数据库电子书

预测分析与数据挖掘集锦

数据挖掘、预测分析以及相关业务建模技术几乎完全是由高技能高工资的统计学家、数学家和定量分析师所使用。但随着商务智能和分析厂商提供更方便用户使用的预测分析工具，这一情况正在发生变化

- *数据挖掘与预测分析：趋势 效益 挑战*
- *传统 BI vs.高级预测分析*
- *预测分析：五大顶级技巧*
- *SNS 数据挖掘技术尚需完善*

TechTarget 中国

预测分析与数据挖掘集锦

数据挖掘、文本挖掘和其他应用建立预测分析模型的开放式框架正在日益普及。主要是 MapReduce 和 Hadoop，已经被分析工具和数据仓库平台厂商广泛采用。——James Kobielus

数据挖掘与预测分析：趋势 效益 挑战

数据挖掘、文本挖掘和其他应用建立预测分析模型的开放式框架正在日益普及。主要是 MapReduce 和 Hadoop，已经被分析工具和数据仓库平台厂商广泛采用。

预测分析软件获得了越来越多的来自技术用户、厂商和分析师的关注。先进的分析技术旨在帮助挖掘数据和建立预测模型的组织分析他们未来的业务场景，如顾客购买行为或拟议的企业投资的资金风险。

到现在为止，数据挖掘、预测分析以及相关业务建模技术几乎完全是由高技能高工资的统计学家、数学家和定量分析师所使用。但随着商务智能和分析厂商提供更方便用户使用的预测分析工具，这一情况正在发生变化。在通过电子邮件进行的采访中，Forrester Research 公司的分析师 James Kobielus 评论了预测分析软件的现状并对预测分析技术的发展趋势、潜在的好处和使用这类技术的挑战进行了概述。

关于预测分析是未来商务智能市场的一大战场有很多的讨论。您是否同意?如果是的话，为什么?

James Kobielus：是的，我同意。随着厂商提供支持丰富的历史数据分析

的解决方案，核心商务智能市场已经变得相当拥挤。传统的 BI 市场已经商品化也太过于简单化。然而，所有进入商务智能领域的厂商都在寻找新型的先进的分析应用作为一种避免“我也是”综合症(即提供的产品功能类似，很容易彼此混淆，无法区分及证明其值得溢价的价格)的方法。预测分析是一条 BI 产品自我发展之路，很多用户期望的东西往往需要与当前的 BI 工具相分离。

从常规观点来看，预测分析软件是否已经为更广泛的使用做好准备?还是存在需要首先解决的限制？

James Kobielus：是的，不存在限制。Forrester 公司看到了令人印象深刻的新一代的用户友好的预测分析工具，该工具满足信息工作者和其他非传统用户的大众市场需求。

传统的预测分析工具仍然十分需要一个在多元统计分析和数据挖掘领域具有统计学和数学建模学术背景的专业化骨干队伍，虽然大多数建立预测模型的厂商在推出更多的用户友好的可视化工具方面取得了很大进展。不过，当我在 2010 年初发表关于预测分析和数据挖掘工具的 Forrester Wave 报告时，我必须认真思考行业现状。我并没有很强调面向业务分析、主题专家和其他“非技术性”信息工作者的特性。今天产品的核心问题是很多留存下来的强大工具都有一条陡峭的学习曲线以及相对高昂的价格。

预测分析软件会怎样?您能不能给我们一个您追踪的关键技术趋势的概述？

James Kobielus：主要趋势是朝着用户友好、自助服务、集成商业智能的预测分析工具发展，这会是更普遍采用的做法。另一个是向着在企业级数据仓库中融入更多预测分析的功能发展，即数据库内分析。这是一个可以在数据准备、统计分析、模型评分和其他先进分析功能方面被并行化的路径，能够在一个或多个数据仓库节点上进行加速。数据库内分析还能够针对资源密集型功能(例如数据挖掘和预测建模)实现范围广泛的弹性部署，可部署到集群、网格或高性能分析数据库云。

我们也看到为数据挖掘、文本挖掘和其他应用建立预测分析模型的开放式框架正在日益普及。主要是 MapReduce 和 Hadoop，已经被分析工具和数据仓库平台厂商广泛采用。在未来一年，我们还将看到一个产业的开始，即推动内联预测模型的开放式开发框架部署到实时数据流应用的复杂事件处理(CEP)环境。另外一个趋势是在客户关系管理(CRM)应用中内置预测分析功能以推动在呼叫中心和多渠道客户服务环境中实现实时的“Next best offer”推荐。

为什么潜在用户对预测分析感兴趣?什么是企业可以从它那里得到的潜在利益或竞争优势？

James Kobielus：商业是关于投注的所有，知道是否赔率对你有利。商业成功依赖于你的公司能够很好地预测未来业务场景然后去准备计划和部署资源，以便你能够抓住机遇、消除威胁和降低风险。显然，预测分析可以在日常商业运

