# Apache Mahout 简介

## Apache Mahout 简介

*通过可伸缩、商业友好的机器学习来构建智能应用程序*

Grant Ingersoll, 技术人员, Lucid Imagination

**简介：** 当研究院和企业能获取足够的专项研究预算之后，能从数据和用户输入中学习的智能应用程序将变得更加常见。人们对机器学习技巧（比如说集群、协作筛选和分类）的需求前所未有地增长，无论是查找一大群人的共性还是自动标记海量 Web 内容。Apache Mahout 项目旨在帮助开发人员更加方便快捷地创建智能应用程序。Mahout 的创始者 Grant Ingersoll 介绍了机器学习的基本概念，并演示了如何使用 Mahout 来实现文档集群、提出建议和组织内容。

**本文的标签：** learning, machine, mahout

在信息时代，公司和个人的成功越来越依赖于迅速有效地将大量数据转化为可操作的信息。无论是每天处理数以千计的个人电子邮件消息，还是从海量博客文章中推测用户的意图，都需要使用一些工具来组织和增强数据。这其中就蕴含着*机器学习* 领域以及本文章所介绍项目的前景：Apache Mahout（见 参考资料）。

机器学习是人工智能的一个分支，它涉及通过一些技术来允许计算机根据之前的经验改善其输出。此领域与数据挖掘密切相关，并且经常需要使用各种技巧，包括统计学、概率论和模式识别等。虽然机器学习并不是一个新兴领域，但它的发展速度是毋庸置疑的。许多大型公司，包括 IBM®、Google、Amazon、Yahoo! 和 Facebook，都在自己的应用程序中实现了机器学习算法。此外，还有许多公司在自己的应用程序中应用了机器学习，以便学习用户以及过去的经验，从而获得收益。

在简要概述机器学习的概念之后，我将介绍 Apache Mahout 项目的特性、历史和目标。然后，我将演示如何使用 Mahout 完成一些有趣的机器学习任务，这需要使用免费的 Wikipedia 数据集。

### 机器学习 101

机器学习可以应用于各种目的，从游戏、欺诈检测到股票市场分析。它用于构建类似于 Netflix 和 Amazon 所提供的系统，可根据用户的购买历史向他们推荐产品，或者用于构建可查找特定时间内的所有相似文章的系统。它还可以用于根据类别（体育、经济和战争等）对网页自动进行分类，或者用于标记垃圾电子邮件。本文无法完全列出机器学习的所有应用。如果您希望更加深入地探究该领域，我建议您参阅参考资料。

可以采用一些机器学习方法来解决问题。我将重点讨论其中最常用的两个 — *监管* 和*无监管* 学习 — 因为它们是 Mahout 支持的主要功能。

监管学习的任务是学习带标签的训练数据的功能，以便预测任何有效输入的值。监管学习的常见例子包括将电子邮件消息分类为垃圾邮件，根据类别标记网页，以及识别手写输入。创建监管学习程序需要使用许多算法，最常见的包括神经网络、Support Vector Machines (SVMs) 和 Naive Bayes 分类程序。

无监管学习的任务是发挥数据的意义，而不管数据的正确与否。它最常应用于将类似的输入集成到逻辑分组中。它还可以用于减少数据集中的维度数据，以便只专注于最有用的属性，或者用于探明趋势。无监管学习的常见方法包括 k-Means、分层集群和自组织地图。

在本文中，我将重点讨论 Mahout 当前已实现的三个具体的机器学习任务。它们正好也是实际应用程序中相当常见的三个领域：

- 协作筛选
- 集群
- 分类

在研究它们在 Mahout 中的实现之前，我将从概念的层面上更加深入地讨论这些任务。

### 协作筛选

*协作筛选* (CF) 是 Amazon 等公司极为推崇的一项技巧，它使用评分、单击和购买等用户信息为其他站点用户提供推荐产品。CF 通常用于推荐各种消费品，比如说书籍、音乐和电影。但是，它还在其他应用程序中得到了应用，主要用于帮助多个操作人员通过协作来缩小数据范围。您可能已经在 Amazon 体验了 CF 的应用，如 图 1所示：

**图 1. Amazon 上的协作筛选示例**

CF 应用程序根据用户和项目历史向系统的当前用户提供推荐。生成推荐的 4 种典型方法如下：

- **基于用户**：通过查找相似的用户来推荐项目。由于用户的动态特性，这通常难以定量。
- **基于项目**：计算项目之间的相似度并做出推荐。项目通常不会过多更改，因此这通常可以离线完成。
- **Slope-One**：非常快速简单的基于项目的推荐方法，需要使用用户的评分信息（而不仅仅是布尔型的首选项）。
- **基于模型**：通过开发一个用户及评分模型来提供推荐。

所有 CF 方法最终都需要计算用户及其评分项目之间的相似度。可以通过许多方法来计算相似度，并且大多数 CF 系统都允许您插入不同的指标，以便确定最佳结果。

### 集群

对于大型数据集来说，无论它们是文本还是数值，一般都可以将类似的项目自动组织，或*集群*，到一起。举例来说，对于全美国某天内的所有的报纸新闻，您可能希望将所有主题相同的文章自动归类到一起；然后，可以选择专注于特定的集群和主题，而不需要阅读大量无关

内容。另一个例子是：某台机器上的传感器会持续输出内容，您可能希望对输出进行分类，以便于分辨正常和有问题的操作，因为普通操作和异常操作会归类到不同的集群中。

与 CF 类似，集群计算集合中各项目之间的相似度，但它的任务只是对相似的项目进行分组。在许多集群实现中，集合中的项目都是作为矢量表示在 *n* 维度空间中的。通过矢量，开发人员可以使用各种指标（比如说曼哈顿距离、欧氏距离或余弦相似性）来计算两个项目之间的距离。然后，通过将距离相近的项目归类到一起，可以计算出实际集群。

可以通过许多方法来计算集群，每种方法都有自己的利弊。一些方法从较小的集群逐渐构建成较大的集群，还有一些方法将单个大集群分解为越来越小的集群。在发展成平凡集群表示之前（所有项目都在一个集群中，或者所有项目都在各自的集群中），这两种方法都会通过特定的标准退出处理。流行的方法包括 k-Means 和分层集群。如下所示，Mahout 也随带了一些不同的集群方法。

### 分类

*分类*（通常也称为*归类*）的目标是标记不可见的文档，从而将它们归类不同的分组中。机器学习中的许多分类方法都需要计算各种统计数据（通过指定标签与文档的特性相关），从而创建一个模型以便以后用于分类不可见的文档。举例来说，一种简单的分类方法可以跟踪与标签相关的词，以及这些词在某个标签中的出现次数。然后，在对新文档进行分类时，系统将在模型中查找文档中的词并计算概率，然后输出最佳结果并通过一个分类来证明结果的正确性。

分类功能的特性可以包括词汇、词汇权重（比如说根据频率）和语音部件等。当然，这些特性确实有助于将文档关联到某个标签并将它整合到算法中。

机器学习这个领域相当广泛和活跃。理论再多终究需要实践。接下来，我将继续讨论 Mahout 及其用法