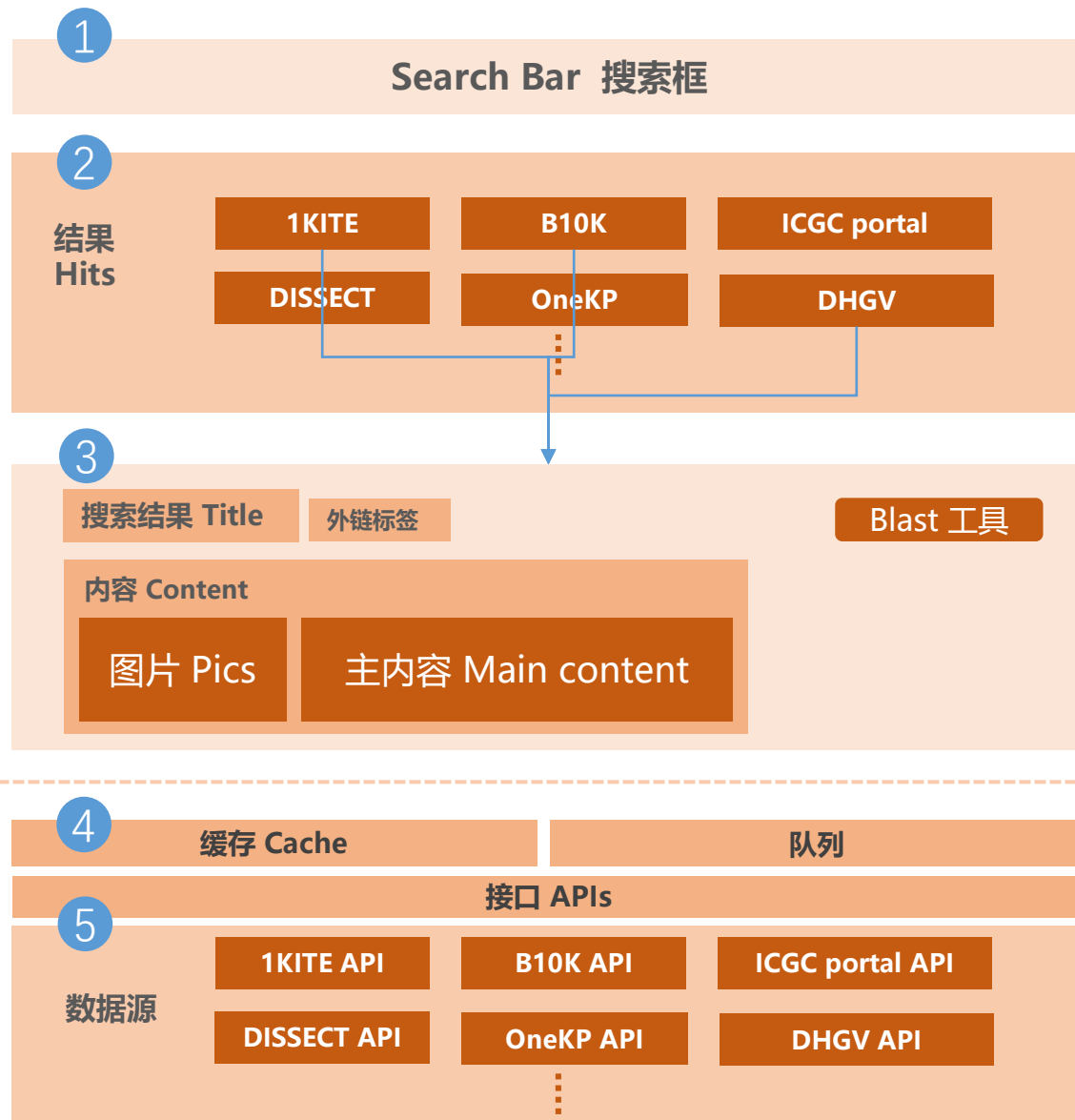


BiomiGO (临时代号) V1.1 文档

Search Lives

I期 搜索页 v1.0版本 db.cngb.org/search



1 Search 搜索框

1. 关键词：基因、突变、染色体位置、样本编号、分类树（物种拉丁名）和序列。
2. 过滤：序列过滤、词根的提取和Catalogue of Life 物种名过滤。
3. 自动补全：检索频次高的关键字，提示的优先级也更高。