作中起到关键作用。它帮助你关注战略并基于实际表现和可能场景不断调整计划。而且，正如我在 Forrester 博客的文章中指出的，这项技术是你面向服务架构战略的核心，你可以将预测逻辑深深嵌入到数据仓库、业务流程管理平台、CEP 流和业务应用中。

预测分析的最大承诺（大多数公司在很大程度上仍然未实现）它会变得无处不在，指导所有的决策、交易和应用。该技术将上升到这一挑战，企业必须走向全面的先进的结合分析数据挖掘、内容分析和数据库内分析的战略。我们已经勾画出“面向服务分析”的愿景，依据你打破数据挖掘和内容分析之间孤岛的倡议，并充分利用这些跨越所有业务流程的池化的资源。

您可能会同意这是正确的理念，但有疑虑这是否是在这个方向上引领你公司的适用的渐进的路线图。实际上就是，重新开始评估它与大多数公司的预测分析能力的核心：你的数据挖掘工具。当你策划你的预测分析倡议时，你应该避免传统的以战术、自下而上、项目的具体要求为焦点的方法。你也应该尽量不要把你的要求硬塞进你目前使用的建模工具的有限功能集合中。

在另一面，当人们衡量预测分析软件的合理部署时，他们应该考虑和准备好接受什么样的挑战或问题？

James Kobielus：预测分析工具的学习曲线、复杂性和成本是主要的挑战。另外，如果你正致力于部署先进的预测分析工具，你需要聘请专业的高薪人才来

处理数据的准备和清理，建立和评价预测模型，并将模型和他们的结果集成到你的 BI、CRM 和其他应用环境中。如果你决定通过数据库内分析把预测分析倡议整合到数据仓库中，你需要将处理这些功能的人组成一个小组，并让他们讲同一种语言。

（作者：Craig Stedman 译者：沈宏 来源：TT 中国）

数据库内嵌式预测分析：消费者行为预测的利器

利用分析技术和数据仓库技术的结合，Catalina 市场营销公司可以在短时间内预测你下一次进入当地的 Safeway 超市或者 Walgreens 连锁药店很可能会买些什么物品。Catalina 公司的业务基础是两种相互独立而又有融合的技术。

第一种是数据仓库技术，允许 Catalina 公司整合、转换和存储近 8000 亿行客户数据，它详细描述了约 2000 万美国人在过去 3 年中的购物记录。

第二种是预测分析技术，驻留在数据仓库中允许 Catalina 公司针对庞大的数据集运行评分模型，而不必将其作为一个独立的应用程序运行。

数据库内嵌式分析融合了数据仓库和先进的分析技术，Netezza 上周在波士顿用户大会上发布了 TwinFin 数据仓库工具的最新版本，包括新的扩展以帮助开发人员使用 SAS 研究院的 MapReduce 框架与 R 预测分析语言来构建先进的分析等。

SAS 公司还与数据仓库供应商 Teradata 和 Aster Data 合作将其分析技术整合进数据库中。此外，IBM 最近收购了先进的分析软件厂商 SPSS，并且已经

开始提供基于 DB2 和 WebSphere 的数据库内嵌式分析产品。

数据库内嵌的分析运行评分模型速度更快

Netezza 市场营销副总裁 Phil Francisco 表示："数据库内嵌式分析技术主要优点是像 Catalina 之类的公司就无需把数据库中的数据移动、转换到一个独立的分析应用程序，从而节约了宝贵的时间和精力。"

Catalina 公司 CIO Eric Williams 赞同并指出使用数据库内嵌式分析技术减少了公司运行预测评分模型所需的大约 30%至 40%的时间，这意味着公司可以更频繁地对其大量的数据运行评分模型，以提高其准确度和客户的期望值。

"为了做到这一点，我需要难以想象的速度流转数据，" Williams 说："来自 Netezza 与 SAS 的数据库内嵌式分析技术，使得公司能够为每个客户每年运行 600 次以上的评分模型。"

"在这种情况下，把超过 50TB 的信息迁入到数据挖掘库，然后运行这些评分模型，最后再移回到它原先的数据仓库，可能需要花费数周时间，而事实上，之前我们的确用掉了这么长时间，" Williams 补充道："我们现在可以在几分钟内完成，因为我们已经把这种技术嵌入到数据库中。"

预测分析工具让企业锁定目标客户

佛罗里达州的 St. Petersburg 公司收集超过 25,000 家杂货店、药店和其他

零售网点的 POS 机客户数据，它给每一个消费者分配一个 ID 号以追踪其购买模式。

数据被汇总放入一个数据仓库，在那里 Catalina 公司运行评分模型从数据中识别出购买行为的趋势。这些模型帮助 Catalina 公司判断消费者下一次到杂货店或药房可能购买哪些东西，然后把结果传递给该公司的客户。

