HIỂN THỊ VÀ KHẢO SÁT KỸ CÁC MẠNG THÔNG TIN TIẾN HÓA TRONG WIKIPEDIA

Ee-Peng Lim[2], Agus Trisnajaya Kwee[1], Nelman Lubis Ibrahim[1], Aixin Sun[1], Anwitaman Datta[1], Kuiyu Chang[1], and Maureen[1]

[1] Trường Kỹ thuật máy tính, Đại học Công nghệ Nanyang, Singapo

[2] Trường Các hệ thống thông tin, Đại học Quản lý Singapo, Singapo

1. Mở đầu

Mạng lưới thông tin ám chỉ việc thể hiện các thực thể thông tin như là các nút, và các mối quan hệ liên thông giữa chúng như là các cạnh. Bài viết này đề cập đến việc hiển thị và khảo sát kỹ các mạng thông tin đang tiến hóa trong Wikipedia. Mỗi bài mục Wikipedia được viết bằng công cụ Wikitext mô tả một thực thể thông tin (ví dụ, một người) và chứa các đường liên kết tới các thực thể thông tin liên quan khác (ví dụ: nơi sinh, công ty nơi đang làm việc, v.v…). Sử dụng trình duyệt web, người dùng có thể lướt xem các bài mục của Wikipedia và di chuyển tới các bài mục khác qua các đường liên kết. Tuy nhiên, phương thức tương tác này của người dùng không hỗ trợ xem thông tin của Wikipedia trên toàn mạng. Do đó, thật khó thu nhận được hiểu biết sâu về các bài mục đã lướt qua trong văn cảnh của toàn bộ mạng thông tin. Wikipedia cũng duy trì các phiên bản cũ của một bài mục dưới phần lịch sử của bài mục đó. Tuy vậy, các phiên bản của bài mục cũ này chỉ được coi như những đơn vị đơn độc mà không là một bộ phận của mạng thông tin đang tiến hóa nào đó.

Hiển thị và khảo sát kỹ các mạng thông tin đang tiến hóa có tầm quan trọng vì một vài lý do sau đây. Trước hết, nó giúp hiểu được mối quan hệ giữa các nút mạng theo chiều thời gian. Khi có một đường liên kết tồn tại giữa một cặp nút trong một thời gian dài, nó được coi như bền vững hơn một đường liên kết khác chỉ tồn tại trong một thời gian rất ngắn. Hai là, việc hiển thị mạng dựa vào thời gian cho phép nghiên cứu những thay đổi đáng chú ý của mạng biểu thị một số khuynh hướng hay sự kiện thú vị. Đây chắc chắn có thể là những khuynh hướng hay sự kiện xảy ra trong thế giới vật chất.

SSNet Viz+ là một công cụ được thiết kế và ứng dụng để khắc phục hạn chế của những trình duyệt web hiện có trong việc phân tích các mạng thông tin của Wikipedia. Như là một phiên bản mở rộng của dự án nghiên cứu SSNet Viz [6] trước đây, vốn tập trung vào việc hiển thị và khảo sát kỹ các mạng ngữ nghĩa không đồng nhất, SSNet Viz+ đưa vào một chiều thời gian mới và một cách thể hiện mạng thông tin đa phiên bản. Nó hỗ trợ việc lưu trữ nhiều phiên bản của một mạng thông tin với nhiều loại nút. Bằng cách đưa vào các toán tử mới để vận hành các mạng thông tin, nó giúp người dùng hiểu rõ hơn mối quan hệ giữa các mạng, các xu hướng và sự kiện mạng. Từ các mục của Wikipedia, trước hết chúng ta tách ra được những mạng thông tin tương thích đa

phiên bản dựa vào nhiệm vụ phân tích mạng phải thực hiện. Hiện nay, bước này được thực hiện bán tự động. Các mạng tách ra được lưu trữ trong một kho nào đó. Sau đó người dùng thực hiện việc phân tích mạng tương tác trên các mạng đa phiên bản bằng cách sử dụng một tổ hợp toán tử vận hành mạng và phép tìm mạng.

Có đôi chút thách thức về mặt thiết kế đối với SSNet Viz+. Trong phần sau đây, chúng tôi sẽ phác họa một số thách thức này và cách tiếp cận chung của chúng tôi để đương đầu với những thách thức đó.

Khi có nhiều phiên bản của các mạng thông tin, làm thế nào để có thể hiển thị và khảo sát chúng kỹ lưỡng mà không làm quá tải người dùng với bộn bề thông tin? Mạng thông tin có thể có kích cỡ rất lớn. Do đó, SSNet Viz+ được thiết kế để cho thấy chỉ một phiên bản mạng vào một thời điểm. Hơn nữa, đối với một phiên bản nhất định của mạng thông tin, chúng tôi cho phép người dùng lựa chọn một tập hợp con của các nút, gọi là các nút neo, và các nút lân cận của chúng để hiển thị. Cách tiếp cận khảo sát kỹ gia tăng này làm giảm số lượng thông tin được xem cùng một lúc, cho phép người dùng tập trung nhiều hơn vào những mạng con và các phiên bản hay.

- Làm thế nào để có thể thực hiện phân tích mạng dựa vào thời gian một cách dễ dàng? Một mục đích quan trọng của việc sử dụng SSNet Viz+ là tìm được những mạng con và các phiên bản hay thông qua các phương tiện hiển thị. Chúng tôi thực hiện khả năng này bằng cách đưa vào: (a) thanh cuộn thời gian để xem xét các nút được tạo lập vào các thời điểm khác nhau; (b) con số thống kê các nút cũ để hướng dẫn người dùng trong khi khai thác dữ liệu về hoạt động của người sử dụng Wikipedia, có thể nhận dạng được các phiên bản mạng hay; (c) toán tử đồ thị delta để so sánh các phiên bản khác nhau của mạng; và (d) khảo sát kỹ mạng có thể tìm, cho phép bổ sung những nút đáng chú ý mới vào các mạng con được hiển thị để mở rộng phạm vi phân tích mạng.

