Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00042

LAN TRUYỀN CHỦ ĐỀ KHOA HỌC TRÊN MẠNG TRÍCH DẪN

Nguyễn Trác Thức [1], Phạm Thế Anh Phú [1], Đỗ Phúc [1]

[1] Trường Đại học Công nghệ thông tin, Đại học quốc gia TP. HCM

thucnt@uit.edu.vn, phamtheanhphu@gmail.com, phucdo@uit.edu.vn

TÓM TẮT: *Bài báo này trình bày nghiên cứu của chúng tôi về việc kết hợp giữa phân tích mạng trích dẫn và phân tích nội dung để nắm bắt sự tiến hóa, lan truyền của các chủ đề nghiên cứu giữa các bài báo khoa học trong lĩnh vực khoa học máy tính. Đóng góp chính của nghiên cứu này là đề xuất mô hình lưu trữ mạng trích dẫn sử dụng cơ sở dữ liệu (CSDL) đồ thị thay cho cách thức lưu trữ kinh điển dưới dạng CSDL quan hệ; kết hợp các kỹ thuật phân tích mạng trích dẫn với thuật toán mô hình hóa chủ đề (topic models) để có sự phân tích đầy đủ, toàn diện về bài báo khoa học nhằm giúp phát hiện nguồn gốc của các chủ đề nghiên cứu và theo dõi sự lan truyền của các chủ đề trên mạng trích dẫn.*

Từ khóa: *mạng trích dẫn, mô hình hóa chủ đề, sự tiến hóa của chủ đề, cơ sở dữ liệu đồ thị, dữ liệu lớn.*

I. GIỚI THIỆU

Khi triển khai một công trình nghiên cứu khoa học, bài báo khoa học chính là nguồn dữ liệu đầu vào và cũng là sản phẩm đầu ra ghi nhận kết quả xuyên suốt quá trình nghiên cứu. Khi bắt đầu một nghiên cứu, người nghiên cứu sẽ tìm đọc các bài báo của tác giả khác về lĩnh vực dự định nghiên cứu nhằm: học những kiến thức nền tản và nắm bắt xu thế nghiên cứu hiện tại. Trên cơ sở đó, nhà nghiên cứu định ra con đương đi của mình, tìm hướng nghiên cứu riêng của mình. Bắt đầu từ việc đọc và thu thập tri thức từ các bài báo của các nhà nghiên cứu đi trước và kết thúc ở việc công bố bài báo của bản thân mình, đó là một chu trình bắt buộc của quá trình nghiên cứu.

Trong toàn bộ quá trình trên hoạt động phân tích và đánh giá một bài báo nghiên cứu là thao tác rất quan trọng. Việc phân tích đánh giá các bài báo khoa học đã được thực hiện từ rất lâu dựa trên phân tích mạng trích dẫn (citation network), kết quả tiêu biểu của cách tiếp cận này chính là chỉ số trích dẫn (citation index), hệ số ảnh hưởng (Impact Factor - IF) hoặc EF hay gần hơn chỉ số H. Chỉ số trích dẫn được Garfield đưa ra vào năm 1995, đó là toàn bộ số lần một bài báo được trích dẫn trong các tài liệu khác [1]. Đây là chỉ số đơn giản nhất và quan trọng nhất vì nó là nền tảng để tính các chỉ số khác. Việc phân tích các chỉ số liên quan Chỉ số trích dẫn cho chúng ta nhiều thông tin bổ sung về chất lượng của bài báo cũng như tầm ảnh hưởng của các bài báo. Tuy nhiên, nếu thuần túy dựa vào việc phân tích mạng trích dẫn chúng ta không thể nắm bắt được bài báo nghiên cứu về vấn đề gì và nội dung nghiên cứu của bài báo đó được lan truyền như thế nào đến các bài báo khác. Trong nghiên cứu này, chúng tôi cho rằng một bài báo nghiên cứu khoa học gồm hai thành phần: một tập hợp các từ (bag of words) và một tập hợp các trích dẫn (bag of citations). Như vậy để phân tích, nghiên cứu một bài báo khoa học đầy đủ, chúng ta cần phải kết hợp việc phân tích cả hai thành phần này. Đây chính là động lực nghiên cứu của chúng tôi.

Trong thời đại dữ liệu lớn hiện nay, việc tổng hợp và phân tích các bài báo cùng mạng trích dẫn liên quan để có một cái nhìn tổng quát về vấn đề nghiên cứu đã vượt ra khỏi khả năng thực hiện của con người. Do đó, nhu cầu về việc phải có những công cụ hỗ trợ là hết sức cần thiết. Trong bài báo này, chúng tôi đã xây dựng một hệ thống có khả năng hỗ trợ việc lưu trữ và trích xuất thông tin từ mạng trích dẫn. Nhằm khắc phục những hạn chế của phương pháp phân tích mạng trích dẫn trong việc phân tích , đánh giá bài báo, chúng tôi đã tiến hành xây dựng một hệ thống cho phép kết hợp giữa việc phân tích mạng trích dẫn với phân tích chủ đề bài báo. Trên cơ sở kết hợp hai phương pháp này, chúng tôi phân tích và biết được điểm xuất phát của chủ đề nghiên cứu cũng như sự tiến hóa, lan tỏa của chủ đề từ bài báo nguồn đến các bài báo khác trong mạng trích dẫn.

Trong quá trình thực nghiệm, chúng tôi đã sử dụng kho dữ liệu mạng xã hội học thuật (Academic Social Network [2]) được cung tấp tại địa chỉ https://aminer.org/data với 2,092,356 bài báo cùng 8,024,869 trích dẫn. Tất cả dữ liệu về mạng trích dẫn được chúng tôi chuyển sang lưu trữ trong CSDL đồ thị Neo4j [3] để phục vụ cho việc thực hiện các phân tích liên quan đến mạng trích dẫn. Tựa đề và tóm tắt nội dung bài báo được chúng tôi rút trích, lưu trữ sử dụng HDFS (Hadoop Distributed File System) để phục vụ cho bài toán tính toán phân tán trên dữ liệu lớn. Sau đó, toàn bộ nội dung được đưa vào xử lý trong mô hình phân tích chủ đề LDA [4] để phát hiện các chủ đề ẩn trong bài báo. Cuối cùng chúng tôi kết hợp cả hai kết quả phân tích lại để đưa ra các thông tin về nguồn gốc của chủ đề nghiên cứu trong bài báo; sự tiến hóa của các chủ đề nghiên cứu qua thời gian, và sự lan tỏa của một chủ đề nghiên cứu trên mạng trích dẫn.

