Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00040

MÔ HÌNH PHÂN TÁN CHO THUẬT GIẢI XÁC ĐỊNH ĐỈNH CÓ SỨC ẢNH HƯỞNG LỚN NHẤT TRONG ĐỒ THỊ MẠNG XÃ HỘI

Nguyễn Hồ Duy Tri, Ngô Thanh Hùng

Phòng Thí nghiệm Hệ thống Thông tin, Trường Đại học Công nghệ Thông tin – ĐHQG TPHCM

tringuyen@uit.edu.vn, hungnt@uit.edu.vn

TÓM TẮT—*Vấn đề xác định key player trong các mạng xã hội đang thu hút sự quan tâm của nhiều nhà nghiên cứu trên thế giới. Trong một nghiên cứu trước đây, chúng tôi đã đề xuất một phương pháp mới để xác định key player dựa vào tổng sức ảnh hưởng của mỗi đỉnh tới tất cả các đỉnh còn lại. Tuy nhiên việc cài đặt thuật toán trên các nền tảng tuần tự hoặc đa luồng không thể áp dụng được với các mạng xã hội có từ hàng trăm, hàng ngàn node trở lên, trong khi các mạng xã hội thông thường có số lượng node là lớn hơn rất nhiều. Chính vì vậy chúng tôi xây dựng và trình bày trong bài báo này một thuật toán xác định key player trên nền tảng phân tán. Thuật toán đã được cài đặt trên Spark. Bài báo cũng trình bày hiệu quả của thuật toán phân tán so với thuật toán tuần tự qua một số thử nghiệm.*

Từ khóa— *Scalable algorithm, Spark, key player, the most influence node, social network.*

