

# 商品推薦系統

## -基於內容推薦

### Lecture 3

講師：周光宇博士

2018.08

# 音樂基因組項目（潘多拉） Music Genome Project (Pandora)

- 一項“以最基本的層次捕捉音樂本質”的成就，使用超過450種音樂特徵來描述歌曲，並採用複雜的數學算法來組織它們。
- 目前由5個子基因組組成：流行/搖滾，嘻哈/電子樂，爵士，世界音樂和古典音樂。
- 每一首歌曲由包含大約450個“基因”值的向量表示，諸如主唱歌手的性別，持續重複的節奏模式，電吉他上的失真(distortion)程度，背景聲音的類型等等。
- 每個基因被分配一個0到5之間的數字，以半個整數(0.5)為單位。使用公司所稱的“匹配算法”，以一個或多個歌曲的向量，來建立其他類似歌曲的列表。
- 音樂基因組項目是潘多拉(Pandora)使用的核心技術，該公司根據用戶喜好，以基於內容的方法推薦的在線上廣播平台播放音樂。。



Google News

Secure | https://news.google.com.sg/nwshp?hl=en&tab=wn&ei=2AHpWom8MlyHvQT4tYCoAw&ved=0EKkuCAYoBQ

Norton


THIS PAGE IS SAFE

SHOPPING GUARANTEE

ACCESS VAULT

SHARE VIA FACEBOOK


Business »



TODAYonline - 9 hours ago

**45 fake gold bars seized, six men arrested**


SINGAPORE - Forty-five fake gold bars were seized and six men were arrested for trying to sell the counterfeit items, said the Singapore Police in a press release on Saturday (April 8).



Channel NewsAsia - 5 hours ago

**Porsche-Piech clan to stay out of VW management: Porsche chairman**


Members of the Porsche-Piech clan that controls Volkswagen will no longer be eligible to serve as executives of the carmaker, Porsche Automobil Holding SE Chairman Wolfgang Porsche told a German newspaper.



The Online Citizen - 8 hours ago

**Domestic worker runs away from employer after subjected to slave like treatment at flat in Sengkang West**

Domestic worker stands between a locked gate and the door of a flat that she doesn't want to return to. Domestic worker runs away from employer after subjected to slave like treatment at flat in Sengkang West.



Business Standard - 2 hours ago


**Cuban winery uses condoms for production**

A 65-year-old Cuban has discovered an unusual way to help ferment home-made tropical fruit wines. At his house in the southern Havana neighbourhood of El Canal, Orestes Estevez and his family fill glass jugs with wine made from grapes, raisins, beets, ...

NewsX

[More Business stories](#)


Technology »



Vox - 29 minutes ago

**Google's epic legal battle with Uber over self-driving technology, explained**


Uber and Google are locked in a legal battle that could have huge implications for the future of the self-driving car industry. If Uber loses its lawsuit, it could cost the company millions and set back Uber's self-driving car effort by months ...



ValueWalk - 2 hours ago

**Samsung Galaxy S8 vs. Galaxy S7: Specs Comparison**


Samsung is hoping to convince consumers to upgrade from the Galaxy S7 to its new Galaxy S8 flagship. But how does the new device in this smartphone range compare to its cousin, and what do you get for the extra investment?



Hot Hardware - 51 minutes ago

**Google Makes Fact Check Tool For Search Available Globally To Stomp Out Fake News**

One of the biggest problems on the web is the proliferation of fake news. Online services and Internet users alike face the same challenge—trying to discern between what is a legitimate news story and one that is outright false (or a satire).



Financial Express - 6 hours ago

**Broadband labelling crucial to empower consumer experience: TRAI**

Addressing the conference, Ram Sevak Sharma, Chairman, TRAI said the quality of broadband services in India remains a huge concern.

