**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CỦA SINH VIÊN**

**KHOA TOÁN ỨNG DỤNG NĂM 2022**

**TÌM HIỂU THUẬT TOÁN PAGERANK VÀ ỨNG DỤNG**

**Người thực hiện:**

***Trần Thị Lan Hương***

*Mai Ngọc Kiều*

*Trần Thị Thu Trang*

*Lê Quang Vũ*

**Sinh viên lớp Toán ứng dụng 1**

**Khoa Khoa học cơ bản**

**Người hướng dẫn:**

**PGS. TS**: *Trần Văn Long*

**1**

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CỦA SINH VIÊN**

**KHOA TOÁN ỨNG DỤNG NĂM 2022**

**TÌM HIỂU THUẬT TOÁN PAGERANK VÀ ỨNG DỤNG**

**Người thực hiện :**

***Trần Thị Lan Hương******<nữ>***

*Mai Ngọc Kiều <nam>*

*Trần Thị Thu Trang<nữ>*

*Lê Quang Vũ<nam>*

**Sinh viên lớp: Toán ứng dụng 1**

**Khoa: Khoa học cơ bản**

**Ngành học: Toán tin**

**Số năm đào tạo: 4 năm**

**Người hướng dẫn:**

**PGS. TS**: *Trần Văn Long*

**2**

**0363358757**

***Mục lục***

**TÌM HIỂU THUẬT TOÁN PAGERANK VÀ ỨNG DỤNG**

**Chương 1: Thuật toán PAGERANK**

1. Phát triển một công thức để xếp hạng các trang.

1.1. Ý tưởng cơ bản.

1.2. Những thiếu sót

1.2.1. Xếp hạng Nonunique.

1.2.2. Nút nguy hiểm.

1. Một phương thức của dim(V\_1(A))>1.
   1. Một sửa đổi với ma trận liên kết A.
   2. Phân tích ma trận M.
2. Tính toán điểm vecto riêng quan trọng.

**Chương 2: Ứng dụng của thuật toán PageRank trong xếp hạng đội tuyển bóng đá quốc gia**

1. Giới thiệu
2. Dữ liệu và phương pháp
3. Dữ liệu.
4. Phương pháp.
5. Ví dụ.
6. Kết quả và thảo luận
7. Phần kết luận và tài liệu tham khảo

**3**

**Chương I: Thuật toán PageRank**

1. Phát triển một công thức để xếp hạng các trang.

1.1. Ý tưởng cơ bản.

Trong phần tiếp theo, chúng tôi sẽ sử dụng cụm từ “Điểm quan trọng” hoặc “Điểm số” cho bất kì xếp hạng định lượng nào về mức độ quan trọng của trang Web. Điểm quan trọng cho bất kì trang Web nào sẽ luôn là một số thực không âm. Ý tưởng cốt lõi trong việc ấn định điểm cho bất kì trang Web nhất định nào là điểm của trang đó được lấy từ các liên kết đến trang đó từ các trang Web khác. Các liên kết đến một trang nhất định được gọi là liên kết ngược cho trang đó. Do đó, web trở thành một nền dân chủ nơi các trang bỏ phiếu cho tầm quan trọng của các trang khác bằng cách liên kết đến chúng.

3

1

4

2

Hình 1: Một ví dụ chỉ có bốn trang. Mũi tên từ trang A đến trang B chỉ ra một liên kết từ trang A đến trang B.

Giả sử trang Web chứa n trang, mỗi trang được lập bởi một số nguyên k, 1 ≤ k ≤ n. Một ví dụ điển hình được minh họa trong hình 1, trong đó một mũi tên từ trang A đến trang B chỉ ra một liên kết từ trang A đến trang B. Một trang Web như vậy là một ví dụ về biểu đồ có hướng. Chúng ta sẽ sử dụng xₖ để biểu thị điểm quan trọng của trang k trong Web. xₖ là không âm và chỉ ra rằng trang j quan trọng hơn trang k (vì vậy chỉ ra rằng trang j có điểm ít quan trọng nhất có thể).

Một cách tiếp cận rất đơn giản là lấy là số lượng liên kết ngược cho trang k. Trong ví dụ ở hình 1, chúng ta có , do đó trang 3 là trang quan trọng nhất, trang 1 và 4 xếp thứ hai và trang 2 là ít quan trọng nhất. Một liên kết đến trang k trở thành bình chọn cho tầm quan trọng của trang k.

Cách tiếp cận này bỏ qua một tính năng quan trọng mà người ta mong đợi một thuật toán xếp hạng có, cụ thể là một liên kết đến trang k từ một trang quan trọng sẽ nâng cao điểm số quan trọng của trang k nhiều hơn một liên kết từ một trang không quan trọng. Ví dụ, một liên kết đến trang chủ của bạn trực tiếp từ Yahoo! phải nâng cao điểm số cho trang của bạn hơn nhiều so với một liên kết [www.kurtbryan.com](http://www.kurtbryan.com) (không liên quan đến tác giả). Trong trang Web của hình 1, trang 1 và 4 đều có hai liên kết ngược: mỗi liên kết đến trang khác, nhưng liên kết ngược thứ hai của trang 1 là từ trang 3 có vẻ quan trọng, trong khi liên kết ngược thứ hai của trang 4 là từ trang 1 tương đối không quan trọng. Do đó, có lẽ chúng ta nên đánh giá mức độ quan trọng của trang 1 cao hơn mức độ quan trọng của trang 4.

Như một lỗ lực đầu tiên trong việc kết hợp ý tưởng này, hãy tính điểm của trang j là tổng điểm của tất cả các trang liên kết đến trang j. Ví dụ, hãy xem xét trang Web của hình 1. Điểm của trang 1 sẽ được xác định bởi quan hệ . Vì và sẽ phụ thuộc vào , nên lược đồ này có vẻ tự tham chiếu một cách kì lạ, nhưng đó là cách tiếp cận mà chúng tôi sẽ sử dụng, với một sửa đổi nữa. Cũng như trong các cuộc bầu cử, chúng tôi không muốn một cá nhân nào có được ảnh hưởng chỉ bằng cách bỏ phiếu bầu. Tương tự như vậy, chúng tôi tìm kiếm một kế hoạch trong đó có một trang Web không có thêm ảnh hưởng chỉ đơn giản bằng cách liên kết đến nhiều trang khác. Nếu trang j chứa liên kết , một trong số đó liên kết đến trang k lên thay vì . Trong lược đồ này, mỗi trang web nhận được tổng số một phiếu bầu, được tính theo điểm của trang web đó, được chia đều cho tất cả các liên kết đi của nó. Để định lượng điều này cho một trang Web n trang, đặt biểu thị tập hợp các trang có liên kết đến trang k, nghĩa là, là tập hợp các liên kết ngược của trang k. Đối với mỗi k chúng tôi yêu cầu:

Trong đó là số liên kết đi từ trang j (phải là số dương vì nếu , liên kết ít nhất trang k). Chúng tôi sẽ giả định rằng một liên kết từ một trang đến chính nó sẽ không được tính. Trong “nền dân chủ của web” này, bạn không được bỏ phiếu cho chính mình!

Hãy áp dụng cách tiếp cận này cho trang 4 của hình 1. Đối với trang 1, chúng ta có +, vì trang 3 và 4 là các liên kết ngược cho trang 1 và 3 chỉ chứa một liên kết, trong khi trang 4 chứa hai liên kết (chia đôi phiếu bầu của nó). Tương tự, , . Các phương trình tuyến tính này có thể được viết Ax=x, trong đó và

Điều này biến bài toán xếp hạng trang Web thành bài toán “tiêu chuẩn” để tìm vecto riêng cho ma trận vuông! (Nhớ lại rằng các giá trị riêng và các vecto riêng x của ma trận A thỏa mãn phương trình Ax=x, , theo định nghĩa). Do đó chúng tôi tìm kiếm một vecto riêng x với giá trị riêng 1 cho ma trận A. Chúng tôi sẽ gọi A là “ma trận liên kết” cho Web nhất định.

