InfOnCall XML 网关

第 1 版

因科技术（上海）有限公司

地址：上海市淮海中路 381 号中环广场 2201-2208 室

邮编：200020

电话：021-63916988

传真：021-63915988

E-mail：contact@InfOnCall.com

www.InfOnCall.com

声明

本文件作为因科技术(上海)有限公司的知识产权，仅供被授权方内部参阅，不得外传或用于其他途径。

InfOnCall 是因科技术(上海)有限公司的注册商标。其他品牌分别属于其注册者。

1. 术语

HTML——HyperText Markup Language，超文本链接标识语言

ICP——Internet Content Provider,互联网内容运营商

ISP——Internet Service Provider，互联网服务运营商

IVIP——InfOnCall Voice Internet-access Platform，因科语音网络服务平台

HTML2XML——HTML 文件到 XML 文件转换器

XML2DB——XML 文件到关系型数据库转换器

DB2XML——关系型数据库到 XML 文件转换器

IXG——InfOnCall XML Gateway，因科 XML 网关

XML——Extensible Markup Language，可扩展标识语言

XSL——Extensible Stylesheet Language，可扩展样式单语言

VoiceXML——Voice Extensible Markup Language,语音可扩展标识语言

2. 简介

因科 XML 网关(IXG)是因科技术(上海)有限公司研制的技术创新产品，用于 HTML 文件的实时抓取、自动分析和转化，XML 和传统关系数据库的实时转化和交互，以及 XML 文件和 VoiceXML、WML 等之间的实时自动翻译。

因科公司专注于以 XML 为基础的网络软件技术的开发。为了帮助用户将大量的以 HTML 文件形式存在的信息发布到更多的用户，帮助用户将大量基于传统关系数据库的孤立应用（“数据孤岛”）发挥更大的价值，因科公司研制了 IXG 产品。

根据要求分析和转换的问题对象的不同，IXG 中目前包含三个部件：

1） **HTML2XML** 用以将 HTML 文档自动转换为 XML 文档。

2） **DB2XML** 让您不必编写复杂代码就可以从关系数据库中得到结构化的 XML 文档，它是数据库整合、数据库交换以及数据库转换的必备工具。

3） **XML2DB** 实现了从 XML 数据向传统关系数据库的输入。

3. IXG的背景和需求

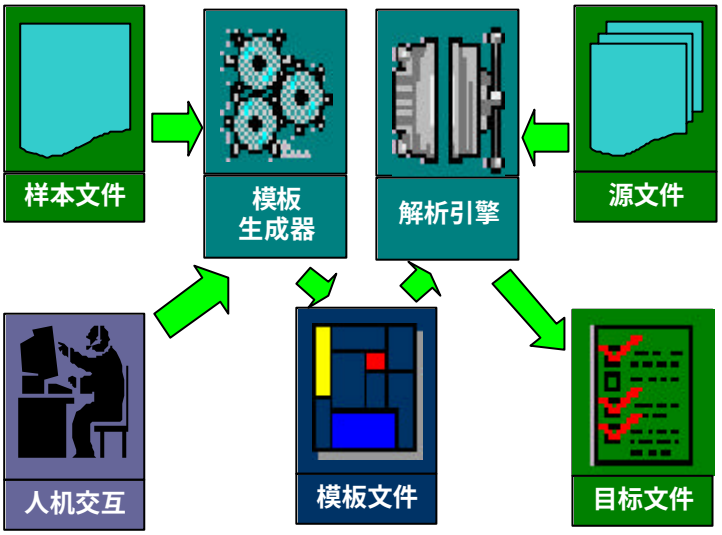
可扩展标记语言(XML，Extensible Markup Language)目前正在成为各种数据特别是文档的首选格式。由于它具有标记不同字段(field)的能力，使得搜索变得更简单和动态化，从而把企业准备扔进废纸篓的文件变成了进行数据挖掘的宝藏。XML 把内容从演示格式中解放出来，使材料可以多次重复使用。这样一来，同样的内容可以分别用于新闻发布、白皮书、宣传册、演示和 Web 页面。对那些需要把不兼容

的系统融合在一起的企业，XML 可以充当公共传输工具，以中性格式进行数据传输。此外，XML 还可以处理各种数据，包括文本、图像和声音，并且可以由用户进行扩展以处理任何特殊类型的数据。XML 的特性使之成为在线和离线数据的共同语言。