2 Hits 结果

1. Hits：各数据库能够匹配到关键词的条目数，包括>0，0，loading三种Hits；loading无响应部分（超时一段时间），会将搜索请求放入队列中进行搜索，并将返回的数据存入缓存，以便下次检索

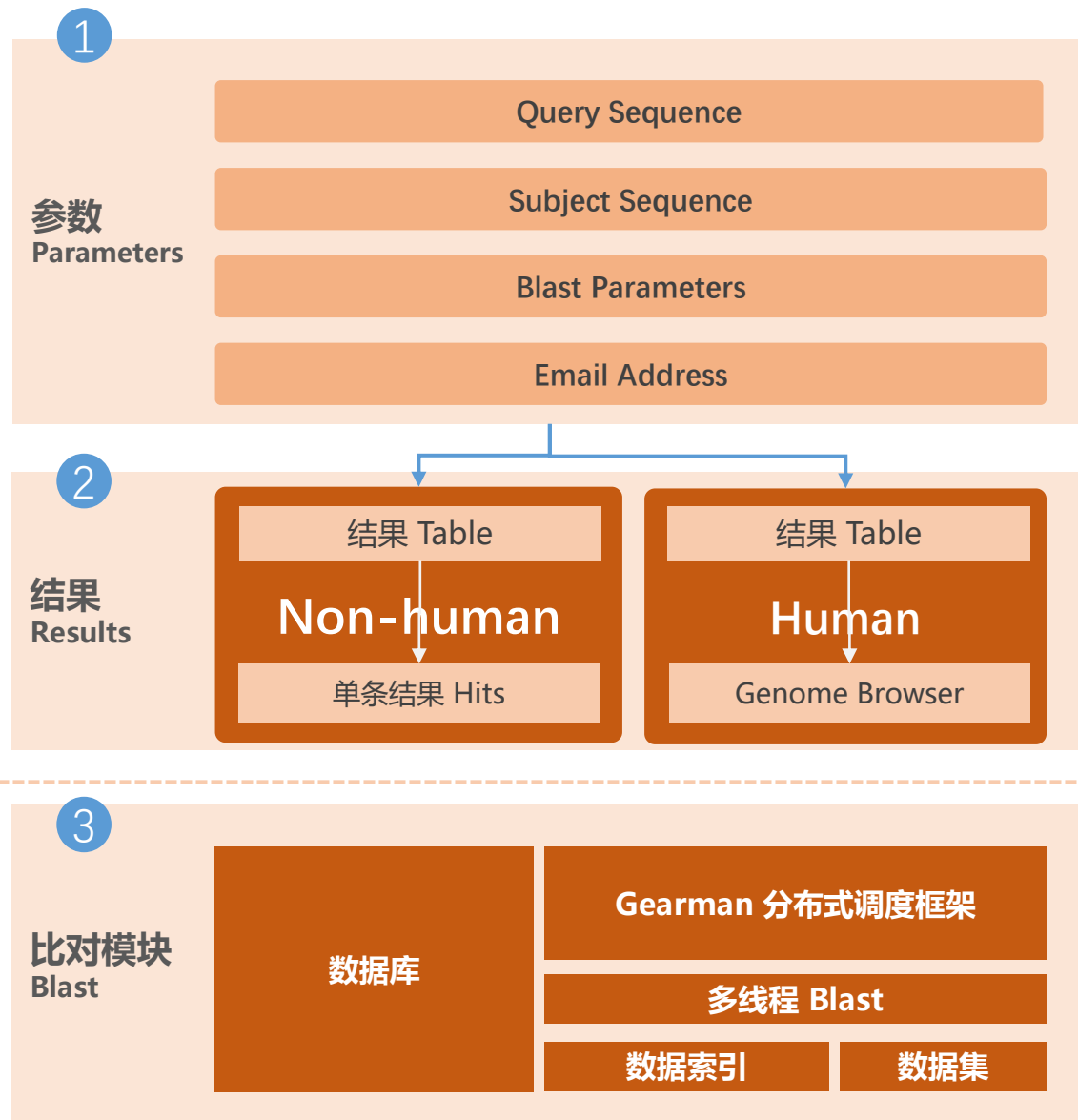
3 Detail 各数据库详细页面

1. 详细条目：点击Hits结果页中的数据库进入
2. 搜索结果：物种、基因、突变和疾病部分附上NCBI, GeneCards, dbSNP, WIKI的外链标签和部分缩略内容
3. 图片：部分物种附上WIKI的图片
4. 主内容：各个条目在数据库中的核心元数据
5. 比对链接：点击进入Blast比对页面，目前会根据所在详情页对应的数据库类型，在Blast工具中预选相应的数据库

4 Data 数据源

1. 数据库缓存
2. 队列
3. 各数据库API

I期 比对模块 v1.0版本 <http://db.cngb.org/search/blast/>



1 参数 Parameters

1. 提交序列：主要以fasta格式
2. 比对序列：根据选择的比对工作类型，筛选可选的数据集及相关项目
3. 比对参数：目前参数只有E-value、Alignment view和Alignments Number
4. 邮箱通知：可选择在比对程序跑完后，通知用户