Hóa ra là ma trận liên kết A trong (2) thực sự có các vecto riêng với giá trị riêng 1, cụ thể là, tất cả các bội của vecto (nhớ lại rằng mọi bội số của một vecto riêng lại chính là vecto riêng). Hãy đồng ý chia tỉ lệ “vecto riêng có điểm quan trọng” này để các thành phần tổng bằng 1. Trong trường hợp này chúng ta thu được . Lưu ý rằng xếp hạng này khác với xếp hạng được tạo bằng cách chỉ đếm các liên kết ngược. Có vẻ ngạc nhiên rằng trang 3, được liên kết bởi tất cả các trang khác, không phải là trang quan trọng nhất. Để hiểu điều này, hãy lưu ý rằng trang 3 chỉ liên kết đến trang 1 và do đó, bỏ toàn bộ phiếu bầu cho trang 1. Điều này, với phiếu bầu của trang 2, kết quả là trang 1 nhận được điểm quan trọng cao nhất.

Nói một cách tổng quát hơn, ma trận A cho bất kỳ trang web nào phải có 1 làm giá trị riêng nếu trang web được đề cập không có các nút treo (các trang không có liên kết đi). Để thấy được điều này, trước tiên hãy lưu ý rằng đối với một web tổng quát gồm n trang, công thức (1) tạo ra ma trận A với nếu trang j liên kết đến trang i và nếu ngược lại. Sau đó, cột thứ j của A chứa mục nhập khác không, mỗi mục bằng và cột do đó tổng bằng 1. Điều này thúc đẩy định nghĩa sau, được sử dụng trong nghiên cứu chuỗi Markov.

Định nghĩa 1. Ma trận vuông được gọi là ma trận ngẫu nhiên cột nếu tất cả các phần tử của nó không âm và các phần tử trong mỗi cột tổng bằng 1.

Ma trận A cho một trang web không có nút treo là ngẫu nhiên cột. Chúng tôi không chứng minh những điều sau đây.

Mệnh đề 1. Mọi ma trận ngẫu nhiên cột đều có 1 là giá trị riêng.

Bằng chứng. Gọi A là ma trận ngẫu nhiên cột n × n và gọi e là vectơ cột thời gian với tất cả các mục bằng 1. Nhớ lại rằng A và phép chuyển vị của nó có cùng giá trị riêng. Vì A là ngẫu nhiên cột nên dễ dàng thấy , do đó 1 là giá trị riêng của và với A.

Theo cách sau, chúng tôi sử dụng để biểu thị không gian riêng cho giá trị riêng 1 của ma trận ngẫu nhiên cột A.

1.2. Những thiếu sót.

Một số khó khăn nảy sinh khi sử dụng công thức (1) để xếp hạng các trang Web. Trong phần này chúng ta thảo luận về 2 vấn đề: Web có thứ hạng không đơn nhất và Web có các nút treo lơ lửng.

1.2.1. Xếp hạng Nonunique.

Đối với thứ hạng của chúng tôi, điều mong muốn là thứ nguyên của bằng 1, để có một vecto riêng x với mà chúng tôi có thể sử dụng cho điểm quan trọng.

Điều này đúng trong trang web của Hình 1 và nói chung luôn đúng cho trường hợp đặc biệt của một trang web được kết nối mạnh mẽ (nghĩa là bạn có thể chuyển từ bất kỳ trang nào đến bất kỳ trang nào khác trong một số bước hữu hạn); xem Bài tập 10 dưới đây.

3

1

5

4

2

Hình 2: Một trang Web gồm 5 trang, bao gồm hai “mạng con” (trang 1 và 2) và W2 bị ngắt kết nối (trang 3,4,5).

Thật không may, không phải lúc nào ma trận liên kết A cũng đúng sẽ mang lại thứ hạng duy nhất cho tất cả các trang web. Hãy xem xét trang web trong Hình 2, ma trận liên kết là

Ở đây chúng ta thấy rằng là hai chiều; một cặp vecto cơ sở có thể là

. Nhưng lưu ý rằng bất kỳ sự kết hợp tuyến tính nào của hai vectơ này đều tạo ra một vectơ khác trong , ví dụ . Không rõ là cái nào, nếu có, trong các vecto riêng này chúng ta nên sử dụng cho bảng xếp hạng!

Không phải ngẫu nhiên mà đối với trang Web của hình 2, chúng ta thấy rằng . Đó là hệ quả của thực tế là nếu một trang web W, được coi như một đồ thị vô hướng (bỏ qua hướng mà mỗi mũi tên chỉ) , bao gồm r mạng con bị ngắt kết nối, sau đó , và do đó không có vectơ điểm quan trọng duy nhất . Điều này có ý nghĩa trực quan: nếu một web W bao gồm r các trang con bị ngắt kết nối , thì người ta sẽ gặp khó khăn trong việc tìm khung tham chiếu chung để so sánh điểm số của các trang trong một trang web con với các trang trong một mạng con khác.

Thật vậy, không khó để hiểu tại sao một Web W bao gồm r các trang con bị ngắt kết nối lại buộc . Giả sử một Web W có n trang và r mạng con thành phần . Gọi là số trang trong , Lập chỉ mục các trang trong với chỉ số từ 1 đến , các trang trong với chỉ số đến , các trang trong với chỉ số , v.v. Nói chung, đặt với , do đó chứa các trang từ Ví dụ, trong trang Web của hình 2, chúng ta có thể lấy , vì vậy chứa các trang 1 và 2 và chứa các trang 3,4, 5. Trang web trong Hình 2 là một ví dụ cụ thể về trường hợp tổng quát, trong đó ma trận A giả định cấu trúc đường chéo khối

Trong đó biểu thị ma trận liên kết cho Trên thực tế, có thể được coi là một trang Web theo đúng nghĩa của nó. Mỗi ma trận là cột ngẫu nhiên và do đó sở hữu một số vecto riêng với vecto riêng 1. Với mỗi i từ 1 đến r xây dựng một vecto trong đó có 0 thành phần cho tất cả các phần tử tương ứng với các khối khác với khối i. Ví dụ

Sau đó dễ dàng nhận thấy các vecto , độc lập tuyến tính vecto riêng A với giá trị riêng 1 bởi vì

Như vậy có thứ nguyên nhỏ nhất là r.

1.2.2. Nút nguy hiểm.

Một khó khăn khác có thể nảy sinh khi sử dụng ma trận A để tạo thứ hạng. Một web với các nút treo lơ lửng tạo ra một ma trận A chứa một hoặc nhiều cột có tất cả các số không. Trong trường hợp này A là cột tính ngẫu nhiên, nghĩa là, các tổng cột của A đều nhỏ hơn hoặc bằng 1. Một ma trận như vậy phải có tất cả các giá trị riêng nhỏ hơn hoặc bằng 1 về độ lớn, nhưng 1 không cần thực sự là một giá trị riêng cho A. Tuy nhiên, các trang trong web với các nút treo lơ lửng vẫn có thể được xếp hạng bằng cách sử dụng một kỹ thuật tương tự. Ma trận dấu ngẫu nhiên tương ứng phải có một giá trị riêng dương λ ≤ 1 và một giá trị riêng x tương ứng với các mục nhập không âm (được gọi là vecto riêng Perron) có thể được sử dụng để xếp hạng các trang web. Xem Bài tập 4 bên dưới. Tuy nhiên, chúng tôi sẽ không xem xét thêm vấn đề của các nút treo ở đây.

Bài tập 1. Giả sử những người sở hữu trang 3 trong trang web của Hình 1 đang tức giận bởi thực tế là điểm quan trọng của nó, được tính bằng công thức (1), thấp hơn điểm của trang 1. Trong một nỗ lực để tăng điểm của trang 3 , họ tạo ra một trang 5 liên kết đến trang 3; trang 3 cũng liên kết đến trang 5. Điều này có thúc đẩy điểm số của trang 3 cao hơn điểm của trang 1 không?

Bài tập 2. Xây dựng một web bao gồm ba hoặc nhiều mạng con và xác minh rằng bằng (hoặc vượt quá) số lượng thành phần trong web.

Bài tập 3. Thêm một liên kết từ trang 5 đến trang 1 trong trang web của Hình 2. Trang web kết quả, được coi như một đồ thị vô hướng, được kết nối với nhau. Chiều của là gì?

Bài tập 4. Trong trang web của Hình 1, xóa liên kết từ trang 3 đến trang 1. Trong trang web kết quả, trang 3 bây giờ là một nút treo. Thiết lập ma trận điểm ngẫu nhiên tương ứng và tìm giá trị riêng dương (Perron) lớn nhất của nó. Tìm một ký hiệu riêng Perron không âm cho giá trị riêng này và chia tỷ lệ vectơ để các thành phần của nó tổng thành 1. Xếp hạng kết quả có hợp lý không?

Bài tập 5. Chứng minh rằng trong bất kỳ trang web nào, điểm quan trọng của một trang không có liên kết ngược là 0.

