

KHOA CNTT & TRUYỀN THÔNG
BM KHOA HỌC MÁY TÍNH

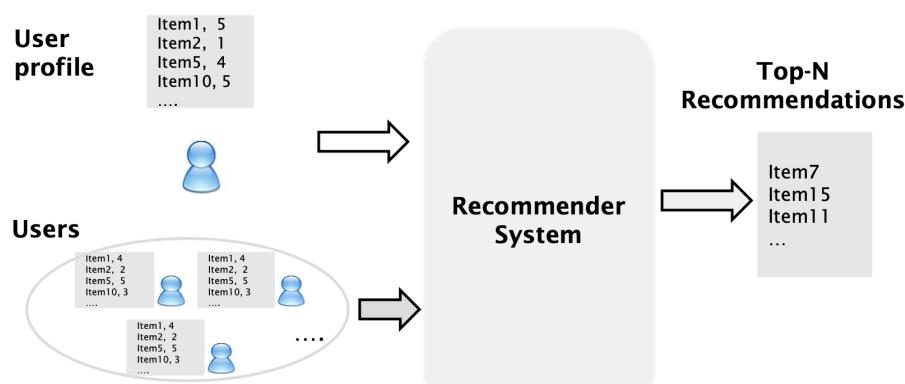
PHƯƠNG PHÁP XÂY DỰNG HỆ THỐNG GỢI Ý Dựa trên nội dung

Giáo viên giảng dạy:
TS. TRẦN NGUYỄN MINH THỦ
tnmthu@ctu.edu.vn

1

Phương pháp lọc cộng tác

Lọc cộng tác đưa ra các gợi ý cho người dùng dựa trên những người dùng có “profile” tương tự



2

Phương pháp lọc cộng tác

Hạn chế của phương pháp lọc cộng tác

- Dữ liệu thừa
- Sản phẩm mới

				✓	
		✓			
			✓		
					✓
	✓				

Collaborative Filtering issues: sparsity

	✓	✓		✓	
		✓	✓		
			✓		
				✓	?
	✓	✓			

Collaborative Filtering issues: new item problem

3

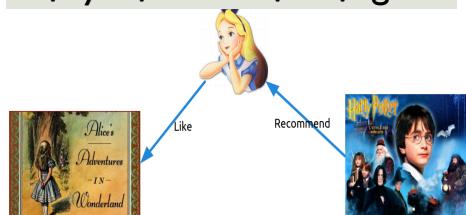
3

Phương pháp lọc cộng tác

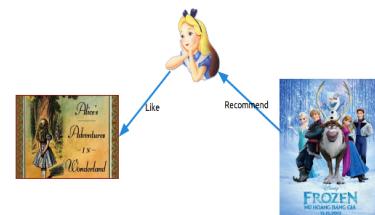
Hạn chế của phương pháp lọc cộng tác

- Gợi ý thiếu tính chính xác vì không quan tâm đến đặc điểm của sản phẩm mà người dùng đã chọn

