi Syllables 彝文音节
46.【A490-A4CF】 Yi Radicals 彝文字母
47.【A700-A71F】 Modifier Tone Letters 声调符号
48.【A720-A7FF】 Latin Extended-D 拉丁字母扩充-D
49.【A840-A87F】 Phags-pa 八思巴字母
50.【AC00-D7AF】 Hangul Syllables 谚文音节
51.【FF00-FFEF】 Halfwidth and Fullwidth Forms 半角及全角字符
52.【10300-1032F】 Old Italic 古意大利文
53.【10330-1034F】 Gothic 哥特文
54.【10380-1039F】 Ugaritic 乌加里特楔形文字
55.【103A0-103DF】 Old Persian 古波斯文
56.【10400-1044F】 Deseret 犹他大学音标
57.【10450-1047F】 Shavian 肃伯纳字母
58.【10800-1083F】 Cypriot Syllabary 塞浦路斯音节文字

59.【10900-1091F】 Phoenician 腓尼基字母
60.【10A00-10A5F】 Kharoshthi 佉卢字母
61.【12000-123FF】 Cuneiform 楔形文字
62.【12400-1247F】 Cuneiform Numbers and Punctuation 楔形文字数字及标点
63.【1D000-1D0FF】 Byzantine Musical Symbols 东正教音乐符号
64.【1D100-1D1FF】 Musical Symbols 音乐符号
65.【1D200-1D24F】 Ancient Greek Musical Notation 古希腊音乐谱记号