第 34 卷 第 7 期 计 算 机 工 程 2008 年 4 月
Vol.34 No.7 Computer Engineering April 2008
·软件技术与数据库· 文章编号：1000—3428(2008)07—0070—03 文献标识码：A 中图分类号：TP391

基于网络的 ERP 实施风险评价信息挖掘模型

朱宗乾[1]，姬 浩[1,2]，杨冬民[1]
(1. 西安理工大学工商管理学院，西安 710054；2. 西安工业大学经济管理学院，西安 710032)

摘 要：针对 ERP 实施风险评价信息获取的高困难性问题，利用网络存储的海量信息，在对 ERP 实施风险评价信息的分类基础上，构建基于网络的 ERP 风险评价信息挖掘模型，探讨模型中的关键技术，并给出一个典型用例。
关键词：风险评价信息；Web 信息挖掘；向量空间模型

Model of Mining in Risk Estimation Information of ERP Implementation Based on Web

ZHU Zong-qian[1], JI Hao[1,2], YANG Dong-min[1]
(1. School of Business and Management, Xi'an University of Technology, Xi'an 710054;
2. School of Economics and Management, Xi'an Technological University, Xi'an 710032)

【Abstract】 Aiming at questions about the difficulty of data acquisition of risk estimation information of ERP implementation, the paper builds the model of Web mining in risk estimation information of ERP implementation, on the base of utilizing massive Web information, using methods of Web information mining and analyzing classification of risk estimation information of ERP implementation, discusses the key technologies of the model, and gives a typical example.
【Key words】 risk estimation information; Web information mining; vector space model

1 概述

ERP是涉及企业不同管理层次和不同业务部门的复杂项目，具有高投入、高风险、实施周期长的特点，据统计[1]，一般只有 10%～20%能按期、按预算成功实施，其失败率居高不下的根本原因在于ERP实施过程中存在着各类风险，如何对风险进行有效的评价与控制就成为关键。

目前，国内外许多学者对 ERP 实施过程的风险问题进行了研究，也有学者建立了各种风险评价模型，提出了相关的风险评价指标体系和评价算法。但分析发现大部分研究对于如何有效获取风险评价中所需的具体的风险评价信息探讨得不够，主要表现为：(1)未对支撑风险辨识、风险评价的相关信息的获取渠道进行深入研究；(2)未对风险评价信息的获取方法进行系统研究。而风险评价信息的有效获取是实现 ERP 风险管理的基础。由于 ERP 实施过程的复杂性决定了风险评价的复杂性，进一步又导致了风险评价信息的复杂性、多源性。ERP 实施风险评价信息，有源于企业内部的，也有源于企业外部的；有显性的，也有隐性的；有可量化的，也有非量化的；其获取的渠道也呈现多样化，不同的获取渠道其获取的方法也不尽相同。随着互联网的快速发展，网络信息以几何级数增长，已使其成为海量信息最重要的集中地。它是一个巨大的、广泛分布的、高度异构的、半结构化的、超文本、超媒体的、相互联系且不断进化的信息仓库。对企业 ERP 实施风险评价而言，面对网络上海量的、复杂的各类信息和数据，如何准确地识别、挖掘和获取相关的风险评价信息变得至关重要。

2 ERP 实施风险评价信息分析

ERP 实施风险评价信息是指用于描述、评价和衡量 ERP 实施风险的相关数据。从其信息的来源可以分为外部评价信息和内部评价信息。相对风险的外部评价信息，内部评价信息主要是源于企业内部，故较易获取。而风险的外部评价信息由于其多源性、不确定性等特点，获取的难度要大得多。

风险的外部评价信息分为 2 大部分：社会环境风险评价信息和各参与方协作风险评价信息。

社会环境风险主要表现为 2 个方面：环境变化给 ERP 实施所带来的风险和环境因素对 ERP 实施的不利约束所带来的风险。这 2 类风险评价信息可以通过政治风险、经济风险、社会风险和法律法规风险等评价信息等来全面反映。