Phần còn lại của bài báo được chúng tôi gồm các nội dung sau: (1) Các nghiên cứu liên quan; (2) Đề xuất hệ thống giúp theo dõi sự lan truyền của các chủ đề nghiên cứu trên mạng trích dẫn; (3) Kết quả – Đánh giá.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Việc sử dụng mạng trích dẫn để nghiên cứu về lịch sử phát triển và sự tương quan giữa các chủ đề nghiên cứu khoa học đã được Garfield đề cập đến từ rất sớm trong [5] và [6]. Trong bài báo [6], Garfield cho rằng chỉ số SCI có hai chức năng chính: cho biết chủ đề các nhà nghiên cứu đã công bố và cho biết mối liên kết tác động qua lại giữa các

chủ đề nghiên cứu. Mặc dù còn nhiều hạn chế nhưng thông tin mà mạng trích dẫn mang lại vẫn có giá trị cho biết mức độ tác động của một bài báo khoa học. Do đó, phần lớn các đánh giá về bài báo khoa học hiện nay như chỉ số IF, chỉ số H đều là các cải tiến từ khái niệm ban đầu.

Nghiên cứu mạng trích dẫn tập trung vào các bài toán suy diễn trên mạng trích dẫn mà thực chất là một đồ thị lớn. Trong bài báo [7], các tác giả đã tiến hành phân tích mạng trích dẫn các bài báo trong hội nghị khoa học LAK và tìm ra sự xuất hiện của hướng nghiên cứu nổi trội, tác giả đã lập ra bản đồ cấu trúc các ngành khoa học. Cho và cộng sự trình bày nghiên cứu mạng trích dẫn các bài báo trong lĩnh vực công nghệ giáo dục nhằm thúc đẩy sự hợp tác liên ngành khoa học. Kỹ thuật phân tích mạng xã hội đã được dùng để phân tích mạng trích dẫn nhằm tìm ra các thuộc tính quan hệ trên mạng trích dẫn [8]. Việc phân tích mạng trích đãn cũng đã thực hiện trong [9] để phân tích các bài báo đề cập đến lĩnh vực điốt phát quang hữu cơ (OLED) sử dụng phương pháp gom cụm cấu trúc để điều tra cấu trúc của các bài báo nghiên cứu và phát hiện lĩnh vực nghiên cứu nổi trội. Trong [10], các tác giả đã phân tích mạng trích dẫn các bài báo như một mạng xã hội; sau đó khảo sát sự lan truyền thông tin trong mạng trích dẫn bằng các kỹ thuật phân tích mạng xã hội qua các phân tích quan hệ giữa các trích dẫn và tác động của các bài báo. Nhóm tác giả N. J. van Eck và L. Waltman cũng đã xây dựng phần mềm CitenetExplorer cho phép xây dựng mạng trích dẫn từ kho dữ liệu bài báo khoa học Web of Science. Hệ thống cho phép phân tích và biểu diễn trực quan mạng trích dẫn trong kho dữ liệu bài báo khoa học [11].

Như đã phân tích bên trên, mạng trích dẫn của các kho dữ liệu bài báo khoa học là các đồ thị cực lớn do đó việc lưu trữ mạng trích dẫn trong các hệ quản trị cơ sở dữ liệu truyền thống trở nên không hiệu quả. Trong bài báo [12], các tác giả đã giới thiệu cơ sở dữ liệu đồ thị để lưu trữ các thông tin có tính chất liên kết với nhau. Thông qua việc sử dụng cơ sở dữ liệu đồ thị Neo4j, thông tin có tính chất liên kết được tổ chức lưu trữ, phân tích và truy vấn một cách dễ dàng nhờ ngôn ngữ truy vấn CSDL đồ thị Cypher. Cypher là một ngôn ngữ đơn giản và hiệu quả trong việc thực hiện các thao tác trên đồ thị như: duyệt đồ thị, tìm miền liên thông, tìm đường đi ngắn nhất giữa hai đỉnh…

Bên cạnh mạng trích dẫn, một phần không thể thiếu đó chính là nội dung của bài báo. Trong các bài báo [13] và [14], nhóm tác giả đề cập đến các kho sưu tập lớn các bài báo khoa học. Nhà khoa học phải truy cập hàng triệu bài báo trong lĩnh vực nghiên cứu của mình để tìm các bài báo liên quan đến nội dung nghiên cứu đang quan tâm. Tác giả đã sử dụng mô hình LDA để khám phá các chủ đề trong kho bài báo khoa học. Tác giả đã thực hiện khám phá 50 chủ đề trong tạp chí Science từ 1980 – 2002. Phần mềm TopicNets trên Web cho phép phân tích và biểu diễn trực quan kho văn bản lớn dùng mô hình chủ đề để khám phá tri thức được giới thiệu trong bài báo [15]. Trong bài báo [16], nhóm tác giả đã trình bày khả năng biểu diễn trực quan kho tài liệu nhằm khám phá tri thức từ các kho văn bản lớn. Tác giả sử dụng mô hình chủ đề để khám phá các xu thế phát triển của các chủ đề nghiên cứu, quan hệ giữa các chủ đề trong quá trình tiến hóa.

Qua các công trình đã giới thiệu chúng ta có thể thấy rằng, việc chỉ sử dụng đơn lẻ các phương pháp phân tích mạng trích dẫn hoặc phân tích nội dung bài báo đều cung cấp cho chúng ta những thông tin hữu ích. Tuy nhiên, việc kết hợp cả hai nhóm kỹ thuật để tạo ra các thông tin thú vị hơn chỉ mới được triển khai gần đây. Trong [17], X. Ren và cộng sự đã đề xuất một mạng học thuật không đồng nhất (heterogeneous bibliographic network) gồm các thông tin: bài báo, tác giả, hội nghị và từ khoá. Với đề xuất này, các tác giả đã có quan tâm đến nội dung bài báo trong quá trình phân tích mạng trích dẫn và đề xuất thuật toán gom cụm trên mạng trích dẫn ClusCite. Bài báo của chúng tôi cũng hướng tới việc kết hợp kết quả của hai cách tiếp cận trên. Tuy nhiên, trong mạng học thuật không đồng nhất của chúng tôi không chỉ dựa trên từ khoá, chúng tôi đưa thêm vào khái niệm chủ đề để nhằm tìm ra những thông tin mới như sự lan truyền của chủ đề nghiên cứu trong cộng đồng khoa học cũng như quá trình tiến hóa của một chủ đề nghiên cứu.

