国际科技资源监测与服务体系构建[1]

刘云[1]，王小黎[2]，樊威[1]

（1 北京理工大学管理与经济学院，北京 100081；2 中原工学院经济管理学院，郑州 450001）

摘要：本文通过对国际科技资源监测内容、方法、技术等的系统研究，构建可视化的国际科技资源的监测和服务体系，提出“三个一流”的国际科技资源概念，采用科学计量学、数据挖掘等方法，针对不同类型的国际科技信息数据库，研究设计了一套行之有效的国际科技资源监测框架体系，为把握国际科技资源的分布状况、寻找高水平的国际合作伙伴、有效利用国际科技资源提供了一个信息化的支撑手段。

关键词：科技资源；监测；服务；数据挖掘

Construction of the International S&T Resources Monitoring System

Yun LIU[1], Xiao-Li WANG[2], Wei Fan[1]

(1 School of Management & Economics, Beijing Institute of Technology, Beijing, 100081. 2 School of Economics & Management, Zhongyuan University of Technology, Zheng-zhou, 450001)

Abstract: Based on the systematic research in contents, methods and technology of the in-ternational S&T resources, the article established the visual monitoring & service system of inter-national S&T resources, proposed the “three first-classes" concept of international S&T resources. For different types of international science and technology information databases, by use of sci-entometrics, data mining, visual technology and other methods, we have designed and built a set of effective resource monitoring framework for international S&T resources, which provided the information support for the grasp of international distribution of technology resources, searching for a high level of international cooperative partners and more effective use of international S&T resources.

Keywords: International S&T resources; monitoring system; Scientometrics; data mining

1 引言

在我国国际科技合作战略决策和计划项目管理中，决策者、管理者、评审专家经常遇到的难题是无法准确地判断合作伙伴的质量和水平，中方合作者寻找国际合作伙伴也存在很大的随意性，缺乏一个高水平的国际科技合作资源数据中心和动态监测决策支持系统为管理者和执行者提供行动指南，导致我国国际科技合作难以把握战略上的主动权和针对性，不能有效利用全球科技资源提升我国的自主创新能力。随着大型科技文献、专利数据库系统在科学研究中的广泛应用，全

本文得到国家科技部国际科技合作项目（2007DFA10950）和国家自然科学基金重点项目（71033001）资助

作者简介：刘云（1963-），北京理工大学管理与经济学院教授、博士生导师，研究方向：科技评价、技术创新管理、数据挖掘。Email:liuyun@bit.edu.cn

球范围内的科技数据信息量急剧增大。面对庞大的科技信息数据库，需要快捷的从中提取出有用和有效的知识。由此，人们提出采用数据挖掘的方法对海量的科技信息进行处理，挖掘出隐藏在海量信息背后的知识，服务于科技工作。

2 国际科技资源监测的概念及必要性

2.1 国际科技资源监测的概念

科技资源监测是指以科学技术信息、数据分析为基础、以数据挖掘、信息萃取、知识发现、数据可视化技术等信息科学前沿技术为手段，对科学技术活动进行动态监测、分析及评估的方法[1]。

科技资源监测是科学学、科学计量学、科技管理以及信息科学等多学科交叉形成的新方向，是一种定性与定量相结合的方法。它充分综合利用了现代先进的信息技术和各方面专家的战略性智力，对过去、现在和未来的科学、技术、经济和社会发展进行系统研究。其目的是为技术管理及决策提供动态、准确的科学技术发展状态，从而了解科技热点，把握技术机会，降低投资风险，提高效率。

科技资源监测的核心思想是利用信息科学的前沿技术，结合技术预测、技术评估和专家知识等，对技术活动的载体进行分析、挖掘并利用数据可视化技术形象表达所得到的知识，为政府、科研部门和企业的技术管理者和科研工作者提供有效的支持[1]。科技资源监测强调的是从大量数据中发现那些尚未发现的知识，实现的是从科学技术活动中大量原始数据中自动获得重要信息的过程。

2.2 国际科技资源监测的必要性

科学技术发展能有利推动经济的发展和综合国力的提高，促使世界各国争相发展应用科学技术，为有效的制定科技政策，建立健康的科技发展环境，进行科学的技术评价至关重要。由于特定的历史条件和国情及定量分析在数据和方法方面的困难，目前科研管理活动主要以同行评议等专家评估方法为主。即利用评估人(专家)的集体智慧，通过评估人对某技术领域现状的把握和对未来的预测来评估项目的创新性和科学性[2]。但传统的以专家评估为主要手段的科技现状评价方法有以下缺陷：

