

面向企业创新服务的大数据 智能分析机制探讨*

□ 李颖 朱礼军 农国武 / 中国科学技术信息研究所 北京 100038

摘要: 大数据智能分析是企业技术创新的新希望,也是科技信息机构在服务转型的压力下所面临的新机遇。应用大数据智能分析,企业可以从海量多样的数据中,深入挖掘并呈现满足创新需求的各种关联信息,获得洞悉,并做出更为准确的创新决策。文章面向企业技术创新服务,讨论大数据智能分析系统的机制,并以大数据类型——结构化和非结构化数据为中心,介绍它们不同的处理技术。由于计算机处理结构化数据的简易性和精准性,文章将数据的结构化处理技术作为主要的议题进行讨论,介绍实用的标准技术DITA。最后,以作者研发项目为例,介绍与大数据分析有关的工作及展望。

关键词: 大数据, 智能分析机制, 技术创新, 数据结构化, DITA

DOI: 10.3772/j.issn.1673—2286.2012.11.006

1 大数据智能分析的意义

当中国的企业,尤其是中小型企业着手于新技术创新之际,当我们的科技信息数字图书馆服务正在全力策划支持企业技术创新的时候,不管我们高兴与否,“大数据大象”已经当道(参见图1)。盲人摸象似乎是人类的魔咒,如果我们对大数据了解片面,又会重演盲人摸象的故事。而如果我们能鸟瞰大数据,并能借助当今计算机的运算能力,就可以做到充分地收集数据、有效地处理和分析数据、正确地挖掘数据间的关系、语义化地呈现数据关系,以此让我们企业用户获得洞悉(insight),进行决策,促进技术创新。

对于企业,数据意味着如下的数学等式:

企业拥有的数据信息=企业业务活动的DNA

参见图2,也就是说,当今的企业,如果离开数据信息,就不可能成长壮大。有了数据,等于有了生命。但是,企业只有高效地将数据应用于其业务,才能赋予她们不一样的生命力。从海量的、不同类型、不同结构的数据中挖掘信息价值,是现代企业获得洞悉、保证竞争优势



图1 “大数据大象”



图2 企业数据信息的战略价值

* 基金项目: 十二五科技支撑计划项目技术创新服务平台关键技术研究与应用示范(编号: 2011BAH30B00)。

势和效益、脱颖而出的唯一出路。为此,大数据环境下的企业数据智能分析是企业战略之本。

什么是大数据的智能(smart,或称智慧)分析?可以有各种诠释,对本文而言,可以简单地理解为:通过应用“企业内数据+大数据(Web数据)”,开创出全新的业务。以研发型企业为例,如果将其拥有的研究成果和Web上的各种信息相整合,发现业务上的合作伙伴,并将研发出的各种技术与新的业务结合,能实现可持续发展,参见图3。

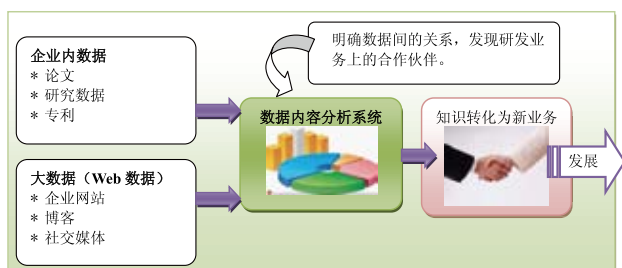


图3 企业数据分析的概念示意

数据分析对企业如此重要,然而,从面向全球的调研中明显地看出,企业对剧增的数据,在收集、选择、仓储、分析等处理的各个环节上,都面临着挑战,参见表1的调查结果。如何及时地收集和整合非结构化、结构化的各种数据,准确地分析和理解这些数据之间的关联,最终获得洞悉,是企业大数据智能分析的目标,为此,本研究进行了初步的探索。

表1 有关大数据分析的调查结果

1 in 3	三人业务主管中的一人认为,在重要决策时,无法信任数据信息内容,或认为信息还不足以决策。
60%	60%的业务主管认为信息需要及时地搜集、理解。
56%	56%的管理者感到,必须处理的数据量过多。
83%	83%的CIO在考虑“BI+分析”才能为竞争中脱颖而出提供洞悉。

