基于 UIMA 的企业非结构信息资源管理系统研究①

Research on UIMA Based Enterprise Unstructured Information Resources Management System

张明宝 马 静 （南京航空航天大学 经济与管理学院 信息管理与电子商务系 江苏 南京 210016）

摘 要: 由于当前非结构信息管理技术的落后，企业对其内部越来越多的非结构信息资源的利用效率非常低。此外，伴随着知识管理、协同商务等新兴管理思想的发展，企业对其非结构化信息资源的高效管理和应用的需求激增。论文提出了一种新的非结构信息资源管理的方法并且介绍了实现这一方法的关键技术 UIMA。介绍了面向用户的企业非结构信息资源管理系统的应用模式及其系统体系结构，通过一个原型系统实例证明了基于 UIMA 的企业非结构信息资源管理系统的可行性。

关键词: 信息资源管理 信息管理 非结构信息管理架构 信息管理系统 体系结构

随着信息技术的广泛应用，企业内部积累的非结构化数据资源正在急剧膨胀。但是，由于当前非结构信息管理技术的落后，企业对其非结构信息资源的利用效率非常低。如何充分的利用企业所拥有的大量非结构数据资源为企业竞争服务成为企业信息管理研究的一个重要问题。

非结构信息资源管理研究涉及的领域主要包括内容管理、企业级信息检索和知识管理。内容管理[1,2]强调借助信息技术实现对内容的创建、储存、分享、应用和更新的全过程的管理。企业级信息检索[3]主要研究在分布、异构的企业信息资源中利用信息检索技术快速寻找所需要的信息内容，它与传统的 web 检索在检索需求、检索对象、检索方法等方面具有本质的不同。知识管理[4,5]研究的内容涵盖知识的表示、发现、存储、利用和再生的全过程。知识管理的基础是信息管理，特别是非结构化信息的管理。

非结构信息资源管理的全生命周期过程包括非结构信息资源的获取、分析、标注、存储和应用。尽管在内容管理、企业级检索以及知识管理系统研究领域都涉及到了对非结构信息进行管理的问题，但是这些研究都各有侧重，无法解决面向全生命周期过程的非结构信息资源管理问题。内容管理是非结构信息资源管理的第一阶段，如何在快速变化的环境中实现面向用户个性化需求的、柔性的、智能的、集成的内容管理仍然有待解决 企业级检索研究涉及的面较窄，没有涉及非结构信息资源的综合应用问题，企业级检索是非结构信息资源管理的一个具体内容 知识管理系统的研究侧重于知识的发现、管理和应用，缺乏对非结构信息资源管理的专门研究，非结构信息资源管理是知识管理的基础。

论文以企业非结构信息资源管理系统为研究目标。首先提出一种以自动分析为基础的非结构信息资源管理的新方法，并且分析了面向用户的企业非结构信息资源管理系统的应用模式及其所应具备的特征，指出企业非结构信息资源管理系统的实现需要合适的底层支撑技术，在此基础上介绍了企业非结构信息资源管理系统实现的关键技术 UIMA。论文重点描述了基于 UIMA 的企业非结构信息资源管理系统体系结构以及实现这一体系结构的原型系统。

① 基金项目 南京航空航天大学引进人才项目（1009－234039）

1　以分析为基础的非结构信息资源管理方法

对企业非结构信息资源的管理主要依赖于其外表特征和内容特征。外表特征包括作者、时间、出处、序号等,内容特征包括关键词、主题、分类等等。传统的信息资源管理主要依靠人工标注获得信息资源对象的外表特征和内容特征,从而进行管理。采用人工标注的方法将非结构化数据转换为结构化数据需要耗费大量的人力、物力和财力,效率低下。此外,在企业应用环境中,由于用户需求的多样性,往往需要对数据资源内容作深度挖掘,譬如挖掘文档内容中包含的人、机构、产品、地点以及它们之间的关系等。这就使得传统信息资源管理方法完全不适合于企业应用。

