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

- Cải tiến không gian vector
- Phương pháp phân loại văn bản đơn giản
- Phương pháp đơn giản: K lân cận gần nhất
- Phương pháp Rocchio

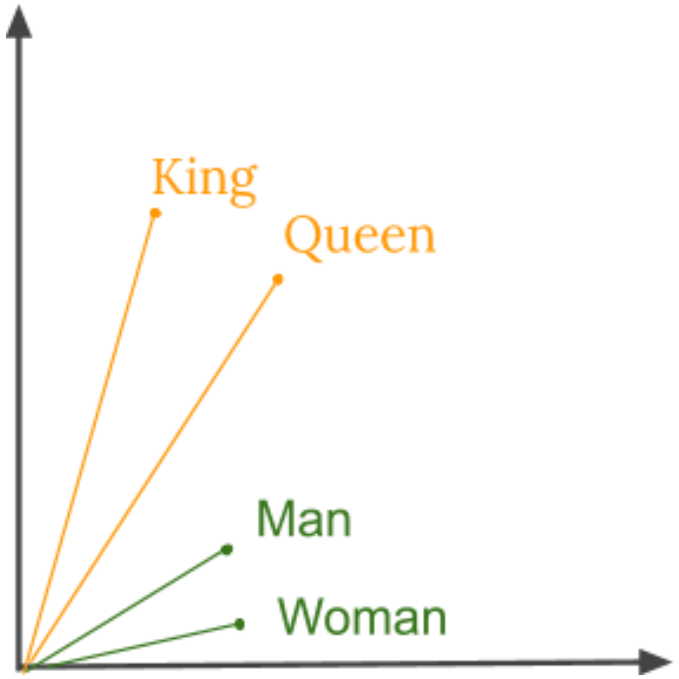
Cải tiến không gian vector – Word Embedding

❖ Word Embedding là gì ?

- **Word Embedding** là một không gian vector dùng để biểu diễn dữ liệu có khả năng miêu tả được mối liên hệ, sự tương đồng về mặt ngữ nghĩa, văn cảnh(context) của dữ liệu.
- Không gian này bao gồm nhiều chiều, các từ trong không gian đó mà có cùng văn cảnh hoặc ngữ nghĩa sẽ có vị trí gần nhau.

Cải tiến không gian vector – Word Embedding

- **Word Embedding là gì ?**



Cải tiến không gian vector – Word Embedding

❖ Tại sai chúng ta cần Word Embedding?

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}

Cải tiến không gian vector – Word Embedding

- **Tại sai chúng ta cần Word Embedding?**

Document	Index	One-hot encoding
a	1	[1, 0, 0, ..., 0](9999 số 0)
b	2	[0, 1, 0, ..., 0]
c	3	[0, 0, 1, ..., 0]
....	.....	.....
mẹ	9999	[0, 0, 0, ..., 1, 0]
vân	10000	[0, 0, 0, ..., 0, 1]

Cải tiến không gian vector – Word Embedding

Document	Index	One-hot encoding
a	1	[1, 0, 0, ..., 0](9999 số 0)
b	2	[0, 1, 0, ..., 0]
c	3	[0, 0, 1, ..., 0]
....	.....	.....
mẹ	9999	[0, 0, 0, ..., 1, 0]
vân	10000	[0, 0, 0, ..., 0, 1]

- **Chi phí tính toán lớn:** data 100 từ lên 10000 từ thì không gian trở nên rất lớn.
- **Mang ít giá trị thông tin:** Các vector hầu như toàn số 0. Không thể biểu diễn vị trí
- **Độ khái quát yếu:** Ví dụ ta có ba từ cùng chỉ người mẹ: **mẹ, má, bầm.** Không thể khái quát chung ba từ này dù có chung nghĩa.

Cải tiến không gian vector – Word Embedding

TF_IDF

- **TF** là tần suất xuất hiện của một từ trong data
- **IDF** là một hệ số giúp làm giảm trọng số của những từ hay xuất hiện trong data
- **TF-IDF** phương pháp này có thể giảm bớt trọng số của những từ xuất hiện nhiều nhưng lại không có nhiều thông tin.