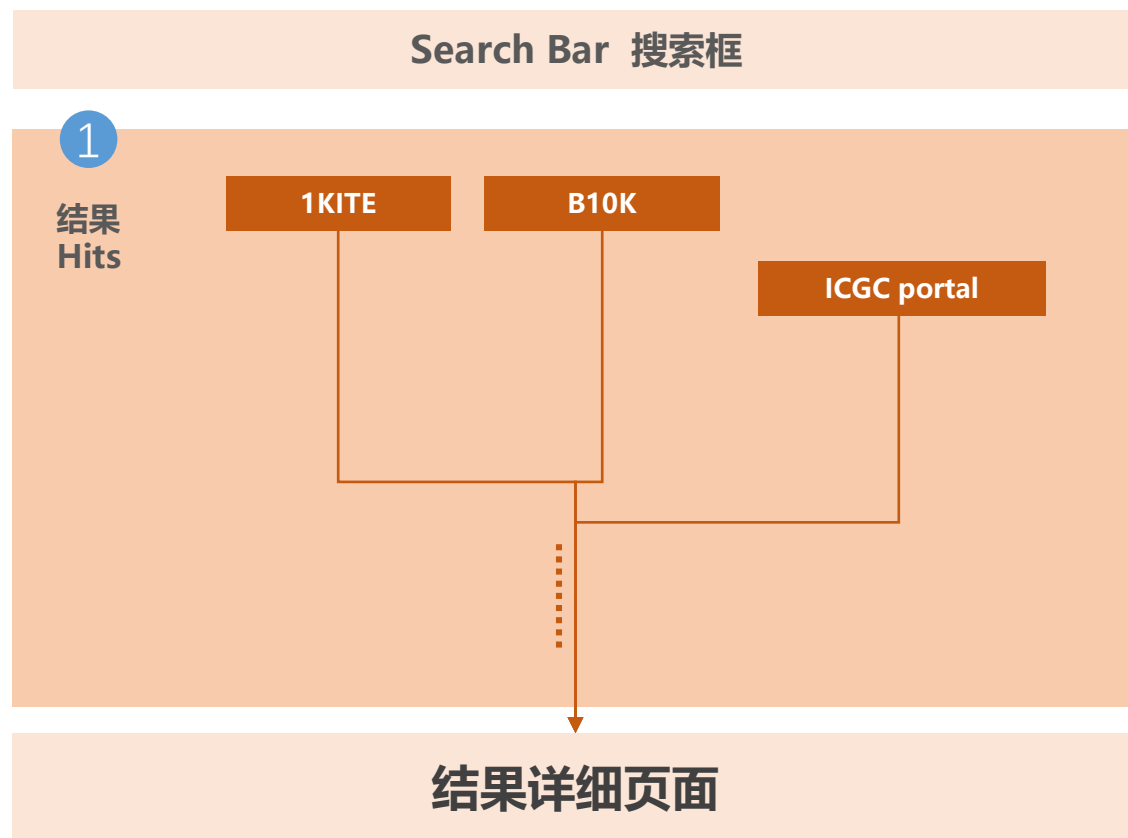
2 结果 Results

1. 人方向数据集：点击结果，可以以Genome Browser形式呈现前10条Hits
2. 非人方向数据集：点击结果，直接以Blast 结果文件的形式呈现

3 比对模块 Blast

1. 数据库：保存用户的比对信息
2. Gearman：分布式调度比对任务
3. Blast：多线程，Blast数据索引和数据集

I期 搜索页 v1.1版本



1 Hits 结果

1. 数据库 LOGO : 更新较清晰LOGO 或 直接去除LOGO (db.cngb.org/search版本)



2 Data 数据源



1. 更新研究院数据库APIs (部门数据库已经更新和添加搜索功能)

I期 比对模块 v1.1版本 <http://db.cngb.org/search/blast/>

1

参数设置页 Parameters

2

ID	Date	Tool	Status	Query Match	Time	Dataset	Output
				 			Browser Download

3

对应Id的详细比对参数

结果页 Results

4

Hits	No.	<<	<	1	2	3	...	12	13	14	>	>>	Go to	No.
Score	27462 bits(14871)	Expect	0.0	Identities			Gaps			Strand				
				14871/14871(100%)			0/14871(0%)			Plus/Minus				
Query	1698	ACGAAATGTTAAAATGATGTATATACACACACTTAGGGCATGTAAATATACCATTAGAGA	1757											
Sbjct	16483	ACGAAATGTTAAAATGATGTATATACACACACTTAGGGCATGTAAATATACCATTAGAGA	16424											
Query	1758	TTTTCTGTTTTCGGGGGTTTTTCAGGAACGTTAGTGATCTCAGgggtatagggggtagg	1817											
Sbjct	16423	TTTTCTGTTTTCGGGGGTTTTTCAGGAACGTTAGTGATCTCAGGGGTATAGGGGGGTAGG	16364											