2 企业数据分析的变革——Watson引发的大数据智能分析时代的到来

2011年2月16日, Watson (沃森) 成为新一代人机大战的冠军,如图4和图5所示,以此开创了企业可以应战大数据挑战的新时代。作为IBM“智慧运算”理念的重要支柱, Watson以其海量数据的实时分析、负载优化系统方面的能力征服了全世界。Watson不仅拥有“渊博”

的知识面,还能理解分析复杂的人类语言,甚至俚语,并以高置信度(confidence)快速地回答问题。Watson用其分析的“法宝”助力企业应对“大数据”的挑战,抓住大数据的机遇,给企业带来了新的希望。Watson的法宝是:

(1) 超强的自然语言理解能力。这对企业的分析应用很重要。Watson证明了通过自然语言与计算机的交互不再是科幻,其超强的自然语言理解技术完全可以根据自然语言上下文的内容和语境来准确理解词义。

(2) 非结构化数据的处理。Watson根据一些文本形式知识库来回答问题,这样的文本库是非结构化的信息,而不是基于结构化的知识库,这样的知识库可以使用知识表示语言(如CYC)来表示。

(3) 快速、高效地处理非结构化数据。尽管Watson在Jeopardy(危险边缘)节目中所分析的数据集相对于“大数据”的标准来说微不足道,但是,能够快速、高效地对非结构化数据进行分析是许多大数据分析情况中非常具有代表性的需求。

(4) 基于文本的预测和分析算法。这为大数据分析提供了另外一种很重要的方法。Watson目前使用了大量的预测模型来分析大数据,并且实时地得出结论。这些基于文本的算法的应用十分重要,因为大部分的客户都面临着这样的文本数据。

Watson的成功,进一步促进了人机合作,这些前沿科技还能应用到日常业务中,如医疗、金融、电信、制造业等。Watson是大数据时代的“先驱英雄”。它让人们相

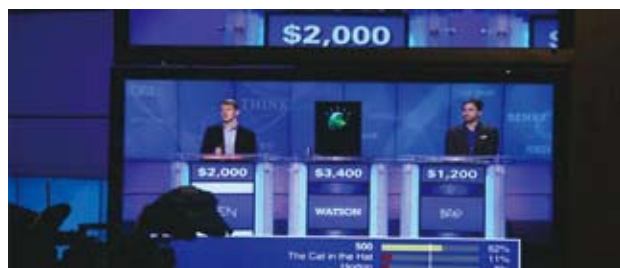


图4 新一代人机大战冠军Watson



图5 IBM的Watson开创了企业深度智能分析的新时代

出处: http://www-03.ibm.com/systems/cn/diHub/products-innovation/watson/index3.shtml?crs=apch_dst2_20110627_1309168038647&cm=edm&cr=uedm&ct=101helge&ck=nur&cmp=101he

