Mô hình tích hợp khám phá, phân lớp và gán nhãn chủ đề tiếp cận theo mô hình chủ đề

- **Hồ Trung Thành**[a]
- **Đỗ Phúc**[b]

[a] Trường Đại học Kinh tế - Luật, ĐHQG-HCM, thanhht@uel.edu.vn

[b] Trường Đại Học Công Nghệ Thông Tin, ĐHQG-HCM, phucdo@uit.edu.vn

(Bài nhận ngày 03 tháng 09 năm 2014, hoàn chỉnh sửa chữa ngày 24 tháng 11 năm 2014)

TÓM TẮT

Trong bài báo này, chúng tôi đề xuất một mô hình tích hợp khám phá chủ đề, phương pháp phân lớp văn bản và gán nhãn chủ đề tiếp cận theo mô hình chủ đề với mục tiêu phân tích và hiểu được nội dung các thông điệp trao đổi của người dùng trên mạng xã hội. Trong đó, phương pháp gán nhãn chủ đề được thực hiện theo phương pháp học máy trên tập dữ liệu huấn luyện trong lĩnh vực giáo dục đại học. Tập dữ liệu huấn luyện này đã được xây dựng dựa trên ontology. Toàn bộ thành phần mô hình đề xuất được tích hợp trên hệ thống phân tích mạng xã hội. Thử nghiệm mô hình đề xuất trên tập ngữ liệu văn bản tiếng Việt được thu thập từ diễn đàn và mạng xã hội trong giáo dục đại học.

Từ khóa: *mô hình chủ đề, khám phá chủ đề, phân lớp chủ đề, gán nhãn chủ đề.*

1. GIỚI THIỆU

Phương pháp tiếp cận theo mô hình chủ đề được nhiều công trình nghiên cứu quan tâm trong lĩnh vực phân tích mạng xã hội và rút trích thông tin [1][3][4][5][15]. Chủ đề thể hiện động cơ, sở thích của cá nhân khi tạo lập văn bản, do đó khám phá chủ đề là khám phá được thuộc tính quan trọng, sở thích, thói quen, hành vi của cá nhân hay cộng đồng trên mạng xã hội [2][4][18][19][20]. Tuy nhiên, đối với tính chất của mạng xã hội, chủ đề của nội dung thông điệp trao đổi trên mạng xã hội chưa được tạo trước chủ đề hay nói cách khác chủ đề được trao đổi trên mạng xã hội là tiềm ẩn. Chính vì vậy, việc khám phá chủ đề và hiểu được nội dung thông điệp trao đổi của người dùng là một thách thức lớn và là bài toán khó, đặc biệt là đối với tập ngữ liệu văn bản tiếng Việt [18][20]. Trong bài báo này, chúng tôi khảo sát, nghiên cứu phương pháp phân tích mạng xã hội tiếp cận theo mô hình chủ đề và đề xuất mô hình kết hợp khám phá chủ đề, phân lớp và gán nhãn chủ đề từ nội dung trao đổi trên mạng xã hội.

Mô hình khám phá chủ đề từ tập ngữ liệu văn bản được đề xuất giai đoạn đầu bởi [13], trong một công trình nghiên được công bố trên tạp chí "Journal Of The American Society For Information Science". Bài báo đề xuất mô hình tiếp cận mới LSI (latent semantic indexing)

trong việc tự động rút trích thông tin và lập chỉ mục từ dựa trên việc xác định mối quan hệ ngữ nghĩa giữa các từ xuất hiện trong tài liệu văn bản hay chủ đề tồn tại trong văn bản đó và dựa trên mô hình không gian vector. Bên cạnh đó, khái niệm ma trận từ-tài liệu cũng được đề xuất trong nghiên cứu này. Tuy nhiên, mô hình LSI chưa quan tâm vấn đề phân bố xác suất của từ trên tài liệu [1] .

