搜索结构研讨

第五次讨论课

搜索引擎技术研讨

目录

CONTENTS

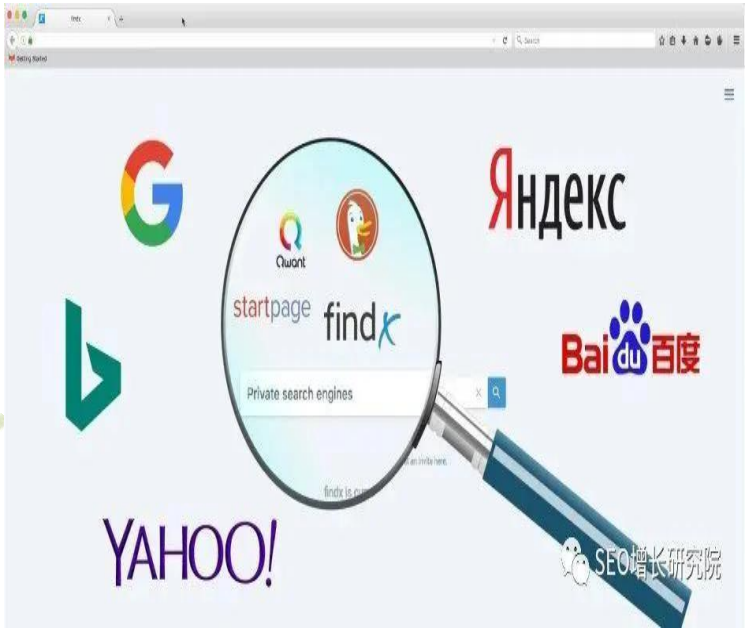
垂直搜索

引入

什么是搜索系统：

搜索系统的产生，来源于人们对高效的在海量内容中寻找信息的需求。常见的搜索系统有百度、Google等，搜索也成为用户使用互联网时绕不开的一个功能。

看新闻时，有什么要了解的顺手用百度搜一下，如果想找某个领域的高质量解读文章，会去微信搜一搜有没有权威公众号发的文章。



当然，在图书馆找书籍，在word中用ctrl+f查找，也可以看成是个搜索系统。同时搜索也成为了产品中必不可少的重要功能。

搜索系统改变了用户和信息之间的关联方式，用户和信息在哪里，搜索的行为就在哪里。除了常见的用文本去搜索，一副图像，一段语音甚至一段视频也是一种搜索的输入形态。

搜索引擎的分类

1、全文索引型

全文搜索引擎，国内是著名的百度搜索引擎。国内著名的有百度（Baidu）国外则是Google。它们从互联网提取各个网站的信息(以网页的文字为主)，建立起数据库，并能检索与用户查询条件相匹配的记录，按一定的排列顺序返回结果。

2、目录索引型

目录索引虽然有搜索引擎功能，但严格意义上不能称为真正的搜索引擎。用户完全不需要依靠关键词（Keywords）查询，只是按照分类目录找到所需要的信息。目录索引中，国内具代表性就是新浪、搜狐、网易分类目录和Yahoo网站雅虎。其他著名的还有Open Directory Project（DMOZ）、LookSmart、About等。

3、元数据索引型

元搜索引擎接受用户查询请求后，同时在多个搜索引擎上搜索，并将结果返回给用户，著名的元搜索引擎有360搜索、infoSpace、Dogpile、VIsisimo等，在搜索结果排列方面，有的直接按来源排列搜索结果，如Dogpile，有的则按自定的规则将结果重新排列组合，如Vivisimo。

4、垂直索引型

垂直搜索引擎适用于有明确搜索意图情况下进行检索。例如，用户购买机票、火车票、汽车票时，或想要浏览网络视频资源时，都可以直接选用行业内专用搜索引擎，以准确、迅速获得相关信息。