信：只要企业能先人一步实时整合数据，从数据分析中获得行动的洞悉，了解业务动向和预测市场的趋势，优化业务，企业就将成功地获得转型，成就未来。

3 企业大数据智能分析机制——以IBM为例

3.1 IBM大数据智能分析机制^[1]

与传统的分析方法对比（参见图6），大数据智能分析的优势为更快、更全面、成本更低。它基于相互补充的方法，与各种数据应用场景对应。

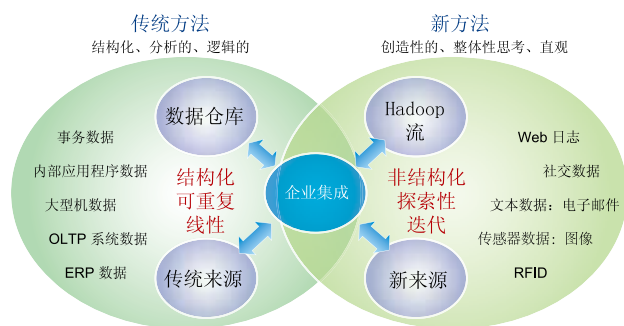


图6 传统与新分析方法的对比

出处：2012 IBM智慧的分析洞察论坛

大数据智慧分析结合了多种技术：

☆ 预处理

摄取并分析非结构化数据类型，并将其转换为结构化数据。

☆ 综合结构化和非结构化分析

利用社交媒体等其他外部源补充数据仓库。

☆ 结合高速和历史分析

对持续变化的数据进行分析，基于深入的履历分析调整模型。

☆ 重用结构化数据，实现探索性分析

使用结构化数据进行实验和即席分析。

以IBM大数据平台为例，其大数据战略使分析与数据更近。由于至今为止所没有的分析应用程序的出现，需要如下的大数据平台，参见图7。

☆ 集成并管理大容量且不同种类的数据

☆ 将高级分析应用于信息并且不改变信息的原本格式

☆ 将所有可用信息可视化，供即席（ad hoc）分析使用

☆ 为新型分析应用程序建立开发环境

☆ 优化工作负载并安排进度（scheduling）

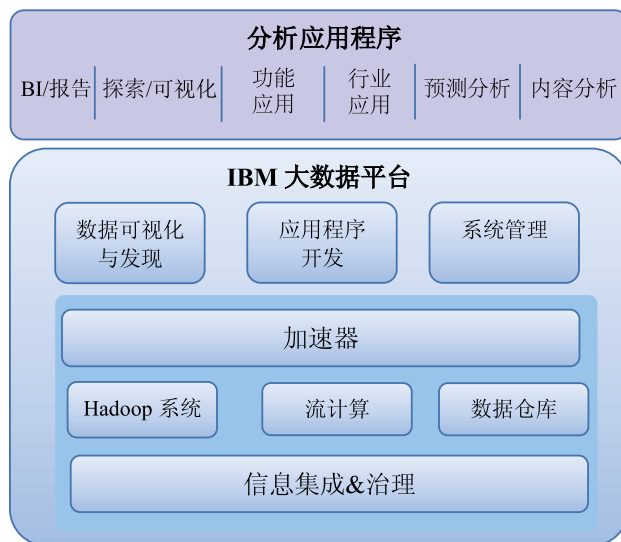


图7 对应新型分析的大数据平台

☆ 安全和治理（governance）

大数据平台要点如下：

（1）Hadoop系统

☆ 可管理多种、大量的数据

☆ 利用企业功能可增强开源的Hadoop系统

☆ 工作负载优化——大数据分析工作负载的性能优化（参见图8）

* 自适应的MapReduce

◆ 优化多个小任务执行时间的算法

◆ 性能提升30%，减少了任务启动的开销

* Hadoop System Scheduler

◆ 利用以往的经验识别小型和大型的作业

◆ 顺序工作，减少开销

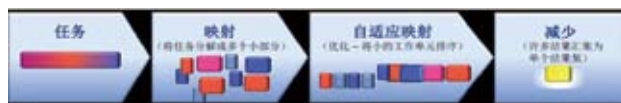


图8 工作负荷优化方法

（2）流计算

☆ 专为分析变化数据而构建

* 多个并发的输入流

* 大规模可扩展性

☆ 处理和分析多种数据

- * 结构化内容、非结构化内容、视频、音频
- * 高级分析运营程序

(3) 数据仓库

- ☆ 工作负载优化的系统
 - * 深度分析设备
 - * 可配置的运营分析设备
 - * 数据仓库软件
- ☆ 功能
 - * 大规模并行处理引擎
 - * 高性能OLAP
 - * 混合的运营和分析工作负载

(4) 加速器

- ☆ 分析加速器
 - * 分析、运营程序、规则集
- ☆ 行业和横向的应用程序加速器
 - * 分析
 - * 模型
 - * 可视化/用户界面
 - * 适配器

(5) 信息集成和管理

- ☆ 将任何类型的数据集成到大数据平台
 - * 结构化
 - * 非结构化
 - * 流
 - * 管理大数据
 - * 保护敏感数据的安全
 - * 生命周期管理, 以控制数据增长
 - * 主数据, 建立事实的单一版本(保证一致性和

质量)

(6) 用户界面

- ☆ 业务用户
 - * 大量和多种数据的可视化
- ☆ 开发人员
 - * 工具和语言的相似性
 - * 企业功能成熟的开源工具
 - * 集成不同的环境
- ☆ 管理员
 - * 协助进行系统管理的控制台