（1） 受限于个体专家对相关技术领域最新成就了解的广度、深度和及时程度；基于“直觉”和“有限知识范围”的评估、预测，往往缺乏有力度的佐证，缺乏数据和知识支持。

（2）涉及资源分配或利益相关的评议工作，会部分受人为和非公正因素的影响，有随机、不稳定问题。

（3）主要是依靠来自系统外部的专家知识，凭借专家经验对未来较长时期的科学、技术、经济和社会发展进行评估、预测，对外部资源(专家)过分依赖，难以为科研决策和管理提供系统性的客观依据。

这样就导致我国科研管理在科研评估、科研立项决策、科研资源管理、科研项目管理、科研规划等方面存在诸多问题。而且随着大型科技文献、专利数据库系统在科学研究中的广泛应用，全球范围内的科技数据信息量急剧增大。面对庞大的科技信息数据库，需要快捷的从中提取出有用和有效的知识。由此，我们需要采取科技资源监测(Science and Technology Monitoring)的方法结合专家智能来解决这一问题。

3 国际科技资源监测的目标与对象

本文所涉及的国际科技资源监测的对象主要包括五类：世界科学前沿监测、世界技术前沿监测、世界一流研发企业监测、世界著名科学家监测、国际科技合作（本文以中美科技合作为例）监测。

3.1 世界科学研究前沿监测

科学技术的“研究前沿”代表的是某一个研究领域的思想现状，它是一个相对的概念。在科研活动中，前沿是根据研究对象当前在学科领域中所处的地位而定的，研究对象在学科领域中所处的地位领先，就可以被称作前沿。

对研究前沿的监测，有助于科研人员和科研管理人员迅速了解某个领域的研究前沿和研究热点，即使进入国际新兴主流科研问题研究，抢占科技制高点。同时，对研究前沿的监测能为科学家即使提供高水平的国际合作伙伴，为政府机构对重大科研项目及时提供帮助，或者为单位内部分配科研资源提供决策依据，从而推动科学技术和社会经济的发展。

3.2 世界技术的前沿监测

以美国经济研究局(National Bureau of Economic Research, NBER)确定的美国专利技术分类作为领域划分的依据，针对 Chemicals exc.Drugs，Computers & Comm.，Drugs & Medical，Electrical & Electronics.，Mechanical，Others 六大领域内的“一流技术”、“一流人才”、“一流机构”开展动态的监测。

世界技术前沿监测将利用先进的数据挖掘、整理分析等技术和手段，确定“三个一流”的内涵和评价模型，并且对其进行动态的监测，实现统计信息分析、合作授权专利分析、技术热点分析、监测报告分析等功能。用户可根据自身需求对某一特定领域、技术进行动态的跟踪和分析，以期掌握改领域、技术的发展动态、技术前沿和研究热点[3]。

3.3 世界一流研发企业监测

世界一流研发企业决策将英国工贸部公布的全球一流研发企业作为世界一流研发企业决策对象，重点监测以下主要内容：

（1）世界一流研发企业的研发投资强度和趋势、各领域及各国家（地区）研发投资强度分布情况；

（2）通过专利数量及质量监测出什么技术是世界一流研发企业的“一流技术”？谁是世界一流研发企业的“一流人才”？世界一流研发企业与哪些企业或机构进行合作[5]？

（3）利用世界领先的知识产权管理和分析平台 Aureka，根据不同的主题发布世界一流研发企业的监测信息，例如世界一流研发企业的可视化专利地图、高被引专利的引证树等；

（4）定期发布的监测报告。

3.4 世界著名科学家监测

世界著名科学家是指在国际上相关科学领域具有权威性的学者，将世界著名科学家界定为国际权威奖励的获奖者，国际权威科学院的院士、学者，以及国际大科学组织的科学家。

针对网络上的非结构化的数据进行监测，抽取世界著名科学家的相关信息，包括研究成果等，建立著名科学家动态监测数据库，为用户提供世界著名科学家的资料等信息的综合发布，从而为最终用户在国际科技合作中选择合作伙伴提供参考。