Gợi ý dựa trên lọc cộng tác



Gợi ý dựa trên nội dung



4

4

2

Phương pháp lọc cộng tác

Hạn chế của phương pháp lọc cộng tác

- Khả năng giải thích cho kết quả gợi ý kém
- Thiếu minh bạch

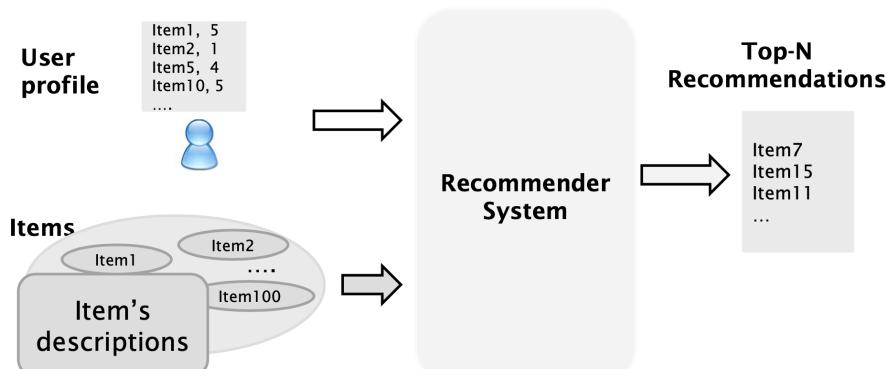


5

5

Gợi ý dựa trên nội dung

Gợi ý dựa trên nội dung đưa ra các gợi ý cho người dùng dựa vào mô tả của item và profile của người dùng



6

6

Phương pháp dựa trên nội dung

Ý tưởng chính: Gợi ý các sản phẩm cho người dùng X tương tự với các sản phẩm đã được đánh giá cao cũng bởi chính **người dùng X** này.

Matching “**user preferences**” with “**item characteristics**”

CB-RSs try to recommend **items similar* to those a given user has liked in the past**

[M. de Gemmis et al. Recommender Systems Handbook. Springer. 2015]

Ví dụ

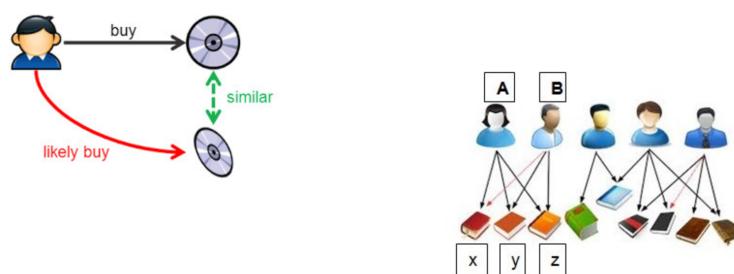
Gợi ý phim: Gợi ý các phim có cùng thể loại, đạo diễn, diễn viên,..

7

7

Phương pháp dựa trên nội dung

Ý tưởng chính: Gợi ý các sản phẩm cho người dùng X tương tự với các sản phẩm đã được đánh giá cao cũng bởi chính **người dùng X** này.