Bài tập 6. Ngụ ý trong phân tích của chúng tôi cho đến thời điểm này là khẳng định rằng cách các trang của web W được lập chỉ mục không ảnh hưởng đến điểm quan trọng được chỉ định cho bất kỳ trang nhất định nào. Chứng minh điều này như sau: Cho W chứa n trang, mỗi trang được gán chỉ số từ 1 đến n, và gọi A là ma trận liên kết thu được. Giả sử sau đó chúng ta hoán vị các chỉ số của trang i và j (vì vậy trang i bây giờ là trang j và ngược lại). Gọi  là ma trận liên kết cho web được gắn nhãn lại.

* Lập luận rằng  = PAP, trong đó P là ma trận cơ bản thu được bằng cách chuyển các hàng i và j của ma trận nhận dạng n × n. Lưu ý rằng phép toán A → PA có tác dụng hoán đổi hàng i và j của A, trong khi A → AP hoán đổi cột i và j. Ngoài ra, P2 = I, ma trận nhận dạng.
* Giả sử rằng x là vecto riêng của A, do đó Ax = λx với λ nào đó. Chứng tỏ rằng y = Px là một hiệu riêng cho với giá trị riêng λ.
* Giải thích lý do tại sao điều này cho thấy rằng việc hoán đổi các chỉ số của hai trang bất kỳ sẽ không thay đổi điểm số quan trọng và sử dụng kết quả này để lập luận rằng bất kỳ đột biến nào của chỉ số trang đều khiến điểm số quan trọng không thay đổi.

1. Một phương thức của

Cần phải có một lượng lớn tài nguyên máy tính để xác định một vecto riêng cho ma trận liên kết tương ứng với một trang web chứa hàng tỷ trang. Do đó, điều quan trọng là phải biết rằng thuật toán của chúng tôi sẽ mang lại một bộ xếp hạng web hợp lý duy nhất. Phân tích ở trên cho thấy rằng nỗ lực xếp hạng trang web đầu tiên của chúng tôi dẫn đến khó khăn nếu web không được kết nối. Và World Wide Web, được coi như một đồ thị vô hướng, chứa nhiều thành phần rời rạc.

Dưới đây chúng tôi trình bày và phân tích một sửa đổi của phương pháp trên được bảo đảm để khắc phục nhược điểm này. Phân tích sau đây về cơ bản là một trường hợp đặc biệt của định lý Perron-Frobenius, và chúng tôi chỉ chứng minh những gì chúng tôi cần cho cation ứng dụng này.

2.1. Một sửa đổi với ma trận liên kết A.

Đối với một trang web n không có nút treo lơ lửng, chúng tôi có thể tạo điểm số quan trọng rõ ràng như sau, bao gồm cả trường hợp web có nhiều mạng con.

Gọi S là ma trận n × n với tất cả các mục là 1 / n. Ma trận S là ngẫu nhiên cột, và dễ dàng kiểm tra rằng là một chiều. Chúng ta sẽ thay thế ma trận A bằng ma trận



Trong đó . M là trung bình có trọng số của A và S. Giá trị của m được Google sử dụng được báo cáo là 0.15 Với bất kì, ma trận M là ngẫu nhiên cột và chúng ta chỉ ra rằng . Do đó, M có thể được sử dụng để tính điểm quan trọng rõ ràng. Trong trường hợp khi m = 0, chúng ta có bài toán ban đầu, khi đó M = A. Ở cực trị khác là m = 1, cho ra M = S. Đây là trường hợp quân bình cuối cùng: biểu tượng chuẩn hóa duy nhất x với giá trị riêng 1 có cho tất cả i và tất cả các trang web đều được đánh giá là quan trọng như nhau. Sử dụng M thay cho A mang lại cho một trang web không có liên kết ngược (nút treo) điểm quan trọng là m / n (Bài tập 9) và ma trận M là hàm ngẫu nhiên cho bất kỳ m <1 nào vì ma trận A là hàm ngẫu nhiên. Do đó, công thức đã sửa đổi mang lại điểm số quan trọng khác không cho các liên kết treo (nếu m> 0) nhưng không giải quyết được vấn đề của các nút treo. Trong phần còn lại của bài viết này, chúng tôi chỉ xem xét các trang web không có nút treo.

Phương trình x=Mx cũng có thể được chuyển thành



Trong đó s là một vectơ cột với tất cả các mục nhập 1 / n. Chú ý rằng

Dưới đây chúng tôi sẽ chứng minh rằng luôn là một chiều, nhưng trước tiên chúng ta hãy xem xét một vài ví dụ.

Ví dụ 1. Đối với trang web gồm bốn trang trong Hình 1 với ma trận A được cho bởi (2), công thức mới cho (với m = 0,15)

và mang lại điểm số quan trọng ≈ 0.368, ≈ 0.142, ≈ 0.288 và ≈ 0.202. Điều này mang lại cùng thứ hạng của các trang như tính toán trước đó, nhưng điểm hơi khác một chút.

Ví dụ 2 cho thấy rõ ràng hơn những lợi ích của việc sử dụng M thay cho A.

Ví dụ 2. Ví dụ thứ hai, đối với trang web của Hình 2 với m = 0.15, chúng ta thu được ma trận



Không gian thực sự là một chiều, với các thành phần mã hiệu chuẩn hóa là = 0.2, = 0.2, = 0.285, = 0.285 và = 0.03. Việc sửa đổi, sử dụng M thay vì A, cho phép chúng tôi so sánh các trang trong các mạng con khác nhau.

Mỗi mục nhập Mij của M được xác định bởi (3) là hoàn toàn dương, điều này thúc đẩy định nghĩa sau.

Định nghĩa 2. Ma trận M là dương nếu Mij > 0 với mọi i và j.

Đây là thuộc tính chính đảm bảo dim(V1(M)) = 1, mà chúng tôi chứng minh trong phần tiếp theo.

2.2. Phân tích ma trận M.

Lưu ý rằng Mệnh đề 1 chỉ ra rằng V1(M) là rỗng vì M là ngẫu nhiên. Mục tiêu của phần này là chỉ ra rằng V1(M) trên thực tế là một chiều. Đây là hệ quả của hai mệnh đề sau.

Mệnh đề 2. Nếu M là rõ ràng và cột ngẫu nhiên, thì bất kỳ dấu hiệu riêng nào trong V1(M) đều có tất cả các thành phần dương hoặc tất cả các thành phần âm.

Bằng chứng. Chúng tôi sử dụng bằng chứng mâu thuẫn. Đầu tiên lưu ý rằng trong tam giác chuẩn bất bình đẳng (với tất cả yi là thực), bất đẳng thức nghiêm ngặt khi yi có dấu hỗn. Giả sử x ∈ V1(M) chứa các phần tử cùng dấu. Từ x = Mx ta và các tổng và Mijxj có dấu hỗn hợp (vì Mij> 0). Kết quả là chúng ta có một sự bất bình đẳng nghiêm ngặt.



Tính tổng cả hai vế của bất đẳng thức (6) từ i = 1 đến i = n, và hoán đổi các tổng i và j. Sau đó, sử dụng thực tế rằng M là cột ngẫu nhiên ( với mọi j) để tìm , một mâu thuẫn. Do đó x không thể chứa cả phần tử âm và dương. Nếu xi ≥ 0 với mọi i (và không phải tất cả xi đều bằng 0), thì xi> 0 ngay sau và Mij> 0. Tương tự xi ≤ 0 với mọi i ngụ ý rằng mỗi xi <0.

Mệnh đề sau đây cũng sẽ hữu ích để phân tích dim(V1(M)).

Mệnh đề 3. Gọi v và w là các vectơ độc lập tuyến tính theo Rm, m ≥ 2.Vậy với một số giá trị của s và t, vectơ x = sv + tw có cả thành phần dương và âm.

Bằng chứng. Độc lập tuyến tính ngụ ý rằng cả v và w đều không bằng 0. Cho . Nếu d=0, thì v phải chứa các thành phần của dấu hỗn hợp, và lấy s = 1 và t = 0 sẽ đưa ra kết luận. Nếu d 0, đặt , t = 1 và x = sv + tw. Vì v và w độc lập nên x 0. Tuy nhiên,. Ta kết luận rằng x có cả thành phần dương và âm.

Bây giờ chúng ta có thể chứng minh rằng việc sử dụng M thay cho A mang lại xếp hạng rõ ràng cho bất kỳ trang web nào không có nút treo.