因此，客户在结账后，就会基于以前他们的购买行为得到一张个性化的优惠券，比如一加仑免费牛奶或者非处方止痛药，Williams 解释道。

能预测一个人未来可能购买什么以及作为第一家向消费品公司和制药公司提供此类服务是很关键的，Williams 说。历史数据表明消费者喜欢一种新产品平均只能坚持大约 18 个月。

“如果我们能预测哪些人对哪些产品有兴趣，就可以给他们参加免费促销活动的机会，这是一个免费的随意的试用活动。如果顾客喜欢它，那就是说我们又多了一个客户。” Williams 说。

数据库内嵌式预测分析的未来：实时分析

数据仓库技术和分析技术的结合甚至可能对大量数据运行近乎实时的预测分析。

比如在 Catalina 公司，对历史数据运行评分模型，然后应用到在结账时满足

消费者的某些需求。Williams 设想有一天对消费者刚刚完成 POS 机交易的数据近乎实时运行评分模型。

“这就是我们要达到的目标，”Williams 说。 “这是我们的信念，我们实际上可以依据该交易数据和历史数据得到实时评分，并在两秒内对该交易作出反应，这就是我们需要的速度。”

（作者：Jeff Kelly 译者：沈宏 来源：TT 中国）

传统 BI vs.高级预测分析

Gregg Hansen 认为在传统商业智能和高级预测分析之间存在一个可望而不可及的“中间地带”。

Hansen 是 AMD 公司的应用程序副总管。他认为 AMD 公司在与英特尔的竞争中，总是寻找各种方法获得竞争优势。

为此，AMD 已对传统的 BI 工具进行了投资，包括报表、SAP Business Objects 和 Microsoft 的可视化软件以及一些预测分析工具。该公司的管理人员和工程师都在使用这些工具发掘相关的数据。

但是，这些程序和工具仅限于对数据库和电子表格中结构化数据的发掘。Hansen 说，AMD 在维基、产品规格和电子邮件上有大量非结构化数据。如果公司能够找到一种方法将非结构化数据集成到它的结构化数据中用于 ad hoc 查询，AMD 的高管和工程师将获得新的视角来帮助他们与 Intel 竞争。

Hansen 说：“合并结构化和非结构化数据到单一查询界面是非常困难的。”

● 非结构化数据探视

AMD 的情况不是孤立的。Forrester 的分析师 Boris Evelson 说这种情况在销售部门或维护一系列与文本文件和其他非结构化数据相关产品的公司是相当普遍的。

如在消费产品公司，产品多种多样且各有独特的维度(大小，颜色，材料等)，用处理规范化、结构化数据的工具来比较、分析产品将是非常困难的。

Evelson 表示，显而易见的改进是以可定制的传统 BI 和搜索工具对不同数据进行分析。这样就占用了宝贵的时间和人力， “建模是一大难关。”

当然，工作人员也可以手工甚至是逐字逐句的将公司文件和电子邮件中大量的非结构化筛选、转换为他们想要的数据。显然，这种劳动密集型的方法在大多数组织中根本站不住脚。

另外还有一种 Hansen 主张的方法。即企业搜索技术，对结构化和非结构化数据的查询进行优化，以表格，图表和其他 BI 接口的形式返回搜索结果以帮助工作人员回答具体的、有针对性的业务问题。

我们的目标是“交付一个统一的信息视图来整合 BI 结构化数据和其他半结构化数据”，Gartner 分析师 James Richardson 在 2009 年的报告中如此写到。

马萨诸塞州 Newton 的 Attivio 提供了这样的一款产品：Active Intelligence Engine(AIE)。当在 AIE 的用户进行搜索时，不同于传统的 BI 工具和搜索引擎，此平台可以对一个组织的结构化和非结构化数据源进行查询以找到最相关的数据。

Attivio 的 CTO Sid Probstein 表示，根据数据之间的关系将数据集成到一

个统一的倒排索引，然后向用户可视化地呈现结果。

比如，搜索一款微处理器将会返回与特定芯片最相关的一系列文件清单，线性图描绘随时间变化的销售数量，饼状图则展示了每种芯片在总收入中所占的百分比。

正如 Gartner 的 Richardson 所解释的："用户受益于 ad hoc 方法的灵活性，将任何数据表和文件集合连接起来动态地生成查询，以用于他们的分析。"

在统一的信息访问方面，分析市场上与 Attivio 竞争的其他供应商包括 Endeca 和 Exaled 以及最近由 Dassault Systèmes 收购的一家巴黎供应商。

