

ỨNG DỤNG TOÁN HỌC TRONG XẾP HẠNG TRANG WEB

● NGUYỄN THỊ PHƯƠNG DUNG

TÓM TẮT:

Bài nghiên cứu tổng quan về thuật toán xếp và mô tả tầm quan trọng cũng như thuật toán ứng dụng Đại số tuyến tính - là một nhánh của Toán học trong thuật toán sắp xếp website. Từ đó, giúp người dùng thấy được vai trò của Toán học trong cuộc sống, chạy thuật toán ra sao để nâng cao khả năng ứng dụng toán học trong việc nâng cao thứ hạng website trong Google.

Từ khóa: toán học, xếp hạng trang web, ứng dụng toán học.

1. Đặt vấn đề

Ngày nay, trong thời đại công nghệ số, việc ứng dụng và sử dụng các kết quả trên internet đóng vai trò rất quan trọng trong nhiều lĩnh vực. Một trong những việc vô cùng quan trọng là tìm kiếm, sử dụng các kết quả đã biết, bằng cách truy vấn vào các trang web có liên quan mật thiết tới kết quả tìm kiếm. Toán học có vai trò to lớn trong việc sắp xếp thứ hạng của các website liên quan mà ta truy cập tới.

Do vậy, bài viết nghiên cứu “ứng dụng của Toán học trong xếp hạng trang web”, cụ thể là một nhánh Đại số tuyến tính trong việc xếp hạng thứ hạng trang web của google, từ đó cho thấy được tầm quan trọng của việc ứng dụng Đại số tuyến tính trong xếp hạng thứ hạng trang web của Google và thuật toán chạy của nó.

2. Tổng quan về thuật toán xếp

Chúng ta đang sống trong thời đại mà máy tính, Internet có vai trò vô cùng quan trọng và ảnh hưởng sâu rộng tới hầu hết mọi lĩnh vực của cuộc sống của chúng ta chỉ thông qua một cái nhấp chuột. Bất kể một thông tin nào chúng ta cần tìm kiếm, chỉ cần mở công cụ tìm kiếm yêu thích của bạn như google,

yahoo..., nhập từ khóa, ngay lập tức công cụ tìm kiếm sẽ hiển thị tất cả các trang web có liên quan cho việc tìm kiếm.

Vấn đề cần là tìm hiểu để làm thế nào cho một công cụ tìm kiếm thực sự hoạt động. Thoạt nhìn có vẻ hợp lý khi tưởng tượng công cụ tìm kiếm làm là giữ một chỉ mục của tất cả các trang web, khi người dùng nhập một truy vấn tìm kiếm, công cụ sẽ duyệt qua chỉ mục của nó và đếm số lần xuất hiện của từ khóa trong đó. Những trang “chiến thắng” trong lần tìm kiếm là những trang có số lần xuất hiện cao nhất của “từ khóa” và trang đó sẽ được hiển thị lại cho người dùng. Đây chính là “bức tranh” mô tả việc tìm kiếm của Internet vào đầu những năm 90, khi các công cụ tìm kiếm đầu tiên sử dụng hệ thống xếp hạng dựa trên văn bản để quyết định trang nào phù hợp với một truy vấn nhất định. Tuy nhiên, có một số vấn đề nảy sinh với cách tiếp cận này. Ví dụ, tìm kiếm một thuật ngữ phổ biến, chẳng hạn như “internet”. Trang đầu tiên được hiển thị bởi một trong những công cụ tìm kiếm đầu tiên được viết bằng tiếng Trung Quốc, với sự xuất hiện lặp đi lặp lại từ “internet” và không chứa thông tin nào khác về internet.

Tính hữu ích của công cụ tìm kiếm phụ thuộc vào mức độ liên quan của tập hợp kết quả mà nó mang lại. Tất nhiên, có hàng triệu trang web bao gồm một từ hoặc cụm từ cụ thể. Tuy nhiên, một số trang trong số chúng sẽ phù hợp hơn, phổ biến hoặc có quyền hơn những trang khác. Người dùng không có khả năng hoặc kiên nhẫn để quét qua tất cả các trang có chứa các từ đã truy vấn. Người ta hy vọng các trang có liên quan sẽ được hiển thị trong khoảng từ 20 - 30 trang đầu tiên do công cụ tìm kiếm trả về.

