Phân tích tầm ảnh hưởng đối tượng theo chủ đề trong mạng Xã hội

- **Phan Hồ Viết Trường**
- **Hồ Trung Thành**
- **Đỗ Phúc**

Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

(Bài nhận ngày 04 tháng 09 năm 2013, hoàn chỉnh sửa chữa ngày 24 tháng 12 năm 2013)

TÓM TẮT:

Trong bài báo này tập trung nghiên cứu các mô hình trong phân tích mạng xã hội và ứng dụng mô hình vào phát triển hệ thống cho phép người dùng tìm kiếm chính xác các chủ đề, công trình nghiên cứu cần thiết có liên quan của các chuyên gia và tác giả. Trong đó, chúng tôi nghiên cứu mô hình Tác giả - Hội Nghị - Chủ đề (Author Conference Topic - ACT) nhằm tìm ra những chủ đề có liên quan đến yêu cầu của người dùng. Ứng với mỗi chủ đề tìm được, hệ thống sẽ xây dựng một đồ thị lan truyền quan hệ theo chủ đề (Topical Affinity Propagation - TAP) nhằm phân tích tầm ảnh hưởng theo chủ đề giữa các tác giả. Kết quả được thử nghiệm trên một phần tập ngữ liệu lấy từ Microsoft Academic Research bao gồm 34.330 tác giả, 19.921 bài báo, 1,335 hội nghị, 424.501 mối quan hệ đồng tác giả.

Từ khóa: *lan truyền ái lực, mạng xã hội, mô hình ACT, phân tích tầm ảnh hưởng*

1.Giới thiệu

Tốc độ phát triển Internet đã giúp con người tiếp cận thông tin trên thế giới một cách nhanh chóng. Tuy nhiên, với khối lượng thông tin khổng lồ và không ngừng phát triển, đã xuất hiện nhiều công cụ hỗ trợ tìm kiếm hữu hiệu hiện nay như Google, Bing, Yahoo!Search . . .Các công cụ này phần nào đáp ứng được yêu cầu của người dùng nhưng kết quả trả về đôi khi không phù hợp với mong muốn. Ví dụ, khi gõ một từ khóa "Data Mining" không chỉ mong muốn tìm những tài liệu có chứa từ này nếu còn muốn tìm các chủ đề có liên quan trong cùng lĩnh vực. Vấn đề tiếp theo, sau khi đã tìm được các chủ đề liên quan, làm sao xác định được chuyên gia theo từng chủ đề. Nhằm mục đích xây dựng công cụ hỗ trợ học tập, nghiên cứu các công trình khoa học của những chuyên gia hàng đầu, nghiên cứu này sử dụng mô hình ACT[5] để khám phá các chủ đề và mô hình TAP[6] để phân tích tầm ảnh hưởng từng tác giả nhằm xác định chuyên gia theo từng chủ đề. Phần còn lại bài báo được tổ chức như sau: Phần 2: Mô hình Tác giả - Hội nghị - Chủ đề. Phần 3: Mô hình đồ thị lan truyền quan hệ theo chủ đề. Phần 4: kết quả thử nghiệm và thảo luận. Phần 5: kết luận và hướng phát triển.

2. Mô hình Tác giả-Hội nghị-Chủ đề

Xuất phát từ mô hình Latent Dirichlet Allocation do Blei, Ng và Jordan[3] đề xuất năm 2003 theo cách tiếp cận mạng Bayes, mô hình LDA dùng để mô hình hóa kho ngữ liệu nhằm phát hiện ra các chủ đề ẩn của kho ngữ liệu. Quá

trình phát sinh của một tập tài liệu gồm ba bước : (i) với mỗi tài liệu có một phân bố xác suất chủ đề của tài liệu đó, phân bố này được lấy mẫu từ phân bố xác suất Dirichlet, (ii) với mỗi từ trong tài liệu, một chủ đề duy nhất được chọn từ phân bố chủ đề trên, (iii) mỗi từ khóa sẽ được rút ra từ phân bố đa thức cho từ khóa theo chủ đề vừa chọn. Đã có rất nhiều mô hình được phát triển từ mô hình LDA như Author-Topic[7],[8] , Author-Recipient-Topic[1], riêng mô hình Tác giả-Hội nghị-Chủ đề (Author-Conference-Topic)[5] là mô hình cải tiến từ mô hình Author-Topic, ngoài việc một chủ đề phát sinh tập từ đặc trưng, nó còn cho phép một chủ đề phát sinh các hội nghị có liên quan. Một số khái niệm liên quan:

a_d : tập tác giả viết tài liệu d

θ_x : phân bố xác suất chủ đề tác giả x quan tâm, phân bố này thỏa phân bố Dirichlet.