- Làm thế nào để có thể tìm kiếm các mạng thông tin có nhiều phiên bản?

SSNet Viz+ tích hợp một máy tìm theo từ khóa cho phép người dùng tìm kiếm những nút đáng quan tâm có chứa những từ khóa trong thời gian tìm cụ thể đối với người dùng. Vì cùng một số từ khóa có thể xuất hiện trong các phiên bản khác nhau của mạng thông tin, chúng tôi phải phát triển một chức năng sắp xếp kết quả của nút mới để xem xét các phiên bản có chứa các từ khóa này.

Trong bài này, chúng tôi sẽ minh họa SSNet Viz+ bằng việc sử dụng thí dụ của một mạng thông tin khủng bố. Chúng tôi tạo lập mạng thông tin này bằng cách tập hợp các bài mục có liên quan đến khủng bố và các đường liên kết giữa chúng. Sử dụng thí dụ thực tế này, chúng tôi nêu bật các điểm mạnh của việc sử dụng SSNet Viz+ để phân tích các mạng thông tin tiến hóa.

2. Công trình liên quan

Có một số công trình nghiên cứu trước đây liên quan đến bài này. Tập hợp đầu tiên của các công trình liên quan có liên hệ

tới phép phân tích hình ảnh trên đồ thị hoặc các dữ liệu mạng. Một công trình khảo sát tốt về hiển thị trên đồ thị hoặc mạng được trình bày ở [4]. Công trình khảo sát này mô tả nhiều kỹ thuật trình bày đồ thị khác nhau có thể áp dụng cho cả cấu trúc hình cây và cấu trúc đồ thị. Nó cũng bao quát cả các kỹ thuật phân nhóm để tổng kết số lượng thông tin mạng được hiển thị. Trong văn cảnh của tài liệu khoa học, Chen cũng đưa ra một số kỹ thuật hiển thị để trích dẫn, đồng trích dẫn và các mạng khác được đề cập tới trong các sưu tập tư liệu [3]. Công trình nghiên cứu của ông cũng đưa ra khái niệm về điểm chốt, ám chỉ một bài mục có giá trị cao, làm trung tâm giữa nhiều bài mục. Một bài mục như thế được tin là quan trọng đáng để người dùng xem xét. Yang và các tác giả khác đã mô tả một bộ dụng cụ phân tích hình ảnh để phát hiện các sự kiện trong một mạng thông tin tiến hóa. Ý tưởng chính là để xác định bảy sự kiện khác nhau được theo dõi cho mạng, để hiển thị và giải thích những thay đổi của mạng [10]. Công trình nghiên cứu này có cùng mục tiêu với chúng tôi nhưng chúng tôi dựa nhiều hơn vào dữ liệu hoạt động của người dùng để tìm ra các sự kiện. Chúng tôi kết hợp hiển thị, tìm kiếm và phân tích trong SSNet Viz+ trong khi bộ dụng cụ này được dùng nhiều hơn cho việc hiển thị.

Trong Wikipedia, cũng có những công trình nghiên cứu về việc tách ra các mạng đề tài hoặc bản thể học từ nội dung của bài mục. Wu và Weld đưa ra bộ sinh Bản thể học Kylin (KOG) để tách bản thể học từ các hộp thoại thông tin của các bài mục trong Wikipedia và kết hợp nó với Wordnet (Mạng văn bản) [9]. DBpedia giới thiệu những nỗ lực đang tiến hành để tách nội dung được định nhãn của Wikipedia, biến chúng thành một cơ sở tri thức của dữ liệu RDF [1]. Kitur, Chi và Suh đã tách các nhãn loại của các bài mục trong Wikipedia, và tìm thấy từ đó, sự phân bố theo đề tài ở mức độ cao của mỗi bài mục khi sử dụng các nhãn loại được tách ra. Bằng cách phân tích theo đề tài dựa vào các đề tài của Wikipedia vào tháng giêng 2008 và tháng 7 năm 2006, họ đã kết luận rằng "Các khoa học tự nhiên và vật lý" và "Toán học và lôgic" là hai đề tài phát triển nhanh nhất [5]. Để hiển thị lịch sử chỉnh lý một bài mục trong Wikipedia, Nuné và các tác giả khác đã đưa ra một cách thể hiện hoạt động chỉnh lý bài mục theo thời gian và thực hiện cách đó như một ứng dụng gọi là WikiChanges (thay đổi Wiki) [7], HistoryViz (Hiển thị lịch sử) là một ứng dụng khác của web cho phép người dùng xem xét các sự kiện và các thực thể liên quan của thực thể người sử dụng dữ liệu trong Wikipedia [8]. Công trình nghiên cứu của chúng tôi khác với các công trình nói trên ở chỗ, chúng tôi sử dụng nhan đề bài mục trong Wikipedia như những nút trong mạng thông tin đó thay vì tách các nút hoặc sự kiện từ nội dung bài mục. Chúng tôi nghiên cứu các bài mục trong Wikipedia trong một mạng tiến hóa để đối lập với một bài mục tiến hóa biệt lập.

3. Mô hình hóa các mạng thông tin tiến hóa

3.1. Mạng thông tin đa phiên bản

Mạng thông tin là cấu trúc dữ liệu cơ bản, SSNet Viz+ được thiết kế để thể hiện và vận hành. Chúng tôi xác định một mạng thông tin G= (V,E) là tập hợp các nút V và tập hợp các cạnh E có định hướng. Mỗi nút thuộc về một loại nút nào đó và các nút cùng loại có chung một tập hợp thuộc tính. Tương tự như vây, các cạnh thuộc về những loại cạnh nhất định, nhưng chúng không mang bất kỳ thuộc tính nào. Đối với Wikipedia, các nút là các bài mục và các cạnh có định hướng là các đường liên kết từ bài mục này đến các bài mục khác. Vì Wikipedia cho phép người dùng biến đổi hoặc chỉnh lý các bài mục, nó tạo ra nhiều phiên bản của mạng thông tin cho các bài mục này.

