# 采用基于 IBM Power Systems 的非结构化文本分析支持关键决策商业案例

Stephen Markham 博士
Michael Kowolenko 博士
北卡罗莱纳州立大学
管理学院

***这是一个问题，不是技术。***
***--Michael Kowolenko***

---

虽然大数据能够极大地改变业务环境，但技术只是决定的推动因素。很多公司利用结构化大数据制定日常运营决策。然而，非程序化的关键战略决策往往涉及非结构化数据。本文介绍使用大数据技术收集和分析非结构化数据的商业价值。这种方法采用融入流程中的重要思考，结合先进的服务器和软件将大数据转化为商业价值。这种流程和结果为需要自适应结构、 文化和专业技术，实现大数据应有能力带来了新机遇。

本文包括三个目的。首先，说明采用结构化和非结构化数据制定决策的差别。然后，举例说明企业如何利用非结构化数据实现商业价值。最后，说明企业如何利用非结构化数据实现类似商业价值，以及为什么选择正确的服务器和软件可以产生不同的结果。

# 1. 采用结构化数据与非结构化数据制定决策的差别

非结构化文本包括现有大约 80%的数据。只使用结构化数据的企业失去大量可用信息产生的收益。非结构化数据包括美国证券交易委员会 (SEC)、国家卫生研究院 (NIH)、国家科学基金会 (NSF) 和能源部 (DOE) 等政府报告，以及学术研究、商业和金融分析报告，咨询调查结果和许多其他来源所含的内容。非结构化文本还存在于无数社会媒体渠道中，如 Facebook、博客、顾客投诉记录和 Twitter，以及新闻报道、大众媒体、专业杂志和许多其他网站。

非结构化数据揭示客户需求、竞争对手的动作、新趋势，以及制定关键业务决策所需的其他信息。结构化大数据方法收集合并行列数字提供给决策者。采用先进的统计技术分析大量数字，可以揭示数据中的重要模式。利用这些技术，决策者可以实时了解发生的情况，或将要发生的情况。分析可以进行加、减、乘、除的数据依赖于结构化方法收集数据，对于定价、分销和库存等业务决策是必不可少的。

另一方面，非结构化办法解析重要信息。例如，非结构化大数据发现一家竞争对手将建新厂的公告，或者一家客户将扩展运营的公告。从而使决策者能够快速做出反应 – 在结构化数据最终显示销售收入下降之前。要想真正做到“大海捞针”，企业必须采用大数据收集大量非结构化文本，利用专用程序通过计算机搜索，在数以千万计的文档中找出特定信息。

## 1.1. 重点思考推动大数据的使用

基于数据制定决策需要人们知道提出的问题，但经验表明，情况并非总是如此。许多组织缺少采用重点思考的流程。在创新管理研究中心 (CIMS) [1] 进行的每个行业赞助项目中，从初创企业到《财富500强》成员，各类公司都面临开展战略调查的困扰。

[1] CIMS 由工业大学合作研究中心 (IUCRC) 设立，成立于1984年，是唯一由美国国家科学基金会 (NSF) 提供资金支持的研究中心，主要调查创新在组织和管理方面产生的成果。

重点思考肯定推动大数据的使用 – 大数据不能推动思考。重点思考是寻找数据源和工具，了解非结构化文本潜在含义的基础。例如，“我们公司需要调查社交媒体进行情感分析”这句话，可以分解成一系列小问题，例如：

- 与我们产品相关的情感
- 与竞争相关的情感
- 客户喜欢和不喜欢产品之类的问题
- 我们公司的新产品如何解决客户不喜欢
- 提高哪方面的竞争力可以解决客户不喜欢

根据重点思考原理，我们帮助开发了利用非结构化文本实现商业价值的流程。实际上，大量公司利用重点思考成功开发了业务决策流程。

### 1.2. 利用重点思考分析非结构化数据的过程

这个过程 (图 1) 需要相关各方跨部门团队从一开始就共同参与，并已在许多行业证明是成功的。这个过程采用重点思考的方法：1) 定义提问内容，组成需要调查的具体问题; 2) 确定信息源; 3) 确定搜索条件，并定义条件之间的关系 (称为规则); 4) 根据条件和规则采用大数据技术，存储并分析由外部和内部信息源收集的大量数据; 5) 评估数据的充分性、适用性和准确性; 最后; 6) 数据进行过滤之后，评估证据支持还是反对决策所需的假设或条件。在此必须强调，这是一个反复的过程。收集过滤数据之后往往产生新的洞察结果，需要进一步调查。这要求团队面对新证据克服个人和集体的偏见。