为了解决企业环境中非结构化信息资源的有效管理和应用问题,提出图 1 所示的以自动分析为基础的非结构信息资源管理方法。该方法的最大特征在于强调采用各种分析技术自动抽取非结构数据源对象的特征,然后根据这些特征对数据源对象进行自动标注,将非结构数据自动转换为结构化信息,在此基础上进行非结构化信息资源的高效管理和应用。

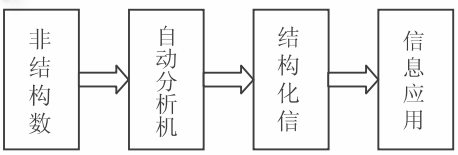

图 1　非结构信息资源管理的方法

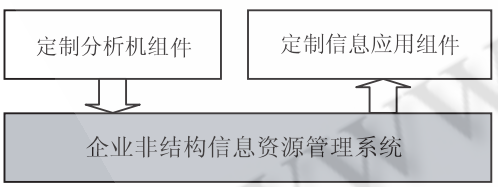

图 2　企业非结构信息资源管理系统应用模式

上述方法完全依赖于各种自动分析机对数据源对象的准确分析。对文本进行分析需要应用自然语言处理、信息检索、智能文本分析以及知识挖掘等领域的各类技术,而这些研究领域经过多年的发展已经积累了大量的技术成果可供我们使用。现在的主要问题在于如何将现有的各类成熟的术综合应用于企业环境中的非结构化信息资源的有效分析。

2　面向用户的企业非结构信息资源管理系统

研究企业环境中的非结构化信息资源管理系统必须要考虑三个问题:一、不同企业的用户对非结构信息资源的管理需求是不同的;二、用户对非结构信息资源管理的需求是随着应用环境的变化而快速变化的;三、单一的文本分析技术是无法实现对信息源对象的深度分析的,需要综合使用各种文本分析技术。

为了解决这三个问题,提出了图 2 所示的企业非结构信息资源管理系统应用模式。一个具体的企业非结构信息资源管理系统应用是由定制文本分析机组件、定制信息应用组件和企业非结构信息资源管理系统三部分组成。其中定制分析机组件和定制信息应用组件是面向用户特殊需求的,根据用户应用需要灵活设计的。定制分析机组件又是由定制信息应用组件的具体功能所决定的。企业非结构信息资源管理系统是面向企业非结构信息资源管理共性需求所设计的,为定制分析机组件和定制信息应用组件提供运行的基础环境,同时提供标准的对外接口方便用户定制分析机组件和信息应用组件。

面向用户的企业非结构信息资源管理系统必须具备如下特征:(1)必须对非结构信息资源获取、转换、分析、管理、应用的全过程进行分析,提供基于标准工作过程的支持环境;(2)必须提供标准的对外接口、信息描述方法和定制规范降低定制分析机组件和信息应用组件的复杂性;(3)必须提供灵活的信息资源描述模式简化结构化信息资源库的构建;(4)采用自然语言处理技术以支持高质量的‘拉式’信息服务和知识抽取;(5)提供对外的标准接口以支持非结构信息资源管理系统与企业其它应用系统的集成;(6)提供界面友好的工具方便用户进行系统管理和应用;(7)其本身应具有易于扩充、动态发展的能力。

可见,面向用户的企业非结构信息资源管理系统非常复杂,我们需要仔细选择合适的支撑技术和研究柔性的系统体系结构来实现上述系统目标。

3　非结构信息资源管理架构 UIMA[6,7]

按照图 1 的思想分析企业非结构信息资源要求综合应用各类分析技术,这需要合适的集成框架技术。按照图 2 所示的应用模式来开发企业非结构信息资源管

理系统要求研究柔性的系统体系结构,这也和选择合适的集成框架密不可分。IBM 于 2005 年 12 月发布的非结构信息管理架构 UIMA(Unstructured Information Management Architecture)为我们解决上述问题奠定了坚实的技术基础。

