SAS Text Miner
文本挖掘

利用隐藏在文本信息中的价值。

SAS 文本挖掘可以做什么？

SAS Text Miner 结合了 Teragram 所提供的先进语言学技术，将其整合到 SAS Enterprise Miner 的核心数据挖掘解决方案之中。这是一套全面的文本挖掘解决方案，可以帮助您将非结构化数据（自由形式文本）和结构化数据整合在一起，从而在集成的预测建模环境中，为企业提供完整的视图和有价值的分析结果。自动化理解文本数据源（而不需要手动分析），结合交互式的深入钻取报告，同时利用严谨先进的数学分析算法，让您可疑抓住未来趋势，在面对新时机时及时有效地采取行动，同时降低风险。

SAS 文本挖掘为什么重要？

SAS Miner 可以自动化阅读和解析文本这些原本非常耗时的任务，节约成本和资源。同时，通过整合结构化数据和非结构化的文本信息，你就可以获得更加精准的企业视图。可以对这两种类型的数据展开分析，生成描述性模型和预测性模型，发现更多的业务机会，更准确地识别出趋势，从而能够制定出更好的决策来指导行动。

SAS 文本挖掘为谁而设计？

SAS Text Miner 主要面向必须查看大量的文本来提取信息、获得创意和了解趋势的业务分析师和统计人员。该软件可应用在所有的行业和政府部门，而对于积极营建预测模型的企业用户来说尤其重要。

企业每天都会产生大量文本形式的信息。包括用户回馈、e－mail、Web 文档、博客、微博、备忘录、保修索赔、调查问卷、期刊文章、调查研究、简历、客户记录、竞争分析等等…而且这一清单还在不断增长。没有人会有足够的时间来阅读所有的文档，更不用说对这些文本信息进行组织和分类，也就很难充分利用这些信息。

要想从这些已收集的数据中获取最大化的价值，就需要对它们进行有效地分析。但是由于会话语言的模糊性和多样性，导致很难识别、量化、分析或利用隐藏在文本数据中的信息。而且，大多数企业在决策过程中都缺乏整合文本信息与结构化数据的能力。

借助 SAS Text Miner，您可以将文档分类到预先定义的或由数据驱动的类别中，找出主题之间隐含的关系或关联，整合文本数据和结构化数据。交互式探索能够帮您发现文档集合中那些以前未知的模式，然后将这些知识用在预测性模型中，从而发掘出所有信息源的最大价值。

主要优点

- **通过自动化处理，缩短决策时间。**通过应用智能算法和自然语言处理技术，那些以往需要耗费大量人力和时间才能完成的工作，如分类、标记或建立主题库及文档索引等，都可以持续有效地自动完成。
- **揭示以前未知的关联关系，加强发现处理能力。**为什么还要仅限于搜索词条和查询已知内容这些简单的文本分析能力呢？SAS Text Miner 提供了独特的数据驱动方法，来识别出新的概念，以其特有的交互式用户界面来高亮显示路径和链接，让您可以对文档做更深入的分析。
- **直观显示高层次的数据视图，深入挖掘文档中的特定词组。**SAS Text Miner 在整个数据挖掘过程中都提供了可视化显示的能力，并且可以深入挖掘文档集合中各个词条之间的关联关系。
- **以一整套预测模型工具帮助您识别趋势和把握商机。**对消费者来信和呼叫中心记录的信息分析，能够提供有关用户不满或服务及产品需求等方面的重要信息。