Các công cụ tìm kiếm hiện đại sử dụng các phương pháp xếp hạng kết quả để cung cấp kết quả “tốt nhất” trước tiên. Phương pháp này phức tạp hơn so với chỉ số xếp hạng văn bản thuần túy. Một trong những thuật toán nổi tiếng và có ảnh hưởng nhất để tính toán mức độ liên quan của các trang web là thuật toán xếp hạng trang được sử dụng bởi công cụ tìm kiếm Google. Nó được phát minh bởi Larry Page và Sergey Brin, khi họ còn đang là sinh viên tại Stanford và nó đã trở thành nhãn hiệu của Google vào năm 1998. Ý tưởng mà xếp hạng trang đưa ra là tầm quan trọng của bất kỳ trang web nào cũng có thể được đánh giá bằng cách xem các trang liên kết tới. Nếu chúng ta tạo ra một trang web i nào đó, mà chứa đựng các siêu liên kết tới trang web, điều này có nghĩa chúng ta xét j cũng là quan trọng và nó liên quan tới chủ đề tìm kiếm của chúng ta.

Nếu còn có nhiều trang web liên quan tới trang j , điều này cũng nói lên rằng trang web j rất quan trọng. Mặt khác, nếu trang web j chỉ có 1 liên kết, nhưng đến từ trang có thẩm quyền k (như Google.com; cnn.com;...), thì điều đó nói lên rằng trang web k chuyển quyền hạn của mình cho trang web j , nói cách khác trang web k khẳng định trang web j là quan trọng. Chúng ta đang quan tâm tới mức độ phổ biến hay quyền hạn của

trang web, ta có thể lặp lại việc xếp hạng mỗi trang web dựa trên thứ hạng của các trang liên kết ngược (trở đến) tới nó.

3. Tầm quan trọng của ứng dụng toán học trong xếp hạng trang web

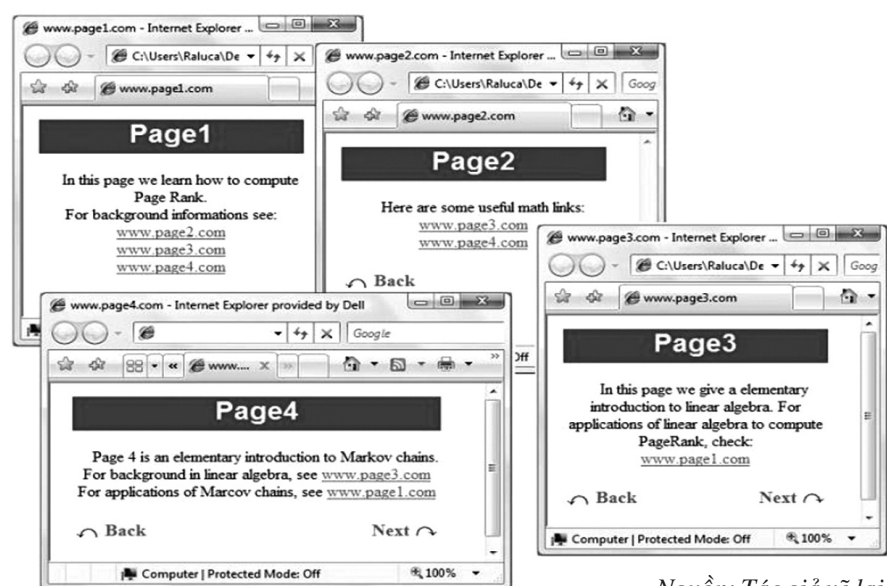
Với mỗi truy vấn mà người dùng truy cập vào để tìm kiếm thông tin, nếu như không có cách sắp xếp thứ hạng các trang web liên quan tới thông tin ta tìm kiếm, chúng ta không thể tìm thấy thông tin giá trị nhất và cũng không thể tìm thấy được những giá trị cốt lõi mà ta quan tâm. Nhờ có thuật toán sắp xếp thứ hạng trang web của Google, chúng ta có thể thấy và tìm hiểu kỹ lưỡng những trang có thông tin cần thiết nhất, gần nhất với những gì chúng ta muốn tìm hiểu. Vì vậy, việc xếp hạng trang web là vô cùng quan trọng trong thời đại công nghệ số.