Cải tiến không gian vector – Word Embedding

Word2Vec:

- Là một công cụ được phát minh để giải quyết vấn đề trên.
- Nó biểu diễn mỗi từ bằng một vector có độ dài cố định và những vector này biểu thị tốt hơn độ tương tự ngữ nghĩa giữa các từ.
- **Word2Vec gồm hai mô hình:**
  - Túi từ liên tục (*continuous bag of words* – CBOW) [Mikilov et al., 2013a]
  - skip-gam [Mikilov et al., 2013b]

Cải tiến không gian vector – Word Embedding

Túi từ liên tục (*continuous bag of words* – CBOW)

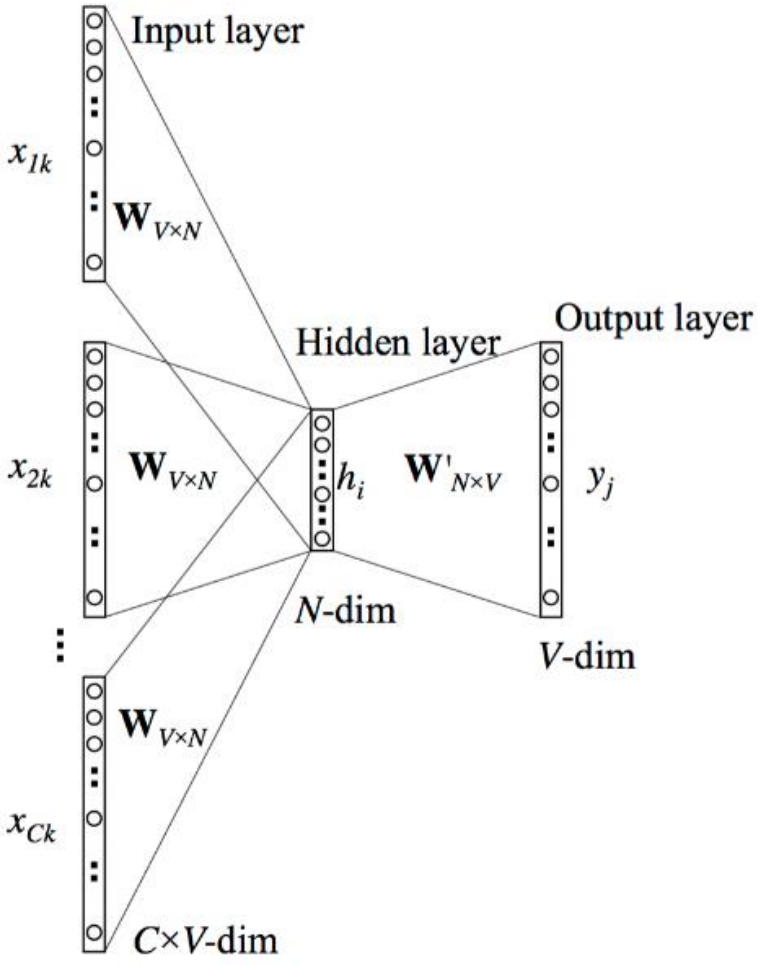
- Phương pháp này lấy đầu vào là một hoặc nhiều từ context word và cố gắng dự đoán output từ đầu ra
- Ví dụ ta có một câu tiếng anh như sau : "I love you". Ta có:

- Input context word : love

- Output target word: you

Cải tiến không gian vector – Word Embedding

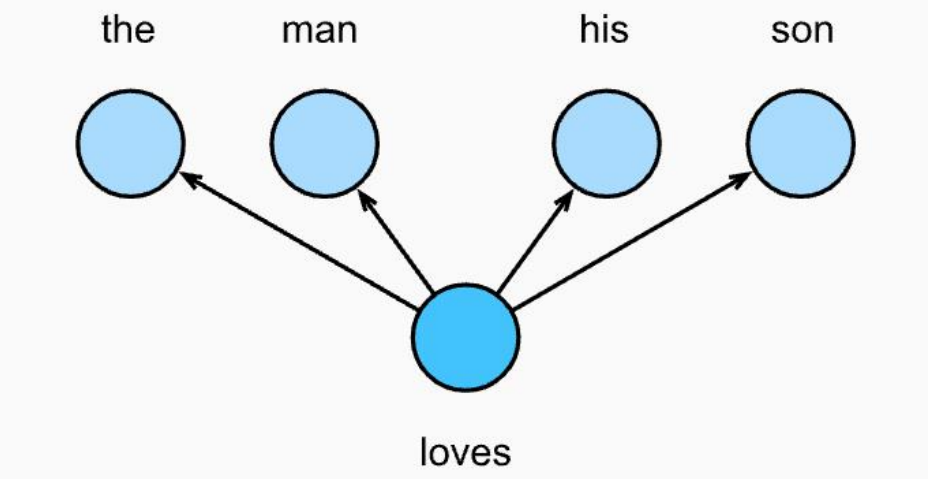
CBOW



Cải tiến không gian vector – Word Embedding

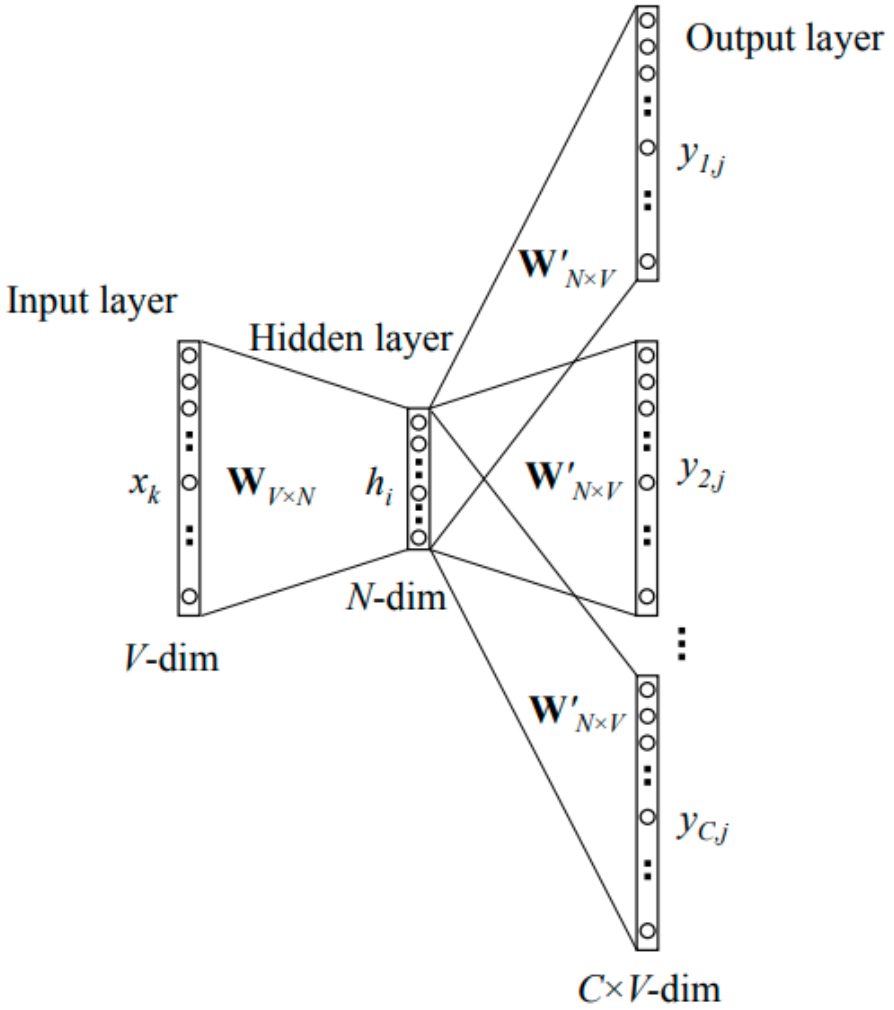
Skip-gam

- Sử dụng input là target word và cố gắng dự đoán ra các từ hàng xóm của nó.
- Ví dụ: The man **loves** his son