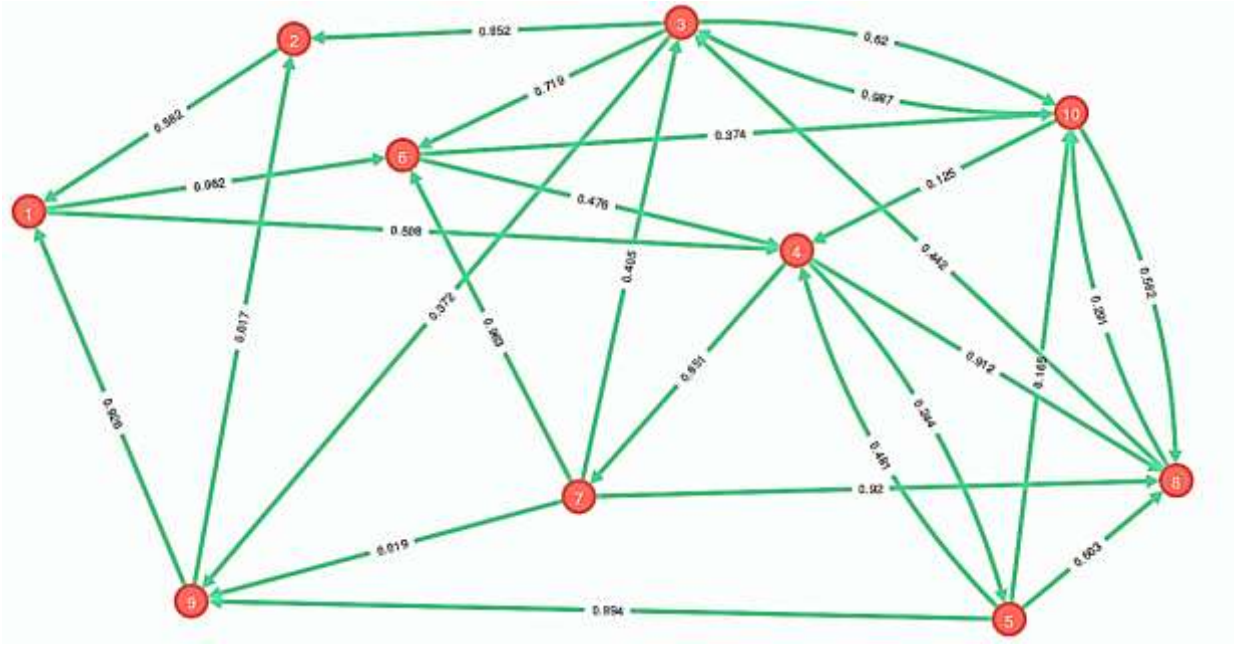
I. GIỚI THIỆU

Tại Việt Nam, đối tượng gia nhập, sử dụng mạng xã hội ngày càng nhiều, với nhiều những hoạt động và mục đích khác nhau. So với các phương tiện thông tin, liên lạc truyền thống như hệ thống phát thanh, truyền hình, báo giấy thì mạng xã hội có nhiều ưu điểm vượt trội, tuy nhiên nội dung lại khó kiểm soát và có thể gây ảnh hưởng lớn đến nhiều người. Nước ta hiện đang là nước có lượng truy cập mạng xã hội Facebook nhiều thứ 15 trên thế giới của Facebook, có khoảng 30 triệu người dùng thường xuyên trên Facebook, và con số này còn tăng trưởng mỗi ngày. Phương pháp xác định sức ảnh hưởng của một cá nhân trong mạng xã hội đã được tác giả Ngô Thanh Hùng nghiên cứu trong một công trình khác trước đây, tuy nhiên, đối với dữ liệu mạng xã hội hiện nay, một đòi hỏi cấp thiết là phải được xử lý nhanh với một khối lượng vô cùng lớn để có thể mang lại hiệu quả và giá trị cao. Trong bài báo này, chúng tôi cố gắng thử nghiệm phương pháp song song hóa giải thuật xác định sức ảnh hưởng của một cá nhân trong mạng xã hội đối với cộng đồng, qua đó tìm ra người dẫn dắt dư luận.

II. THUẬT GIẢI XÁC ĐỊNH ĐỈNH CÓ SỨC ẢNH HƯỞNG LỚN NHẤT TRONG ĐỒ THỊ MẠNG XÃ HỘI

A. Mô hình hóa đồ thị mạng xã hội

Dữ liệu mạng xã hội sử dụng trong bài báo được mô hình hóa dưới dạng đồ thị có hướng, với đỉnh của đồ thị biểu diễn những người dùng trong mạng, cạnh nối giữa các đỉnh thể hiện sự lan truyền thông tin giữa các người dùng. Hướng của cạnh cho biết chiều hướng lan truyền của thông tin và trọng số của cạnh thể hiện sự ảnh hưởng của một người đến một người khác khi xảy ra sự lan truyền thông tin trên mạng.



Hình 1. Mô hình hóa đồ thị mạng xã hội

B. Người dẫn dắt dư luận trong mạng xã hội

1. Sức ảnh hưởng trong mạng xã hội

Sức ảnh hưởng hay còn gọi là sức thuyết phục của một đỉnh A đối với một đỉnh B được đánh giá bởi xác suất lan truyền thành công ý tưởng từ một đỉnh A đến đỉnh B. Nói cách khác nếu trong mạng chỉ có A là người đầu tiên chấp nhận ý tưởng thì B sẽ chấp nhận ý tưởng với xác suất là bao nhiêu (không quan tâm đến thời điểm B chấp nhận).

Trong đó, sức ảnh hưởng trực tiếp từ đỉnh A đến đỉnh B là xác suất lan truyền thành công ý tưởng trực tiếp từ A sang B thông qua cạnh trỏ từ A đến B, được biểu diễn bằng trọng số của cạnh v_{A-B}. Ngoài ra, sức ảnh hưởng riêng phần từ đỉnh A đến đỉnh B thông qua đường đi P lại được đánh giá bởi xác suất lan truyền thành công ý tưởng một cách gián tiếp từ A sang B thông qua đường đi P trong đồ thị bắt đầu từ A và kết thúc tại B, nó sẽ được tính bằng tích các cạnh nằm trên đường đi P.

$$P_{A-B} = v_{A-X} \times v_{X-Y} \times \ldots \times v_{Z-B}$$

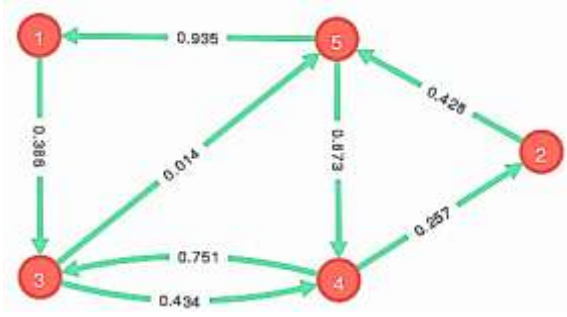
Tổng quát ta có công thức tính sức ảnh hưởng (hay xác suất truyền ý tưởng thành công) giữa hai đỉnh trong đồ thị mạng xã hội như sau:

Gọi hai đỉnh cần tính là A, B. Giữa A và B có n đường đi khác nhau đôi một, nghĩa là từng đôi một có ít nhất một điểm khác và xác suất truyền thành công trên mỗi đường là: $P^1_{A-B}, P^2_{A-B}, \ldots, P^n_{A-B}$

Công thức tính xác suất lan truyền thành công ý tưởng từ A đến B là:

$$P_{A-B} = 1 - (1 - P^1_{A-B}) \times (1 - P^2_{A-B}) \times \ldots \times (1 - P^n_{A-B})$$

Các công thức đã được kiểm nghiệm bằng mô hình xác suất và cho độ chính xác rất cao.