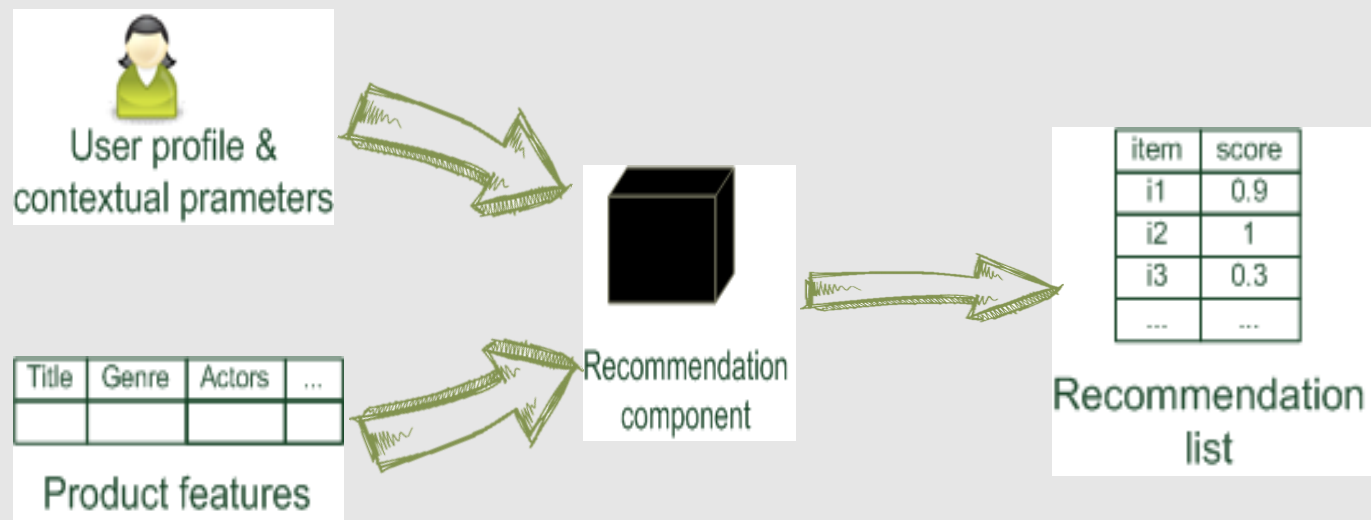
Financial Ex...

[More Technology stories](#)

# 基於內容的推薦系統

- 基於內容的推薦系統利用商品介紹內的描述性特徵來設計。
- 這種推薦系統使用用戶過去喜歡的商品之商品特徵來將用戶與用戶可能喜歡的商品進行匹配。
- 在這種情況下，不需要利用其他用戶的評分來提出推薦。 例如，用戶對於自己所觀看的電影其本身的評分和行動足以發現有意義的推薦。
- 這個方法在新商品評分很少的時候是特別有用的。
- 通常，文字分類和迴歸模型方法仍然是建立基於內容的推薦系統之最廣泛使用的工具。

"show me more of the same what I've liked"



# 基於內容的推薦系統的特點



## 優點

- 只要有足夠的用戶偏好資訊，就可以幫助緩解冷啟動(cold-start)問題。
- 只要可以摘錄新商品的特徵，就可以對新商品進行推薦。



## 限制

- 無法解決新用戶的冷啟動問題。
- 因為不使用其他用戶的評價，推薦商品的多樣性和新穎性會減少。因此，推薦的商品對於用戶來說可能是已經知道並感興趣的，或者是用戶之前已經消費過的其他商品。

基於內容的推薦系統特別適合在文本豐富和非結構化的領域提供推薦

# 應用

- 基於內容的系統主要用於有大量商品特徵訊息、資訊的場景。
- 事實上，多數的基於內容系統從商品敘述中摘錄文字特徵。
- 在許多情況下，這些特徵是從商品描述中提取的關鍵字。
- 因此，基於內容的系統特別適合在文本豐富和非結構化的領域提供推薦。
- 一個經典的例子是在網頁的推薦。例如，用戶以前的瀏覽行為可用於建立基於內容的推薦。
- 在這種情況下，來自商品描述的關鍵字被用於建立商品特徵和用戶個人興趣資料 (User Profile)。另外，除了關鍵字之外，還可以使用諸如製造商、類別和價格的關係屬性。

# 數據來源

- 第一個數據來源是商品介紹內各種商品的特徵。 一個例子是製造商對商品的文字描述。
- 第二個數據來源是根據用戶對各種商品的回饋獲取的用戶個人興趣特點資料。 用戶回饋可能是顯式或隱式的。 顯式反饋可以從用戶的評分中獲取，而隱式反饋可以從用戶的行動中(例如購買)獲取。



# 基於內容系統的基本組件

- 主要組件包括（離線）預處理部分，（離線）訓練部分和線上預測部分。
- **預處理和特徵提取**：在大多數情況下，從網頁、商品描述、新聞及音樂中摘錄特徵，轉換為基於關鍵字和詞的推薦系統。正確摘錄最豐富的訊息、資訊對於有效的推薦至關重要。
- **基於內容的用戶個人特點資料學習**：構建用戶特定的模型，以基於他們過去的購買或評分來預測用戶對商品的興趣。由隱式或顯式評分與商品屬性結合來建構訓練數據。這個訓練數據構建了一個學習用戶個人興趣特點資料的模型。
- **篩選和建議**：所訓練好的模型可以用於推薦商品給特定的用戶。