ϕ_z : phân bố xác suất từ khóa được phát sinh bởi một chủ đề z, phân bố này thỏa phân bố Dirichlet.

w_{di} : tập từ khóa được chọn từ phân bố đa thức Multinomial(ϕ_z)

α : tham số tập trung cho phân bố xác suất chủ đề

β : tham số tập trung cho phân bố xác suất từ đặc trưng

c_{di} : tập hội nghị được chọn từ phân bố đa thức Multinomial(ϕ_z)

μ : tham số tập trung cho phân bố xác suất hội nghị

ψ_z : tập conference (hội nghị, bài báo) thuộc về chủ đề z

w_{di} : từ khóa thứ i được phát sinh bởi chủ đề z_{di} được chọn từ phân bố đa thức(ϕ_z)

z_{di} : chủ đề z gán cho từ khóa thứ i được phát sinh bởi tác giả x_{di} chọn phân bố đa thức(θ_x)

x_{di} : tác giả được chọn cho mỗi từ thứ i trong tài liệu d được phát sinh bởi phân bố đồng nhất(a_d)

Quá trình tạo sinh của mô hình ACT như sau:

Với mỗi chủ đề z, sẽ tính được phân bố xác suất cho các từ khóa ϕ_z và các hội nghị ψ_z tương ứng theo phân bố xác suất Dirichlet(β) và phân bố xác suất Dirichlet(μ)

Với mỗi từ khóa w_{di} trong bài báo d:

Chọn một tác giả x_{di} từ tập phân bố tác giả a_d, phân bố này thỏa phân bố đồng nhất(a_d).

Chọn một chủ đề z_{di} từ phân bố chủ đề của tác giả x_{di}, phân bố này thỏa phân bố đa thức ($\theta_{x_{di}}$). Với θ là phân bố xác suất chủ đề Dirichlet(α).

Chọn một từ khóa w_{di} từ phân bố từ khóa của chủ đề z_{di}, phân bố này thỏa phân bố đa thức ($\phi_{z_{di}}$).

Chọn một hội nghị c_{di} từ phân bố hội nghị của chủ đề z_{di}, phân bố này thỏa phân bố đa thức($\psi_{z_{di}}$).

Tương tự như mô hình LDA, mô hình ACT cần ước lượng 2 tham số là (1) phân bố θ là ma trận tác giả-chủ đề (Author-Topics), phân bố ϕ là ma trận chủ đề-từ khóa (Topic-Words), phân bố ψ là ma trận chủ đề-hội nghị (Topic-Conferences); tham số (2) là xác suất chủ đề z_{di} và tác giả x_{di} gán cho từ w_{di}. Tiếp theo sẽ sử dụng lấy mẫu Gibbs[11] để ước lượng tham số cho 2 biến x và z như sau:

$$P(z_{di}, x_{di} \mid z_{-di}, x_{-di}, w, c, \alpha, \beta, \mu) \propto \frac{m^{-di}_{x_{di}z_{di}} + \alpha}{\sum_z (m^{-di}_{x_{di}z} + \alpha)} \frac{n^{-di}_{z_{di}w_{di}} + \beta}{\sum_v (n^{-di}_{z_{di}v} + \beta)} \frac{n^{-di}_{z_{di}c_{di}} + \mu}{\sum_c (n^{-di}_{z_{di}c} + \mu)} \quad (1)$$

Trong quá trình thực hiện sẽ tính được 3 ma trận A x T (tác giả – chủ đề), T x V (chủ đề – từ khóa) và T x C (chủ đề – hội nghị). Từ 3 ma trận này tính được xác suất phát sinh một chủ đề z bởi một tác giả x là θ_{xz}, xác suất phát sinh một từ v bởi một chủ đề z là ϕ_{zv} và xác suất một hội nghị c được phát sinh bởi một chủ đề z là ψ_{zc} theo các công thức Gibbs[11] tương ứng sau:

$$\theta_{xz} = \frac{m_{xz} + \alpha}{\sum_{z'}(m_{xz'} + \alpha)} \quad (1)$$

$$\phi_{zv} = \frac{n_{zv} + \beta}{\sum_{v'}(n_{zv'} + \beta)} \quad (2)$$

$$\psi_{zc} = \frac{n_{zc} + \mu}{\sum_{c'}(n_{zc'} + \mu)} \quad (3)$$

Với m_{xz} là số lần tác giả x phát sinh chủ đề z, $m_{xz'}$ là số lần tác giả x phát sinh chủ đề khác z ($z \neq z'$). Và n_{zv} là số lần chủ đề z phát sinh từ đặc trưng v, $n_{zv'}$ là số lần chủ đề z phát sinh từ đặc trưng khác v ($v \neq v'$). Cuối cùng, n_{zc} là số lần chủ đề z phát sinh hội nghị c và $n_{zc'}$ là số lần chủ đề z phát sinh hội nghị khác c ($c \neq c'$).

