ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

DS300
HỆ KHUYẾN NGHỊ

KHUYẾN NGHỊ DỰA TRÊN NỘI DUNG (Content based Recommendation System)

Giảng viên: ThS. Nguyễn Văn Kiệt
CN. Huỳnh Văn Tín
Bộ môn Khoa học Dữ liệu
Khoa Khoa học và Kỹ thuật Thông tin

Nội dung

- Giới thiệu CB
- Nội dung là gì?
- Tiền xử lý nội dung
- Biểu diễn nội dung
- Tính toán tương tự

Giới thiệu CB

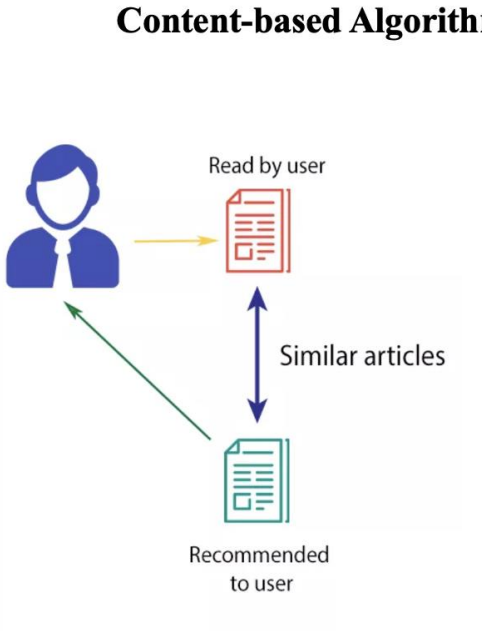
- Nguồn gốc từ nghiên cứu về Information Retrieval, Chẳng hạn, người dùng tìm tài liệu liên quan một vài từ khóa.
  - Người dùng thường tìm đọc những bài viết về ResSys, về NLP.
  - Người dùng đọc tin về thể thao, chính trị,…
  - Sở thích người dùng
  - Tìm sản phẩm tương tự sở thích → recommend

Giới thiệu CB

- Lọc cộng tác(Colaborative Fittering): Không cần thông tin về đối tượng
- Dựa trên nội dung:
  - Khai thác thông tin về đối tượng như: nội dung, chủ đề, thể loại, …
  - Không cần thêm thông tin rating của người dùng như CF.
  - Ứng dụng: Nhiều trong Text doc recommendation.

Giới thiệu CB

- **Giả định**: Sở thích của người dùng phải khớp với mô tả về các mặt hàng cái mà mà người dùng nên được hệ thống đề xuất.
  - Mô tả của mặt hàng càng giống với mô tả mà người dùng quan tâm, thì người dùng càng có nhiều khả năng thấy đề xuất của mặt hàng đó thú vị.
- **Mục tiêu**: tìm ra điểm tương đồng giữa người dùng và tất cả các mặt hàng hiện có là cốt lõi của loại hệ thống khuyến nghị này.



Nội dung?

- **Nội dung là t**hông tin của các đối tượng khuyến nghị (sách, phim, tin tức…) được biểu diễn dưới dạng văn bản.
- Mỗi đối tượng được biểu diễn cùng một tập các thuộc tính thể hiện nội dung

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
...					

Nội dung?


TIN MỚI
21:41 Mười hòn đảo được 'cai trị' hoàn toàn bởi các loài động vật
21:20 Ăn chay hay ăn thịt sẽ giúp chúng ta sống lâu hơn?
21:00 Ngôi chùa cổ hàng trăm năm tuổi bên dòng Kỳ Cùng
20:54 Cảng hàng không Sa Pa, Quảng Trị khởi công năm 2022 hoặc đầu 2023
20:52 Nữ sinh liệt hai chân 10 năm tìm chữ "trên đôi chân của mẹ"
20:37 Đám cưới trong rừng đẹp như phim cổ tích của cặp đôi thanh mai trúc mã
CafeBiz
TRANG THÔNG TIN ĐIỆN TỬ TỔNG HỢP
XÃ HỘI
KINH TẾ VĨ MÔ
KINH DOANH
CÔNG NGHỆ
SỐNG
Apple sẽ chuyển 25% sản lượng iPhone sang Ấn Độ, 20% iPad và Apple Watch sang Việt Nam
26/09/2022 10:24 AM | KINH DOANH
Apple đã bắt đầu sản xuất một số thiết bị của mình ở Ấn Độ và Việt Nam từ cách đây vài năm, dần dần cắt giảm sự phụ thuộc vào Trung Quốc. Theo các nhà phân tích của JP Morgan, gã khổng lồ Cupertino hiện đang chuẩn bị đưa hai quốc gia này trở thành những trung tâm sản xuất quan trọng trên toàn cầu.
Chia sẻ
Từ khóa: gã khổng lồ , khoản trợ cấp , đáng mơ ước