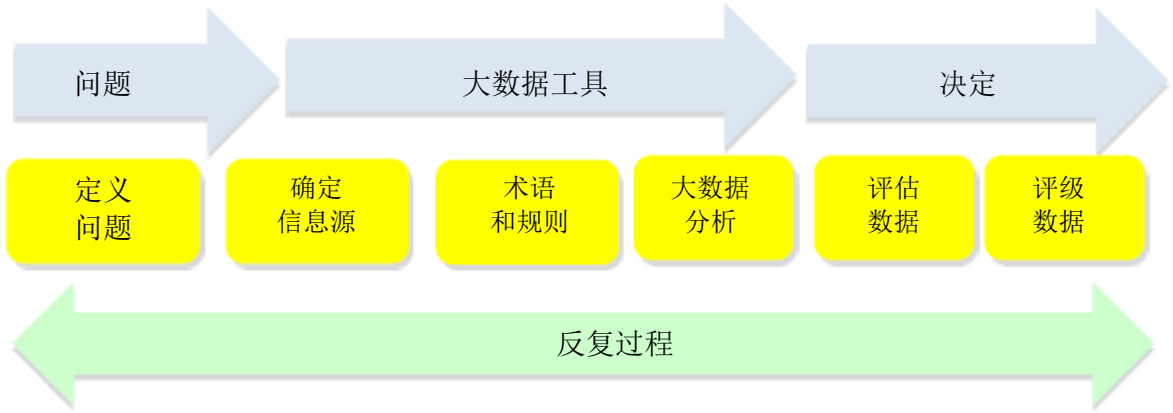


图1. 基于非结构化数据制定决策的过程

# 2. 企业利用非结构化数据实现商业价值举例

下表列出的只是大量利用这个过程成功解决业务问题的五个例子。每个示例代表不同类型的常见问题。另外，从这些例子可以看出，所做决定不是根据结构化数据的量化分析，而是以通过分析非结构化数据得出的事实为依据。所选问题、使用的术语、创建的词典和规则便于企业配置软件，解答各种关键决策问题。

| 行业:<br>公司 | 问题 | 信息来源 | 结果 |
|---|---|---|---|
| **人力资源服务:<br>Kelly Services** | 开发医疗保健业人力资源服务新产品 | SEC, URLs, 商贸期刊, 专业杂志, 保险提供商 | 决定进军未曾想过的医疗保健领域 |
| **工业气体:<br>Air Products** | 发现新客户和市场机遇 | SEC, 新闻报道, 行业出版物, 营建许可 | 确定了计划建造新设施的新客户 |
| **大学:<br>北卡州立大学** | 识别新技术的商业合作伙伴 | SEC, URLs, 行业出版物 | 确定了协作的潜在合作伙伴 |
| **临床研究组织:<br>PRA International** | 为新的临床实验提供商业情报 | Clintrials, PubMed | 确定具有临床实验专业技术的新医师/医院 |
| **非政府组织:<br>克林顿基金会** | 发现疾病诊断新技术与市场之间的契机 | Clintrials, PubMed 风投机构 | 确定了从事尖端诊断研究的研究实验室 |

表2. 行业举例

### 2.1. Kelly Services: 开发医疗保健业新的人力资源服务产品

Kelly Services 是全球人力资源服务领导者，经过产品扩充已成长为人力资源解决方案全面服务的公司。他们希望了解为全球医疗保健行业提供人力资源服务是否可行。通过与内部和外部主题专家合作，这种新服务需求以及竞争和监管规定的相关非结构化数据，不仅表明存在大量机会，而且提供了如何推进这项工作的指导。

研究表明，各类客户特别需要护士远程提供医疗评估和建议，包括医院、护理机构、保险公司和自我保险组织。研究还发现医护人员短缺，且当地或地区不能满足这种需求。Kelly 国内人力资源与地区需求相结合，Kelly 可以独家满足竞争对手尚未发现的市场需求。

