第 14 卷第 7 期
2 0 0 8 年 7 月
计算机集成制造系统
Computer Integrated Manufacturing Systems
Vol. 14 No. 7
July 2 0 0 8

文章编号:1006 - 5911(2008)07 - 1297 - 09

一种文档自动生成模型的构建及其应用

曲明成,廖明宏,吴翔虎,刘志强
(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

摘 要:针对企业在编辑数据汇总文档(Word 格式)时因手动计算、人工校验造成效率低下且容易出错等问题,通过分析数据汇总文档的特点及其所包含的基本数据类型,提出了一个文档自动生成数学模型。该模型对文档生成过程中一些必要的操作及文档模板进行了定义,并应用模型经过迭代计算,将文档模板转化为特定日期的数据汇总文档。该模型已应用于某发电厂的经营决策文档自动生成系统,并成功地集成于工作流系统中。应用实验表明,该模型是可行且有效的。

关键词:文档自动生成;数学模型;数据分析;工作流;发电厂;系统集成
中图分类号: TP311 **文献标识码**: A

Construction of an automatic document generation model and its application

QU Ming-cheng, LIAO Ming-hong, WU Xiang-hu, LIU Zhi-qiang
(School of Computer Science & Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: To solve the problem of low efficiency, frequently occurred mistakes, and the necessity of manual calculation and validation in the compilation of data-gather documents (Word format) in enterprises, a mathematic model to automatically generate document was put forward by analyzing the characteristics and basic data types of the documents. This model defined essential operations and document template in the process of document construction. Furthermore, the applied model translated the document template into data-gathering documents with specific date through iterative calculation. Then, this model was applied in the management and automatic document generation system in a power plant and it was successfully integrated with workflow systems. It enabled integration process possible for automatic generation document to checkout. Experimental results showed that this model was feasible and effective.

Key words: document auto-building; mathematic model; data analysis; work-flow; power plant; system integration

0 引言

工作流是在多个参与者之间按照某种预定义的规则,自动地传递文档、信息或任务的过程,用以实现某个预期的目标,或促使此目标的实现[1]。越来越多的企业希望将日常工作融合到自身工作流系统中。如果单独搭建满足特定企业工作需求的系统,企业将面临与工作流系统集成的复杂问题[2]。

当前信息化建设良好的企业对各种数据的分析处理提出了更高要求。数据的分析方式很多,初级分析方式主要为报表,高级分析方式主要为联机分析处理(On Line Analytical Processing,OLAP)。报表展示给决策者的是数据和表格,而 OLAP 为决策者提供了多角度、多层次、高效率的数据探查方

收稿日期:2007-09-19;**修订日期**:2007-11-22。Received 19 Sep. 2007; accepted 22 Nov. 2007.
基金项目:黑龙江省电力总公司科技攻关资助项目。**Foundation item**: Project supported by the State Power Company of Heilongjiang Province, China.
作者简介:曲明成(1980 -),男,黑龙江哈尔滨人,哈尔滨工业大学计算机科学与技术学院博士研究生,主要从事商务智能系统、嵌入式计算等的研究。E-mail: qumingcheng@126.com。

式[3-4]。但是无论是报表还是 OLAP,展现给决策者的只是各种表格或图形化的知识[5],而决策者更希望看到的是将这些数据按照一定的业务规则,经过复杂的计算和分析得到的综合性数据汇总文档(Word)。现有的数据展现技术和分析方式已经不能满足这种需求[6-7]。目前,国外对于文档自动生成技术的研究处于起步阶段,尚缺乏完善的理论;国内的研究只是简单地用工作人员在运行时手动输入的内容替换模板内预先编辑好的标签,从而生成规范的工程文档[8]。

另一方面,企业各部门在制作各种数据汇总文档时,都要用到企业数据库(或数据仓库)中的基础数据,如计划、生产、销售、工资等。虽然各部门根据自己的业务在使用文档的格式、样式和内容上可能不同,但是对于不同时期同一个部门的专有数据汇总文档,在格式或样式上基本不会有大的变化,只需要针对特定的日期汇总数据即可[9]。综合数据汇总文档可能包含如下各种数据:从数据库直接可提取的、需要引入常数计算的、需要应用公式对多个直接提取的数据进行复杂公式计算的、需要引入工程或统计函数进行计算的、需要分析指标变化具体原因的数据类型等。目前,绝大部分企业都是由专门人员根据各种报表提供的数据进行汇总、计算、编辑和审核,不仅工作效率低,而且带来错误的可能性也很大[10]。倘若单独开发一套系统来完成文档的自动生成工作,则该系统与原有工作流系统的集成又会变得很困难。在工作流管理系统中,过程定义负责将企业的实际经营过程和生产过程转化为计算机可处理的工作流模型。因此,各种文档的审核流程完全可以由工作流完成,如果文档自动生成系统产生的文档在工作流中的各审核节点可以读取,集成问题将迎刃而解[11-12]。

因此,问题的关键是如何设计一个能够对文档进行多样式处理、格式可灵活编辑、计算公式自定义、具有保存指标历史变化原因的知识库和通用性强的文档自动生成系统,以及该系统如何与工作流系统进行集成等,以满足上述需求。

对于文档自动生成理论,目前国内外尚无完善的解决方案。本文在研究文档自动生成系统的组成及实现方式的基础上,建立了一个文档自动生成模型,并给出相应算法,最后对模型进行应用验证。该模型已应用于某电力制造企业的经营决策文档自动生成系统中,并成功地将该系统与工作流系统进行了整合。为了阐述的准确性,文中的业务公司及业务指标均以电力制造企业为例。

