frlily / July 21, 2025 02:09PM

资料探勘的10大演算法我用大白话讲清楚了，新手一看就懂

一個優秀的資料分析師，除了要掌握基本的統計學、資料庫、資料分析方法、思維、資料分析工具技能之外，還需要掌握一些資料探勘的思想，幫助我們挖掘出有價值的資料，這也是資料分析專家和一般資料分析師的差距之一。

資料探勘主要分為分類演演算法，聚類演演算法和關聯規則三大類，這三類基本上涵蓋了目前商業市場對演演算法的所有需求。而這三類裡又包含許多經典演演算法。市面上很多關於資料探勘演演算法的介紹深奧難懂，今天就給大家用簡單的大白話來介紹資料探勘十大經典演演算法原理，幫助大家快速理解。
演演算法分類

連線分析：PageRank

關聯分析：Apriori

分類演演算法：C4.5，樸素貝葉斯，SVM，KNN，Adaboost，CART

聚類演演算法：K-Means，EM

1. 一、PageRank

當一篇論文被引用的次數越多，證明這篇論文的影響力越大。
一個網頁的入口越多，入鏈越優質，網頁的質量越高。

1. 原理

網頁影響力=阻尼影響力+所有入鏈集合頁面的加權影響力之和

1.一個網頁的影響力：所有入鏈的頁面的加權影響力之和。

2.一個網頁對其他網頁的影響力貢獻為：自身影響力/出鏈數量。

3.使用者並不都是按照跳轉連結的方式來上網，還有其他的方式，比如直接輸入網址訪問。

4.所以需要設定阻尼因子，代表了使用者按照跳轉連結來上網的機率。

2. 比喻說明

1.微博:一個人的微博粉絲數不一定等於他的實際影響力，還需要看粉絲的質量如何。如果是殭屍粉沒什麼用，但如果是很多大V或者明星關注，影響力很高。

2.店鋪的經營顧客比較多的店鋪質量比較好，但是要看看顧客是不是託。

3.興趣在感興趣的人或事身上投入了相對多的時間，對其相關的人事物也會投入一定的時間。那個人或事，被關注的越多，它的影響力/受眾也就越大。

3. 關於阻尼因子

1.透過你的鄰居的影響力來評判你的影響力，但是如果不能透過鄰居來訪問你，並不代表你沒有影響力，因為可以直接訪問你，所以引入阻尼因子的概念。

2.海洋除了有河流流經，還有雨水，但是下雨是隨機的。

3.提出阻尼係數，還是為了解決某些網站明明存在大量出鏈（入鏈），但是影響力卻非常大的情形。

（1）出鏈例子：hao123導航網頁，出鏈極多入鏈極少。

（2）入鏈例子：百度谷歌等搜尋引擎，入鏈極多出鏈極少。

2. 二、Apriori（關聯分析）

關聯關係挖掘，從消費者交易記錄中發掘商品與商品之間的關聯關係。

1. 原理

1.支援度:某個商品組合出現的次數與總次數之間的比例。
5次購買，4次買了牛奶，牛奶的支援度為4/5=0.8。5次購買，3次買了牛奶+麵包，牛奶+麵包的支援度為3/5=0.6。

2.置信度:購買了商品A，有多大機率購買商品B，A發生的情況下B發生的機率是多少。買了4次牛奶，其中2次買了啤酒，(牛奶->啤酒)的置信度為2/4=0.5。買了3次啤酒，其中2次買了牛奶，(啤酒->牛奶)的置信度為2/3-0.67。

3.提升度:衡量商品A的出現，對商品B的出現機率提升的程度。提升度(A->=置信度(A->/支援度(。提升度>1，有提升；提升度=1，無變化；提升度