Một bài mục $v_i \in V$ trong mạng thông tin có thể có nhiều phiên bản $\{V_{i1}, V_{i2}, \cdots, V_{i\,|Ti|}\}$ được tạo ra ở những thời điểm khác nhau $T_i = \{t_{i1}, t_{i2}, \ldots t_{i\,|Ti|}\}$. Mỗi phiên bản bài mục V_{ik} có thể có các đường liên kết tới các bài mục khác và chúng tôi chỉ các đường liên kết này bằng $E_{ik} \subseteq V$. Với một tập hợp các bài mục V như vậy, chúng tôi cho rằng, các phiên bản của chúng có thể được tuyến tính hóa theo thời điểm và chúng tôi xác định phiên bản của mạng thông tin cho từng thời điểm. Về hình thức, chúng tôi định nghĩa một **mạng thông tin đa phiên bản** (V, E) như là một chuỗi mạng thông tin được thể hiện bằng (**V, E, T**), trong đó $\mathbf{V}= \{(v_i\ t_{ij}) \mid v_i \in V \wedge t_{ij} \in T_i\}$, $\mathbf{E} =\{ (v_i', v_j') \mid \exists_{vi,vj} \in V,\ \exists\ t_{ik} \in T_i, \exists\ t_{ik'} \in T_j, v_i' =(v_i\ t_{ij}) \wedge v_j' = (v_j, t_{jk'}) \wedge t_{ik} \geq t_{jk'} \wedge v_j \in E_{ik}\}$, và T là liên hợp thời điểm của tất cả các bài mục, nghĩa là, $\mathbf{T}= \cup vi \in v\ T_i$.

Với một mạng thông tin đa phiên bản (V, E, T) như vậy, chúng tôi có thể đưa vào một mạng thông tin đa phiên bản khác với những nút nhất định $Vt \subseteq V$ ở một thời điểm *t* (*t* có thể hoặc không nằm trong **T**) được xác định bằng ($\mathbf{V}_t$, $\mathbf{E}_t$), trong đó: $\mathbf{V}_t = \{ (v_i\ t_{ij}) \in V \mid v_i \in V \wedge t_{ij}$ là thời điểm cuối cùng ở thời điểm *t* hoặc trước t $\}$, $\mathbf{E}_t = \{ (v_i\ v_j) \in E \mid v_i\ v_j \in V_t \}$. Do đó mạng thông tin cảm ứng này là một chọn động của mạng thông tin đa phiên bản ở thời điểm *t*. SSNet Viz+ được thiết kế để giới thiệu các mạng thông tin với tập hợp nút được cụ thể hóa cho người dùng vào những thời điểm đều nhau, ví dụ, hằng tuần hoặc hằng tháng, để hiển thị và khảo sát kỹ lưỡng. Điều này đòi hỏi phải có một kích cỡ dữ liệu nhỏ hơn nhiều được SSNet Viz+ vận hành so với mạng thông tin đa phiên bản gốc, từ đó giảm được mã phụ trên đầu (overhead) trong khi khảo sát kỹ mạng. Cho nên, từ đây, chúng tôi sẽ sử dụng luân phiên các mạng thông tin và các mạng thông tin cảm ứng.

3.2. Vận hành các mạng thông tin

Dựa vào một tập hợp các mạng thông tin cảm ứng với các thời điểm cách đều, chúng tôi có thể vận hành các mạng bằng cách sử dụng một tập hợp toán tử để khảo sát kỹ các mạng. Trong phần tiếp theo, chúng tôi chính thức xác định một số toán tử phục vụ cho mục đích này.

Thêm vào một nút bài mục. Về cơ

bản, toán tử này thêm một nút mới v' vào một mạng thông tin cảm ứng hiện có (V_t , E'_t) bằng cách đem lại một mạng thông tin cảm ứng mới (V'_t , E'_t)), trong đó (a) $V'_t = V_t \cup \{ (v', t') \in V | t'$ là dấu ấn thời gian cuối cùng vào hoặc trước $t\}$, (b) $E'_t = E_t \cup \{ (v', v_j) \in E | v' = (v', t') \in V'_t \wedge v_j \tilde{} E | V'_t\} \cup \{ v_j , v') \in E | v' = (v', t') \in V'_t \wedge v_j \in V'_t \}$

Khi một nút mới được thêm vào, các đường liên kết của nó tới mạng thông tin cảm ứng hiện có cũng được đưa vào mạng mới nếu các đường liên kết dẫn tới các nút trong mạng hiện có. Tương tự như vậy, chúng tôi có thể loại bỏ một nút bài mục trong mạng thông tin cảm ứng bằng cách loại bỏ nút và các đường liên kết tương ứng của nó.

Loại bỏ một nút bài mục. Toán tử loại bỏ nút xóa một nút v' từ một mạng thông tin cảm ứng hiện có (V_t , E_t) bằng cách đưa vào một mạng thông tin cảm ứng mới (V'_t , E'_t), trong đó (a) $V'_t = V_t$ - { $(v', t') | (v', t') \in V_t$ }, (b) $E'_t = E_t - \{(v', v_j) \grave{} (v', v_j) \in E_t \}$ - $\{(v_j v') | (v_j v') \in E_t \}$

Vi sai delta của hai mạng thông tin cảm ứng. Vi sai delta của hai mạng thông tin cảm ứng ở vào thời điểm $t1$ và $t2$ ($t1 < t2$), (V_{t2}, E_{t2}) - (V_{t1}, E_{t1}), đáp ứng ba mạng thông tin, cụ thể là:

- Mạng Δ^+, (V^+ , E^+), được xác định bằng (a) $V^+ = \{v | (v, t) \in V_{t2} \wedge \exists (v, t') \in V_{t1}$, và (b) $E^+ = \{ (v, w) | (v, t_k) (w t'_k) \tilde{} E_{t2} \wedge \exists (v,t_l), (w, t'_l) \in E_{t1}$.

