使用实体分析快速且轻松地显著提高您模型的准确性

Redguides

面向业务负责人

Dr. Lisa Sokol
Jeff Jonas

- 了解实体分析如何为您的业务提供价值
- 了解 IBM SPSS Modeler Premium 如何支持实体分析
- 从实时实体分析中获取洞察见解

Redbooks

实体分析概述

分析人员在尝试集成大量涉及整个企业的数据时，会频繁遇到各种严峻的挑战。尤其是当此类数据包含个体差异 （例如，Bob 与 Robert）、意外错误 （例如生日中月份和日期的顺序颠倒）以及专业伪造的谎言 （例如，假身份）时，这种挑战变得更加明显。错误或不完整的集成可能对于通过使用数据构建的任何分析解决方案具有负面的影响。

通过实施 “ 实体分析 ”，分析人员可以比以往任何时候都更轻松地克服部分最艰难的数据准备难题。通过使用 “ 实体分析 ”，分析人员可以生成更高品质、更准确的分析模型，从而取得更好的业务成果。无论目标是检测和预防风险，还是识别和响应商机，此功能均可以完成。

其中一个关键的数据准备功能是，当同一实体对相同实体进行多次引用 （在相同的数据源内和多个数据源之间）时进行识别。例如，由三个不同人执行三次交易与由一个人执行全部三次交易之间是完全不同的概念，理解这一点很重要。

确定实体相同 （已解析）之后，识别出这些已解析的实体互相之间存在关联 （例如，共享一个家庭地址），那么就可以对实体进行更深入的了解。“ 实体分析 ” 并不局限于以往简单的匹配或者合并技术，而是提供一些新的思想：真实的**环境累积**。环境累积是将新数据与原先的数据相关联并牢记其间关系的一种渐进过程。您可以通过对事物相关信息的了解来更好地对其加以理解。该过程可改善数据准确性。

例如，当您仅仅观察一片独立的拼图本身时，将难以评估其重要性 （第二页图 1 中所示）。但是，通过首先将这一片拼图与整个拼图进行比较以查看其与先前所见的拼图碎片之间的关联，可以更全面地掌握整个图片并更好地进行预测。

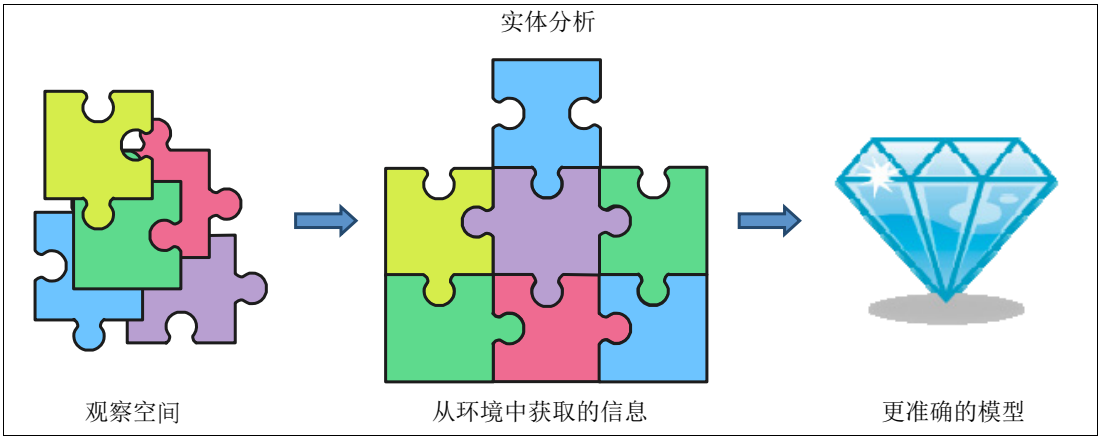

图 1 “ 实体分析 ” 通过多样化数据累积环境

本 IBM® Redguide ™ 阐述了 “ 实体分析 ” 如何通过其创建的模型来帮助分析人员取得更好的业务成果。本指南对 IBM SPSS® Modeler Premium 产品 （其中集成 “ 实体分析 ” 功能）进行了概述。同时还提供了 “ 实体分析 ” 如何作出判断并帮助解决业务问题的示例。此外，本指南通过假想银行的贷款场景，向您展示了 “ 实体分析 ” 的应用环境。