UIMA 提供了在企业级的环境中处理各类非结构化的信息资源的通用解决方法和支撑技术。目前 IBM 已将其 UIMA 技术规范提交 OASIS 讨论希望能够作为非结构信息处理的统一标准发布。为了促进 UIMA 的快速发展,IBM 将其 UIMA 源代码提交给 Apache 作为其一个开放源代码项目,Apache UIMA 目前提供支持 JAVA 和 C/C++的免费软件包下载。

UIMA 框架是由 UIMA AE FACTORY 和各种 AE(Analysis Engines)构成。UIMA AE FACTORY 负责各种 AE 的运行、调度和管理。AE 负责具体的数据源分析工作,并且向用户提供标准的编程接口,使用户可以编程访问分析的结果。

AE 的核心是分析算法,它承担分析文本和记录分析结果的所有工作。UIMA 提供一个基础的组件类型 type 来封装 AE 中运行的核心分析算法。这个组件的实例被称为 Annotators。UIMA 框架提供合适的方法来使用 Annotator 和创建 AE。最简单的 AE 只含有一个 annotator,复杂的 AE 则包含一组其它的 AE,这些子 AE 中可能还包括其它子 AE。Annotator 的工作是找到特定类型数据的实例并创建 Annotation 或者该数据的实例。Annotation 包含实际的数据以及其在文档中的位置。每一个 Annotators 的分析结果都是由 CAS(Common Analysis Structure)精确描述的。CAS 是一个基于对象的数据结构,它逻辑地包含了被分析的文档,并将分析的结果创建为一个相应的对象。这些对象可以有复杂的继承结构。AE 将其信息存储在 CAS 对象中。一个 CAS 对象包含被分析文件的所有 Annotation。使用 CAS 我们可以采用标准方式对最终结果进行分析。

UIMA 提供的过程框架可以将各种简单的或复合的 AE 按照用户定义的分析过程集成为一个满足特定分析需求的复杂 AE,这为 UIMA 提供了广阔的扩展空间。用户使用 UIMA 的关键在于开发分析算法,封装分析算法、定义 CAS 结构以及定义分析流程。UIMA 为采用通用方法进行非结构信息管理奠定了坚实的基础。

4　企业非结构信息资源管理系统体系结构

图 3 所示为基于 UIMA 设计的一种企业非结构信息资源管理系统的体系结构。该平台分为三大部分:分析资源库、运行环境和组件库。分析资源库中的资源组件以及组件库中的分析组件和应用组件都是通过标准接口集成于 UIMA 之上的。UIMA 是实现本系统的核心。

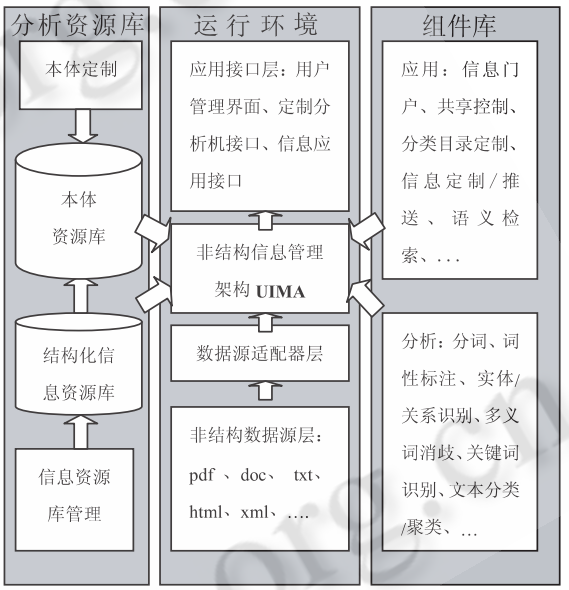

图 3　基于 UIMA 的企业非结构信息资源管理系统体系结构

运行环境分为四层,是在 UIMA 基础之上扩充而成。最底层的非结构数据源层表明了本系统处理的非结构数据源对象的类型。目前企业中主要的非结构数据源对象格式有 pdf、doc、txt、html、xml 等,其中最基本的文件格式是 txt 文件格式。数据源适配器层针对前述主要的文件类型提供映射工具,将各种不同格式的数据源转换成上层非结构数据管理集成框架所使用的标准数据源格式。对于其他任意的数据格式可开发适配器直接向标准格式映射,也可采用多级映射的方法,先将其转换成上述几种主要的文件格式之一,就可以利用系统提供的相应适配器再向标准格式转换。将数据源适配器层独立出来增加了对各类非结构数据源处理的灵活性。UIMA 是运行环境的核心,它是各种组件运行的‘容器’,通过标准接口将非结构信息资源获