4. Mô tả thuật toán thông qua tình huống cụ thể

Để đạt được mục đích này, chúng ta hình dung mạng web dưới dạng một biểu đồ có hướng, với các nút được biểu thị bằng các trang web, các cạnh được biểu thị bằng các liên kết giữa các trang web.

Ví dụ: giả sử chúng ta có một mạng Internets nhỏ chỉ gồm 4 trang: www.page1.com; www.page2.com; www.page3.com; www.page4.com, các trang này tham chiếu lẫn nhau theo cách được gợi ý bởi các hình như mô tả Hình 1.

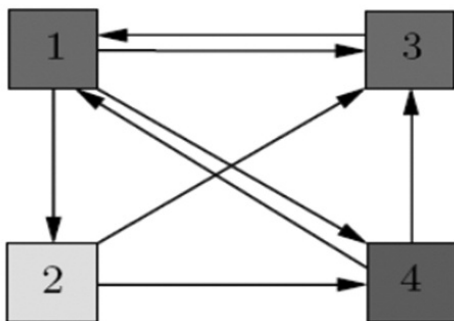
Hình 1: 4 website giả định



Nguồn: Tác giả vẽ lại

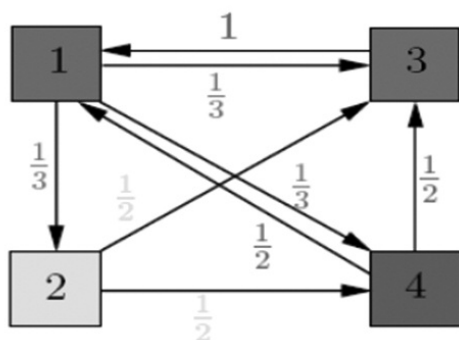
Ta có thể mô tả mối liên quan các trang web trên các bức tranh trên bởi một biểu đồ có hướng với 4 nút, mỗi nút đại diện cho 1 trang web. Khi trang web i tham chiếu tới trang web j , ta sẽ có một cạnh có hướng từ i tới j trong biểu đồ. Với mục đích tính toán thứ hạng của các trang, chúng ta sẽ bỏ qua mọi liên kết điều hướng như nút quay lại, nút tiếp theo, vì ta chỉ quan tâm đến các kết nối giữa các trang web khác nhau. Trên Hình 1 ta thấy, trang 1 liên kết tới tất cả các trang khác, vì vậy nút 1 trong biểu đồ sẽ có các cạnh có hướng đi đến tất cả các nút khác. Trang 3 chỉ có 1 liên kết tới trang 1, do đó nút 3 sẽ chỉ có 1 cạnh có hướng đi tới nút 1. Sau khi phân tích từng trang web, chúng ta nhận được mối liên kết như Biểu đồ 1 mô tả.

Biểu đồ 1: Mối liên quan của 4 trang web



Trong Biểu đồ 1, mỗi trang sẽ chuyển đều tầm quan trọng của nó sang mỗi trang khác mà nó liên kết đến. Nút 1 có 3 cạnh hướng ra ngoài, vì vậy sẽ truyền tầm quan trọng của nó cho 3 nút còn lại. Nút 3 chỉ có 1 cạnh hướng ra ngoài, vì vậy nó sẽ chuyển toàn bộ tầm quan trọng của nó

Biểu đồ 2: Đánh số Mối liên quan của 4 trang web



cho nút 1. Tổng quát, nếu một nút có k cạnh ra, nó sẽ chuyển tầm quan trọng của nó cho k nút ra mà nó liên kết đến. Ta sẽ mô tả một cách dễ hiểu, thông qua việc đánh trọng số thể hiện như tại Biểu đồ 2.

Ta ký hiệu A là ma trận chuyển tiếp cho sơ đồ trên, khi đó:

Dưới góc nhìn của Toán học, ta sẽ đi tìm vector page rank của các mô tả trên bằng một số phương pháp sau đây:

$$A = \begin{bmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

- Góc nhìn trên quan điểm Hệ động lực:

Ta luôn giả sử rằng, tầm quan trọng ban đầu được phân bổ đều trên 4 nút, vì vậy, mỗi nút sẽ có giá trị về “tầm quan trọng” là $1/4$. Xét vector $v = (1/4; 1/4; 1/4; 1/4)$ là vector biểu thị tầm quan trọng của việc xếp hạng ban đầu. Mỗi liên kết đến nút i sẽ làm tăng tầm quan trọng của 1 trang web, vì vậy ở bước 1, ta sẽ cập nhật tầm quan trọng của từng trang web bằng cách thêm giá trị hiện tại của “tầm quan trọng” của các liên kết đến. Việc này đồng nghĩa với phép nhân ma trận $v_1 = Av$; $v_2 = A(Av) = A^2 v$. Ta hoàn toàn có thể lặp lại quy trình cho tới khi $A^k v = A^{k+1} v$ thì dừng lại. Khi đó $A^k v$ được gọi là vector page rank.