关于 IBM SPSS Modeler Premium

IBM SPSS Modeler Premium 是高性能的预测和文本分析工作平台，有助于您从数据中发掘出史无前例的洞察见解。它可提供一系列分析功能，其中包括：

- 数据的可视化和探索
- 数据操作
- 数据清理和转换
- 预测模型的创建和评估
- 部署以生产 （运行时间）模型或评分形式展示的结果

在 IBM SPSS Modeler Premium 中应用 “ 实体分析 ” 功能

SPSS Modeler Premium 包含 “ 实体分析 ” 功能，该功能可供分析人员用于快速轻松地将身份、行为和操作数据与其各自的实体进行实时关联或批量关联。SPSS Modeler Premium 中的这些 “ 实体分析 ” 功能代表着一种突破性的技术，这是其同类技术中迄今为止唯一投入商用的功能。不仅如此，这些功能便于使用，因此，您可以即刻加以利用。

过去，分析人员将 80% 的时间用于准备和清理要用于分析的数据。通过使用 “ 实体分析 ”，用户可以根据在更短时间内获得的更加彻底的清理过的数据来构建更准确的模型。
“ 实体分析 ” 的用户可以获得以下独特优势：

- 更准确的构想

  为实体累积的标识符越多，“ 实体分析 ” 技术准确性越高。

- 更好的模型

  从环境中获取的信息 （了解数据如何关联）可提供更高质量的模型。

- 更好的成果

  应用于环境增强型交易，这种更高质量的模型可促成更好的决策 （例如，风险评分）。

  例如，根据常规，银行需要就超过 5,000 美元的所有现金交易进行报告。按此法规，银行必须了解五笔看似不相关的 1,000 美元现金存款交易与某个人存入 5,000 美元现金存款的交易之间的差异。如果银行无法准确确定该个人的累积 （历史）交易次数，那么将无法确定是否超过了 5,000 美元限额。

实体分析 （如图 2 中所示）提供一种简单的方式 （通过使用环境累积功能），在即使不使用公共密钥的情况下也可以将交易与正确的实体相关联。（帐户不共享同一个税务标识号。）因此，当在该环境中发生交易时，评分模型会基于 5,000 美元的数字进行操作，而非基于多笔看似不相关的 1,000 美元交易进行操作。