下面我们详细介绍垂直搜索引擎

垂直搜索

1

垂直搜索-是什么技术

一、什么是垂直搜索：

论搜索系统的发展史：

从产品形态上，搜索引擎可以分为三代。第一代是早期PC互联网的产物，以Google、百度为代表的通用搜索。自身没有内容，而是通过爬虫去爬取全网海量信息进行索引，提供搜索功能。第二代则是基于内容平台和移动互联网的产物，爆发出了拥有丰富优质内容生态的垂直搜索，以Facebook、微信、知乎、高德地图为代表。搜索结果也不仅限于文档，也可以搜索朋友、公众号、位置等。第三代引擎是基于人工智能的产物，不限于有框输入，从搜索的本质出发，相比于返回相关文档列表，而是转变到返回更直接有效的信息答案。目前还没有比较有革命性的产品，但是在通用搜索中，搜索直达、直接问答（google也叫精选摘要）的触发率越来越高，这也是算是向第三代搜索引擎的尝试。

我们总喜欢用一个搜索引擎搜索所有的动议，最常用的就是百度，以为这样很方便，但是这就导致搜到的信息量过大，而且不够专、精、深。

利用垂直搜索就能很好的避免这个问题，而且近些年，越来越多的垂直搜索网站出现，极大的方便了我们的生活。

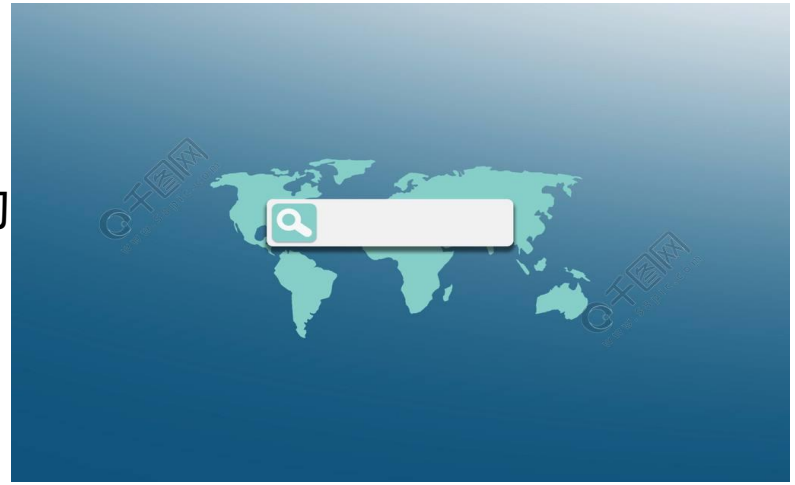
那么什么是垂直搜索呢？垂直搜索就是搜索范围并不是包罗万象，是针对某一领域、某一方面进行的资源统一整理管理。我们很多时候都是模糊的知道自己想搜什么，搜到什么完全是碰运气，垂直搜索更能满足我们在某一方面的需求。

垂直搜索

2

垂直搜索-为什么用垂直搜索

二、为什么用垂直搜索(1)



垂直搜索存在的意义：

很多人，其实也包括我在一段时间，在质疑垂直搜索引擎的存在意义：都有了那么好的通用搜索引擎（如google baidu），还需要所谓垂直搜索引擎干吗?

其实，这里面有误解。

一、所谓通用搜索引擎，并不能够囊括所有的网页。
据googl的人说，也就猜测覆盖了40%不到的网页，也就是说，更多的网页是没有被通用搜索引擎收录的，也就谈不上被搜到了。那些没有机会收录的网页，有些是需要身份验证等之后才可以看到，有些是根本未被通用搜索引擎的蜘蛛爬到。这些信息却往往是宝贵的，更有价值的。

