

Search is
FASTER &
SMARTER



Cooling Search

Cooling Search (“酷灵搜索引擎”) 是 “上海埃帕信息科技有限公司” 在自然语言处理、数据挖掘、人工智能以及搜索等领域历经多年研究与积累的成果，软件产品具有完全的自主知识产权。

语义级的搜索产品

Cooling Search 是一个具备语义分析以及数据挖掘能力的互联网搜索引擎。从传统的搜索引擎的角度来看，它能够提供：

- 快速高效的非结构化数据分析与抓取，保证了向整个互联网探索的能力。
- 支持多种语言，以及各种文本格式。
- 高效的索引机制，保证对海量数据的快速检索能力。
- 高性能，高可用以及可扩展的分布式运行与存贮技术，保证了互联网级的海量信息的存贮能力。

在传统的搜索引擎之的功能之上，它还能够提供：

- 自然语义理解能力，能够区分出自然语言的真实含义，更精准地定位搜索结果。
- 数据挖掘能力，能够对互联网的数据进行进一步的分析，建立起更具备商业价值的数据模型。

Cooling Search 的愿景

Cooling 致力于将计算机变得更人性化、更智能，同时不断改变人类与机器的交互方式。

在目前的企业信息化系统中，关系型数据库成为保存企业信息的主要手段。关系型数据库的优势在于结构化数据的管理，如企业中已经预选定义好的，具备明确标准与格式的各类表单，单据。但实际情况是，大量的人性化的非结构化信息存放于各类办公文档、网页、邮件、即时消息、图象、音频、视频以及一些特定格式的文件中。随着企业规模的日益增大，信息化系统的日益复杂，非结构化信息大量涌现出来，根据 Gartner 研究显示，此类数据以每月翻一倍的速度增长。不少企业都面临到了如何整合与管理这些信息，并发现这些信息中隐含的更有价值的东西。

Cooling Search 认为，消除企业中的信息孤岛固然是一个非常重要的行为，但如何对这些信息进行分析，并得出更有价值的模型，才是重中之重。

非结构化信息的处理

在目前的信息化系统中，信息可划分为两大类。一类信息能够用数据或统一的结构加以表示，我们称之为结构化数据，如数字、符号。这类数据通常可以用二维表结构的形式，存放在关系型数据库中。

而另一类信息无法用数字或统一的结构表示，如文本、图像、声音、网页等，我们称之为非结构化数据，这类数据无法使用传统的关系型数据库来进行管理。

随着信息化建设的不断推进，使得目前的企业中，非结构化信息所占的比重也越来越大。但是，绝大多数企业虽然拥有对结构化信息的处理经验，但对于非结构化信息，还是缺乏足够的理解与经验，使得这些信息虽然丰富，但无法提供更多的价值。整个市场都迫切需要一种非结构化信息的解决方案。

Cooling Search 就是整个解决方案中最核心的产品。Cooling Search 是一个能够同时处理结构化与非结构化信息的信息处理平台，不但能够识别与处理更多格式与标准的信息，同时也能兼容传统的关系型数据库。Cooling Search 的网络爬虫，能够自动抓取、分析并存储企业内部的各类信息，不仅挖掘出更多的信息，也大大提升了信息处理的效率。任何企业都可能通过 Cooling Search 降低成本，并获得更大的竞争力。

自然语言分析

Cooling Search 具备的自然语义分析能力，能够理解非结构信息包含的真实语义，将大大提升搜索服务的质量。目前主流的互联网搜索产品中，都是基于关键字严格匹配的方式来实现，这些搜索产品并不能读懂信息的语义。因而，根据这种方式得出的搜索结果，往往会和用户原始搜索意图完全不匹配。Cooling Search 的搜索服务，能够通过自然语言分析，理解每一篇文档的真实语义，并通过语义结合关键字去搜索特定的信息，使得搜索结果能够准确地符合用户的搜索意图。

机器学习

Cooling Search 的语义理解，建立在机器学习理论之上。

自动化

Cooling Search 将企业内的信息抓取、分析与存取工作完全自动化，大大降低了企业的信息化成本。企业内的信息管理人员，可以方便的通过 Cooling Search 的控制台，管理整个自动化的工作流程，查看信息的流向、状态，并定义最终信息的展现格式。