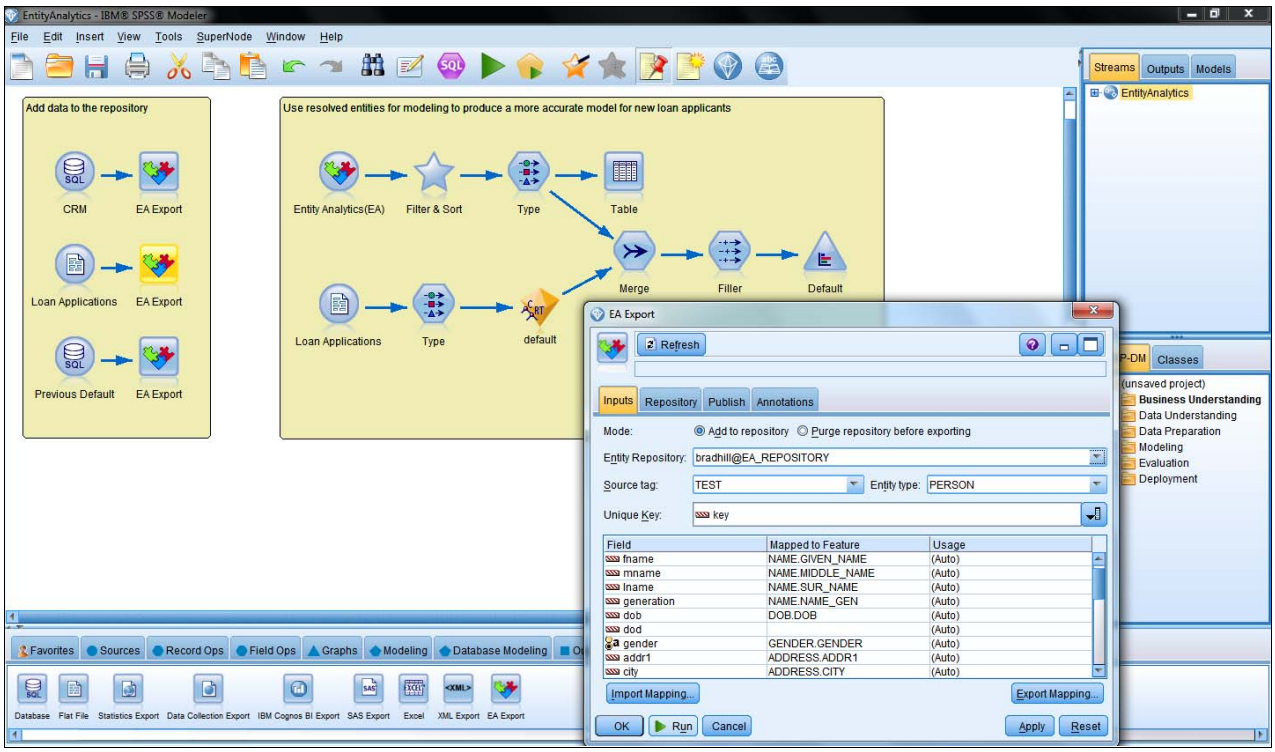

图 2 使用 “ 实体分析 ” 的 SPSS Modeler Premium 流示例

在 SPSS Modeler Premium 中，可通过以下方法来使用 “ 实体分析 ”：

- “ 实体分析 ” 导出客户机节点可执行环境累积操作。该节点确定两个实体（例如，个人、企业或车辆）是否相同。尽管这些实体是分别记录的，甚至在某种程度上是以不同方式记录的，但是此操作仍可实现。如果确定两个实体相同，那么会积累该实体的标识 （例如，名称、地址或电话号码）和衡量信息 （例如，平均余额或贷记限额）。该节点会自动应用复杂的模糊匹配技术。例如，它利用基于 8 亿多个人名的内部数据库来提供世界顶级水平的、并且考虑到文化差异的姓名比较结果。随着实体解析过程的不断深入，关于每个实体的了解也会不断增加。导出客户机节点通常用于将历史数据与新的渐进数据相集成。

- 通过使用 “ 实体分析 ” 源节点（用于读取已解析的标识），分析人员可以访问环境中的信息。该节点通常用于分析环境中的历史信息，并用于创建数据视图以支持构建新的模型。

- 流 “ 实体分析 ” 节点用于将新记录成批或实时应用于历史信息。它会在实体相同或者相关联时即刻识别。该功能与拿到一个数据片段 （一片拼图）并寻找是否存在其他关联的数据片段 （关联的拼图碎片）的情况略有相似。此节点可发现与实体相关的新数据，而此新发现可发送至合理化实体数据的流程中。

  例如，您的另一个银行帐户中新发现的新数据可用于更新财富变化。下游模式评估过程可通过相关模式和模型再次评估该实体，以确定该实体目前是否达到相关阈值。由于环境得到了增强，模式或评分评估结果将更准确。

应用 “ 实体分析 ” 的银行贷款场景

要了解 “ 实体分析 ” 的工作方式，请参考此假设示例，其中包含银行向客户提供贷款的典型流程。

预测分析可用于帮助银行确定哪些客户可能偿还其贷款以及哪些客户可能拖欠其贷款。为确定个人偿还贷款的能力，通过来自以下不同数据源的可用数据创建了多个模型：

- 历史客户数据 （例如，收入、债务或者先前的拖欠情况）
- 过去的贷款结果 （例如，贷记限额、平均付款额或者逾期未还款情况）
- 其他常用数据点

图 3 显示了银行客户贷款数据示例。前两行是历史客户数据。第三行包含正在申请新贷款客户的数据。表中最后一列显示该客户是否存在申请中的贷款。

对抓 署吃	媪凍	偻趿 惚刑	澄倪 趁傖	负债 收入比	产助 挚歼	疏豔争
102	8000	5359	2009	92.1	是	否
343	9000	6000	3000	100	是	否
642	31000	1362	4001	17.3	否	是