III. KIẾN TRÚC HỆ THỐNG

Trong thời đại dữ liệu lớn hiện nay, thông tin về các công trình khoa học và mạng trích dẫn liên quan đã vượt qua khả năng xử lý của con người. Do đó, nhu cầu cần phải có những công cụ để lưu trữ và xử lý các thông tin liên quan đến bài báo khoa học là một nhu cầu thiết yếu. Hệ thống Topic Citation Evolution (TCE) chúng tôi xây dựng sẽ gồm 04 module (Xem hình 1): (1) Module xây dựng và lưu trữ mạng trích dẫn; (2) Module phân tích mạng trích dẫn; (3) Module phân tích chủ đề bài báo; (4) Module trực quan hóa sự lan truyền của chủ đề khoa học trên mạng trích dẫn.

Hình 1. Kiến trúc hệ thống Topic Citation Evolution

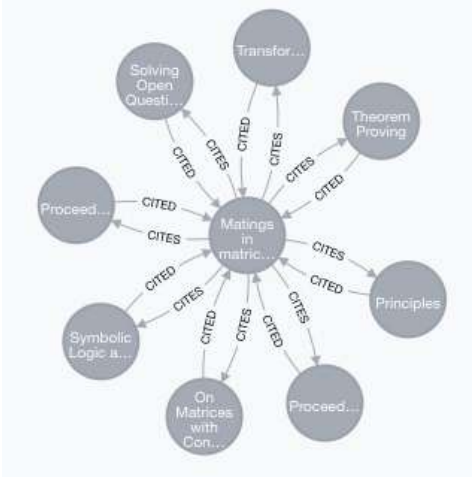
Hệ thống của chúng tôi gồm có 03 tính năng chính:

- Phân tích sự lan tỏa của các chủ đề nghiên cứu trong 1 bài báo khoa học đến tất cả các bài báo liên quan trong mạng trích dẫn.
- Phân tích sự lan tỏa của một chủ đề nghiên cứu trong một khoảng thời gian.
- Tìm kiếm các bài báo có tác động lớn (core documents) trong chủ đề nghiên cứu.

A. Module xây dựng và lưu trữ mạng trích dẫn

Tập dữ liệu về thông tin bài báo khoa học chúng tôi thu thập từ trang web https://aminer.org/data. Dữ liệu này được một nhóm nghiên cứu tại đại học Tsinghua, Trung Quốc sưu tầm và công bố. Dữ liệu về mạng trích dẫn được nhóm tác giả rút trích từ trang kho quản lý thông tin bài báo khoa học DBLP. Tóm tắt của các bài báo được các tác giả sưu tầm và công bố chung với thông tin mạng trích dẫn dưới định dạng tập tin văn bản. Chúng tôi đã xây dựng module phân tách dữ liệu và tiến hành xây dựng mạng trích dẫn lưu trữ trong CSDL đồ thị Neo4j với 2,092,356 đỉnh và 8,024,869 cạnh. Phần thông tin về tóm tắt bài báo được lưu trữ trên hệ thống HDFS phục vụ cho việc phân tích chủ đề bài báo.

Mạng trích dẫn lưu trong CSDL đồ thị Neo4j được chúng tôi tổ chức như một đồ thị G<V,E>. Mỗi đỉnh v ∈ V trong đồ thị là một bài báo khoa học với các thông tin gồm ID bài báo, tiêu đề bài báo và năm xuất bản. Mỗi cạnh e ∈ E trong đồ thị cho biết mối quan hệ trích dẫn ("CITES" và ""CITED") giữa hai bài báo (Xem hình 2).



Hình 2. Lưu trữ mạng trích dẫn trong CSDL đồ thị Neo4j

B. Module phân tích mạng trích dẫn

Mạng trích dẫn thực chất là một đồ thị có hướng không tuần hoàn G(V,E). Mỗi đỉnh trong đồ thị là một bài báo và mỗi cạnh nối từ đỉnh p_i đến đỉnh p_j thể hiện quan hệ trích dẫn hoặc được trích dẫn giữa hai bài báo p_i và p_j. Trong module phân tích mạng trích dẫn, chúng tôi đưa ra khái niệm CitingSet và CitedSet của một bài báo p như sau:

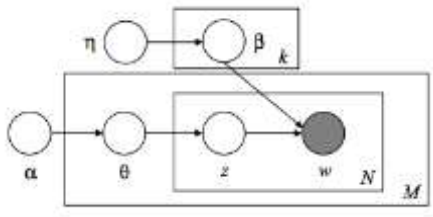
- CitingSet(p) = {m ∈ V | Tồn tại quan hệ trích dẫn (CITES) từ p đến m}. CitingSet(p) là tập các bài báo được bài báo p trích dẫn trực tiếp hoặc gián tiếp.
- CitedSet(p) = {m ∈ V | Tồn tại quan hệ được trích dẫn (CITED) từ bài báo p đến m}. CitedSet(p) là tập các bài báo đã trực tiếp hoặc gián tiếp trích dẫn bài báo p.

Với hai khái niệm trên, module phân tích mạng trích dẫn được xây dựng dựa trên thư viện GraphX của nền tảng Apache Spark [17] kết hợp với CSDL đồ thị Neo4j để tìm rút trích các bài báo cùng cây trích dẫn liên quan. Module này cho phép chúng tôi rút trích cây trích dẫn từ một bài báo bất kỳ sử dụng các thuật toán duyệt theo chiều sâu (DFS)/duyệt theo chiều rộng (BFS). Bên cạnh đó, chúng tôi còn sử dụng thuật toán phát hiện thành phần liên thông Connected Component để phát hiện ra các cụm bài báo có liên quan với nhau. Trong mỗi cụm, chúng tôi sử dụng thuật toán PageRank để phát hiện các bài báo quan trọng (core documents) trong từng cụm.