Hình 2. Sức ảnh hưởng trực tiếp giữa các đỉnh trong đồ thị mạng xã hội

Đối với các số liệu đã cho như trong hình 2, xác suất lan truyền ý tưởng thành công từ đỉnh 1 và 2 sẽ được tính như sau:

Xác định các đường đi giữa hai đỉnh 1 và 2, ta được đường đi thứ nhất thông qua các đỉnh 1 – 3 – 5 – 4 – 2. Từ đó, tính được sức ảnh hưởng riêng phần của đường đi này là:

$$P^1_{1-2} = v_{1-3} \times v_{3-5} \times v_{5-4} \times v_{4-2} = 0{,}386 \times 0{,}014 \times 0{,}873 \times 0{,}257 = 0{,}001212446844$$

Đường đi thứ hai từ đỉnh 1 đến đỉnh 2 là: 1 – 3 – 4 – 2. Sức ảnh hưởng riêng phần của đường đi này được tính như sau:

$$P^2_{1-2} = v_{1-3} \times v_{3-4} \times v_{4-2} = 0{,}386 \times 0{,}434 \times 0{,}257 = 0{,}043053668$$

Sức ảnh hưởng của đỉnh 1 đến đỉnh 2 sẽ bằng:

$$P_{1-2} = 1 - (1 - P^1_{1-2}) \times (1 - P^2_{1-2}) = 1 - (1 - 0{,}001212446844) \times (1 - 0{,}043053668)$$

$$= 0{,}0442139145601108$$

Vậy xác suất lan truyền thành công ý tưởng từ đỉnh 1 đến đỉnh 2 là 4,42%.

2. Người dẫn dắt dư luận

Trong mạng xã hội, người dẫn dắt dư luận thường là người khởi phát các ý tưởng, các quan điểm của bản thân hay từ những nguồn khác bên ngoài mạng xã hội và thông qua những mối liên kết, những mối quan hệ của mình, họ có thể lan truyền chúng và gây được ảnh hưởng lớn đến cộng đồng. Từ đó, ta có thể xác định rằng người dẫn dắt dư luận là người có tổng sức ảnh hưởng tới tất cả các đỉnh là lớn nhất trong đồ thị mạng xã hội.

Thuật giải tìm người dẫn dắt dư luận được trình bày như sau:

Foreach (đỉnh I trong đồ thị):

Tổng sức ảnh hưởng của đỉnh I := 0;

Foreach (đỉnh J ≠ I trong đồ thị):

Tính sức ảnh hưởng của đỉnh I đến đỉnh J;

Cộng sức ảnh hưởng của đỉnh I đến đỉnh J vào tổng sức ảnh hưởng của đỉnh I;

End;

End;

Đỉnh có tổng sức ảnh hưởng lớn nhất là người dẫn dắt dư luận trong mạng xã hội;