1 问题分析及解决方案

文档自动生成系统与工作流系统的整体架构如图 1 所示,文档自动生成系统从数据库(或数据仓库)中提取需要的数据生成文档,并存入数据库中,而后工作流系统读取数据库中的文档,并在各个流转节点显示,经过流转完成文档的审核。从结构中可以看出,生成的文档必须直接存入数据库中,才能实现各个节点的灵活读取。目前主流的数据库系统基本都支持二进制流数据的读取,包括语音、视频、图片、文档等,因此将生成的文档以二进制流格式存入数据库是可行的。在各工作流节点将二进制流格式文件转化并显示成正常的文档,从而实现系统的集成。

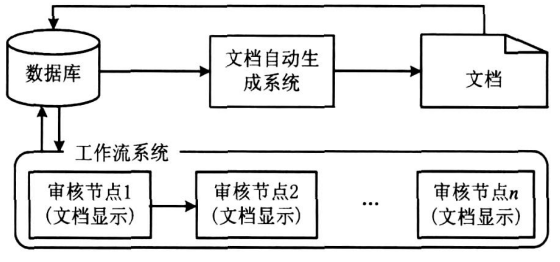


图1 系统集成结构图

现在要研究的关键问题是如何实现文档自动生成系统。文档应包含固定不变或变化很少的文字部分,以及随文档生成日期变化的数据部分两部分内容。如果将一个已有文档中变化的数据进行编辑,保留固定的文字,将各种变化的数据转化成文档全局唯一关键字,针对这些关键字进行其他计算规则的封装(可以将其称为文档模板),在运行时根据输入的日期等条件,结合各关键字的计算规则,进行数据搜索和计算,并用生成的数据替换全局唯一关键字,文档将可以自动生成。

上面分析中的每一步经过实际证明都可以实现,因此这种文档生成系统与工作流系统的结合策略是可行的。那么文档应包含哪些元素呢?一个文档的组成除固定的文字部分外,其余数据均是随日期变化的。这些变化的部分统称为变元,变元应包括:日期型、增长下降类文字型、仅一次性数据库搜索可得到结果的、需要多次数据库搜索并结合一定的业务公式进行计算的、需要重用已经计算过的数据并可能需要多次数据库搜索和一定业务公式进行

综合计算的,以及影响指标变化的具体原因等。如果数据需要根据业务公式计算得出,则在各级文档操作者那里动态查看指标相应的业务公式,以及具体的算术公式,就可以核查数据变化的具体原因。同时生成的文档应是动态可编辑的(系统中直接可编辑),这样才能进行批注、原因分析等其他操作。

2　模型构建与算法研究

经研究发现,文档自动生成系统应涉及到三种公式: ①算术公式(如 4 - 3); ②变量公式(如 I - J); ③业务公式(如:全口径售电量 - 过网电量)。系统将业务指标展现给普通操作人员(如:全口径售电量),系统自动将业务指标映射为系统变量(如:全口径售电量 I,过网电量 J),业务员根据业务指标和与其对应的变量、业务指标和与其对应的业务公式来编写变量公式,如业务公式为"全口径售电量 - 过网电量",对应的变量公式就为 I - J。系统将变量及变量公式进行数据库持久化存储。算术公式是系统根据变量公式在运行时针对特定的输入条件自动生成的,业务公式是系统根据变量公式在运行时自动生成的。特别地,对于只有一个变量的情况,如 A,(A + 2)/ 1 000,也是变量公式。

模型的构建从文档关键字的定义开始,由关键字所对应数据的计算公式引出公式计算引擎的定义。本文首先介绍如何实现从变量公式到算术公式和业务公式的转化,然后讨论变量到物理字段的映射,以及变量数据的运行时条件生成,最后针对原因关键字阐述原因知识库的构造方法和原因知识的提取策略。为进行从模板到文档的迭代计算,又定义了关键字元组、关键字对应数值元组、文档元组和模板元组,以此为基础应用模型理论进行迭代计算,最终推导出文档自动生成系统的数学模型。可以看出,模型的推导过程为采用通常求解问题的逆向思维方式,由下至上逐渐推导。下面推导中涉及的变量、公式、定义、定理适用于全文。

2.1　文档关键字

不同时期的文档包含的业务指标类型基本相同,只是指标数值不同。可以将文档中每一个随时间变化的指标转化成文档中唯一的全局关键字,如"2 月份过网电费 30 万元/ kW ·h"中的 2,30 是变化的,可以将其转化为关键字 K_1, K_2。这样可以针对 K_1, K_2 进行文档位置的精确定位,同时也可以为 K_1, K_2 封装一些其他的计算逻辑和约束条件。

定义 1　函数 $K = GK(C)$ 为将文档中的文字 C 转化为全局唯一关键字 K,即 $K = GK(C)$。

2.2　公式计算引擎

如果系统可以实现算术公式的自动计算,则就需要有一个能够完成公式计算的计算引擎。

定义 2　函数 $RS = G(X)$ 可以将算术公式 X 计算出结果,X 中可以包含任意数量的括号、常用的数学函数(如 sin,log 等),RS 为结果。

设有算术公式 $X = (5 - 2)/(40 - 30) \times 10 + \log(100)$,当执行操作 $G(X)$ 时,函数返回算术公式 X 的结果 302。假设有变量公式 $B = (X_1 - X_2)/(X_3 - X_4) \times 10 + \log(X_5)$; $X_1 = 5$, $X_2 = 2$, $X_3 = 40$, $X_4 = 30$, $X_5 = 100$,将数据代入便将变量公式转化为算术公式,执行 $G(X)$ 函数将返回算术公式的计算结果。这里 X_1, X_2, X_3, X_4 和 X_5 都是变量。对于计算引擎 $G(X)$ 的实现有多种方式,可以采用栈和逆波兰表达式,通过括号匹配实现对算术公式的计算。