Genoverse Genome Browser

1

参数 Parameters

1. 完善 Blast参数：参考NCBI
2. 完善 数据集：公开的数据集和IZBox上可用数据，还有GigaDB数据库

2

3

4

结果 Results（左边设计仅做说明之用）

1. 结果表格：Title做部分的文字调整；Query Match部分可换成更易辨识的带颜色icon
2. 详细比对参数：点击表格，在底部呈现相应ID的详细比对参数
3. Hits 导航：尝试将Hits做成导航（可自定义填写），加载的性能部分可以根据实际进行优化
4. Hits结果：比对上线粒体（chrM）数据时，暂时不显示 Browser，只有Blast结果部分（目前的Browser直接把chrM部分的hit直接过滤，不太科学）；或者是将Browser中绘制染色体的数据补全

比对模块 Blast

II期 搜索页 V2.0 版本



- 1

搜索框 Search Bar

1. 搜索语法：提供Syntax，类似 AR@mutation (AR 是基因名称)，AR@sample ... 以便定位搜索。2.0 版本会先引入NCBI部分数据做尝试，需先了解NCBI的检索语法

2. 序列搜索：对序列（核酸和氨基酸）关键字的过滤，并提供以下两种跳转方式

2.1 尝试序列搜索，主要针对NCBI页面中的序列（例如基因页面GBFF格式中的CDS）

2.2 跳转至Blast页面
- 2

3

详细结果页

1. 结果Hits：左侧Hit结果呈现与目前的V1.0版本的类似。在第一个结果，呈现的信息可以更为详细和具体（如果该结果具备详细信息）

2. 结果过滤：右侧提供结果过滤的高级选项（涉及多个领域，选项部分需要再讨论商榷），以便在左侧的搜索结果中筛选出用户感兴趣的部分
- 4

搜索引擎 数据

1. 利用I期的数据和NCBI的数据搭建全文本的搜索引擎

2. 其中涉及的技术较多，需做详细调研和数据测试

II期 比对模块 v2.0 版本

参数设置页 Parameters

1

结果页
Results

1

结果页 Results

1. 增加更多的注释信息，来自华大研究院已有数据和外部的信息，尝试整合更多开源的可视化工具。

2

比对模块 Blast

2

比对模块

1. 比对效率优化：Blast集群

II期 搜索页 V2.1 版本



1



1

搜索引擎 数据

1. 引入更多的外部数据源，整合到搜索引擎中

II期 搜索页 V2.2 版本

Search Bar 搜索框

详细结果页

搜索框

1

结果 Hits

Filter 结果过滤 ▼

1

结果Hits

1. 数据的开放与否：很多科研工作者本身没有数据，想通过外部的数据来进行分析。可以在结果的数据部分注释 open 或者 restrict，方便用户查询和跳转到感兴趣的数据。目前有公司 repositive.io 将此工作做成独立的产品。

搜索引擎

内部数据源

外部数据源

2

2

外部数据源

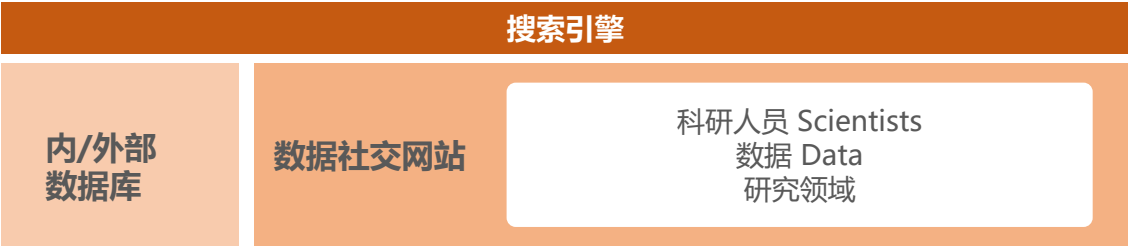
1. 添加数据的权限标签

III期 搜索页 V3.0 版本 (设想)

(在其他小组将数据社交网站开发完成后)