各参与方选择与协作风险，包括各参与方选择风险和各参与方协作风险。对于这 2 类风险可以通过软件供应商、实施咨询商和第三方监理方的选择风险评价信息及协作风险评价信息等来反映。这些评价信息涉及到许多方面。如对于软件供应商选择风险评价信息可以通过软件的完善性、集成性、稳定性、软件成熟度、可扩展性、价格水平、技术支持能力、资产规模、财务状况、成功案例数、信誉度等来反映[2]。

对于上述风险的评价信息，常常分散在成千上万的网页之中，使风险评价信息具有多源性、不确定性、分散性、模糊性等特征，导致了 ERP 实施风险评价信息获取比较困难，也致使对 ERP 实施风险的评价往往缺乏必要的信息支持，所以本文将重点研究 ERP 实施过程中风险的评价信息的 Web 挖掘方法和模型。

基金项目：陕西省教育厅基金资助项目(071C087)；西安市软科学研究基金资助项目(YF07201-07)；西安理工大学特色计划基金资助项目(107-210612)
作者简介：朱宗乾(1962－)，男，副教授、硕士，主研方向：ERP 应用，数据挖掘；姬 浩，讲师、硕士；杨冬民，副教授、博士
收稿日期：2007-11-02 **E-mail：**wwwjihso_78@163.com

3 ERP 实施风险评价信息的 Web 挖掘模型

模型应用 Web 信息挖掘的思想、技术和算法，来实现对 ERP 实施风险评价信息的获取，为建立 ERP 项目风险评估提供必要的信息支撑。

3.1 ERP 实施风险评价信息的 Web 挖掘模型建立过程

模型建立首先是以ERP项目实施的风险评价和风险管理需求为出发点，通过对 ERP 风险评价信息的类型进行划分，确定并细化不同实施风险评价信息的特征、特点；其次，应用 Web 信息挖掘的思想、方法和工具，结合 ERP 实施风险评价信息的特性，确定 ERP 风险评价信息挖掘的策略，进而确定ERP风险评价信息挖掘模型的基本构造和采用的关键技术，具体过程如图 1 所示。

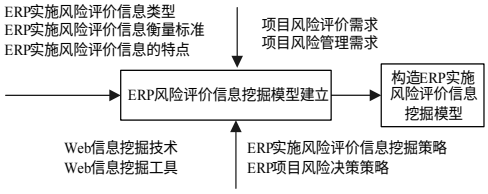

图 1 风险评价信息网络挖掘模型建立过程

3.2 模型的构成和基本流程

3.2.1 模型构成

模型主要由 3 部分构成：

(1)基于 WWW 的 ERP 实施风险评价信息搜索模块，主要功能是通过应用元搜索引擎来实现对风险评价信息的初步搜索。

(2)ERP 风险评价信息分词处理模块，主要功能是对初步搜索的网页格式的风险信息和ERP实施风险评价信息目标样本进行格式化处理，并建立 ERP 实施风险评价信息索引数据库，为下一步挖掘做准备。