Một mô hình tiếp theo được gọi là PLSI (probabilistic latent semantic indexing) [14], mô hình tiếp cận dựa trên cải tiến mô hình LSI và mô hình chủ đề. Ý tưởng chính là mỗi tài liệu là sự pha trộn của nhiều chủ đề và mỗi chủ đề là một phân bố xác suất trên nhiều từ. Mô hình PLSI là một mô hình cải tiến hữu ích trong việc mô hình hoá tài liệu văn bản. Tuy nhiên, mô hình PLSI vẫn chưa đưa ra lý thuyết và mô hình xác suất tại mỗi cấp của các tài liệu và theo PLSI, mỗi tài liệu được xem là sự pha trộn tập những tỉ lệ của những chủ đề (mức độ cố định) trong tài liệu mà không đưa ra mô hình sinh xác suất cho những tỉ lệ đó. Điều này dẫn đến nhiều vấn đề: 1) số tham số trong mô hình sẽ gia tăng tuyến tính theo kích thước của tập ngữ liệu có nhiều tài liệu, điều này dẫn đến vấn đề "quá vừa dữ liệu – overfitting" trong quá trình học; 2) mô hình PLSI chưa giải quyết được làm thế nào có thể gán xác suất đến tài liệu bên ngoài tập tài liệu huấn luyện. Những hạn chế được [1] đề xuất cải tiến trong việc phát triển mô hình chủ đề sẽ được trình bày trong phần 2.2. Nội dung còn lại của bài báo được tổ chức như sau: Phần 2: trình bày các nghiên cứu liên quan. Phần 3: giới thiệu mô hình đề xuất tổng quát. Phần 4: kết quả thử nghiệm và thảo luận. Phần 5: kết luận và hướng phát triển.

2. Các nghiên cứu liên quan

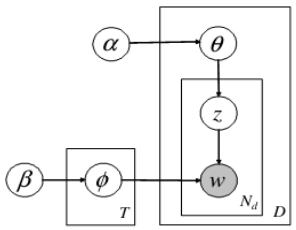
a. Tách từ và gán nhãn từ loại tiếng Việt

Công cụ vnTokenizer: công cụ tách từ được phát triển trong đề tài VLSP [16] độ chính xác được thẩm định là đạt 97%. Công cụ vnTagger là công cụ gán nhãn từ loại tiếng Việt dựa trên Conditional Random Fields (Lafferty et al.,2001) và Maximum Entropy (Nigam et al., 1999). vnTagger được xây dựng trong khuôn khổ nhánh đề tài VLSP [16] với dữ liệu huấn luyện khoảng 10.000 câu và 20,000 câu của Viet Treebank. Thử nghiệm của nhóm tác giả [16] với phương pháp 5-fold cross validation trên VTB-10,000 cho thấy kết quả gán nhãn với CRFs có thể đạt giá trị F1 lớn nhất là 93.45% và 10-fold cross validation với Maxent trên VTB-20,000 đạt giá trị F1 lớn nhất là 93.32%. Trong quá trình áp dụng công cụ này, kết quả của quá trình tách từ và gán nhãn từ loại vẫn còn tồn tại những hư từ (stopwords), các ký hiệu đặc biệt, các thẻ HTML tồn tại trong văn bản... làm ảnh hưởng đến kết quả phân tích từ đặc trưng và gán nhãn chủ đề. Vì thế, chúng tôi đã đề xuất cải tiến quá trình tiền xử lý dữ liệu đầu vào, nội dung được đề cập trong phần 3b.

b. Mô hình LDA

Sau quá trình tách từ và gán nhãn từ loại theo phương pháp được trình bày trong phần 2a, chúng tôi đề xuất áp dụng mô hình LDA nhằm mô hình hoá tài liệu và khám phá chủ đề trong tập ngữ liệu văn bản tiếng Việt. Latent Dirichlet Allocation (LDA) là một mô hình sinh xác suất (generative probabilistic model) cho các tập ngữ liệu rời rạc theo nhóm [1] và là mô hình cải tiến khắc phục những hạn chế của [14]. Về bản chất, LDA là một mô hình mạng Bayes theo 3 cấp [1][3][4], trong đó mỗi tài liệu (document) được mô tả dưới dạng kết hợp ngẫu nhiên của một tập

các chủ đề. Mỗi chủ đề (topic) là một phân bố rời rạc của một tập các từ đặc trưng (word). Theo đó, mỗi tài liệu tồn tại nhiều chủ đề tiềm ẩn và mỗi chủ đề có thể sẽ tồn tại trong nhiều tài liệu khác nhau, đây có thể được xem là mối quan hệ nhiều - nhiều giữa tài liệu và chủ đề.



