# 企业核心竞争力的 Web 挖掘研究①

## The study of the enterprise key competitiveness's Web ming

邵良杉　付曙光　薛立军　（辽宁工程技术大学电子与信息工程系　阜新 123000）

**摘要**：介绍了虚拟企业、企业核心竞争力的概念，分析了企业核心竞争力信息在虚拟企业合作伙伴选择过程中的重要性，提出了一种企业核心竞争力信息的 Web 挖掘模型，并给出了企业核心竞争力信息 Web 特征提取的算法，最后，取得了良好的结果。

**关键词**：虚拟企业　企业核心竞争力　合作伙伴　Web 挖掘　特征提取

## 1　引言

虚拟企业，又称动态联盟，是指当市场出现新的机遇时，具有不同资源与优势的企业为了共同开拓市场，共同对付其他的竞争者而组织起来的暂时性联盟体。这个联盟体以信息技术和网络技术为基础，在合作过程中各成员互不干涉、共担风险、共享资源、公担费用、共享利益、联合开发、互惠互利。当预期目标达成后，此组织即告解体。

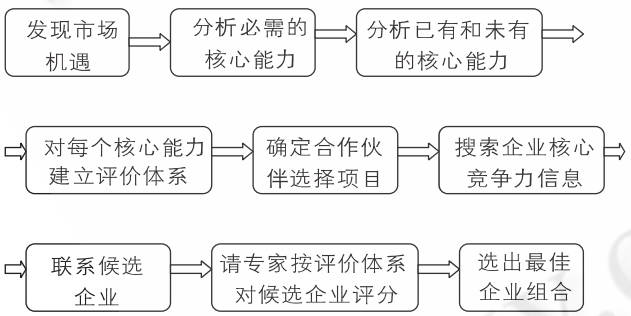


图 1　虚拟企业合作伙伴的选择过程

在虚拟企业建立的过程中，合作伙伴的选择至关重要。虚拟企业成功与否很大层度上取决于合作伙伴的正确选择，只有匹配的合作伙伴，才能进行富有成效的动态资源重组和最佳配置。虚拟企业需要发起企业认真地分析合作目标，并为整体任务和各个子任务选定合适的成员企业。一般地，虚拟企业合作伙伴的选择过程如图 1 所示。

分析该图我们发现，发起企业所掌握的企业核心竞争力信息的多寡对合作伙伴的选择至关重要。如果发起企业掌握较少的企业核心竞争力信息，即使发起企业发现绝佳的市场机遇，由于可选择的候选企业很少，因而选择出来合作伙伴的竞争力不是很强，所建立的虚拟企业整体的竞争力也就不是很强；另一方面，如果发起企业掌握了大量的企业核心竞争力信息，那么，当某一市场机遇来临时，该企业就可以根据核心能力筛选选出大量的候选企业。可以预见，在此基础上选择出来的合作企业的竞争力必然很强劲，那么，所建立的虚拟企业整体竞争力也就很强劲。

那么，企业核心竞争力的具体含义是什么？包含哪些方面？如何搜索到大量的企业核心竞争力信息？

1990 年西方战略学家普拉哈拉德（C. K. Prahalad）和哈迈尔（Gary Hamel）在《哈佛商业评论》首次正式论述了企业核心竞争力理论。他们认为核心竞争力是："组织中的积累性学识，特别是如何协调不同的生产技能和有机结合多种技术流派的学识。"其中特别强调"学识、协调和结合"。并以学识——知识的拥有程度或能力为前提，以此获得竞争势的核心能力。