多语言的支持

对自然语言的识别，是计算机普及后，一个永恒的主题。目前有种主要的识别方法：第一类是希望建立一个计算机可以识别的自然语言语法规则；第二类是希望通过数学方法，建立一个基于统计概率的语法模型。Cooling Search 从最初设计时，就定位于是一个全球性的，能够识别各类自然语言的搜索产品。设计团队为了能够识别更多的自然语言，使用了第二类基于统计概率的语法识别模型。在 Cooling Search 真正实现了独立于语言特性的自然语言处理，在整个处理过程中，词更象是一个非常抽象的符号。这种方式，避免了为世界上的每一门语言建立一个语法规则模型，而是利用搜索产品对海量数据的处理能力，运用字词出现以及共现的可能性来推导出其含义，实现用运算能力来提升准确性。另外，词干提取、“分词”库、非检索用词列表以及 n-gram 算法等专有技术进一步优化了整体的性能与结果的准确率。

多种格式的支持

Cooling Search 能够识别信息化系统中多种格式的信息。由此，Cooling Search 使企业能够发挥各种数据格式和不同来源信息的作用，有效利用

- 非结构化信息。HTML 页面、办公文档、电子邮件、压缩文档多媒体内容等。
- 半结构化信息。自定义的 XML 格式。
- 结构化信息。Oracle、Lotus Notes、ODBC 等关系型数据库。
- 未知格式信息。即使对于一些未公布的信息格式，也仍有可能去识别、分析，并获取一些有价值的信息。

此外，Cooling Search 还提供了整合各种不同类信息源的功能，包括 Lotus Notes、RDBMS、File Server、Web Server 等。

Condition of Search 搜索条件

关键字搜索条件

把关键字作为搜索条件，是目前最主流的一种搜索方式，Google 使用的正是这种方式。用户在搜索框中输入自己需要搜索的关键字，提交后，所有包含该关键字的文档便被当作搜索结果返回给用户。

准确性

当用户能够准确地使用关键字表达搜索意图时，这种方法的正确性是可以保证的。但这种方式是非常机械的匹配，即搜索引擎并不明白这篇文章表达了一个什么样的语义。

有些时候，关键词在文章中出现的位置，可能出现在文章的标题，也可能出现在文章的末尾；关键词在句子中的角色，主语、宾语；关键词的词性，动词、名词，都将影响到该关键词对文章主题的表达能力。针对这一点，关键词搜索方法通常对关键词赋予权重来对搜索结果进行排列。当对比关键词出现在不同位置的两篇文档，其中一篇关键词出现在标题，另一篇关键词出现在文章末尾，搜索方法会认为出现在标题或者其他显著位置，比如文章的前两段的重要性比出现在末尾要高，并给予其较高的权重值。此外关键词出现多次的文档得到的权重值也会比较高。但有时权重也不能解决所有问题。

让我们来看以下这个例子：

“因为英格兰受到暴风雪的侵袭，周中的两场联赛杯半决赛及阿森纳对阵博尔顿的英超补赛都被迫延期。”

全句中没有提到一个“足球”，但任何人都知道这是相关足球的句子。当用户输入“足球”这个关键字时，含有这个语句的文章应该出现在搜索结果中，Cooling Search 完全可以做到这点。

Cooling Search 并不是单纯的只对关键词的严格匹配。在长期的计算机学习中，让计算机了解一句句子真正的含义是 Cooling Search 的优势所在。

学习能力

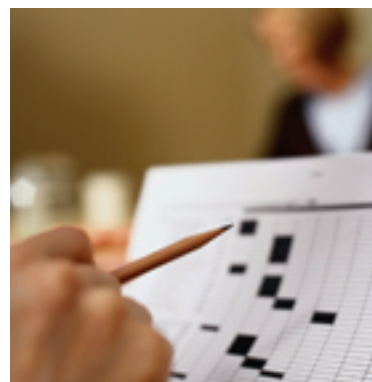
对于语义级的搜索产品，他就像一个小孩子一样，需要大人去指导。但若是等到用户搜索的时候再去学习，那是绝对来不及的。它必须经过一段时间的学习和训练，当中需要人工参与，培训机器认识术语，熟悉特征，并提供机器样本训练，最终实现计算机可以独立准确的分析出句子的语义。