Hình 1. Mô hình LDA

Mô hình LDA [1] phù hợp với tập ngữ liệu (corpus) với các dữ liệu rời rạc nhau được phân nhóm và được áp dụng để mô hình hóa kho ngữ liệu nhằm phát hiện ra các chủ đề tiềm ẩn của tập ngữ liệu. Trong đó, θ là phân bố xác suất các chủ đề của tài liệu d (phân bố này thỏa phân bố Dirichlet), α là tham số tập trung [0,1], z là một chủ đề được rút ra từ phân bố đa thức Multinomial(θ), ϕ là phân bố xác suất các từ khóa cho chủ đề z, phân bố này thỏa phân bố Dirichlet(β), w là tập từ đặc trưng được rút ra từ phân bố đa thức phát sinh bởi chủ đề z.

Quá trình sinh một tập tài liệu [1] 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. Trong biểu thức (1), chúng ta quan tâm đến các biến tham số z, θ, ϕ. Với mỗi θ_j là vector chứa các chủ đề của tài liệu j, mỗi z_i là chủ đề được gán cho từ w_i, mỗi $\phi^{(k)}$ là ma trận K x V với $\phi_{i,j} = p(w_i \mid z_j)$. Mục đích của mô hình LDA là phát hiện các từ thuộc về một chủ đề để từ đó ta có thể suy diễn chủ đề đó là chủ đề gì. Vấn đề quan trọng trong mô hình chủ đề là suy diễn hậu nghiệm. Đây được xem như là quá trình tạo sinh và suy diễn phân bố hậu nghiệm cho các biến ẩn là tập từ khóa cho chủ đề. Trong LDA, việc tạo sinh và suy diễn được tính theo công thức sau :

$$p(\theta, \phi, z \mid w, \alpha, B) = \frac{p(\theta, \phi, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \quad (1)$$

Tuy nhiên, không thể tính chính xác biểu thức (1) với nhân tố chuẩn hóa $p(w \mid \alpha, \beta)$ trong quá trình suy diễn và tạo sinh từ đặc trưng theo chủ đề [2]. Vì vậy, trong nghiên cứu [2] đề xuất dùng kỹ thuật lấy mẫu Gibbs (Gibbs Sampling).

c. *Phương pháp lấy mẫu Gibbs*

Việc hiện thực kỹ thuật lấy mẫu Gibbs cho mô hình LDA [2]: (1) thiết lập các biến đếm, (2) khởi tạo ngẫu nhiên các biến này, (3) sau đó thực hiện vòng lặp với số bước lặp mong muốn (thường từ 1000 tới 2000 lần lặp). Trong mỗi bước lặp, một chủ đề được lấy mẫu cho mỗi từ trong tập ngữ liệu. Qua (3), các biến đếm được sử dụng để tính các phân bố tiềm ẩn θ và ϕ. Kỹ thuật lấy mẫu Gibbs được nhiều nghiên cứu [3][4][5][15] quan tâm áp dụng.

Gibbs sampling [2] là một trong những họ giải thuật của Markov Chain Monte Carlo [1][2][4]. Mục đích của các giải thuật này là tạo ra xích Markov có phân bố hậu nghiệm như là phân bố ổn định. Nói một cách khác, sau khi lặp lại nhiều lần trên xích Markov, mẫu từ phân bố trở nên hội tụ giống với mẫu từ xác suất hậu nghiệm mong muốn. Gibbs sampling thực hiện

việc lấy mẫu dựa trên các phân bố điều kiện của các biến theo phương pháp tính xác suất hậu nghiệm [2].