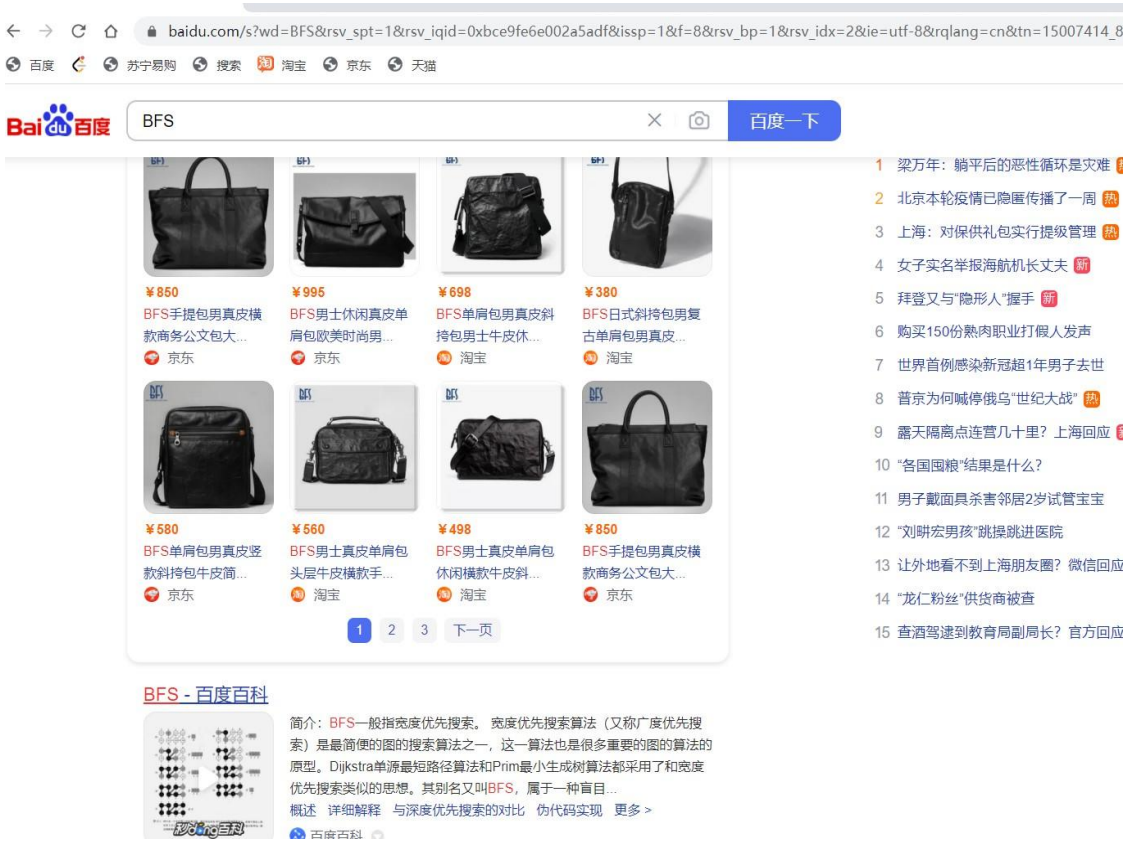
二、每一个行业都是复杂的，从目前计算技术来讲，还是遵循冯.诺依曼的体系，也即是说还是依靠图灵未实现的人工智能之下的计算机逻辑来处理信息，在搜索收录的分析过程中，如果不加上行业特点和特性 进行分析，很难说会更准确分析到网页的重要性和分析的准确。这个也是垂直的意义所在。当然，这里面也需要注意到，并非你垂直了，你的搜索收录和搜索结果就一定比通用搜索更准确。

二、为什么用垂直搜索(2)

三、搜索由通用到专业目前来看是一趋势，都分了什么图片搜索、mp3搜索之类，这也好理解，用户输入关键字的时候，可能并不需要其他行业的内容，仅仅一个关键字不结合其他补充信息，是无法准确分析用户的搜索要求的，但是通用搜索引擎只能够一股脑给你信息。从这个角度讲，信息多了会造成一部分搜索结果是垃圾，而这垃圾会影响用户的感受，以及继续试用搜索的兴趣。而垂直搜索引擎应该可以更好的做到理解垂直用户的需求，从而给出更好的结果。

四、从搜索信息的结果来看，除了上面的垃圾会过多外，还会存在信息不符合要求的情况，有时候用户搜索某类事物，并以此作为关键字，他需要的是关于这个事物的数量、价格等 甚至相关比较信息，而通用搜索引擎只能给你线索，给你网页。通用搜索引擎由于自身巨大，他做不到更深入分析后给出更符合行业、用户需求的结果。