Cooling Search 的语义就是通过人工训练结合机器学习来实现。

我们的优势

Cooling Search 不但可以进行多关键字结合搜索这样的传统搜索，最让人期待的是语义搜索。

Cooling Search 可以完全理解语句的含义给出最符合用户搜索意图的文档。用户不用在众多文档中通过人工选择合适的文档，真正体现搜索的意义。



Semantic Analysis 语义分析

自然语言的复杂性

近几年来，人们一直在研究如何通过语义分析来处理人类的自然语言。但自然语言十分复杂，存在着大量的歧义与不确定性。

举例一：

“我把羽毛球拍卖了。”

这句话可以理解为“我把 羽毛球拍 卖了”。或者理解为“我把 羽毛球 拍卖了”。

举例二：

“An aggressive policy was raised at meeting.”

“aggressive”有侵略性的意思，也有积极的意思，整句话可以理解为在会议上提出了一个侵略政策或者理解为在会议上提出了一个积极的政策。

举例三：

“衣服被放在沙发上，它很干净。”

这句话所要表达的含义到底是“衣服很干净”还是“沙发很干净”呢？

单看三句话，无论是人还是计算机都很难理解。人们依靠上下文关系可以看出正确的组合，可是对于机器来说就很难判别。Cooling Search 在传统的语义分析方法之上，做到了能够像人一样，依据上下文语境来消除歧义，准确地判定句子结构理解句子真正的含义。

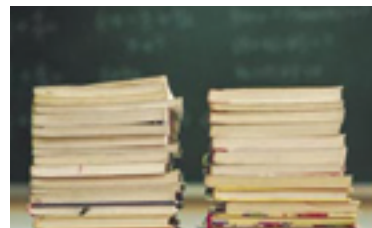
APE 有着多年汉语输入法的开发经验，将输入法的整句输入技术以及多年积累的语法模型应用到 Cooling Search 中。

语言的多样性

世界上有多少种语言？一说七千多种，一说五千多种，一说两千多种，无一定论。使用人口超过 100 万的语言也有 140 多种，语言用语法来连贯，就算在相似的语言，例如：德语与英语。之间的语法也有着细微的不同。如果每一种语言都按照语法来建立模型分析语义的话，无论对计算机或人，都是相当大的工作量。而 Cooling Search 使用的统计模型，不依赖于语法。Cooling Search 所能识别的不是一种文字，而是一类文字，如东亚语系（汉语、日语、韩语），拉丁语系（英语、法语、德语）。

我们的优势

Cooling Search 无论是从语义分析来说，还是从语言的判断分析能力来讲，所使用的理念都是应用最适合的模型解决多个问题。所以 Cooling Search 在能够正确的分析出语句含义的同时也对语言没有任何的限制。这给搜索带来更多的便利。



欲知更多有关酷灵搜索产品及解决方案，请发送邮件至 contact@ape-tech.com

Cooling 酷灵

Sort Order 排序方式

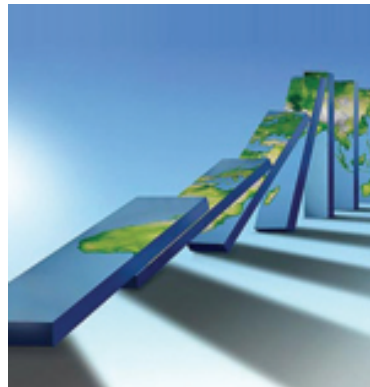
现在大多数提供搜索的网站都会应用排序技术，如淘宝、太平洋等。这可以作为用户搜索的一个辅助条件，帮助用户更直接地搜索到自己所需要的结果。Cooling Search 也同样可以满足用户这样的需求，提供升序和降序两种排序方式。

互联网主要是依靠 Page Ranking，它是一种排列搜索结果的方法。该技术根据链接到某特定页面的其他页面的数量，以及这些其他页面本身的重要程度来判断该网页的重要性。之后该重要性将与用户输入的关键词相结合来搜索“最为相关”的结果。

Cooling Search 可以根据相似度进行排序。这里的相似度主要是指搜索出来的文章和预设的主题之间的相似程度。使用传统搜索引擎，用户往往会觉得搜索出来的文档杂乱无章，有 100% 关联的文档和与仅仅只有 6% 关联的文档排列在一起。而 Cooling Search 则会按照相似的程度对所有搜索出来的文档排序，使用户对自己想要的结果一目了然。