C. Thuật giải xác định đỉnh có sức ảnh hưởng lớn nhất trong đồ thị mạng xã hội

3. Giới thiệu Apache Spark

Một trong những mô hình xử lý dữ liệu lớn rất phổ biến được sử dụng nhiều trong các tính toán phân tán là MapReduce. Đây là một mô hình luồng dữ liệu, nó thích hợp và được ứng dụng với đa số các công cụ xử lý dữ liệu lớn hiện nay. Nhưng cũng có những ứng dụng không thích hợp khi áp dụng mô hình này, đó là những ứng dụng có dạng mô hình lặp. Trong mô hình này, quá trình xử lý cứ được lặp đi lặp lại. Lúc đó mô hình MapReduce sẽ bộc lộ nhiều hạn chế thể hiện qua việc mỗi lần thực thi sẽ là một lần truy vấn lại dữ liệu từ đĩa cứng, điều này làm cho cả quá trình bị chậm đi rất nhiều. Bên cạnh đó, những dữ liệu được sử dụng nhiều lần trong quá trình thực thi không được tải sẵn lên bộ nhớ đệm để truy vấn mà nó được tải lại đối với mỗi thành phần công việc riêng biệt gây nên độ trễ lớn. Chính vì thế chúng tôi chọn tìm hiểu và cài đặt xử lý dữ liệu lớn trên framework Apache Spark [1]. Được cải tiến và khắc phục những khuyết điểm từ mô hình Hadoop MapReduce, Apache Spark sử dụng một đối tượng bộ nhớ đặc biệt gọi là RDD (Resilient Distributed Dataset), nó là một tập hợp chỉ đọc chứa các loại đối tượng dữ liệu trong các ngôn ngữ lập trình hay các lớp mà người dùng tự định nghĩa, được phân tán lưu trữ ở các nút tính toán (các máy con trong mạng tính toán). Tập hợp này cũng có khả năng mở rộng một cách mềm dẻo, tự cân bằng và khả năng chịu lỗi, phục hồi khi có sự cố xảy ra giống như Hadoop. Khi thao tác RDD sẽ được Spark tải lên bộ nhớ đệm của những nút tính toán để sử dụng nhiều lần qua các quá trình tính toán song song, chính vì thế tốc độ của Spark có thể nhanh hơn Hadoop đến gấp 10 lần.

Các đối tượng RDD trong Apache Spark hỗ trợ hai loại phép tính đặc biệt là: phép biến đổi (transformations) và phép tác động (actions) [2]. Các phép biến đổi trên RDD thường trả về một RDD mới, nó sẽ bao gồm các phép tính cơ bản sau: *map* – hàm tính toán trên từng phần tử trong RDD, tương ứng với mỗi phần tử sẽ trả về một kết quả, *flatMap* - hàm tính toán trên từng phần tử trong RDD, tuy nhiên đối với mỗi phần tử, kết quả trả về có thể là rỗng hoặc có nhiều hơn một kết quả, *filter* – hàm lọc các phần tử của RDD theo điều kiện… Bên cạnh đó, trên RDD còn có những phép tính tác động, các phép tính này thường trả về một giá trị hoặc ghi dữ liệu ra hệ thống lưu trữ bên ngoài. Các phép tính tác động thường dùng bao gồm: *collect* – hàm trả về danh sách tất cả các phần tử trong RDD, *count* – hàm đếm số lượng các phần tử trong RDD, *top* – hàm trả về một số lượng cho trước các phần tử nằm ở đầu của RDD, *reduce* – hàm tính toán song song trên các phần tử của RDD… Do cơ chế "lazy evaluation", một phép tính biến đổi trên RDD sẽ không được thực thi ngay lập tức mà chỉ được Spark ghi nhận vào trong metadata. Sau này, khi chương trình cần thực thi một phép tác động trên RDD, lúc đó Spark sẽ tìm lại trong metadata các phép biến đổi đã được yêu cầu trước đó trên RDD này và lần lượt thực thi chúng, sau đó sẽ thực thi phép tác động. Nguyên nhân khiến Apache Spark sử dụng cơ chế "lazy evaluation" là để giảm thiểu được số quy trình tính toán song song phải thực thi, giúp thời gian xử lý được rút ngắn hơn.

4. Thuật giải tuần tự xác định đỉnh có sức ảnh hưởng lớn nhất trong đồ thị mạng xã hội

Đầu vào của thuật giải: tập đỉnh và tập cạnh của đồ thị mạng xã hội.

Đầu ra của thuật giải: đỉnh có sức ảnh hưởng lớn nhất trong đồ thị.

Ý tưởng: dữ liệu đồ thị mạng xã hội bao gồm tập đỉnh và tập cạnh được đọc vào hệ thống từ các tập tin lưu trữ. Do đặc tính dữ liệu dễ gây bùng nổ số lượng khi tìm kiếm đường đi giữa các đỉnh, mỗi khi tính toán sức ảnh hưởng của một đỉnh phải vét cạn tất cả các đường đi từ đỉnh đó đến tất cả các đỉnh còn lại, bằng cách xét lần lượt từng đỉnh, từng cạnh một của đồ thị. Những điều đó làm cho việc tính toán tốn rất nhiều chi phí cũng như thời gian thực thi. Trong một dữ liệu thực nghiệm mà chúng tôi đã tiến hành khảo sát, đồ thị mạng xã hội bao gồm 600 đỉnh, mỗi đỉnh có tối đa 30 cạnh trỏ đến các đỉnh khác, sau khi tính toán số lượng đường đi tổng cộng trong đồ thị phải xử lý lên tới hơn 3,5 triệu, một con số rất lớn. Vì vậy, để tiết kiệm chi phí, chúng tôi đã cài đặt thuật giải theo phương pháp tổ hợp tất cả các cạnh để tìm ra toàn bộ các đường đi có thể của đồ thị. Lần lượt các đường đi này sẽ được sử dụng để tính toán sức ảnh hưởng của từng đỉnh và qua đó tìm ra đỉnh có sức ảnh hưởng lớn nhất trong đồ thị.

