



*MỞ ĐẦU*

Truy xuất thông tin từ Internet và từ CSDL lớn ngày nay thường khó khăn và tốn nhiều thời gian, đặc biệt là với dữ liệu không có cấu trúc. Một ví dụ điển hình là thư viện sách và hệ thống tìm kiếm tài liệu, dữ liệu liên tục bị thay đổi khi các văn bản được thêm, sửa hoặc xóa.

Khi xây dựng một hệ thống truy xuất thông tin hay cũng như cung cấp cho người dùng chức năng tìm kiếm, kiểm tra các tài liệu đã trả về. Nó phải chứa những CSDL trong phạm vi hàng GB, đáp ứng các yêu cầu của người dùng trong vòng 1 đến 10s. Nhiều thuật toán và kĩ thuật được phát triển nhưng vẫn chưa đáp ứng đủ những nhu cầu này cho đến khi các nhà nghiên cứu đề cập đến phương pháp ứng dụng lí thuyết không gian vector. [1]

Mô hình không gian vector là một mô hình đại số thể hiện thông tin văn bản như một vector, các phần tử của vector này thể hiện mức độ quan trọng của một từ và cả sự xuất hiện hay không xuất hiện của nó trong một tài liệu. Mô hình này biểu diễn văn bản như những điểm hay những vector trong không gian n chiều, mỗi chiều tương ứng với một từ trong tập hợp các từ. Chúng ta có thể đáp ứng như cầu truy vấn và xếp hạng bằng việc tính khoảng cách giữa điểm, hoặc là góc giữa vector truy vấn và vector văn bản. Nhiều kĩ thuật tính toán được áp dụng để hiện thực hóa lí thuyết này.

Tuy nhiên trong bài tiểu luận, chúng tôi chỉ đề cập đến việc sử dụng tần số từ – tần số nghịch đảo (term frequency–inverse document frequency) là một phương pháp phổ biến được phát triển bởi Gerard Salton vào đầu thập niên 60 của thế kỉ 20. [1]

Nội dung bài tiểu luận chỉ đưa ra những khái niệm và những ví dụ căn bản, đơn giản nhất để người đọc tiếp cận dễ dàng với đề tài. Nội dung gồm hai phần là cơ sở môn học (tổng quan về kiến thức của Không gian véctơ) và Ứng dụng (trong sắp xếp, truy vấn dữ liệu).

*Bảng phân công nhiệm vụ thành viên nhóm*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Họ và tên | Dẫn dắt nhóm | Tổng hợp tài liệu | Tìm hiểu cơ sở môn học | Tìm hiểu ứng dụng thực tế | Làm file trình chiếu |
| Vũ Tuấn Hải | X | X |  |  |  |
| Nguyễn Phi Khang |  |  | X |  |  |
| Nguyễn Hồng Khoa |  | X |  |  |  |
| Nguyễn Hoàng Trường |  |  | X |  |  |
| Lê Huỳnh Thăng |  |  |  | X | X |

*Bảng kế hoạch thực hiện*

|  |  |  |
| --- | --- | --- |
| Họ và tên | Thời gian | Thực hiện |
| Vũ Tuấn Hải | 01/11/2017 | Tổ chức họp mặt các thành viên trong nhóm và phân công công việc |
| Nguyễn Phi Khang  Nguyễn Hoàng Trường | 01/11–15/11/2017 | Tìm hiểu cơ sở kiến thức môn học |
| Lê Huỳnh Thăng | 15/11-30/11/2017 | Tìm hiểu ứng dụng thực tế |
| Nguyễn Hồng Khoa  Vũ Tuấn Hải | 30/11-02/12/2017 | Hoàn thành bài tiểu luận và chuẩn bị bài trình chiếu |
| Lê Huỳnh Thăng | 30/11-02/12/2017 | Hoàn thành bài trình chiếu |

*MỤC LỤC*

Nội dung

[MỞ ĐẦU 2](#_gjdgxs)

[Bảng Phân công nhiệm vụ thành viên nhóm 3](#_30j0zll)

[Bảng kế hoạch thực hiện 3](#_1fob9te)

[MỤC LỤC 4](#_3znysh7)

[CƠ SỞ MÔN HỌC 5](#_2et92p0)

[1.](#_tyjcwt) Định nghĩa về không gian vector 5

[2.](#_3dy6vkm) Cơ sở không gian vector 5

[3.](#_1t3h5sf) Chiều của không gian vector 5

[ỨNG DỤNG THỰC TẾ 6](#_4d34og8)