- Mạng Δ^-, (V^- , E^-), được xác định bằng (a) $V^- = \{\grave{v} (v, t) \in V_{t1} \wedge \exists (v, t') \in V_{t2}$, và (b) $E^- = \{ (v, w) | (v, t_k) (w t'_k) \in E_{t1} \wedge \exists (v,t_l), (w, t'_l) \in E_{t2}$.

- Mạng Δ^0, (V^0 , E^0), được xác định bằng (a) $V^0 = \{v | (v, t) \in V_{t1} \wedge (v, t') \in V_{t2}$, và (b) $E^0 = \{ (v, w) | (v, t_k) (w t'_k) \in E_{t1} \wedge (v,t_l), (w, t'_l) \in E_{t2}.\}$

Các mạng thông tin Δ^+, Δ^- và Δ^0 thể hiện các phần được thêm vào, bị xóa đi và còn lại, so sánh các mạng thông tin mới hơn và cũ hơn một cách tương ứng. Chú ý rằng, phép toán này không đem lại một mạng cảm ứng. Cả ba mạng hình thành không đặc thù về thời gian.

3.3. Lấy mạng thông tin khủng bố làm thí dụ

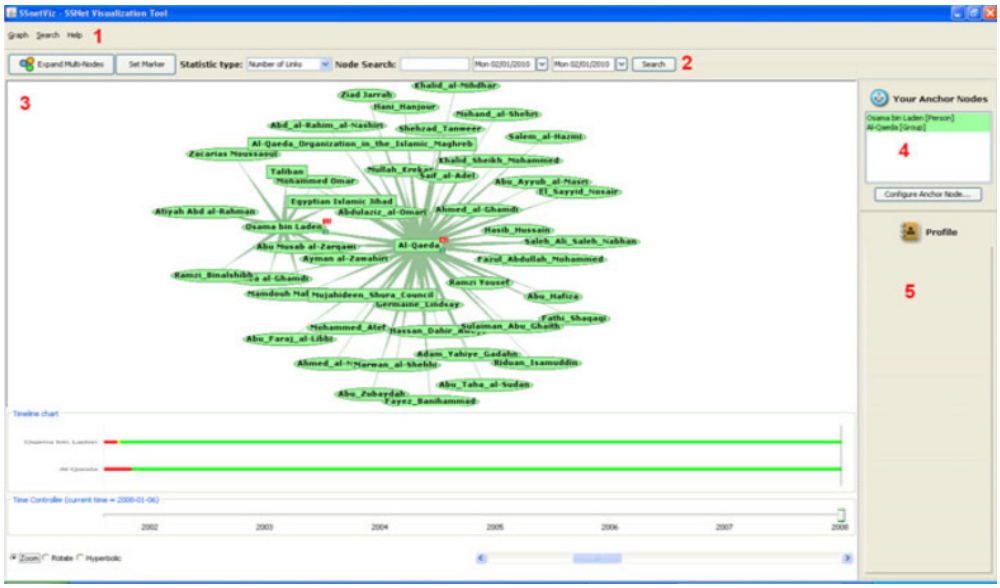
Trong dự án nghiên cứu SSNet Viz+, chúng tôi xây dựng một mạng thông tin khủng bố với nhiều phiên bản từ Wikipedia. Mạng này gồm có 813 bài mục về khủng bố, 439 bài mục về nhóm khủng bố và 1797 bài mục về sự kiện. Những bài mục này được nhận dạng thoạt đầu bằng cách sử dụng các thực thể liên quan đến khủng bố, do cơ sở tri thức về khủng bố (TKB) của Viện tưởng niệm quốc gia ngăn chặn khủng bố (MIPT) cung cấp, để định vị các bài mục tương thích của Wikipedia, tiếp theo là một chu trình kiểm tra khác của con người để loại trừ các bài mục ánh xạ sai. Một khi các bài mục về khủng bố được tìm thấy, chúng tôi nhận dạng bằng phương pháp thủ công các đường liên kết giữa các thực thể do các bài mục này giới thiệu. Tổng cộng tìm được 1922 đường liên kết. Chúng tôi tách ra các phiên bản cũ

của tất cả các bài mục được lựa chọn và xây dựng các mạng thông tin cảm ứng từ các phiên bản cũ với các thời điểm hằng tháng. Các mạng thông tin cảm ứng cuối cùng bao quát thời kỳ từ ngày 5 tháng 8 năm 2001 đến ngày 6 tháng giêng năm 2008.

4. Hiển thị và khảo sát kỹ mạng

4.1. Tổng quan về giao diện người dùng

Trong mục này, chúng tôi trình bày các đặc điểm của việc hiển thị và khảo sát kỹ SSNet Viz+. SSNet Viz+ có một giao diện tương tác với người dùng như được trình bày trong Hình 1.



Hình 1. *Giao diện người dùng của SSNet Viz+*

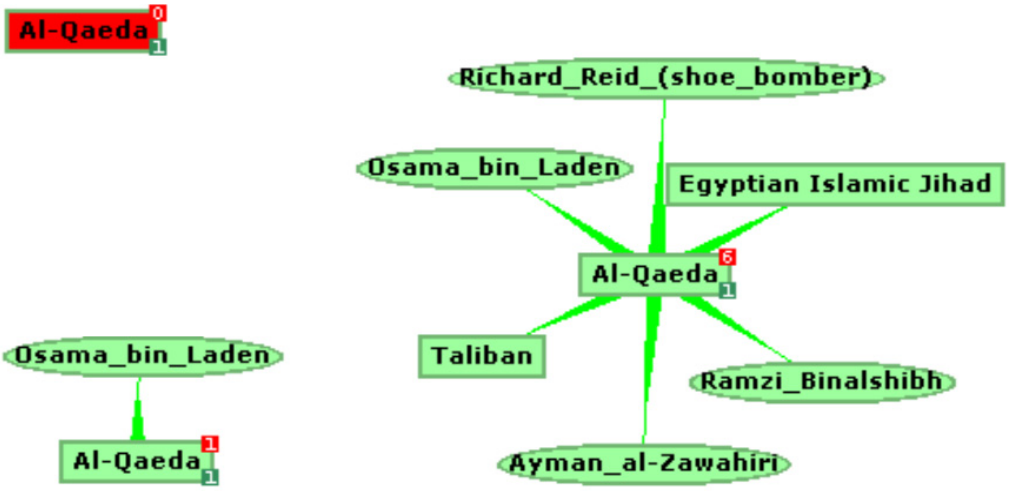
Giao diện này gồm có: (1) một thanh menu cho phép tải về các mạng thông tin để hiển thị; (2) một thanh tìm kiếm để truy vấn các nút mạng; (3) Một bảng (panel) hiển thị để hiển thị các mạng và khảo sát chúng kỹ lưỡng; (4) một bảng nút neo để duy trì một danh sách các nút neo được người dùng đánh dấu là hay; và (5) một bảng diện để hiển thị bài mục Wikipedia của bất kỳ nút nào được lựa chọn.