3. Mô hình Lan truyền Ái lực theo Chủ đề

Mô hình lan truyền ái lực theo chủ đề (Topical Affinity Propagation) do Jie Tang, Jimeng, Chi Wang và Zi Yang[6] đề xuất năm 2009, sử dụng mô hình đồ thị nhân tố (factor graph)[4] kết hợp với mô hình lan truyền ái lực (affinity propagation)[2] để phân tích tầm ảnh hưởng của từng nút trong đồ thị. Tuy nhiên, khác với 2 giải thuật đã được sử dụng để phân tích ảnh hưởng từng nút trong đồ thị là giải thuật Vector riêng và PageRank, mô hình TAP (Topical Affinity Propagation) [6] có thể cho biết mức độ ảnh hưởng cụ thể giữa các nút. Để phân tích được tầm ảnh hưởng trong mạng đồng tác giả, thông tin quan trọng sau phục vụ cho việc nghiên cứu là cần thiết.

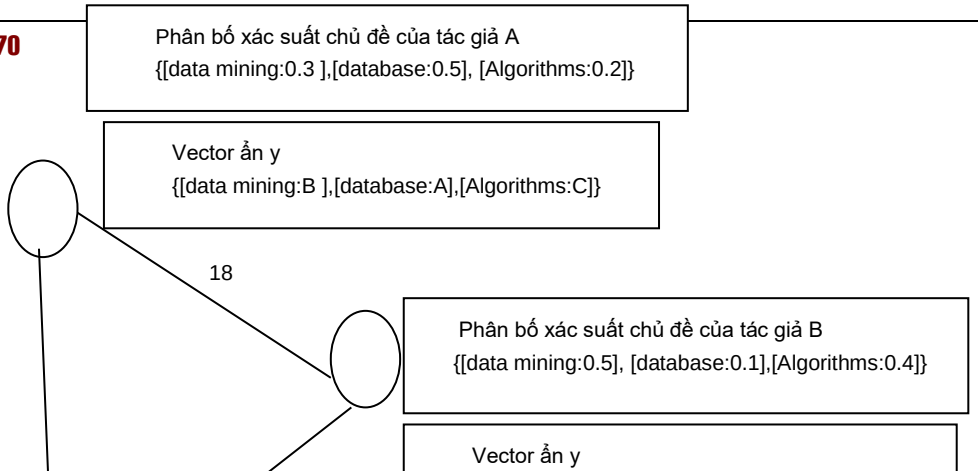
Một tập mối quan hệ giữa các tác giả với nhau trong mạng

Số bài báo mà tác giả viết cho chủ đề họ quan tâm.

Phân bố xác suất theo chủ đề của một tác giả x. Ví dụ: tác giả "Jiawen Han" quan tâm đến 3 chủ đề có phân bố xác suất là{"Data Mining"=> 0.642857, "Database"=> 0.285714, "Information Retrieval"=> 0.071429}

Ảnh hưởng theo chủ đề : tầm ảnh hưởng của nút *s* tới nút *t* (ký hiệu μ_{st}) gồm hai biến *r(t,s)* và *a(t,s)*.

Việc xây dựng đồ thị G=(E, V) với V là tập nút (các tác giả), E là tập cạnh biểu diễn liên kết giữa các tác giả, mỗi cạnh có trọng số tương ứng số bài báo mà 2 tác giả cùng tham gia. Một tập phân bố xác suất chủ đề cho từng nút tham gia trong G. Từ các dữ liệu này, hình thức hóa việc phân tích tầm ảnh hưởng vào trong đồ thị các nhân tố theo chủ đề (Topical Factor Graph)[6] như sau: Cho một tập các nút cần phân tích $\{v_i\}_{i=1}^{N}$, với N là số nút trên mạng và một tập vector ẩn $\{y_i\}_{i=1}^{N}$ dùng để biểu diễn các nút có xác suất quan tâm đến các chủ đề cao nhất ảnh hưởng đến nút v_i và y_i^z là nút có ảnh hưởng cao nhất đến chủ đề z lên nút v_i. Ví dụ trong hình 1, ở nút A với vector ẩn cho biết nút B quan tâm nhiều nhất đến chủ đề "Data Mining" có xác suất cao nhất. Một cách trực quan, khi quan tâm đến chủ đề nào đó, thường sẽ chọn những tác giả có nhiều đóng góp nhất cho chủ đề đó để tham khảo, và những tác giả này sẽ có tầm ảnh hưởng liên quan. Một tác giả càng được nhiều người khác tham khảo đến (tin cậy) sẽ có tầm ảnh hưởng càng cao. Để thực hiện việc phân tích tầm ảnh hưởng mỗi tác giả trên đồ thị nhân tố, kết hợp sử dụng giải thuật TAP do Jie Tang và cộng sự [6] mở rộng từ giải thuật AP (Affinity Propagation) [2] và sử dụng các công thức trong [6] để giải quyết vấn đề tìm điểm số ảnh hưởng giữa các nút trong đồ thị.