Tiền xử lý nội dung văn bản

- **Loại bỏ thẻ HTML** (Nếu có)
  - <html> **Đây là nội dung cần lấy** </html>
- **Chuẩn hóa bảng mã**: Đưa văn bản về cùng bản mã tiếng Việt (Unicode)
- **Chuẩn hóa kiểu gõ** dấu tiếng Việt (Khóc oà khóc òa, trơi trời ,…)
- **Xử lý teencode** (khum không, nyc người yêu cũ, móa mẹ, …)
- **Tách câu, tách từ** (tokenization, segmentation)
  - Một số thư viện như vncorenlp, underthesea, pyvi,…
- **Chuyển về LowerCase, UpperCase**.
- Xem xét **bỏ các ký tự đặc biệt.**
- **Loại bỏ stopword**, những từ xuất hiện thường xuyên trong hầu hết các văn bản, không có ý nghĩa.

Tiền xử lý nội dung văn bản

Ví dụ 1:

Document 1

<html>
<body>
Trí tuệ Nhân tạo (AI) là cuộc đua của những gã khổng lồ như Amazon, Google, Facebook, etc.
</body>
</html>

Tiền xử lý nội dung văn bản

Ví dụ 2:

Document 2

<html>
<body>
Nguồn nhân lực CNTT nói chung và AI nói riêng thật sự khan hiếm.
</body>
</html>

Tiền xử lý nội dung văn bản

- Loại bỏ HTML
- Chuẩn hóa bộ mã, kiểu gõ
- Tokenization
  - *Trí tuệ Nhân tạo (AI) là cuộc đua của những gã khổng lồ như Amazon, Google, Facebook, etc.*
  - Nguồn nhân lực CNTT nói chung và AI nói riêng thật sự khan hiếm.

→ {

Trí, tuệ, Nhân, tạo, (AI), là, cuộc, đua, của, những, gã, khổng, lồ, như, Amazon, Google, Facebook, etc,

Nguồn, nhân, lực, CNTT, nói, chung, và, AI, nói, riêng, thật, sự, khan, hiếm

}

- Lowercase

Tiền xử lý nội dung văn bản

- **Lowercase**

→ {

trí, tuệ, nhân, tạo, (ai), là, cuộc, đua, của, những, gã, khổng, lồ, như, amazon, google, facebook, etc, nguồn, nhân, lực, cntt, nói, chung, và, ai, nói, riêng, thật, sự, khan, hiếm

}

- **bỏ các ký tự đặc biệt.**

→ {

trí, tuệ, nhân, tạo, ai, là, cuộc, đua, của, những, gã, khổng, lồ, như, amazon, google, facebook, etc, nguồn, nhân, lực, cntt, nói, chung, và, ai, nói, riêng, thật, sự, khan, hiếm

}

Tiền xử lý nội dung văn bản

- Tokenization (theo từ đơn và từ ghép)

 {trí_tuệ_nhân_tạo, ai, là, cuộc_đua, của, những, gã_khổng_lồ, như, amazon, google, facebook, etc,
nguồn, nhân_lực, cntt , nói, chung, và, ai, nói, riêng, thật, sự, khan_hiếm
}

Tiền xử lý nội dung văn bản

- Loại bỏ stopword

{trí_tuệ_nhân_tạo, ai, là, cuộc_đua, của, những, gã_khổng_lồ, như, amazon, google, facebook, etc,
nguồn, nhân_lực, cntt, nói, chung, và, ai, nói, riêng, thật, sự, khan_hiếm
}

{trí_tuệ_nhân_tạo, ai, cuộc_đua, gã_khổng_lồ, amazon, google, facebook
nhân_lực, cntt, ai, khan_hiếm
}

Tiếp cận CB

Để tính f(u,p) dựa trên tiếp cận nội dung (CB), (dựa trên bộ nhớ)

- Bước 1: Biểu diễn nội dung đối tượng, **content(p)**.
- Bước 2: Biểu diễn User Profile, **UserProfile(u)**.
- Bước 3: **Tính toán tương tự** về nội dung, f(u,p)

Biểu diễn nội dung - content(p)

Ví dụ 1:

Bài viết	Từ khóa (rút từ nội dung)
D1: BKAV sẽ ra mắt hai mẫu Bphone mới trong năm nay, sẽ có cả flagship hỗ trợ 5G?	nguyễn tử quảng, phát triển công nghệ, smartphone
D2: iPhone lần đầu có dung lượng 1TB	nhà phân tích, dây chuyền sản xuất, apple
...	