(3)ERP 实施风险评价信息挖掘模块，通过数据挖掘的算法实现对 ERP 实施风险评价信息的挖掘，如图 2 所示。

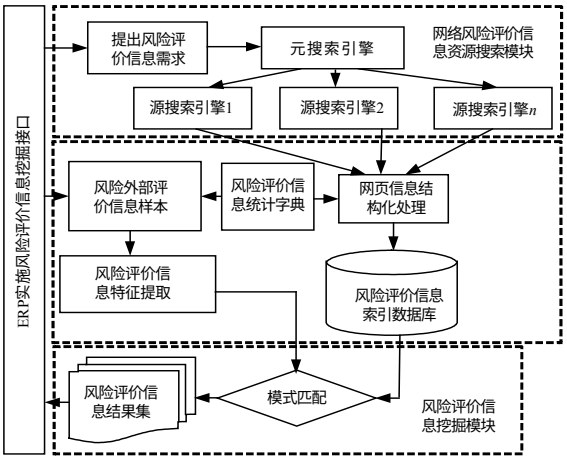

图 2 基于 Web 的 ERP 风险评价信息挖掘模型结构

3.2.2 模型基本流程

模型基本流程如下：

(1)建立风险信息统计词典。建立用于 ERP 风险评价信息特征提取和词频统计的统计词典，包括主词典、同义词词典和近义词词典等，作为特征信息提取的依据。

(2)提出 ERP 实施风险评价信息需求，完成信息检索。应用元搜索引擎技术，通过 WWW 初步获取 ERP 实施风险评价信息。元搜索引擎的优势在于，它通过对多个独立搜索引擎的整合、调用、控制和优化来实现 ERP 实施风险信息相关网页的检索。

(3)建立 ERP 实施风险评价信息索引数据库。对元搜索引擎检索所得信息进行结构化处理，参照风险评价信息统计词典，通过词频分布计算，从中提取出风险评价信息的特征向量并计算出相应权值，进而建立 ERP 实施风险评价信息索引数据库，作为挖掘对象。

(4)提取 ERP 实施风险评价信息目标样本特征。通过统计词典，提取 ERP 实施风险评价信息目标样本的特征向量并计算出权值。

(5)实现 ERP 实施风险评价信息挖掘。提取风险评价信息索引数据库中信息的特征向量，并与目标风险评价信息样本的特征向量进行模式匹配，将符合阈值条件的结果集提交给用户处理，为用户实施 ERP 实施风险评价提供支持。

3.3 模型的关键技术

模型建构过程中，采用的关键技术主要有：风险评价信息的目标样本的特征提取，分词处理，模式匹配计算。

3.3.1 风险评价信息文本的特征信息提取和权重计算

对ERP实施风险评价信息文本的特征信息的提取采用向量空间模型(Vector Space Model, VSM)来实现，该模型是由Salton教授等人[3]在 20 世纪 50 年代提出并发展起来的。其基本思想是将ERP实施风险评价信息文本看作由相互独立的词条组$\boldsymbol{D}(t_1,t_2,\cdots,t_n)$构成，其中$n$为特征词的总个数，对于每一词条$t_i$，都根据其在文本中的重要程度赋予一定的权值$w_i$。将$\boldsymbol{D}(t_1,t_2,\cdots,t_n)$看成一个$n$维坐标系中的坐标轴，$(w_1,w_2,\cdots,w_n)$为对应的坐标值，从而转化为一个向量空间，如图 3 所示。

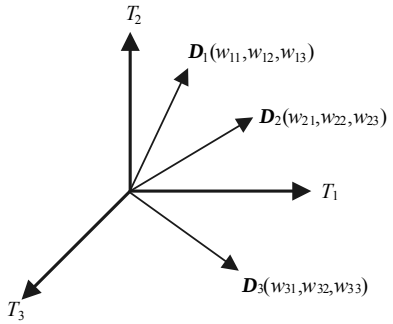

图 3 风险评价信息文本的向量空间模型

词条t_i在文本d_i中的权值W_i通常由 2 部分计算获得：一部分是词条t_i在文本d_j中出现的次数，即tf_{ij}，另一部分是整个文本集合中包含词条t_i的文本个数，即df_j。对于一个给定的检索单元，其权值是tf_{ij}与df_j的乘积。这样有

$$W_{ij}=tf_{ij}\times idf_j=tf_{ij}\times \mathrm{lb}(N/df_j) \tag{1}$$

由式(1)可知，tf_{ij}越大，W_{ij}值越大；同样df_j越小，W_{ij}值也越大，说明该特征项t_i更能够代表文本d_j的内容。由于文本集中的文档长度常常不相同，有时文档长度差别很大，因此很有必要使用归一化技术[4]。归一化技术就是将文档集中的所有文档长度都归到一个单位球的球面上，从而克服了文本长度不一甚至差别很大的缺点。采用归一化技术可以大大提高文本信息权重计算的准确度，归一化后的权重计算如式(2)所示：

$$W_i=tf_{ij}\times idf_i=\frac{tf_{ij}\times \mathrm{lb}(N/n_i+0.01)}{\sqrt{\sum_{t\in d}[tf_{ij}\times \mathrm{lb}(N/n_i+0.01)]}} \tag{2}$$