Hàm đặc tính cục bộ của mỗi nút[6]: hay còn gọi là ảnh hưởng của mỗi tác giả theo từng chủ đề được tính theo công thức sau:

$$g(v_i, y_i, z) = \begin{cases} \dfrac{w^z_{iy^z_i}}{\sum_{j \in NB(i)} (w^z_{ij} + w^z_{ji})} & y^z_i \neq i \\ \dfrac{\sum_{j \in NB(i)} w^z_{ji}}{\sum_{j \in NB(i)} (w^z_{ij} + w^z_{ji})} & y^z_i = i \end{cases} \quad (4)$$

Trong đó v_i là nút đang xét theo chủ đề z, y_i là nút có xác suất ảnh hưởng cao nhất ứng với chủ đề z đối với nút v_i. $w^z_{ij} = y^z_j.\alpha_{ij}$ với y^z_j là xác suất ảnh hưởng cao nhất trong chủ đề z lên nút v_i và α_{ij} là trọng số tương đương số bài báo giữa 2 nút v_i và v_j (chung của 2 tác giả). Còn $j \in NB(i)$ là những nút kề với nút v_i. Trường hợp nút v_i đang xét cũng là nút có ảnh hưởng cao nhất trong chủ đề z ($y^z_i = i$) thì ta tính ảnh hưởng của v_i đến các nút kề với nó $v_j \in NB(i)$, ngược lại ($y^z_i \neq i$) và chỉ cần tính ảnh hưởng của nút y^z_j trực tiếp lên bản thân của v_i. Hàm đặc tính cục bộ của mỗi nút trong công thức (5) chỉ mới tính tầm ảnh hưởng của đối tượng đối với chủ đề. Tiếp theo ta sẽ tính điểm số ảnh hưởng giữa các nút trong đồ thị. Trong mạng đồng tác giả, r^z_{ij} là thông tin trao đổi từ nút i tới nút j theo chủ đề z, a^z_{ij} là thông tin gởi từ nút *j* tới nút *i* được khởi tạo bằng 0. Nhưng trước tiên cần có hàm đo mức độ ảnh hưởng của từng cặp tác giả trong mỗi chủ đề:

$$b_{ij}^{z} = \log \frac{g(v_i, y_i, z) \mid y_i^z = j}{\sum_{k \in NB(i) \bigcup \{i\}} g(v_i, y_i, z) \mid y_i^z = k} \quad (6)$$

Công thức (6) chuẩn hóa mức độ ảnh hưởng của từng cặp tác giả trong mỗi chủ đề, hàm log dùng để chuyển từ cách tính tổng-tích (Sum-Product) sang tính Max-Sum do giải thuật Sum-Product khi áp dụng trên đồ thị nhân tố, giá trị các hàm sẽ quá nhỏ (số lẻ thập phân gồm nhiều số 0)[10]. Giải thuật TAP[6] mở rộng giải thuật lan truyền theo quan hệ[2] có đầu vào gồm ba tập biến, b_{ij}^{z} là mức độ ảnh hưởng của nút *i* và nút *j* trong chủ đề *z*, biến r_{ij}^{z} là mức độ nút *i* chịu ảnh hưởng bởi nút *j*, biến a_{ij}^{z} là mức độ ảnh hưởng của nút *j* lên nút *i*. Hàm tính thông tin trao đổi từ nút i đến nút j[6]

$$r_{ij}^{z} = b_{ij}^{z} - \max_{k \in NB(j)} \{b_{ik}^{z} + a_{ik}^{z}\} \quad (7)$$

Công thức (7) mô tả nút *i* gởi thông tin đến nút *j* nhằm xác định mức độ chịu ảnh hưởng của nút *i* bởi nút *j* bằng hiệu mức độ ảnh hưởng của nút *i* và nút *j* trong chủ đề *z*, b_{ij}^{z}, với giá trị lớn nhất ảnh hưởng của các nút *k* kề với nút *j* mà nút *i* cũng bị ảnh hưởng bởi theo Frey và Dueck[2] nếu giá trị mức độ chịu ảnh hưởng của nút *i* bởi nút *j* là số dương thì chứng tỏ nút *j* có ảnh hưởng lên nút *i*. Biến a_{ik}^{z} gởi thông tin từ nút *k* (kề với nút j) đến nút i để phản ánh mức độ ảnh hưởng của nút *k* lên nút *i*, ban đầu khởi động bằng 0. Hàm tính thông tin trao đổi từ nút j đến nút i [6]

$$a_{ij}^{z} = \min(\max\{r_{jj}^{z}, 0\}, -\min\{r_{jj}^{z}, 0\} \quad - \max_{k \in NB(j) \setminus \{i\}} \min\{r_{kj}^{z}, 0\}), i \in NB(j) \quad (8)$$