8

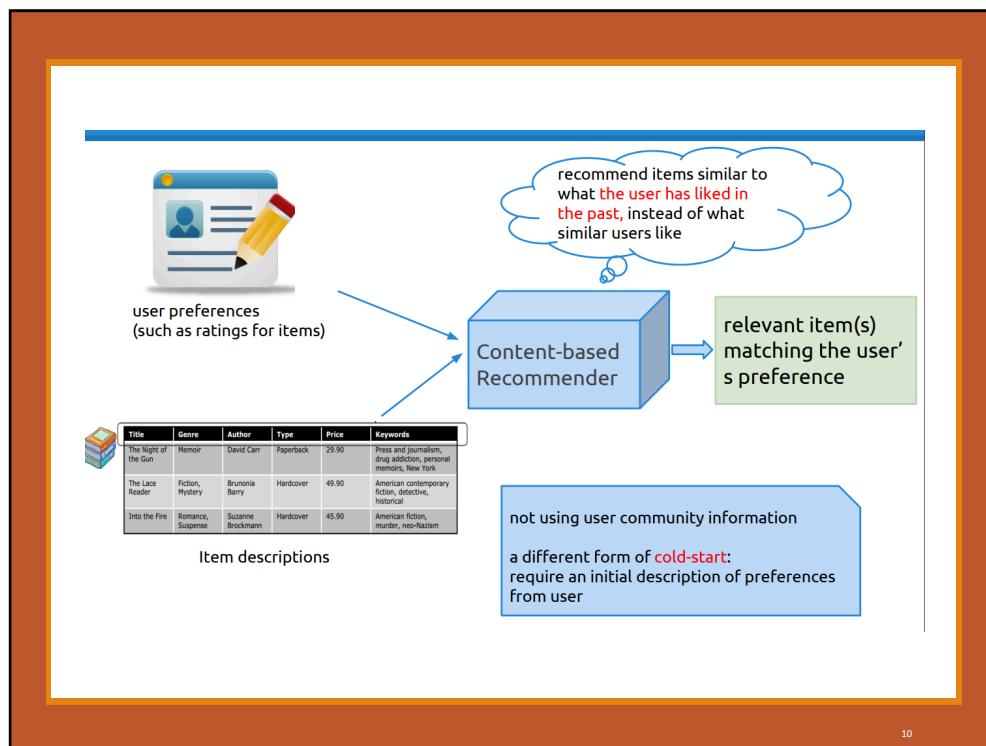
8

Hệ thống gợi ý dựa trên nội dung

Tìm kiếm các mục dữ liệu tương tự với mục dữ liệu mà **người dùng** đã mua/chọn lựa/ xem/ đánh giá trước đó **dựa trên thuộc tính** của các mục dữ liệu (ví dụ như màu sắc, giá cả, mô tả mục dữ liệu,...) và/ hoặc **dựa trên hồ sơ người dùng**

Title	Genre	Author	Type	Price	Keywords
The Night of the Gun	Memoir	David Carr	Paperback	29.90	Press and journalism, drug addiction, personal memoirs, New York
The Lace Reader	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
Into the Fire	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism

9



10

Hệ thống gợi ý dựa trên nội dung

Các bước chính:

- ❖ Biểu diễn mỗi “item” dưới dạng **một vector thuộc tính**.
- ❖ Gợi ý các “item” **tương tự** các “item” trong quá khứ **của chính người dùng**.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

- ❖ Hoặc xây dựng “profile” người dùng theo các thuộc “item” và gợi ý “item” có thuộc tính phù hợp với “profile” người dùng

11

11

Phương pháp dựa trên nội dung

Dữ liệu xây dựng gợi ý -> item description

➤ Có cấu trúc

- Phim: thể loại, đạo diễn, diễn viên
- Sách: tác giả, thể loại, nhà xuất bản,...

➤ Phi cấu trúc

- Nội dung bản tin
- Nội dung email

12

12

Phương pháp dựa trên nội dung

Xây dựng profile cho item và người dùng.

Profile - Tập các thuộc tính/ đặc trưng

- **Dữ liệu có cấu trúc :**

- Phim: thể loại, đạo diễn, diễn viên
- Sách: tác giả, thể loại, nhà xuất bản,...

- **Dữ liệu không cấu trúc:**

- **Văn bản:** tập các từ khóa quan trọng trong văn bản

Chỉ số TF-IDF

13

13

Phương pháp dựa trên nội dung

Dữ liệu xây dựng gợi ý -> item description

➤ Có cấu trúc

- Phim: thể loại, đạo diễn, diễn viên
- Sách: tác giả, thể loại, nhà xuất bản,...

Title	Genre	Author	Type	Price	Keywords
<i>The Night of the Gun</i>	Memoir	David Carr	Paperback	29.90	press and journalism, drug addiction, personal memoirs, New York
<i>The Lace Reader</i>	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
<i>Into the Fire</i>	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism
...					

Item of books

14

14

Dữ liệu có cấu trúc

Title	Genre	Author	Type	Price	Keywords
<i>The Night of the Gun</i>	Memoir	David Carr	Paperback	29.90	press and journalism, drug addiction, personal memoirs, New York
<i>The Lace Reader</i>	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
<i>Into the Fire</i>	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism
...					

Title	Genre	Author	Type	Price	Keywords
...	Fiction, Suspense	Brunonia Barry, Ken Follett	Paperback	25.65	detective, murder, New York
...					

Item of books

Alice's User profile

15

15

Dữ liệu có cấu trúc

Title	Genre	Author	Type	Price	Keywords
<i>The Night of the Gun</i>	Memoir	David Carr	Paperback	29.90	press and journalism, drug addiction, personal memoirs, New York
<i>The Lace Reader</i>	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
<i>Into the Fire</i>	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism
...					