世界著名科学家的数据为基于 WEB 进行定向数据挖掘，因为数据库类型是非结构化数据库，数据具有复杂性和不规范性等特征，结合最终用户需求和非结构化数据库数据特点，数据的最终发布形式往往采用新闻传播的形式[4]。

3.5 中美科技合作监测目标

本部分主要针对中美合作的SCI论文情况进行监测，通过对SCI中美科技合作论文检索结果进行数据挖掘，分析近五年来中美科技合作的参与国家、机构、期刊、学科领域和高质量论文的情况及发展趋势[4]。主要监测内容如下：

（1）各年度合作论文数演变。监测各年度中美合作论文的数量及其在近五年的演变情况。

（2）各学科合作论文数演变。监测五年中17个学科中每个学科合作论文数的变化情况

（3）各学科合作机构演变。监测五年中每个学科内合作机构参与论文的情况及其演变过程、中美机构合作网络。

（4）各学科合作科学家演变。监测五年中每个学科内参与合作的科学家及其论文数、中美科学家合作网络。

（5）各学科合作资助计划或机构情况。监测五年中各学科领域主要资助机构或计划及其变化情况。

（6）各学科第三方国家参与情况。监测参与中美科技合作的第三方国家及其在各个学科领域的分布、五年中的演变情况。

（7）各学科合作高被引频次论文。监测各个年度每个学科内的高被引频次论文。

4 国际科技资源监测的手段和方法

4.1 科技资源监测的手段

资源监测基础理论和方法、所需要的监测目标的要求，科技资源监测的一般过程及步骤如下：

（1）确定科技资源监测的目标，划定研究主题的范围。例如：要监测电子信息行业的最新热点及主要研究人员。就要选择电子信息行业数据库，电子信息行业期刊论文，会议文献、学位论文及专利文献等作为研究对象。

（2）搜集数据。在涉及研究主题范围内尽可能的搜集全的文献信息，使研究结果更具有权威性和说服力，把收集到的信息放在一个专门的数据库中。

（3）数据集成、数据预处理，去除噪音词、清洗和转换。从目标数据集中除去明显错误数据和冗余的数据，去除噪声或无关数据，进行数据清洗。并通过各种转换方法将数据转换成有效形式，为今后的数据挖掘做好准备工作[6]。

（4）数据挖掘与分析。根据第一步所确定的可视化数据挖掘的目标，选择特定的数据挖掘算法(如技术组(群)自动识别、技术创新指标、自然语言处理和竞争分析、模糊聚类等)，在数据库中提取数据模式，并用一定的方法表达成某种易于理解的模式(知识)[7]。根据某种兴趣度度量，对发现的模式(知识)进行解释、评估和价值评定，必要时需要返回前面处理中的某些步骤以反复提取[8]。

（5）知识表示，形成报告和一系列图表。通过数据收集、整理、挖掘等步骤而得到的结论结合可视化技术写成报告，以用户容易理解的方式展示出来。

4.2 国际科技资源监测的技术方法

4.2.1 文献计量学

文献计量学是用数学和统计学的方法，定量地分析一切知识载体的交叉科学。它是集数学、统计学、文献学为一体，注重量化的综合性知识体系。其计量对象主要是：文献量(各种出版物，尤以期刊论文和引文居多)、作者数（个人集体或团体)、词汇数(各种文献标识,其中以叙词居多)文献计量学最本质的特征在于其输出务必是“量”[9]。

文献计量学指标包括：关键词词频统计、来源机构统计、来源国际统计等。文献计量学指标的可视化方法本身并不指出哪些主题是研究趋势，而是通过图形展现各个统计指标的在时序上的变化情况，以此来反映客观的统计数据事实，最后的判定则交由用户或专家，此仅作为判定的依据之一[7]。

文献计量学的可视化评价技术和方法按照不同地域和领域的情况，形成了许多不同的评价技术和方法。如引文分析法、共被引分析法、多元统计分析法、词频分析法、社会网络分析法。应用的软件平台主要有：Bibexcel、SPSS、WordSmith Tools、Pajek、Ucinet、CiteSpace II 等。