Công thức (8) mô tả nút *j* gởi thông tin đến nút *i* nhằm xác định mức độ ảnh hưởng của nút *j* lên nút *i*. Biến r_{jj}^{z} cho biết mức độ ảnh hưởng bản thân nút *j* lên chủ đề *z*. Do giải thuật lan truyền theo quan hệ là giải thuật gom cụm nên Frey và Dueck[2] cho rằng nếu giá trị của mỗi điểm dữ liệu quá lớn thì bản thân mỗi điểm có khả năng trở thành một cụm riêng, ngược lại nếu giá trị mỗi điểm dữ liệu quá nhỏ thì tất cả các điểm này có khuynh hướng gom về cùng một cụm. Cũng theo Frey và Dueck[2] thì giá trị tốt nhất của nút j ảnh hưởng lên nút *i* nằm trong khoảng 2 giá trị nhỏ nhất và lớn nhất của nút *j*, trong đó giá trị lớn nhất là mức độ ảnh hưởng bản thân nút j đối với chủ đề z xác định bằng hàm $\max\{r_{jj}^{z}, 0\}$, còn giá trị nhỏ nhất là tổng giá trị nhỏ nhất của mức độ ảnh hưởng bản thân của j với giá trị lớn nhất trong các các nút k kề với j chịu sự ảnh hưởng của nút j được xác định bằng $-\min\{r_{jj}^{z}, 0\} - \max_{k \in NB(j) \setminus i} \min\{r_{kj}^{z}, 0\}$. Do đang xét mức độ ảnh hưởng của j lên nút i nên chỉ chọn giá trị ảnh hưởng của những nút k kề với j mang giá trị âm bằng hàm $\min_{k \in NB(j) \setminus i} \{r_{kj}^{z}, 0\}$. Hàm min của cả công thức nhằm tránh trường hợp giá trị a_{ij}^{z} quá lớn dẫn đến mỗi nút trở thành một cụm riêng biệt[2]. Hàm tính điểm số ảnh hưởng [6]

$$\mu_{st}^{z} = \frac{1}{1 + e^{-(r_{ts}^{z} + a_{ts}^{z})}} \quad (9)$$

Hàm sigmoid này chuyển giá trị của hai biến r_{ts}^{z}, a_{ts}^{z} thành xác suất.

4.Kết quả thử nghiệm

4.1. Thu thập Dữ liệu

Việc đầu tiên là phải có thông tin các tác giả liên quan lĩnh vực công nghệ thông tin. Hiện nay, hầu như các nhà khoa học đều đang công tác tại các trường đại học hàng đầu thế giới, mỗi tác giả đều công bố thông tin cá nhân trên Web. Có nhiều trang Web đã tập hợp thông tin tác giả cũng như các bài báo khoa học như trang DBLP chuyên cung cấp thông tin về bài báo, hay như trang CiteseerX là một hệ thống tìm kiếm theo tài liệu và tác giả. Hiện nay, Microsoft đã đưa ra trang Web Academic Search cung cấp khá đầy đủ thông tin về tác giả, chủ đề, sở thích, hội nghị, bài báo khoa học . . . Vì vậy, chúng tôi sẽ chọn

trang Web này để thu thập dữ liệu hình thành nên mạng đồng tác giả.

Dữ liệu được lựa chọn, trích lọc có 34.330 tác giả, 19.921 bài báo, 1,335 hội nghị, 424.501 mối quan hệ đồng tác gia (được lấy từ trang web Microsoft Academic Research do Microsoft cung cấp theo địa chỉ trang web http://academic.research.microsoft.com/?Search Domain=2 và sau đó tinh chỉnh cho phù hợp với hệ thống chương trình). Từ các dữ liệu này, chúng tôi sẽ xây dựng hệ thống tạm đặt tên là hệ thống học thuật công nghệ thông tin nhằm phục vụ tìm kiếm chủ đề, chuyên gia.

4.2. Tìm những Chủ đề Liên quan đến Từ khóa

Chúng tôi sẽ sử dụng mô hình ACT[5] do phù hợp với hệ thống được xây dựng, để mô hình tất cả tài liệu về công nghệ thông tin. Sau khi áp dụng mô hình ACT, ta sẽ có được ma trận T x V cho biết xác suất một từ khóa v $\in$ V được phát sinh bởi chủ đề t $\in$ T. Từ ma trận trên dễ dàng xác định chủ đề liên quan đến từ khóa cần tìm. Cách làm như sau : Gọi từ khóa nhập vào là q (ở đây q là từ kép trở lên), đầu tiên loại bỏ những từ thường xuyên xảy ra, thường gọi là stopword. Sau đó sẽ ngắt q thành từng từ đơn $\{q_1, q_2, \ldots, q_n\}$ để thực hiện tìm kiếm. Kết quả sẽ xảy ra trường hợp một chủ đề có nhiều từ khóa xuất hiện phù hợp với q. Từ đây, ta chỉ việc cộng xác suất của mỗi từ khóa theo từng chủ đề và sắp xếp theo thứ tự giảm dần. Cuối cùng, ta chọn 5 chủ đề có xác suất cao nhất trả kết quả về cho người dùng.