图 3 银行贷款数据

通过使用历史数据，SPSS Modeler Premium 可以生成预测模型，来评估新贷款申请的偿还能力。生成的评分规则示例：“ 如果某个人的负债收入比高于 24.6，并且曾经拖欠还款，那么可能在将来贷款时拖欠还款。” 在图 3 所显示的示例中，实体 #642 正在申请贷款。此人称未曾有拖欠还款，并且负债收入比较低。将先前定义的规则用作评估标准，那么该个人可能会被批准接收贷款。

如果您仔细观察图 3 ，您可以想象有关三个不同客户的三个数据点与有关同一客户的三个数据点之间的差异。假设客户 #642 与客户 #102 和 #343 是同一人。如果您明明知道此人过去曾两次拖欠还款，那么您是否会认为此客户（拥有申请中的信贷）存在信贷风险吗？

如果客户始终一致地使用其真实姓名、地址和标识，并完整且明白无误地提供所有详细信息，那么确定此信息属于同一客户将轻而易举。不幸的是，由于疏忽的数据质量问题和时而存在的犯罪意图，确定此信息代表相同的客户说起来容易做起来难。但又幸运的是，利用 “ 实体分析 ”，用户可以快速轻松地执行环境累积以准确检测此类情况。

图 4 显示实体 #102、#343 和 #642 共享大量的标识，以强烈佐证这些实体属于相同客户。

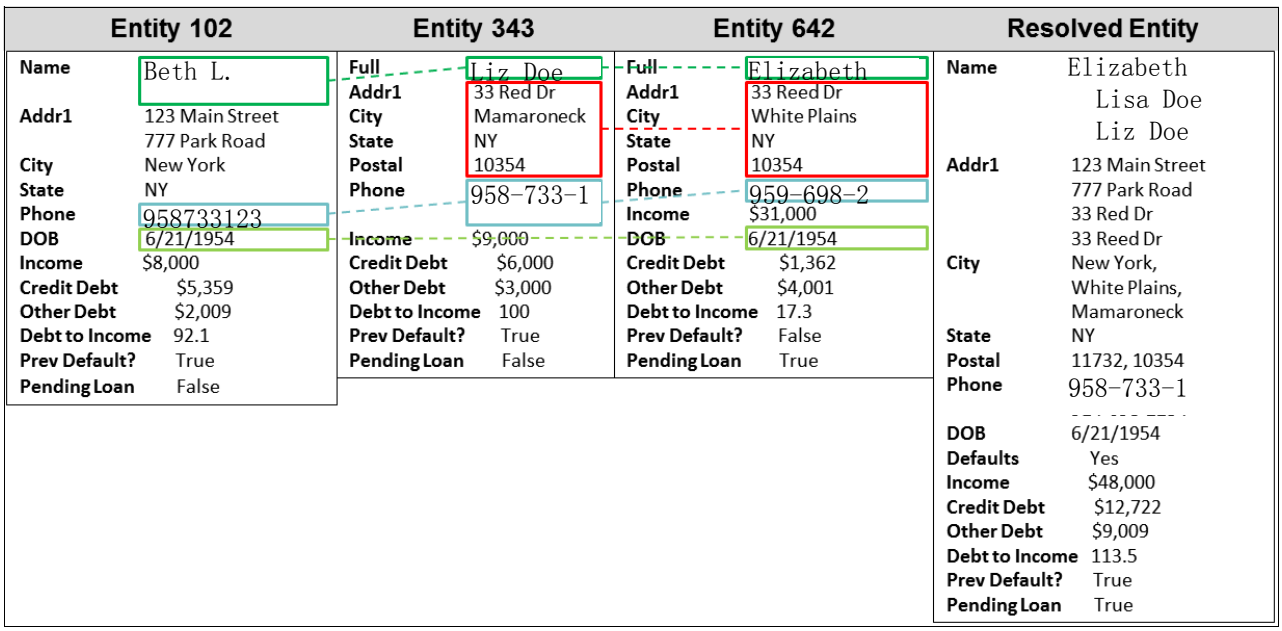

图 4 用于构造环境的多条记录之间的共通属性

利用 “ 已解析实体 ” 列中所收集的这些事实数据，更加突显了帮助对实体 #642 申请中贷款进行正确评分的环境的重要性。通过使用 “ 实体分析 ” 源节点创建的实体数据，分析人员可以对真实的借贷情况加以总结。分析人员可以确定已解析实体的借贷额为 12,722 美元，并且负债收入比为 113.5。当评分算法应用于已解析的实体时，评分结果显示实体 #642 不应得到贷款。此示例展示了 “ 实体分析 ” 的真实价值，即更快提供更准确的决策。