4.2. Hiển thị và khảo sát kỹ mạng

Giống như phiên bản SSNet Viz trước, SSNet Viz+ hiển thị các nút của một mạng thông tin dưới các hình thể khác nhau tùy thuộc vào các loại nút của chúng. Như Hình 1 cho thấy, các nút về nhóm khủng bố (ví dụ, Al-Qaeda) và các nút về tên khủng bố (ví dụ, Osama Bin Laden) được trình bày trong các hình chữ nhật và hình bầu dục một cách tương ứng. Các đường kẻ giữa các nút thể hiện mối quan hệ giữa các nút nguồn ở các đầu nét đậm và các nút đến ở các đầu nét thanh của đường kẻ. Người ta có thể thực hiện phóng to, thu nhỏ bình thường hay theo hình hyperbol, xoay ngược xuôi và xác định vị trí nút trên mạng thông tin bằng cách sử dụng các lựa chọn hiển thị và các thanh trượt ở đáy của cửa sổ chính. Một tập hợp các nút quan trọng luôn luôn được đưa vào trong bảng hiển thị có thể được người dùng đánh dấu như là các nút neo. Các nút neo được duy

trì trong bảng nút neo và các nút tương ứng trong bảng hiển thị được tô điểm bằng các hộp thoại xanh ở góc dưới cùng bên phải. Bất cứ khi nào một nút được lựa chọn, thì bài mục mà nút đó thể hiện sẽ có nội dung bằng văn bản xuất hiện trong bảng diện. Một hộp thoại con màu đỏ ở góc trên bên phải của một nút chỉ ra rằng nút này là một nút neo. Con số trong hộp thoại đỏ cho biết số lượng của các nút lân cận chưa hiển thị.

Du hành theo thời gian bằng cách sử dụng thanh cuộn. SSNet Viz+ tích hợp một số đặc điểm hiển thị độc đáo khác. Để giúp người dùng lựa chọn được một mạng thông tin cảm ứng với một thời điểm cụ thể để hiển thị, chúng tôi sử dụng một thanh cuộn thời gian ở cuối cửa sổ. Một khi thời điểm được lựa chọn, một mạng thông tin cảm ứng thích hợp sẽ được hiển thị với những màu sắc phù hợp được ấn định cho các nút chỉ *trạng thái tồn tại* của chúng. Trạng thái tồn tại của nút có thể là *tiền hé lộ* (bài mục chưa được tạo ra và trích dẫn), *hé lộ* (bài mục chưa được tạo ra nhưng đã được trích dẫn) hoặc đã được *tạo lập hoàn chỉnh* (bài mục đã được tạo ra và trích dẫn), chúng tôi sử dụng màu đỏ, màu vàng và xanh lá cây để mã hóa ba giá trị trạng thái của các nút một cách tương ứng. Để giúp người dùng hiển thị khi các bài mục của nút được trích dẫn và tạo lập, SSNet Viz+ có thể hiển thị các thanh vòng đời của các nút đã chọn, hiển thị ba giai đoạn của nút màu đỏ, vàng và xanh lá cây. Theo cách này, người dùng có thể dễ dàng nhận biết khi nào một bài mục trong nút được trích dẫn và tạo lập, và chuyển dịch thanh cuộn thời gian tới một thời điểm để xem xét mạng và các bài mục trong nút tại thời điểm đó. Thí dụ, Hình 2(a) cho thấy mạng *Al-Queda* trước khi bài mục của nó được tạo lập. Khi chúng ta chuyển dịch thanh cuộn thời gian tới các thời điểm muộn hơn, thì mạng *Al-Queda* tiến hóa để có nhiều mạng lân cận hơn như được trình bày trong Hình 2 (b) và 2 (c).