Thuật giải tuần tự:

Dựa theo những phân tích trên, thuật giải tuần tự được chúng tôi đưa ra để giải quyết bài toán xác định đỉnh có sức ảnh hưởng lớn nhất trong đồ thị mạng xã hội như sau:

Đọc dữ liệu đầu vào là tập đỉnh và tập cạnh;

Khởi tạo ma trận sức ảnh hưởng $T_{n \times n}$;

Cập nhật sức ảnh hưởng trực tiếp từ các cạnh vào $T_{n \times n}$;

do

Tạo đường đi có m cạnh từ đường đi có m – 1 cạnh và các cạnh đơn, tính sức ảnh hưởng riêng phần của từng đường đi mới;

Cập nhật sức ảnh hưởng vào ma trận $T_{n \times n}$;

while (m < n – 1 or không tạo được đường đi mới);

Tính sức ảnh hưởng của từng đỉnh dựa vào ma trận $T_{n \times n}$;

So sánh và kết luận đỉnh có sức ảnh hưởng lớn nhất trong đồ thị mạng xã hội;

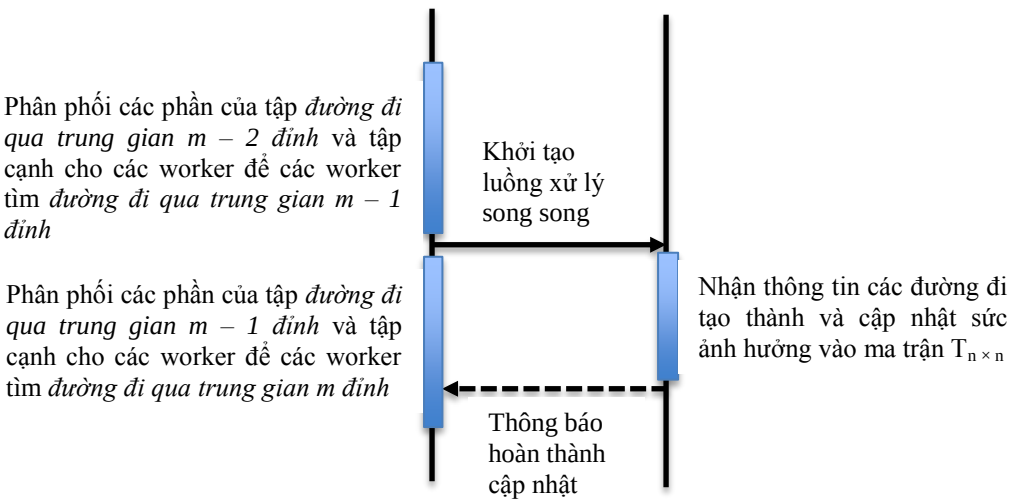
5. Song song hóa thuật giải xác định đỉnh có sức ảnh hưởng lớn nhất trong đồ thị mạng xã hội bằng Apache Spark

Trên nền tảng Apache Spark với các máy tính con đóng vai trò là các đơn vị xử lý trong hệ thống phân tán, gọi là các worker và một máy tính với vai trò quản lý tài nguyên, thu thập các dữ liệu kết quả từ các đơn vị và tính toán kết quả chung gọi là master, chúng tôi đã đưa ra phương án song song hóa giải thuật tuần tự ở trên như sau:

- Bước 1: Tại máy master, tập tin lưu trữ thông tin của đồ thị sẽ được đọc vào bộ nhớ. Khởi tạo ma trận sức ảnh hưởng 2 chiều $T_{n \times n}$.
- Bước 2: Cũng tại máy master, một luồng xử lý riêng biệt sẽ được tạo ra, có nhiệm vụ nhận thông tin các đường đi tạo thành và cập nhật sức ảnh hưởng vào ma trận $T_{n \times n}$. Luồng xử lý này sẽ được thực thi song song, cùng lúc với việc master phân chia tài nguyên cho các worker và chờ nhận lại kết quả các đường đi từ các worker.