一般的企业核心竞争力指企业在所从事的行业中自身所拥有的占优势地位的资源和能力。主要包括：

（1）企业所从事的行业；

（2）与其他竞争者相比占优势地位的资源与能力。

① 本论文为教育部博士点基金项目，基金号：20041047006

以前,发起企业只能掌握以往有过合作经验的,或在同一个地域范围内的企业的核心竞争力信息。随着互联网的迅猛发展,越来越多的企业在互联网上建立了自己的商务站点或企业主页。它们或是在互联网上开展电子商务活动,或是将企业的竞争力信息发布到网上,或是在互联网上举行活动。所有这些,都为我们挖掘企业核心竞争力信息提供了一个前所未有的机遇。互联网上的企业信息数量庞大,种类繁多,而且更新速度非常快。如果我们能找到一种对互联网上企业信息进行挖掘的方法,进而抽取出包含企业核心竞争力的信息,那么,我们就拥有了一个无比巨大而且随时更新的企业核心竞争力信息库。这样一个信息库无疑对虚拟企业合作伙伴的选择将起到巨大的推动作用。

## 2　Web 挖掘模型的建立

分析一个企业的核心竞争力信息 Web 挖掘的一般过程,主要包括文档采集、信息抽取、用户交互这三个主要步骤。其中,信息抽取是 Web 挖掘的中心环节,它又可以细分为分词处理、特征提取、企业核心竞争力信息结构化这三个子步骤。

因此,建立企业核心竞争力的 Web 挖掘模型如图 2 所示。

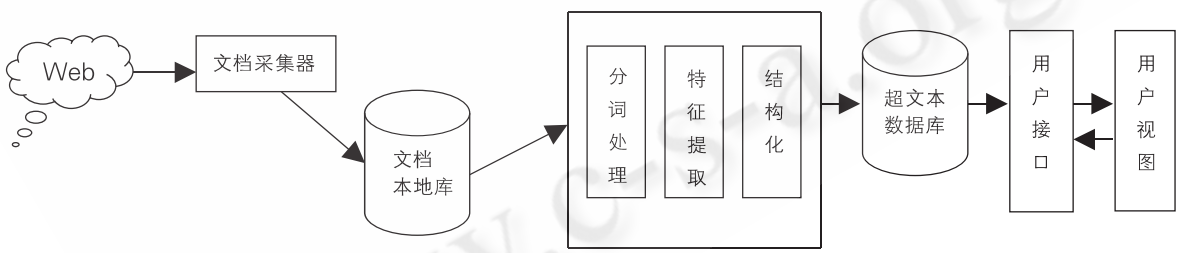


图 2　企业核心竞争力的 Web 挖掘模型

下面对模型的各部分加以说明:

首先需要一个企业 Web 文档采集器将 Web 上大量的、分散的企业信息采集到本地计算机上存储,以便于后继的相关处理。

其次,企业 Web 文档本身是半结构化甚至是无结构的,且缺乏机器可理解的语义。要理解其语义,要对企业的 Web 文档进行信息抽取,就必须首先对其进行分词处理。分词处理是进行信息抽取的第一步。

特征提取是企业核心竞争力信息 Web 抽取的关键步骤。对 Web 文本中出现的词条,及其权值的选取称为特征提取。企业 Web 文档特征提取的重点是对文档中出现的构成企业核心竞争力信息的词条进行特征提取,目标是实现提取过程的自动化。

企业核心竞争力信息结构化是特征提取的后继步骤。其目的就是要实现将企业的超文本文件转换成数据库中结构化数据的目标。

最后,为了便于用户的使用,需要设计一个用户使用系统的接口。

由于 Web 挖掘模型所涉及的技术太多,本文仅对 Web 挖掘模型的中心环节,特征提取的关键技术加以论述,其他技术请参阅 Web 文本挖掘的相关论文。

## 3　企业核心竞争力的特征提取

文本特征指的是关于文本的元数据,分为一般文本特征和专有文本特征。一般文本特征指文本中出现的所有词汇;专有文本特征指文本中对某些应用有特殊意义的词汇,如企业核心竞争力特征项等等。对一般文本特征的描述,目前常用的特征表示法是向量空间模型,即用向量 $(t1,d1;\cdots;ti,di;\cdots;tn,dn)$ 来表示文档。其中 n 是文档中所有词的数目,ti 为词条项,di 指第 i 词的权。对 Web 文本中出现的词条 ti 及其权值 di 的选取称为一般文本特征项提取。其目的是从文本中抽取出来一些能代表文本内容的词条,通过分析这些特征词,达到分析 WEB 文本内容的目的。