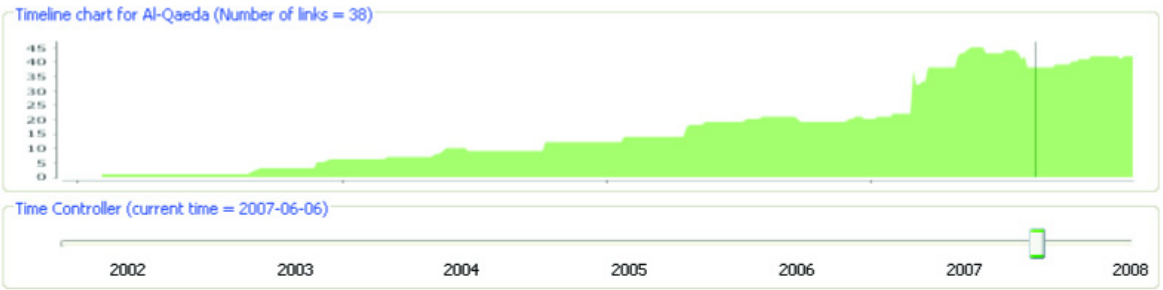


(a) Initial Al-Qaeda network (b) At June 6, 2002 c) At April 18, 2003

Hình 2. *Mạng Al-Queda ở những thời điểm khác nhau*

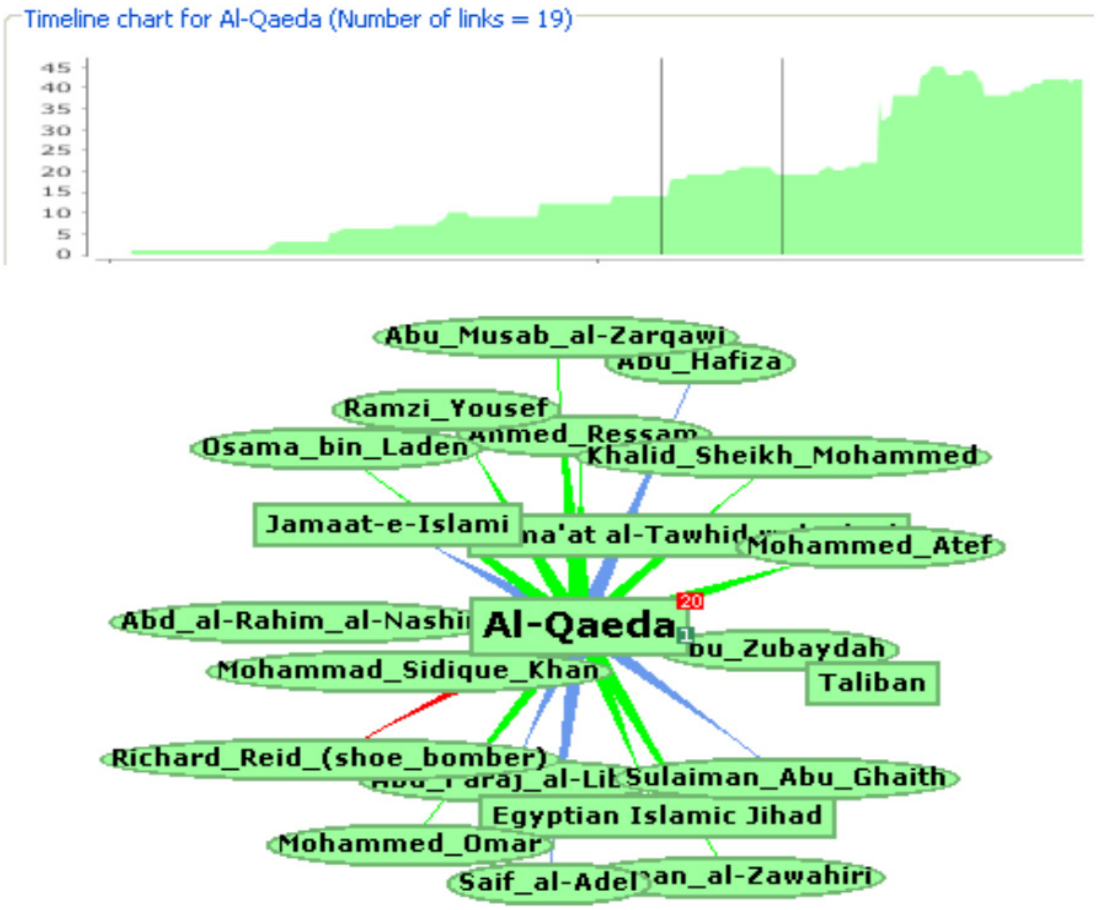
Lướt tìm các bài mục dựa vào các hoạt động trong quá khứ. Để hiển thị hoạt động của người dùng ở hậu trường một mạng thông tin, SSNet Viz+ cho phép thống kê hoạt động của người dùng có liên quan tới các nút đã chọn để hiển thị trong biểu đồ thời gian. SSNet Viz+ có thể hiển thị ba loại biểu đồ thời gian, cụ thể là (a) số lượng các đường liên kết, (b) tổng số các lần xem lại và (c) số lượng từ. Một lần nữa, thanh cuộn thời gian có thể chuyển dịch tới các thời điểm thích hợp để hiển thị các bài mục của nút và mạng vào các thời điểm mà các bài mục cho thấy những thay đổi đáng chú ý trong các con số thống kê hoạt động. Thí dụ, Hình 3 cho thấy các dữ liệu hoạt động liên kết cho nút *Al-Queda*.

Thêm và loại bỏ nút. Một nút mới có thể được thêm vào hoặc loại bỏ từ một mạng thông tin hiện hữu bằng cách gọi các toán tử thêm và loại bỏ nút được xác định trong Mục 3.2. Ngoài ra, SSNet Viz+ còn hỗ trợ bổ sung nhiều nút cùng một lúc khi một nút hiện hữu (được thể hiện bằng một hộp thoại đỏ) được lựa chọn để mở rộng sang các vùng lân cận. Việc mở rộng này sẽ bao gồm tất cả các nút được liên kết với nút được chọn lựa để bổ sung vào mạng thông tin. Việc loại bỏ nhiều nút cùng một lúc được thực hiện để xóa các nút lân cận của nút được chọn lựa, giải tỏa các vùng lân cận của các nút được lựa chọn này.



Hình 3. *Dữ liệu hoạt động liên kết*

Phân tích đồ thị Delta. Phép phân tích đồ thị Delta trong SSNet Viz+ được thực hiện bằng cách đặc tả núm đánh dấu thời điểm (sử dụng núm này ở thanh menu) và chuyển dịch thanh cuộn thời gian tới một thời điểm khác để so sánh các mạng thông tin ở hai thời điểm này. Phép toán vi sai delta trong mục 3.2 sẽ được gọi để đem lại các nút và cạnh trong các mạng Δ^+, Δ^- và Δ^0. Các nút và cạnh trong Δ^+ và Δ^- được trình bày bằng màu xanh và đỏ một cách tương ứng, còn những nút và cạnh trong Δ^0 có màu sắc không thay đổi. Hình 4 trình bày vi sai delta của mạng Al-Qaeda giữa ngày 11 tháng 4 năm 2005 và 22 tháng giêng năm 2006.