C. Module phân tích chủ đề bài báo

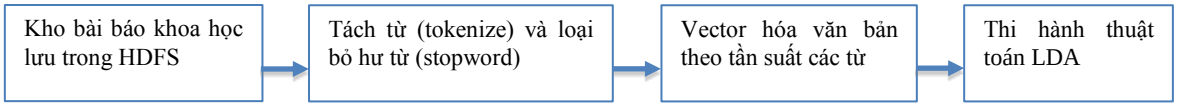
Các nghiên cứu trên mạng trích dẫn dựa trên đồ thị và đường đi trên đồ thị trong mạng trích dẫn do đó chưa chú ý nhiều vào nội dung bài báo, cụ thể là các chủ đề nghiên cứu đang được quan tâm trong bài báo. Một bài báo khoa học thường chứa nhiều chủ đề nghiên cứu và công cụ, phương pháp để giải quyết vấn đề. Ví dụ, bài báo về mô hình chủ đề của David Blei trình bày các kiến thức liên quan đến xác suất thống kế, mạng Bayes, chuỗi Markov.. Chúng tôi sẽ dùng mô hình LDA [18] để phám phá các chủ đề có trong kho dữ liệu bài báo khoa học. Việc dùng mô hình LDA để tìm các chủ đề ẩn trong các bài báo kết hợp với phân tích mạng trích dẫn sẽ cho phép khám phá nhiều thông tin và kết quả hơn so với việc chỉ phân tích nội dung bài báo hoặc chỉ phân tích mạng trích dẫn của bài báo.

Tiêu đề và tóm tắt bài báo được phân tích dùng thuật toán mô hình hóa chủ đề LDA để phát hiện ra các chủ đề tiềm ẩn. Mô hình LDA được David Blei đề xuất năm 2003. LDA là một mô hình sinh xác suất cho tập dữ liệu rời rạc như kho ngữ liệu dạng văn bản thô. Mô hình này dựa trên ý tưởng: mỗi tài liệu là sự pha trộn của nhiều chủ đề; mỗi chủ đề là một phân bố rời rạc của một tập các từ. Về bản chất mô hình LDA là một mô hình Bayes 3 cấp (cấp Tập ngữ liệu, cấp văn bản và cấp từ) trong đó mỗi phần của mô hình được coi như một mô hình trộn hữu hạn trên cơ sở tập các xác suất chủ đề (Xem hình 3). Mô hình LDA phù hợp với tập ngữ liệu với các dữ liệu rời rạc nhau được phân nhóm. Mô hình LDA dùng để mô hình hóa kho ngữ liệu nhằm phát hiện ra các chủ đề tiềm ẩn của tập ngữ liệu.



Hình 3. Mô hình LDA

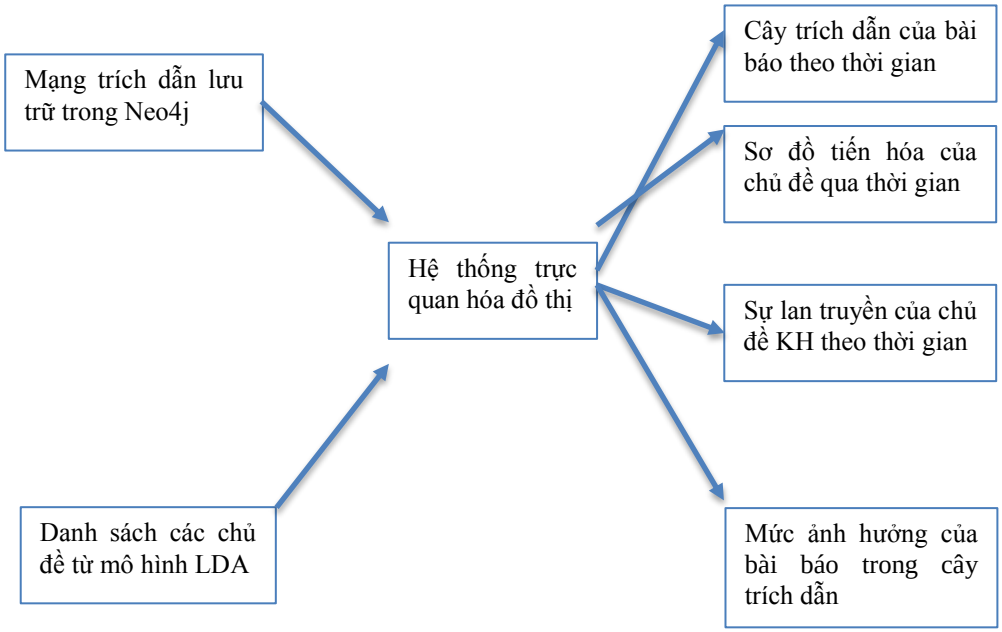
Theo định nghĩa của mô hình LDA thì chủ đề *z* là một hoặc một số từ *w* mô tả nội dung chính của một văn bản, một chủ đề gồm một nhóm các từ xuất hiện thường xuyên cùng nhau. Để phân cụm các văn bản theo LDA, mỗi văn bản được biểu diễn bằng khái niệm túi từ (bag of word). Mỗi chủ đề được biểu diễn bởi một tập các từ trong kho dữ liệu bài báo khoa học với một xác suất. Mỗi văn bản khi đó có thể chứa một tập các chủ đề với một xác suất. Để triển khai khám phá chủ đề ẩn trong các bài báo chứa trong kho ngữ liệu Arminer ASN, chúng tôi đã sử dụng thuật toán LDA được cung cấp trong thư viện tính toán phân tán Apache Spark kết hợp với các thư viện xử lý ngôn ngữ tự nhiên để tiền xử lý các tiêu đề và tóm tắt của bài báo (Xem hình 4).