● 减少建模问题，增加对 BI 的采用

Richardson 认为，如 AIE 这类技术除了帮助工作人员获得新的洞见，也可以使得传统的 BI 更易接受。由于它的用户界面仿照熟悉的网络搜索格式，那些过度担心而不愿尝试 ad hoc 工具的商业用户更乐于使用它。

他说："希望使 BI 更加普及的 CIO 们可以考虑 AIE，那些不习惯用传统的 BI 工具来找到他们作决策所需信息的工作人员变得更容易接受 BI。"

Forrester 的 Evelson 同意上述观点并指出如 AIE 的这些工具很像谷歌，但内含更多的分析。

该工具的统一索引方法也无需创建复杂的数据模型和模式，使得数据架构师

可以专注于更为紧迫的工程。

“Endeca 和 Attivio 通过获得的数据动态推断数据模式来实现它，”Evelson 在最近的一份报告中写道：“每一个处理的元素都可以潜在的被用作一个事实或维度，因此，在源系统模式中发生的改变可以自动地以各种方式传播到 BI 应用中。”

不过，这并不意味着 AIE 和类似的平台可以取代传统的 BI 和分析工具。相反，相互配合的意义更大，甚至 Attivio 的 Probstein 都强调这一点。例如，报表构建最好由一个更加传统的 BI 工具如 Crystal 报表而不是如 AIE 这类的平台。

这种那这种那个 这种结合并不是对所有组织都是良好的。例如，有些公司基本上是存储在传统数据库中的结构化数据将会发现在 AIE 这类统一信息访问和分析平台上将获得很少的利益。

Evelson 表示，也存在与供应商锁定相关的风险。例如，AIE，作为一个专有平台则意味着客户不能定制低价格的数据库或硬件。

他说，有些组织适从这种平台架构，它们围绕多样化内容而拥有大量的结构化和非结构化数据集则相当受益。

AMD 最近购买 Attivio 的平台以解决如何提高公司网站的搜索功能这一特殊问题。为此平台的分析功能的潜在价值深感兴奋。

他说："我们认为这是 Attivio 的一个机会，给人们一个非常简单的方式，以简化的语法得到数据并获得想要的东西。"

（作者：Jeff Kelly 译者：宋广磊 来源：TT 中国）

NBA 也疯狂：体育业高级预测分析软件选型案例

IT 的触角已经伸向了每一个行业当中，当然体育界也不能例外。美国著名 NBA 球队奥兰多魔术在全世界有着数量众多的球迷粉丝，而球队除了拥“魔兽”霍华德之外，他们成功的背后还有预测分析软件作为强大的后盾。

新的赛季，奥兰多魔术也启用了他们最新的球场安利中心(Amway Center)，该球场位于奥兰多市中心最为繁华的地段。魔术队高层也开始利用数据仓库以及预测分析软件，来帮助决策者更加理解球迷的消费行为，从而做出更加明智的决策，为球队获取最高的利润。

根据魔术队业务战略团队总裁助理 Anthony Perez 的介绍，自从 2009-2010 赛季以来，球队就开始使用分析软件来对其动态票价项目进行决策支持。动态票价项目的初衷，就是为了给不同的球票制定出不同的价格，从而吸引更多的球迷前来购买，带动球票销售利润以及主场气氛。而决定球票价格的因素有很多，其中包括周几、对手的情况以及最近球队表现等等。

从今年的赛季开始，球队部署了一个新的数据仓库来对项目进行扩展，其中使用了一些高级技术，包括预测模型、客户分化以及实时决策系统来帮助团队捕捉范围更宽的业务变量和市场信息。Perez 表示，项目的扩展最终将帮助魔术队更好地销售球票，决策层将认清楚哪些因素驱使球迷购买季票，以及球迷零售商店又有哪些商品将热销等等。

● 高级分析项目遭遇数据仓库挑战

IT 团队新添加的数据仓库将从内外的数据源系统抽取大量的信息，包括球票销售、市场运作、零售店销售、赞助商情况以及球迷分布特征等。但是 Perez 表示，利用一些厂商比如 Ticketmaster 提供的技术来满足安利中心的需求，在数据仓库方面我们还是遇到了很大的问题。

球队首先遇到的困难就是试图将信息从 Ticketmaster Archtics 抽取出来，Ticketmaster Archtics 是一个软件平台，一般从事体育事业的组织可以利用它来管

理售票相关的数据，它可以提供 CRM 系统功能、市场工具以及报表功能等。当球队试图从 Archtics 的数据库中下载将近 60 万条客户数据的时候，一共花去了 70 小时的时间，Perez 对此表示这是无法容忍的，根本无法满足实时信息访问的需求。他说：“准对安利中心，我们面临了一些挑战，正是因为我们使用的技术。举个例子，Ticketmaster 处理球队信息时，通常不会提供 GB 级别的网络连接，因此我们提出了需求添加一个 GB 的路由器。”