[1.](#_2s8eyo1) Thể hiện tài liệu và truy vấn như vector 6

[2.](#_17dp8vu) Tính trọng số nghịch đảo văn bản (Inverse document weighting - ) 7

[3.](#_3rdcrjn) Thể hiện không gian Vector dưới dạng ma trận 7

[4.](#_26in1rg) Sử dụng cosin để tính sự tương đồng giữa những cặp tài liệu – truy vấn 8

[TÀI LIỆU THAM KHẢO 10](#_35nkun2)

*CƠ SỞ MÔN HỌC*

1. **Định nghĩa về không gian vector**

Trong toán học, không gian vectơ là một tập hợp mà trên đó hai phép toán, phép cộng vectơ và phép nhân vectơ với một số, được định nghĩa và thỏa mãn các tiên đề.

Các không gian vectơ quen thuộc là không gian Euclid hai chiều và ba chiều. Các vectơ trong các không gian này là các cặp số thực hay các bộ 3 số thực, có trật tự, và thường được biểu diễn như là một vectơ hình học với độ lớn và phương hướng. [2]

1. **Cơ sở không gian vector**

Cơ sở B = của không gian vector là một hệ các phần tử sinh độc lập tuyến tính, tức là B thỏa 2 điều kiện sau:

1. được biểu diễn dưới dạng
2. Phương trình chỉ thỏa mãn khi

Các vô hướng được gọi là các tọa độ của vector trong cơ sở B = [3]

1. **Chiều của không gian vector**

Nếu tồn tại số nguyên dương n sao cho không gian vector V có một cơ sở gồm n vector, số nguyên này là duy nhất và được gọi là số chiều của không gian vector V, Kí hiệu

Chiều là số các vector của mọi cơ sở của V và cũng là số tối đại các vector độc lập tuyến tính của không gian vector V.

Không gian vector V có số chiều vô hạn chứa vô số vector độc lập tuyến tính. [3]

*ỨNG DỤNG THỰC TẾ*

Để tìm một cụm từ trong tập hợp các tài liệu sẵn có thì chúng ta cần so sánh cụm từ đó với các tài liệu sẵn có. Ý tưởng là chúng ta cũng xem các câu truy vấn như là một vector và chúng ta sẽ xếp hạng các tài liệu mà chúng ta có dựa vào sự tương đồng với câu truy vấn. Tài liệu nào càng gần với câu truy vấn thì có điểm cao hơn.

Khi so sánh hai vector chúng ta có thể tính khoảng cách giữa hai vector, hoặc tính góc tạo ra bởi hai vector. Tuy nhiên cách tính khoảng cách có nhược điểm không chính xác, bởi vì khoảng cách lớn với các vector có chiều dài khác nhau (ví dụ: tài liệu d′ là nhân đôi nội dung tài liệu d) mặc dù nội dung chúng tương đồng nhau. Do đó phương pháp sử dụng góc là tối ưu hơn.

Một ví dụ đơn giản khi bạn tìm kiếm một tài liệu trên hệ thống lên quan đến gà sử dụng phương pháp góc.

: Gà đẻ và ấp trứng như thế nào?

Giả sử hệ thống đang chứa 3 tài liệu (documents – D)

: Gà thuộc họ trĩ

: Gà chỉ đẻ một quả trứng mỗi ngày

: Gà đẻ trứng và ấp sau 21 ngày sẽ nở

Để trả về tài liệu có liên quan đến câu hỏi nhiều nhắt thì hệ thống sẽ thực hiện các bước sau đây.

1. **Thể hiện tài liệu và truy vấn như vector**

Hệ thống sẽ cắt tài liệu và truy vấn thành các từ, trong đó mỗi từ khác nhau được biểu diễn thành một chiều không gian.

Số lần từ t xuất hiện trong tài liệu D được gọi là tần suất của t trong D – kí hiệu là term frequency - . Những khác nhau trong cùng một D sẽ giúp mỗi D thể hiện dưới dạng một vector đếm.

1. **Tính trọng số nghịch đảo văn bản (Inverse document weighting - )**

Từ hiếm thì quan trọng hơn những từ có cao. Trong mỗi ngôn ngữ có những từ lặp đi lặp lại nhiều lần nhưng vô nghĩa. Ví dụ trong tiếng Anh là a, the, to, of v.v) hay trong tiếng việt là thì, là, mà, ...

Do đó chúng ta cần một cách đánh giá khác với các từ hiếm, vì nó sẽ mang nhiều thông tin hơn là những từ phổ biến trong văn bản.

