Atitit 知识发现 艾提拉总结

目录

[1. 知识发现](#_Toc20439) [1](#_Toc20439)

[1.1. 与数据挖掘的区别与联系 数据挖掘"的一种更广义的说法 2](#_Toc13330)

[1.2. 基本任务 3](#_Toc23247)

[1.2.1. 数据分类 3](#_Toc19326)

[1.2.2. 数据聚类 3](#_Toc27951)

[1.2.3. 衰退和预报 3](#_Toc1425)

[1.2.4. 关联和相关性 3](#_Toc31229)

[1.2.5. 顺序发现 3](#_Toc562)

[1.2.6. 描述和辨别 3](#_Toc7827)

[1.2.7. 时间序列分析 4](#_Toc14880)

[2. 知识类型 4](#_Toc9820)

[2.1. 1)广义型知识(Generalization)。 4](#_Toc12850)

[2.2. 2)分类型知识(Classification&Clustering)。 4](#_Toc13695)

[2.3. 3)关联型知识(Association)。 4](#_Toc11742)

[2.4. 4)预测型知识(Prediction)。 4](#_Toc6622)

[2.5. 5)偏差型知识(Deviation)。 4](#_Toc695)

[3. 知识发现的分类 5](#_Toc32020)

[3.1. 按挖掘的方法分有数据驱动型、查询驱动型和交互型 5](#_Toc31989)

[3.2. 按知识类型分有关联规则、特征挖掘、分类、聚类、总结知识、趋势分析、偏差分析、文本采掘。 5](#_Toc18261)

[3.3. 知识发现技术可分为两类：基于算法的方法和基于可视化的方法 。 5](#_Toc1367)

[4. 关联技术 5](#_Toc16991)

[4.1. 基于算法的方法和基于可视化 6](#_Toc31438)

[4.2. Web和搜索引擎 6](#_Toc17627)

[4.3. 分类 6](#_Toc1753)

[4.4. 数据结构化 标注 属性标准化化 6](#_Toc12766)

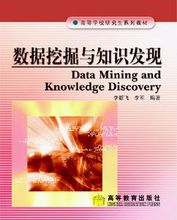
# 知识发现

[编辑](https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/javascript:;) [讨论](https://baike.baidu.com/planet/talk?lemmaId=1407266" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)

本词条缺少****概述图****，补充相关内容使词条更完整，还能快速升级，赶紧来编辑吧！

知识发现是从各种信息中，根据不同的需求获得知识的过程。知识发现的目的是向使用者屏蔽原始数据的繁琐细节，从原始数据中提炼出有效的、新颖的、潜在有用的知识，直接向使用者报告。

知识发现(Knowledge Discovery in Database, KDD)，是所谓"[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)"的一种更广义的说法，即从各种媒体表示的信息中，根据不同的需求获得知识。知识发现的目的是向使用者屏蔽[原始数据](https://baike.baidu.com/item/%E5%8E%9F%E5%A7%8B%E6%95%B0%E6%8D%AE/10119290" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)的繁琐细节，从原始数据中提炼出有意义的、简洁的知识，直接向使用者报告。

[](https://baike.baidu.com/pic/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/1407266/0/964b2e4e300518f6d1c86af3?fr=lemma%26ct=single)相关书籍

基于数据库的知识发现([KDD](https://baike.baidu.com/item/KDD/1407460" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank))和[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)还存在着混淆，通常这两个术语替换使用。KDD表示将低层[数据转换](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E8%BD%AC%E6%8D%A2" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)为高层知识的整个过程。可以将KDD简单定义为：KDD是确定数据中有效的、新颖的、潜在有用的、基本可理解的模式的特定过程。而数据挖掘可认为是观察数据中模式或模型的抽取，这是对数据挖掘的一般解释。虽然数据挖掘是知识发现过程的核心，但它通常仅占KDD的一部分(大约是15% 到25%) 。因此数据挖掘仅仅是整个KDD过程的一个步骤，对于到底有多少步以及哪一步必须包括在KDD过程中没有确切的定义。然而，通用的过程应该接收原始数据输入，选择重要的[数据项](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E9%A1%B9/3227309" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)，缩减、预处理和浓缩数据组，将数据转换为合适的格式，从数据中找到模式，评价解释发现结果

## 与数据挖掘的区别与联系 [数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)"的一种更广义的说法

**中文名**

知识发现

**外文名**

Knowledge Discovery in Database, KDD

**知识发现**

[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)"的一种更广义的说法

**数据分类**

[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)研究的重要分支之一

## 基本任务

[编辑](https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/javascript:;)

### 数据分类

分类是[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)研究的重要分支之一，是一种有效的数据分析方法。分类的目标是通过分析训练数据集，构造一个分类模型(即[分类器](https://baike.baidu.com/item/%E5%88%86%E7%B1%BB%E5%99%A8" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank))，该模型能够把数据库中的[数据记录](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E8%AE%B0%E5%BD%95" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)映射到一个给定的类别，从而可以l立用于数据预测。

### 数据聚类

当要分析的数据缺乏必要的描述信息，或者根本就无法组织成任何分类模式时，利用[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)函数把一组个体按照相似性归成若干类，这样就可以自动找到类。聚类和分类类似，都是将数据进行分组。但与分类不同的是，聚类中的组不是预先定义的，而是根据实际数据的特征按照数据之间的相似性来定义的。