无关信息过多



垂直搜索

1. 数据特点
2. 搜索特点
3. 垂直搜索与通用搜索的区别
4. 关键技术
5. 引擎的优缺点
6. 应用领域
7. 价值

三、垂直搜索的数据特点：

（1）数据来源

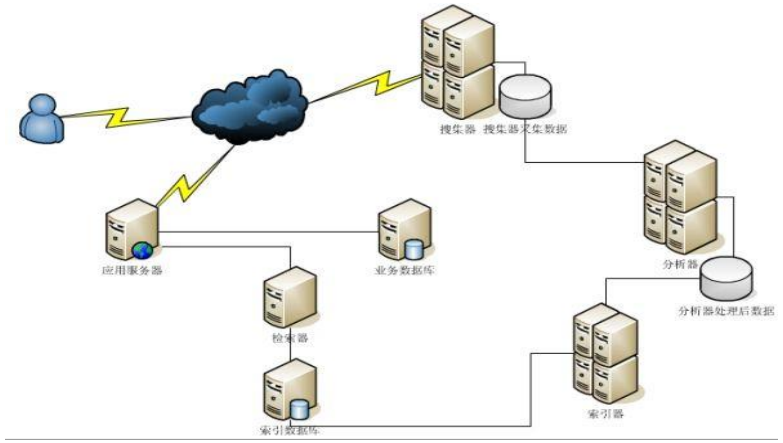
垂直搜索引擎的数据来源有两个方面：

①来源于所处行业的相关站点。

②来源于自身平台：来源于自身平台的搜索常被认为是“站内搜索”。但是，当某一平台上的信息达到足够量大的时候，其实就是一种垂直搜索。况且，垂直搜索本身就是从这些行业站点提取出数据的。

（2）数据特性

垂直搜索引擎的数据倾向于结构化和格式化。例如，在某个购物类的垂直搜索引擎上输入“MP3”就会出现该产品的相关属性，如内存、尺寸、大小、电池型号、价格、生产厂家等相关技术属性，有的还提供比价服务。在某餐饮搜索引擎的高级搜索针对一家餐馆的搜索属性设置多达300个选项，把想到和没想到都列出来了，这就把搜索服务专业化、细致化、个性化了。



四、垂直搜索的搜索特点(1)：

①实时性：垂直搜索引擎需要获取的信息来自于某一特定领域，这比起通用搜索引擎漫无边际的信息抓取，有一个非常大的优势，那就是信息的实时性。由于互联网上的信息量非常巨大，通用搜索引擎的数据更新周期短则十几天，长则几个月，而垂直搜索引掌的数据更新完全可以以秒为单位。

②数据挖掘分析、BI、报表：行业的历史发展、最新动向、趋势都是行业从业人员非常关注的话题。垂直搜索引擎集中了行业海量的信息和数据，基于这些信息和数据的商务智能分析，将为行业创造非常有价值的信息增值服务。

③个性化、社会化；查询服务只是垂直搜索引擎的一部分，垂直搜索引掌在用户的个性化方向的发展非常重要。垂直搜索引擎不能只提供一个窗口，它应该是一个用户高度参与交互的社会化平台。这不光是用户粘度、忠诚度的问题，更为重要的是，垂直搜索引擎需要能够获取并且分析用户的偏好信息，从而提供更加完善而且准确的数据服务。

四、垂直搜索的搜索特点(2)：

④智能化语义网：语义网（semantic web）将有可能成为下一代互联网，此类网络上的数据和信息将被计算机程序所理解。这将为垂直搜索引擎提供一个巨大的机会，Spider程序如果能理解网络上的数据，将对信息的收集和整理更加准确和专业，搜索服务的查全率和查准率将更高。

⑤多元化查询：目前的搜索引擎，都只局限于关键字搜索，其中主要的原因是，对用户的查询需求无法建模，无法模式化。而关键字搜索带来的问题是，搜索结果过多，并且不准确。互联网信息量越大，这种情况越严重，可以说是灾难

个性化搜索，根据用户偏好提供相应信息，在某领域内多元化，全面覆盖

五、垂直搜索与通用搜索的区别(1)：

1、信息处理的区别：

垂直搜索引擎和普通的网页搜索引擎的最大区别是对网页信息进行了结构化信息抽取，也就是将网页的非结构化数据抽取成特定的结构化信息数据，好比网页搜索是以网页为最小单位，基于视觉的网页块分析是以网页块为最小单位，而垂直搜索是以结构化数据为最小单位。