1



1

搜索引擎 数据

1. 将数据设计网站的数据整合进搜索中，主要是搜索 核心的科研人员

III期 比对模块 v3.0 版本 可视化模块（设想）

参数设置页 Parameters

1

结果页
Results

比对模块 Blast

1 结果页 Results

1. 比对结果注释的可视化：将生物信息中的可视化图进行整合成一个工具集，细节还需要再调讨论。可以针对人数和非人数数据做开发。

III期 文本（文献）注释 v1.0 版本（设想）

（对外提供类似 Google translate 的工具，可以独立于搜索引擎，也可以整合进搜索中）

提交文本

1

We further evaluated the mechanism underlying the top hit, PCAT1, and found that a risk-associated variant at rs7463708 increases binding of ONECUT2, a novel androgen receptor (AR)-interacting transcription factor, at a distal enhancer that loops to the PCAT1 promoter, resulting in upregulation of PCAT1 upon prolonged androgen treatment.

注释结果

API

2

We further evaluated the mechanism underlying the top hit, **PCAT1**, and found that a risk-associated variant at **rs7463708** increases binding of ONECUT2, a novel **androgen receptor (AR)**-interacting transcription factor, at a distal enhancer that loops to the **PCAT1** promoter, resulting in upregulation of **PCAT1** upon prolonged androgen treatment.

PCAT1 Gene(RNA Gene)
Prostate Cancer Associated
Transcript 1 (Non-Protein Coding)

1 提交文本

1. Pubmed ID: 提取文献中的摘要作为输入
2. 直接复制黏贴文档
3. 上传文档（格式问题待讨论）

2 注释结果

1. 注释的部分包括基因、蛋白、疾病、SNP等，注释信息来源于各个数据源
2. 对外提供注释的API

附录：早期 概念草图



附录：II, III期 搜索页 V2 版本 关键字 Keywords

(下面关键词以及语法待更详细调研和商榷)

单词：

1. 纵向搜索 (Horizontal)

提供类似 Tag 的方式，例如
AR@mutation (AR 是基因名称)，
AR@sample ...

2. 横向搜索 (Vertical)

搜索框提供 模糊匹配 和 自动补全

句 / 段落：

1. 语义分词 (Semantic) & 正则匹配

提供句/段落的分词搜索，例如 We further evaluated the mechanism underlying the top hit, **PCAT1**, and found that a risk-associated variant at **rs7463708** increases binding of ONECUT2, a novel **androgen receptor (AR)**-interacting transcription factor, at a distal enhancer that loops to the **PCAT1** promoter, resulting in upregulation of **PCAT1** upon prolonged androgen treatment.

搜索结果将是对这段文字关键词的注释

1

Search Bar 搜索框

1. Words 单词：

1.1 Phylogenetic Tree 分类树

界 (Kingdom)、门 (Phylum)、纲 (Class)、目 (Order)、科 (Family)、属 (Genus)、种 (Species)

1.2 Mutations 突变

DNA Level、RNA Level、Protein Level (genomic, cDNA, mitochondrial, RNA, protein) | SNP rs No.

1.3 Diseases 疾病

1.4 Samples 样本

1.5 Gene 基因 | Protein(Amino Acids) 蛋白 (氨基酸) | RNA ...

1.6 Research area 研究领域

Researcher/ Scientist 相关科研工作者，Projects 相关大小型科研项目

1.7 生物序列

2. Sentence / Paragraph 句/段落 (语义、分词、正则)：

2.1 Papers 文献

Title 文献名，Abstract 摘要

2.2 Others 其他

附录：排序 (Rank 网页质量和相关性) (仅作为讨论的基础)：

1. 高质量的网页 (Authority)

Authority页面，某个领域或者某个话题相关的高质量网页，类似NCBI，GeneCard，Wiki，COSMICS ...