我们的优势

Cooling Search 对相似度的排序是搜索的一大亮点。没有一个用户不需要完美的搜索结果，Cooling Search 就做到了这点，使用语义搜索结合相似度排序，最后呈现在用户面前的是近乎 100% 的准确度。



Vocabulary Self-learning 词汇自学习

词汇自学习常用的方法就是字典方法，这种方法就是收集一组行业特定的术语及其同义词并提供给系统，这样当不常见的词语出现时，系统就能匹配到这些词并了解这些词的含义。字典方法往往收集术语、缩写这类专业强度很高的词。比如航天航空，当有人要搜索关于“神舟一号”具体的重量，材质这些专业信息的时候，系统就会根据收集的术语给出相对应的文档。但字典方法成本昂贵，创建过程过于耗时，且有时其中的定义并不符合语境。字典中的词语列表要靠专家编订。这十分费时，费用也十分昂贵，但是不采用这种方式，无法准确的定义术语和放入准确的类别。Cooling Search 权衡了各种利弊，采取了让机器自己学习词汇的方式。当用户打开一个文档，系统在分析了文章以后，发现新词，并添加到词典中。随着用户使用搜索的内容越来越多，系统所学到的知识也不断在扩展。词典每天都处在更新过程中。

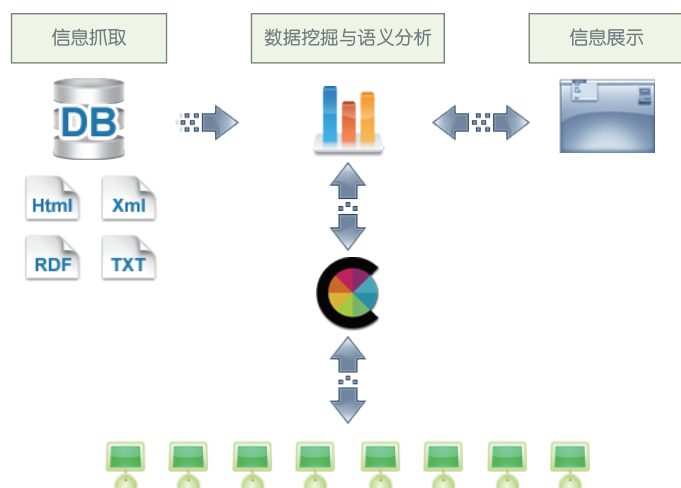
Cooling 输入法已经首先应用了这个方法。比如用户在浏览有关《阿凡达》这部电影的网页，之后用户使用 Cooling 输入法输入“a 'fan' da”时，“阿凡达”将以一个独立词语的形式出现。



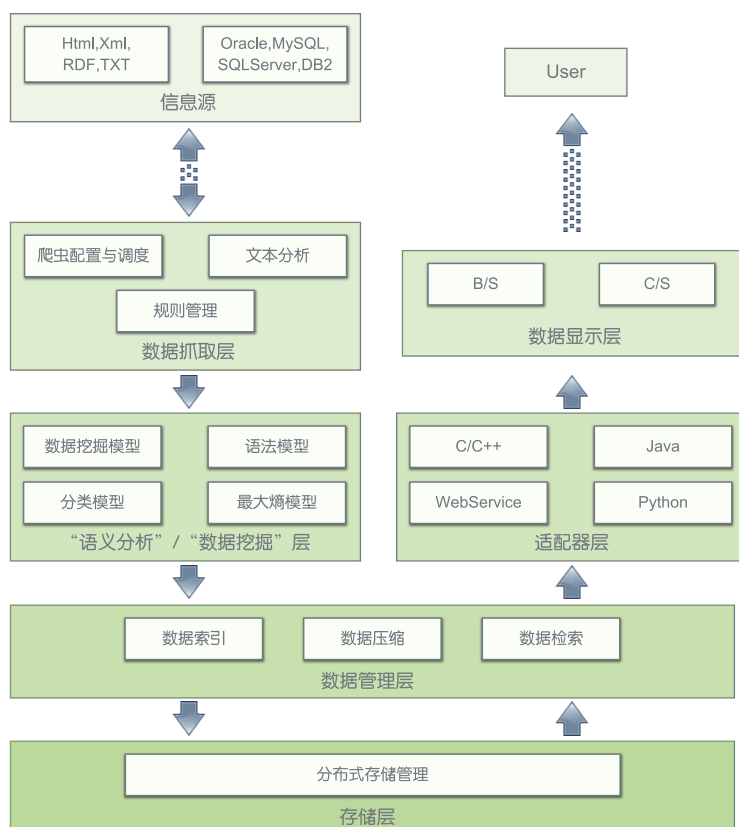
欲知更多有关酷灵搜索产品及解决方案，请发送邮件至 contact@ape-tech.com