然后将这些数据存储到数据库，进行进一步的加工处理，如去重、分类等，最后分词、索引再以搜索的方式满足用户的需求。举个例子来说明会更容易理解，比如购物搜索引摹，整体流程大致如下：抓取网页后，对网页商品信息进行抽取，抽取出商品名称、价格、简介……甚至可以进一步将笔记本式计算机简介细分成品牌、型号、CPU、内存、硬盘、显示屏……然后对信息进行清洗、去重、分类、分析比较、数据挖掘，最后通过分词索引提供用户搜索、通过分析挖掘提供市场行情报告。

在整个过程中，数据由非结构化数据抽取成结构化数据，经过深度加工处理后以非结构化的方式和结构化的方式返回给用户。

五、垂直搜索与通用搜索的区别(2)：

2、信息采集的区别：

垂直搜索引擎技术同信息采集技术不同的是，信息采集主要是将采集的信息导入本地数据库，而垂直搜索引擎主要是以网页的形式展现给用户。

通用搜索引擎主要是利用Spider程序到网络上搜索，一般是某个特定的周期派出一次将网页更新，垂直搜索引擎同样应有Spider程序，但该程序只在一些特定的网络上爬行，并不会对每一个链接都感兴趣。

相对来说，垂直搜索引擎的收录范围大大缩小了，但并不意味着内容的缩小，通用搜索引擎对一些动态脚本是不敏感的。另外，由于目前网页中的链接形式非常多，不但有动态脚本也有Flash做链接，这些链接方式通过传统的Spider程序是很难解析出来的，在垂直搜索引掌中也应该解决

五、垂直搜索与通用搜索的区别(3)：

其他区别：

1.用户结构的区别

一般来说通用搜索引擎包括了垂直搜索引擎的用户，通用搜索引擎的用户使用率更高，但其用户的忠诚度没有垂直搜索引擎高，其原因是垂直搜索引擎的用户对于专业知识的搜索，要高于通用搜索的泛需求。

2.搜索精度的区别

当然了对于搜索精度而言，仁者见仁智者见智，比如你在一个比较大的网站上使用了其站内搜索，可以得到比较精确的答案，但你所获取的答案只限于该网站，而其他网站的相关信息还是需要根据通用搜索引擎来获取，所以搜索精度是根据搜索者的需求而定的。

3.应用范围的区别

正常来说通用搜索引擎的应用范围要大于垂直搜索引擎，但这又涉及到了用户结构问题，用户如果只是针对某一专业领域的搜索需求，垂直搜索引擎可以满足其大部分要求，当然利用通用搜索引擎的机会就会减少。

六、垂直搜索的关键技术(1)：

由于垂直搜索引擎服务具有其自身的特性，因此其技术要求特点上与通用搜索引擎有很多不同之处，主要有四大关键技术。

1、聚焦、实时和可管理的网页采集技术：

一般互联网搜索面向全网信息，采集的范围广、数量大，但往往由于更新周期的要求，采集的深度或说层级比较浅，采集动态网页优先级比较低，因而被称为水平搜索。

而垂直搜索带有专业性或行业性的需求和目标，所以只对局部来源的网页进行采集，采集的网页数量适中。但其要求采集的网页全面，必须达到更深的层级，采集动态网页的优先级也相对较高。

在实际应用中，垂直搜索的网页采集技术能够按需控制采集目标和范围、按需支持深度采集及按需支持复杂的动态网页采集，即采集技术要能达到更加聚焦、纵深和可管控的需求，并且网页信息更新周期也更短，获取信息更及时。

2、从非结构化内容到结构化数据的网页解析技术：

水平搜索引擎仅能对网页的标题和正文进行解析和提取，但不提供其时间、来源、作者及其他元数据的解析和提取。由于垂直搜索引擎服务的特殊性，往往要求按需提供时间、来源、作者及其他元数据解析，包括对网页中特定内容的提取。

例如，在论坛搜索、生活服务、订票服务、求职服务、风险信用、竞争情报、行业供需、产品比较等特定垂直搜索服务中，要求对于作者、主题、地区、机构名称、产品名称以及特定行业用语进行提取，才能进一步提供更有价值的搜索服务。