4.3. Xác định Chuyên gia theo Chủ đề

Việc xác định chuyên gia theo từng chủ đề chính là việc phân tích tầm ảnh hưởng của tác giả a đối với các tác giả còn lại có quan hệ với a. Trong nghiên cứu này sẽ sử dụng mô hình TAP[6] do có thời gian chạy và độ chính xác cao hơn giải thuật Vector riêng và PageRank. Khác với đồ thị truyền thống chỉ bao gồm các nút, cạnh và trọng số, mô hình này là sự kết hợp đồ thị truyền thống với xác suất thống kê. Để tính được mức độ ảnh hưởng của tác a đối với tác giả b theo chủ đề z, μ_{ab}^{z}, TAP sẽ lần lượt xét mức độ ảnh hưởng với các nút kề với nút b, sau đó , một lần nữa, TAP lại lần lượt xét mức độ ảnh hưởng các nút kề với nút a để xác định tầm ảnh hưởng của nút b đối với a, μ_{ba}^{z}. Ngoài ra, TAP còn phải dựa thêm vào việc tính tầm ảnh hưởng của tác giả có xác suất quan tâm đến chủ đề đang xét là cao nhất. Sau khi đã lần lượt tính tầm ảnh hưởng giữa các cặp nút trong đồ thị, việc tiếp theo chỉ việc cộng xác suất theo từng nút để cho ra tầm ảnh hưởng của nó theo chủ đề.

4.4. Kiến trúc Hệ thống

Hệ thống gồm các thành phần sau:

Thành phần lưu trữ : Đây là thành phần lưu trữ dữ liệu mạng xã hội đồng tác giả, bao gồm hồ sơ tác giả, các bài báo khoa học, các hội nghị dựa theo tiếp cận mô hình ACT.

Thành phần tìm kiếm chủ đề: Thành phần này có nhiệm vụ tiếp nhận yêu cầu từ người dùng, sau đó hệ thống sẽ tìm những chủ đề có liên quan đến từ khóa mà người dùng đã cung cấp. Ví dụ : người dùng nhập từ khóa “Data Mining” thì hệ thống sẽ trả về những chủ đề liên quan như “Database”, “Information Retrieval”. Đối với thành phần này dùng mô hình ACT để giải quyết. Từ những chủ đề kết quả, người dùng có thể chọn một chủ đề để biết tác giả nào có tầm ảnh hưởng nhất.

Thành phần phân tích tầm ảnh hưởng: Thành phần này thực hiện việc phân tích dựa theo từng chủ đề, bằng cách áp dụng mô hình đồ thị kết hợp xác suất thống kê và thuật toán TAP, để đánh giá mức độ ảnh hưởng giữa các tác giả cũng như tìm được tác giả ảnh hưởng nhất trong chủ đề. Cuối cùng, thành phần cho biết tập tác giả trung tâm có mối quan hệ với nhiều tác giả nhất.

4.5. Kết quả Thử nghiệm và Đánh giá

Áp dụng mô hình ACT để thử nghiệm trên hệ thống mạng đồng tác giả gồm 34.330 tác giả, 19.921 bài báo, 1.335 hội nghị và 5.258 từ vựng cho lĩnh vực công nghệ thông tin để phát sinh 24 chủ đề và được gán nhãn bằng tay với kết quả như sau:

Bảng 1. Kết quả thực hiện mô hình ACT phát sinh vài chủ đề

Chủ đề 6 "Data Mining"		**Chủ đề 7 "Database"**	
data mining	0.0784	primary key	0.0184
association rule	0.0781	raw data	0.0175
machine learning	0.0775	reference data	0.0171
information retrieval	0.0679	remote backup	0.0163
time series	0.0648	roll back	0.0154
social network	0.0529	query processing	0.0151
high dimensionality	0.0525	structured query	0.0149
text categorization	0.0503	stored procedure	0.0149
knowledge discovery	0.0472	cartesian product	0.0149
feature selection	0.0408	decision support	0.0142
KDD	0.4352	SIGMOD	0.2634
TKDE 0.1829		ADBT 0.1465	
CIKM	0.1638	FODO 0.1155	
IDA	0.1093	TKDE 0.0933	
WWW	0.0811	CIKM	0.0914
Chủ đề 12 "Information Retrieval"		**Chủ đề 24 "Word Wide Web"**	
search engine	0.0849	sematic web	0.1156
text categorization	0.0782	xml document	0.1073
social network	0.0744	query answering	0.0928
feature selection	0.0689	client server	0.0801
decision tree	0.0647	document frequency	0.0642
naive bayes	0.0613	social networks	0.0613
pattern recognition	0.0594	answering queries	0.0518
text classification	0.0589	ontology language	0.0488
clustering method	0.0517	web ontology	0.0410
machine learning	0.0493	markov model	0.0400
SIGIR	0.2667	WWW0.6759	
TOIS	0.2098	ISWC 0.4114	
CIKM	0.1946	Hypertext	0.0492
ISCI	0.0374	ISEC 0.0038	
IPL	0.0300	WISE 0.0009	