4.2.2 文本挖掘

科技资源监测是数据挖掘的一种，它以科技信息为挖掘对象，利用信息技术对数据进行深度挖掘，可以对科技热点和发展方向等做出预测和评价。

随着数据量的递增和数据信息的复杂化和多元化，通过数据挖掘算法挖掘出的信息可能不易理解或不一定正确，因而提出了使用可视化的数据挖掘技术，即利用人们容易理解的图形、图表等直观的表现方式来表示复杂的数据信息，或要求用户参与到数据挖掘过程中，通过设置参数控制挖掘进度和质量，从而能够加

深用户对复杂数据信息的理解和保证数据结果信息的正确性[10]。

可视化数据挖掘是知识发现(KDD:Knowledge Discovery in Databases)过程中的一个特定步骤，提供了用户与计算机之间的一个通讯接口，以便帮助用户从数据库或数据仓库中发现未知的、潜在的、有使用价值的信息的方法、理论和技术。可视化数据挖掘可与 KDD 过程中的数据挖掘和模式评估相关，先通过挖掘算法从数据库或数据仓库中挖掘出信息，然后以容易理解的形式通过人机接口显示出来，从而使用户对挖掘结果有更清楚的认识，也可以是用户通过人机接口与数据挖掘过程充分交互，实时观察挖掘出的信息，以便及时纠正错误的数据模式[11]。

文本挖掘是近几年数据挖掘领域的一个分支，在国际上，是一个非常活跃的研究领域。从技术上说，它实际上是数据挖掘和信息检索两门学科的交叉。

传统的数据挖掘技术，主要针对的是结构化的数据，如关系数据库。随着信息技术飞速发展，大量形式各异的复杂数据(如结构化和半结构化数据、超文本数据和多媒体数据)不断涌现。大部分文本数据没有结构，转换为特征数据后特征数将达到几万甚至几十万。如何快速地从来自异构的数据源的大规模的文本信息资源中提取符合需要的简洁、精炼、可理解的知识，就涉及到文本知识挖掘[12]。

目前，主要的文本挖掘工具分为两种，一种是基于 web 的文本挖掘工具，如：Chemical Abstracts Service、Web of Science 的 Results Analysis 功能子系统等；一种是基于桌面的文本挖掘工具，如：Vantage Point、Aureka、TDA(Thomson Data Analyzer)等。

5 国际科技资源监测与服务框架设计

5.1 国际科技信息服务现状

长期以来，包括科技文献、专利数据等科技信息海量存储于各种专业数据库，而最终的信息用户在使用中与这些数据库没有直接的渠道，没有办法获取直接的信息；或是获取了部分原始数据，由于没有科学便捷的信息处理工具，难以形成有用的，可以直接用于辅助决策的数据信息。主要的问题有以下几个方面：

（1）科技信息数据库购买

目前，国内购买了各种科技信息数据库的主要是一些科研院所、大型的研究机构和企业，购买使用权限的用户需要付出大量的信息使用费。而普通的企业用户是承受不起，但是又非常需要这部分的科技信息。

（2）科技信息数据库的使用

科研工作者通过授权的科技信息数据库中获取了需要的科技数据信息后，又出现了新的问题，他们需要学习新的信息处理软件工具，来分析处理目标数据，以形成有用的结果。如社会网络分析软件等尽管有些是开源的，但在获取软件和学习软件上，是需要一定的过程，对部分用户来说，甚至是不可能的事情。

（3）科技信息处理工具的使用

目前，广泛使用的科技信息处理工具，要灵活地使用，必须要掌握一定的理论基础知识和使用方法，这或许对于非专业人员，是个大难题。另一方面，大部分工具软件是单机的形式来处理数据库的信息，尽管可以处理网络数据库。而经过工具软件处理过的科技信息可能需要面向网络用户的需要，或是用户的网络处理需要，这就需要进行一定的转换。对于这种科技信息的转换对于用户来说，有不同程度的难度，也有可能由于专业的限制，无法完成。

也有一些上述没有提到的问题因素，如各种科技信息可视化等的需求得不到解决，使得目前的科技信息没有被大规模和快速的使用，而科技工作者也在这方面进行这不断地探索。其中 Thomason 公司的 TDA 在这些方面有着突出的成绩，但是也存在价格高昂、非结构化处理、数据库范围和格式限制、数据信息结果没有个性化，如：科技信息资源的实时地理分布等情况。其中最关键的因素是所有的科技信息需要用一种统一的数据库格式进行处理，或在一种应用软件平台下，进行多种数据格式的转换。