六、垂直搜索的关键技术(2)：

3、精、准、全的全文索引和联合检索技术：

水平搜索引擎并不能提供精确和完整的检索结果，只是给出预估的数量和排在前面部分的结果信息（TOPN），但响应速度是水平搜索引擎所追求的最重要因素。在文本索引方面，它也仅对部分网页中特定位置的文本而不是精确的网页正文全文进行索引，因而其最终检索结果是不完全的。

垂直搜索由于在信息的专业性和使用价值方面有更高的要求，因此能够支持全文检索和精确检索，并按需提供多种结果排序方式，例如按内容相关度排序（与水平检索的page rank不同）或按时间、来源排序。

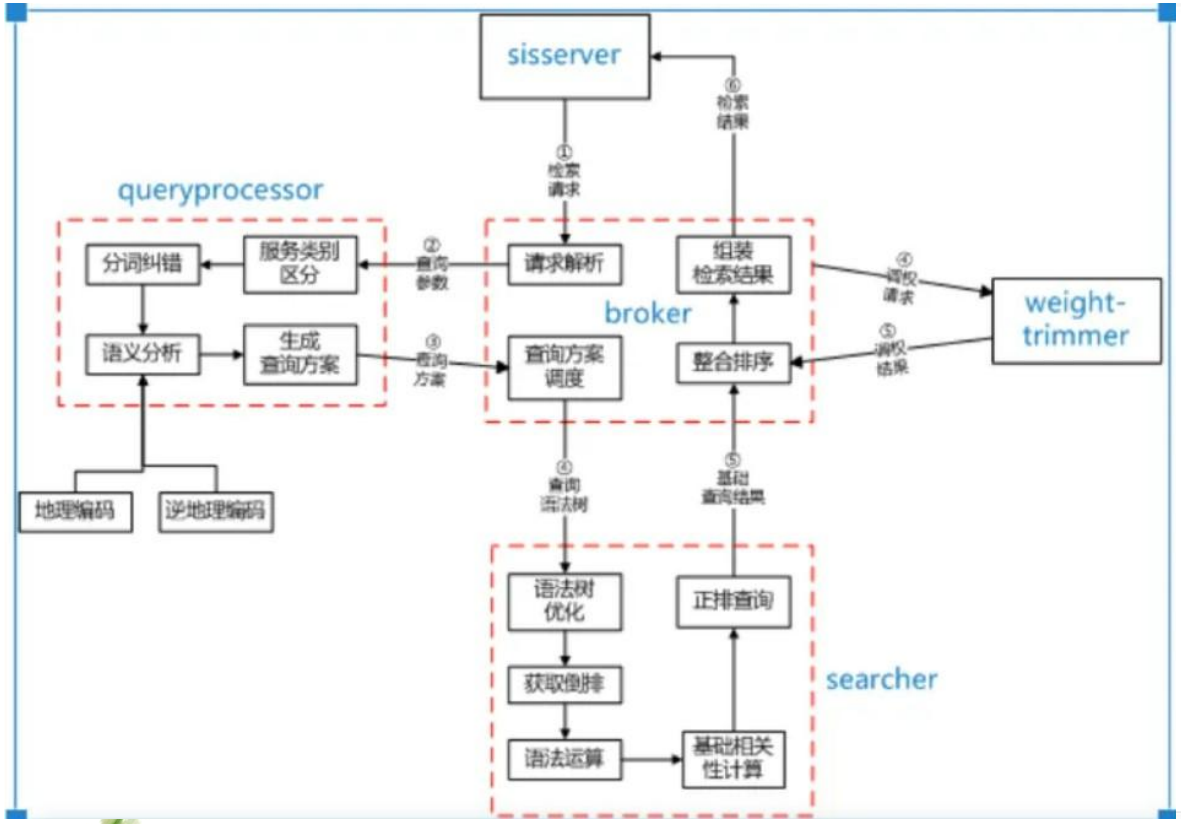
另外，一些垂直搜索引擎还要求按需支持结构化和非结构化数据联合检索，如结合作者、内容、分类进行组合检索等。