# 關鍵字表示

- 商品或用戶使用一組關鍵字表示。每個維度表示與關鍵字存在相關聯的二進制隨機變量（1：存在，0：文檔中不存在）
- 文件1: This is a sample
- 文件2: This is another example

	this	is	a	sample	another	example
文件1	1	1	1	1	0	0
文件2	1	1	0	0	1	1

# 預處理和特徵提取

- 第一階段是摘錄用於表示商品具區分性的特徵。具區分性的特徵是對用戶興趣有高度預測性的特徵。
- 這個階段的程序和所針對的具體應用有關。例如：網頁推薦系統與商品推薦系統就非常不同。
- 最常用的商品特徵摘錄方法是從商品敘述中摘錄關鍵字。以網頁來說，則需要在網頁中找到相關部分來摘錄關鍵字。
- 為了便於在分類過程中使用，需要對各個特徵領域，根據其重要性，進行加權。

# 電影推薦的例子

- 考慮電影推薦網站，如IMDb。 每部電影通常都與電影的描述相關聯，如劇情簡介，導演，演員，類別等。
- 對於關鍵字選擇，可以從各個領域中的文字描述建立關鍵字。但是，並非每個關鍵字都具有同等重要性。
- 為了盡量減少問題，可以使用以下方法：
  - 可以使用特定領域的知識來確定關鍵字的相對重要性。 例如，電影的標題和主要演員可以被賦予比電影簡介中的詞更多的權重。 這是通過試驗和錯誤來完成的。
  - 在許多情況下，自動化學習各種功能的相對重要性是可能的。

IMDb

Google Code Archive - Lo... X IMDb: Highest Rated Sci-Fi... X

www.imdb.com/search/title?genres=sci-fi&sort=user\_rating,desc&title\_type=feature&num\_votes=25000&pf\_rd\_m=A2FGELUUNOQJNL&pf\_rd\_p=2406822102&pf\_rd\_...

Find Movies, TV shows, Celebrities and more... All Search

IMDbPro Help Sign in with Facebook Other Sign in

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist

## Highest Rated Sci-Fi Feature Films With At Least 25000 Votes

1 to 50 of 567 titles | [Next »](#) View Mode: [Compact](#) | [Detailed](#)

Sort by: [Popularity](#) | [Alphabetical](#) | [IMDb Rating ▼](#) | [Number of Votes](#) | [US Box Office](#) | [Runtime](#) | [Year](#) | [Release Date](#)



**1. Inception** (2010)

PG-13 | 148 min | Action, Adventure, Sci-Fi

★ **8.8** ☆ [Rate this](#) **74** Metascore

A thief, who steals corporate secrets through use of dream-sharing technology, is given the inverse task of planting an idea into the mind of a CEO.

Director: [Christopher Nolan](#) | Stars: [Leonardo DiCaprio](#), [Joseph Gordon-Levitt](#), [Ellen Page](#), [Ken Watanabe](#)

Votes: 1,567,645 | Gross: \$292.57M

# 關鍵字和詞的向量代表和清理

- 當非結構化文件被轉換成向量時，這個過程特別重要。
- 向量中關鍵字和詞的摘錄需要清理並以適當格式的文字袋 (Bag of Words) 以進行處理。
- 清潔過程中有幾個步驟：
  - Stop-word removal: articles, prepositions, conjunctions, and pronouns, are treated as stop-words, e.g. “a”, “an”, “the”.
  - Stemming: e.g. “went” -> “go”, “hoping” -> “hop”. Off-the-shelf tools such as Open NLP and Google Code are available.
  - Phrase extraction: “United States”, “hot dog”.
  - Size cut-off: e.g. use top 100 keywords
- 執行這些步驟後，關鍵字和詞被轉換成向量代表。文件成為文字袋及其頻率所代表的向量。
- 例如：This is a **cat** **born** in the **United States**. => (1,1,1) with first ‘1’ representing “cat”, second “1” represent “bear”, and third “1” represent “United States”. 所有其他的字可以被去除。