Ở bước này, máy master sẽ khởi tạo luồng xử lý để lấy thông tin từ các cạnh cập nhật vào ma trận $T_{n \times n}$.

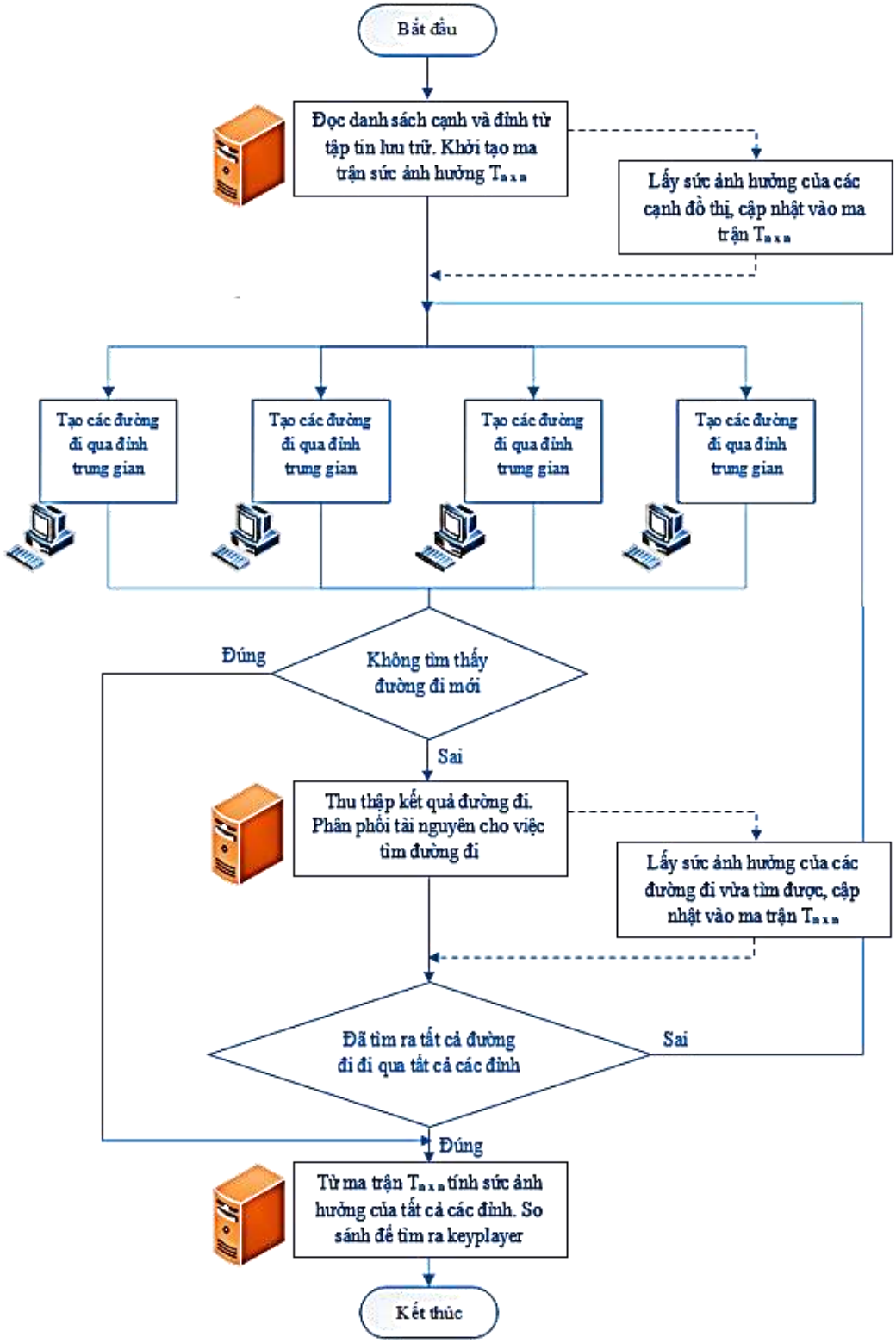
- Bước 3: Song song với luồng xử lý ở bước 2, thông qua RDD, master sẽ đưa danh sách cạnh đến tất cả các máy worker để tìm ra *đường đi qua trung gian một đỉnh*. Kết quả từ các worker được trả về lại master. Tại đây master sẽ khởi tạo luồng xử lý riêng để ghi nhận kết quả vào ma trận $T_{n \times n}$.
- Bước 4: Máy master được lập trình để lặp lại việc phân phối các phần của tập *đường đi qua trung gian m – 1 đỉnh* và tập cạnh cho các worker để các worker tìm *đường đi qua trung gian m đỉnh*. Kết quả đường đi được tìm thấy mà các worker trả về cho master sẽ được luồng xử lý đồng thời cập nhật vào ma trận $T_{n \times n}$.