Cooling 酷灵

Cooling 搜索引擎产品模型



Cooling 搜索引擎产品架构



Cooling Search 由四个核心模块组成

Spider

Spider 是 Cooling Search 的非结构与结构化数据的抓取与分析工具，更多时候它也被称为“网络爬虫”与“网络蜘蛛”。要建立一个高效的搜索引擎，最首要的任务是提高网络资源的抓取速度与效率，这样才能跟得上互联网信息增长的速度，Spider 在 Cooling Search 中就承担着这么一个角色。在非结构化数据方面，Spider 包含了完整的 HTTP/1.1, FTP, HTML/4, XML, RDF 的实现，能够识别与分析各类互联网文本。在结构化数据方面，Spider 能够支持对 Oracle, SQL Server, DB2, MySQL 等主流关系型数据的抓取与分析。

Egg

Egg 是 Cooling Search 的索引平台，用来保存海量的非结构化数据，并提供基于关键字以及语义的高效检索。Egg 的核心理念是构建一套高效的索引机制，把原始的非结构化数据转化成可供检索的数据结构，并提升检索的效率。

Schola

Scholar 是 Cooling Search 的“语义分析”以及“数据挖掘”平台。Scholar 通过对互联网信息的分析与挖掘，构建了一套完整的自然语言语料库以及基于统计观点的自然语言语法模型，为语义分析提供了可靠的基石。Scholar 目前还在不断地对互联网信息进行分析与挖掘，构建出各类有用的数据模型，力争为互联网应用带来更便捷更准确、更高效的搜索服务。

Platform

Platform 为 Cooling Search 提供了分布式的存贮支持。Platform 架构了一个集群，将搜索引擎前端采集，分析并索引到的信息与模型，切分成数据块，并将这些数据块的多个副本分发到集群中的不同节点之上。Platform 能够为 Cooling Search 提供可靠，更高效以及更易于扩展的存贮结构。

欲知更多有关酷灵搜索产品及解决方案，请发送邮件至 contact@ape-tech.com

Function 功能

关键字搜索

语义搜索

相似度排名

信息整合

自动分类

自定义搜索展现

通用爬虫

搜索建议

自动信息补全

控制台

自动问答

敏感词过滤

关键字搜索

这是一种最常见的搜索方式，把一个或多个关键字作为搜索条件。同时支持多种逻辑运算操作符。Cooling Search 关键词搜索支持多国语言，不受地域的限制。

语义搜索

语义搜索将理解语句含义作为搜索前提。机器通过不断学习，正确对句子进行分词并通过概率计算准确得出切割语句的最佳结果，理解用户搜索语句的含义。Cooling Search 除了对中文语句可以进行语义搜索外，外文也同样得心应手。

对权重的调节

在语义搜索中，Cooling Search 会调节权重，同样一个关键字，在一篇文章的所在位置不同，出现次数不同，使得这个关键字在每一篇文章所占的权重也会有很大的差异。权重是衡量文档是否符合用户搜索的一个很重要的标志。

对词性的认识

Cooling Search 能够对一个关键词在语句中的磁性做出标注，这样系统就能更准确地理解语句的含义，具有以下优点：

1. 更准确地理解句子含义，消除歧义。
2. 给出的搜索结果更加符合用户的需求。
3. 结合权重调节和词性标识，可以有效增加搜索效率。

相似度排名

系统将利用生成索引时为各种概念动态计算出的理论值来评估内容的相似度。相似度可以视为查询文字和结果文档中文字在概念上的吻合度。对相似度进行排序，查找出需要搜索的内容与正文最接近的文章，提高用户搜索的准确性。