定义 3　函数 $X = C(B, E, R, K)$,其中 B 是变量公式,E 是变量公式中的具体变量,R 是变量 E 的真实值,K 是每个公式对应的文档中全局唯一关键字,函数 C 将变量公式转化为算术公式 X。

其中属于同一个公式的 E 有相同的 K,不同公式的 E 包含不同的 K,且 $K \leftarrow\!\!\!\rightarrow B$。此外 E 的类型为: $E.\text{type} = \text{enum}\{DB, AH\}$,即 E 包含两种类型,DB 为从数据库查询得出,AH 为已经算出结果的某业务指标的关键字。

2.3　变量公式与业务公式的转化

在问题分析及解决方案研究中,阐述了文档操作人员应能在系统运行时,动态地查看某个指标具体的业务公式及对应的算术公式,从而可以进行数据批注及指标升降原因分析。

定义 4　函数 $A = P(B)$,B 为变量公式,A 为业务公式,函数 P 将变量公式转化为业务公式。假设求解网售平均单价的业务公式为 $A = ($全口径正常电费 - 过网正常电费$)/($全口径售电量 - 过网电量$) \times 1\,000$。设有变量 $M =$ 全口径正常电费, $N =$ 过网正常电费, $O =$ 全口径售电量, $P =$ 过网电量,变量公式 $B = (M - N)/(O - P) \times 1\,000$。当执行函数 P 后,可以得出 $A = P(B) = ($全口径正常电费 - 过网正常电费$)/($全口径售电量 - 过网电量$) \times 1\,000$。

通过上述定义可以看出,定义 3 和定义 4 分别将变量公式转化为算术公式和业务公式。

2.4 变量与物理字段的映射

定义 5 定义业务指标与物理字段的映射关系为一个三元组∂〈*BN*, *TB*, *Ω*〉,其中 *BN* 为业务指标名称, *TB* 为物理表,*Ω*为数据库字段。

定义 6 函数 *Ω* = *F*(*E*)。其中:*Ω*为数据库字段; *E* 为变量公式中的某一变量,并且 *E* 为"数据库查询"类型; *F* 实现将 *E* 转化为 *Ω*。

函数 *F* 的映射可以按如下方式实现:当业务员选择业务指标编辑变量公式时,每一个业务指标选择框固有一个变量名称,即输入框标签名称,如两个选择框 *I* 过网电量 和 *J* 过网电费(输入框可随意选择,上面的选择也可以为 *I* 过网电费, *J* 过网电量)。待指标选择结束时,系统将业务指标名称 *BN* 与变量 *E* 进行自动关联。如果此时要求解电价,则变量公式为 *B* = *J*/*I*(按本段括号中的方式进行编辑则为 *B* = *I*/*J*);执行函数 *A* = *P*(*B*)后,生成业务公式为 *A* = 过网电费/过网电量。

业务指标与物理字段的映射方式如图 2 所示,图中的的业务指标归类为电力营销总表,关联的物理表名为 DL YXZB,则在选择某一个业务指标(如全口径正常电费)时,程序可以直接将该业务指标映射为:全口径正常电费→DL YXZB→Q KJ ZCDF。

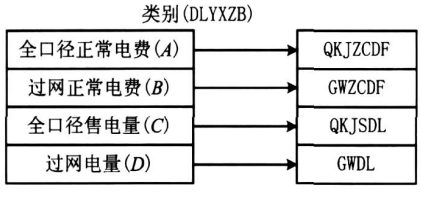


图2 业务指标与物理字段的映射

由于 *BN* 与 *Ω* 之间的关联存在于元组 ∂ 中,则有关系 *Ω*←*BN*→*E*, *F* 是正确的。

2.5 文档数据的条件生成

定义 7 函数 *R* = *f*(*Ω*),*Ω*为各类基本业务指标(变量)对应的物理字段, *R* 为通过一次数据库操作查询出的结果,若 *E* 为"数据库查询"类型,在运行时 *f* 根据具体的输入条件将字段封装成查询语句。利用变量与物理字段的映射函数 *F* 可以将 *f* 表示为 *R* = *f*(*F*(*E*))。

假设要实现下述运算:变量 *E* 对应的业务指标为"全口径正常电费",要求解的数值 *R* 为"2005 年 6 月全口径正常电费值",则首先应进行变量与物理字段的映射,然后进行时间条件封装,即 *R* = *f*(*F*(*E*)) = 执行"select Q KJ ZCDF from DL YXZB where ny = 200506"返回的结果,其中字段 Q KJ ZCDF 到表 DL YXZB 的映射由函数 F 完成,查询条件由函数 *f* 完成。

定义 8 函数 *O*(*E*, *K*′)返回变量 *E* 引用的已经计算出结果的指标的数值,且被引用指标的关键字为 *K*′。

2.6 原因知识库的建立

对于某一些指标的变化,影响因素可能会持续,也可能存在一定的重复间隔。根据影响因素的特点、影响范围、影响时间、影响指标的变化趋势等,进行人为分类,并将其按分类规则进行存储,这样就将专业人员对问题的分析及看法转化为各种约束和过滤条件,并与具体影响因素进行了绑定。随着时间的延续,积累的影响因素越多,能够准确描述当前指标变化原因的概率就越大,这些被存储的影响因素集合称为原因知识库[13]。根据当前文档的生成时间及某类指标的变化方向(升高、降低、增长、下降等),结合约束及过滤条件,将该类指标的影响因素按发生概率由大到小排序,并将发生概率最大的一项置于文档中,同时可以对该指标的影响因素进行手动选择。

