Web数据挖掘

Web挖掘的目标是从Web的超链接结构、网页内容和使用日志中探寻有用的信息。虽然Web挖掘使用了许多数据挖掘技术，但它不仅仅是传统数据挖掘的一个简单的应用。在过去的20年中，许多新的挖掘任务和算法被相继提出。依据在挖掘过程中使用的数据类别，Web挖掘任务可以分为三种类型：Web结构挖掘、Web内容挖掘和Web使用挖掘。

1. Web结构挖掘
   ：Web结构挖掘从表征Web结构的超链接中寻找有用的知识。例如：从这些链接中，我们可以找出哪些是重要的网页，这是一项搜索引擎采用的重要技术。我们也可以发掘具有共同兴趣的用户社区。这些任务在传统的数据挖掘中并不存在，因为在关系型表格中没有链接结构。
2. Web内容挖掘
   ：Web内容挖掘从网页内容中抽取有用的信息和知识。例如：根据网页的主题，我们可以进行自动的聚类和分类。虽然这些任务与传统数据挖掘的任务相似，但是我们依然可以为了各种不同的目的从网页中根据模式抽取有用的信息，例如商品描述、论坛回帖等。这些信息可以被用作进一步分析挖掘用户态度，这些任务也不是传统的数据挖掘任务。
3. Web使用挖掘
   ：Web使用挖掘从记录每一位用户点击情况的使用日志中挖掘使用的访问模式。这项任务也使用了许多数据挖掘的算法。其中一项重要的议题是点击流数据的预处理，以便生成可以用来挖掘的合适数据。

Web挖掘过程和数据挖掘过程十分相似，区别通常只是数据收集。在传统数据挖掘中，这些数据经常是收集并存储在数据仓库中的；而对于Web挖掘而言，数据的收集是一项艰巨的任务，实际上，Google创始人Sergey Brin和Lawrence Page在他们一篇颇具影响力的论文（The anatomy of a large-scale hypertextual Web search engine,1998）中指出：收集的收集是Web搜索引擎最薄弱而复杂的模块。在进行Web结构挖掘和内容挖掘的时候，需要爬取大量的网页。一旦数据收集完毕，我们就可以进行相同的三步工作了，那就是：数据预处理、Web数据挖掘和数据后续处理。但是每一步涉及到的具体技术又会与传统的数据挖掘大相径庭。