1. **Khai phá dữ liệu là gì? Hãy phân biệt khái niệm khám phá tri thức trong cơ sở dữ liệu và khai phá dữ liệu?**

Khai phá dữ liệu là quá trình tìm kiếm, phát hiện các tri thức mới, hữu ích tiềm ẩn trong cơ sở dữ liệu lớn.

Khám phá tri thức trong CSDL (Knowledge Discovery in Databaes – KDD) là mục tiêu chính của khai phá dữ liệu, do vậy hai khái niệm khai phá dữ liệu và KDD được các nhà khoa học xem là tương đương nhau. Thế nhưng, nếu phân chia một cách chi tiết thì khai phá dữ liệu là một bước chính trong quá trình KDD.

Khám phá tri thức trong CSDL là lĩnh vực liên quan đến nhiều ngành như: Tổ chức dữ liệu, xác suất, thống kê, lý thuyết thông tin, học máy, CSDL, thuật toán, trí tuệ nhân tạo, tính toán song song và hiệu năng cao,... Các kỹ thuật chính áp dụng trong khám phá tri thức phần lớn được thừa kế từ các ngành này.

2. **Hãy cho biết những thách thức và các loại dữ liệu trong khai phá dữ liệu**

Những thách thức:

- Số đối tượng trong cơ sở dữ liệu thường rất lớn
- Số chiều (thuộc tính) của cơ sở dữ liệu lớn
- Dữ liệu và tri thức luôn thay đổi có thể làm cho các mẫu đã phát hiện không còn phù hợp.
- Dữ liệu bị thiếu hoặc nhiễu
- Quan hệ giữa các thuộc tính phức tạp
- Giao tiếp với người sử dụng và kết hợp với các tri thức đã có.
- Tích hợp với các hệ thống khác…

Các loại dữ liệu có thể khai phá (về cơ bản, khai phá dữ liệu có thể ứng dụng cho bất kỳ kho thông tin nào), ví dụ:

- Các cơ sở dữ liệu quan hệ
- CSDL đa chiều
- CSDL không gian và thời gian
- CSDL đa phương tiện
- Các cơ sở dữ liệu giao tác
- Các hệ thống cơ sở dữ liệu tiên tiến
- …

3. **Trình bày các loại cơ sở dữ liệu dạng Text? Tại sao việc phân loại trên không thật rõ ràng?**

Các loại cơ sở dữ liệu dạng Text:

- Dạng không có cấu trúc (unstructured): Những văn bản thông thường mà chúng ta thường đọc hàng ngày được thể hiện dưới dạng tự nhiên của con người và nó không có một cấu trúc định dạng nào. Ví dụ: tập hợp sách, tạp chí, bài viết được quản lý trong một mạng thư viện điện tử.
- Dạng nửa cấu trúc (semi-structured): Những văn bản được tổ chức dưới dạng cấu trúc không chặt chẽ như bản ghi các ký hiệu đánh dấu văn bản và vẫn thể hiện được nội dung chính của văn bản, ví dụ như các dạng HTML, email,…

Tuy nhiên việc phân làm hai loại cũng không thật rõ ràng, trong các hệ phần mềm, người ta thường phải sử dụng các phần kết hợp lại để thành một hệ như trong các hệ tìm tin (Search Engine), hoặc trong bài toán tìm kiếm văn bản (Text Retrieval), một trong những lĩnh vực qua tâm nhất hiện nay. Chẳng hạn trong hệ tìm kiếm như Yahoo, Altavista, Google... đều tổ chức dữ liệu theo các nhóm và thư mục, mỗi nhóm lại có thể có nhiều nhóm con nằm trong đó. Hệ Altavista còn tích hợp thêm chương trình dịch tự động có thể dịch chuyển đổi sang nhiều thứ tiếng khác nhau và cho kết quả khá tốt.

4. **Trình bày khái niệm về khai phá Web ? Hãy cho biết các quá trình khai phá Web?**

Khai phá Web là việc sử dụng các kỹ thuật KPDL để tự động hóa quá trình khám phá và trích rút những thông tin hữu ích từ các tài liệu, các dịch vụ và cấu trúc Web.

Các quá trình khai phá Web:

- Tìm kiếm nguồn tài nguyên: Thực hiện tìm kiếm và lấy các tài liệu Web phục vụ cho việc khai phá.
- Lựa chọn và tiền xử lý dữ liệu: Lựa chọn và tiền xử lý tự động các loại thông tin từ nguồn tài nguyên Web đã lấy về.
- Tổng hợp: Tự động khám phá các mẫu chung tại các Web site riêng lẻ cũng như nhiều Website với nhau.
- Phân tích: Đánh giá, giải thích, biểu diễn các mẫu khai phá được.