产品概述

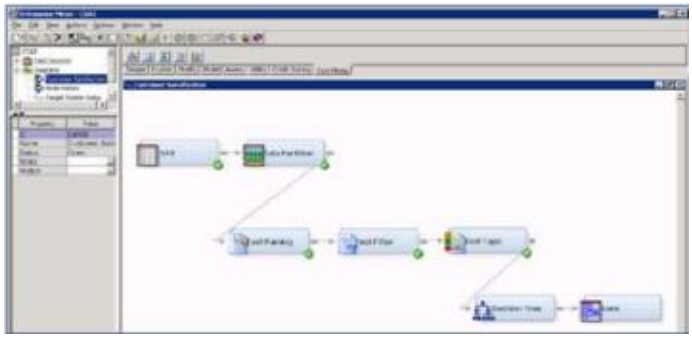
SAS 文本挖掘软件提供了丰富的语言学和分析建模工具，用来从繁多的文本文档中发现提炼有用的信息并进行预测。文本经过转换后会变成结构化数据，可以提供给后续的数据挖掘引擎进行挖掘。主题和话题也被识别出来，形成明确的关联关系，这样就可以对文档进行聚类，划分到相关的群组中，用于后续的评分或者预测型模型。提供高性能的搜索功能，增强型的拼写检查和处理单文档中包含多个话题的能力。SAS Enterprise Content Categorization（SAS 企业内容分类）的分析结果，或者是 SAS 文本挖掘软件中的概念创建插件产生的分析结果，都可以直接整合在您的文本挖掘系统中，用以补充您创建的自定义实体。

访问多种格式和语言的文档

SAS Text Miner 可以读取以各种文档格式存储的文本。它还可以进行预处理，将多种格式的文档转换成 SAS 数据集，以便输入到 SAS 文本挖掘系统中。这使您可以在单个系统中整合多种数据源（包括通过网页抓取功能获得的互联网和社会媒体网络信息），进行信息分析。它还包括自定义的程序和阿拉伯语、中文、英语、法语、德语、意大利语、葡萄牙语、日语、韩语、波兰语、瑞典语、荷兰语和西班牙语等字典。它还能对所支持的语言进行实体抽取。对于那些目前尚不支持的语言，您可以使用 Unicode UTF-8 编码来进行处理。

友好灵活的用户界面

Java 客户端/SAS 服务器架构，提供了信息摘要图形，使您可以更轻松地进行文本文档挖掘，获得深入的洞察。通过服务器分层，可以将计算过程和用户界面分离开来。用户们在自己的桌面上操作时，强大的 UNIX 和 Windows 服务器可以处理密集型挖掘任务。这一功能提供了前所未有的灵活性，方便您从单用户平台扩展到企业级解决方案。此外，当模型创建后，界面可自动生成评分代码。评分代码可以导出和部署到常用的商业智能客户端软件中，包括 Microsoft Excel、SAS Enterprise Content Categorization、SAS Enterprise Guide 和 JMP。



SAS Text Miner 4.2 包括三个新的节点（文本解析节点、文本过滤节点和文本主题节点）、多词短语、全文搜索功能和与用户自定义实体之间的协作，这些都是新的功能。

全面的文本解析

文本解析用于分解文本数据，通过量化的表达方式，用于后续的数据挖掘。SAS Text Miner 4.2 增加了新的文本解析节点，将文本数据解析成为有意义的词性、地址、电话号码以及公司名称，包括词干或词根。这种增强型的解析器，让您可以选择忽略一些停用词或者指定同义词。在原有的解析功能上，又新增了多词短语和用户自定义实体等功能。

降维

通过 SAS Text Miner 成熟的降维技术，可以进行高级过滤（使用权重信息），集成了拼写检查和将定量数据转化为紧凑格式的功能。通过奇异值分解、上卷词或者两者结合的技术，可将解析后的文档转化成数值表达形式。

文本主题识别和聚类

通过先进的算法，可以根据文本内容将文档自动分组成多个常见话题和主题。与以往将文档硬性划分到某个单一主题（也称‘硬聚类’）不同，SAS Text Miner 4.2 提供了新的文本主题节点，任何给定文档都可以和多个兴趣主题相关联，或者是不和任何主题关联。这些主题可以由用户定义，或者由工具自动判别。文本主题节点的交互式界面，让用户能够查看文档聚类以及与之相关联的不同主题，并随时调整主题定义。当然，如果您需要硬聚类，可以使用文本挖掘节点，将主题对应到聚类层次或聚类列表上。期望最大化聚类，应用了空间聚类技术来组织文档分组。您还可以在原始文本文档旁边，以一种易于解释的方式显示聚类摘要。交互式可视化环境，使分析人员能够探索文档间的概念和关系，并进行动态修改，以便后续的处理和分析。