定义 9 函数 *GR*(*K*, *K*′)的功能是返回依赖于关键字 *K* 的、且关键字为 *K* 的原因型指标的最大可能影响因素,令 *K_DB* 为所有知识的集合。

例如:3 月份全口径电量增长了 20 万/kW·h,原因是"天气持续寒冷,电暖气使用增多"。这里引号中部分是原因知识,它参照于"增长"这项指标,且应归属于"全口径电量"的影响因素类别中。其他的约束和过滤条件这里省略,因为根据行业的不同定义的约束类型、影响情况和发生概率不同。

2.7 模板与文档的定义

定义 10 定义文档模板为一个二元组 *T*〈*C*, *K*〉, *C* 为模板中不变的文字部分, *K* 为文档中各种指标变量的唯一关键字。*T* 为 RTF 文档——不损失任何 Word 文档的文字格式,是绝大部分富文本(Richedit)控件支持的格式,Word 不能直接导入到富文本控件中。

定义 11 定义文档关键字 *K* 的类型为一个四元组 *K*〈*μ*,*σ*,*ω*, *r*〉。其中:*μ*为根据定义 2 和定义 3 中的函数 *C*, *G* 可以算出数值结果的关键字;*σ*为参考于 *μ* 的"升高、降低、增长、减少"型关键字(例如有业务文档片段如下:"全口径售电量比同期的 20 万

kW ·h,增长了 5 万 kW ·h"。在该片断中文字"增长"对应于σ,"5"对应于μ,如果μ为"-5",则σ应变为"下降");ω为日期型关键字(如 2005 年 6 月);r 为参考于σ的原因型关键字。

定义 12　定义文档关键字 K 的最终生成值为一个四元组 $KV\langle\mu_V,\sigma_V,\omega_V,rv\rangle$,四个变元分别对应 $K\langle\mu,\sigma,\omega,r\rangle$中四类关键字的最终返回值。

定义 13　定义文档为一个二元组 $W\langle C,U\rangle$,C 为模板中不变的文字部分,U 为文档指标变量唯一关键字在给定输入条件下的最后取值。

定义 14　函数 $M(K,U)$ 将 K 替换成 U。其中 K 为文档关键字,U 为文档指标变量唯一关键字在给定输入条件下的最后取值。

定义 15　函数 $L(\sigma,K')$ 将变量σ转化为具体的文字。K 为被参考的指标的关键字 K,函数 $S(\omega)$ 将变量ω转化为运行时输入的日期时间。

2.8　计算逻辑的存储

在定义 1～定义 15 中,很多计算规则、函数、基础元素及约束条件需要与关键字进行绑定,并进行持久化存储,而其他计算逻辑则在运行时施加。

定义 16　集合 STO 为进行了持久化存储的各种计算逻辑。

持久化集合 $STO=\{K,B,E,E.\text{type},\partial,F,K\_DB,T,K\langle\mu,\sigma,\omega,r\rangle\}$。其中变量分别为:关键字、变量公式、变量公式中的变量、变量类型、业务指标与物理字段的映射关系、变量与物理字段的映射规则、原因知识、RTF 文挡模板、关键字的类型。

定义 17　操作 $\text{Get}(STO)\rightarrow T$ 实现从 STO 中取已存储的计算逻辑提供给模板进行处理。

2.9　文档数学模型的推导

根据前面的定义可以将模板转化为文档,具体推导如下。

步骤 1　根据定义 13 得 $U=KV\langle\mu_V,\sigma_V,\omega_V,rv\rangle$。

步骤 2　根据定义 14、定义 16 和定义 17 得 $W\langle C,U\rangle=\text{Get}(STO)\rightarrow T\langle C,M(K,U)\rangle$。

步骤 3　根据步骤 1 和步骤 2 得 $W\langle C,U\rangle=\text{Get}(STO)\rightarrow T\langle C,M(K,KV\langle\mu_V,\sigma_V,\omega_V,rv\rangle)\rangle$。

步骤 4　根据定义 1 得 $K=GK(C')\rightarrow\mu=GK(C_1')$,$\sigma=GK(C_2')$,$\omega=GK(C_3')$,$r=GK(C_4')$。

这里 C_1',C_2',C_3',C_4' 代表几种关键字对应的文档中的文字。

步骤 5　根据定义 1～定义 4、定义 7、定义 8 和定义 11 得:$\mu_V=G(C(B,E,f(F(E)))\,||\,O(E,K'),\mu)=G(C(B,E,f(F(E)))\,||\,O(E,K'),GK(C_1'))$,其中"||"表示或。

步骤 6　根据定义 15 得 $\sigma_V=L(\sigma,K')=L(GK(C_2'),K')=L(GK(C_2'),GK(C'))$,$\omega_V=S(\omega)=S(GK(C_3'))$。