Chúng tôi đã cài đặt ba giải thuật Vector riêng, PageRank và TAP trên chủ đề "Data Mining" gồm 2.825 tác giả và 40.110 mối quan hệ trên cùng máy tính có cấu hình CPU core i3 dual 2.4GHz, RAM 2GB. Việc đánh giá độ chính xác của 3 giải thuật này sẽ được thực hiện như sau: chọn danh sách 10 tác giả hàng đầu trong lĩnh vực "Data Mining" lấy từ trang Web Microsoft Academic Research làm tiêu chuẩn đánh giá độ chính xác các thuật giải phân tích. Ứng với mỗi kết quả từ 3 giải thuật sẽ tiếp tục đếm những tác giả nào có xuất hiện trong mẫu, sau đó chia cho tổng số tác giả.

Bảng 2. Kết quả thực hiện giải thuật vector riêng

Eigenvector	Microsoft Academic Search
Jiawei Han	**Jiawei Han**
Phillip S. Yu	**Philip S. Yu**
Wei Wang	Rakesh Agrawal
Vipin Kumar	Christos Faloutsos
Michael Stonebraker	Hans-Peter Kriegel
Jian Pei	Eamonn J. Keogh
Ke Wang	Geogre Karypis
Joydeep Ghosh	Andrew McCallum
H. V. Jagadish	Heikki Mannila
Ugur Cetintemel	**Jian Pei**

Bảng 3. Kết quả thực hiện giải thuật PageRank

PageRank	Microsoft Academic Search
Jiawei Han	**Jiawei Han**
Hans-Peter Kriegel	**Philip S. Yu**
Heikki Mannila	Rakesh Agrawal
Hannu T. T. Toivonen	Christos Faloutsos
Jian Pei	**Hans-Peter Kriegel**
Johannes Gehrke	Eamonn J. Keogh
Vipin Kumar	Geogre Karypis
Philip S. Yu	Andrew McCallum
Alexander Tuzhilin	**Heikki Mannila**
Guozhu Dong	**Jian Pei**

Bảng 4. Kết quả thực hiện giải thuật TAP

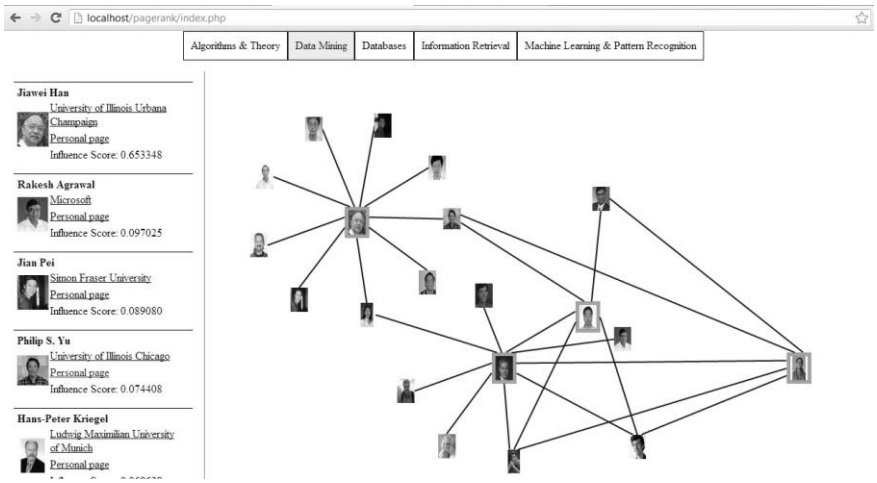
TAP	Microsoft Academic Search
Jiawei Han	**Jiawei Han**
Jian Pei	**Philip S. Yu**
Christos Faloutsos	**Rakesh Agrawal**
Rakesh Agrawal	**Christos Faloutsos**
Philip S. Yu	**Hans-Peter Kriegel**
H. V. Jagadish	Eamonn J. Keogh
Hans-Peter Kriegel	Geogre Karypis
Johannes Gehrke	Andrew McCallum
Heikki Mannila	**Heikki Mannila**
Raghu Ramakrishnan	**Jian Pei**

Bảng 5. Tổng hợp đánh giá độ chính xác

Giải Thuật	Tác giả	Thời gian (giây)	Chính xác (%)
Eigenvector	2.825	2564,94	30
PageRank	2.825	2478,19	50
TAP	2.285	149,20	70

Từ bảng 5 cho thấy rằng giải thuật TAP có thời gian chạy nhanh hơn và độ chính xác cao hơn 2 giải thuật còn lại. Do đó, chúng tôi chọn giải thuật TAP cho hệ thống chương trình.