# Term-Frequency - Inverse Document Frequency (TF – IDF)

- 簡單的關鍵字表示有其問題
  - in particular when automatically extracted as
    - **not every word has similar importance** (e.g. “recommender” vs. “course”)
    - **longer documents have a higher chance to have an overlap with the user profile**
- 標準的度量：TF-IDF
  - Encodes text documents in multi-dimensional Euclidian space
    - weighted term vector
  - TF: Measures, how often a term appears (**density in a document**)
    - assuming that important terms appear more often
    - **normalization has to be done in order to take document length into account**
  - IDF: Aims to reduce the weight of terms that appear in all documents



# TF-IDF

- Given a keyword  $i$  and a document  $j$
- $TF(i, j)$ 
  - term frequency of keyword  $i$  in document  $j$
- $IDF(i)$ 
  - inverse document frequency calculated as  $IDF(i) = \log \frac{N}{n(i)}$ 
    - $N$ : 所有可推薦文件的數量
    - $n(i)$ : 出現關鍵字 $i$ 的文件數量
- $TF-IDF$ 
  - is calculated as:  $TF-IDF(i, j) = TF(i, j) * IDF(i)$

基於TF-IDF的概念，哪些關鍵字是最重要的關鍵字？

# TF-IDF的例子

Document 1

Term	Term Count
This	1
Is	1
a	2
cat	1

$$\text{tf}(\text{"This"}, d1) = 1/5 = 0.2$$

$$\text{tf}(\text{"This"}, d2) = 1/7 \approx 0.13$$

$$\text{idf}(\text{"This"}, D) = \log(2/2) = 0$$

$$\text{tfidf}(\text{"This"}, d1) = 0.2 * 0 = 0$$

$$\text{tfidf}(\text{"This"}, d2) = 0.13 * 0 = 0$$

Document 2

Term	Term Count
This	1
Is	1
a	2
train	3

$$\text{tf}(\text{"train"}, d1) = 0/5 = 0$$

$$\text{tf}(\text{"train"}, d2) = 3/7 \approx 0.429$$

$$\text{idf}(\text{"train"}, D) = \log(2/1) = 0.301$$

$$\text{tfidf}(\text{"train"}, d1) = 0 * 0.301 = 0$$

$$\text{tfidf}(\text{"train"}, d2) = 0.429 * 0.301 \approx 0.13$$

這兩個文件中有多少個非零的TF-IDF關鍵字？

# 收集用戶的喜歡和不喜歡

- 關於用戶喜好和不喜歡的數據可以採取以下任何形式：
  - **評分**：它們可以是二進制(binary)，基於區間(interval-based)，或序數(ordinal)。
  - **隱式反饋**：基於用戶的行為，如購買或瀏覽一個商品。
  - **文字發表的意見**：透過意見挖掘和情感分析提取隱含評分。
  - **案例**：用戶指定的例子或他們感興趣的商品。
- 在所有情況下，用戶對商品的喜好或不喜歡最終轉換為一元，二元，基於區間或實數(real number)的評價。
- 這個評分也可以被看作是一個提取的種類標籤或因變量(Dependent Variables)，最終的目的是用於學習。

# 受監督的字和詞選擇和權重

- 字和詞選擇和加權的目標是確保只有對商品最具代表性的字和詞才被保留在向量中。
- 根據實驗結果，提取的字和詞的數量應該在50-300之間。
- 在代表文件的向量中提取具代表性的字和詞有兩個截然不同的方法。一個是選擇，另一個是加權。
- 停止詞的去除和TF-IDF的使用分別是無監督(unsupervised)的字和詞選擇和加權的例子，其中用戶的評分沒被考慮。
- 受監督的(supervised)字和詞選擇和加權是考慮用戶評分以評估這些字和詞對商品的代表性。
- 大多數的這些方法評估因變量(客戶的評分)對特定字和詞的敏感性，以評估其代表性。方法有很多，例如，基尼指數，熵(entropy)， $\chi^2$ 統計等。然後，根據敏感性給予額外的權量。

# 基尼指數和熵 (Gini Index and Entropy)

- 基尼指數是特徵選擇最常用的方法之一。
- 設 $t$ 為評分可能出現值的總數。在包含特定單詞 $w$ 的商品文件中，令 $p_1(w)$ ,  $p_2(w)$ ,  $\dots$ ,  $p_t(w)$ 為這些可能出現的評分值中每一個的評分的百分比。 $w$ 這個詞的基尼係數定義如下：

$$\text{Gini}(w) = 1 - \sum_{i=1}^t p_i(w)^2$$

- $\text{Gini}(w)$  的值總是在  $(0, 1 - 1/t)$  的範圍內，較小的值表示更大的判別力。
- 例如，當單詞 $w$ 的存在總是導致文檔被評為第 $j$ 個可能值時(i.e.,  $p_j(w)=1$ ), 那麼基尼指數是0並且該單詞是非常區別的。
- 除了用資訊理論的原則來設計指標之外，熵(entropy)在原理上與基尼指數非常相似：  
$$\text{Entropy}(w) = - \sum_{i=1}^t p_i(w) \log(p_i(w))$$
- $\text{Entropy}(w)$  的值總是處於  $(0, 1)$  的範圍內，較小的值更具有區分性。