(7) 大数据平台的方向

- ☆ 成熟的企业功能
 - * 可扩展性和管理性
 - * 强大的文件系统和信息生命周期管理

- * 部署选项: 软件、设备、云
- * 深度集成企业系统和应用程序

☆ 生态系统支持

- * 所有类型的用户都易于使用: 开发人员、业务用户、合作伙伴和数据科学家
- * 增强的开发环境
- * 自助服务应用程序开发和可视化工具
- ☆ 加速器驱动更快实现价值
 - * 面向不同用途的广泛分析技术
 - * 行业特定的模型和用例
- ☆ 更加关注“4V”中的第四个V: 准确性 (Veracity)
 - * 管理数据、流程和模型的不确定性

3.2 企业大数据智能分析的商务变革^[2,3]

4年前, IBM提出了智慧的地球(smarter planet), 通过数千个项目, 学到了通向成功分析不可欠缺的东西。之后, 分析在不断地进化, 从开创新业务到必要的业务事项, 从企业数据到大数据, 从一个组织的发展到行业的整体变革。企业大数据智能分析带来了商务变革:

☆ 分析的进化: 从开创新业务到必要的业务事项, 显示出实施高端分析的企业其业绩明显优于竞争企业。

☆ 分析在扩展: 从企业数据到大数据。

☆ 分析在变化: 从一个组织的发展到行业的整体变革。

☆ 分析在进化: 从可能性向现实变化。

☆ 促进机构变革的出发点为如下4个重要动机之一:

A. 客户的获得与维护、提高客户的满意度

B. 改善业务

C. 财务过程的变革

D. 风险、不当以及合规性管理(compliance management)

☆ 将信息变为insight(洞察或洞见、远见)、再将insight变成业务成果的整体方法。

☆ 将信息作为基础设施(infrastructure), 维护机构的整合性。基于业务战略的信息和大数据战略的展开。

☆ OODA – Observe、Orient、Decide、Act(观察、适应、决定、行动)。

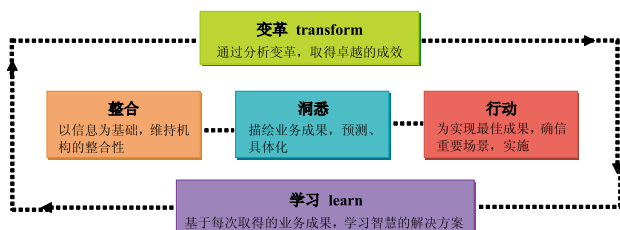


图9 智慧分析将信息转化为企业成果

由此看出，企业大数据的智慧分析，通过将信息转化为企业成果（参见图9），带来了商务变革。

4 企业大数据智能分析中的结构化与非结构化数据的处理技术

4.1 大数据时代企业数据的构成

以下这组全球总数据的指数增长预示着大数据时代的到来：

2005年：数据量突破了1Peta Byte

2009年：数据量突破了1Exa Byte

2010年：数据量突破了1Zetta Byte

2020年：预计将突破35 Zetta Byte

在这种大数据时代下，企业面临着如下的课题：

- 1) 庞大的信息
- 2) 发现不了的必要信息
- 3) 缺乏可信性的内容
- 4) 处理数据的预算有限
- 5) 商务活动需要最优化

而智能数据分析架构建立，解决了上述难题。



图10 约80%的企业数据为非结构化数据

为进行数据的智能分析，需要了解企业数据的构成情况，数据有结构化与非结构化数据之分，参见图10。

表2 结构化和非结构化数据的处理方式

结构化数据	非结构化数据
☆ 通过元数据定义含义 例如，RDB等表形式数据	☆ 通过上下文定义含义 例如，文本数据
◆ 通过字段名表现含义	◆ 通过词的前后关系表现含义
机构名称 ISTIC 职工数量 1000人 公益部门 六个	ISTIC约有职工1000人，有六个公益部门向社会提供信息科技服务。

结构化和非结构化数据特性不同，其处理也采取不同的方式，如表2所示。

结构化数据，其值的含义明确，计算机处理比较容易。而同时，处理非结构化数据时，多数需要进行解析（语句分析）等前处理。

那么，企业80%的数据为非结构化数据，如何“变垃圾为金山”？一种方式是将其进行结构化，再进行处理；另一种方式是，利用现有的NLP技术，进行处理。4.2和4.3分别介绍这两种理念不同但又互相补充的技术。论及前者，我们将展开介绍DITA体系，因为我们认为国内企业对其结构化的认识和实践还远远不够。