Perez 认为魔术队利用高级分析技术是为其他 NBA 球队开了先河，在这期间他们遭遇的问题将更好地指引其他球队，让他们在未来避免出现同样的问题。

● 高级分析软件选型过程

当决定启用分析技术的时候，球队高层首先选择了 IBM SPSS，他们购买了一个基础的 SPSS 许可证和一个 SPSS 决策树模块。Perez 表示球队使用这些软件的主要目的就是决定季票购买者在下赛季是否会继续购买。

但是不久前，Perez 参加了一个研讨会，在会议上，他听了 SAS 公司一个演讲者的讲座，探讨了其新产品 SAS for Sport。从讲座中，Perez 认识到奥兰多魔术队使用高级分析软件能解决的问题，肯定不仅限于球票销售。

SAS 的技术对 Perez 的吸引力巨大，特别是因为数据仓库与分析软件的结合方面。SAS 希望从更加广泛的数据源抽取各种各样的信息，其广度是 IBM SPSS 以及

其他一些产品所无法比拟的。Perez 表示 SAS 的技术将开启一扇门，通过这扇门，球队将更加了解其球迷的方方面面。

“我们将抽取零售信息到现有的 Microsoft Dynamics CRM 平台，” Perez 说道：“这能让我们获取每一个客户的视图信息，这是我们之前没有想到的，而且是 SPSS 无法做到的。”

魔术队最终决定从 SPSS 软件转投 SAS for Sports。该软件包含了 SAS Marketing Automation 模块(测试和竞标)、SAS Enterprise BI Server 模块(包括自定义仪表盘和报表功能)、还有数据统计工具和数据挖掘工具。

● 魔术队未来如何利用高级分析技术

随着奥兰多魔术队的战绩越来越好，貌似球队没有必要担心自身的球票销售情况，但是有了高级分析软件作为双保险也不为过。

当分析软件实施到位之后，它将用来分析到场球迷分布情况，利用特殊处理将他们进行精准的定位以便确保下一次主场比赛，他们还会买票看球。该技术还将针对那些季票购买者，为他们制定更为合理的购票套餐。

对此，SAS for Sports 产品主管 Craig Duncan 表示：“这就是所谓的在独立的级别上捕捉数据。一些情况下，如果一个球队能让他们的球迷把每赛季看两场球提升到看三场球，这部分增加的利润将是极为可观的数字。”

（作者：Mark Brunelli 译者：孙瑞 来源：TT 中国）

预测分析：五大顶级技巧

在刚刚结束的全球市场分析 2010 大会上，专业技术人士讨论了一些关于高效的预测分析策略，并对如何执行给出了相关的意见与建议。

在大会上，演讲者描述了认清业务挑战、着眼于数据、跟随基本的流程以及有条理地执行项目的重要性。其中一位专家还为与会者演示了一款新的预测分析软件，该软件可以帮助企业充分利用目前十分火爆的社交网络来提高效率。

本次的大会有一个核心的思想，即积极向上的企业通常会在预测分析方面取得一定的成功。换句话说，就是那些使用分析工具来进行预测的公司，必须要时刻关注和利用这些信息。说起来容易做起来难，专家认为想要普通的公司业务人员充分地利用预测分析信息，那需要经过长时间的培训、决心和耐心。

为此，TechTarget 网站特意总结出关于预测分析的五个顶级技巧，来帮助企业用户了解从规划到执行预测分析策略的相关内容。那些有决心的企业一定会从本文中得到一些启示和帮助。

● 辨别业务挑战并报告结果

来自一家知名在线招聘机构的分析师 Jean-Paul Isson 在大会上表示，任何规划一个预测分析项目的企业，都应该以认清自身的业务挑战为基础。Isson 作为

Monster 公司的分析师和副总裁，他主持了公司的全球分析项目，他说懂得公司业务存在哪些挑战是极为重要的，这可以指导随之而来的每一步决策，包括技术和个人决策。一个公司使用预测分析软件往往希望了解潜在的销售对象、认清消费者的消费习惯或者下一年的购买成本预算。但无论处于什么原因，Isson 表示一定要循序渐进，千万不要冒进，每一项挑战都要认真分析，而不要寄希望于一次性解决许多问题。

Isson 表示，在规划阶段，相关的 IT 员工和公司业务人员需要每一步都紧密地协作，从认清业务挑战到设计报表。IT 部门还应该逐渐地计划生成有规律的预测分析结果。企业切忌制定过度费时费力的项目，这将让他们丢失更多的交易。

Isson 说：“至少每三个月，我们都需要向业务部门提供一定的分析成果，你知道，业务是不等人的。”