當每個值 $p_j(w)$ 取相同的值 $1/t$ ，基尼指數是多少？

# 基尼指數的例子

- 文件1: I like those beautiful dogs (Rating = 3)
- 文件2: Puddles are beautiful dogs (Rating = 5)
- 文件3: I like dogs than cats (Rating = 1)
- Let's calculate the Gini Index of “Dog” and “Cat”
- Dog:  $p_3(\text{Dog}) = 0.33$ ,  $p_5(\text{Dog}) = 0.33$ ,  $p_1(\text{Dog}) = 0.33$  ;
- $\text{Gini}(\text{Dog}) = 1 - ((0.33)^2 + (0.33)^2 + (0.33)^2) = 0.67$
- Cat:  $p_1(\text{Cat}) = 1$ ,  $\text{Gini}(\text{Cat}) = 1 - 1 = 0$



# 監督特徵的權重

- 我們已經討論過如何使用TF-IDF等措施來衡量文件的權重。但是，逆文件頻率是一種無監督(unsupervised)的度量，不依賴於用戶喜歡或不喜歡。
- 我們還可以使用監督(supervised)的度量來進一步來加權文件中的向量值，以便對不同的單詞產生不同的重要性。
- 例如，對於單詞 $w$ ，考慮以下加權函數 $g(w)$ ，其中 $a$ 是大於1的參數： $g(w) = a - \text{Gini}(w)$
- 然後將向量中每個單詞 $w$ 的權重乘以 $g(w)$ 。

# 學習用戶興趣特點資料 (User Profile) 和過濾

- 用戶興趣特點資料 (User Profile) 的學習與分類和回歸問題密切相關。
- 假設我們使用的內容是商品的描述文件。在裡面我們有一套已經由特定用戶給過評分商品的描述文件。這些文件， $D_L$ ，也被用作訓練文件。這些訓練文件在預處理和特徵選擇階段中被轉換成文字袋及其頻率代表的商品向量，並且包含由用戶給的評分。所有這些文件用於建立訓練模型。
- 測試集 $D_U$ 中的文件則未被評分。我們可以用 $D_L$ 訓練出來的模型從 $D_U$ 中選擇商品向用戶推薦。
- 一些常用來訓練 $D_L$ 的方法是：最近鄰分類，貝葉斯分類器，關聯分類器和基於回歸的模型。

# 最近鄰的分類

- 第一步是定義一個相似度函數。最常用的相似函數是餘弦函數。
- 餘弦相似函數定義如下：Let  $\bar{X} = (x_1, \dots, x_d)$  and  $\bar{Y} = (y_1, \dots, y_d)$ , then

$$\text{Cosine}(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^d x_i y_i}{\sqrt{\sum_{i=1}^d x_i^2} \sqrt{\sum_{i=1}^d y_i^2}}$$

- 隨後，對於 $\mathbf{D}_U$ 中的每個文件，使用餘弦相似度函數在 $\mathbf{D}_L$ 中找出的k個最近鄰居。這個文件的預測評分（如果是分類，則為多數選票）是 $\mathbf{D}_U$ 中k個最近鄰居相應文件的平均評分。
- 然後基於評分的預測值對文件進行排序，並向用戶推薦最佳的商品。

# 討論和總結

- 與協同過濾方法相反，基於內容的技術不需要利用用戶群的數據來推薦
- 所用的方法利用用戶的顯式或隱式反饋來學習用戶興趣偏好的模型
- 評估證明，借助機器學習技術可以達到良好的推薦準確性，而且這些技術不需要用戶群的數據
- 經常推薦類似的物品，然而，如果推薦列表包含太多類似的商品風險將增加(為什麼?)
- 純粹的基於內容的系統很少在商業環境中找到

補

充



時

間

# 電影推薦系統的例子

- <https://www.kaggle.com/rounakbanik/movie-recommender-systems>



# 推薦閱讀：推薦系統 in Python 101

- <https://www.kaggle.com/gspmoreira/recommender-systems-in-python-101/notebook>