2. 高质量Hub页面

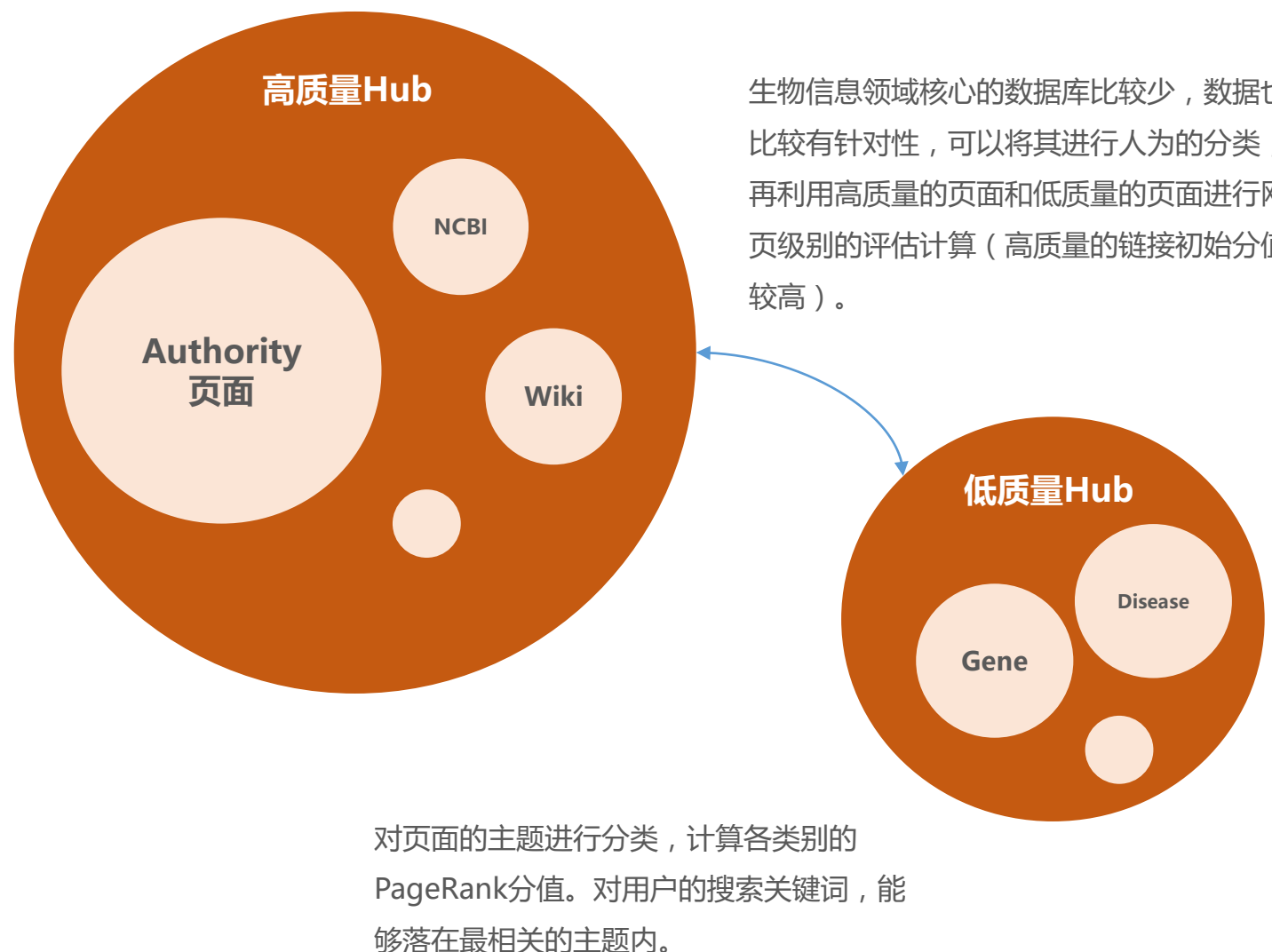
高质量的页面组成Hub页面

3. 根集(Root Set)