同时，研究结果证明，医疗保健提供商为各国提供大量不同类型的远程医疗服务可以获得回报。利用非结构化文本分析明确掌握每个国家的管理规定，Kelly 知道如何根据监管规定，在正确的国家开展正确的服务。这项新服务现在已成为 Kelly服务产品中两大创新举措之一。

### 2.2. Air Products and Chemicals: 发现新客户和市场机遇

Air Products 为大量商业市场提供工业气体和特种化学品。他们希望能够识别潜在的新客户，并为他们的销售队伍提供与这些客户有关的信息。金属加工行业中，及早确定客户，甚至在生产开始前确定客户是十分重要的。金属加工厂需要专用气源，因此尽快与气体供应商沟通，以便满足新客户的需求也是至关重要的。

数据采集过程从“读取”所有 SEC 数据开始，采用自然语言处理规则确定计划投资新厂和新设备的公司。同时，登录各种公开出版物的网站，查找发布新开工消息的公司。此外，搜索包括读取国外的外文报纸 (22 种语言)，金属加工设备厂商宣布在这些国家新销售的设备。最后，引擎读取美国所有营建许可，确定建设新厂的金属加工公司。

研究发现一家金属加工公司将在 Air Products 运营地区建立新厂，这家公司订购了特种设备，并宣布一定数量新的工作岗位。Air Products 不仅确定了新的潜在客户，而且知道他们将要使用的设备 (从而确定了所需的气体)，以及公司计划加工的金属数量 (换算出所需气体的体积) -- 所有这些早在开工建设十八个月之前就掌握了。因此，Air Products 能够在竞争中提前争取这家新客户。此外，分析还显示大量准备拓展运营的现有客户，这种信息是以前不知道的。

### 2.3 Pentair: 发现新产品的客户

Pentair 提供应用化学处理和曝气设备，用于湖泊、 池塘和集水区等受损水体。公司最近开发了一种新产品，能自动将化学品均匀地分撒到湖泊或池塘。州和地方政府及行政机构确定水体是否受损。负责补救的单位各不相同，从郡县和直辖市到业主协会和高尔夫

球场等。这些单位存在于所有地理区域，但集中在某几个州。

业务挑战：发现新型水处理系统的潜在销售机遇。需要确定的问题包括：

- 在哪些州重点开展销售工作?
- 地方机构中参与补救的决策人?
- 什么样的水体视为水质不宜?
- 水体责任人?

利用几个关键单词列表和自然语言处理 (NLP) 规则，开发了从数据集中提取感兴趣信息的模型。州名字典与受损水身体术语组合使用确定重点地区。通过搜索这些州的监管机构，发现机构名称和联系信息。利用受损水体的机构列表，发现具体目标单位的名称。此外，这个模型建立的一项规则包括“公司查找”。生成的NLP 规则还可以根据命名规范模式提取责任单位名称。利用列表和规则，开发的模型可用于查询检索的数据集。

这个项目的完成大约用了三个月的时间，其中包括爬网检索数以千计的文件，设置过滤模型选择解答特定问题的相关信息并分析结果。项目组与客户主题专家紧密合作选择信息源，审查结果，通过多次反复改进信息过滤模型。受损水体的名称、它们的负责单位、以及联系人姓名由客户查找并跟踪。

### 2.4 PRA: 为新的临床实验提供商业情报

PRA Internal 是一家为医药公司进行药品实验的临床研究组织。该公司希望更好地了解其他临床研究组织的工作，更准确地预测客户未来需求，从而更好地为他们提供服务。

采用非结构化文本分析读取含有7500多万网页的400万个数据文件。然后，将这些非结构化信息源与 www.clinicaltrails.gov 所含的结构化信息，以及公司特定数据文件结合，提供更加深入的行业内部情况。

在一个项目中，PRA 想知道骨髓瘤的研究进展。公司领导想了解未普遍公开的各种情况，例如，过去骨髓瘤实验失败的原因，哪些公司可能在从事骨髓瘤研究活动，而这些活动可能未在政府和行业报告中公布。这项研究发现了几个有关的调查结果。大多数实验失败是由于缺少后备力量，而且还发现一些其他重要原因。有趣的是，调查发现七家公司各种适应症采用相同的基因靶向，如哮喘、乳腺癌和肺癌、帕金森病、阿尔茨海默