式中，N代表评价信息索引数据库中文本的数量；n_i为索引库中出现词t_i的文本数量。要特别注意的是：与普通的文本文档相比，Web文档中有明显的标识符，文档结构信息更加清晰而且对象的属性丰富。在计算特征词条权重时，应该充分考虑Web文档的特点，对于标题和特征信息较多的文本可以赋予较高权重。为了提高计算效率，在设计实现时对特征向量进行降维处理，仅保留权重较高的词条作为文档的特征项，从而构成维数比较少的目标特征向量。

3.3.2　风险评价信息分词处理

简单而有效的词条切分方法是直接使用大型通用词库的机器分词法，但通用词库中包含了大量非本专业特征项的常用词汇，为了提高处理速度，可根据 ERP 实施风险评价信息挖掘目标建立相关的分词表，保证特征提取准确性的前提下显著提高机器的运行效率。词条切分时，先用标点符号进行粗略切分，然后分别使用正向和逆向最大匹配算法进行细切分。词频统计时，考虑到自然语言的多样性，可建立相应的同义词、近义词等辅助词典，以扩大 ERP 实施风险评价信息匹配的范围。例如，对于 ERP 软件供应商选择风险评价信息而言，可根据对评价指标的重要性，建立与之相应的风险评价信息分词词库：{软件供应商名称，软件成熟度等级，成功案例，信誉度，软件功能，技术支持，价格水平等}。另外对具体的软件供应商可建立同义词、近义词词典，比如：对于软件供应商名称=“A 公司”与软件供应商名称=“A 企业”，可建立同义词词典，认为是等效名称，避免漏掉重要的文本信息。

3.3.3　风险评价信息模式匹配计算

为了实现ERP实施风险评价信息最大化挖掘，就要通过对其目标文本的特征信息与索引数据库中的信息进行匹配程度的运算，以去除大量的无关信息，保留相关性高的信息，可通过模式匹配的运算来实现。本文通过计算ERP实施风险评价信息目标文本向量与索引数据库风险评价信息向量的相似度$Sim(\boldsymbol{D}_i, \boldsymbol{Q}_i)$来实现模式匹配，其相似度的计算公式为

$$Sim(\boldsymbol{D}_i, \boldsymbol{Q}) = \cos\theta = \frac{\sum_{k=1}^{n} w_{ik} \times q_k}{\sqrt{(\sum_{k=1}^{n} w_{ik}^{\ 2})(\sum_{k=1}^{n} q_k^{\ 2})}} \tag{3}$$

对于相似度的算法通常采用余弦法，首先设ERP实施风险评价信息索引数据库信息空间的维数为n，风险评价信息向量表示为$\boldsymbol{D}_i = (W_{1i}, W_{2i}, \cdots, W_{ni})$，ERP实施风险评价信息目标文本向量$\boldsymbol{Q}_j = (W_{1j}, W_{2j}, \cdots, W_{nj})$，相似度计算的概念模型如图 4 所示。夹角$\theta$的大小反映了相似度的大小。夹角越小，说明相似性越大，当余弦值接近于 1 时，向量之间的相似度最大，其夹角接近于 0；反之当余弦值接近于 0 时，则相关度最小。