Hình 4. Quy trình tiến hành khám phá chủ đề bài báo

D. Module trực quan hóa sự lan tỏa của chủ đề nghiên cứu

Module này phục vụ việc trực quan hóa các kết quả thu được từ module phân tích mạng trích dẫn và phân tích chủ đề bài báo. Bên cạnh việc hiển thị các kết quả thu được riêng lẻ từ việc phân tích mạng trích dẫn và phân tích chủ đề, hệ thống còn hướng đến việc tích hợp kết quả của hai quá trình phân tích. Việc tích hợp này giúp chúng ta dễ dàng thấy sự lan tỏa của một chủ đề nghiên cứu trên mạng trích dẫn hoặc sự lan tỏa của các chủ đề, mức độ ảnh hưởng của 1 bài báo trên cây trích dẫn của nó (Xem hình 5).



Hình 5. Module trực quan hóa kết quả phân tích mạng trích dẫn và khám phá chủ đề khoa học

Module trực quan hoá được xây dựng dựa trên thuật toán chính là suy diễn sự lan truyền chủ đề trên mạng trích dẫn. Thuật toán này gồm các bước cơ bản sau:

Thuật toán: Suy diễn sự lan truyền của chủ đề nghiên cứu trên mạng trích dẫn

Input: ID bài báo cần phân tích

Output: Đồ thị trích dẫn cùng thông tin về chủ đề trong bài báo

Bước 1: Rút trích CitingSet(p) hoặc CitedSet(p) sử dụng câu lện truy vấn Cypher

Bước 2: Với mỗi đỉnh V trong CitingSet/CitedSet

+ Suy diễn phân bố chủ đề của bài báo sử dụng mô hình LDA.

+ Cập nhật phân bố chủ đề vào thuộc tính của đỉnh V

Kết thúc lặp

Bước 3: Trực quan hoá đồ thị trích dẫn với sự phân bố theo năm xuất bản và theo chủ đề nghiên cứu.

Để tiến hành suy diễn sự lan truyền của các chủ đề nghiên cứu, chúng ta cần sử dụng thuật toán duyệt cây trên đồ thị để tiến hành rút trích CitedSet(p) và CitingSet(p) của một bài báo p ở bước 1. Thao tác này sẽ tốn khá nhiều thời gian xử lý nếu mạng trích dẫn được lưu dưới dạng cơ sở dữ liệu quan hệ. Trong bài báo của chúng tôi, công việc này được thực hiện rất nhanh chóng và đơn giản với việc sử dụng câu lệnh Cypher của hệ quản trị cơ sở dữ liệu đồ thị Neo4j với cú pháp cơ bản gồm 03 thành phần: đỉnh – cạnh – đỉnh như sau:

MATCH p = (:PAPER {idPaper:"54"}) - [r:CITES*1..5] - > ( ) RETURN p

Trong lý thuyết đồ thị, tập CitedSet và CitedSet có thể xem như một cấu trúc cây. Do đó, khi rút trích hai tập này, cơ sở dữ liệu đồ thị sẽ có thể hiện vượt trội khi cần xây dựng các cây có độ cao lớn. Với đề xuất sử dụng cơ sở dữ liệu đô thị thay thế cho cơ sở dữ liệu quan hệ, hệ thống của chúng tôi đã nâng cao được hiệu năng hệ thống đồng thời có thể giải quyết được bài toán dữ liệu lớn hiện nay.

Sự khác biệt của hệ thống chúng tôi so với [17] chính là việc tích hợp mô hình chủ đề LDA khám phá chủ đề nghiên cứu của các bài báo. Với mô hình chủ đề LDA, mỗi bài báo có thể chứa nhiều chủ đề khác nhau và do đó phản ánh đúng nội dung bài báo hơn so với các sử dụng vector từ khoá đề gom cụm bài báo. Qua đó, hệ thống có thể suy diễn chi tiết hơn sự lan truyền của từng chủ đề trên cây trích dẫn (xem kết quả thực nghiệm).

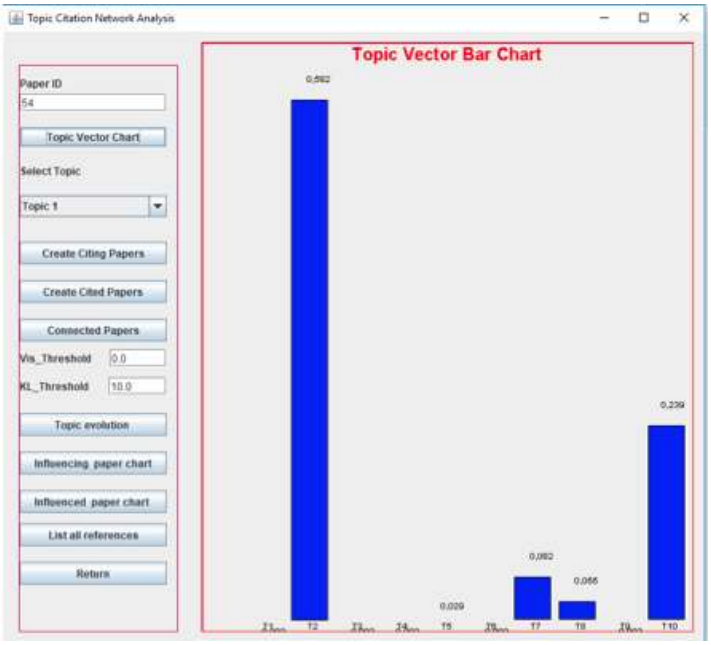
IV. KẾT QUẢ THỰC NGHIỆM

A. Khám phá chủ đề khoa học trong kho dữ liệu các bài báo khoa học trong lĩnh vực Khoa học máy tính:

Kho dữ liệu Aminer ASN chứa 2.092.356 bài báo khoa học được chúng tôi rút trích tiêu đề và tóm tắt bài báo phục vụ cho việc khám phá chủ đề khoa học. Qua quá trình tiền xử lý, chúng tôi đã rút trích ra bộ từ vựng gồm 824.578 từ với tổng số 94.360.726 từ có trong kho ngữ liệu (sau khi đã loại bỏ các từ vô nghĩa – stopword). Quá trình tiền xử lý dữ liệu được chúng tôi thử nghiệm trên hệ thống Apache Spark Cluster gồm 03 máy chủ Linux có cấu hình chung: 08 VCPUs và 16 GB RAM. Sau quá trình tiền xử lý, chúng tôi tiến hành khám phá các chủ đề tìm ẩn trong các bài báo với số chủ đề cho trước là 50. Kết quả thu được trong bảng 1 và được trực quan hóa như trong hình 6.