5.2 智能化科技信息服务解决方案

针对上述提到的科技信息服务存在的问题，本文提出了基本的解决方案设想。本文的解决方案旨在建立一种成本低，容易形成的解决模式。具体思路如下：

（1）建立用户服务平台

可以由科技部门或科研机构建立网络服务平台，用户可以使用会员注册的方式进入平台，按照自身的需要选择不同的科技服务信息。这对于很多企业用户，可能就省下了高昂的战略咨询费用。用户服务平台的结构图如图 1，针对图 1 中的应用服务器的运行模式，详见图 2。

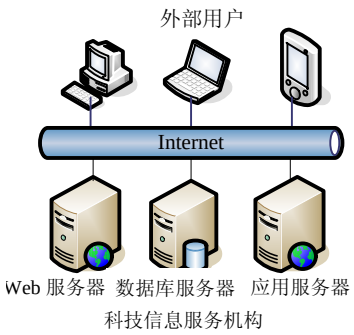

图 1 科技信息服务平台结构

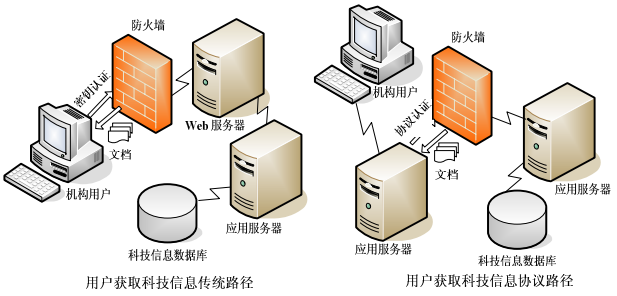

图 2 用户获取科技信息模式比较图

（2）建立应用服务渠道

机构在拥有科技信息数据库的使用权限后，能够随时从中获得离散的数据信息，这是目前很多科研院校和研究机构可以做到的。但是能够自动的从科技信息数据库中读取，并且能将其自动的导入机构自己的数据库的服务还没有完全建立，就目前的网络和计算机技术而言，完全可以解决。

解决上述问题的关键就是建立自适应的应用服务器，在机构的应用服务器与专业科技信息提供商之间建立证书许可服务，允许机构服务器直接通过应用服务器访问科技信息数据库，进而直接将其转换为自身可以处理的结构化数据，这也只是对目前专业科技信息提供商对机构提供服务的一种改变，两种服务模式比较，如图 2。

（3）开发设计自适应处理软件

目前的软件，如社会网络分析软件 UCINET、Pajek、CiteSpace 等均是单机处理形式，无法满足网络用户的便捷需要。但是由于他们当中很多都是开源结构，对软件行业而言，是比较容易开发出基于网络服务的处理软件。

目前，也存在一些网络化的处理软件如：Web of Science 的 Results Analysis 功能子系统、Vantage Point、TDA，它们的致命缺点是针对的数据库来源是有限制的，因为它们是服务于自身而专业的数据库，而目前，国内的科研机构购买的数据库资源往往是十几家，甚至更多的数据库供应商。

既然很多软件是开源的，参照的模式是既有的，在此基础上进行自行的开发，再参照已经实现网络处理的专业软件的运行和服务模式，应该是完全可行的。据此，本文提出自适应软件及机构应用服务器功能结构，如图 3。