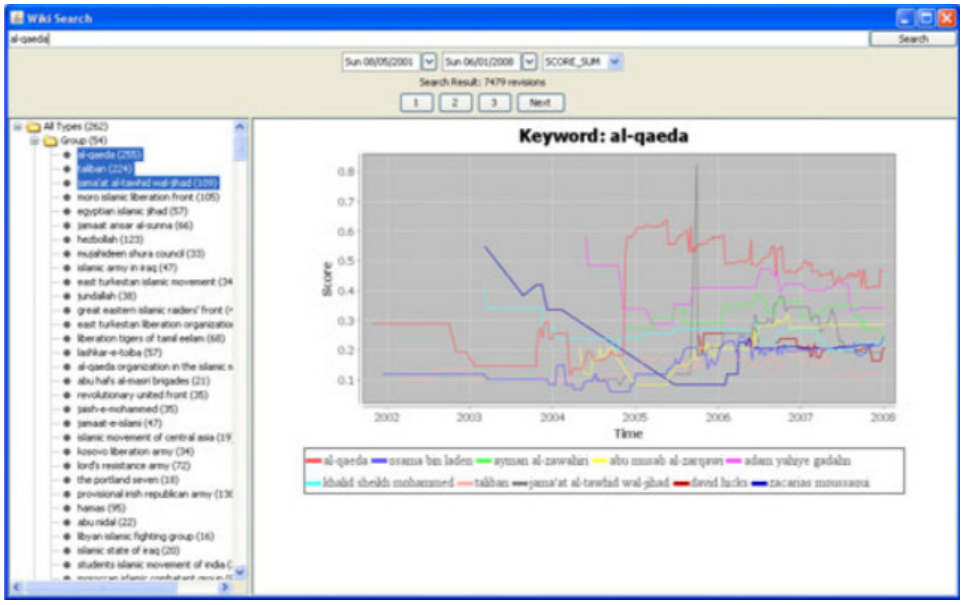


Hình 4. *Thí dụ về Vi sai Delta*

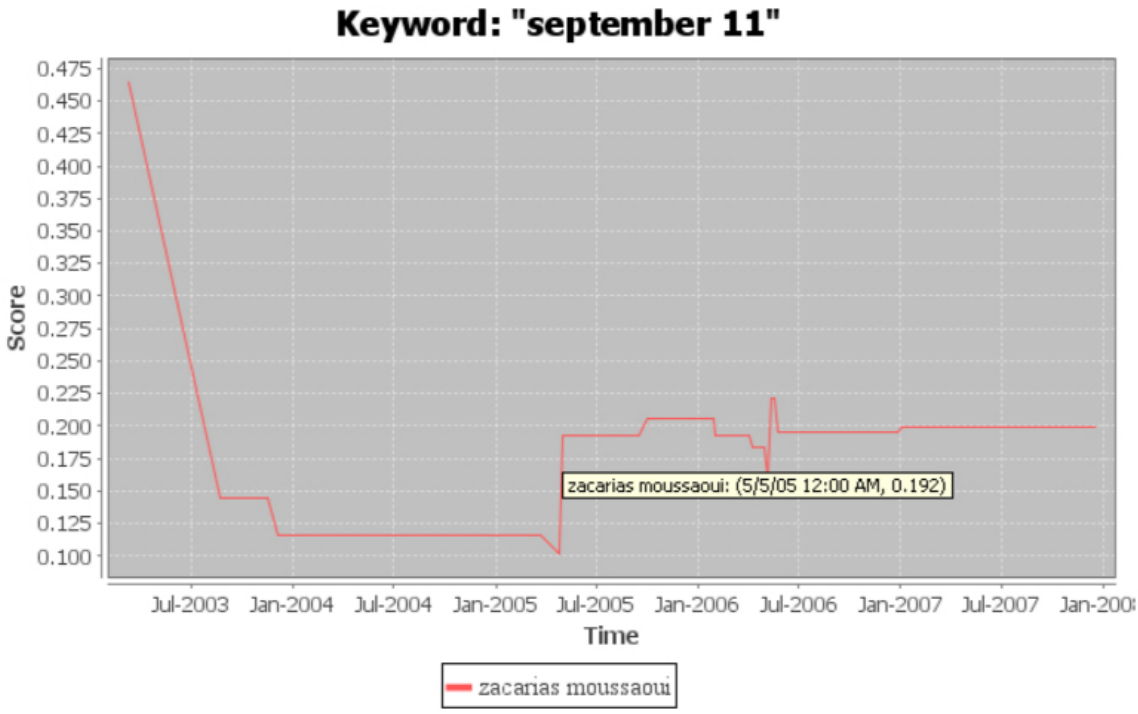
5. Tìm theo thời gian

SSNet Viz+ hỗ trợ tìm theo từ khóa thời gian với ngày tháng bắt đầu và ngày tháng kết thúc như được trình bày trong cửa sổ giao diện chính. Phép tìm đem lại các bài mục của nút với các lần chỉnh lý giữa ngày tháng bắt đầu và ngày tháng kết thúc và có chứa các từ khóa đặc tả. Sau khi sử dụng Lucene[1] để ghi điểm sự tương thích của mỗi lần chỉnh lý, SSNet Viz+ hiển thị, trong cửa sổ kết quả, các điểm số chỉnh lý qua các thời điểm của mỗi bài tương thích trong nút như một đường kẻ màu trong một biểu đồ tuyến như được trình bày trong Hình 5. Hình này cho thấy các nút có chứa từ khóa tìm kiếm «al-qaeda» và 10 nút tương thích đứng đầu được nêu bật (bao gồm cả ba nút nhóm khủng bố al-qaeda, taliban và jama'at al-tawhid wal-jihad). Chỉ trình bày điểm số chỉnh lý của 10 nút này để làm cho biểu đồ dễ đọc. Những lần chỉnh lý mới đây của nút al-qaeda (tô màu đỏ) được cho là tương thích nhất. Tiếp theo là các lần chỉnh lý nút Bin Laden (màu xanh). Biểu đồ tuyến có thể được tiếp tục phóng to, thu nhỏ để xem chi tiết hơn.