文本过滤

SAS Text Miner 4.2 新增了文本过滤节点，提供了集成的全文搜索、自动拼写检查，概念链接、以及对词条和文档抽取子集等功能。交互式查询，可以让您按照自己指定的搜索参数检索各个匹配文档。过滤器可以基于任何特征进行过滤，包括是否包含某些词条等。而且交互式可视化使您能够进行深入探查，直到您找到所需的文档和词条。在概念图中将词条、短语和实体以可视化的方式链接起来，并提供交互式的操作，让您可以识别以前未曾检测出的模式。

直接应用其它 SAS 分析软件的分析结果

SAS Text Miner 可以与 SAS 的主流预测建模软件或其它新的 SAS 文本分析产品无缝集成，提供一整套的文本和结构化数据的挖掘工具，以及数据处理、评分和部署工具。采用 SAS 获得高度好评的分析软件，企业能够在他们的运营环境中部署分析系统，有效地解决关键的商业问题。

主要特点

通用数据访问

- 访问各种格式的文本数据，包括 PDF、扩展的 ASCII 文本、HTML 和
- Microsoft Office 文档、电子表格、演示文稿、e-mail 和数据库格式。
- 网页抓取功能，包括社会媒体讨论，如 Twitter 和 news feeds。
- 可以抽取、转换和加载文本数据到 SAS 数据集，用于后续的分析挖掘。
- 支持多种语言。

支持多种语言

- 支持 Latin-1，双字节字符和 UTF-8 编码。
- 欧洲语言（Latin-1 编码）：荷兰语、英语、法语、德语、意大利语、波兰语***、
- 葡萄牙语、西班牙语和瑞典语。
- 东方语言（支持双字节字符）阿拉伯语、汉语、日语和韩语。

友好灵活的用户界面

- 文本挖掘功能被封装到 4 个不同的节点，各自负责相应的常见任务。您
- 可以根据任务需要，将这些节点任意组合使用。这些文本挖掘节点
- 可以与 SAS Enterprise Miner 中的其它节点一起直接使用，也可通过自定义
- 算法或者自定义业务规则来进行扩展，包括预测建模、
- 聚类、可视化和报表等，并可以像评分代码那样轻松部署。
- 您可以修改、保存文本挖掘分析流程图，并与他人分享。
- 灵活的报告功能，允许结果以简洁的 HTML 格式发布。
- 概念链接图，可以直观显示词条之间的关系。

文本解析节点***

- 您可以使用缺省的或自定义的停用词表，从您的分析中移除那些只有很少信息量或没有信息量的词条。
- 自动拼写校正。
- 抽取词条的词干。
- 基于上下文来标注词性。
- 名词词组抽取，用于识别短语概念，例如“竞争智能”。
- 支持多种不同的实体类型的自动识别，包括人名和公司名称、位置、日期、地址、度量、e-mail 以及 URL 地址。这些实体可以在任一种支持的语言中进行客户化。
- 用户定义的多字短语，如“点击”。
- 用户定制的和默认的同义词表。
- 将复合词分解成不同的子项。

降维技术

- 上卷词条，自动识别文档中的高权重词条。
- 奇异值分解将每篇文档转换到 n 维空间内，其中越接近的两个文档就越相似。

文本主题节点***

- 分类浏览器显示自动生成的初始主题，以及手动创建的用户自定义主题。
- 文档可以对应零主题、单个主题或多个主题。
- 主题可以在一个易于理解和直观的环境中交互式定义。

文本聚类算法

- 最大期望值聚类法，使用空间聚类技术将文档分组成离散的不相交的群组（也称为硬聚类）。
- 分层聚类便于文档自动分组归类。
- 结合结构化数据和原始文档的聚类和主题，进行特征刻画，从而强化整体分析效果。（例如考虑年龄、购买偏好等因素）。

