**CHƯƠNG 2 : BÀI TOÁN RECOMMENDER SYSTEM**

2.1 Giới thiệu

**Recommender Systems**: Là các hệ thống đề xuất,gợi ý thông tin tới người dùng dựa trên mối quan tâm của họ về một lĩnh vực nào đó hoặc mức độ liên quan của các thông tin mà họ đưa ra. Recommender Systems được sử dụng rỗng rãi trong thực tế như gợi ý phim , gợi ý bài viết , gợi ý đia điểm ăn uống ,du lịch , …

2.1.1 Hai nhóm chính của recommendation system

Các Recommendation Systems thường được chia thành hai nhóm lớn:

1. **Content-based systems**: đánh giá đặc tính của items được recommended. Ví dụ: một user xem rất nhiều các bộ phim về cảnh sát hình sự, vậy thì gơi ý một bộ phim trong cơ sở dữ liệu có chung đặc tính hình sự tới user này, ví dụ phim Người phán xử. Cách tiếp cận này yêu cầu việc sắp xếp các items vào từng nhóm hoặc đi tìm các đặc trưng của từng item. Tuy nhiên, có những items không có nhóm cụ thể và việc xác định nhóm hoặc đặc trưng của từng item đôi khi là bất khả thi.
2. **Collaborative filtering**: hệ thống gợi ý items dựa trên sự tương quan (similarity) giữa các users và/hoặc items. Có thể hiểu rằng ở nhóm này một item được recommended tới một user dựa trên những users có hành vi tương tự. Ví dụ: users A, B, C đều thích các bài hát của Noo Phước Thịnh. Ngoài ra, hệ thống biết rằng users B, C cũng thích các bài hát của Bích Phương nhưng chưa có thông tin về việc liệu user A có thích Bích Phương hay không. Dựa trên thông tin của những users tương tự là B và C, hệ thống có thể dự đoán rằng A cũng thích Bích Phương và gợi ý các bài hát của ca sĩ này tới A.

Document-term matrix hoặc term-document matrix là một ma trận số biểu diễn tần suất xuất hiện các từ trong một tập các văn bản (collection). Trong ma trận term-document matrix, hàng tương ứng với từng văn bản trong collection đó và cột tương ứng với mỗi từ thuộc các văn bản đó. Có nhiều cách khác nhau để xác định từng giá trị trong ma trận này, một trong số những cách phổ biến là sử dụng TF-IDF

Mô hình chung:

Giả sử ta có 2 văn bản

* D1 = "I like databases"
* D2 = "I hate databases",

Khi đó, term-document matrix chỉ ra văn bản này chứa những từ nào và tần suất xuất hiện của các từ đó là bao nhiêu:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | I | like | hate | database |
| D1 | 1 | 1 | 0 | 1 |
| D2 | 1 | 0 | 1 | 1 |

Chú ý rằng ngoài tần suất, ta có thể đánh trọng số cho các từ này theo cách khác, ví dụ như ***tf-idf***