取、转换、分析、应用和管理的全过程所涉及的各个功能部件集成起来满足用户的应用需求。处于最上层的是应用接口层，它包含三部分：用户管理界面提供 GUI 界面的系统管理功能，方便用户配置、运行、管理和维护系统；定制分析机接口提供标准的外部组件调用接口、分析过程描述语言和资源描述语言来定制各类自动分析机组件；信息应用接口提供标准的编程接口、应用过程描述语言和结构化信息资源描述语言来定制各类满足用户特定需求的信息应用组件。

分析资源库主要包括本体资源库和结构化信息资源库两大部分。本体资源库用来存储各类自动分析机所需要使用的本体资源。本体资源在自然语言处理与知识挖掘领域中的应用越来越广泛，特别是对于语义级别的文本分析更是不可缺少。本体资源包括通用语言本体和领域本体。通用语言本体可以直接采用各类语义词典，如知网、WordNet 等。领域本体是与用户的应用需求密切相关的，可以直接采用现有的成果，如企业本体 TOVE 等，也可以使用本体生成工具定制和管理特殊的领域本体。目前关于本体定制的工具比较丰富，譬如斯坦福的 protégé 等。结构化信息资源库用来存储经自动标注以后产生的各种结构化了的信息资源。非结构数据源经自动分析以后被抽取出各种特征，可以用这些特征描述源数据对象。特征以及特征之间的关系是可以采用规范化的方法定义的，可以按照这些特征及其关系重新组织数据源，使其从非结构化转换为结构化。定义资源库的数据库模式是比较复杂的，资源库管理组件提供相关的功能支持。此外图 3 中结构化信息资源库指向本体资源库的箭头表明我们可以从大量结构化的信息源中通过统计的方法或者学习的方法自动的扩充本体库。

组件库是指本系统提供的可供用户直接使用的组件，这些组件应该实现通用的、基础的信息资源分析和应用功能。分析组件库提供的组件最少应包括汉语分词、词性标注、命名实体/关系识别、多义词消歧、关键词识别、文本分类、文本聚类等组件；信息应用组件库至少应提供信息门户服务、信息共享控制、分类目录定制、信息定制/推送、语义检索服务等组件。用户可以在应用的过程中开发可重用的功能组件，不断地扩充组件库。将组件库独立出来表明系统具有很强的自我扩充能力。

5　原型系统实现

图 4 为我们实现的原型系统结构图。UIMA 框架采用 Apache UIMA 的 JAVA 软件包，使用 Eclipse 作为开发环境。按照 UIMA 规范在系统中设计了一些常用组件。其中分析机组件有人名识别分析机、产品名识别分析机、地名识别分析机、机构名识别分析机以及语义识别分析机。这些分析功能的实现是通过封装我们自主开发的自然语言处理软件包 WSDisambiguation-Toolv1.0 的相关功能实现的。我们使用知网 2.0 作为中文通用语言的语义词典。信息应用组件有本体库构建工具、知网词义查询工具、信息推送定制工具、语义级检索工具和检索匹配工具。其中，本体库构建工具使用了斯坦福的 protégé 软件包，检索匹配工具使用了 Lucene 软件包。