Bổ đề 3. Nếu M dương và cột ngẫu nhiên thì V1(M) có thứ nguyên là 1.

Bằng chứng. Chúng tôi một lần nữa sử dụng bằng chứng bằng cách mâu thuẫn. Giả sử có hai ký tự riêng độc lập tuyến tính v và w trong không gian con V1(M). Đối với bất kỳ số thực s nào, vectơ x = v + sw phải nằm trong V1(M) và do đó có các thành phần là tất cả âm hoặc tất cả dương. Nhưng theo Mệnh đề 3, đối với một số lựa chọn của s và t, vectơ x phải chứa các thành phần của dấu hỗn hợp, một điều mâu thuẫn. Chúng tôi kết luận rằng V1(M) không thể chứa hai vectơ độc lập tuyến tính và do đó nó có chiều 1.

Bổ đề 3 cung cấp “đường đột” cho phân tích của chúng ta về thuật toán xếp hạng sử dụng ma trận M (với 0 <m <1). Không gian V1(M) là một chiều, và hơn nữa, các vecto riêng liên quan có các thành phần hoàn toàn dương hoặc âm. Do đó, chúng tôi được đảm bảo sự tồn tại của một vecto riêng duy nhất x ∈ V1(M) với các thành phần dương sao cho .

Bài tập 7. Chứng minh rằng nếu A là ma trận ngẫu nhiên cột n × n và 0 ≤ m ≤ 1 thì M = (1 - m) A + mS cũng là ma trận ngẫu nhiên cột.

Bài tập 8. Chứng tỏ rằng tích của hai ma trận ngẫu nhiên cột cũng là ngẫu nhiên cột.

Bài tập 9. Chỉ ra rằng một trang không có liên kết ngược được cho điểm quan trọng theo công thức (4).

Bài tập 10. Giả sử A là ma trận liên kết của một web liên kết mạnh gồm n trang (bất kỳ trang nào cũng có thể được truy cập từ bất kỳ trang nào khác bằng cách tuân theo một số lượng liên kết hữu hạn). Chứng tỏ rằng dim(V1(A)) = 1 như sau. Gọi (Ak)ij biểu thị mục tiêu (i, j) của Ak.

* Lưu ý rằng tôi có thể truy cập trang từ trang j trong một bước khi và chỉ khi Aij> 0 (vì Aij> 0 có nghĩa là có một liên kết từ j đến i). Chỉ ra rằng (A2)ij> 0 khi và chỉ khi có thể truy cập trang i từ trang j trong đúng hai bước. Gợi ý: (A2)ij =; tất cả Aij đều không âm, do đó (A2)ij> 0 ngụ ý rằng đối với một số k thì cả Aik và Akj đều dương.
* Chỉ ra một cách tổng quát hơn rằng (Ap)ij> 0 khi và chỉ khi có thể truy cập trang tôi từ trang j trong p bước chính xác.
* Lập luận rằng (I + A + A2 + ··· + Ap)ij> 0 khi và chỉ khi có thể truy cập trang i từ trang j trong p hoặc ít bước hơn (lưu ý rằng p = 0 là một lựa chọn hợp pháp — bất kỳ trang nào cũng có thể đạt được từ chính nó trong không bước!)
* Giải thích tại sao I + A + A2 + ··· + An-1 là ma trận dương nếu web được kết nối mạnh.
* Sử dụng phần cuối cùng (và Bài tập 8) để chứng tỏ rằng là rõ ràng và cột ngẫu nhiên (và do đó theo Bổ đề 3, dim(V1(B))=1).
* Chứng tỏ rằng nếu x ∈ V1(A) thì x ∈ V1(B). Tại sao điều này ngụ ý rằng dim(V1(A)) = 1?

Bài tập 11. Hãy xem xét lại trang web trong Hình 1, có thêm một trang 5 liên kết đến trang 3, trong đó trang 3 cũng liên kết đến trang 5. Tính thứ hạng mới bằng cách tìm ký hiệu riêng của M (tương ứng với λ = 1) có các thành phần dương tổng bằng 1. Sử dụng m = 0.15.

Bài tập 12. Thêm trang thứ sáu liên kết đến mọi trang của web trong bài tập trước, nhưng không có liên kết trang nào khác. Xếp hạng các trang bằng cách sử dụng A, sau đó sử dụng M với m = 0.15 và so sánh kết quả.

Bài tập 13. Xây dựng một trang web bao gồm hai hoặc nhiều mạng con và xác định thứ hạng theo công thức (3).

Tính đến tháng 1 năm 2005, web chứa ít nhất tám tỷ trang — làm thế nào để người ta tính toán một vecto riêng cho ma trận tám tỷ x tám tỷ? Một cách tiếp cận hợp lý là một thủ tục lặp lại được gọi là phương pháp lũy thừa (cùng với các sửa đổi), bây giờ chúng ta sẽ kiểm tra cho trường hợp đặc biệt. Cần lưu ý rằng có nhiều phân tích bổ sung mà người ta có thể làm và có nhiều phương pháp được cải tiến để tính thứ hạng của PageRank. Tài liệu tham khảo [7] cung cấp một ví dụ điển hình và các tài liệu tham khảo bổ sung.

3. Tính toán điểm vecto riêng quan trọng.

Ý tưởng sơ lược đằng sau phương pháp lũy thừa để tính toán ký hiệu riêng của ma trận M là: Người ta bắt đầu với vectơ “điển hình” x0, sau đó tạo ra chuỗi xk = Mxk-1 (do đó xk = Mkx0) và để k tiến tới vô cùng. Vectơ xk, để xấp xỉ tốt, là một vrcto riêng cho giá trị riêng (độ lớn lớn nhất) của M. Tuy nhiên, tùy thuộc vào độ lớn của giá trị riêng này, vectơ xk cũng có thể phát triển mà không bị ràng buộc hoặc phân rã thành vectơ 0. Do đó, một thứ thường thay đổi tỷ lệ ở mỗi lần lặp, chẳng hạn, bằng cách tính, trong đó có thể là bất kỳ chuẩn vectơ nào.

Tuy nhiên, để đảm bảo rằng phương pháp lũy thừa hội tụ ở một tốc độ hợp lý, người ta thường yêu cầu giá trị riêng ưu thế λ = 1 đối với M phải đơn giản. Nghĩa là, đa thức đặc trưng cho M phải có dạng p (λ) = (λ - 1) q (λ) đối với một đa thức q (λ) bậc n - 1, trong đó (λ - 1) không chia q (λ). Ngoài ra, M không được sở hữu “các bộ định vị tổng quát” cho giá trị riêng λ = 1 ngoài các bộ định vị trong V1(M) (xem [1] để biết thêm về các bộ định vị tổng quát).

Trên thực tế, mệnh đề sau đây cung cấp những gì chúng ta cần để chứng minh rằng phương pháp lũy thừa hội tụ trong trường hợp này, nhưng không tham chiếu rõ ràng đến bản chất đại số của giá trị riêng λ = 1 chi phối. Nó cũng có thể được sử dụng để cho thấy rằng giá trị riêng này là đơn giản.

Định nghĩa 4: Chuẩn 1 của vecto v là 1

Mệnh đề 4. Gọi M là ma trận n × n cột ngẫu nhiên dương và gọi V là không gian con của Rn bao gồm các vectơ v sao cho. Khi đó Mv ∈ V với mọi v ∈ V, và 1 với v ∈ V bất kỳ, trong đó .

Bằng chứng. Để thấy rằng Mv ∈ V là đơn giản: Cho w = Mv, sao cho và

Do đó w = Mv ∈ V. Để chứng minh sự ràng buộc trong mệnh đề, hãy lưu ý rằng

trong đó ei = sgn(wi). Lưu ý rằng các ei không phải là tất cả cùng một dấu, vì (trừ khi w ≡ 0, trong trường hợp đó thì ràng buộc giữ rõ ràng). Đảo ngược tổng gấp đôi để có được



Trong đó Vì các ei có dấu hỗn hợp và với dễ thấy rằng

Do đó chúng tôi có thể ràng buộc

Cho Quan sát rằng c <1 và | aj | ≤ c với mọi j. Từ (7) chúng ta có mà chứng minh mệnh đề.

Mệnh đề 4 tạo tiền đề cho mệnh đề sau.

Mệnh đề 5. Mọi ma trận ngẫu nhiên cột dương M có một vectơ q duy nhất với các thành phần dương sao cho Mq = q với = 1. Vectơ q có thể được tính là q = limk → ∞ Mkx0 với bất kỳ dự đoán ban đầu nào x0 với các thành phần rõ ràng như = 1.