● 将目光放在数据质量上

根据 Isson 的观点，将更多注意力集中在数据质量上，是任何一个高级分析项目成功的前提条件。

想要找到“正确的数据”，需要的是从相关的企业内部和外部数据源进行数据挖掘和映射，当然与此同时还需要将相关业务挑战记在心里。Isson 表示，企业应该组建一个跨职能团队(IT+业务人员)来执行这一过程。

“数据是基础，” Isson 说：“当你想要造一间房子，打好地基是关键。”

● 执行分析策略时，要遵循基本的流程

来自一家大型技术咨询公司的分析师 Alberto Roldan 回忆起自己在十几年前的一段经历，那时候 BI 项目如雨后春笋一般涌现出来，但是过度地追求速度，让这些项目最终以失败告终。当 Roldan 谈到失败的原因时，他认为没有遵循基本流程是罪魁祸首。

从这些失败中学到的教训，就是那些想要实施预测分析策略的企业，需要谨记遵循基本流程否则等待他们的一样是失败。Alberto Roldan 说：“实施预测分析是有它特定的流程的，而无论何种产品，其流程也都是一样。你需要有一个登记点，你需要有数据治理，你需要有元数据，你需要有变更请求，你需要有合理的文档和捕捉这些元素的方法。”

● 有条理地选择预测分析计划

纽约德勤咨询公司的经理 Charlie Veers 认为一次实施一个小型的预测分析项目是非常重要的。他还对如何选择正确的分析项目给出了一点建议。

根据 Charlie Veers 的观点，最好为每一个潜在的分析项目制定一个业务案例，而不要将它们进行对比，然后再挑选出价值最高的那个。“当你寻找什么才是达成目标的利益驱动时，总会有各种各样的质量成分，” Veers 解释道：“当使用到某

种质量成分时，你就可以站在更高的位置上去看待一个项目，然后做出更明智的决策，哪个会给你带来更多的回报。”

● 准备将社交网络功能引入到预测分析软件

一些文本分析软件厂商提供非结构化数据的分析功能，比如 Facebook 和其他社交网络来源的数据。但是根据 Roldan 的观点，还没有哪家厂商拥有成熟的技术，来将社交网络数据转化成有用的数字变量，因此这些数据无法进入到预测的模型之内。至少现在还没有这样的软件。

Roldan 说：“我们这在为此而努力，而且我们正在以协作的方式来完成这一任务。我认为一个解决方案就是信号检测，如何将信号检测转化成数值。我们可以将这一方法运用在社交媒体上。给我们一年的时间，我想到时候你们会看到成果。”

（作者：Mark Brunelli 译者：孙瑞 来源：TT 中国）

成长的烦恼：社交媒体分析特别报道

随着越来越多的人加入到诸如 Facebook 以及 Twitter 等社交网络的行列，许多企业也从中看到了商机，他们加入社交网络并不是为了去玩一些线上游戏，上传照片，而是选择利用社交网络的优势进行市场调查、收集客户数据，并且利用一些分析软件和文本挖掘工具来看看公众对于他们业务的口碑如何。但是我们需要看到，利用 SNS 进行这一工作的大多数还是属于零售行业，那么其他行业是否也能使用这一新兴的商业智能技术呢？

为此，TT 数据库网站推出了这次社交媒体分析技术的特别报道，来帮助您决定是否需要这一技术。在这次的报道中，您还可以看到不同社交媒体分析软件产品的成熟度如何，以及利用好社交媒体分析技术的企业该如何去规划这一项目。希望能对您未来的项目起到一定的启迪作用。。

● 社交媒体分析的黄金时代已经到来？

社交网络的魅力已经无需证明，如果你还没有开始微博那么就证明你已经 out 了。放眼全球，Facebook、Twitter 的成功让无数互联网工作者赞叹不已，而现在也有更多的企业从社交网络中看到了机遇，BI 和分析软件厂商也开始竞争社交网络这一领域，这让社交媒体分析软件得到了前所未有的发展。

比如去年，SAS 就推出了一款社交媒体分析套装，旨在帮助企业了解互联网上用户对于他们品牌的口碑如何。而 IBM 也发布了一个升级的 SPSS 功能模块(数据挖掘厂商 SPSS 在 2009 年被 IBM 所收购)，该模块对特定的 200 多个行业进行了细致的分类设计，可以帮助企业挖掘出社交媒体中出现的关键字，比如微博等。就连 Facebook、Twitter 自己都推出了相应的分析工具。

这是不是意味着社交媒体分析软件的黄金时代已经到来?这一技术是否已经做好了大量消费的准备呢?如果是问厂商，那么答案当然是肯定的，但并不是所有的人都那么乐观。

咨询公司 KDPaine and Partners LLC 的 CEO Katie Paine 认为，目前市面上见到的大多数社交媒体分析软件都处在非常不成熟的阶段，它们能给企业带来的信息是非常糟糕的。收集数据是容易的，困难的是如何从大量的信息中整理出头绪，因为大部分信息是无用的灌水信息。

传统的分析技术经过几年的时间已经逐渐走向成熟，商业智能厂商 Liquidnet 的总裁 Angela Chen 表示，社交媒体分析技术的潜力是巨大的，在未来几年中，它还将不断得到发展，也许真正爆发的一天已经离我们并不远了。

● SNS 分析：企业是否做好了准备?