Biểu diễn nội dung - content(p)

Nội dung đối tượng có thể biểu diễn:

- Tập từ khoa
  - D1 = {nguyễn tử quảng, phát triển công nghệ, smartphone}
  - D2 = {nhà phân tích, dây chuyền sản xuất, apple}

- Vector nội dung: Boolean vectors, TF-IDF vectors, word-embedding,…
  - D1 = {1,1,1,0,0,0}
  - D2 = {0,0,0,1,1,1}

Biểu diễn nội dung - content(p)

Ví dụ: Document1, Document2. Sử dụng boolean vector biểu diễn
Sau khi tiền xử lý:
D1 = {trí_tuệ_nhân_tạo, ai, cuộc_đua, gã_khổng_lồ, amazon, google, facebook}
D2 = {nhân_lực, cntt, ai, khan_hiếm}

- *Từ điển = {trí_tuệ_nhân_tạo, ai, cuộc_đua, gã_khổng_lồ, amazon, google, facebook, nhân_lực, cntt, khan_hiếm}*

- D1 = {1, 1, 1, 1, 1, 1, 1, 0, 0, 0}
- D2 = {0, 1, 0, 0, 0, 0, 0, 1, 1, 1}

Biểu diễn nội dung - content(p)

$$Content(p) = \overrightarrow{w_p} = (w_{1,p}, w_{2,p}, \ldots, w_{k,p})$$

Trong đó,

- K: là tổng số đặc trưng dùng để diễu diễn nội dung đối tượng.
- $w_{i,p}$: trọng số đặt trưng thứ i của đối tượng p.

Biểu diễn nội dung - content(p), TF-IDF

Đưa vào một từ t và một văn bản d

IF-IDF(t,d) = TF(t,d)*IDF(t)

- IF: Term Frequency
  - Tần suất xuất hiện của từ t trong văn bản d
  - Những từ quan trọng xuất hiện thường xuyên hơn
  - TF(t,d) = count of t in d / number of words in d (hoặc có thể không chia)
- IDF: Inverse Document Frequency
  - Giảm trọng số cho những từ xuất hiện trong hầu hết các tài liệu.

$$\text{IDF(t)} = \frac{N}{df(t)} \quad \text{Hoặc} \quad \text{IDF(t)} = log\left(\frac{N}{df(t)}\right) \quad \text{Hoặc} \quad \text{IDF(t)} = log\left(\frac{N}{df(t)+1}\right)$$

  - N: Tổng số tài liệu
  - df(t): Số tài liệu mà từ t xuất hiện trong N tài liệu

Biểu diễn nội dung người dùng – UserProfile(u)

- Có thể biểu diễn
  - Tập các từ thể hiện nội dung, sở thích của mỗi user.
  - Vector thể hiện sở thích của người dùng u.

- Ví dụ: u đọc cả D1 và D2

D1 = {nguyễn tử quảng, phát triển công nghệ, smartphone}

D2 = {nhà phân tích, dây chuyền sản xuất, apple}

  - u = {nguyễn tử quảng, phát triển công nghệ, smartphone, nhà phân tích, dây chuyền sản xuất, apple}
  - u = {1, 1, 1, 1, 1, 1}
  - U = ($if*idf_{w1}$, $if*idf_{w2}$, $if*idf_{w3}$, $if*idf_{w4}$, $if*idf_{w5}$, $if*idf_{w6}$)

Biểu diễn nội dung người dùng – UserProfile(u)

$$UserProfile(u) = \overrightarrow{w_u} = (w_{1,u}, w_{2,u}, \dots, w_{k,u})$$

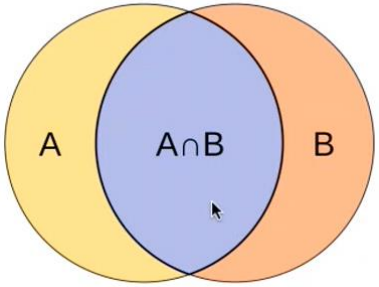
Trong đó,

- K: là tổng số đặc trưng dùng để diễu diễn nội dung đối tượng.
- $w_{i,p}$: trọng số đặt trưng thứ i của đối tượng p.

Tính toán tương tự – f(u,p)

- Tính toán tương tự Jaccard: dựa trên tập từ khóa chung.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



- Tính toán tương tự dựa trên góc giữa 2 vector (Cosine) hoặc khoảng cách Euclide giữa 2 vector trong không gian n chiều.

$$f(u,p) = Sim(\overrightarrow{w_u}, \overrightarrow{w_p}) = \frac{\overrightarrow{w_u} \bullet \overrightarrow{w_p}}{\| \overrightarrow{w_u} \| * \| \overrightarrow{w_p} \|}$$

Q&A

Thank you!