Title	Genre	Author	Type	Price	Keywords
...	Fiction, Suspense	Brunonia Barry, Ken Follett	Paperback	25.65	detective, murder, New York
...					

Item of books (not yet seen by Alice)

Dice coefficient

 $i: \text{a not-yet-seen item}$
 $u: \text{user profile}$

$$\text{sim}(i, u) = \frac{2|\text{keyword}(i) \cap \text{keyword}(u)|}{|\text{keyword}(i)| + |\text{keyword}(u)|}$$

Alice's User profile

Measure similarity between items and user profile to make recommendations

16

16

Dữ liệu phi cấu trúc

Dữ liệu xây dựng gợi ý -> item description

➤ Phi cấu trúc

- Nội dung bản tin
- Nội dung email

- Boolean term vector

	team	coach	play	ball	score	game	win	lost	...
document1	1	0	1	0	0	1	0	1	
document2	0	0	0	1	1	0	1	0	
document3	1	1	1	0	0	1	0	1	
...									

17

17

Dữ liệu phi cấu trúc

➤ Phi cấu trúc

- Nội dung bản tin
- Nội dung email

Example TF-IDF Representation

Instead of a vector of Boolean values, the vector for each document is represented as the computed TF-IDF weights

id	men	entered	bank	charlotte	missiles	masks	aryan	guns	witnesses	reported	silver	suv	august
seg1.txt	0.239441	0	0.153457	0.195204	0	0.237025	0	0.395204	0.237029	0.140004	0.195243	0.237029	0
seg3.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg14.txt	0	0.192197	0	0	0	0	0	0	0	0	0	0	0.172681
seg15.txt	0	0	0	0	0	0	0	0	0	0	0	0	0.149052
seg16.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg17.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg18.txt	0	0.158432	0	0	0	0	0	0	0	0	0	0	0
seg19.txt	0	0	0	0.153457	0	0	0	0	0	0	0	0	0.135088
seg20.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg21.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg22.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg23.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg24.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg25.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg26.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg27.txt	0	0	0	0.235411	0	0	0	0	0	0	0	0	0
seg28.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg29.txt	0	0	0	0	0	0	0	0	0	0	0	0	0.142329
seg30.txt	0.078262	0	0	0	0	0	0	0	0	0	0	0	0
seg31.txt	0	0	0.213409	0	0	0	0.194701	0	0	0	0	0	0
seg32.txt	0	0	0	0	0	0	0	0	0	0	0	0	0

<http://jcsites.juniata.edu/faculty/rhodes/ida/textDocViz.html>

18

18

TF.IDF (Tần số từ khoá/ nghịch đảo tần số văn bản)

Phương pháp dựa trên tần số từ khoá (TF-Term Frequency)

Giá trị của một từ khoá w_{ij} được tính dựa trên tần số xuất hiện của từ khoá trong văn bản. Gọi tf_{ij} là số lần xuất hiện của từ khoá t_i trong văn bản d_j , khi đó có thể chọn cách tính w_{ij} theo một trong ba công thức dưới đây:

$$w_{ij} = \sqrt{tf_{ij}}$$

$$w_{ij} = 1 + \log(tf_{ij})$$

$$w_{ij} = tf_{ij}$$

19

TF.IDF (Tần số từ khoá/ nghịch đảo tần số văn bản)

Phương pháp dựa trên nghịch đảo tần số văn bản (IDF-Inverse Document Frequency)

Gọi df_i là số lượng văn bản có từ khoá t_i trong tập m văn bản đang xét, thì giá trị trọng số từ w_{ij} được tính bởi công thức:

$$w_{ij} = \log \frac{m}{df_j} = \log(m) - \log(df_i)$$

TF.IDF score : $w_{ij} = TF_{ij} * IDF_i$

Doc profile = set of words with highest TF.IDF scores, together with their scores

20