是否一些特定行业的企业在使用社交媒体分析软件的时候，能够将它的优势发挥到极致呢?Baseline 咨询公司的创始人 Jill Dyche 表示，这要分成两部分来看。对于那些需要更多客户服务的行业来说，它们的确比其他的一些行业有着更加天然的优势，但是这并不意味着其它的行业就不适合使用社交网络分析软件，只要是那些希望更好地了解客户，并想从公众那里得到反馈的企业，都可以充分利用这一技术。

对于 Dyche 来说，最大的问题是如何利用好 SNS 以及收集来的社交媒体数据。在一份最新的报告中，Dyche 写道：“社交媒体分析现在存在的问题，同传统 BI 之前的问题是一样的：你该如何使用它?一个社区是否对你的旗舰产品说三道四?如果有的话，那就好了，现在该怎么办呢?”一个成功的社交媒体分析项目可以帮助你解决这些问题。

● 社交媒体分析项目成功的关键因素

社交媒体分析目前还是一个巨大的未知领域，现在的用户都是在不断探索着这一技术，记录着点点滴滴并将自己的经验传授给其他想要部署软件的企业。这些社交媒体分析的先驱者们一直都在寻找着项目成功的关键因素。

举个例子，广告公司 DraftFCB 的副总裁 Pradeep Kumar 坚信他的社交媒体分析项目最终会走向成功，尽管目前他还不太确定如何去做以及何时才能看到最终的成功。Kumar 表示挖掘与分析社交媒体数据需要使用各种各样的工具，需要在不断地失败中总结出经验。

根据 Kumar 的观点，企业在使用社交媒体分析软件时，还不能够忽视最重要的一点，那就是人。情感分析(sentiment analysis)并不会那么准确，它经常会对于那些讽刺挖苦的语言得出错误的分析结果。

● 社交媒体分析：实现无处不在的 BI

最近一份英国商业应用研究中心的 BI 用户调查报告中显示，只有 11%的用户表示在他们的企业中最多只有超过半数的员工在使用 BI 工具，尽管如此，咨询师、BI 厂商和终端用户都在寻找将商业智能无处不在的良方。

而社交媒体分析软件也许能够解决这一难题。

Forrester 高级分析师 Jim Kobielus 认为社交媒体分析在未来必定会更多地整合到传统 BI 平台上去，这意味在企业中只要你是 SNS 的用户，就可以使用 BI 工具

来进行分析，并帮助自己更好地完成工作。也许，无处不在的 BI 并不是遥不可及的梦。

（作者：Justin Aucoin 译者：孙瑞 来源：TT 中国）

SNS 数据挖掘技术尚需完善

根据一份新的报告，社交媒体分析的重要性对旅游及酒店业来讲很难被低估。

最成功的品牌将是那些包容并学会利用社交媒体的，而不是那些低估其影响的。德勤事务所顾问 Robert Bryant 写道。

但百万级供应商，如 SAS 软件研究所和 IBM 公司，都投注于能够受益于社交媒体分析的行业，而不仅仅是旅游及酒店业。这两个公司都将注意力转向基于博客、维基和受欢迎的社交网站(如 Facebook 和 Twitter)数据的挖掘，目的是为一些行业企业发现有价值的数据。

然而有一些分析师和行业观察家担心和质疑其准确性，最终从目前的立场来看这种基本分析技术是有益的。

例如，今年 4 月 SAS 软件研究所发布了社交媒体分析套件，帮助其零售、金融、制药和其他行业客户去了解线上讨论对其企业品牌和设计来说是一种更有效的营销活动。

去年夏天，IBM 收购了 SPSS，其文本分析软件是作为精确进行社交媒体分析的关键部分而被关注。今年 5 月，IBM 公司发布了最新版本的 SPSS 建模软件，其

中包括了超过 180 个特定行业的分类法旨在识别各种垂直市场所特有的在博客和微博帖子上发现的语言。

即便是社交网站本身也进入了数据分析市场。今年 6 月，Twitter 收购了 Smallthought System，一家网络数据分析公司。同样在上个月，Facebook 发布了一款增强版本的分析工具(Insights Dashboard)，它向 Facebook 内容所有者提供衡量他们内容的度量指标。