氏症、镰状细胞贫血、失明等。此外，还发现这些项目经理的姓名，便于 PRA 建立关系，更好地推进自己的研究。

### 2.5. 克林顿基金会 (CHAI): 发现疾病诊断新技术与市场之间的契机

CHAI 希望评估诊断方面的医疗保健投资是否可以取得成效。经理们支持大量计划，并希望实现成果最大化。

大数据评估不仅用来评估新诊断方法的有效性，而且包括新诊断工具可能产生的影响。通过评估成千上万文章和报告中的非结构化数据，发现开发新的诊断方法对新疾病的影响小于常见病症现有诊断的影响。这一发现促使重新评估 CHAI 策略并调整资源，以提高更多人的诊断效果。

### 小结

这些例子证明，采用非结构化文本分析发现、评估新商机，发现合乎条件的新客户，提供显著改进的、更加详细的商业情报，以及制定战略资源分配决策可以实现商业价值。非结构化数据可用于许多其他应用，这种应用条件下，决策需要了解不能进行加、减、乘、除的特定信息。

## 3. 公司如何利用非结构化数据实现商业价值

使用非结构化大数据的优点是经过优化的工具和技术，现在可供人们利用重点思考能力解答重要业务问题，而不需要软件程序员或统计人员。这个过程初看起来很复杂，软件相当专业，但它已不再是数据科学家的专属范畴。

非结构化大数据项目需要整个 IT 部门全面协作，支持软件并帮助收集和存储数据，结合统计分析运行大数据流程。但到目前为止，成功的非结构化数据分析最重要的是公司主题专家 (SME) 的重点思考能力。这是一种跨部门的决策能力 -- 不是一次性活动。因此，公司必须保证执行这一过程和制定决策所需的 SME 资源，并授予高级管理层采取行动的权利。

这种业务人员大数据应用民主化不仅有利，甚至是这些工具的目的，更确切地说是一种必要。业务内容必须转化为实现业务价值的具体问题、条件、信息源和规则。

选择正确的软件进行非结构化文本分析是至关重要的。可供使用的商业和开源程序很多。如果不选择含有图形用户界面，可与大型数据集无缝交互的程序，哪怕只是为了运行技术部分的软件，也仍然需要数据科学家。

IBM Content Analytics Studio (ICA) 软件是唯一符合所有标准的产品，可供业务内容人员自行完成非结构化数据的大数据分析。经过几天培训，大部分业务人员完全可以学会 ICA 应用到自己工作中的使用方法。因此，各类人员可在日常工作中，而不是特定的项目上使用大数据。

使用正确的服务器平台支持大数据也是十分重要的。尽管我们设法解析出高度特定的决策信息，而不是合并大型数据集，但如果我们可以收集并及时处理的话，我们只能确定关键信息。因此，选择正确的服务器平台是部署成功的非结构化大数据项目的一个重要方面。

一些公司只是刚决定使用 x86，因为他们已经安装了服务器和基于 x86 运行的大数据软件，包括 ICA。不过，这种方法有很大局限性。x86 服务器可能满足小规模论证，但这种低可靠性平台难以采用大数据。在 x86 服务器上，您可能无法实际论证大数据的充分价值。我们使用 x86 和 IBM Power Systems 服务器处理大数据。毫无疑问，Power 服务器具有不可比拟的优越性。x86 服务器常常崩溃，所以我们的客户端不选择使用这种服务器。

Power System 服务器采用基于 POWER8 处理器的技术，可在每内核含有多线程的多个内核之间，以并行方式更加快速地进行数量更多的大数据并发查询。同时，增加了内存带宽和IO 速度，可以更加快速地提取、移动和访问数据，从而提高公司执行海量数据分析的查询速度。

### 3.1. 利用大数据建立竞争优势

恰当部署流程、工具、功能和结构，非结构化数据分析可以成为创造更大商业价值的动力。如果这些要素结合不当，等待您的将是令人沮丧和失望的结果。失败的代价是落后于成功利用大数据的竞争对手。

成功实施的关键是建立业务内容专家和决策人员日常使用的简单流程。这意味着，必须确定决策过程输入和输出的正确信息来源、角色和职责。必须确立决策人员的上下级关系，以保证落实责任制。必须制定公司预期的大数据绩效水平和成果。必须采用保证符合性能和可靠性水平的基础架构。只有决策人员掌握更加具体的信息，才能实现预期商业价值和竞争优势。