使用相似度排名有以下主要优点：

1. 通过排序可以将更符合用户期望搜索的结果提前，从而更加符合用户的满意度。
2. 管理者可以针对各自业务的需求和每个员工的各自角色定义相关度。
3. 准确地实施相似度可以提供搜索的准确率，节省时间。

信息整合

传统的多个数据库搜索，目前给搜索带来了很多问题。传统的数据库关系复杂，使用多个接口连接多个数据库。搜索所用的时间长，往往会出现网络延迟，这是在搜索中致命问题。Cooling Search 采用信息整合方式，将多个数据库整合在一起，只用一个接口搜索各个不同的数据库里的信息。这样做的优点有：

1. 实现搜索的批量性。
2. 搜索的速度快，这是传统数据库所不能达到的。
3. 只需要一个接口，进入所有的数据库。

Function 功能

关键字搜索

语义搜索

相似度排名

信息整合

自动分类

自定义搜索展现

通用爬虫

搜索建议

自动信息补全

控制台

自动问答

敏感词过滤

自动分类

自动生成分类

如今现代企业的类别越来越多，企业内部对专业知识的要求越来越高，对系统的分类要求也越来越高。现在搜索的目录也应该能够处理各种信息格式，包括结构化数据和非结构化数据、HTML、XML等。Cooling Search 能够分析各种格式的信息，同时也创建了各种细化的分类，能够让专业人士节省时间，集中在自己所需要的文档中。

Cooling Search 采用了一种整体式方法生成分类目录，实现了机器自动化和人工手动加入之间的平衡。主要优点有以下几点：

1. 人工参与和修改与机器自动生成并行。
2. 可以分析不同的语言。
3. 具有灵活的管理能力。
4. 用户可以准确了解自己所需要的板块信息。
5. 提高搜索效率。

行业目录库

系统需要提供多个子系统满足用户需要非常精确的搜索结果的需求。Cooling Search 通过不断地调研当前流行的类别划分，将原本的大类例如：经济，体育等细化成更小的单元，例如：微观经济，宏观经济等。使用户搜索到的内容更符合自己的理想标准。Cooling Search 又根据行业分成自己的行业库。将经济分成 29 个小行业，体育分成 32 个小分类，另外还有旅游，天气，科技这些细小分类 12 个。精细的分类和高准确率的优势已经在上海日报等分类系统应用中得以体现。

自定义搜索展现

由于非结构化数据的存储格式非常多样，搜索的结果往往会以多种形态展现，这给不同搜索客户端的显示带来困难；而且，在企业信息化门户建设中，往往会要求数据以统一、规范的形式进行展现。Cooling Search 把搜索结果与展现通过 MVC 的模式进行分离。在服务端，搜索结果通过多种接口形式分发到客户端；在客户端，用户可以通过自定义的模板，自行定义各类展现方式。

搜索建议

在具体的搜索应用中，用户往往会通过一个或多个关键词来表达搜索意愿。由于自然语言天然存在的歧义性，尤其在没有上下文语境的前提下，一个关键词往往会有多种含义。搜索建议能够在遇到歧义的时候，主动提供建议，方便用户缩小搜索结果的范围，提供更精确的搜索结果。

如：用户在搜索框中输入“D700”，这时“搜索建议”会根据该关键词对应的不同含义，询问用户的搜索意图是，照相机、笔记本电脑还是手机。

自动信息补全

当用户在输入搜索条件的同时，自动信息补全会根据目前存在的关键字，自动提示可能匹配用户输入的搜索条件。

Function 功能

关键字搜索

语义搜索

相似度排名

信息整合

自动分类

自定义搜索展现

通用爬虫

搜索建议

自动信息补全

控制台

自动问答

敏感词过滤

通用爬虫

企业信息化系统中往往存在各类异构数据源，这些数据源中的信息多种多样，往往会包含结构化数据、半结构数据以及非结构化数据在内的多种格式。通用爬虫能够整合这些数据源中以不同方式保存的信息。

数据源多样

目前支持的数据源有

• 关系型数据库 • 文档数据库 • 本地文件系统 • 网络文件系统 • Web 服务器 • FTP服务器 •

完全可配置

Cooling Search 自带图形化配置工具。可以新建索引，新建索引组，并对索引和索引组进行管理，可以对索引组选择运行时间，并且查看和下载日志。随时随地监控整个系统的运行过程。