Hình 3. Phân luồng xử lý tại máy master

- Bước 5: Sau khi tìm ra tất cả các đường đi của đồ thị và cập nhật kết quả vào ma trận $T_{n \times n}$. Máy master sẽ tính sức ảnh hưởng của từng đỉnh trong đồ thị. So sánh sức ảnh hưởng của các đỉnh và kết luận đỉnh nào là đỉnh có sức ảnh hưởng lớn nhất trong đồ thị mạng xã hội.

Và trong những bước thực hiện của giải thuật trên, có những bước được chia nhỏ cho các máy worker trong hệ thống phân tán xử lý, có những bước sẽ do máy master đứng ra quản lý, thu thập các dữ liệu từ các đơn vị và tính toán kết quả chung. Thuật giải trên được chúng tôi mô tả trực quan trong sơ đồ sau với các công việc do master xử lý sẽ được đặt cạnh một hình máy chủ màu cam, còn worker sẽ là các máy tính con màu xanh dương:



Hình 4. Sơ đồ xử lý phân tán

III. THỬ NGHIỆM

A. Cấu hình thử nghiệm

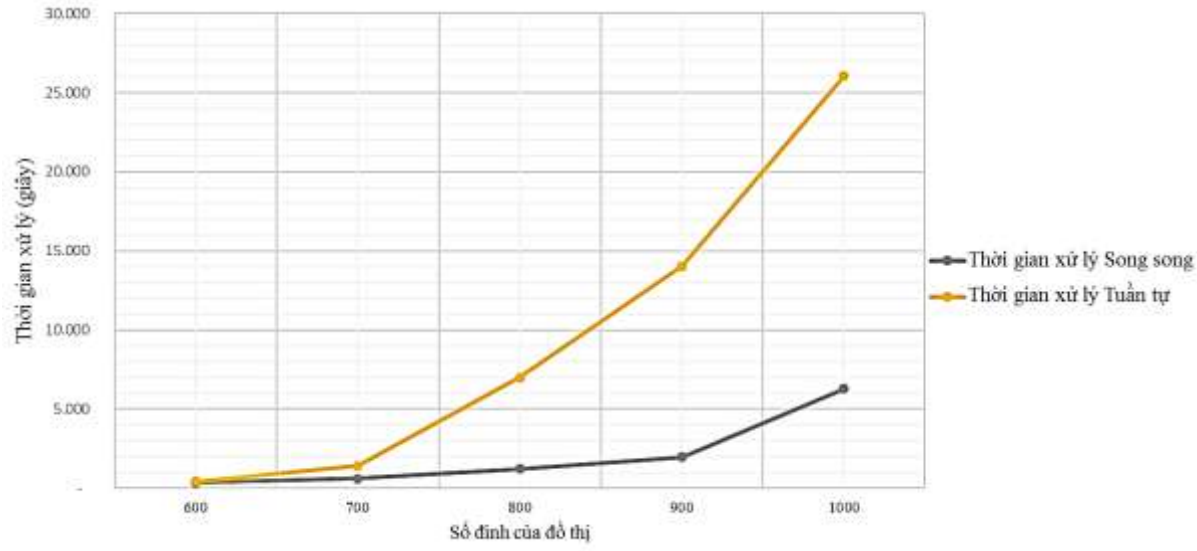
Chúng tôi đã cài đặt hệ thống phân tán thử nghiệm trên hệ thống máy chủ của trường Đại học Công nghệ Thông tin, các máy ảo được cấp có cấu hình giống nhau với tổng số nhân xử lý là 80 nhân và bộ nhớ RAM có dung lượng 80GB. Hệ thống bao gồm 10 máy được kết nối với nhau, trong đó có một máy vừa là máy con với vai trò xử lý, vừa là máy chủ với vai trò quản lý cấp phát tài nguyên, dữ liệu; thu thập, tổng hợp kết quả, xử lý những tính toán cục bộ. Các máy chạy hệ điều hành Ubuntu 16.04, được cài đặt Apache Spark 1.6.2.