在一般文本特征项的提取中,di 的计算方法有很多种,常用的有 TF,TF * IDF 和布尔方法。这些方法没有考虑和利用 HTML 文档中的格式信息。与普通文档不同,由于不同 HTML 的逻辑结构由超文本标签表达,这些标签清楚地标明了哪些文字属于标题、哪些文字属于正文等。不同标签中出现的检索字,其表达文档

内容的能力是有差别的。若两个文档 d1 和 d2 都包含检索字 t，且 t 在 dl 和 d2 中出现的次数均为一次，但是，在 d1 中 t 出现在文档标题中；而在 d2 中 t 出现在正文中，运用传统的信息检索系统，会认为检索字 t 表达两个文档的能力相同。但是统计数据表明，标题比正文更具有对文章内容的概括性，所以出现在标题中的 t 比出现在正文中的 t 更能确切表达文档内容。基于以上考虑，本文采用扩展的 TF * IDF 方法。它的思想是：先将标签根据重要度分类，并将这些标签分组。然后令 di = (TFV? TIFV) * IDF。TFV 由 TF 扩展而来，假设标签分成 n 类，TFV = (tfv1, …, tfvn), tfvg (g = 1, …, n) 代表文档 d 中 ti 在 g 类标签中出现的次数。TIFV = (tifv1, …, tifvn) 为标签重要度因子向量，tifvg 为正整数，它越大，表示 g 类标签越重要。其中，"?" 代表向量积运算。

如前所述，企业核心竞争力信息主要包括：

(1) 企业所从事的行业；

(2) 与其他竞争者相比占优势地位的资源与能力。另外，企业的名称、企业的联系方式等也是发起企业在选择合作伙伴时需要的信息，我们也把它们提取出来。因此，我们要提取的企业核心竞争力特征项包括：企业名称特征项、企业从事行业特征项、企业优势特征项、企业联系方式特征项。

另一方面，这些信息隐藏在企业 Web 文档中的某个或某些词汇的组合中，如何识别出这些信息？如何进行这些专有文本特征的提取？

首先，我们从详细地分析各专有特征项的汉语表达规律入手。

### 3.1 企业核心竞争力的汉语表达规律分析

考察大量的企业 Web 文档，我们发现，企业核心竞争力具有如下特定的表达方式：

(1) 企业的名称。在汉语当中，企业名称的表达方式有很多种，但最基本的表达方式如下：

企业名 = 地名 + {企业名关键字} + {企业名类型} + 企业名后缀

如：中国船舶工业集团公司、上海山地计算机系统有限公司、哈尔滨天宝科技发展有限公司、太原化工股份有限公司等等。其中，大括号的项为可选项。可作为企业名后缀项的常用词有："公司"、"银行"、"集团"、"企业"、"工厂"等，它们是识别一个词表达意义是企业名称的关键。我们把这类词集中到一块，建立一个企业名后缀词库，用以识别企业名称。

可作为企业名类型的常用词有："投资"、"开发"、"有限责任"等。

企业名关键字通常采用名词。其中，有一类企业名关键字包含有企业所属行业的信息，如：中国船舶工业集团公司、上海山地计算机系统有限公司、太原化工股份有限公司等等，有助于识别企业所属的行业。

另一方面，绝大多数的企业 Web 主页以企业的名称作为主页的标题。

因此，我们得出如下判断：如果企业 Web 主页的标题是以企业名后缀词库中的词结尾的，则该标题就是企业的名称。

(2) 企业所从事行业。企业所从事行业在汉语中的表达方式复杂而多样。在企业的名称当中，有一大部分企业的企业名关键字采用该企业所属的行业名来表达，这是最常见的企业所从事行业的描述方式。

为了精确地描述企业所从事行业，我们考察了大量的企业 Web 主页，把出现频率比较高的用以描述企业所从事行业的专有名词集中起来，建立企业行业关键词库，作为提取企业所从事行业特征项的依据。