现在的问题是，怎样既利用到 XML 的优点和好处，又充分地与现有的应用结合。一种方法是根据原有的数据格式，重新构造 XML 格式的数据，这种方式所需要的工作量和代价是很大的；另一种方式是通过工具直接对现有的数据文件进行分析并转换出 XML 格式的数据,进而可以通过 XSL 技术转化为任何扩展类型的 XML 格式，如 VoiceXML 用于语音，WML 用于 WAP，XHTML 用于 WEB 等等。

因科 XML 网关（IXG）正是针对这样的背景和应用需求而应运而生的。IXG 实现了 HTML 到 XML 数据的自动抓取和转换，支持目前市场上几乎所有的关系数据库到 XML 的转换，支持 XML 数据到传统关系数据库的输入。

4. 系统架构



图表 1 IXG 原理结构图

IXG 的原理结构如上图所示。IXG 中包含的每个系统（HTML2XML，DB2XML，XML2DB）都是按照上述原理结构图来构造的。系统主要有两个模块组成：模板生成器和解析引擎。模板生成器提供图形化界面，让用户有机会指明要转化的数据模块，这通过用户输入数据样本，指明数据模块来完成，结果用户会得到指导解析引擎工作的模板文件。解析引擎根据用户预先配置好的指导性文件（模板文件），对实时更新的旧数据文件进行解析，分析出用户感兴趣的数据，组装成 XML 文件，供其他

应用程序进行进一步处理。IXG 解析引擎支持两种用户界面：Service 和 API。Service 界面不需要用户有较深的编程经验；API 界面为开发人员提供更灵活的编程接口，帮助开发人员把 XML 转化功能集成到用户的系统软件中。

5. IXG的功能和特性

5.1 HTML2XML

用以将HTML文档自动转换为XML文档。目前主要针对以表格数据为核心（data-centric）的HTML格式文件。这是由于XML标准主要是用以精确标识所包含的数据,而有进一步应用需求的HTML文件多以含有Table的 Data-Centric文件为主。目前该工具功能主要包括：

- 提供基于XML的语言来表达如何从HTML网页获取复杂结构；
- HTML到XML声明性文档的映射，可以根据相应的解析模板自动产生XML；
- 提供可视化工具使得开发更加的迅速和便捷。

5.2 DB2XML

让您不必编写复杂代码就可以从关系数据库中得到结构化的XML文档，它是数据库整合、数据库交换以及数据库转换的必备工具。它可以工作在任何的平台（UNIX/NT）中，连接到任何存在的数据库(MS SQL Server, IBM DB2, Oracle, MySQL, MS Access, Informix, Sybase 等等)。除此以外，DB2XML工具还提供了非常灵活、容易使用的树型结构的查询工具，让你能够从复杂的数据中选取出你所需要的数据，并且非常方便地发布为与应用有关的XML或者HTML格式的数据。

5.3 XML2DB

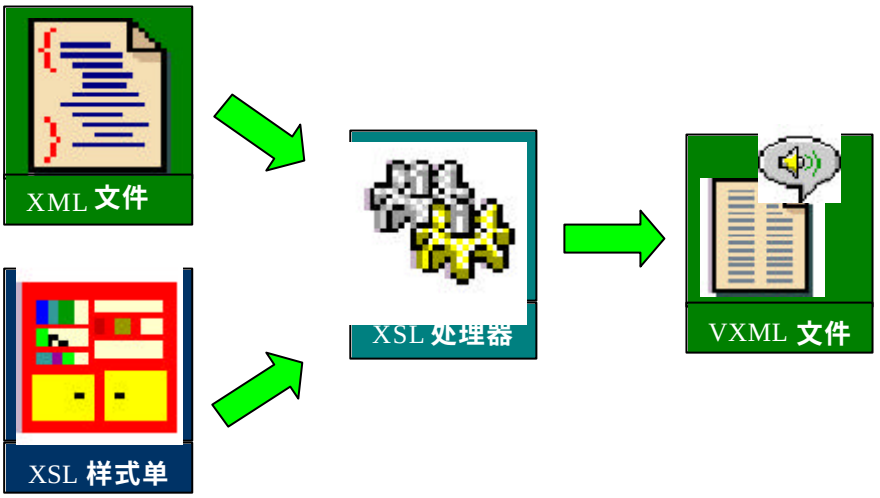
实现了从 XML 数据到传统关系数据库的转化。XML2DB 的主要原理是通过一个 map 文件来描述数据库中的域或表和 XML 树型结构之间的对应关系。在作 XML 数据到 Database 之间转化时通过对 map 文件的分析为 XML 中的每一个节点找到对应的字段或表。这种转化工具支持多种后台数据库类型，可以采用的数据库包括 Oracle，DB2 等。