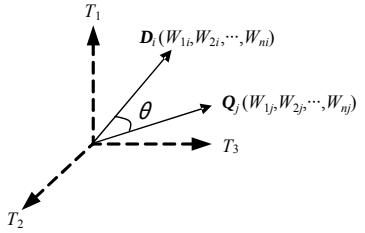

图 4　相似度计算模型

在相似度计算之前，要先设立阈值作为取舍和排序的依据。阈值的设置，通常采用预定初始值，以此为依据进行风险评价信息的挖掘，但这种方法的效率比较低，所以本文采用了“平均值”法[5]，基本思想是：对于每个ERP实施风险评价信息目标文本向量$\boldsymbol{Q}_j$，计算在索引数据库中的评价信息向量$\boldsymbol{D}_i$与$\boldsymbol{Q}_j$相似度的算术平均值：

$$k_0 = \sum_{i=1}^{n} Sim(\boldsymbol{D}_i, \boldsymbol{Q}_j)/n$$

用户还可以根据实际情况对阈值进行微调，公式为$k'=k\pm\Delta k$，其中，k为当前阈值；$\Delta k = \sum_{i=1}^{n} |k_0 - k_i|/n$；$k_i=Sim(\boldsymbol{D}_i, \boldsymbol{Q}_j)$。确定好阈值后，就可以计算文本的相似度，当相似度大于等于预设的阈值时，可以认定文本包含的信息符合ERP实施风险评价所需要的信息，小于阈值则舍弃。最后对符合要求的文本信息进行排序并输出，提供给用户进行ERP项目实施风险的评价。

4　ERP 实施风险评价信息挖掘模型的应用

下面以ERP软件供应商选择风险评价信息挖掘为例说明ERP 风险评价信息挖掘模型应用：

(1)提出 ERP 软件供应商风险评价信息需求，根据需求通过元搜索引擎技术在 Internet 上检索到 4 篇(即 N=4)与之相关的文档。

(2)建立分词词典(ERP 软件供应商，CMM 等级，软件功能，软件价格，资产规模，成功案例，信誉度等级，技术支持)。

(3)计算 ERP 实施风险评价信息的挖掘样本 $\boldsymbol{Q}$ 特征项权值，其风险信息词条为“ERP 软件供应商，软件功能，软件价格”，令向量空间维数 k=4，根据前面算法，统计计算则其特征项权值为 $\boldsymbol{Q}$={4, 2, 2, 0}。

(4)处理风险评价信息索引数据库中所包含的 4 个待挖掘的潜在文本，经过与分词词典比较统计，向量空间的特征信息词条提取如表 1 所示。

表 1　ERP 实施风险评价信息特征提取统计表

特征信息词条	W_i
文档 1	(2.8, 1.4, 2.1, 0)
文档 2	(3.5, 1.1, 4.2, 0)
文档 3	(0, 0.9, 1.4, 0)
文档 4	(0, 0, 0, 0)

对上述ERP实施风险评价信息文本向量的特征词条的df_j统计和权值计算如表 2、表 3 所示。

表 2　df_j统计

索引文档	特征信息词条组合	tf_{ij}
文档 1	{ERP 软件供应商，软件功能，软件价格}	$\boldsymbol{D}_1$={4, 5, 3, 0}
文档 2	{ERP 软件供应商，软件功能, CMM 等级}	$\boldsymbol{D}_2$={5, 4, 3, 0}
文档 3	{软件供应商，软件功能，信誉度等级}	$\boldsymbol{D}_3$={0, 3, 1, 0}
文档 4	{硬件供应商，硬件价格，硬件功能}	$\boldsymbol{D}_4$={0, 0, 0, 0}

表 3　ERP 实施风险评价信息文本权值计算

特征信息词条	df_j
ERP 软件供应商	2
软件供应商	1
硬件供应商	1
软件功能	3
软件价格	1
CMM 等级	1
信誉度等级	1
硬件价格	1
硬件功能	1

(5)计算相似度和确定阈值：

$Sim(\boldsymbol{D}_1, \boldsymbol{Q})$=0.98, $Sim(\boldsymbol{D}_2, \boldsymbol{Q})$=0.92, $Sim(\boldsymbol{D}_3, \boldsymbol{Q})$=0.56, $Sim(\boldsymbol{D}_4, \boldsymbol{Q})$=0

（下转第 75 页）