4.2 数据结构化处理技术DITA

上一节提到，数据的结构化，可以使计算机的处理变得容易，使用户获得精准信息，是企业运营的关键任务之一。

除了RDB数据之外，国内企业比较缺乏对其他方式的信息结构化技术的重视与积累。本节介绍一种重要的结构化处理技术——DITA。通过DITA，可构建结构化信息并发布，企业有必要理解DITA理念和应用方法。

DITA的结构化优势如下：

☆ 易于同时更新、编排，易于多语种的数个版本同时处理

- * 容易同时操作；
 - * 不锁定：多数人进行操作，处理单元文档规模小时，容易同时并行作业；
 - * 容易再利用：进行结构化处理，容易再利用。
- ☆ 容易输出多种格式
- * 多种媒体、格式的输出：PDF、Web、印刷媒体、其他；
 - * 与多种版本对应：对应产品的不同版本、操作系统及多种语言环境。

-
- 文档(手册等)
- 热创作
- 转换
- 文档<-> 细分组件和关联信息
- 组件和关联信息
- 主视图
- 主图
- 主图
- 主图
- 主图
- 保存
- 综合文档管理系统
- 检索
- 用各种轴工具一次性转换
- 既有文档
- 修订重用
- 编译
- 文档部分为组件:
- 转化为 DITA
 - 转化为网页的 XML 格式等

工具开发
结构化手册创作

主题单位方向变化

内容管理和过程管理整合

1990年代
 纯文本—扁平化 (SOML)
 * 大幅降低成本
 * 处理时间显著增长
 * 检索性提高

2003年~
 SOML→CETA XML
 * 从单据单位转向主题单位
 * 重用性显著提高

2005年~
 信息中心
 * 成为顾客随时随地都能获取最新型信息的大门

2010年~
 9604m的应用
 * 检索时间缩短
 * 重用性提高
 * 工作流的合理化

- * 一个信息源以多种格式
- * 发布 (PDF、HTML、e-pub等)

- ### 4.3 非结构化数据的处理技术

Watson针对提问文中的内容进行分析,从大量的文档中计算答案的候选和依据以及置信度,在获得高置信

合规性 (compliance)

低媒体电子化 电子表格的管理 内容的制作和管理 业务流程的管理 与外部系统间的交互

PowerPoint, Excel, Word, 智能信息的制作和发布 DITA

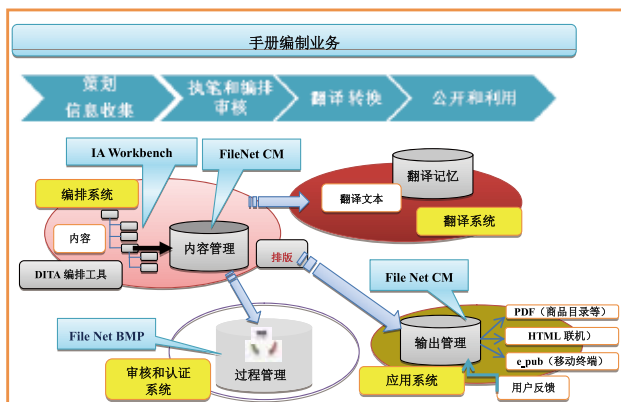
```
graph TD; A[提问] -- "文本输入" --> B[问句和主题的分析]; B --> C[答案候选的抽出]; C <--> |检索| D[(信息源)]; C --> E[答案打分]; E <--> D; E --> F[答案排序]; F <--> G[(统计模型)]; F -- "回答时语音输出" --> H[解答+置信度];
```

The diagram illustrates the QA process flow:

- 提问** (Question) leads to **问句和主题的分析** (Question and Topic Analysis) via **文本输入** (Text Input).
- 问句和主题的分析** leads to **答案候选的抽出** (Answer Candidate Extraction).
- 答案候选的抽出** involves **检索** (Retrieval) from the **信息源** (Information Source).
- 答案候选的抽出** leads to **答案打分** (Answer Scoring).
- 答案打分** involves interaction with the **信息源** (Information Source).
- 答案打分** leads to **答案排序** (Answer Ranking).
- 答案排序** involves interaction with the **统计模型** (Statistical Model).
- 答案排序** leads to **解答+置信度** (Answer + Confidence) via **回答时语音输出** (Voice Output during Answering).