Chẳng hạn để lấy mẫu x từ phân bố liên hợp $p(x) = p(x_1, x_2, \ldots, x_m)$, sử dụng một Gibb Sampling sẽ thực hiện các bước sau [2]:

1. Khởi tạo x_i ngẫu nhiên

2. Với mỗi t = 1 . . . T

b.1. $x_1^{t+1} = p(x_1 \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_m^{(t)})$

b.2. $x_2^{t+1} = p(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_m^{(t)})$

b.m. $x_m^{t+1} = p(x_m \mid x_1^{(t+1)}, x_2^{(t)}, \ldots, x_{m-1}^{(t)})$

Quá trình trên được lặp đi lặp lại nhiều lần cho đến khi các mẫu hội tụ. Tính hội tụ sẽ đảm bảo tính đúng đắn của Gibb Sampling. Tuy nhiên, trong [2] cũng chưa chứng minh phải cần chính xác bao nhiêu bước lặp thì các mẫu sẽ hội tụ. Trong thực tế, chúng ta có thể ước lượng ở mức độ chấp nhận được cho độ hội tụ có thể đạt được bằng việc tính log-likelihood [17] hoặc thậm chí thông qua việc kiểm tra các xác suất hậu nghiệm. Dưới đây là giải thuật đầy đủ để hiện thực lấy mẫu Gibbs [1][2][4][19][20].

Giải thuật 1. Gibbs Sampling cho LDA [1][2]

đầu vào: tập các từ w của tập văn bản d

đầu ra: các phép gán chủ đề và các biến đếm $n_{d,k}$, $n_{k,w}$, n_k

bắt đầu

khởi tạo ngẫu nhiên tập z và tăng các biến đếm

foreach bước lặp do

for i = 0 → N – 1 do

từ ← w[i]

chủ đề ← z[i]

$n_{d,\text{chủ đề}}$-= 1; $n_{\text{từ, chủ đề}}$-= 1; $n_{\text{chủ đề}}$-= 1

for k =0 → K -1 do

$$p(z = k \mid \cdot) = (n_{d,k} + \alpha_k) \frac{n_{k,w} + \beta_w}{n_k + \beta \times W}$$

[1]

end

chủ đề ← lấy mẫu từ $p(z \mid \cdot)$

z[i] ← chủ đề

$n_{d,\text{chủ đề}}$ += 1; $n_{\text{từ, chủ đề}}$ += 1; $n_{\text{chủ đề}}$ += 1

end

end

return z, $n_{d,k}$, $n_{k,w}$, n_k

end

Trong đó, d là một văn bản trong tập ngữ liệu nhiều văn bản D, w là một từ (word), $n_{d,k}$ số các từ w được gán vào chủ đề k trong văn bản d, $n_{k,w}$ số lần từ w được gán váo chủ đề k., n_k tổng số lần từ w được gán vào chủ đề k.

*d. **Phân lớp văn bản***

Phân lớp văn bản là một tiến trình đưa các văn bản chưa biết chủ đề vào các lớp văn bản đã biết tương ứng. Mỗi lĩnh vực được xác định bởi một số tài liệu mẫu của lĩnh vực đó được thể hiện dưới dạng văn bản [6][7]. Để thực hiện quá trình phân lớp, các phương pháp huấn luyện được sử dụng để xây dựng tập phân lớp từ các tài liệu mẫu, sau đó dùng tập phân lớp này để dự đoán lớp của những tài liệu mới. Trong bài báo này, chúng tôi chọn phương pháp phân lớp văn bản SVM (Support Vector Machine) [6][7] để thực hiện. Đặc điểm của các phương pháp phân

hai lớp tương tự SVM đến các thuật toán phân lớp đa lớp đều có đặc điểm chung là thường yêu cầu văn bản phải được biểu diễn dưới dạng vector đặc trưng, tuy nhiên các thuật toán khác SVM đều phải sử dụng các ước lượng tham số và ngưỡng tối ưu, trong khi đó thuật toán SVM có thể tự tìm ra các tham số tối ưu này. SVM là phương pháp sử dụng không gian vector đặc trưng lớn nhất hơn 10.000 chiều trong khi đó phương pháp khác có số chiều nhỏ hơn như Naive Bayes là 2000 và k-Nearest Neighbors là 2415 chiều [7].