Cải tiến không gian vector – Word Embedding

Skip-gam



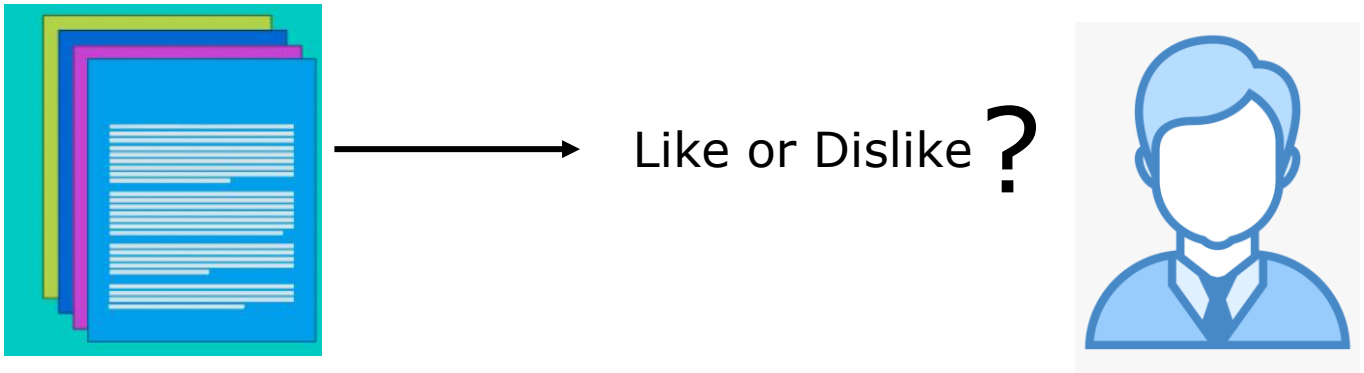
Word2Vec với bài toán Recommender System

Một số cách tiếp cận khi sử dụng Word Embedding cho bài toánRecommender System:

- **Fuzzy Search**: Giả sử khi bạn google, bạn search: **tà liệu về recommenderrr system cho newbiee**, rõ ràng là bạn gõ sai từ nhưng google sẽ tự động gợi ý với câu đúng: **tài liệu về recommender system cho newbie.**
- **Word Similarity hoặc Document Similarity**: Sử dụng các mô hình liên quan đến Word Embedding hoặc Document Embedding để xử lí các task cụ thể. Với word thì có khá nhiều các mô hình như: Word2Vec (CBOW, skip-gram), Glove, FastText, ... sử dụng như là bộ **Pre-train Word Embedding** để tính toán mức độ tương đồng hoặc đầu vào ban đầu cho các bài toán sử dụng mạng NN.

Phân loại văn bản

- **Phân loại văn bản:** là quá trình phân các tài liệu văn bản thành hai hay nhiều loại.
- Phân loại văn bản là một phương pháp được sử dụng trong hệ khuyến nghị dựa trên nội dung.
- Phân loại văn bản là một bài toán học giám sát (supervised learning) trong học máy. Nội dung của văn bản đã được gán nhãn, và được sử dụng để thực hiện phân loại.



Mô hình dựa trên xác suất

- Naive Bayes là một thuật toán phân lớp được mô hình hoá dựa trên định lý Bayes trong xác suất thống kê.

$$P(Y|X) = \frac{\prod_{i=1}^{d} P(X_i|Y) \times P(Y)}{P(X)}$$

P(Y|X): Xác suất xảy ra Y với điều kiện là X
P(X|Y): Xác suất của đặc trưng X khi biết Y
P(Y): Xác suất của lớp Y
P(X): xác suất của đặc trưng X
Vì P (X) là một giá trị không đổi, chúng ta có thể bỏ qua nó trong các phép tính của mình. P

Mô hình dựa trên xác suất

Ví dụ:

Doc-ID	recommender	intelligent	learning	school	**Label**
1	1	1	1	0	**1**
2	0	0	1	1	**0**
3	1	1	0	0	**1**
4	1	0	1	1	**1**
5	0	0	0	1	**0**
6	1	1	0	0	**?**

Mô hình dựa trên xác suất

Ví dụ:

$$P(Y|X) = \frac{\prod_{i=1}^{d} P(X_i|Y) \times P(Y)}{P(X)}$$

Doc-ID	recommender	intelligent	learning	school	**Label**
1	1	1	1	0	**1**
2	0	0	1	1	**0**
3	1	1	0	0	**1**
4	1	0	1	1	**1**
5	0	0	0	1	**0**
6	1	1	0	0	**?**