步骤 7　根据定义 9 得 $rv=GR(r,K'')=GR(GK(C_4'),GK(C''))$。

步骤 8　由步骤 1～步骤 7 推出文档数学模型为:

$$W\langle C,U\rangle=\text{Get}(STO)\rightarrow T\langle C,M(GK(C'),KV\langle\mu_V,\sigma_V,\omega_V,rv\rangle)\rangle$$

$$=\text{Get}(STO)\rightarrow T\langle C,M(GK(C'),KV\langle G(C(B,E,f(F(E)))\,||\,O(E,K'),GK(C_1')),L(GK(C_2'),GK(C'),S(GK(C_3')),GR(GK(C_4'),GK(C''))\rangle)\rangle。$$

模型的形式化描述:

目标为实现将 $T\langle C,K\rangle$ 转化为 $W\langle C,U\rangle$,$\text{Get}(STO)\rightarrow T$ 为模板提供必要的计算逻辑,基本过程为分别将 $T\langle C,K\rangle$ 中的四元组 $K\langle\mu,\sigma,\omega,r\rangle$ 转化为最终数据值 $KV\langle\mu_V,\sigma_V,\omega_V,rv\rangle$。

(1) $\mu\rightarrow\mu_V$ 的转化　将 μ 对应的变量公式翻译成算术表达式,即 $C(B,E,f(F(E)))\,||\,O(E,K')$,$GK(C_1')$,该步骤对应 $X=C(B,E,R,K)$ 过程,即根据输入条件,将公式 B 中的所有变量 E 计算出数值 $f(F(E))\,||\,O(E,K')$,根据 B 中 E 的类型 $E.\text{type}=\text{enum}\{DB,AH\}$,当 E 为 DB 类型时 $R=f(F(E))$,E 为 AH 类型时 $R=O(E,K')$,此过程中所需关键字由 $GK(C_1')$ 产生。当公式 B 转化为 X 后,调用 $G(X)$,即实现 $\mu\rightarrow\mu_V$ 过程。

(2) $\sigma\rightarrow\sigma_V$ 的转化　查找关键字 σ 所参照的关键字 K',用 $L(\sigma,K')$ 生成具体数据,即增长下降、升高降低,而 $\sigma=GK(C_2')$,$K'=GK(C')'$。

(3) $\omega\rightarrow\omega_V$ 的转化　判断关键字 ω 的日期类型,$S(\omega)$ 根据运行时输入条件,将特定格式的日期数据赋予 ω_V,即完成 $\omega_V=S(\omega)=S(GK(C_3'))$ 过程。

(4) $r\rightarrow rv$ 的转化　根据 r 参照的 K 查找 r 知识库的类别,根据原因知识搜索策略,$GR(r,K')$ 将搜索出最可能原因,并将原因结果赋予 rv,完成 $rv=GR(r,K')=GR(GK(C_4'),GK(C')')$ 的操作过程。

待完成 $K\langle\mu,\sigma,\omega,r\rangle$ 到 $KV\langle\mu_V,\sigma_V,\omega_V,rv\rangle$ 的全部转化,调用 $M(K,U)$,实现 $T\langle C,K\rangle$ 到 $W\langle C,U\rangle$

的转化,并输出最终文档 W。

模型的执行流程如图 3 所示,整个系统的执行过程如图 4 所示。文档自动生成算法描述如下:

步骤 1 处理完全由数据库(数据仓库)提供基础数据的关键字,执行函数 C, P,将变量公式转化为算术公式 X 和业务公式 A,调用函数 $G(X)$ 计算出公式 X 的结果。

步骤 2 处理需要重用已生成数据值的关键字(执行函数 C, P),调用函数 $G(X)$ 计算出公式 X 的结果。

步骤 3 处理日期型和增长下降型关键字。

步骤 4 调用 $GR(K, K')$ 处理原因型关键字。

步骤 5 待所有关键字处理完毕后,调用搜索替换函数,将模板中所有关键字进行替换(为了简化,图中并没有画出四趟循环的处理过程,而只进行关键字类型的判断)。

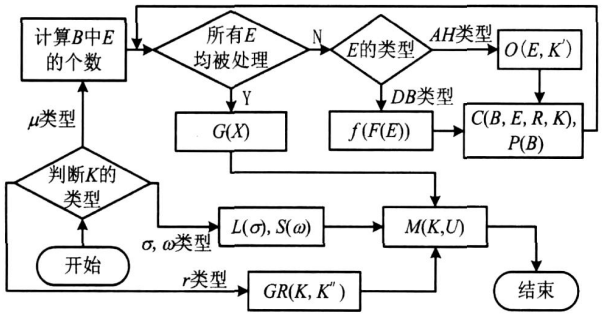


图3 模型的执行流程

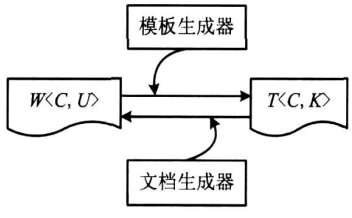


图4 系统执行过程

算法由五趟扫描完成,其中第二趟扫描即重用数据,搜索策略为从当前关键字所在链表节点向前搜索,因为被重用数据的链表节点离当前节点往往较近,如计算同比增长率、同比增长量、增长下降型关键字等,被重用的数据节点与当前节点一般只有 2～5 个间隔;最坏的情况下可能需要扫描到链表初始节点,但这样的情况非常少。因此,在一般情况下重用阶段的时间复杂度为 $O(n)$。假设出现最坏情况的次数为 m,出现的节点位置为 i,则扫描次数大致为 $m \times i$。取 $i = kn$, $m = jn$,其中 k, j 为比例系数,在平均情况下取 $k = 0.5$(即最坏情况出现在 $n/2$ 节点处),则 $m \times i = 0.5jn^2$。j 的取值需要根据文档情况确定,经统计,在本系统中 n 为 1 500 左右,m 不超过 20,即 $j = 20/1500 \approx 0.013$,而 $m \times i \approx 20 \times (n/2) = 10n$,时间复杂度大致为 $O(n)$。在最坏的情况下令 $m = n$, $i = n$, $mi = n^2$,时间复杂度为 $O(n^2)$。其余四趟扫描时间复杂度为 $O(n)$,因此整体的时间复杂度在一般情况下为 $O(n)$,最坏情况下为 $O(n^2)$。