Bằng sơ đồ trên, ta có các phép nhân ma trận sau đây:

$$\begin{aligned} v &= \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.15 \end{pmatrix}; Av = \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix}; A^2 v = \begin{pmatrix} 0.43 \\ 0.12 \\ 0.27 \\ 0.16 \end{pmatrix} \\ A^3 v &= \begin{pmatrix} 0.35 \\ 0.14 \\ 0.29 \\ 0.20 \end{pmatrix}; A^4 v = \begin{pmatrix} 0.39 \\ 0.11 \\ 0.29 \\ 0.19 \end{pmatrix}; A^5 v = \begin{pmatrix} 0.39 \\ 0.13 \\ 0.28 \\ 0.19 \end{pmatrix} \\ A^6 v &= \begin{pmatrix} 0.38 \\ 0.13 \\ 0.29 \\ 0.29 \end{pmatrix}; A^7 v = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.29 \end{pmatrix}; A^8 v = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.29 \end{pmatrix} \end{aligned}$$

Dãy các tính toán trên sẽ dừng tại A^7v , vector đó được gọi là vector page rank:

$$v^* = (0.38; 0.12; 0.29; 0.29).$$

- Để tính vector page rank, ta có thể dùng một công cụ khác của toán học, đó là dùng Đại số tuyến tính. Ta sẽ minh họa cách tính vector page rank trên, cụ thể như sau:

+ Ký hiệu x_1, x_2, x_3, x_4 là tầm quan trọng của 3 trang cần quan tâm, với sơ đồ trong Hình 1, ta có hệ phương trình tuyến tính sau đây:

$$\begin{cases} x_1 = & 1x_3 + 0.5x_4 \\ x_2 = & 1/3x_1 \\ x_3 = 1/3x_1 + 1/2x_2 + 1/2x_4 \\ x_4 = & 1/3x_1 + 1/2x_2 \end{cases}$$

Nếu viết dưới dạng ma trận, bài toán trên trở thành tìm vector $x = (x_1, x_2, x_3, x_4)$ sao cho $Ax = x$. Điều đó có nghĩa là tìm vector riêng ứng với giá trị riêng là 1 của ma trận A. Bằng kiến thức của môn đại số tuyến tính, ta có thể dễ dàng tính toán được vector riêng đó có giá trị là:

$$V^* = 1/31(12; 4; 9; 6) \approx (0.38; 0.12; 0.29; 0.29)$$

Vector v^* đó được gọi là page rank vector của ví dụ trong sơ đồ của Biểu đồ 2.

+ Ngoài ra, chúng ta có thể tìm cách xếp hạng trang web theo quan điểm của xác suất:

Tầm quan trọng của mỗi trang web được đo bằng mức độ phổ biến của nó - có bao nhiêu trang web liên kết đến nó, nên ta có thể xem tầm quan trọng của trang i là xác suất mà một người lướt web ngẫu nhiên trên internet mở trình duyệt tới bất kỳ trang nào và bắt đầu theo dõi các siêu liên kết, truy cập trang i. Ta có thể diễn giải các trọng số mà ta gán cho các cạnh của biểu đồ theo cách của xác suất. Nếu một người lướt ngẫu nhiên và hiện đang xem trang 2, có xác suất là 0.5 sẽ truy cập trang 3 và 0.5 sẽ truy cập trang 4. Ta có thể lập mô hình quá trình như một bước đi ngẫu nhiên trên đồ thị. Mỗi trang có xác suất 0.25 được chọn làm điểm bắt đầu. Vì vậy, phân bố xác suất ban đầu là vector $v = (0.25; 0.25; 0.25; 0.25)$. Xác suất trang mà ta sẽ truy cập ở bước thứ 2 là Av , bước 3 là $A^2x, \dots$ dãy