“通过对用户的增长趋势和用户数统计、内容消费和内容创建等数据的了解和分析，内容提供商和平台开发者有更好的依据去改善其在 Facebook 的业务，”一个社交网站的软件工程师 Alex Himel 在电子邮件采访中说：“他们可以使用这些信息来调整自己的网页，以增加点击率或者进一步改进应用程序。”

此外，像供应商所说的，客户都获得了成功。

IBM 介绍道，荷兰的 RTL Nederland 电视制作公司一直在使用社交媒体分析技术去了解观众在 Facebook 和 Twitter 上对播出的“So You Think You Can Dance?”节目说些什么。发现观众对节目的投票过程不太感兴趣，因此 RTL 对节目的形式进行了改变

负责 SPSS 软件产品营销的 Marcus Hearne 说，情感分析技术的最新进展使得像 RTL 一样的公司不仅能从博客上观察到对他们的节目是怎么评价的，而且还能基于这些说法见解采取行动。

Hearne 说，十年前的文本分析技术唯一能做的就是找出独立的词汇，仅能提供极小范围的上下文环境关联。即使有的话也很少应用于新兴的博客。

但随着情感分析(sentiment analysis)技术的来临，“我们已经从计算词汇数量发展到理解在博客和社交网站上发帖人背后的情感是什么。” Hearne 说。

但是，并非所有人都信服这个说法。KDPaine 事务所首席执行官 Katie Paine 顾问认为，大多数社交媒体分析技术仍然在做一件“可怕”的工作，即停留在如何确定 Facebook 帖子、微博和其他网上讨论的情绪。

“很容易收集这些与社交媒体有关的数据，” Paine 说：“但是要使这些数据变得有意义是非常困难的，因为 93%的数据都是无序或者不相关的。”

Paine 引用了近期与 SAS 软件研究所合作的经历。该公司使用自己的技术来收集和分析与其本身有关的社交媒体数据。它收集了约 3500 件社交媒体的内容，然后剔除输入错误或其它主题后削减到 250 件左右实际上与 SAS 软件研究所相关的内容。

Paine 说，“利用其情感分析技术，SAS 软件研究所试图确定每个帖子的情绪是积极的还是消极的。它能够确切地标识为积极的还是消极的只有大约 50 件。情感分析引擎把其余的大约 200 件标识为“中立”。虽然其中一些很可能是真正的中立，但对大多数内容来说，这项技术根本无法确定它的情绪，因此得到了一个默认的‘中性的’分类。”

其结果是，许多公司仍然不信任依靠情感分析技术做出重要决策。

“这项技术只是尚未成熟，” Liquidnet 金融交易公司的商业智能总监 Anglela Chen 说。但是她认为情感分析技术具有很大潜力，可以帮助企业更好地了解他们的客户。

“我会说这将是一个无论在那里也不得不挖掘数据的废墟，” Chen 说，她指的是社交媒体相关数：“我希望厂商对增强这项技术给予更多的关注。”

Forrester 研究公司的 James Kobielus 分析师认为，寻找天才的 BI 和 IT 专业人员去设计和管理社交媒体分析项目也是一个问题。

“这些都不是被广泛采用的，或熟知的，或由传统的专业人员使用的 BI 工具，” Kobielus 说：“因此那些具有所需技能的分析专家可以得到一个不错的薪水。”

然而，尽管它的缺点和高昂成本，情感分析技术在过去几年中得到明显地改善 -- 即便供应商、分析师和客户不认同有多大的改善。

同样地，社交媒体分析有可能随着博客本身不断地扩大其重要性也将继续增长，这意味着对 SAS 和 IBM 来说还是有用武之地的。

然而对客户来讲这也许并不一定是一件好事。

“随着社交媒体向先进的分析应用程序提供更多的内容，假的积极情绪和假的消极情绪将成为一个更大的问题，” Kobielus 说：“而这个还没有被语言学家制定出来。”

（作者：Jeff Kelly 译者：沈宏 来源：TT 中国）

我们的编辑团队

您若有何意见与建议，欢迎与我们的编辑联系。

诚挚感谢以下人员热情参与 TechTarget 中国《数据库电子书》的内容编辑工作！

诚邀更多的数据库专业人士加入我们的内容建设团队！

沈宏
TechTarget中国特邀技术编辑。具有丰富的软件开发及测试经验，多年以来一直致力于数据库优化、性能测试等方向的研究与探索。

宋广磊
TechTarget中国数据库论坛版主。毕业于中科院，获得计算机硕士学位，擅长数据库和数据仓库领域。曾经就职于千橡互动公司，担任高级DBA。

孙瑞
TechTarget 中国数据库网站编辑，三年网络媒体从业经验。负责“TT 数据库”网站的内容建设，熟悉数据库以及商业智能等企业信息化领域，拥有计算机学士学位。