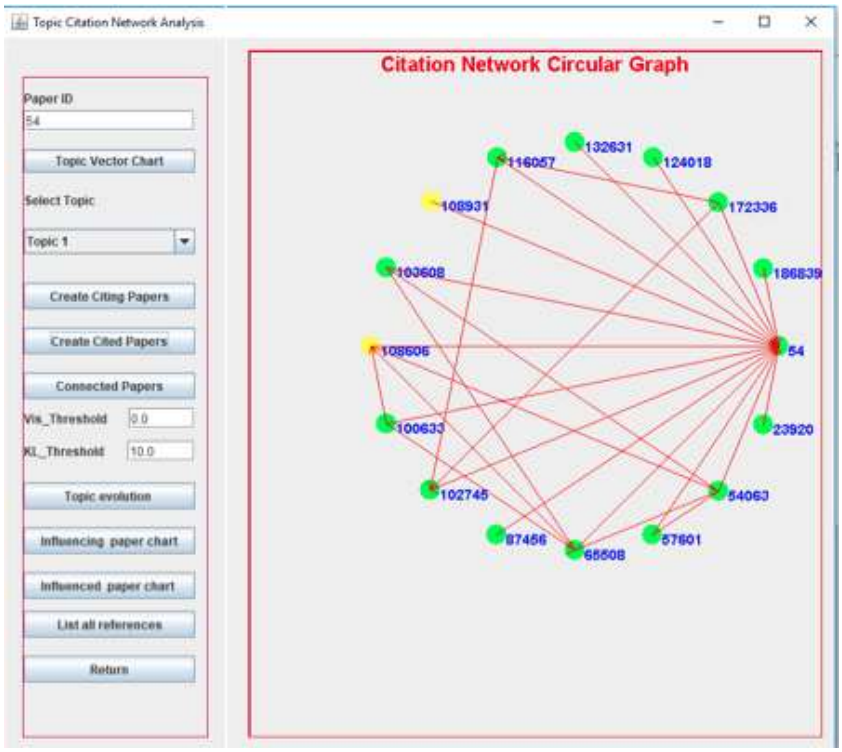
Bảng 1. Danh sách 5 từ phổ biến của 10 chủ đề được rút trích trong kho ngữ liệu Aminer ASN

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Từ	Tần suất	Từ	Tần suất	Từ	TTần suất	Từ	TTần suất	Từ	TTần suất
Program	S0.04	Code	0.03	Software	00.06	Theory	00.02	System	00.04
Time	00.02	Instruction	0.02	Development	00.03	Logic	00.02	Object	00.03
Process	00.02	Performance	0.02	Computer	00.02	Proof	00.01	Language	00.02
Method	00.02	Compiler	0.02	Engineering	00.02	Unification	00.01	Programming	00.02
State	00.01	Machine	0.02	System	00.02	Calculus	00.01	Implementation	00.02
Topic 6		Topic 7		Topic 8		Topic 9		Topic 10	
Graph	00.03	Paper	00.02	Dependency	00.05	Storage	00.04	Memory	00.05
Algorithm	00.02	Model	00.02	Database	00.03	Colleciton	00.03	Cache	00.04
Problem	00.02	Data	00.02	Relational	00.03	Garbage	00.02	Performance	00.02
Complexity	00.02	Problem	00.01	Schema	00.02	Performance	00.02	Clustering	00.02
Tree	00.01	Information	00.01	Entity-relationship	00.01	Realtime	00.01	Multiprocessor	00.01



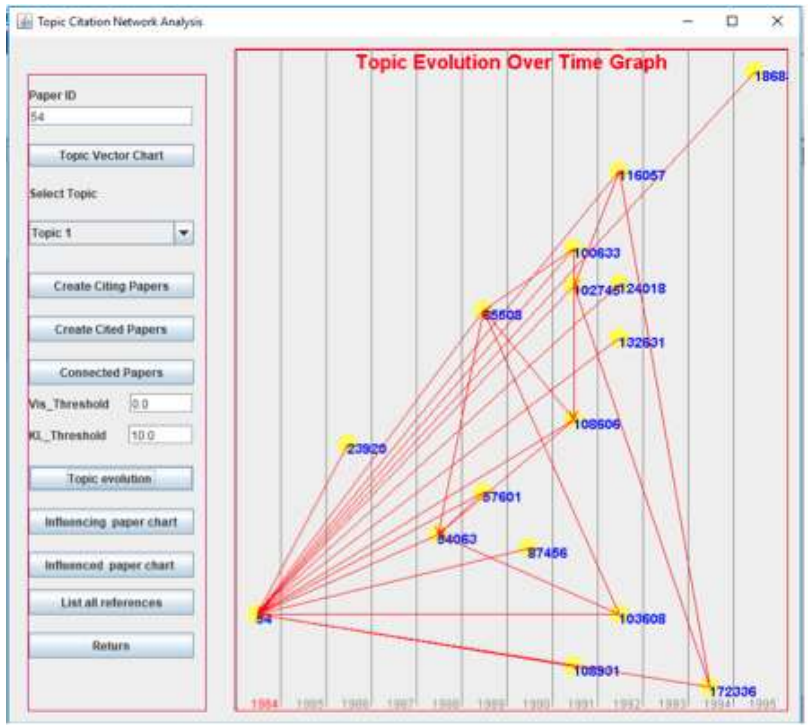
Hình 6. Sự phân bố chủ đề trong một bài báo khoa học

Trong hình 6, chúng ta có thể thấy nội dung bài báo có ID 54 nhắc đến 04 chủ đề nghiên cứu khác nhau, trong đó chủ đề số 2 và 10 có thể được xem như chủ đề chính của bài báo. Tuy nhiên như đã phân tích bên trên, nếu chỉ phân tích và đánh giá bài báo khoa học dựa vào nội dung được trình bày chúng ta có thể chưa thấy hết phạm vi nghiên cứu bởi vấn đề mà tác giả bài báo quan tâm còn được thể hiện thông qua danh sách các bài báo được tác giả trích dẫn trong công trình của mình. Do đó, hệ thống chúng tôi đã tích hợp việc phân tích chủ đề của một bài báo riêng lẻ vào mạng trích dẫn để người tham khảo có thể thấy rõ hết các vấn đề nghiên cứu mà bài báo muốn đặt ra cũng như nguồn gốc xuất phát của các chủ đề mà bài báo quan tâm (Xem hình 7).