新建索引组中的步骤

基本信息

索引组名称

显示名称

描述

保存

下一步

添加索引

选择索引

index1

→

index1

生成列表

选择	索引名称	索引组名称	更改索引组
<input type="checkbox"/>	索引名称1	索引组名称1	更改索引组

删除

共25条数据 当前第1/3页 首页 上一页 | 下一页 末页

上一步

保存

下一步

新建索引中的步骤

基本信息

索引名称

索引组名称

选择索引组名称

显示名称

描述

保存

下一步

常规设置

爬虫数量

分词规则

☐ 默认

选择字符集

utf-8

选择分词法

中文

☐ 自定义

添加字符集

utf-8

→

utf-8

生成列表

选择	字符集	分词法
<input type="checkbox"/>	utf-8	中文

删除

共25条数据 当前第1/3页 首页 上一页 | 下一页 末页

上一步

保存

下一步

支持的格式

Cooling Search 支持多种关系型数据库、文档、传输协议和编码方式。下表为 Cooling Search 目前所支持的文档格式，传输协议和编码方式。

序号	属性名	描述
1	文档格式	html、htm、asp、jsp、php、pdf、doc、xsl、txt、ppt、不带文件后缀名、xml、aspx、rar、zip、tar、tar.gz、gz
2	传输协议	http、ftp、file只访问本地文件
3	关系型数据库	Oracle、MySql、SqlServer。

欲知更多有关酷灵搜索产品及解决方案，请发送邮件至 contact@ape-tech.com



Function
功能

关键字搜索

语义搜索

相似度排名

信息整合

自动分类

自定义搜索展现

通用爬虫

搜索建议

自动信息补全

控制台

自动问答

敏感词过滤

控制台

Cooling Search 控制台使用户能够有效掌控 Spider 的各种服务。它提供界面向导和图形界面的结合可以查看各种信息。Cooling Search 控制台使每个用户无论角色或任务的差异，都能够对使用模式中的变化做出快速响应，并为最终用户提供高针对性的服务。控制台还提供了日志的报告和统计报表，实时给出各种数据的统计，让用户对系统所进行的系统进行一目了然。包括：出错的日志信息，处理中的日志信息。运行状态等。

选择	名称	索引大小	索引位置	更新日期	选择运行	页面模板	状态	操作
<input type="checkbox"/>	jack	128M	/ape/ImRoBot/index/jack	Wed Dec 30 10:37:23 2009	增量索引 ▾ ▶	预览模板	更新完毕	查看日志 下载日志
<input type="checkbox"/>	zhongwen	128M	/ape/ImRoBot/index/zhongwen	Wed Dec 30 15:46:41 2009	增量索引 ▾ ▶	预览模板	更新完毕	查看日志 下载日志

应用一：选择索引类型后，进行索引运行，状态栏将改变状态。

应用二：可查看索引的日志，选择全部日志，出错日志或处理中的日志。也可下载规定日期的日志。

敏感词过滤

Cooling Filter 能够通过机器学习或预先设定的规则，判定自然语言中是否包含敏感内容。支持敏感词列表、向量机、经典后验概率，及语义分析等几种方式的过滤方法。

分成四个方式：

敏感词列表

用户通过预先设定的一组敏感词，作为过滤条件。这种方式的优势是简单高效，是唯一不需要学习的方法，当语境简单，缺乏上下文环境时，非常有效。但缺点也很明显，机器不能区别分章的感情色彩与任何倾向性。

向量机

分析文章包含的特征，将特征作为向量，并组合在一起成为一个向量空间。在使用前需要预先准备大量包含了敏感内容的文章供机器学习，建立敏感特征向量空间。未知文章到来后，都要建立自己的特征向量空间，并通过同敏感特征向量空间进行比较，得出是否包含敏感内容的概率。这种方法的优势是以一组特征，通过相似度进行比较，不以一个词决定文章是否敏感。缺点是比较的时间随样本空间的大小呈线性增长。