4.6. Chương trình Thử nghiệm



Hình.2. Minh họa hoạt động của hệ thống khi người dùng gõ từ khóa "Data Mining", hệ thống trả về 5 chủ đề có liên quan. Sau đó, người dùng sẽ chọn một chủ đề để tìm chuyên gia trong chủ đề đó. Trong ví dụ này là danh sách các chuyên gia theo chủ đề "Data Mining" được sắp xếp theo điểm số ảnh hưởng.

5.Kết luận và hướng phát triển

Trong bài báo này đã trình bày hai vấn đề chính để xây dựng một hệ thống hỗ trợ học thuật nghiên cứu thônng qua các công trình khoa học của các tác giả hàng đầu trong lĩnh vực công nghệ thông tin. Vấn đề thứ nhất là tìm các chủ đề liên quan đến từ khóa mà người dùng cung cấp bằng việc sử dụng mô hình Tác giả-Hội nghị-Chủ đề (Author-Conferences-Topic), vấn đề thứ hai là tìm chuyên gia trong từng chủ đề bằng việc sử dụng mô hình lan truyền quan hệ theo chủ đề (Topical Affinity Propagation) để phân tích tầm ảnh hưởng của từng tác giả đối với chủ đề mà họ quan tâm. Hệ thống tìm chủ đề và phân tích ảnh hưởng đối tượng là một ví dụ minh họa cho việc sử dụng mạng xã hội. Tuy nhiên, hệ thống này còn phải cải tiến nhiều vấn đề như sau: (1) Tìm hiểu các giải thuật song song, phân tán để có thể xử lý dữ liệu ngày càng lớn của mạng xã hội đồng tác giả; (2) Mô hình phân tích chỉ áp dụng cho mạng đồng nhất, vô hướng và có trọng số. Giữa hai nút chưa phân biệt được vai trò của chúng (chẳng hạn một bài báo gồm có người hướng dẫn, sinh viên nghiên cứu hay hai tác giả cùng nghiên cứu . . .)

Topic based object Influence analysis in Social networks

- **Phan Ho Viet Truong**
- **Ho Trung Thành**
- **Do Phuc**

University of Information Technology, VNU-HCM

ABSTRACT:

In this article, we focus on researching models in social networks analysis and use them to develop the system which enables users to exactly find topics as well as necessary research of authors. In this research, we study Author – Conference – Topic (ACT) model to find out the topic relevant to users' requirements. With each of these topics, the system will create a graph of Topical Affinity Propagation (TAP) in purpose of analyzing the topic based influence of authors. Experimental results on the corpus extracted from Microsoft Academic Research include 34,330 authors, 19,921 articles, 1,335 conference and 424,501 co-author relations.

Keyword: *ACT model, influent extent analysis, social networks,TAP model*

TÀI LIỆU THAM KHẢO

[1]. Andrew McCallum, Andrés Corrada-Emmanuel, Xuerui Wang, *The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email*, Technical Report UM-CS-2004-096 (2004)

[2]. Brendan J.frey, Delbert Dueck, *Clustering by Passing Messages Between Data Points*, SCIENCE, vol 315, pp. 972-976 (2007)

[3]. David M.Blei, Andrew Y.Ng, Micheal I.Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, pp. 993 – 1022 (2003)

[4]. Frank R. Kschischang, Brendan J. Frey, Hans-Andrea Loelige, *Factor graphs and the sum-product algorithm*. IEEE Transactions on Information Theory (TIT) 47(2), pp. 498-519 (2001)

[5]. Jie Tang, Ruoming Jin, Jing Zhang, *A Topic Modeling Approach and its Integration into the Random Walk Framework for Academic Search ,ICDM 2008*, pp. 1055-1060 (2008)

[6]. Jie Tang, Jimeng Sun, Chi Wang, Zi Yang, *Social Network Analysis in large-scale networks*, KDD 2009, pp. 807 – 816 (2009)

[7]. Michal Rosen-Zvi, Thomas L.Griffiths, Mark Steyvers, Padhraic Smyth, *The Author-Topic Model for Authors and Documents*, *UAI 2004*, pp. 487 – 494 (2004)

[8]. Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas L.Griffiths, Mark Steyvers, Padhraic Smyt , *Learning Author-Topic Models from Text Corpora*, ACM Transaction on Information Systems, 28(1) (2010)

[9]. Thomas P. Minka, *Estimating a Dirichlet Ditribution* (2000)

[10]. William M. Darling, *A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling* (2011)

[11]. Xiaojin Zhu, *Inference in Graphical Models*, CS769 Spring 2010 Advanced Natural Language Processing (2010)