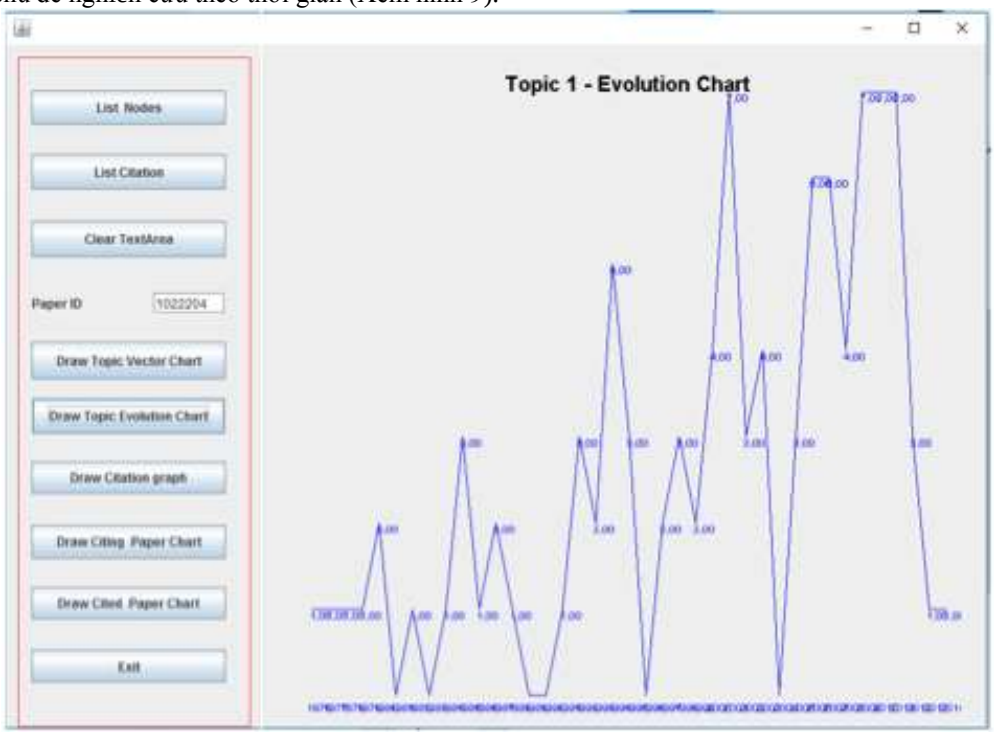
Hình 7. Sự lan tỏa và tác động của chủ đề trên mạng trích dẫn

Theo kế quả trong hình 7, chúng ta có thể nhận ra nội dung bài báo có ID 54 có thể liên quan đến 2 chủ đề thông qua mạng trích dẫn của bài báo. Trong đó chủ đề thứ nhất được tham khảo từ các bài báo có "màu xanh" và chủ đề thứ hai từ các bài báo có nút "màu vàng". Việc truy vấn nguồn gốc của các chủ đề mà bài báo quan tâm có thể được mở rộng ra trên cây trích dẫn của bài báo. Ngoài ra qua hệ thống chúng tôi có thể biết được sự lan truyền các chủ đề nghiên cứu từ một bài báo thông qua mạng trích dẫn (Xem hình 8).



Hình 8. Sự lan truyền của chủ đề trên cây trích dẫn

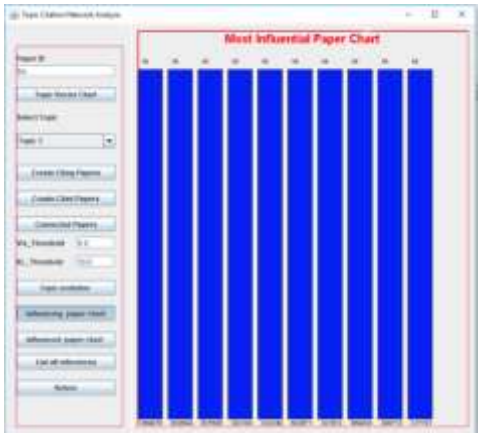
Qua kết quả của hệ thống, chúng ta có thể dễ dàng nhận ra rằng chủ đề "Topic 1" của bài báo có ID 54 được viết năm 1984 đã lan tỏa đến 15 bài báo từ năm 1984 đến năm1995. Trong đó, hai năm 1991 và 1992 có sự đột biến về số bài báo (8 bài báo) quan tâm đến chủ đề này. Bên cạnh đó, hệ thống của chúng tôi cũng cho phép theo dõi sự tiến hóa của các chủ đề nghiên cứu theo thời gian (Xem hình 9).



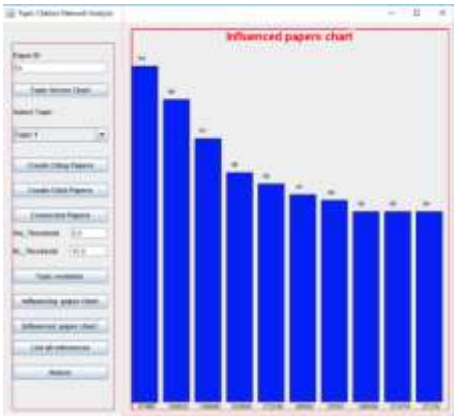
Hình 9. Đồ thị tiến hóa của chủ đề theo thời gian

Biểu đồ trong hình 9, cho chúng ta thấy được sự quan tâm của cộng đồng nghiên cứu đối với chủ đề "Topic 1" qua thời gian. Mỗi điểm trên biểu đồ chính là số bài báo trong năm có quan tâm đến "Topic 1". Biểu đồ đã thể hiện đúng thực tế nghiên cứu là chu kỳ phát triển của mỗi chủ đề nghiên cứu cũng có sự thăng trầm qua thời gian.