Bằng chứng. Từ Mệnh đề 1, ma trận M có 1 là giá trị riêng và theo Bổ đề 3, không gian con V1(M) là một chiều. Ngoài ra, tất cả các vectơ khác không trong V1(M) đều có thành phần hoàn toàn dương hoặc âm. Rõ ràng là tồn tại một vectơ duy nhất q ∈ V1(M) với các thành phần dương sao cho = 1.

Gọi x0 là vectơ bất kỳ trong Rn có các thành phần dương sao cho = 1. Ta có thể viết x0 = q + v, trong đó v ∈ V (V như trong Mệnh đề 4). Ta thấy rằng Mkx0 = Mkq + Mkv = q + Mkv. Kết quả là

1. Mkx0 – q = Mkv

Quy nạp đơn giản và Mệnh đề 4 cho thấy với 0 ≤c <1 (c như trong Mệnh đề 4) và do đó limk → ∞ = 0. Từ (8), chúng tôi kết luận rằng limk → ∞ Mkx0 = q.

Ví dụ 3. Gọi M là ma trận được xác định bởi (5) cho web của Hình 2. Ta lấy x0 = [0.24, 0.31, 0.08, 0.18, 0.19]T làm dự đoán ban đầu; nhớ lại rằng chúng ta có q = [0.2, 0.2, 0.285, 0.285, 0.03]T. Bảng dưới đây cho thấy giá trị của cho một số giá trị của k, cũng như tỷ số. So sánh tỷ lệ này với c từ Mệnh đề 4, trong trường hợp này là 0.94.

|  |  |  |
| --- | --- | --- |
| k |  |  |
| 0  1  5  10  50 | 0.62  0.255  0.133  0.0591  8.87x10-5 | 0.411  0.85  0.85  0.85 |

Rõ ràng rằng giới hạn là khá bi quan (lưu ý rằng 0.85 là giá trị 1 - m và 0.85 là giá trị riêng lớn thứ hai của M). Người ta có thể chỉ ra rằng phương pháp lũy thừa sẽ hội tụ tiệm cận theo, trong đó λ2 là giá trị riêng lớn thứ hai của M. Hơn nữa, với M có dạng M = (1 - m) A + mS với A cột ngẫu nhiên và mọi Sij = 1/n, có thể chỉ ra rằng |λ2| ≤ 1 - m. Kết quả là, phương pháp lũy thừa sẽ hội tụ nhanh hơn nhiều so với được chỉ ra bởi . Tuy nhiên, giá trị của c trong Mệnh đề 4 cung cấp một ràng buộc rất đơn giản về sự hội tụ của phương pháp lũy thừa ở đây. Dễ dàng thấy rằng vì tất cả các mục của M ít nhất là m / n, nên chúng ta sẽ luôn có c ≤ 1 - 2m / n trong Mệnh đề 4.

Như một vấn đề thực tế, lưu ý rằng ma trận M dương n × n không có các phần tử khác không, vì vậy phép nhân Mv với v ∈ Rn thường sẽ nhận O (n2) phép nhân và phép cộng, một phép tính đáng kể nếu n = 8,000,000,000. Nhưng (4) cho thấy rằng nếu x dương với 1 = 1, thì phép nhân Mx tương đương với (1 − m) Ax + ms. Đây là một phép tính hiệu quả hơn nhiều, vì A có thể được mong đợi chứa hầu hết các số 0 (hầu hết các trang Web chỉ liên kết đến một vài trang khác). Bây giờ chúng tôi đã chứng minh định lý chính của chúng tôi.

Định lý 5. Ma trận M được xác định bởi (3) cho một Web không có nút treo sẽ luôn là một ma trận ngẫu nhiên cột dương và do đó có một q duy nhất với các thành phần dương sao cho Mq = q và . Vectơ q có thể được tính là giới hạn của các lần lặp xk = (1 - m) Axk − 1 + ms, trong đó x0 là vectơ ban đầu bất kỳ có thành phần dương và 1 = 1.

Vecto riêng x được xác định bởi (4) cũng có cách diễn giải theo xác suất. Hãy xem xét một người lướt Web trên một trang Web gồm n trang không có nút treo. Người dùng bắt đầu tại một số trang Web (không quan trọng ở đâu) và di chuyển ngẫu nhiên từ trang Web này sang trang Web khác theo quy trình sau: Nếu người dùng hiện đang ở một trang có r liên kết đi, anh ta chọn ngẫu nhiên bất kỳ một trong những liên kết này với xác suất đồng nhất hoặc anh ta nhảy đến bất kỳ trang nào được chọn ngẫu nhiên trên Web, mỗi trang có xác suất  (lưu ý rằng r + n = 1, vì vậy điều này tính cho tất cả những gì anh ta có thể làm). Người dùng lặp lại quy trình nhảy trang này trong phần nhỏ. Thành phần xj của vectơ chuẩn hóa x trong (4) là phần thời gian mà người dùng sử dụng, về lâu dài, trên trang j của Web. Các trang quan trọng hơn có xu hướng được liên kết với nhiều trang khác và do đó người dùng truy cập những trang đó thường xuyên nhất.

Bài tập 14. Đối với trang Web trong Bài tập 11, tính các giá trị của và cho k = 1, 5, 10, 50 bằng cách sử dụng dự đoán ban đầu x0 không quá gần với vecto riêng q thực tế (để bạn có thể xem sự hội tụ). Xác định và giá trị tuyệt đối của giá trị riêng lớn thứ hai của M.

Bài tập 15. Để biết tại sao giá trị riêng lớn thứ hai lại đóng vai trò trong giới hạn , hãy xem xét một n × n của ma trận ngẫu nhiên cột dương M có thể phân chia theo đường chéo. Gọi x0 là bất kỳ vectơ nào có thành phần không âm có tổng bằng 1. Vì M có thể theo đường chéo, chúng ta có thể tạo cơ sở của các vector riêng {q, v1, ..., vn − 1}, trong đó q là vectơ trạng thái ổn định, và sau đó viết . Xác định Mkx0, và sau đó chỉ ra rằng a = 1 và tổng các thành phần của mỗi vj phải bằng 0. Tiếp theo, áp dụng Định đề 4 để chứng minh rằng, ngoại trừ giá trị riêng không được tăng λ = 1, các giá trị riêng khác đều nhỏ hơn 1 trong giá trị tuyệt đối. Sử dụng điều này để đánh giá limk → ∞.

Bài tập 16. Xét ma trận liên kết

Chứng tỏ rằng M = (1 - m) A + mS (mọi Sij = 1/3) không có đường chéo với 0 ≤ m <1.

Bài tập 17. Giá trị của m phải chọn như thế nào? Lựa chọn này ảnh hưởng như thế nào đến thứ hạng và thời gian tính toán?

NGƯỜI GIỚI THIỆU

[1] A. Berman and R. Plemmons, Nonnegative Matrices in the Mathematical Sciences, Academic Press, New York, 1979.

[2] M. W. Berry and M. Browne, Understanding Search Engines: Mathematical Modeling and Text Retrieval, 2nd ed., SIAM, Philadelphia, 2005.

[3] M. Bianchini, M. Gori, and F. Scarselli, Inside PageRank, ACM Trans. Internet Tech., 5 (2005), pp. 92–128.

[4] S. Brin and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, http://www-db.stanford.edu/∼backrub/google.html (August 1, 2005).

[5] A. Farahat, T. Lofaro, J. C. Miller, G. Rae, and L. A. Ward, Authority rankings from HITS, PageRank, and SALSA: Existence, uniqueness, and effect of initialization, SIAM J. Sci. Comput., 27 (2006), pp 1181–1201.

[6] A. Hill, Google Inside Out, Maximum PC, 9 (4) (2004), pp. 44–48.

[7] S. Kamvar, T. Haveliwala, and G. Golub, Adaptive methods for the computation of PageRank, Linear Algebra Appl., 386 (2004), pp. 51–65.

[8] A. N. Langville and C. D. Meyer, A survey of eigenvector methods of Web information retrieval, SIAM Rev., 47 (2005), pp. 135–161.

[9] A. N. Langville and C. D. Meyer, Deeper inside PageRank, Internet Math., 1 (2005), pp. 335–380.

[10] C. D. Meyer, Matrix Analysis and Applied Linear Algebra, SIAM, Philadelphia, 2000.