### 衰退和预报

这是一种特殊类型的分类，可以看作是根据过去和当前的数据预测未来的数据状态。通过对用衰减统计技术建模的数字值的预测，学习一种([线性](https://baike.baidu.com/item/%E7%BA%BF%E6%80%A7" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)或非线性)功能将[数据项](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E9%A1%B9/3227309" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)映射为一个数字预测变量。

### 关联和相关性

是指发现大规模数据集中项集之间有趣的关联或相关关系。[关联规则](https://baike.baidu.com/item/%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)是指通过对数据库中的数据进行分析，从某一数据对象的信息来推断另一数据对象的信息，寻找出重复出现概率很高的知识模式，常用一个带有[置信度](https://baike.baidu.com/item/%E7%BD%AE%E4%BF%A1%E5%BA%A6" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)因子的参数来描述这种不确定的关系。

### 顺序发现

通常指确定数据组中的顺序模式。当数据的特定类型的关系已被发现时，这些模式同关联和相关性相似。但对关系基于[时间序列](https://baike.baidu.com/item/%E6%97%B6%E9%97%B4%E5%BA%8F%E5%88%97" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)的数据组，顺序发现和关联就不同了。概括总结：顺序发现是将数据映射为有关数据组的简练描述的子集或映射为数据库中一组特定用户数据的高度概括的数据。

### 描述和辨别

是指发现一组特征规则，其中的每一条都是或者显示数据组的特征或者从对比类中区别试验类的概念的命题。

### 时间序列分析

其任务是发现[属性值](https://baike.baidu.com/item/%E5%B1%9E%E6%80%A7%E5%80%BC" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)的发展趋向，如从股票价格指数的金融数据、客户数据和医学数据等。它是用来搜寻相似模式以发现和预测特定模式的风险、因果关系和趋势。

# 知识类型

[编辑](https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/javascript:;)

## 1)广义型知识(Generalization)。

是根据数据的微观特性发现其表征的、带有普遍性的、高层次概念的、中观或宏观的知识。

## 2)分类型知识(Classification&Clustering)。

反映同类事物共同性质的特征型知识和不同事物之间差异型特征知识。用于反映数据的汇聚模式或根据对象的属性区分其所属类别。

## 3)关联型知识(Association)。

是反映一个事件和其他事件之间依赖或关联的知识，又称依赖(Dependency)关系。这类知识可用于数据库中的[归一化](https://baike.baidu.com/item/%E5%BD%92%E4%B8%80%E5%8C%96" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)，查询优化等。

## 4)预测型知识(Prediction)。

通过[时间序列](https://baike.baidu.com/item/%E6%97%B6%E9%97%B4%E5%BA%8F%E5%88%97/1389644" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)型数据，由历史的和当前的数据去预测未来的情况。它实际上是一种以时间为关键属性的关联知识。

## 5)偏差型知识(**[Deviation](https://baike.baidu.com/item/Deviation/10802284" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)**)。

通过分析标准类以外的特例、[数据聚类](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E8%81%9A%E7%B1%BB" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)外的[离群值](https://baike.baidu.com/item/%E7%A6%BB%E7%BE%A4%E5%80%BC" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)、实际[观测值](https://baike.baidu.com/item/%E8%A7%82%E6%B5%8B%E5%80%BC/558500" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)和[系统预测](https://baike.baidu.com/item/%E7%B3%BB%E7%BB%9F%E9%A2%84%E6%B5%8B/12609292" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)值间的显著差别，对差异和极端特例进行描述。

# 知识发现的分类

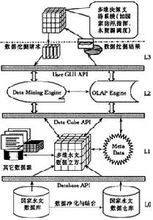
## 按挖掘的方法分有数据驱动型、查询驱动型和交互型

知识发现已经出现了许多知识发现技术，[分类方法](https://baike.baidu.com/item/%E5%88%86%E7%B1%BB%E6%96%B9%E6%B3%95" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)也有很多种，按被挖掘对象分有基于[关系数据库](https://baike.baidu.com/item/%E5%85%B3%E7%B3%BB%E6%95%B0%E6%8D%AE%E5%BA%93" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)、[多媒体数据库](https://baike.baidu.com/item/%E5%A4%9A%E5%AA%92%E4%BD%93%E6%95%B0%E6%8D%AE%E5%BA%93/5389699" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)；按挖掘的方法分有数据驱动型、查询驱动型和交互型；

## 按知识类型分有[关联规则](https://baike.baidu.com/item/%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)、特征挖掘、分类、[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)、总结知识、趋势分析、偏差分析、文本采掘。

## 知识发现技术可分为两类：基于算法的方法和基于可视化的方法 。

大多数基于算法

[](https://baike.baidu.com/pic/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/1407266/0/e865a699d21ac1346e068cfd?fr=lemma%26ct=single)知识发现

的方法是在人工智能、信息检索、数据库、统计学、模糊集和[粗糙集理论](https://baike.baidu.com/item/%E7%B2%97%E7%B3%99%E9%9B%86%E7%90%86%E8%AE%BA" \t "https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E5%8F%91%E7%8E%B0/_blank)等领域中发展来的

、

# 关联技术

## 基于算法的方法和基于可视化

## Web和搜索引擎

## 分类

## 数据结构化 标注 属性标准化化