3 模型实现及应用验证

根据模型中定义的各种操作和算法的执行过程,构建文档自动生成系统,将模型定义中涉及到的所有函数封装成具体的功能模块,将功能模块按照算法执行过程进行整合,最终实现模板维护和文档生成两大子系统。

3.1 文档自动生成系统模块设计

文档自动生成系统包括模板维护和文档生成两个主要部分,如图 5 所示。

模板维护阶段主要包括关键字的生成、原因知识库维护、公式与变量编辑器、变量与物理字段映射、公式正确性验证、鼠标位置计算等主要模块。执行过程将 RTF 文档导入到模板编辑器中,对需要变

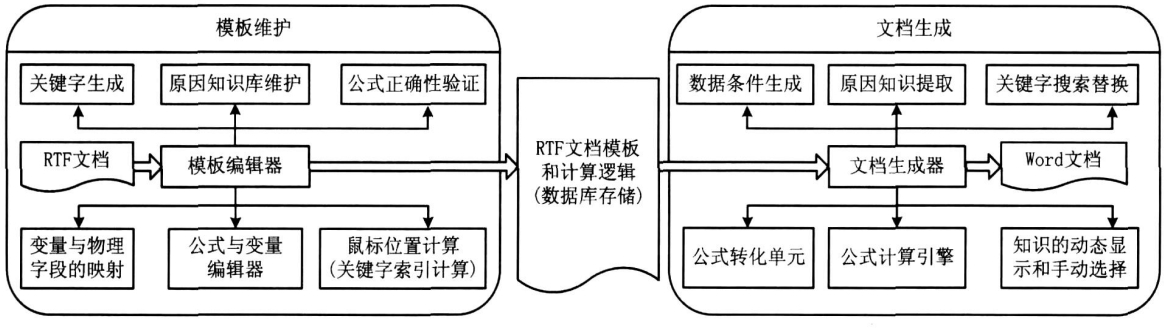


图5 系统体系结构及主要功能模块

化的数据进行计算逻辑编辑,包括关键字、公式、变量、数据单位、参考类型、重用关系、编辑合法性检测、关键字索引存储、变量与字段映射关系等主要操作过程,该过程输出 RTF 格式的文档模板和相应的计算逻辑(数据库存储)。

文档生成阶段主要包括数据的条件生成、公式转化单元、原因知识提取、公式计算引擎、知识的动态显示和手动选择、关键字搜索替换等主要模块。该阶段以 RTF 格式的文档模板和相应的计算逻辑作为系统输入,主要执行过程为:构造数据链表;处理由数据库直接提供基础数据的关键字,对公式中的变量进行运行时条件封装,并生成真实数据,将变量公式转化为算术公式和业务公式,用公式计算引擎计算出算术表达式的数值,更新链表数据;处理需要重用数据值的关键字,搜索已有数值,进行公式转化,生成最终数据(包括"增长下降型");最后处理知识型关键字,按照前文提到的搜索策略和搜索方法,进行最大可能性原因提取,并提供操作员手动校验功能;之后调用搜索替换模块,将所有关键字替换成最终数值,最后输出特定日期数据汇总文档。

3.2　数据存储结构

图 3 中 $M(K,U)$ 执行的前提条件为:模板中所有的 K 全部处理完成。这就要求建立一个存储链表,链表节点中应包含的信息为 K,U,A,X,K',Y,T,I。其中:K 为指标的关键字;U 为关键字的值;A 为业务公式;X 为算术公式;K 为被重用或被参考的指标的关键字(没有参考项时该项为空),如果 K 的类型为 r(原因型),则 Y 中存储影响该指标变化的最可能因素,如果 r 为其他类型,则该项为空;T 代表 K 的类型(μ,σ,ω,r);I 为指标在模板中的位置顺序索引(作用在 3.2 节中说明)。在文档的数学模型中并没有包含 A,X,这两项在运行时并不显示在文档中。

链表构造过程为:首先根据模板中每一个 K 的计算逻辑和约束条件,进行链表节点的初始化,之后扫描链表并根据 T 判断 K 的类型,由 K 的类型来确定指标的计算顺序——先计算 μ 中不需要重用已有数据的指标和 ω,然后计算 μ 中需要重用已有数据的指标,最后计算 σ,ω。待链表处理结束,循环调用 $M(K,U)$。

3.3　公式的展现策略

定义 1 中的函数 P 可以将变量公式转化为业务公式,在系统运行时可以生成某个业务指标的算术公式。经过上述推导已经生成并在结构链表中存储了这两个公式。操作者点击文档中某个业务指标时,可以显示该指标相应的业务公式和算术公式,从而使系统更加智能化和人性化。问题是:如何才能判断鼠标是否点到特定的业务指标,及如何将这个业务指标与公式进行关联。实现策略如下:

步骤 1　将模板中的 K 在模板中的索引存入数据库。

步骤 2　在每个链表节点中存储指标的位置索引。

步骤 3　生成文档时在业务指标值两边加上特殊符号,如"⊦全口径售电量⊣"。

步骤 4　将特殊符号"⊦,⊣"的字号设置为最小,并将其颜色设置为白色(不可见)。

步骤 5　判断鼠标是否点入两个符号之间,如果是,计算在这个指标之前有多少个"⊦"符号,从而确定指标的索引。

步骤 6　根据索引搜索链表中该指标节点,根据 T 判断指标类型

步骤 1　如果 T 类型为 μ,则读取指标对应的公式。

步骤 8　弹出对话框显示业务公式和算术公式。

实际效果如图 6 所示,图中黑框内的指标是鼠标点下的位置。

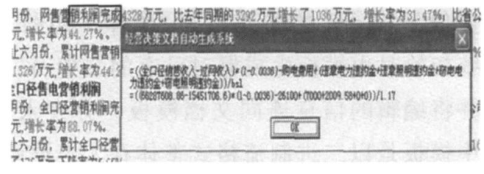


图6　公式显示

3.4　知识的手动选择

对于函数 $GR(K,K')$ 生成的影响因素可能存在偏差,则在文档生成时,操作员应可以手动选择,或录入最新的影响因素,在选择时仍需要处理指标位置索引及关键字查找等问题。实现算法与公式展现算法基本相同,只是在上面算法的步骤 7 判断 T 是否为 r 类型,如果是,则将该指标的影响因素按照发生概率大小全部显示给操作者,以供手动选择。实现效果如图 7 所示,在操作员点到影响因素时(图中"大丰造纸 1 - 2 月份……"),按发生概率大小,以列表形式显示知识库中该类指标的所有影响因素,待选取具体影响因素后替换当前影响因素。

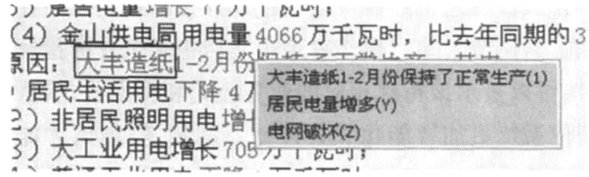


图7 知识选择

3.5 应用验证

定义一个符合文档审核过程的工作流程,下面以电力制造企业的经营数据分析月报审核流程为例,流程中涉及到五个角色(如图 8)。其中业务员是文档的操作者,使用文档自动生成系统生成文档,并将生成的文档以二进制流格式整体存入数据库。在处长、主任、经济师、主管局长各审核节点,除正常的工作流操作外还具有查看月报功能,查看功能负责将二进制流文件显示成正常的文档(如图 9a),并且在各审核节点均可以将文档以 Word 格式输出。

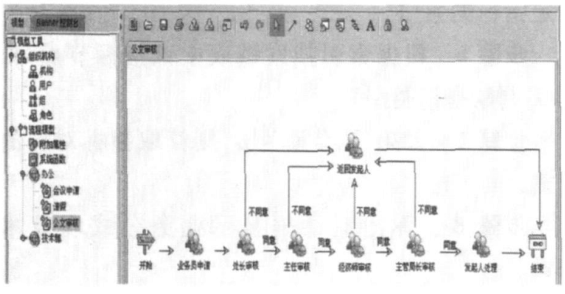
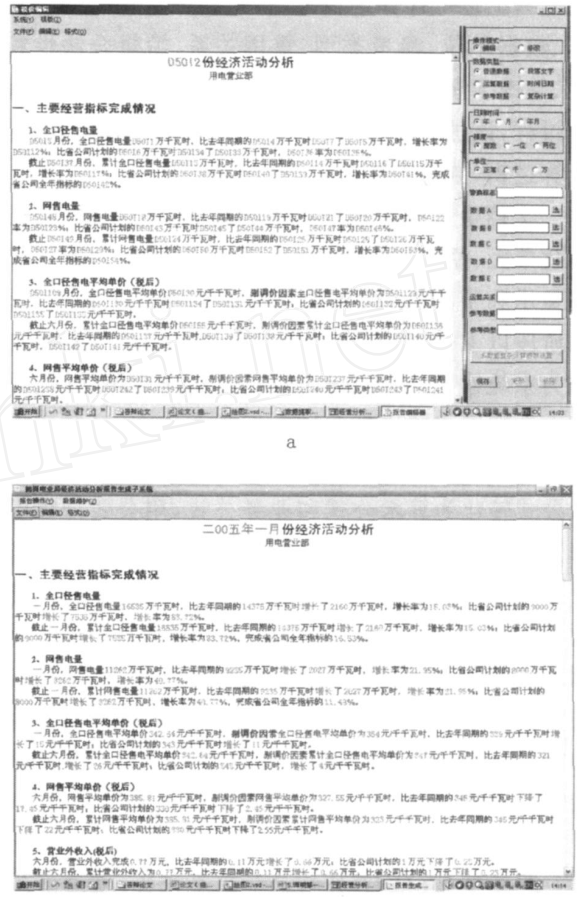


图8 工作流程定义

文档自动生成系统包含文档、模板两个 RTF(一种类似于 Word 的文档格式)实体。如图 4 所示,系统初始时由模板生成器对已有的文档 *W* 进行编辑,并将编辑的信息连同文档模板一并存入数据库,其中模板是以二进制流格式整体存入数据库,这样可以保证文档格式不会丢失,模板的编辑只在系统初始化时进行,如果文档内容或格式需要改动,可以通过模板编辑器对模板进行改动。在使用系统时由文档生成器根据输入的条件,从数据库(或数据仓库)中提取数据填充模板生成特定日期的数据汇总文档。