企业内容数据管理,即ECM(Enterprise Content Management)实现的主要功能如图12所示。

* 内容与格式分离



度之际,作出回答。它是可高速进行一系列的智能化处理的问答系统。原理如图15所示。

支撑Watson的技术被众多应用领域看好，主要包括：

- (1) 迅速准确地进行医疗诊断支持的系统
- (2) 监测药物间潜在相互作用的系统
- (3) 为律师及法官服务的已有判例的参考系统
- (4) 金融领域假想场景和法规遵从的系统

其中，基于循证（evidence）医疗诊断支援系统已经在国外实施，即基于症状、家族病史、双亲病史、药物治疗、检查发现、所见和假设，以及大规模医疗相关信息（文档、论文、报道、数据库等）广范围的证据，利用诊

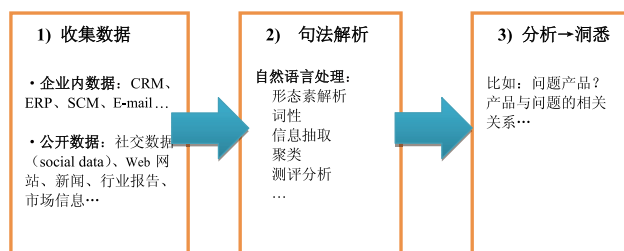


图16 非结构化数据分析处理流程

断模型，计算疾病的置信度，来支持诊断，实现提高医疗质量和减低医疗成本。

在非结构化数据分析处理中，文本挖掘是从大量文本中抽取信息，获得洞察的关键技术。虽然人工手段可以获得较高的精度，但是费时费力，不具有规模性。文本挖掘可以消除低效及对人的依赖，适合于大量数据分析。主要应用如下：

- (1) 抽取主题、原因和结果等，具有统计倾向的分析
- (2) 基于文本内容相似性进行文献检索（概念检索）
- (3) 基于文本内容的相似性对文献进行分组（聚类）
- (4) 自动摘要、自动抽取其所含的人名、地名等实体
- (5) 自动生成分析报告、生成与行为（action）相关的知识
- (6) 预定领域文献的分类（自动分类）

5 中信所在企业创新服务项目上的相关研发

5.1 智能化企业培训系统

自2008年，中国科学技术信息研究所开展面向企业技术创新的知识服务，很多实施项目都依赖基于海量的数据分析，为企业的转型提供解决方案。其中，在数据仓储、分析工具及可视化工具方面进行了较为深入的研究。如图17所示，在某一金融企业智能化培训系统中，结合了结构化和非结构化数据处理技术，即文本挖掘技术和主题化（文本结构化处理后，再应用）技术，来构建智能化的系统。本文重点介绍基于主题的信息结构化处理的相关工作。见下节内容。



图17 银行智能化培训系统架构

5.2 主题的半自动生成体系

将文献基于主题进行结构化处理，会提高企业信息应用的生产力，给企业带来巨大的收益。基于主题的内容被认为是一种智慧的信息。比如，DITA技术可以构建信息的语义元数据、互相关联的主题图，实现信息利用层面的语义导航。而这种结构化信息体系的建立，最为挑战的部分为主题概念的确立。为此，我们也有有机地融合了NLP技术，进行了初步的尝试。

5.3 支持铝行业技术创新的智能化数据分析架构

中国铝行业正面临着严冬，创新转型成为行业可持续发展的必由之路。在此背景下，中信所承担了科技部技术创新服务关键技术研究与应用示范工作。其中，创新机会的发现就是典型的大数据分析的任务。通过与韩国科技信息研究院（KISTI）的合作及智能化分析系统的技术转移工作，将加快我们的研究产出。我们将采用KISTI如下的架构：