[1] http : //lucene.apache.org/



Hình 5. *Cửa sổ tìm kiếm theo thời gian*



Hình 6. *Phát hiện sự kiện trong kết quả tìm kiếm*

Một biểu đồ tuyến như thế cung cấp cách diễn giải hữu ích kết quả tìm kiếm theo thời gian.

Bảng phía trái của cửa sổ kết quả, liệt kê tất cả các nút kết quả tìm kiếm được nhóm hợp theo loại nút sắp xếp thứ bậc theo mức độ tương thích. Mỗi nút kết quả có một số lượng lần chỉnh lý thỏa mãn các tiêu chí tìm kiếm được trình bày như một con số trong ngoặc đơn. Mỗi lần, chỉ có 10

nút tương thích được nêu lên và người sử dụng lúc nào cũng có thể chọn 10 nút tương thích, trước hoặc sau, để xem xét.

Có thể sử dụng phép tìm theo thời gian kết hợp với biểu đồ tuyến để định vị các sự kiện tương thích ở những nút tương thích. Thí dụ, khi sử dụng từ khóa "ngày 11 tháng 9" giữa ngày 5 tháng 8 năm 2001 đến ngày 1 tháng 6 năm 2008, chúng tôi tìm được điểm số chỉnh lý tương thích của tên khủng bố Zakarias Moussaoui xuất hiện đột ngột vào khoảng 5 tháng 5 năm 2005 như Hình 6 cho thấy. Sau khi kiểm tra đối chiếu, chúng tôi phát hiện ra rằng, Moussaoui đã làm tòa án phải ngạc nhiên khi nhận tội trước mọi cáo buộc hắn liên quan đến ngày 11 tháng 9 trong khoảng thời gian đó. Việc này đã làm cho từ khóa "ngày 11 tháng 9" được sử dụng thường xuyên hơn trong bài mục về hắn.

Những nút được lựa chọn trong kết quả tìm kiếm có thể được bổ sung vào bảng hiển thị của SSNet Viz+ như những nút neo bằng cách dịch chuyển và thả nút vào bảng hiển thị. Điều này kết hợp có hiệu quả kết quả tìm kiếm với việc hiển thị và khảo sát kỹ.

6. Kết luận

Trong bài này, chúng tôi mô tả việc thiết kế và sử dụng SSNet Viz+ , một công cụ để hiển thị, khảo sát kỹ và truy vấn các mạng thông tin đang tiến hóa trong Wikipedia. SSNet Viz+ đã được thực hiện bằng cách sử dụng Java với các dữ liệu của mạng thông tin được tách ra lưu trữ trong cơ sở dữ liệu MySQL. SSNet Viz+ quản trị nhiều phiên bản của các mạng thông tin và hỗ trợ các thao tác để vận hành các mạng này. Sử dụng mạng khủng bố làm thí dụ, chúng tôi cho thấy các khả năng khác nhau của SSNet Viz+, bao gồm việc khảo sát kỹ các mạng thông tin theo mốc thời gian, so sánh các mạng thông tin qua các mốc thời gian, và trình bày kết quả tìm kiếm theo các biểu đồ tuyến.

SSNet Viz+ hiện đang được lấy ý kiến đánh giá của người sử dụng, đó là các chuyên gia về chủ nghĩa khủng bố từ Trung tâm quốc tế nghiên cứu bạo lực chính trị và khủng bố (ICPVTR). Các vấn đề nghiên cứu trong tương lai về SSNet Viz+ gồm có nhận dạng và hiển thị tự động các nút và đường liên kết đáng chú ý dựa trên cơ sở dữ liệu hoạt động của người dùng và tổng kết kết quả tìm kiếm bao gồm các mục của các nút được liên kết.

Vũ Văn Sơn *(Dịch)*

Nguồn: *"the Role of Digital Libraries in a Time of Global Change": ICADL 2010, LNCS 6102, pp. 50-60*

Tài liệu tham khảo

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia – a crystallization point for the web of data. Web Semantics: Science, Service and Agents on the World Wide Web 7(3) (September, 2009)

2. Brin S., Page L. : The anatomy of a large scale hypertextual web search engine. Computing Network and ISDN Systems 30(1-7) (1998)

3. Chen C.: Visualizing semantic spaces and author co-citation networks in digital libraries. Information Processing and Management, 401-420 (1999)

4. Herman, L., Melancon, G., Marshall, M.S.: Graph visualization and navigation in information visualization: A survey. IEEE Transactions on Visualization anf Computer Graphics 6(1), 24-43 (2000)

5. Kittur, A., Chi, E., Suh, B. : What's in wikipedia ? mapping topics and conflict using socially annotated category structure. In: ACM CHI (2009)

6. Lim, E.-P., Maureen, Ibrahim, N., Sun, A,. Datta, A., Chang, K., Ssnetviz : a visualization engine for heterogeneous semantic social networls. In: International Conference on Electronic Commerce (2009)

7. Nunes, S., Ribeiro, C., Gabriel, D.: Wikichanges – exposing wikipedia revision activity. In: WikiSym (2008)

8. Sipos, R., Bhole, A, Fortuna, B, Grobelnik, M., Mladenic, D.: Historyviz- visualizing events and relations extracted from wikipedia. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvonen, E., Mizoguchi, R., Oren, E., Sabou M., Simperl, E., (eds.) ESWC 2009. LNCS, vol. 5554. Springer, Heidelberg (2009)

9. Wu, F., Weld, D.: Automatically refining the wikipedia infobox ontology. In: WWW (2008)

10. Yang, X., Asur, S., Parthasarathy S., Mehta, S.: A visual-analytic toolkit for dynamic interaction graphs. In KDD, pp. 1016-1024 (2008)