实时实体分析

通过使用 SPSS Modeler Premium 中的 “ 实体分析 ”，企业可以在环境中对交易进行实时分析以制定最优决策。基于所有全局化的信息，模型可以更准确地预测结果，以便立即做出决策（例如，实时欺诈检测）。

想象某欺诈调查员刚发现一个与正在进行中的内部犯罪调查相关的新地址。利用此信息，几秒钟后，“ 实体分析 ” 会警告此调查员，在该调查员的信用部门中某个员工具有相同的地址。通过这一环境累积过程，“ 实体分析 ” 将新数据 （新地址）与先前的数据 （调查情况、客户和员工）相关联，以提供此类卓越的 “ 内部威胁 ” 洞察力及更多其他洞察见解。

总结

通过在 IBM SPSS Modeler Premium 中应用 “ 实体分析 ” 功能，分析人员可以将多样化的企业数据统一结合到环境中。随后，无论以减轻风险还是识别商机作为目标，组织都可以使用环境中的此类信息来改善模型质量、制定更好的决策，并最终实现更大的成功。

组织可以对其所掌握的情况更深入理解，并就此比竞争对手更快速地开展行动，从而提升竞争力。利用这一激动人心的新技术，各种规模的组织可立即获取竞争优势。

IBM Business Analytics 软件可提供决策者实现更理想的业绩所需的切实可行的洞察力。IBM 可提供由商业智能、预测和高级分析、财务业绩、战略管理、监管、风险与合规性以及分析应用程序构成的全面统一的产品服务组合。

利用 IBM 软件，企业可识别趋势、模式和异常状况，比较 “ 假设 ” 场景，预测潜在威胁和商机，识别并管理关键的业务风险，对资源进行规划、预算和预测。利用这些深入的分析功能，我们的全球客户可以更好地了解、预测并实现业务成果。

有关更多信息的其他资源

有关 SPSS Modeler Premium 的更多信息，请参阅位于以下地址的产品页面：

ibm.com/software/analytics/spss/products/modeler/premium.html

以下团队负责编写了本指南

本指南是由来自全球的专家团队与 International Technical Support Organization (ITSO) 合作制作的。

Lisa Sokol 博士是 IBM 政府服务 CTO 办公室的架构设计师。其主要工作领域为辅助政府机构处理决策过载问题以及使用分析来发现埋藏在大量数据中的可操作信息。设计了多个系统用于检测和评估与欺诈、恐怖主义、反间谍活动和犯罪活动相关的威胁风险。并且是毕业于 University of Massachusetts 的运筹学博士。

Jeff Jonas 是 IBM 员工，也是 IBM Entity Analytics Group 的首席科学家。在加入 IBM 之前，他领导自己的公司 Systems Research and Development 完成了多个独一无二的系统的设计和开发。他设计的下一代技术有助于组织更好地利用其企业范围内的信息资产。他走遍世界，就创新、国家安全和隐私与政府领导、行业主管、全球各大智囊团、隐私维护团体和策略研究组织进行讨论。他是负责编写 National Security in the Information Age（信息时代中的国家安全）的 Markle Foundation Task Force 成员之一，同时也是 US Geospatial Intelligence Foundation (USGIF) 委员会、EPIC Advisory Board 和 Privacy International Advisory Board 的成员。他还是 Center for Strategic and International Studies (CSIS) 的高级助理以及 Singapore Management University 的杰出的信息系统工程师（助手）。