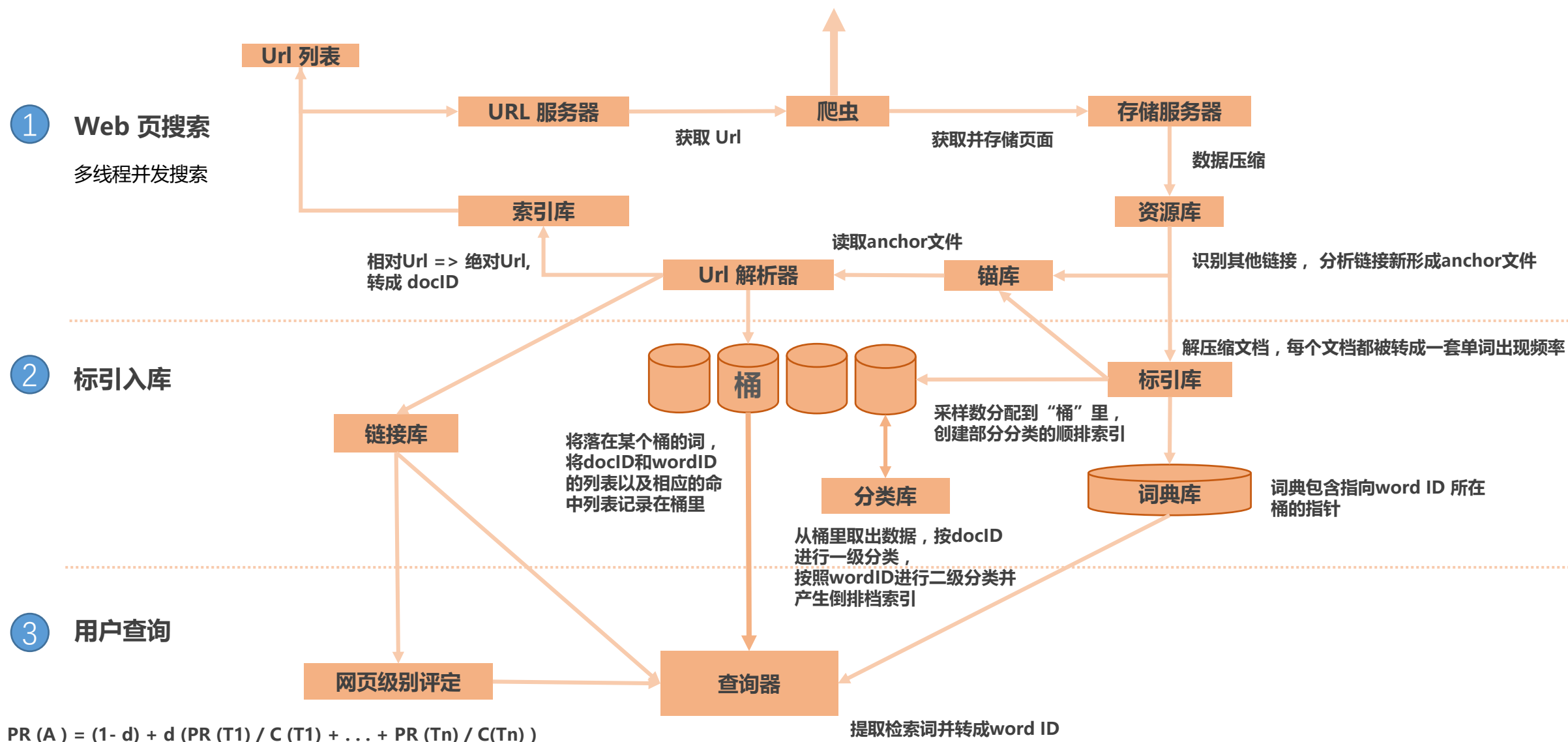
HITS 算法收到用户查询后，将查询提交给搜索引擎，并在返回的搜索结果中，提取排名靠前的网页，得到一组和用户查询高度相关的初始网页集合，成为根集。

4. 主题敏感

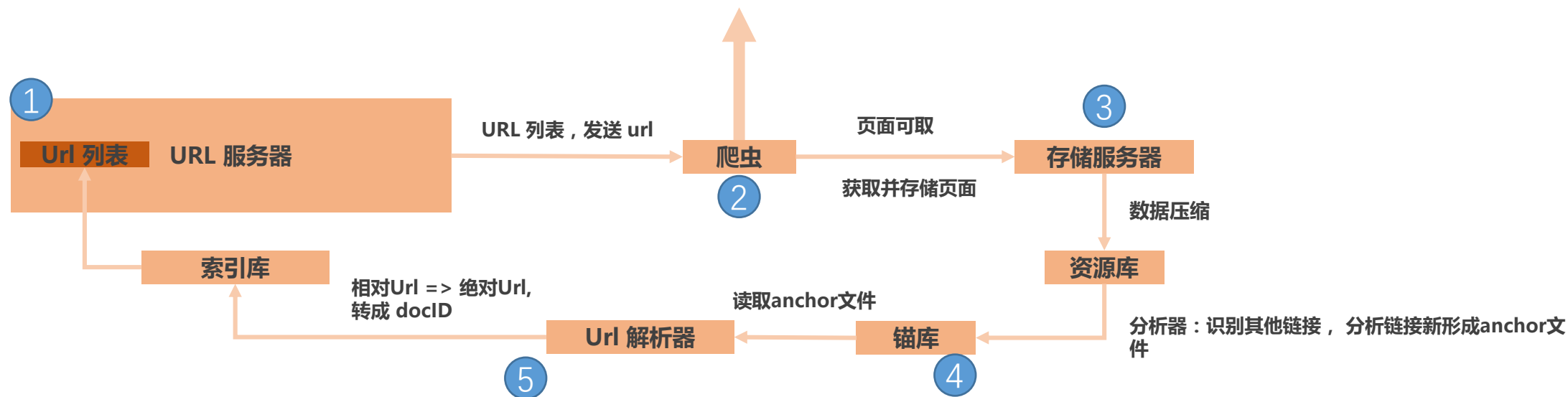
用户的研究领域，搜索关键词的领域、项目等。对各个类别主题的内容，分别依次计算该类别的PageRank分值。



