ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH



**BÁO CÁO THỰC TẬP TỐT NGHIỆP**

**ĐỀ TÀI**

**Recommendation Systems.**

Giáo viên hướng dẫn: PGS. TS Quản Thành Thơ

Nhóm sinh viên thực hiện: Lê Tuấn Anh - 51300080

TP. Hồ Chí Minh, tháng 4 năm 2017

Contents

[1. Overview 3](#_Toc480822979)

[1.1: Stereotyping 4](#_Toc480822980)

[1.2: Content-based Filtering(CBF) 4](#_Toc480822981)

[1.3: Collaborative Filtering(CF) 4](#_Toc480822982)

[1.4: Co-occurrence recommendations 4](#_Toc480822983)

[1.5: Graph based 4](#_Toc480822984)

[1.6: Global relevance 4](#_Toc480822985)

[1.7: Hybrid 4](#_Toc480822986)

[2. Content-based filtering(CBF) 5](#_Toc480822987)

[2.1: Item Profiles 6](#_Toc480822988)

[2.2: Features. 6](#_Toc480822989)

[2.3: Representing Item Profiles 6](#_Toc480822990)

[2.4: Users Profiles 7](#_Toc480822991)

[2.5: Recommending Items to Users Based on Content 7](#_Toc480822992)

[2.6: Classification Algorithms: 7](#_Toc480822993)

[3. Collaborative Filtering(CF) 8](#_Toc480822994)

[3.1: Measuring Similarity 9](#_Toc480822995)

[3.1.1: Jaccard Distance 9](#_Toc480822996)

[3.1.2: Cosine Distance 9](#_Toc480822997)

[3.1.3: Rounding Data 9](#_Toc480822998)

[4. Comparing between Content-based and Collaborative Filtering 11](#_Toc480822999)

**RECOMMENDATION SYSTEMS**

1. **Overview**

\_ Recommendation systems changed the way inanimate websites communicate with their users.

\_ Increase interaction to provide a richer experience.

\_ Recommedation systems:

+ Customers of an on-line retailer suggestions about what they might like to buy, based on their past history of purchases and/or product searches.

\_ Recommendation concepts and approaches :

Recommendation Systems

## **1.1: Stereotyping**

* Allow the definition of a set of differentiating characteristics for a group of users; when a new user is introduced into the system, they can be assigned to a predefined stereotype, based on their personal data, which allows the activation of a set of default preferences.
* Another way to understand this approach is: define an abstract user that has general properties similar to a set (community) of real users.
* Used for varying purposes, ranging from initial user profile creation to generating recommendations.

## **1.2: Content-based Filtering(CBF)**

* Users ’s interests based upon the contents of the papers and a profile of the user ‘s interests.

## **1.3: Collaborative Filtering(CF)**

* Collaborative Filtering methods identify similarity between users based on items they have rated and recommend new items similar users have liked.

## **1.4: Co-occurrence recommendations**

* The items those are recommended frequently co-occur with some source items.
* Distance measure between tags it is straightforward to derive methods to analyze user interest and compute recommendations.

## **1.5: Graph based**

* In the data is represented in the form of a graph in graph based approach where nodes are users, items or both, and edges encode the interactions or similarities between the items and users. In this method various similarity approaches are used.
* Distance between two nodes is valuated as:
  + Path-based similarity: a function of the number of paths connecting the two nodes, as well as the length of these paths.
  + Random walk similarity: a probability of reaching these nodes in a random walk.
    - Which a random walk is actually a transition from a vertex to another vertex.
* Item rank: approach ranks the preferences of a user u for new items i as the probability of u to visit i in a random walk of a graph in which nodes correspond to the items of the system, and edges connect items that have been rated by common users.

## **1.6: Global relevance**

* One fits all approach and recommends items that have the highest global relevance.

## **1.7: Hybrid**

* To get better results, an approach of hybrid is combining different techniques of CF and CBF.

1. **Content-based filtering(CBF)**

* The user modeling process, in which the interests of users are inferred from the items that users interacted with.
* ’Interaction’ is typically established through actions, such as downloading, buying, authoring, or tagging an item.
* Items are represented by a content model containing the items’ features. Features are typically word-based, i.e. single words, phrases, or n-grams.
* **Advantages:**
* Allows a user-based personalization => the recommender system can determine the best recommendations for each user individually, rather than being limited by stereotypes.
* Requires less up-front classification work, since user models can be created automatically.
* **Disadvantages:**
  + - Requires more computing power than stereotyping.
      * Ex: Each item must be analyzed for its features, user models must be built, and similarity calculations must be performed.
* **Content description**: In some domains generating a useful description of the content can be very difficult.
  + Ex: video, music.
    - Low serendipity(may mắn) and **overspecialization**(vượt quá) leading it to recommend items as similar as possible to the ones a user already knows.
      * Not select items if the previous user behavior does not provide evidence.
    - Ignores quality and popularity of items.
    - Dependent on access to the item’s features.
    - **Subjective domain:** difficulty in distinguishing between subjective information such as points of views and humor.

**Example: Film saw today: **

**Recommended film:** 

## **2.1: Item Profiles**

* Construct for each item a profile, which is a record or collection of records representing important characteristics of that item.
* Products often have descriptions written by the manufacturer, giving features relevant to that class of product.