图 9a 为编辑好的文档模板(只显示出一部分),图 9b 为根据模板生成的 2005 年 1 月的经营决策文档,从图中可以看出,所有需要随生成日期变化的数据全部自动生成,并可实现原因知识的手动选择和公式的动态显示,如图 6 和图 7 所示。与原来该月由专工手动计算编辑的文档进行逐个数据比较发现,对于个别不相等的数据存在人工计算错误,并且



a b

图9 文档自动生成系统运行效果图

原因知识分析能够达到 80 %左右正确率,其余指标影响因素可以进行手动选择矫正,省去重新录入的工作量。专工确认无误后进入工作流的流转审核过程,在每个流转节点查看文档时其效果与图 9a 相同(对于权限的限制不在本文讨论范围之内)。使用该系统后,可以将 15 页涵盖企业经营中各类业务指标的数据汇总文档(Word 格式),由原来四五个人两天的工作量降低到 5 min 内完成,包括校正原因知识时间(刨除工作流审核过程,文档生成时间视机器配置一般不超过 10 s),极大地提高了办公自动化程度、工作效率和数据准确性,大大削减了人力资源的开销。

图 10 所示为编辑模板时变量公式的编辑,当变量超过五个时,可以使用另一个界面单独处理。根据需求该系统可以实现一个指标最多 24 个数据的同时运算(变量以英文 24 个字母命名),当然如果需要更多的基础数据进行公式计算,可以改动程序添加更多的选择框。图 10 中的三个黑色矩形框分别显示了变量与业务指标名的关联方法及公式的编辑

方法。

图10　公式编辑

4　结束语

本文针对企业在编辑数据汇总文档时存在的问题,提出了一个文档自动生成模型及其文档自动生成算法,结合某电力制造企业的经营决策系统,设计并实现了该模型。应用验证表明,该文档自动生成系统能够满足企业对文档化数据分析方式的需求,有效地与企业的工作流集成,提高了工作效率和数据的准确性。

参考文献:

[1] LI Weiping. Research on implementation technology of workflow management system[J]. Computer Integrated Manufacturing Systems, 2002, 8(3):1-4(in Chinese). [李伟平. 工作流管理系统实现技术研究[J]. 计算机集成制造系统, 2002, 8(3):1-4.]

[2] GUAN Tingzhao, WANG Qianping. Research on integration of project drawing design systems based on workflow[J]. Computer Integrated Manufacturing Systems, 2006, 12(5):2-4(in Chinese). [管廷昭, 王潜平. 基于工作流的工程图纸设计系统的集成研究[J]. 计算机集成制造系统, 2006, 12(5):2-4.]

[3] ZHENG Hongzhen, LIU Yang, ZHAN Dechen. Multiple nearest neighbor algorithm based on data mining[J]. Computer Engineering, 2007, 33(3):1-2(in Chinese). [郑宏珍, 刘　扬, 战德臣. 基于数据挖掘的组合近邻模型算法[J]. 计算机工程, 2007, 33(3):1-2.]

[4] PAN Ding, SHEN Junyi. Similarity discovery techniques in temporal data mining[J]. Journal of Software, 2007, 18(2):1-5(in Chinese). [潘　定, 沈钧毅. 时态数据挖掘的相似性发现技术[J]. 软件学报, 2007, 18(2):1-5.]

[5] LEE K S, LEE K W. Frame of an evolutionary design system incorporating design information and history[J]. Computers in Industry, 2001, 44(3):205-227.

[6] ZHUGE Hai. A knowledge grid model and platform from global knowledge sharing[J]. Expert System with Applications, 2002, 22(6): 313-320.

[7] KWAN M M, BALASUBRAMANIAN P. KnowledgeScope: managing knowledge in context[J]. Decision Support Systems, 2003, 35(4): 467-486.

[8] GE Fen, WU Ning. A platform for automatically producing Word document based on multiple techniques[J]. Journal of University of Electronic Science and Technology of China, 2007, 36(2):1-4(in Chinese). [葛　芬, 吴　宁. 基于多种技术的 Word 设计文档自动生成平台[J]. 电子科技大学学报, 2007, 36(2):1-4.]

[9] AHN H J, LEE H J, CHO K. Utilizing knowledge context in virtual collaborative work[J]. Decision Support Systems, 2005, 39(4):563-582.

[10] NISSEN M E. An extended model of knowledge-flow dynamics[J]. Communications of the Association for Information Systems, 2002, 8:251-266.

[11] WANG Zhongjie, ZHAN Dechen, XU Xiaofei, et al. Code generator for enterprise software and applications based on business object association model[J]. Computer Integrated Manufacturing Systems, 2007, 13(5):1-4(in Chinese). [王忠杰, 战德臣, 徐晓飞, 等. 基于对象关联模型的企业应用软件代码生成器[J]. 计算机集成制造系统, 2007, 13(5):1-4.]

[12] SUN Chengzhu, XU Xiaofei, LI Xiangyang. Modeling navigation system supporting virtual enterprise building[J]. Mini-Micro Systems, 2006, 27(2):365-369(in Chinese). [孙成柱, 徐晓飞, 李向阳. 支持虚拟企业组织建立的模型化导航系统[J]. 小型微型计算机系统, 2006, 27(2):365-369.]

[13] ZHANG Xiaogang, LI Mingshu. Workflow-based knowledge flow modeling and control[J]. Journal of Software, 2005, 16(2):1-6(in Chinese). [张晓刚, 李明树. 基于工作流的知识流建模与控制[J]. 软件学报, 2005, 16(2):1-6.]