Ví dụ trong một tập hợp các tài liệu về áo khoác, thì từ khóa “áo” sẽ có khả năng có mặt hầu hết trong tất cả các tài liệu. Để hạn chế nhược điểm này, chúng tôi giới thiệu cơ chế để giảm thiểu sự ảnh hưởng và tăng tính chính xác khi quyết định sự phù hợp của cặp tài liệu – truy vấn. Ý tưởng là giảm trọng số của từ nào có tần số từ cao, bằng cách lấy tổng số tài liệu (N) chia cho số tài liệu mà một từ xuất hiện.

Nếu gọi là số văn bản chứa một từ t thì là cách đánh giá ngược sự hữu ích của t. ( bé hơn N, số tài liệu trong tập hợp mà chúng ta có).

Chúng ta định nghĩa trọng số của một từ t bởi . [4]

1. **Thể hiện không gian Vector dưới dạng ma trận**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | *tf* | | | | |  | |
| Từ | Q | D1 | D2 | D3 |  |  |  |
| Gà | 1 | 1 | 1 | 1 | 3 | 1 | 0 |
| Đẻ | 1 | 0 | 1 | 1 | 2 | 1.5 | 0.176 |
| Ấp | 1 | 0 | 0 | 1 | 1 | 3 | 0.477 |
| Trứng | 1 | 0 | 1 | 1 | 2 | 1.5 | 0.176 |

Nhờ vào việc tính mà hệ thống có thể xác định được những từ nào là quan trọng thật sự, điều này vô cùng hữu ích đối với những không gian có đến hàng triệu chiều, những truy vấn – tài liệu dài dòng phức tạp hơn.

1. **Sử dụng cosin để tính sự tương đồng giữa những cặp tài liệu – truy vấn**

Như đã đề cập thì chúng tôi sẽ sử dụng công thức tính góc sau:

, khi tài liệu và truy vấn không có từ chung nào. [4]

Trước hết ta sử dụng công thức Pythagoras để tính độ dài mỗi vector

|V| =

|| = , || =

|| = , |Q| =

Tiếp đó tính tích vô hướng

= 1x1 + 1x0 + 1x0 + 1x0 = 1

= 1x1 + 1x1 + 1x0 + 1x0 = 2

= 1x1 + 1x1 + 1x1 + 1x1 = 4

Việc cuối cùng là tính Cosin

Từ kết quả trên, được xếp hạng 1, và xếp hạng 2 và 3. sẽ xuất hiện đầu tiên khi lấy thông tin bằng công cụ tìm kiếm. [1]

*KẾT LUẬN*

Bài tiểu luận đã hoàn thành mục tiêu đặt ra là đưa ra được các khái niệm cơ bản về ứng dụng của Không gian véctơ trong truy vấn và xếp hạng dữ liệu, nội dung khá rõ ràng và đơn giản giúp người đọc hiểu được tổng quát cách ứng dụng véctơ để làm việc với dữ liệu.

Song, vẫn còn hạn chế về mặt nội dung do quá cơ bản, không nêu được chi tiết làm người đọc chưa nắm rõ cơ chế của việc ứng dụng véc tơ, ngoài ra do tiểu luận còn hạn chế về mặt giới hạn nội dung nên chưa thể giải thích rõ một số thuật ngữ chuyên môn trong tài liệu, làm người đọc chưa hiểu hết được về mặt nội dung tiểu luận.

Đánh giá chung: tiểu luận cơ bản đã hoàn thành được mục tiêu đặt ra đầu bài.

*TÀI LIỆU THAM KHẢO*

|  |  |
| --- | --- |
| [1] | R. B. J. E. E. Ogheneovo, “http://ijarcsse.com,” 6 2016. [Trực tuyến]. Available: http://ijarcsse.com/Before\_August\_2017/docs/papers/Volume\_6/5\_May2016/V6I4-01424.pdf. [Đã truy cập 23 11 2017]. |
| [2] | N. d. Internet, “wikipedia.org,” 13 9 2017. [Trực tuyến]. Available: https://vi.wikipedia.org/wiki/Kh%C3%B4ng\_gian\_vect%C6%A1. [Đã truy cập 11 11 2017]. |
| [3] | N. d. Internet, “wikipedia.org,” 17 7 2015. [Trực tuyến]. Available: https://vi.wikipedia.org/wiki/Kh%C3%B4ng\_gian\_vect%C6%A1. [Đã truy cập 11 11 2017]. |
| [4] | H. Neo, “butchiso.com,” 27 10 2016. [Trực tuyến]. Available: http://butchiso.com/2013/10/tim-hieu-ve-mo-hinh-khong-gian-vector.html. [Đã truy cập 27 11 2017]. |