5.4 IXG特性

- 有效利用现有的信息资源；
- 快速建立和商业伙伴的合作；
- 无缝升级到基于XML的网站系统
- 提供多渠道发布的转换中间件；

- 将原有的信息的内容和表现更好的分离，有利于增加商业机会，提高企业灵活度和竞争力。
- 从任何的数据库自动地创建结构化数据的 XML
- 适用于所有的数据库和平台
- 最大程度支持 XML 标准
- 使用简单方便

6 从XML到VXML



图表 2 XML 转化为 VXML 工作原理图

IXG 最主要的功能在于：从现有的 HTML 文件和传统关系数据库中分析出用户感兴趣的数据，转化为 XML 格式的文件。有了这些良好格式而且自解释的 XML 文件，用户可以通过手头上的各种工具，将 XML 集成到他们的系统软件中。

比如在因科技术公司的 IVIP（因科语音网络服务平台），它是连接互联网和电话语音网的“桥”。这种桥接作用是通过支持 VoiceXML 并提供一个 VoiceXML 运行平台来实现的。对于用户而言，这时可以把注意力集中在用户逻辑上，其中最关心的问题是如何将他们原有的 HTML 文件和数据库内容方便快速地转化为 IVIP 支持的 VoiceXML 格式。考虑到用户将来会把他们的数据发布到语音和 Web 以外的渠道，如 WAP, PDA 等，所以将用户的现有数据先转化为标准的 XML 格式，然后根据用户需要将 XML 文件转化为 VoiceXML、XHTML、WML 或 HDML 等。IXG 目前已经完成了前半部分功能，而后半部分需求可通过 XSL 技术（如图表 2 所示）来完成。XSL 是 W3C 制定的专门用于支持 XML 高级应用的样式单语言，它最大的特性是把 XML 转化为 VXML/HTML/WML/HDML 等的能力。在图表 2 中，用户设计的人

机对话逻辑将体现在 XSL 样式单中，模板合成器按照 XSL 样式单的描述从 XML 文件中读出数据，然后封装成 VXML 文件。

7. 应用前景

1) 电子商务应用中的数据交换

在E-Commerce应用系统中，往往需要在多种应用、平台之间共享、交换数据。为了解决异构应用系统之间的通信问题，IXG将E-Commerce应用中原有系统的数据转化为统一格式的XML。XML的灵活性和扩展性使其可以对不同应用甚至是差异很大的应用间的数据进行描述，尤其是对于那些专用于记录数据的应用。另外，XML具有自我描述的特性，结果是数据可以在不同的应用间进行交换与处理而不必要求相应的应用程序是针对该数据定制的。

2. 强大的网站内容管理

目前HTML的将内容和表现形式捆绑在一起的固有缺点使得原来的网站模式很难符合新的需求，特别是在商务之间相互通信的场合。XML的产生和相关技术的成熟，特别是基于XML的XHTML逐渐更新HTML，使得越来越多的网站逐渐升级到基于XML设计的网站。在这个过程中既要将新的内容以XML的方式存储和发布，同时也要考虑到兼容原来的数据。这就需要将原来的数据进行组织和转换。通过IXG，用户只要在目标XML的DTD文件中定义一系列有意义的标记，这样基于该DTD文件从数据库或HTML文件中产生出来的XML文档就可以按照任意的条件进行查询和检索，甚至实现计算机自动检索，而相应的检索引擎可以是通用的而不必局限于具体的应用。

3. 多种信息发布模式支持

IXG也提供了这样机制，既可以将HTML或数据库转换为独立于应用的XML通用格式，然后通过XSL将XML继续格式化成HTML、WML或VoiceXML等。这样通一次数据库到XML的转化，可以将数据库中的内容发布到Web、 Wap或语音渠道等。这将是新一代网站发展中的重要环节。

4. 网站与增值服务提供商的数据交换。

一般的情形，网站已经通过Internet发布其信息内容(比如汇率、证券信息、气象信息等)，这样的信息通常是通过其服务系统不同的格式和渠道进行发布(比如提供给WAP手机)。在进行实施过程中，要直接开放其原来的后台数据库可能对数据来源的安全性造成影响；或者有可能不同的频道信息来自不同的网站，也就可能来自不同的平台和数据库。这就需要直接针对HTML，通过调用应用服务器而不是访问后台数据库的方式来获取网页信息，并且转换成为统一的基于XML格式。XML具有独立

于平台和发布渠道的特点，可以很好地用于各种不同方式的发布。