Hệ thống Topic Citation Evolution cũng giúp người nghiên cứu biết được danh sách và mức độ tác động của các bài báo trên mạng trích dẫn đến chủ đề nghiên cứu bài báo đang quan tâm đề cập cũng như sự tác động của các chủ đề này đến những bài báo khác trên mạng trích dẫn (Xem hình 10 và hình 11).



Hình 10. Danh sách bài báo đã tác động đến chủ đề nghiên cứu của bài báo đang phân tích



Hình 11. Danh sách bài báo bị tác động bởi bài báo đang phân tích

V. ĐÁNH GIÁ - KẾT LUẬN

Hệ thống chúng tôi đã xây dựng giúp người nghiên cứu triển khai giai đoạn nghiên cứu tổng quan rất thuận lợi. Với phần mềm này, các nhà nghiên cứu có thể dễ dàng rút trích danh sách các bài báo cần nghiên cứu dựa vào cây trích dẫn của một chủ đề quan tâm hoặc một bài báo đang quan tâm. So với hệ thống CiteNetExplorer [11], hệ thống chúng tôi bổ sung thêm một thông tin rất quan trọng trong quá trình khảo luận tổng quan nhờ vào việc kết hợp giữa phân tích nội dung bài báo cùng phân tích mạng trích dẫn. Với sự kết hợp này, hệ thống có thể cung cấp một cái nhìn tổng quan và đầy đủ về một hướng nghiên cứu qua đó giúp người nghiên cứu có thể dễ dàng nắm bắt và hiểu rõ xu hướng nghiên cứu hiện tại. Bên cạnh đó, hệ thống cũng đã đề xuất một hướng tiếp cận mới trong việc lưu trữ mạng trích dẫn sử dụng CSDL đồ thị thay thế cho cách lưu trữ truyền thống với CSDL quan hệ. Qua kết quả thực hiện, chúng tôi thấy rằng việc lưu trữ mạng trích dẫn sử dụng CSDL đồ thị giúp đơn giản hóa và nâng cao hiệu suất của hệ thống.

Hệ thống hiện tại vẫn còn một hạn chế là việc phân tích bài báo dựa trên phương pháp mô hình hóa chủ đề không thể gán nhãn tự động cho các chủ đề nghiên cứu. Do đó, trong tương lai, chúng tôi sẽ tích hợp thêm việc gán nhãn tự động các chủ đề nhằm giúp cho người nghiên cứu thuận lợi hơn trong quá trình sử dụng hệ thống.

LỜI CẢM ƠN

Tôi xin được bày tỏ lòng biết ơn đến PGS-TS. Đỗ Phúc và tập thể nhóm nghiên cứu đã tôi hoàn thành công trình này. Bài báo là kết quả của đề tài nghiên cứu được Trường Đại học Công nghệ Thông tin, ĐHQG-TP. HCM tài trợ kinh phí với mã số D2-2017-03.

TÀI LIỆU THAM KHẢO

[1] E. Garfield, “Citation indexes for science. A new dimension in documentation through association of ideas†,” *Int. J. Epidemiol.*, vol. 35, no. 5, pp. 1123–1127, Oct. 2006.

[2] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, “Arnetminer: extraction and mining of academic social networks,” in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 990–998.

[3] J. J. Miller, “Graph database applications and concepts with Neo4j,” in *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA*, 2013, vol. 2324, p. 36.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[5] E. Garfield, I. H. Sher, and R. J. Torpie, “The use of citation data in writing the history of science,” DTIC Document, 1964.

[6] E. Garfield, “Citation indexing for studying science,” *Nature*, vol. 227, no. 5259, pp. 669–671, 1970.

[7] S. Dawson, D. Gašević, G. Siemens, and S. Joksimovic, “Current State and Future Trends: A Citation Network Analysis of the Learning Analytics Field,” in *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, New York, NY, USA, 2014, pp. 231–240.

[8] Y. Cho and S. Park, “Using Citation Network Analysis,” *Educ. Technol.*, 2012.

[9] Y. Kajikawa and Y. Takeda, “Citation network analysis of organic LEDs,” *Technol. Forecast. Soc. Change*, vol. 76, no. 8, pp. 1115–1123, 2009.

[10] X. Shi, B. Tseng, and L. A. Adamic, "Information diffusion in computer science citation networks," *ArXiv Prepr. ArXiv09052636*, 2009.

[11] N. J. van Eck and L. Waltman, "CitNetExplorer: A new software tool for analyzing and visualizing citation networks," *J. Informetr.*, vol. 8, no. 4, pp. 802–823, Oct. 2014.

[12] I. Robinson, J. Webber, and E. Eifrem, *Graph databases: new opportunities for connected data*. O'Reilly Media, Inc., 2015.

[13] D. M. Blei and J. D. Lafferty, "Topic models," *Text Min. Classif. Clust. Appl.*, vol. 10, no. 71, p. 34, 2009.

[14] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[15] B. Gretarsson *et al.*, "Topicnets: Visual analysis of large text corpora with topic modeling," *ACM Trans. Intell. Syst. Technol. TIST*, vol. 3, no. 2, p. 23, 2012.

[16] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher, "Serendip: Topic model-driven visual exploration of text corpora," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, 2014, pp. 173–182.

[17] X. Ren *et al.*, "Cluscite: Effective citation recommendation by information network-based clustering," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 821–830.

[18] J. E. Gonzalez, R. S. Xin, A. Dave, D. Crankshaw, M. J. Franklin, and I. Stoica, "GraphX: Graph Processing in a Distributed Dataflow Framework.," in *OSDI*, 2014, vol. 14, pp. 599–613.

[19] D. Blei, L. Carin, and D. Dunson, "Probabilistic Topic Models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, Nov. 2010.

SCIENTIFIC TOPICS PROPOGATION ON CITATION NETWORK

Nguyen Trac Thuc, Pham The Anh Phu, Do Phuc

ABSTRACT: *In this paper, we introduce a system to discover scientific topics evolution and propagation in computer science. This paper proposes a new approach to store citation network in graph database in stead of relational database. Besides that, this paper also presents a method to fully analyse a scientific publication by integrating citation analysis methods with topic models to follow evolution and propagation of scientific topics in citation network.*