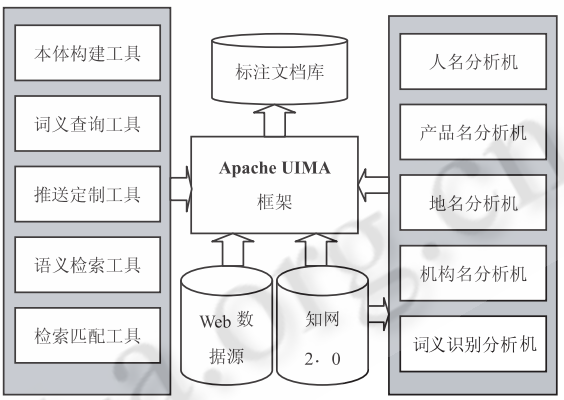

图 4　原型系统结构示意图

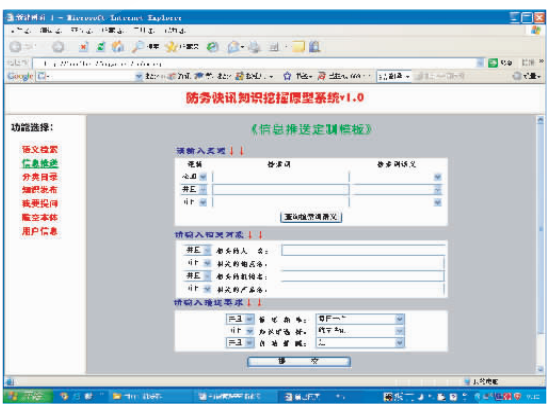

图 5　防务快讯知识挖掘原型系统主界面

使用该原型系统可以灵活的构建各种非结构信息

资源管理的应用。防务快讯挖掘原型系统是南航经管院承担的一项国防技术基础项目的关键内容,该系统的主要目的在于从大量的防务快讯中挖掘有价值的信息为决策人员提供参考。该系统是在图 4 所示原型系统基础上通过功能裁剪和二次开发实现的。图 5 所示为防务快讯挖掘原型系统的信息推送定制工作界面,可以通过输入“ 主题 ”检索词和“ 相关对象 ”检索词按照“ 某某对象的某某主题 ”规则实现语义级的推送信息过滤。

此外,我们还可以利用 UIMA 提供的 GUI 工具定制特殊的分析过程。用户可以根据对文挡集合的特殊分析开发各种具体的应用。譬如,通过非结构信息资源寻找某人的邮箱信息,某单位的电话等等。可以根据这些信息在文档中的表现模式来编程实现具体的分析算法,然后使用 Apache UIMA 提供的 Document Analyzer 工具进行分析获得结果。这些分析机一旦开发出来以后就可以存入组件库供其他用户重用,这样随着原型系统使用的增多,其功能也在不断扩充。

6 结束语

由于当前非结构信息管理技术的落后,企业对其内部越来越多的非结构信息资源的利用效率非常低。此外,伴随着知识管理、协同商务等新兴管理思想的发展,企业对其非结构化信息资源的高效管理和应用的需求正在激增。研究功能强大的企业非结构信息资源管理系统是解决这一矛盾的根本办法。企业非结构信息资源管理系统必须要能够满足企业多样化和个性化的应用需求,要适应用户需求和环境的快速变化,同时要综合应用自然语言处理、文本挖掘、本体论、协同工作等领域的各类技术。IBM 的非结构信息管理架构 UIMA 技术为解决这些问题提供了一个很好的技术规范和集成框架。论文以 UIMA 为基础构建了一个面向用户的企业非结构信息资源管理系统,未来我们的工作在于不断扩充该系统的功能并且开发解决具体问题的非结构信息资源管理应用。

参考文献

1 Browning P, Lowndes M. JISC TechWatch Report: Content Management Systems. http://www.jisc.ac.uk,2008.1.

2 Autonmy,Inc. Introduce to Autonmy. http://www.autonomy.com/,2006.12.

3 Laurent Proulx. Enterprise search as a productivity tool. http://www.nstein.com,2007.8.

4 Washington University. Knowledge ManagementTools. http://courses.washington.edu,2006.12.

5 Ittoolbox Inc. Information Technology Toolbox. http://knowledgemanagement.ittoolbox.com,2006.12.

6 IBM. Unstructured Information Management Architecture (UIMA). http://domino.research.ibm.com/, 2007.12.

7 Apache UIMA Development Community. UIMA Tutorial and Developers´ Guides. http://incubator.apache.or[illegible] 2007.12.