Trong công trình năm 1999 của [6], Joachims đã so sánh SVM với Naive Bayesian, k-Nearest Neighbour, Rocchio, và C4.5 và đến năm 2003 [7], Joachims đã chứng minh rằng SVM làm việc rất tốt cùng với các đặc tính phức tạp của văn bản. Các kết quả cho thấy rằng SVM đưa ra độ chính xác phân lớp tốt nhất khi so sánh với các phương pháp khác. Các công trình nghiên cứu của nhiều tác giả (chẳng hạn như Kiritchenko và Matwin vào năm 2001, Hwanjo Yu và Han vào năm 2003, Lewis vào năm 2004) đã chỉ ra rằng thuật toán SVM đem lại kết quả tốt nhất trong phân lớp văn bản. Những phân tích của các tác giả trên đây cho thấy SVM có nhiều điểm phù hợp cho việc ứng dụng phân lớp văn bản.

Bên cạnh đó, chúng tôi khảo sát và ứng dụng bài toán phân lớp tuyến tính (linear classification) trong phân lớp văn bản. Đây là bài toán tìm ra hàm phân lớp hiệu quả nhất để phân biệt thành phần của các lớp trong việc huấn luyện dữ liệu [12]. Chẳng hạn, trong tập dữ liệu phân chia tuyến tính, hàm phân loại tuyến tính tương ứng f(x) phân chia 2 tập hợp. Khi hàm này được xác định thì bất kỳ một thể hiện mới sẽ được phân lớp đơn giản bằng việc xét dấu của hàm $f(X_n)$. Nếu X_n thuộc về tập các giá trị dương thì $f(X_n) > 0$ ngược lại thì thuộc tập các giá trị âm. Tính chất nổi trội của SVM là đồng thời cực tiểu lỗi phân lớp và cực đại khoảng cách lề giữa các lớp. Giả sử có 1 số điểm dữ liệu thuộc một trong hai lớp, và mục tiêu là xác định xem dữ liệu mới thêm vào sẽ thuộc phân lớp nào.

Ta xem mỗi điểm dữ liệu như một vector *p* chiều và chúng ta muốn biết là liệu có tách được những điểm đó bằng một siêu phẳng p-1 chiều hay không. Cho tập dữ liệu học gồm n dữ liệu gán nhãn[6][7]:

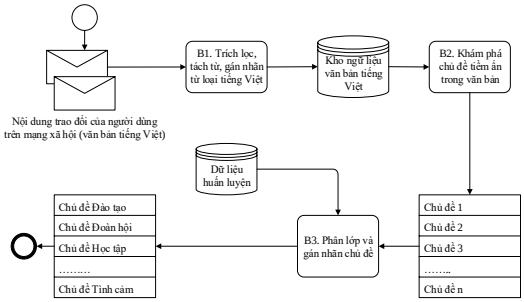
$$D=\{(x_1, y_1), (x_1, y_1),..., (x_n, y_n)\} \quad (2)$$

Với $y_i \in \{-1,1\}$ là một số nguyên xác định lớp của x_i. Mỗi x_i là một văn bản được biểu diễn dưới dạng một vector thực *d* chiều. Bộ phân lớp tuyến tính (mô hình phân lớp) được xác định thông qua một hàm có dạng: $f(x) = w.x - b = 0$, trong đó: w là vector pháp tuyến của f(x) và b đóng vai trò là tham số mô hình. Bộ phân lớp nhị phân $h: R^d \rightarrow \{0,1\}$ được xác định thông qua dấu của hàm sau[6][7]:

