非结构化数据

计算机信息化系统中的数据分为结构化数据和非结构化数据。非结构化数据是数据结构不规则或不完整，没有预定义的数据模型，不方便用数据库二维逻辑表来表现的数据。包括所有格式的办公文档、文本、图片、XML, HTML、各类报表、图像和音频/视频信息等等。

非结构化数据其格式非常多样，标准也是多样性的，而且在技术上非结构化信息比结构化信息更难标准化和理解。所以存储、检索、发布以及利用需要更加智能化的IT技术，比如海量存储、智能检索、知识挖掘、内容保护、信息的增值开发利用等。

非结构化数据的优势

[编辑](javascript:;)

有大量的数据需要处理

非结构化数据在任何地方都可以得到。这些数据可以在你公司内部的邮件信息、聊天记录以及搜集到的调查结果中得到，也可以是你对个人网站上的评论、对客户关系管理系统中的评论或者是从你使用的个人应用程序中得到的文本字段。而且也可以在公司外部的社会媒体、你监控的论坛以及来自于一些你很感兴趣的话题的评论。[2]

蕴藏着大量的价值

有些企业现在正投资几十亿美金分析结构化数据，却对非结构化数据置之不理，在非结构化数据中蕴藏着有用的信息宝库，利用数据可视化工具分析非结构化数据能够帮助企业快速地了解现状、显示趋势并且识别新出现的问题。

不需要依靠数据科学家团队

分析数据不需要一个专业性很强的数学家或数据科学团队，公司也不需要专门聘请IT精英去做。真正的分析发生在用户决策阶段，即管理一个特殊产品细分市场的部门经理，可能是负责寻找最优活动方案的市场营销者，也可能是负责预测客户群体需求的总经理。终端用户有能力、也有权利和动机去改善商业实践，并且视觉文本分析工具可以帮助他们快速识别最相关的问题，及时采取行动，而这都不需要依靠数据科学家。

终端用户授权

正确的分析需要机器计算和人类解释相结合。机器进行大量的信息处理，而终端客户利用他们的商业头脑，在已发生的事实基础上决策出最好的实施方案。终端客户必须清楚的知道哪一个数据集是有价值的，他们应该如何采集并将他们获取的信息更好地应用到他们的商业领域。此外，一个公司的工作就是使终端用户尽可能地收集到更多相关的数据并尽可能地根据这些数据中的信息作出最好的决策。  
　　很明显，非结构化数据分析可以用来创造新的竞争优势。新的前沿可视化工具使用户容易解释，让他们在点击几下鼠标之后就能清楚地了解情况。从非结构化的数据源中挖掘信息从来就没有像现在这样如此简单。

非结构化数据的采集

[编辑](javascript:;)

在很多知识库系统中，为了查询大量积累下来的文档，需要从PDF、Word、Rtf、Excel和PowerPoint等格式的文档中提取可以描述文档的文字，这些描述性的信息包括文档标题、作者、主要内容等等。这样一个过程就是非结构化数据的采集过程。

非结构化数据的采集是信息进一步处理的基础。现在有许多开源库己经实现了从非结构化文档中采集关键信息的功能，但针对不同格式的文档，所用的开源库不尽相同。

例如，Apache POI是Apache软件基金会的开放源码函式库，POI提供API给Java程序对Microsoft Office格式档案读和写的功能。其结构包括：HSSF提供读写Microsoft Excel XLS格式档案的功能；XSSF提供读写Microsoft Excel OOXML XLSX格式档案的功能；HWPF提供读写MicrosoftWord DOC格式档案的功能;HSLF提供读Microsoft PowerPoint格式档案的功能；HDGF提供读写Microsoft Visio格式档案的功能等。

PDFBox是Java实现的PDF文档协作类库，提供PDF文档的创建、处理以及文档内容提取功能，也包含了一些命令行实用工具。主要特性包括:从PDF提取文本；合并PDF文档；PDF文档加密与解密；与Lucene搜索引擎的集成；填充PDFIXFDF表单数据；从文本文件创建PDF文档；从PDF页面创建图片；打印PDF文档。PDFBox还提供和Lucene的集成，它提供了一套简单的方法把PDFDocuments加入到Lucene的索引中去。

另外还有parse-rtf可以对RTF文件处理，SearchWord可对Word和Excel，PPT文件进行处理等等。[1]

非结构化数据的查询

[编辑](javascript:;)

随着计算机、互联网和数字媒体等的进一步普及，以文本、图形、图像、音频、视频等非结构化数据为主的信息急剧增加，面对如此巨大的信息海洋，特别是非结构化数据信息，如何存储、查询、分析、挖掘和利用这些海量信息资源就显得尤为关键。传统关系数据库主要面向事务处理和数据分析应用领域，擅长解决结构化数据管理问题，在管理非结构化数据方面存在某些先天不足之处，尤其在处理海量非结构化信息时更是面临巨大挑战。为了应对非结构化数据管理的挑战，出现了各种非结构化数据管理系统，例如基于传统关系数据库系统扩展的非结构化数据管理系统，基于NoSQL的非结构化数据管理系统等。[3]

在非结构化数据管理系统中，查询处理模块是其中一个重要的组成部分，针对非结构化数据的特性设计合理的查询处理框架和查询优化策略对于非结构数据的快速、有效访问极为重要。传统的结构化查询处理过程是:首先翻译器翻译查询请求生成查询表达式，然后由优化器优化查询表达式，得到优化过的查询计划，最后由执行器选择最优的查询计划执行，得到查询结果。查询处理的主要操作包括选择操作、连接操作、投影操作、聚合函数、排序等。查询优化的方法包括基于代价估算的优化和基于启发式规则的优化等。

非结构化查询处理过程中除了结构化数据查询处理所包含的操作外，还有两个重要的操作相似性检索和相似性连接。相似性检索是指给定一个元素，在由该种类元素组成的集合中寻找与之相似的元素。例如论文查重系统用到文本的相似性检索，谷歌的以图搜图的功能用到图像的相似性检索，手机上根据哼唱匹配音乐是音频的相似性检索等。相似性连接是数据库连接操作在非结构化数据上的一种扩展，它寻找两个元素种类相同的集合之间满足相似性约束的元素对，在数据清洗、数据查重、抄袭检测等领域有着重要的作用。非结构化查询处理框架要针对这两种非结构化数据特有的查询操作对结构化查询处理框架进行改进。

非结构化查询优化，在代价估算上除了要考虑结构化数据的代价估算模型外，还要设法建立相似性查询和相似性连接的代价估算模型，对于针对非结构化数据的全文索引和空间索引，也应该有不同与B树索引的代价估算模型。代价估算模型除了要考虑CPU时间、IO时间外，由于非结构化数据一般都存储在分布式系统之上，还需要考虑到中间结果网络传输所用的时间，所以中间结果的大小估算对于非结构化数据的查询优化比结构化数据的查询优化更为重要。非结构化数据的查询优化中的启发式规则和结构化数据也有所不同。

非结构化数据的存储

[编辑](javascript:;)

据IDC的一项调查报告中指出：企业中80%的数据都是非结构化数据，这些数据每年都按指数增长60%。据报道指出:平均只有1%-5%的数据是结构化的数据。如今，这种迅猛增长的从不使用的数据在企业里消耗着复杂而昂贵的一级存储的存储容量。如何更好的保留那些在全球范围内具有潜在价值的不同类型的文件，而不是因为处理它们却干扰日常的工作？[云存储](http://baike.baidu.com/item/%E4%BA%91%E5%AD%98%E5%82%A8)是越来越多的IT公司正在使用的存储技术。[4]