经典后验概率

在使用前需要预先准备大量包含了敏感内容的文章供机器学习，计算不同特征出现在敏感内容中的概率。对未知文章的判定，需要计算多个特征是否敏感的后验概率，当此概率超过一定的阈值后，即认为内容敏感。优点是判定速度非常地快，仅次于敏感词列表。缺点是只能对已知特征进行判定，而无法对未知情况作出预测。

语义分析

通过理解文章的含义，来判定是否包含敏感内容。这种方法需要对大量的样本进行长时间的学习，样本包含的内容越全，机器建立的语义模型越准确。这种方法的优势是，不仅对特征进行分析，更对特征存在的上下文语境进行分析，能够有效地区分出词的感情色彩，得出最准确的结果。缺点也很明显，对样本的准确性要求非常地高，机器学习的时间非常长，对未知文章的预测速度要比敏感词列表的方式低。

欲知更多有关酷灵搜索产品及解决方案，请发送邮件至 contact@ape-tech.com



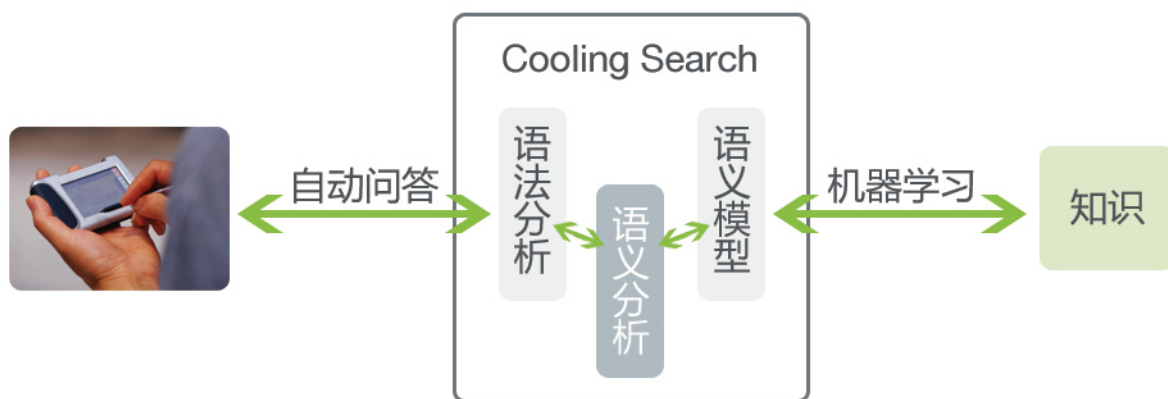
Function 功能

关键字搜索
语义搜索
相似度排名
信息整合
自动分类
自定义搜索展现
通用爬虫
搜索建议
自动信息补全
控制台
自动问答
敏感词过滤

自动问答

Cooling Search 的自动问答从使用上来看，类似于百度知道。两者最本质的区别是，百度知道是由人来回答，而 Cooling Search 自动问答采用自然语言处理技术，一方面完成对用户疑问的分析处理；另一方面通过机器学习建立的模型，得到正确答案的生成，让人们在海量且杂乱的信息中快速、准确地得到想要的信息。

工作流程图如下：



角色描述

知识库

知识库是用于存放人类对已知世界认知的一种特殊的逻辑存储结构。知识库中的知识源于领域专家，它是求解问题所需领域知识的集合，包括基本事实、规则和其它有关信息。从物理数据结构来看，知识库以非结构化数据或结构化数据的形式存在。结构化的信息库通常保存在关系型数据库中，而非结构化信息以各类文本格式（txt，office，pdf，xml等）保存在企业的存贮系统（文件系统，Web，ftp等）中。

语义模型

语义模型表达了原始知识库中各类概念的含义以及这些含义之间的关系，是对知识库中的数据抽象或者更高层次的逻辑表示。该模型通过机器自动学习从 Cooling Search 从知识库中搜索来的数据，用一种机器可以识别方式，保存了知识库中各类抽象的概念。终端用户的所有疑问都将在这里得到解答，这是 Cooling Qa 中最核心的组成部分。

语法分析器

语法分析器的作用是分析用户的提问，并转化成语义模型的输入条件。语法分析器可以对自然语言进行分析，通过分词、词性标注、成分划分、相关词聚类、语法树构建、语义条件生成等多步骤，将自然语言转化成机器可以识别的输入条件，进而由语义模型完成概念的查找。