附录：Google 基础架构草图 1

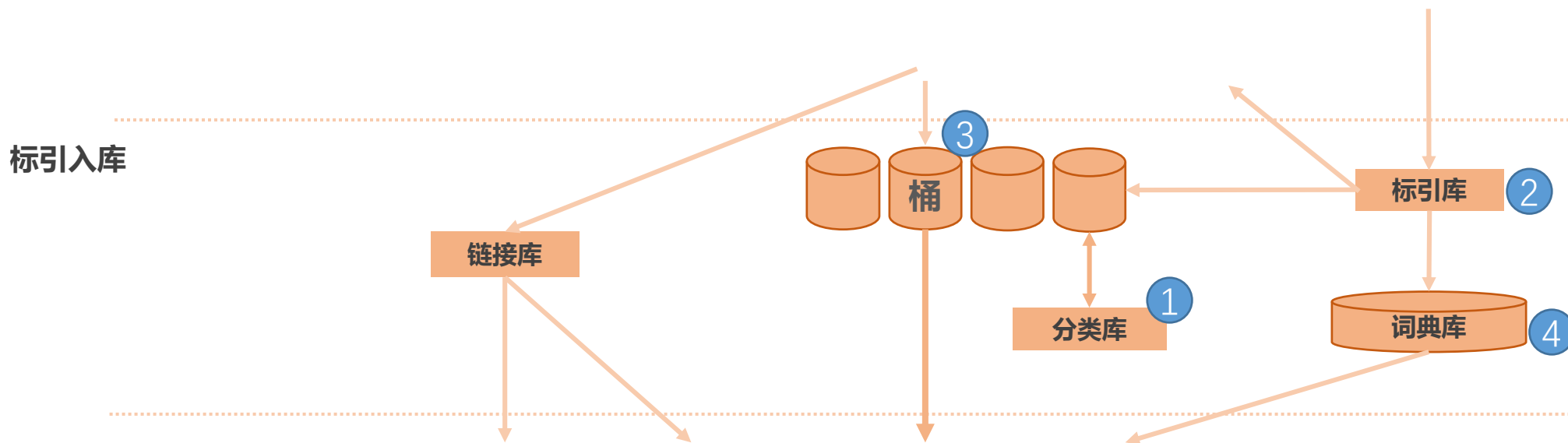


附录：Google 基础架构草图 2



- 1 Url 服务器**：整个Web页搜索模块的开始，用来管理和维护Url 列表。将Url发给爬虫服务器，从文档索引库中不断获取新的Url
- 2 爬虫服务器 Crawler**：分布式爬虫，对爬过的页面添加标记。
- 3 存储服务器 + 资源库**：存储器将爬虫抓取的页面进行压缩并存储到资源库。资源库对文档进行归类，加上docID前缀、文档的长度和Url。文档索引记录着每一个文档的信息，是一个固有长度，经docID排序的ISAM索引。如果文档被爬过，有一个docinfo的文件包括文档的Url和标题，否则指针只包含URL的Url 列表
- 4 分析器 Parser**：分析每个Web页的所有连接并把相关的重要信息存储在Anchors文件里，构成一个锚库。每当从Web页分析出一个新的Url时，为每个Web页分配一个成为docID的关联ID。这个文件包含足够的信息来决定每一个docID的关联ID
- 5 Url 解析器 (Url Resolver)**：把相对Url转成绝对Url，然后依次转成docID。把Anchor文本指向顺排索引，存到文档索引库里，并用Anchor所指向的docID进行关联。把Url转换成docID的文件，是由Url校验和及相应的docID两列组成的一个列表，并以校验和排序

附录：Google 基础架构草图 3



- 1 分类器 Sorter：** 分类器从桶中取出数据，按docID进行以及分类，然后按照wordID进行二级分类并产生倒排档索引。
- 2 标引器 Indexer：** 读取数据库，解压缩文档然后进行分析。每个文档都被转成一套单词出现频率，称之为采样数。采样数记录单词以及在文档中出现的位置，字体的大小以及大小写信息。标引器把这些采样数分配到一套“桶”中，创建一个部分分类的顺排索引。
- 3 桶 Barrels：** 每个桶都存着wordID的归类，包括顺排档和倒排档。如果一个文档包含落在某个桶里的词, docID 和word ID 的列表以及相应的命中列表就被记录到桶里。Google 存储每一个word ID 时, 存储的是与所在桶的最小word ID 的相对差异, 而不是存储实际的word ID。
- 4 桶 Barrels：** 词典由两部分实现，词表和指针的哈希表。对于每个有效的wordID，词典包含指向wordID所在桶的指针。