感谢以下人员对本项目作出的贡献：

LindaMay Patterson
ITSO, Rochester, MN

现在您也可以成为一名发表作品的作家！

这正是同时展现您的技能、发展您的职业生涯并成为发表作品的作者的一个良机。加入 ITSO 实习项目并帮助编写您所擅长领域的书籍，同时使用领先的技能来磨砺您的经验。您的努力将帮助提升产品接受度和客户满意度，与此同时您会扩展自己的技术联系人和关系网络。实习期跨度为两到六周，您可以亲自参与或者通过在家工作形式远程参与实习。

了解有关该实习项目的更多信息，浏览实习索引，并在以下地址在线申请：
ibm.com/redbooks/residencies.html

随时了解 IBM Redbooks 的最新信息

- 在 Facebook 上关注我们：
  http://www.facebook.com/IBMRedbooks
- 在 Twitter 上关注我们：
  http://twitter.com/ibmredbooks
- 在 LinkedIn 上关注我们：
  http://www.linkedin.com/groups?home=&gid=2130806
- 探索新的 IBM Redbooks® 出版物、实习以及 IBM Redbooks 每周时事通讯研讨会：
  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm
- 利用 RSS 订阅源来随时了解最近的 Redbooks 出版物信息：
  http://www.redbooks.ibm.com/rss.html

声明

本信息是为在美国提供的产品和服务编写的。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能特性。有关您当前所在区域的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。但是，评估和验证任何非 IBM 产品、程序或服务，则由用户自行负责。

IBM 公司可能已拥有或正在申请与本文档内容有关的各项专利。提供本文档并未授予用户使用这些专利的任何许可。您可以用书面方式将许可查询寄往：
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

本条款不适用英国或任何这样的条款与当地法律不一致的国家或地区：INTERNATIONAL BUSINESS MACHINES CORPORATION“ 按现状 ” 提供本出版物，不附有任何种类的 （无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某种特定用途的保证。某些国家或地区在某些交易中不允许免除明示或暗含的保证。因此本条款可能不适用于您。

本信息中可能包含技术方面不够准确的地方或印刷错误。此处的信息将定期更改；这些更改将编入本资料的新版本中。IBM 可以随时对本资料中描述的产品和 / 或程序进行改进和 / 或更改，而不另行通知。

本信息中对非 IBM Web 站点的任何引用都只是为了方便起见才提供的，不以任何方式充当对那些 Web 站点的保证。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

此处包含的任何性能数据都是在受控环境中测得的。因此，在其他操作环境中获得的数据可能会有明显的不同。有些测量可能是在开发级的系统上进行的，因此不保证与一般可用系统上进行的测量结果相同。此外，有些测量是通过推算而估计的，实际结果可能会有所不同。本文档的用户应当验证其特定环境的适用数据。

涉及非 IBM 产品的信息可从这些产品的供应商、其出版说明或其他可公开获得的资料中获取。IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。有关非 IBM 产品性能的问题应当向这些产品的供应商提出。

本信息包含在日常业务操作中使用的数据和报告的示例。为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。所有这些名字都是虚构的，若现实生活中实际业务企业使用的名字和地址与此相似，纯属巧合。

版权许可：

本信息包括源语言形式的样本应用程序，这些样本说明不同操作平台上的编程方法。如果是为按照在编写样本程序的操作平台上的应用程序编程接口 （API）进行应用程序的开发、使用、经销或分发为目的，您可以任何形式对这些样本程序进行复制、修改、分发，而无须向 IBM 付费。这些示例并未在所有条件下作全面测试。因此，IBM 不能担保或暗示这些程序的可靠性、可维护性或功能。

本文档 “REDP-4913-00” 创建或更新于 2013 年 1 月 28 日。

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corporation 在美国和 / 或其他国家或地区的商标或注册商标。这些术语和其他 IBM 已注册商标的术语在本信息中首次出现时都使用适当的符号 （® 或 ™ ）标记，以表示在本信息发布时由 IBM 在美国注册或拥有的普通法商标。这些商标也可能是在其他国家或地区的注册商标或普通法商标。在 Web 地址 ibm.com/legal/copytrade.shtml 中包含了 IBM 商标的最新列表。

以下术语是 International Business Machines Corporation 在美国和 / 或其他国家或地区的商标：

IBM®
Redbooks®
Redguide™
Redbooks （标识）®
SPSS®

以下术语是其他公司的商标：

其他公司、产品或服务名称可能是其他公司的商标或服务标记。