[11] C. Moler, The World’s Largest Matrix Computation <http://www.mathworks.com/company/> newsletters/new notes/clevescorner/oct02 cleve.html (August 1, 2005).

[12] J. Mostafa, Seeking better web searches, Sci. Amer., 292 (2005), pp. 66–73.

[13] S. Robinson, The ongoing search for efficient web search algorithms, SIAM News, 37 (9) (November 2004), p. 4.

[14] I. Rogers, The Google Pagerank Algorithm and How It Works, <http://www.iprcom.com/> papers/pagerank/(August 1, 2005).

[15] W. J. Stewart, An Introduction to the Numerical Solution of Markov Chains, Princeton University Press, Princeton, NJ, 1994.

**Chương II: Ứng dụng thuật toán PageRank trong xếp hạng đội tuyển bóng đá quốc gia**

Tóm tắt — Giải vô địch bóng đá thế giới là sự kiện thể thao được yêu thích nhất trên thế giới là nguồn cung cấp cả giải trí và lượng dữ liệu khổng lồ về các trận đấu đã chơi. Trong bài báo này, chúng tôi phân tích dữ liệu có sẵn về các giải vô địch bóng đá thế giới kể từ năm 1930 cho đến ngày nay. Mục tiêu của chúng tôi là xếp hạng các đội tuyển quốc gia dựa trên tất cả các trận đấu trong các giải vô địch. Vì mục đích này, chúng tôi áp dụng Xếp hạng trang với thuật toán khởi động lại cho biểu đồ được xây dựng từ các trò chơi đã chơi trong các giải đấu. Một số thống kê như số trận thắng và số bàn thắng ghi được được kết hợp theo các số liệu khác nhau để gán trọng số cho các liên kết trong biểu đồ. Cuối cùng, kết quả của chúng tôi chỉ ra rằng phương pháp Đi bộ ngẫu nhiên với việc sử dụng các chỉ số phù hợp thực sự có thể tạo ra thứ hạng phù hợp so với bảng xếp hạng chính thức mọi thời đại của FIFA.

1. Giới thiệu

Bóng đá, môn thể thao được yêu thích nhất trên thế giới, thu hút sự chú ý của mọi người trong mọi lĩnh vực, từ phương tiện đơn giản là giải trí đến phương tiện phức tạp hơn là thống kê, nghiên cứu và phân tích dữ liệu. kể từ khi FIFA World Cup diễn ra lần đầu tiên vào năm 1930 cho đến nay, có khoảng 20 giải đấu được diễn ra, mỗi giải đấu gồm 64 trận đấu, không tính vòng loại [1], [2]. Do đó, có một lượng lớn dữ liệu đáng kể mà người ta có thể kiểm tra, phân tích và đưa ra kết luận.

Có ý nghĩ đó, các nhà nghiên cứu đang giải quyết vấn đề liên quan đến chiến lược chơi, sắp xếp các đội hoặc phân tích hiệu suất từ các khía cạnh khác nhau bao gồm các yếu tố kinh tế, nhân khẩu học, văn hóa và khí hậu [3]. Ví dụ chiến lược của một đội có thể được quan sát từ quan điểm lý thuyết đồ thị bằng cách xây dựng một mạng lưới các đường truyền giữa các cầu thủ. Trong bối cảnh này, các biện pháp tập trung khác nhau có thể được sử dụng để xác định tầm quan trọng của các cầu thủ cụ thể [4]-[6]. Đối tượng quan tâm khác có thể là mô hình hóa các trận đấu bóng đá về điểm số trong trận đấu. Ví dụ, trong [7] các tác giả thảo luận về một mô hình thống kê cho thời gian ghi bàn của một trận đấu.

Ở đây chúng tôi đề cập đến vấn đề xếp hạng đội tuyển bóng đá quốc gia. Nhiệm vụ chính của chúng tôi là sử dụng bản thống kê có sẵn, để đưa ra một phương pháp xếp hạng thay thế cho các đội bóng dựa trên thành tích của họ tại World Cups. Có các phương pháp đánh giá khác nhau hiện đang được sử dụng và chúng tạo ra các kết quả phù hợp. FIFA have their own 4-year points dựa trên hệ thống xếp hạng của FIFA/Coca-Cola [8] và xếp hạng mọi thời đại của World Cup [9] bao gồm tất cả các chức vô địch kể từ khi xuất hiện. Ngoài ra còn có xếp hạng Elo bóng đá thế giới dựa trên hệ thống xếp hạng FIDE sử dụng để xếp hạng các kì thủ cờ vua [10].

Một phương pháp xếp hạng tốt không chỉ nên tính đến một đội thắng bao nhiêu lần, mà còn xem xét họ đánh bại một đối thủ mạnh như thế nào. Chiến thắng trước đối thủ mạnh hơn được ưu tiên hơn và do đó có ý nghĩa hơn so với chiến thắng trước đối thủ yếu hơn. Một phương pháp kết hợp logic như vậy là phương pháp PageRank (Random walk), có thể áp dụng cho nhiều loại mạng dựa trên các vấn đề xếp hạng theo một cách nào đó. Ngoài vấn đề đã biết về xếp hạng các trang web [11], nó còn được dùng trong phân tích mạng xã hội, trong các nhiệm vụ như dự đoán liên kết, truyền bá thông tin và phát hiện cộng đồng [12]-[14]. Ngoài ra nó cũng được sử dụng trong NLP với mục đích tóm tắt văn bản và phân biệt từ ngữ [15], [16]. Đối với những lần thử trước về việc sử dụng cơ chế PageRank trong các sự kiện thể thao, chúng tôi giới thiệu người đọc đến [17]-[19].

Phần còn lại của bài báo cáo được tổ chức như sau. Trong mục II, chúng tôi trình bày vấn đề xếp hạng và PageRank dựa trên phương pháp để giải quyết vấn đề đó. Chúng tôi cũng đưa ra mô tả và số liệu thống kê của dữ liệu có sẵn cho chúng tôi. Kết quả thu được được trình bày trong mục III bao gồm một cuộc thảo luận và sự so sánh về bảng xếp hạng chính thức và sau đó chúng tôi kết thúc bài báo cáo trong mục IV.

1. Dữ liệu và phương pháp.
2. Dữ liệu.

Dữ liệu chúng tôi sử dụng được lấy từ 11v11, trang web thống kê bóng đá chứa tất cả các số liệu thời gian về trận đấu của World Cup, bao gồm cả các trận đấu vòng loại [20]. Đối với mỗi đội tuyển quốc gia có thông tin về đội tuyển nào họ đã đối đầu, số trận thắng, hòa và thua, cũng như số bàn ghi điểm và để lọt lưới trong tất cả các trận đấu. Trong suốt bài báo cáo này chúng tôi sử dụng thuật ngữ đối đầu trong bối cảnh của một trận đấu được chơi giữa hai đội. Và một cặp là hai đội đã đấu với nhau. Tệp dữ liệu, bao gồm 210 quốc gia và bản thống kê về 2335 cặp đối đầu đã thi với nhau, hoặc tổng cộng 7141 trận, trong đó có 20298 bàn thắng được ghi. Số trận trung bình của mỗi cặp đối đầu là 30582, và trung bình số bàn thắng của mỗi cặp đối đầu là 43465. Mexico với Mỹ là cặp có số trận đấu với nhau lớn nhất. Khoảng 28 trận được diễn ra trong đó có khoảng 100 bàn thắng, 15 trong số đó thuộc về Mỹ, 6 trận hòa và 7 trận còn lại mang đến chiến thắng cho Mexico. Quốc gia có nhiều trận đấu nhất là Brazil với khoảng 200 trận và cũng là quốc gia có nhiều trận thắng nhất và ghi nhiều bàn thắng nhất.

Bảng I:

|  |  |  |
| --- | --- | --- |
| # | Trọng số thống kê | Nghịch đảo |
| 1 |  | 0,032 |
| 2 |  | 0,038 |
| 3 |  | 0,040 |
| 4 |  | 0,040 |
| 5 |  | 0,041 |
| 6 |  | 0,043 |
| 7 |  | 0,044 |
| 8 |  | 0,044 |
| 9 |  | 0.046 |
| 10 |  | 0,050 |

1. Phương pháp

Phương pháp xếp hạng được khám phá trong suốt bài báo cáo này là PageRank với thuật toán khởi động lại được áp dụng cho việc xây dựng đồ thị xung quanh dữ liệu được cung cấp [11]. Mỗi đội tuyển quốc gia là một điểm nút duy nhất trong đồ thị và hai điểm nút được liên kết nếu hai đội tuyển (cặp đối đầu) đã từng thi đấu trong một giải đấu World Cup. Trọng số liên kết được xác định bởi một trọng số thống kê gồm một hoặc nhiều chỉ số như số trận đấu giữa một cặp đối đầu, số trận thắng, thua và hòa, hoặc số bàn thắng được ghi và để thủng lưới. Các trọng số thống kê khác nhau mà chúng tôi đã thử nghiệm được đưa ra trong bảng 1.