此架构包含四个层：基础设施层、仓储层、核心逻辑和应用层。基础设施层是全部知识处理骨干，比如软硬件的支持。全部目标数据和产生的知识在仓储层系统化地管理。核心逻辑层包含众多的功能，主要与学术知识抽取相关。最后，应用层具有多种应用，基于底层模块和功能开发。

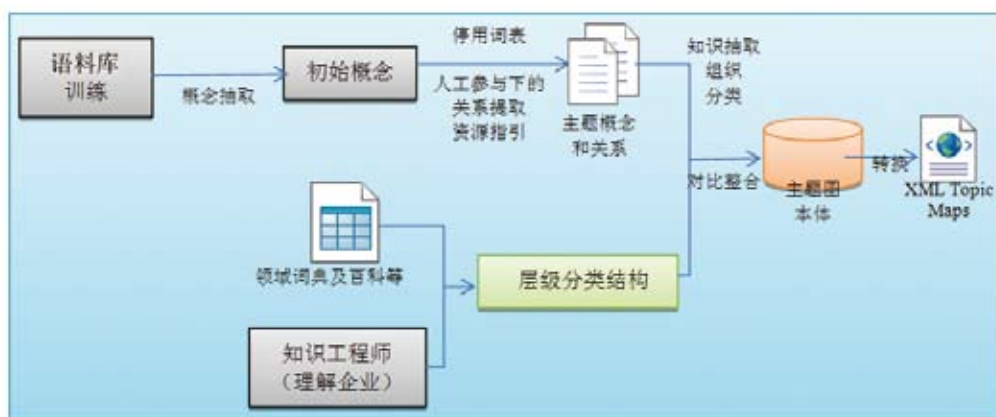
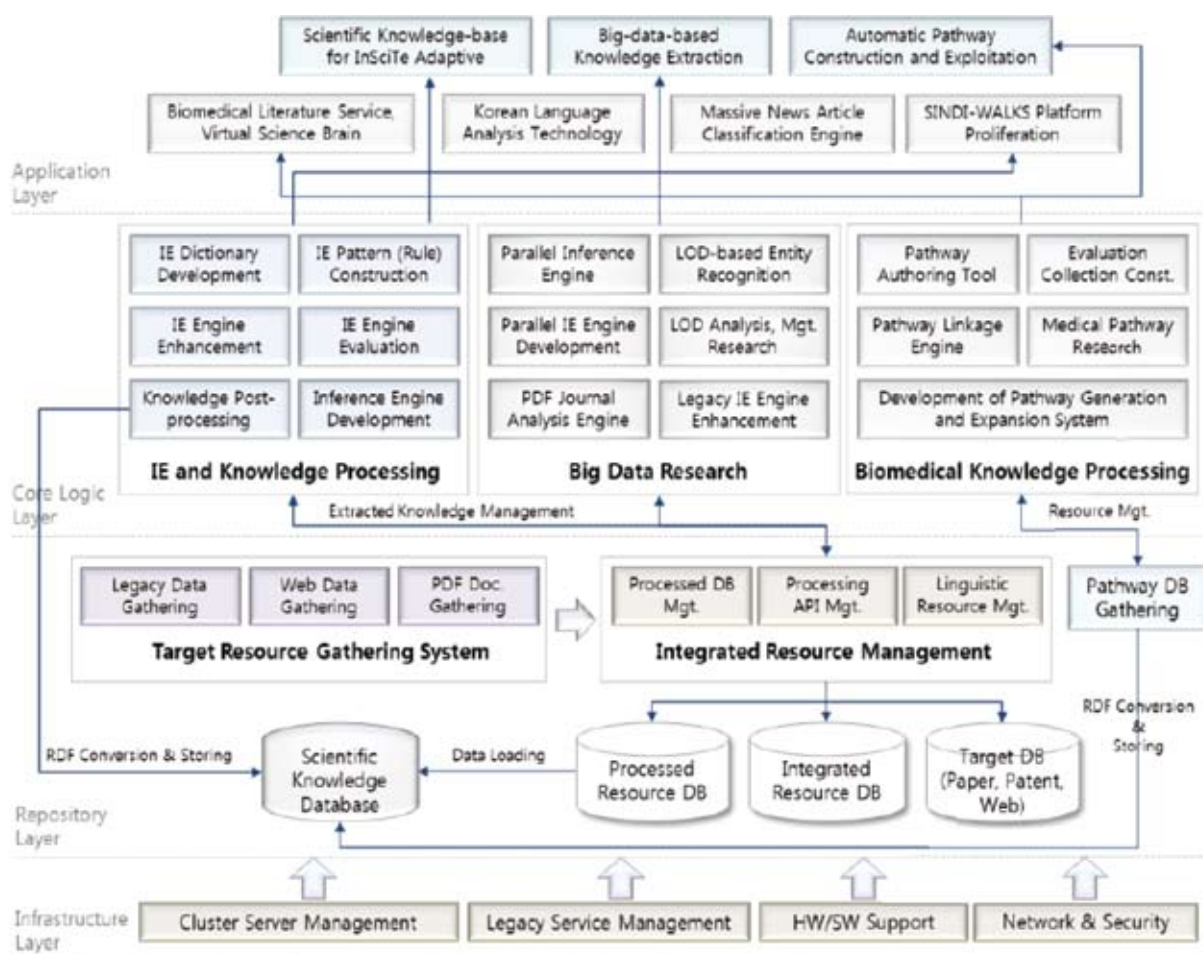
图18 主题topic半自动生成流程^[4]