5. **Hãy cho biết chức năng của các nhóm chính khi tiến hành phân chia các kỹ thuật khai phá dữ liệu?**

Các kỹ thuật khai phá dữ liệu thường được chia làm 2 nhóm chính:

- Kỹ thuật mô tả: Bao gồm các kỹ thuật mô tả các tính chất hoặc các đặc tính chung của dữ liệu trong CSDL hiện có. Các kỹ thuật này gồm có: phân cụm (clustering), tóm tắt (summerization), trực quan hóa (visualiztion), phân tích tiến hoá và độ lệch (Evolution and deviation analysis), phân tích luật kết hợp (association rules analysis)…
- Kỹ thuật dự đoán: Có nhiệm vụ đưa ra các dự đoán dựa vào các suy diễn trên dữ liệu hiện thời. Các kỹ thuật này gồm: Phân lớp (classification), hồi quy (regression), … .

Các kỹ thuật khai phá:

- Phân lớp và dự đoán (classification and prediction): Là việc xếp các đối tượng vào những lớp đã biết trước. Ví dụ, phân lớp các bệnh nhân, phân lớp các loài thực vật, ... Hướng tiếp cận này thường sử dụng một số kỹ thuật của học máy như cây quyết định (decision tree), mạng nơ-ron nhân tạo (neural network), ... Phân lớp và dự đoán còn được gọi là học có giám sát.
- Phân cụm (clustering/segmentation): Là việc xếp các đối tượng theo từng cụm tự nhiên.
- Luật kết hợp (association rules): Là việc phát hiện các luật biểu diễn tri thức dưới dạng đơn giản. Ví dụ: “70% nữ giới vào siêu thị mua phấn thì có tới 80% trong số họ cũng mua thêm son”.
- Phân tích hồi quy (regression analysis): Là việc học một hàm ánh xạ mỗi bộ tập dữ liệu thành một giá trị thực của biến dự đoán. Nhiệm vụ của phân tích hồi quy tương tự như của phân lớp, điểm khác nhau là ở chỗ thuộc tính dự báo là liên tục chứ không phải rời rạc.
- Phân tích các mẫu theo thời gian (sequential/temporal patterns): Tương tự như khai phá luật kết hợp nhưng có quan tâm đến tính thứ tự theo thời gian.
- Mô tả khái niệm (concept description and summarization): Thiên về mô tả, tổng hợp và tóm tắt các khái niệm. Ví dụ tóm tắt văn bản.

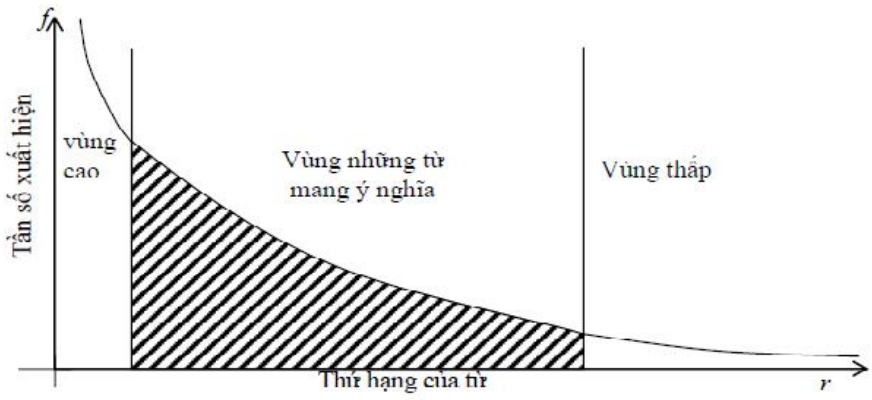
6. Bài tập Zipf

B1: Sắp xếp các từ theo chiều giảm dần của tần số xuất hiện

B2: Sử dụng công thức tính thứ hạng của từng từ r=K/f

- K là hằng số, K = N/10 với N là tổng số từ
- f là tần số xuất hiện

B3: Vẽ biểu đồ thống kê tuần số của từ, trục hoành là thứ hạng, tung là tần số



Lược đồ thống kê tần số của từ theo Định luật Zipf

B4: Nhận xét: Những từ có tần số nhỏ ảnh hưởng rất ít tới toàn bộ văn bản.

7. PageRank

B1: Thứ hạng của 1 tài liệu = tổng giá trị của các liên kết trỏ tới nó.

B2: Sắp xếp theo chiều giảm dần

B3: Nhận xét: PageRank càng cao thì tài liệu sẽ được trả về đầu tiên trong kết quả tìm kiếm.

8. HITS

B1: Xác định tập gốc R (các kq trả về từ Search Engine), khởi tạo tập cơ bản S tương ứng với R, thêm vào S tất cả các trang được trỏ tới từ bất kỳ trang nào trong R.

B2: Xác định Authority (= tổng số liên kết trỏ đến), Hub (tổng số liên kết đi ra)

B3: Vẽ đồ thị

B4: Nhận xét: Các trang có Authority lớn thường cung cấp các thông tin quan trọng, nội dung tin cậy. Một Authority tốt nết nó được trỏ đến bởi nhiều Hub tốt và ngược lại.

Giá trị trọng số các Hub và Authority