## **2.2: Features.**

* Values of features may be document collections and images.
* **With documents:**
* There are many kinds of documents (ex: articles).
* To distinguish among topics, to characterize the topics of documents:
  + - For the remaining words, compute the TF-IDF score.
    - The highest scores of those words are words that characterize the documents => take features of ducuments.
* **With images:**
* Instead of calculating properties of pixel of images, inviting users to tag the items by entering words or phrases to describe them.
* ‘Tag’ make a new method of search available, where users entered a set of tags as their search query, and the system retrieved the Web pages that had been tagged that way.
* But the problem of this way is only work if users accept to create the tags.

## **2.3: Representing Item Profiles**

* Goal for content-based recommendation is to create both an item profile consisting of feature-value pairs and a user profile summarizing the preferences of the user, based of their row of the utility matrix.

## **2.4: Users Profiles**

* Besides create vectors describing items, also create vectors with the same components that describe the user’s preferences.
* To connect between items and users, using matrix.
* Weight the vectors representing the profiles of items by the utility value, then subtracting the average value for a user.

## **2.5: Recommending Items to Users Based on Content**

* With profile vectors for both users and items, estimate the degree to which a user would prefer an item by computing the cosine distance between the user’s and item’s vectors.

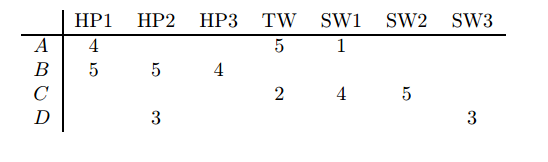
## **2.6: Classification Algorithms:**

* There are lots of algorithms to classify, one of those are using decision tree for classifying that predicts the ratings of all items.

1. **Collaborative Filtering(CF)**

* The concept of CF was that users like what like-minded users like, where two users were considered like-minded when they rated items alike.
* When like-minded users were identified, items that one user rated positively were recommended to the other user, and vice versa.
* The **advance** of CF compared with CBF:
  + CF is content independent, no error-prone(the error-prone hardware is limited to calculation units, a small fraction of processor die area) item processing is required.
  + Because humans do the ratings, CF takes into account real quality assessments.
  + CF is supposed to provide serendipitous(the fact of finding interesting or valuable things by chance) recommendations because recommendations are not based on item similarity but on user similarity.
* **Disadvantages**:
  + ‘Cold-start’ problem:
    - If a new user rates few or no items, the system cannot find like-minded users and therefore cannot provide recommendations.
    - If an item is new in the system and has not been rated yet by at least one user, it cannot be recommended.
    - In a new community, no users have rated items, so no recommendations can be made and as a result, the incentive for users to rate items is low.
* **Solutions:** implicit ratings may be inferred from the interactions between users and items.
  + - Sparsity(thưa thớt).
    - Computing time for CF tends to be higher than for content-based filtering.
* Items profiles and user profiles vector represent by columns ,rows.

**3.1: Measuring Similarity**

****

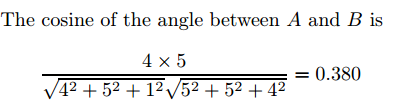
**Fig 1: The utility matrix representating ratings of movies on a 1-5 vscale**

**3.1.1: Jaccard Distance**

* Find intersections and unions of sets of items rated(constraints condition, unless distance loses important information).
* After having intersection and union of each item rated, distance equal 1-count(intersection)/count(union).
* Ex: A&B: 1 /5, A&C: 2/4

### 3.1.2: Cosine Distance

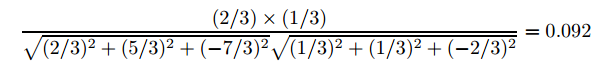
* Set blank as a 0 value.

****

**3.1.3: Rounding Data**

* Try to eliminate the apparent similarity between movies a user rates highly and those with low scores by rounding the ratings

**3.1.4: Normalizing Ratings**

* Subtracting from each rating the average rating of that user, turn low ratings into negative numbers and high ratings into positive numbers
* Cosine of the angle between A and B:****

This Equation based on Figure 2 shows the matrix of Figure 1

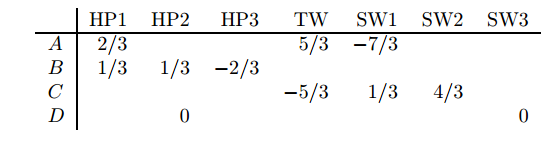


Fig 2: The utility matrix is introduced in Fig 1

# 4. Comparing between Content-based and Collaborative Filtering

|  |  |  |
| --- | --- | --- |
|  | Content-based Filtering | Collaborative Filtering |
| **Attributes of an item** | Belong to | Ignore |
| **Purchase Information** | Ignore | Belong to |
| **Systems** | A rated set of item attributes | A vector of rated items |
| **Problems** | \_ Content description  \_ Over-specialization  \_ Subjective domain problem | Cold-stat |
| **Advantages** | \_ Easily provide valid recommendations to new users, even if they never used the system before.  \_ Useful in environments where new items are constantly added. | \_ No error-prone.  \_ Account real quality assessments. |
| **Computing time** | Slow | Fast |
| **Background** | Features of items in I | Ratings from U of items in I |
| **Process** | Generate a classifier that fits u’s rating behavior and use it on i. | Identify users in U similar to u, and extrapolate(guess based on information) from their ratings of i. |