Trong các hàm số chúng tôi sử dụng ký hiệu sau:

trọng số liên kết từ điểm nút i đến điểm nút j;

số trận đấu đã chơi giữa hai đội;

số trận đấu đã thua của đội i trong tất cả các trận đấu i và j đã chơi;

số trận đấu đã thắng của đội i trong tất cả các trận đấu i và j đã chơi;

số bàn thua của đội i trong tất cả các trận đấu i và j đã chơi;

số bàn thắng của đội i trong tất cả các trận đấu i và j đã chơi;

số trận đấu hòa giữa hai đội;

số trận đấu tối đa đã chơi giữa bất kỳ cặp đối đầu nào;

Một yếu tố khác ảnh hưởng đến PageRank là yếu tố giảm dần. Yếu tố giảm dần tương ứng với xác suất mà một người đi bộ ngẫu nhiên dừng đi bộ và nhảy đến một điểm nút ngẫu nhiên [21]. Yếu tố giảm dần ngoài việc cần thiết để đảm bảo rằng bước đi ngẫu nhiên sẽ hội tụ về một phân bố đứng im, nó cũng trực quan. Trực quan đằng sau việc sử dụng yếu tố giảm dần trong mạng lưới đối đầu của chúng tôi là như sau: mặc dù đồ thị dày đặc nhưng không phải mọi đội đã thi đấu với nhau. Vì vậy khi sử dụng các số liệu trọng số như tỉ lệ thua (hàm số 1 trong bảng 1) yếu tố giảm dần sẽ có nghĩa là thêm một số cơ hội thắng cho tất cả các đội chưa từng đối đầu. Nó cũng tăng thêm một số cơ hội thắng cho một đội chưa bao giờ thắng một trận nào trong một trận đấu.

PageRank được tính bằng cách sử dụng phương pháp lũy thừa [22]. Phương pháp này là một thuật toán lặp lại để tìm ra vecto riêng quan trọng, tương ứng với phân phối bất biến của thời gian mà một người đi bộ ngẫu nhiên dành cho một điểm nút nhất định-PageRank. Bằng cách chuẩn hóa ma trận kề A chúng ta nhận được ma trận xác suất chuyển tiếp Q với các phần tử như đã cho.

(1)

(2)

Lưu ý rằng Q được đảm bảo là tối giản và không tuần hoàn do hệ quả của hệ số giảm chấn d.

Bảng II: Số trận đã chơi và kết quả từng cặp đấu

|  |  |  |
| --- | --- | --- |
| Cặp | Số trận | Kết quả |
| A-B | 3 | A thắng 2, B thắng 1 |
| A-C | 3 | A thắng 2, C thắng 1 |
| A-D | 3 | A thắng 3, D thắng 0 |
| B-C | 3 | C thắng 3, B thắng 0 |
| B-D | 3 | D thắng 3. B thắng 0 |
| C-D | 3 | C thắng 1, D thắng 2 |

Bảng III: PageRank của mỗi đội theo thứ tự giảm dần

|  |  |  |  |
| --- | --- | --- | --- |
| Đội | Số trận | Thắng | PageRank |
| A | 9 | 7 | 0,333 |
| C | 9 | 5 | 0,281 |
| D | 9 | 5 | 9,211 |
| B | 9 | 1 | 0,175 |

1. Ví dụ.

Để minh họa, hãy xem xét một ví dụ về đồ chơi điều đó làm sáng tỏ mục tiêu của chúng ta. Giả sử có 4 đội và thống kê đưa ra cho từng cặp được biểu diễn trong bảng 2. Đồ thị (Hình 1) được xây dựng bằng cách sử dụng tỉ số tổn thất làm thước đo (Hàm 2 trong bảng 1). Do đó trọng số của một liên kết đã cho từ i đến j là phần trong số các trận i đã thua j. Ví dụ có một liên kết từ A đến C với trọng số là và cũng có một liên kết từ C đến A với trọng số là . Điều đó có nghĩa là trong số 3 trận đấu A và C đã đấu với nhau, A thắng 2 trận, C thắng 1 trận và không có trận hòa. Bước tiếp theo là tính toán PageRank. Do đó chúng ta cần ma trận chuyển tiếp được tính theo với giá trị yếu tố giảm dần thường dùng là 0,15.

B

1/3

2/3

3/3

3/3

0/3

0/3

A

3/3

2/3

1/3

0/3

1/3

C

D

2/3

Hình 1:

Kết quả cuối cùng được thể hiện ở bảng 3. A được chỉ ra là có xếp hạng cao nhất và B là đội xếp hạng thấp nhất như mong đợi. Mặt khác, đội C và đội D đều đã thắng 5 trận như thể hiện ở bảng 3. Tuy nhiên, PageRank không chỉ tính đến sức mạnh của đối thủ bị đánh bại mà còn tính đến số tiền thắng. Do đó, đội C được xếp hạng cao hơn vì họ đã thắng một trận đấu với đội A, được coi là một đối thủ mạnh, trái lại đội D chỉ thắng trước các đối thủ yếu hơn.

1. Kết quả và thảo luận.

Để tìm ra thứ hạng chính xác nhất một số trọng số thống kê khác nhau đã được thử và hầu như chúng đều cho kết quả tương tự. Kết quả được đánh giá bằng cách so sánh PageRank với bảng xếp hạng chính thức của World Cup. Chúng tôi đã sử dụng số lần đảo ngược chuẩn hóa làm chỉ số đánh giá [23], lấy bảng xếp hạng chính thức mọi thời đại của FIFA làm thứ tự giới thiệu. Các trọng số thống kê đã thử nghiệm và điểm số của chúng được liệt kê ở bảng I. Điểm thấp hơn có nghĩa là kết quả được tạo ra bằng cách sử dụng số liệu tương ứng giống với xếp hạng chính thức hơn. Chúng tôi chỉ sử dụng 30 đội xếp hạng cao nhất hàng đầu trong cuộc so sánh vì chúng tôi vì chúng tôi muốn giành cho họ mức độ ưu tiên cao hơn và nhận được thứ tự của họ ngay lập tức với cái giá phải trả là đặt sai vị trí một số đội được xếp hạng thấp hơn. Sai số của trọng số thống kê cũng phụ thuộc vào hệ số giảm dần. Mức tối thiểu đạt được khi hệ số giảm dần rất nhỏ, khoảng 0,05. Đó là giá trị chúng tôi đã sử dụng trong đánh giá các chỉ số được hiển thị trong bảng I. Hình 3 cho thấy lỗi (trong số lần đảo ngược chuẩn hóa) cho 5 chỉ số hàng đầu dưới dạng hàm của hệ số giảm dần. Như mong đợi sai số tăng lên cùng với sự phát triển của hệ số giảm dần. Bảng IV cho thấy 20 đội hàng đầu (ngắn gọn), theo trọng số thống kê tốt nhất của chúng tôi. Cột thứ 4 chứa các vị trí của mỗi đội trên bảng xếp hạng chính thức. Vị trí được đánh dấu màu xanh lá cây nếu đội giữ cùng một vị trí trong bảng xếp hạng của chúng tôi và bảng xếp hạng chính thức. Vị trí được đánh dấu màu đỏ nếu có sự dịch chuyển lớn (Đan Mạch và Croatia). Nếu một đội không được tìm thấy trong bảng xếp hạng chính thức (trong trường hợp của chúng tôi là Tiệp Khắc và Nam Tư) thì vị trí của họ sẽ được đánh dấu bằng NA. Hình 2 cho thấy biểu đò phù hợp. Mỗi đội là một điểm nút trong biểu đồ được biểu thị bởi quốc kỳ của họ và kích thước của mỗi điểm nút là tỉ lệ thuận với PageRanl của đội đó. Trong hình, một phần của các liên kết được bỏ qua vì lợi ích của sự rõ ràng, do đó đồ thị thực dày đặc hơn so với biểu đồ xuất hiện.