B. Kết quả và đánh giá

Sau quá trình cài đặt và thử nghiệm trên hệ thống, chúng tôi đã có được những kết quả tương đối khả quan như số liệu được ghi nhận trong bảng sau. Bên cạnh đó, chúng tôi có cài đặt thử nghiệm thuật giải trên đối với nền tảng tuần tự, để qua đó có được những so sánh về kết quả đạt được trên hai hệ thống.

Bảng 1. Kết quả thử nghiệm giải thuật trên nền tảng tuần tự và song song

Số đỉnh của đồ thị	Tổng số đường đi trong đồ thị	Thời gian xử lý (giây)	
		Song song	Tuần tự
600	3 549 444	341,324	426,346
700	5 413 683	608,365	1 417,233
800	9 481 402	1 241,276	7 015,631
900	13 059 927	1 977,202	14 051,420
1000	19 600 515	6 290,760	26 051,420



Hình 5. Biểu đồ Thời gian xử lý trên nền tảng tuần tự và song song

Do điều kiện không cho phép nên chúng tôi chưa thử nghiệm được trên những tập dữ liệu có số lượng lớn hơn nữa. Tuy nhiên, qua những kết quả trên chúng ta thấy được phần nào hiệu quả của thuật giải và tiềm năng mà tính toán song song, phân tán mang lại.

IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Việc phân tán hóa giải thuật xác định đỉnh có sức ảnh hưởng lớn nhất trong đồ thị mạng xã hội đã giúp nâng cao được tốc độ thực thi cũng như tính toán của hệ thống, bên cạnh đó, nó còn giúp cho việc xử lý, tính toán trên một lượng dữ liệu lớn hơn, làm cho kết quả càng chính xác và giá trị hơn. Đặc biệt đối với một vài loại dữ liệu lớn ngày nay, việc tính toán trên những nền tảng nhỏ, tuần tự ngày càng mất dần đi ý nghĩa của nó. Tính toán song song, phân tán, thay vào đó, đã giúp ích rất nhiều trong việc giải quyết những bài toán phức tạp, đòi hỏi tốc độ nhanh và độ chính xác cao.

Tuy nhiên, việc sử dụng thuật giải mang tính vét cạn làm tiêu tốn khá nhiều tài nguyên, chi phí cho việc tính toán, điều này cũng khiến cho việc cài đặt triển khai trở nên khó khăn và phức tạp hơn. Trong thời gian tới, nhóm tác giả sẽ tiếp tục theo đuổi hướng nghiên cứu về đồ thị hóa mạng xã hội bằng những thử nghiệm mới, chẳng hạn như đưa vào thêm những đặc trưng, yếu tố của mạng xã hội mà hiện tại chưa được đồ thị hóa, để phần nào tăng tính chính xác của những tính toán, góp phần đưa bài toán này áp dụng vào thực tế cuộc sống. Ngoài ra nhóm sẽ tìm hiểu và áp dụng các thuật toán mới thay thế thuật toán vét cạn nhằm tối ưu hơn về mặt tài nguyên, chi phí.

V. LỜI CẢM ƠN

Nghiên cứu này là sản phẩm của đề tài "Nghiên cứu các kỹ thuật xử lý dữ liệu lớn, áp dụng cho việc xác định những cá nhân có tầm ảnh hưởng trong mạng xã hội" mã số D2015-07, thuộc Trường Đại học Công nghệ Thông tin – ĐHQG TP.HCM.

VI. TÀI LIỆU THAM KHẢO

[1] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker and I. Stoica, "Spark: Cluster Computing with Working Sets," in *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, Boston, MA, 2010.

[2] H. Karau, A. Konwinski, P. Wendell and M. Zaharia, "Learning spark: lightning-fast big data analysis", O'Reilly Media, Inc., 2015.

A DISTRIBUTED MODEL OF SCALABLE ALGORITHM FOR IDENTIFYING THE MOST INFLUENCE NODE IN A SOCIAL NETWORK

Nguyen Ho Duy Tri, Ngo Thanh Hung

ABSTRACT— *The discovery of key player in social networks has attracted the attention of researchers. In the previous research, we have proposed a method to identify the key player in a social network based on the sum of impact from a given node to all others. When implementing and applying such algorithm as a serial of instructions for a social network, which may be hundreds or thousands of nodes, it can be impractical to solve them on a single computer. To overcome such drawbacks, an algorithm for identifying a key player based on parallel computing is proposed in the paper. We test such approach and conclusions are drawn to describe the encouraging results we achieved.*