P(Label =1) = 3/5

$$\begin{aligned} P(X|Label=1) = {} & P(recommender=1|Label=1) \times \\ & P(intelligent=1|Label=1) \times \\ & P(learning=0|Label=1) \times P(school=0|Label=1) \\ = {} & 3/3 \times 2/3 \times 1/3 \times 2/3 \\ \approx {} & 0.149 \end{aligned}$$

P(Label =1|Doc-ID = 6) = 0.149*3/5 = 0.0894

Mô hình dựa trên xác suất

Ví dụ:

$$P(Y|X) = \frac{\prod_{i=1}^{d} P(X_i|Y) \times P(Y)}{P(X)}$$

Doc-ID	recommender	intelligent	learning	school	**Label**
1	1	1	1	0	**1**
2	0	0	1	1	**0**
3	1	1	0	0	**1**
4	1	0	1	1	**1**
5	0	0	0	1	**0**
6	1	1	0	0	**?**

P(Label = 0|Doc-ID = 6) = ?

- Kết luận?

Mô hình dựa trên xác suất

Với trường hợp các đặc trưng là văn bản

DocID	Words	**Label**
1	recommender intelligent recommender	**1**
2	recommender recommender learning	**1**
3	recommender school	**1**
4	teacher homework recommender	**0**
5	recommender recommender recommender teacher homework	**?**

Mô hình dựa trên xác suất

Với trường hợp các đặc trưng X là văn bản

$$P(Y|X) = \frac{\prod_{i=1}^{d} P(X_i|Y) \times P(Y)}{P(X)}$$

X bây giờ là chuỗi bao gồm các ký tự *v* khác nhau.

$$P(v_i|C = c) = \frac{CountTerms(v_i, docs(c))}{AllTerms(docs(c))}$$

- CountTerms(v_i, docs(c)):Trả về số lần xuất hiện của từ v_i trong các tài liệu có nhãn c.
- AllTerms(docs(c)): Trả về số lượng tất cả các từ trong các tài liệu c này.

Mô hình dựa trên xác suất

X bây giờ là chuỗi bao gồm các ký tự *v* khác nhau.

$$P(v_i|C=c) = \frac{CountTerms(v_i, docs(c))}{AllTerms(docs(c))}$$

Làm min

+1

$$\hat{P}(v_i|C=c) = \frac{CountTerms(v_i, docs(c)) + 1}{AllTerms(docs(c)) + |V|}$$

V là tập từ vựng, các từ có trong tất cả các lớp

$$\hat{P}(v_i|C=c) = \frac{CountTerms(v_i, docs(c)) + 1}{AllTerms(docs(c)) + |V|}$$

Mô hình dựa trên xác suất

DocID	Words	**Label**
1	recommender intelligent recommender	**1**
2	recommender recommender learning	**1**
3	recommender school	**1**
4	teacher homework recommender	**0**
5	recommender recommender recommender teacher homework	**?**

$$\hat{P}(recommender|Label = 1) = (5+1)/(8+6) = 6/14$$

$$\hat{P}(homework|Label = 1) = (0+1)/(8+6) = 1/14$$

$$\hat{P}(teacher|Label = 1) = (0+1)/(8+6) = 1/14$$

$$\hat{P}(Label = 1|v_1 \ldots v_n) = 3/4 \times (3/7)^3 \times 1/14 \times 1/14 \approx 0.0003$$

Mô hình dựa trên xác suất

DocID	Words	**Label**
1	recommender intelligent recommender	**1**
2	recommender recommender learning	**1**
3	recommender school	**1**
4	teacher homework recommender	**0**
5	recommender recommender recommender teacher homework	**?**

$$\hat{P}(Label = 1|v_1 \ldots v_n) = 3/4 \times (3/7)^3 \times 1/14 \times 1/14 \approx 0.0003$$

$\hat{P}(Label = 0|v_1 \ldots v_n)$? ☐ Kết luận

Q&A

Thank you!