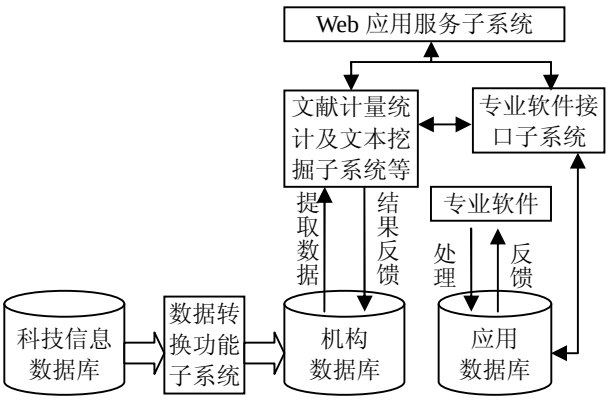

图 3 机构应用服务器功能结构图

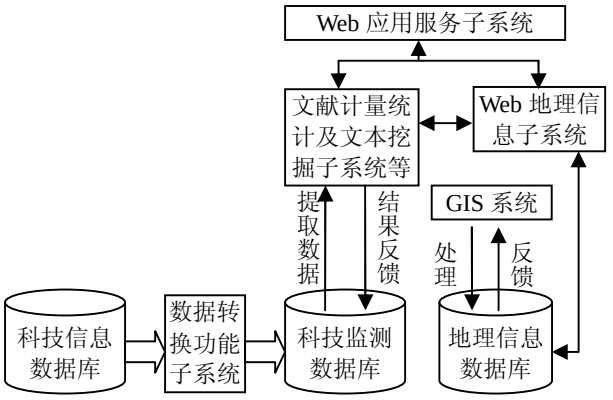

图 4 国际科技资源监测应用服务器功能结构图

5.3 实例分析

以本文所受资助项目为原型需求，首先建立一个门户网站，来达到用户服务的基本平台，如图 1；其次购买或取得相关科技信息数据库的查询使用授权，建立框架协议，如图 2；针对课题研究的特定需要，建立应用服务器功能结构，如图 4。

图 4 中所涉及的课题设计思想主要是：用户通过 Web 服务器发出科技信息服务要求后，服务器通过对大量科技文献、专利数据库和 Web 文本等数据的收集，加工成目标数据，存储于科技监测数据库。一方面运用社会网络分析软件，如 Ucinet 等软件，文本挖掘软件，如 TDA 等软件，进行科技信息统计和分析，；另一方面，借助 Supermap 地理信息处理系统，将科技监测数据库涉及科技信息与其所有者的地理信息联系起来，能让科技信息利用者随时掌握科技信息的聚集分布状态和地理分布信息，来支持企业的创新战略决策。

结论

本文以国际科技资源监测为研究对象，对科技资源监测的目标对象、手段和方法做了系统的分析。并根据目前国际科技资源监测和服务的情况，拟定了一套行之有效的智能化解决方案，并且结合实际的科学研究课题进行了分析。

在国际科技资源监测的研究方面，还有很多技术和方法有待进一步开发、探索，在实际的运用中，还要根据实践情况进行适应的调整和改进，等等，均需要国际科技资源监测的研究人员进行深化研究。

参考文献

[1] April Kontostathis ,et al.Asurvey of Emerging trend Detection in textual Data Mining[M].A

Comprehensive Survey of Text Mining ,chapter 9.Spring-Verlag,2003.

[2] Kostoff R N, et al.Literature-related discovery (LRD): methodology [J].Technological Fore-casting and Social Change, 2008 (75):186-202.

[3] Sun taotao. International Science and Technology Cooperation-Oriented S&T Resource Monitoring[D]. Beijing: Beijing Institue of Technology,2010.

[4] Sun taotao, Liu Yun, Wang Wenping. Competitive Technical Intelligence Analysis Based on Patents Coupling[C]. 2010 International Colloquim on Computing, Communication,Control and Management Proceedings,August 20-22,2010,Yangzhou,China: 102-105.

[5] Sun taotao, Liu Yun. Development of Virtual Reality Technology Rearch via Patents Data Mining[C]. 2010 Third International Symposium on Intelligent Uniquitous Computing and Education Proceedings, September 18-19, 2010,Beijing, China: 428-431.

[6] MENG Xiangping, GAO Yan. Electric systems analysis[M]. Beijing: Higher Education Press, 2004. 3-21.

[7] Small H. Co-citation in the scientific literature: a new measure of the relation between two documents[J]. Journal of American Society for Information Science, 1973, 24(4):265-269.

[8] VARY COATES,MAHUD FARDOQUE,RICHARD KLVANS. On the future of technolog-ical forecasting[J]. Technological Forecasting and Social Change,2001,67(1):1-17.

[9] CHAOMEI CHEN. CiteSpace II Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. Journal of the American Society for Information Science and Technolog[J], 2006,57(3):359-377

[10] JAMES MOODY,DANIEL MCFARLAND,SKYE BENDER DEMOLL.Dynamic Network Visualization[J],2005,110(4):1206 -1241.

[11] Juan Gabriel Cegarra Navarro, Narciso Arcas Lario.Building co-operative knowledge through an unlearning context[J]. Management Research Review, 2011,34(5):1-24.

[12] Mário Franco, André Magrinho, Joaquim Ramos Silva. Competitive intelligence: A research model tested on Portuguese firms[J].Business Process Management , 2011,17(2):1-27.