SAS 文本挖掘的系统需求

支持的平台

- AIX: Version 5.3 and Version 6.1 on POWER architectures
- HP-UX Itanium: HP-UX 11iv2 (11.23), 11iv3 (11.31)
- Linux for x86 (x86-32): RHEL 4 and 5, SuSE SLES 9 and 10
- Microsoft Windows (x86-32): Windows XP Professional, Windows Vista*, Windows Server 2003 family
- Microsoft Windows on x64 (EM64T/AMD64): Windows XP Professional for x64, Windows Vista* for x64, Windows Server 2003 for x64
- Solaris on SPARC: Version 9, 10
- Solaris on x64: Version 10

*注：支持的 Windows Vista 版本包括 Enterprise、Business 和 Ultimate

支持的网页浏览器

- Internet Explorer 6 on Windows XP Pro
- Internet Explorer 7 on Windows XP Pro and Windows Vista*
- Firefox 2.0 on Windows XP Pro, Windows Vista* and Linux x86 (SuSE and RHEL)

中间层要求/可选软件

- SAS 客户端和中间层需要 Sun JRE 1.5

软件要求

- 您必须在同一台机器上安装 SAS Enterprise Miner 与 SAS Text Miner; 或者必须在同一台机器上安装 SAS Enterprise Miner 桌面版与 SAS Text Miner 桌面版

主要特点 (续)

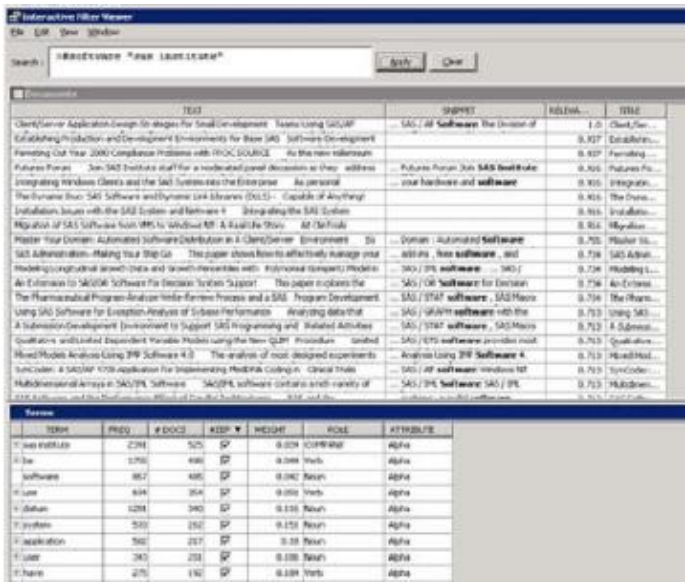
文本过滤节点***

- 包含一个有关文档、词汇或在文本解析中发现的所有词条的简明视图。
- 通过将错误拼写单词映射到正确词条，自动执行拼写检查。
- 应用类似 google 的搜索或 SQL WHERE 子句，进行子集分析(例如：对每个汽车制造商或车型进行单独的保修分析)。
- 可以通过编程或交互方式，区分和过滤出不重要的词条，轻松地映射为缩写和对等的词条。

获得 360 度的数据视图

- 将文本化数据与传统的结构化数据挖掘相结合，自动可视化地分类和部署您的预测模型结果。
- 定量和定性数据和文本分析的无缝结合，提高预测精度。
- 通过 SAS Enterprise Miner 代码节点，可以对多种高级分析技术进行扩展，包括神经网络、记忆推理、回归模型和决策树等，让您可以在低风险下进行更多的创新并且更快的进行部署。
- 可以并排显示多个模型的性能评估，帮助你选择最好的模型，像部署评分代码那样对新的文档分类。
- SAS Enterprise Content Categorization（企业内容分类）的输出可直接集成到您的文本挖掘分析中。而在没有预先分类的情况下，对 SAS Enterprise Content Categorization 来说，由 SAS Text Miner 发现生成的主题是非常有价值的输入。

*** SAS Text Miner 4.2 的新特性（2009 年 12 月发布）



提供强大搜索语法的交互式过滤查看器，帮您搜索包含给定词语或短语的文档，并提供灵活的子集分析能力

SAS 公司，免费咨询电话: 400 818 1081

若要联系您当地的 SAS 分公司，请访问: www.sas.com/china