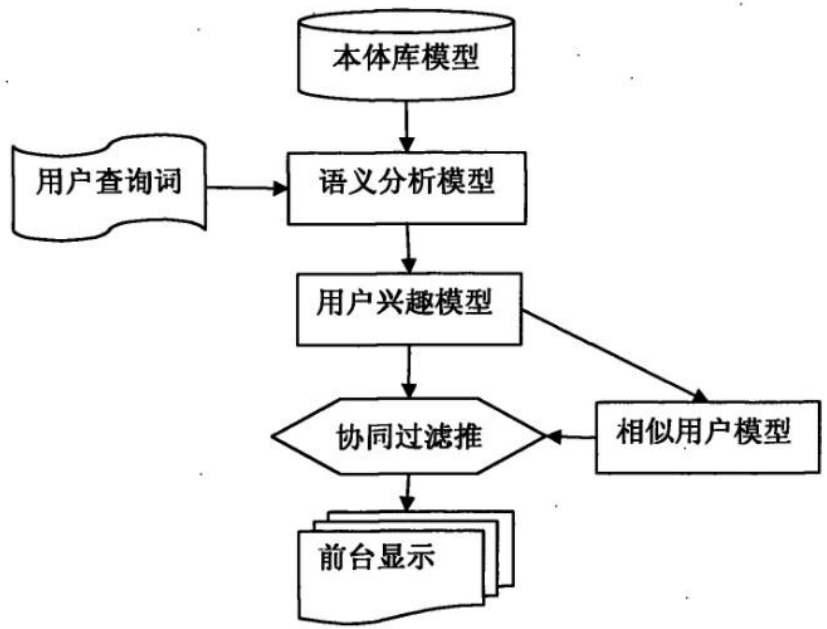
4、高度智能化的文本挖掘技术：

垂直搜索与水平搜索的最大区别是，它对网页信息进行了结构化信息抽取加工，也就是将网页的非结构化数据抽取成特定的结构化信息数据，好比网页搜索是以网页为最小单位，基于视觉的网页块分析是以网页块为最小单位，而垂直搜索是以结构化数据为最小单位。

基于结构化数据和全文数据的结合，垂直搜索才能为用户提供更加到位、更有价值的服务。整个结构化信息提取贯穿从网页解析到网页加工处理的过程。同时，面对上述要求，垂直搜索还能够按需提供智能化处理功能，如自动分类、自动聚类、自动标引、自动排重，文本挖据等。这部分是垂直搜索乃至信息处理的前沿技术，虽然尚不够成熟，但有很大的发展潜力和空间，并且目前在一些海量信息处理的场合已经能够起到很好的应用效果。

六、垂直搜索的关键技术(3)：





七、垂直搜索引擎的优缺点(1)：

在信息多元化时代，垂直搜索的出现及推广必有它过人之处，但是任何事物都有不完美的地方，那么垂直搜索引擎相对于其他类型的搜索引擎有什么优点和缺点呢

垂直搜索引擎的优点：
1，查询速度快，相应时间短。垂直搜索引擎是一种专业类引擎，搜索的信息来自于某一个特定的行业，待搜索的信息范围远远小于通用搜索引擎，再加上拥有先进技术的支持，查询速度快，相应时间短。

2，查询较为准确。垂直搜索引擎的搜索范围局限再某一行业内，信息库在进行录入信息时略去了其他行业的信息，在查询过程中所搜索到的信息均和这个特定行业有关，而通用搜索引擎查询范围大，数据杂，较不为准确

七、垂直搜索引擎的优缺点(1)：

垂直搜索引擎的缺点：

1，我们了解到垂直搜索引擎将采集到的数据存储到数据库，进行进一步的加工处理，如去重、分类等，最后分词、索引再以搜索的方式满足用户的需求。但是这样会导致信息库更新缓慢，垂直搜索引擎采取信息的速度远远比不上网络资源的增长速度，

2，检索到的文献数量有限，对于较为专业偏僻的查询很难提供满意的结果。

3，范畴检索功能有限，不够智能，垂直搜索在某个领域进行搜索，但是大多数情况会出现跨领域的发生，这时，垂直搜索只能进行专业性搜索，而进行范围式检索功能稍有欠缺。

八、垂直搜索的应用领域(1)：

垂直搜索引擎的应用方向很多，比如企业库搜索、供求信息搜索、购物搜索、房产搜索、人才搜索、地图搜索、mp3搜索、图片搜索、工作搜索、交友搜索等，几乎各行各业、各类信息都可以进一步细化成各类垂直搜索引擎。