图19 KISTI科学知识的抽取与揭示架构

出处: CHOI S-P, SONG S-K, JUNG H. Architecture of Scientific Knowledge Extraction and Exploitation [C]// Activity-Centric Architecture of Scientific Knowledge Extraction, Integration and Exploitation for R&D Trends Analyzing. AST2012, Beijing, 2012-09-26

6 结论与展望: 基于跨行业大数据应用的企业技术创新通用基础设施愿景

中国科学技术信息研究所作为国家科技信息收集、长期保存、应用服务的公益部门, 在支持企业技术创新方面, 承载着重大的使命。在大数据环境下, 我们只有

通过跨领域合作、国际合作、进行数据分析高端人才的培养, 才能将分散在不同领域、不同类型的数据进行整合, 基于信息的结构化以及自然语言处理技术, 开创数据的科学, 为企业创新提供有效的数据分析服务, 使企业获得洞悉, 开拓业务和拓宽市场, 强化竞争力。这也是科技信息服务领域的新机会和面临的新挑战。



图20 跨行业大数据应用的企业技术创新通用基础设施愿景

参考文献

- [1] 大数据平台: 构建下一代的分析应用程序[C]//2012 IBM智慧的分析洞察论坛。
- [2] SUH I.C. Smart-analytics: 大数据时代的商务变革[C]//IBM Corporation. Information on Demand Conference, Japan 2012.
- [3] Smarter Analytics [OL]. [2012-09-30]. <http://www.ibm.com/analytics/us/en/>.
- [4] 朱锐. 中国科学技术信息研究所硕士论文, 2012.

作者简介

李颖, 信息系统专业博士。主要研究方向: 知识组织与服务、语义技术等。E-mail: liying@istic.ac.cn

朱礼军, 研究员, 主要研究方向: 智能信息处理, 语义和知识技术在知识管理、竞争情报和电子政务方面的应用。E-mail: zhulj@istic.ac.cn

农国武 (1970-), 中国铝业广西分公司信息中心高级工程师, 主要研究方向: 信息服务。

Mechanism Study of Big Data Smart-analytics System Oriented Enterprise Innovation Service

Li Ying, Zhu Lijun, Nong Guowei / Information Technology Support Center, Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Big Data Smart-analytics is a new hope of enterprise technological innovations, it is also a new opportunity for science & technology institutions under the pressure of service transformation. Adopting Big Data Smart-analytics, enterprises can deep mine and present kinds of related information from a massive and variety of data to meet innovation needs, represent insight, and make more accurate innovation decisions. This article is for enterprises technical innovative services. It discusses the mechanism of Big Data smart-analytics system, focuses on the types of Big Data - structured and unstructured data, and introduces different processing techniques. Since computer processing of structured data is easier and more accurate, this article takes structured data processing technology as the main topic for discussion, and introduces the practical DITA standard technology. Finally, taking our R & D projects as an example, it describes the data analysis work and vision.

Keywords: Big data, Mechanism of smart-analytics, Technological innovations, Structured data, DITA

(收稿日期: 2012-09-22)