因此，如果企业名称当中企业名关键字是企业行业关键词库中的词，则把该词提取出来作为企业业务结构的特征项；如果没有，则空白。

(3) 企业的联系方式。企业的联系方式不外乎有：地址、网址、电话、Email、Fax 这么几类。考察大量的企业 Web 文档，我们发现，企业的联系方式具有如下特定的构成形式：

企业的联系方式 = 企业联系方式前缀词 + 企业联系方式关键字

其中，企业联系方式前缀词通常有："地址"、"网址"、"电话"、"Email"、"传真"等等。我们把这类词集中到一块，建立各类企业联系方式前缀词库，作为提取企业联系方式特征项的依据。企业联系方式关键字依各类联系方式的不同而不同。概括起来，该关键字主要有如下这么几项表达方式：

① 企业的地址关键字 = XX 市 + XX 街(路) + XX 号 + XX 楼 + { XX 室(屋)}。其中，"XX" 代表汉语中的一个词，多数情况下为名词，大括号中的项为可选项；

② 企业的网址关键字 = http://{www.}xxxx.{com.}{cn}。其中,大括号中的项为可选项,xxxx 代表一个特定的字符串;

③ 企业的电话关键字 = XXXX-XXXXXXXX。其中,XXXX 代表 3 到 4 位阿拉伯数字;XXXXXXXX 代表 7 到 8 位阿拉伯数字;

④ 企业的 Email 关键字 = xxxx@xxxxx.com.{cn}。其中,大括号中的项为可选项,xxxx、xxxxx 都代表特定的字符串;

⑤ 企业的传真关键字 = XXXX-XXXXXXXX{-XXXX}。其中,大括号中的项为可选项,XXXX 代表 3 到 4 位阿拉伯数字;XXXXXXXX 代表 7 到 8 位阿拉伯数字。

因此,如果企业 Web 主页中出现了各类企业联系方式前缀词库中的词,并且其后面出现的词符合某一类企业联系方式关键字的表达方式,我们就可以认为,这个几个词表达的是企业的联系方式,我们企业联系方式前缀词和企业联系方式关键字一并提取出来,作为企业联系方式的特征项;如果没有,则空白。

(4) 企业资源与能力优势。企业资源与能力优势在汉语中的表达方式复杂而多样。对这类信息的挖掘目前还没有成型的方法。本文采取的办法是将企业 Web 主页正文中出现频率比较高的一般文本特征项作为对企业资源与能力优势的描述信息。

### 3.2 建立企业核心竞争力知识库

通过以上分析我们发现,对企业核心竞争力各专有特征项进行提取时,一方面需要考查某些辅助词库;另一方面,需要按照特定的汉语组词规则进行判断。为此,我们建立企业核心竞争力专有特征项的知识库。该知识库主要包括辅助词库和判断规则库。包括:

企业名后缀词库、企业行业关键词库、企业地址前缀词库、企业网址前缀词库、企业电话前缀词库、企业 Email 前缀词库、企业传真前缀词库、企业行业判断规则库;企业地址判断规则库、企业网址判断规则库、企业电话判断规则库、企业 Email 判断规则库、企业传真判断规则库共计 14 种。

## 4 结论

本文分析了企业核心竞争力信息在虚拟企业合作伙伴选择过程中的重要性,给出了企业核心竞争力 Web 挖掘的模型,并分析了其中的关键技术。最后取得了良好结果。

## 参考文献

1 郭瑞军,基于 Web 的虚拟企业合作伙伴选择,武汉理工大学硕士学位论文,2003,3:12-20。

2 钱小军,Web 文本挖掘技术研究及其实现,浙江大学硕士学位论文,2002,3:18-25。

3 赵忠华等,虚拟企业合作伙伴的寻找与评价,商业研究,2003,8(248):19-21。

4 许建潮等,Web 文本信息抽取与挖掘方法,长春工业大学学报,2002,8(23)增刊:50-53。

5 刘芳等,有效的检索 HTML 文档[J],小型微型计算机系统,2002,21(9):989-988。