1、音乐搜索

除必应外，各搜索引擎都提供了音乐搜索服务，支持各种格式的音乐文件的搜索，并提供了各种榜单、音乐专题和挑歌功能。

2、视频搜索

除了谷歌，其他的搜索引擎都有其独立的视频搜索页面，并提供了视频分类搜索。

3、新闻搜索

各搜索引擎都提供了分类搜索，例如，百度提供了国际、国内、体育、娱乐等16个分类的新闻搜索。

4、图书搜索

目前仅有百度与谷歌提供了图书搜索的服务，因为涉及到版权的关系，只有那些已不再受版权保护或出版商已授权搜索引擎的图书，才会提供给用户预览。只有在某些情况下，用户才可以查看全文内容，如公众领域的图书。对于那些无法预览或下载的图书，搜索引擎则提供了借阅或购买该书的渠道。

八、垂直搜索的应用领域(2)：

5、地图搜索

地图搜索一般用于公交、行车路线的搜索，但大多数搜索引擎都集合了其他生活信息的搜索，如餐饮、住宿、出游、企业等信息的搜索。

6、财经搜索

财经搜索主要提供股市报价、资讯、货币汇率等信息的搜索，目前仅有谷歌和百度提供了财经信息的垂直搜索。在谷歌财经的首页上，可以看到各个主要板块的当前行情，将鼠标移到条线图上时可以看到该板块的一些详细的涨跌信息。谷歌还提供了“股票筛选器”的服务，为用户选择投资对象提供了便利。

7、图片搜索

各搜索引擎都提供了图片搜索服务，并提供了内容类型、图片尺寸、文件类型、图片颜色、图片版式甚至图片风格等条件的限定搜索。必应、谷歌提供的是一页式浏览结果，其他几款搜索引擎提供的则是传统的分页式浏览。

九，垂直搜索的应用价值：

垂直搜索从海量的商讯中直接选出用户最需要的供求信息、买（卖）家背景资料、交易方式、服务跟踪等，它既是大量相关产品、企业信息的展示平台，又是行业网站、电子商务的聚合平台，中小企业通过它可获得传统门户网站、通用搜索无法提供的闭合式网络体验，这种附加值就是细分市场巨大的商业价值所在。总结起来，垂直搜索引擎在企业中的应用价值包括：

1.整合企业内外资源，打造企业竞争情报系统的核心引擎企业的竞争情报信息既包括外部的互联网信息、商业数据库信息等，也包括内部的办公文档资料、内部交流信息等。垂直搜索引擎是整合这些内外信息资源的有效手段之一，在资源整合的基础上，形成以情报规划、情报采集、情报加工、情报服务、评估反馈为全生命周期的、完善的、统一的企业竞争情报平台，为企业的风险预警和决策支持提供信息服务。

2.高效采集和组织管理企业内外网门户信息，使信息共享更加便捷、有序随着企业信息化的发展和深入，为了提高企业内部、企业和客户、企业和供应商之间的信息传递和共享速度，加速企业的业务进程，大部分企业（特别是分支机构较多的大型集团性企业）都建立了内外网服务门户，以便通过垂直搜索引擎高效地采集内外网门户信息，为企业职工、客户、供应商提供统一的信息检索入口，并通过权限控制实现安全的检索服务，使得信息的传递和共享更加便捷和有序。

垂直搜索-升华：

“绝对的光明与绝对的黑暗，对一个人来说，结果都是一样的——什么也看不见”，同样，没有信息与拥有无限多的信息，结果也一样——在无限多的信息中，你就无法或难以找到对你真正有用的东西，

垂直搜索引擎的诞生，成为搜索引擎发展史上的一块里程碑。

参考资料

简书https://www.jianshu.com/p/b38675c87b08骨雕

CSDN-shanhe-https://blog.csdn.net/shanhe/article/details/1659002

https://baike.baidu.com/item/%E5%9E%82%E7%9B%B4%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E/210198?fromtitle=%E5%9E%82%E7%9B%B4%E6%90%9C%E7%B4%A2&fromid=214492&fr=aladdin

https://wiki.mbalib.com/wiki/%E5%9E%82%E7%9B%B4%E6%90%9C%E7%B4%A2

https://baike.sogou.com/v62384.htm

THANKS