Sự cố có thể xảy ra khi sử dụng PageRank làm phương pháp xếp hạng có thể là như sau: một điểm nút có thể đạt được điểm PageRank cao nếu nó có một quốc gia láng giềng xếp hạng cao mà từ đó nó có thể nhận được lượng phiếu bầu đáng kể hoặc nếu nó có nhiều quốc gia láng giềng xếp hạng thấp. Trong ví dụ của chúng tôi, nếu một đội tuyển quốc gia được xếp hạng cao thì họ phải đánh bại nhiều đội xếp hạng thấp hoặc giành được kết quả đáng kể trước đối thủ có xếp hạng cao. Tính chất này của bước đi ngẫu nhiên ảnh hưởng đến kết quả của chúng tôi đặc biệt là vì chúng tôi đối xử bình đẳng với tất cả các trận đấu mà không tính đến việc đó là trận đấu vòng loại hay trận đấu cuối cùng. Do đó, có thể có các đội chỉ nhận được xếp hạng cao vì họ đã thi đấu và giành chiến thắng trước nhiều đối thủ xếp hạng thấp trong các trận đấu kém quan trọng hơn.

Bảng IV:

Top 20 đội tuyển quốc gia được đánh giá cao nhất

|  |  |  |  |
| --- | --- | --- | --- |
| # | Quốc gia | PageRank | Chính thức |
| 1 | Brazil | 0,040375 | 1 |
| 2 | Ý | 0,037992 | 3 |
| 3 | Đức | 0,033801 | 2 |
| 4 | Hà Lan | 0,031052 | 8 |
| 5 | Argentina | 0,029159 | 4 |
| 6 | Anh | 0,029100 | 6 |
| 7 | Tây Ban Nha | 0,027904 | 5 |
| 8 | Pháp | 0,025670 | 7 |
| 9 | Tiệp Khắc | 0,025155 | NA |
| 10 | Thụy Điển | 0,022882 | 10 |
| 11 | Mexico | 0,022034 | 13 |
| 12 | Hungary | 0,022014 | 16 |
| 13 | Uruguay | 0,020660 | 9 |
| 14 | Bỉ | 0,020255 | 14 |
| 15 | Bồ Đào Nha | 0,020211 | 17 |
| 16 | Ba Lan | 0,019528 | 15 |
| 17 | Đan Mạch | 0,019206 | 25 |
| 18 | Croatia | 0,018993 | 27 |
| 19 | Thụy Sĩ | 0,016650 | 21 |
| 20 | Nam Tư | 0,016466 | NA |

Hình 2: Biểu đồ của các trận đấu Chart, diagram

Description automatically generated

Hình 3:

Chart

Description automatically generated

1. Phần kết luận và tài liệu tham khảo.

Trong suốt bài báo cáo này chúng tôi đã khám phá phương pháp PageRank để xếp hạng các đội tuyển bóng đá quốc gia. Kết quả của chúng tôi cho thấy ngay cả với các trọng số thống kê đơn giản như tỉ số bàn thắng được ghi hoặc các trận thắng, thuật toán PageRank cho kết quả đầy hứa hẹn. Bảng xếp hạng mà phương pháp này tạo ra tương tự như bảng xếp hạng chính thức mọi thời đại của FIFA. Tuy nhiên, rất khó để đánh giá PageRank với việc sử dụng chức năng trọng số phức tạp hơn và nhiều tính năng hơn trong tập dữ liệu có thể dẫn đến sơ đồ xếp hạng chính thức tốt hơn hay không. Dù sao, với giả định rằng hệ thống xếp hạng FIFA là phù hợp và chính xác, RandomWalk mặc dù có tập dữ liệu đơn giản và các chỉ số trọng số có thể sao chép rất nhiều kết quả của nó.

THAM KHẢO

[1] F. I. de Football Association et al., “Fifa competitions and olympic football tournaments 1908-2017,” 2014. [Online]. Available: http://www.fifa.com/worldcup/organisation/documents/index.html

[2] ——, “Fifa world cup comparative statistics 1982-2014,” 2014. [Online]. Available: <http://www.fifa.com/worldcup/organisation/documents/index>. html

[3] R. Hoffmann, L. C. Ging, and B. Ramasamy, “The socio-economic determinants of international soccer performance,” Journal of Applied Economics, vol. 5, no. 2, pp. 253–272, 2002. [4] J. L. Pena and H. Touchette, “A network theory analysis of football ˜ strategies,” arXiv preprint arXiv:1206.6904, 2012.

[5] J. Duch, J. S. Waitzman, and L. A. N. Amaral, “Quantifying the performance of individual players in a team activity,” PloS one, vol. 5, no. 6, p. e10937, 2010.

[6] M. Hughes and I. Franks, “Analysis of passing sequences, shots and goals in soccer,” Journal of Sports Sciences, vol. 23, no. 5, pp. 509–514, 2005.

[7] M. Dixon and M. Robinson, “A birth process model for association football matches,” Journal of the Royal Statistical Society: Series D (The Statistician), vol. 47, no. 3, pp. 523–538, 1998.

[8] “Fifa/coca-cola world ranking,” <http://www.fifa.com/fifa-world-ranking/> ranking-table/men/, accessed: 2015-01-25.

[9] F. I. de Football Association et al., “Fifa world cup all-time ranking,” 2014. [Online]. Available: <http://www.fifa.com/worldcup/organisation/> documents/index.html

[10] “World football elo ratings,” http://www.eloratings.net, accessed: 2015-02-11.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web.” 1999.

[12] L. Backstrom and J. Leskovec, “Supervised random walks: predicting and recommending links in social networks,” in Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011, pp. 635–644.

[13] M. Kimura and K. Saito, “Tractable models for information diffusion in social networks,” in Knowledge Discovery in Databases: PKDD 2006. Springer, 2006, pp. 259–271.

[14] A. Stanoev, D. Smilkov, and L. Kocarev, “Identifying communities by influence dynamics in social networks,” Physical Review E, vol. 84, no. 4, p. 046102, 2011.

[15] G. Erkan and D. R. Radev, “Lexrank: Graph-based lexical centrality as salience in text summarization,” J. Artif. Intell. Res.(JAIR), vol. 22, no. 1, pp. 457–479, 2004.

[16] E. Agirre and A. Soroa, “Personalizing pagerank for word sense disambiguation,” in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009, pp. 33–41.

[17] J. P. Keener, “The perron-frobenius theorem and the ranking of football teams,” SIAM review, vol. 35, no. 1, pp. 80–93, 1993.

[18] S. Mukherjee, “Identifying the greatest team and captain—a complex network approach to cricket matches,” Physica A: Statistical Mechanics and its Applications, vol. 391, no. 23, pp. 6066–6076, 2012.

[19] F. Radicchi, “Who is the best player ever? a complex network analysis of the history of professional tennis,” PloS one, vol. 6, no. 2, p. e17249, 2011.

[20] “11v11 - home of football statistics and history,” <http://www.11v11.com>, accessed: 2015-01-25.

[21] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” Computer networks and ISDN systems, vol. 30, no. 1, pp. 107–117, 1998.

[22] A. N. Langville and C. D. Meyer, “Deeper inside pagerank,” Internet Mathematics, vol. 1, no. 3, pp. 335–380, 2004.

[23] D. E. Knuth, The art of computer programming: sorting and searching. Pearson Education, 1998, vol. 3

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**THÔNG TIN KẾT QUẢ NGHIÊN CỨU ĐỀ TÀI**

1. **Thông tin chung:**

- Tên đề tài:TÌM HIỂU THUẬT TOÁN PAGERANK VÀ ỨNG DỤNG

- Sinh viên thực hiện:

* Trần Thị Lan Hương
* Trần Thị Thu Trang
* Mai Ngọc Kiều
* Lê Quang Vũ

- Lớp: Toán ứng dụng 1

- Khoa: Khoa học cơ bản

- Năm thứ : 1

- Số năm đào tạo: 4 năm

- Người hướng dẫn: PGS. TS Trần Văn Long

**2. Mục tiêu đề tài**

Để người đọc có thể hiểu rõ hơn về thuật toán Pagerank và những ứng dụng trong thực tiễn của nó

**3. Tính mới sáng tạo**

**4. Công bố khoa học của sinh viên từ kết quả nghiên cứu đề tài**

Ngày tháng năm 2022

**Sinh viên chịu trách nhiệm chính thực**

**hiện đề tài**

**Nhận xét của người hướng dẫn về những đóng góp khoa học của sinh viên thực hiện đề tài**

Ngày tháng năm 2022

**Người hướng dẫn**