vector này sẽ hội tụ tới 1 vector duy nhất v^* , v^* được gọi là page rank vector. Thành phần thứ i của v^* là xác suất để tại mỗi thời điểm, một người lướt web ngẫu nhiên có thể truy cập vào trang i. Các tính toán giống hệt tính toán trong việc tìm page rank vector khi dùng quan điểm hệ động lực. Tuy nhiên, phép gán giá trị ta có thay đổi một chút. Ta sẽ minh họa công việc một cách tường minh, bằng chính ví dụ mô tả sơ đồ trong Biểu đồ 2.

Vector xếp hạng trang v^* mà chúng ta tính toán ở các phần trên, cho ta thấy trang 1 là trang có liên quan nhất (có thứ hạng cao nhất của truy vấn). Điều này có vẻ hơi lạ vì trang 1 có 2 liên kết ngược, trong khi trang 3 có 3 liên kết ngược. Nhìn vào sơ đồ ta thấy nút 3 chỉ có 1 cạnh đi ra nút 1 nên nó chuyển toàn bộ tầm quan trọng sang nút 1. Tương tự, khi một người lướt web chỉ theo dõi các siêu liên kết truy cập trang 3, khi đó người dùng chỉ có thể sẽ truy cập trang 1. Lưu ý rằng thứ hạng của mỗi trang không chỉ đơn giản là tổng trọng số của các cạnh đi vào nút. Theo cảm nhận ở bước thứ 1, một nút nhận được nhiều phiếu bầu quan trọng từ các nút lân cận trực tiếp của nó, ở bước 2 lại từ các nút lân cận của lân cận, ... Vì vậy, việc thay đổi biểu đồ web có thể dẫn tới những vấn đề khá ý nghĩa.

Với những diễn giải về ứng dụng của toán học trong xếp hạng trang web qua ứng dụng của một lĩnh vực Toán học vô cùng quan trọng - Đại số tuyến tính - thông qua một số ví dụ tường minh, giúp chúng ta phần nào hiểu được Toán học có vai trò vô cùng quan trọng trong mọi lĩnh vực của đời sống con người, đặc biệt là trong kỷ nguyên của máy tính, thời đại công nghệ 4.0.

5. Kết luận

Bài viết cho chúng ta thấy Toán học có vai trò to lớn trong cuộc sống của chúng ta. Đặc biệt cho chúng ta hiểu được tường minh trong một trường hợp cụ thể, cách thức Toán học được ứng dụng trong xếp hạng trang web nó thể hiện như thế nào. Điều đó có vai trò rất tích cực trong việc tạo hứng khởi cho người trẻ yêu thích lĩnh vực khoa học tự nhiên hơn nữa ■

TÀI LIỆU THAM KHẢO:

1. Beel, J., Gipp, B. (2008). The Potential of Collaborative Document Evaluation for Science. In: Buchanan, G., Masoodian, M., Cunningham, S.J. (eds) Digital Libraries: Universal and Ubiquitous Access to Information. ICADL 2008. Lecture Notes in Computer Science, vol 5362. Springer, Berlin, Heidelberg.
2. Meyer, C. (Ed.) (2000). Matrix Analysis and Applied Linear Algebra. Philadelphia, PA: SIAM: Society for Industrial and Applied Mathematics; Edition Unstated.
3. Zack, L., Lamb, R., & Ball, S. (2012). An application of Googles PageRank to NFL Rankings. Involve, 5(4), 463-471.

Ngày nhận bài: 14/8/2023

Ngày phản biện đánh giá và sửa chữa: 27/8/2023

Ngày chấp nhận đăng bài: 14/9/2023

Thông tin tác giả:

NGUYỄN THỊ PHƯƠNG DUNG

Học viện Ngân hàng

USING MATHEMATICS IN WEBSITE RANKING ALGORITHMS

● **NGUYEN THI PHUONG DUNG**

Banking Academy

ABSTRACT:

This paper provides an overview of the sorting algorithm and describes its importance as well as the application of linear algebra in website sorting algorithms. This paper is expected to help users understand the role of mathematics in life and how to better use mathematics in sorting algorithms to improve website rankings on Google.

Keywords: mathematics, website ranking, math apps.