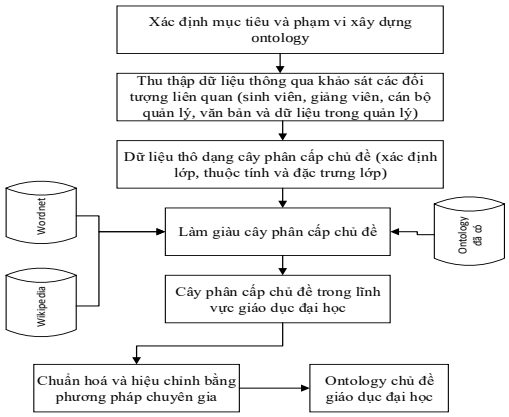
$$f(x).h(x) = \begin{cases} 1, f(x) > 0 \\ 0, f(x) \leq 0 \end{cases} \quad (3)$$

3. Mô hình khám phá chủ đề, phân lớp và gán nhãn chủ đề

a. Mô hình tổng quát



Hình 1. Mô hình tổng quát hệ thống khám phá, phân lớp và gán nhãn chủ đề



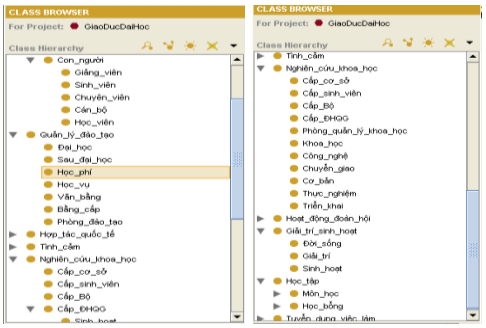
Hình 2. Xây dựng ontology trong lĩnh vực giáo dục đại học

Hình 1 là một mô hình tổng quát do chúng tôi đề xuất, kết hợp từ các công cụ, phương pháp tiền xử lý dữ liệu mô hình rút trích đặc trưng văn bản và phát hiện chủ đề, mô hình máy học phân lớp nội dung văn bản. Sự kết hợp xử lý của các giai đoạn trong mô hình này giúp tự động khám phá và gán nhãn chủ đề cùng với tập các từ đặc trưng cho chủ đề. Mô hình được thử nghiệm trên tập ngữ liệu văn bản tiếng Việt trong lĩnh vực giáo dục đại học.

b. Xây dựng ontology lĩnh vực giáo dục đại học

Để xây dựng bộ dữ liệu huấn luyện, trước tiên chúng tôi lựa chọn phương pháp dựa trên khái niệm ontology để xây dựng tập các khái niệm. Mục đích xây dựng ontology để xác định tập các khái niệm (các lớp) tương ứng với các chủ đề đề cập trong trường đại học [18]. Phương pháp xây dựng ontology gồm các bước (Hình 2).

Việc xác định các khái niệm (trong nghiên cứu này, xem khái niệm là chủ đề) trong giáo dục đại học được thực hiện dựa trên mô hình nghiên cứu do chúng tôi đề xuất (xem hình 2) [18]. Chúng tôi thực hiện khảo sát các đối tượng liên quan bằng nhiều phương pháp như phiếu khảo sát đối tượng liên quan như giảng viên, sinh viên, nhà quản lý; khảo sát ontology trong lĩnh vực đại học đã có bằng tiếng Anh thuộc chương trình DAML[1] và tiến hành kết nạp thêm các khái niệm mới vào ontology phù hợp với giáo dục Việt Nam [18]. Kết quả sau khi xây dựng ontology bao gồm các lớp (chủ đề): lớp cấp 0 (Giáo dục đại học), lớp cấp 1 (Đào tạo, Con người, Giáo trình, Hoạt đồng đoàn hội, Nghiên cứu khoa học, Doanh nghiệp, Hợp tác quốc tế,...) và các lớp con khác. Tất cả các lớp được tổ chức và quản lý trên phần mềm Protégé (xem hình 3).

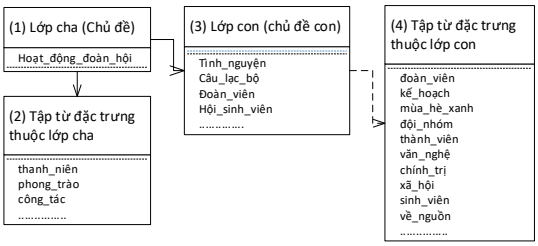


Hình 3. Một phần ontology giáo dục đại học được cài đặt và quản lý trên phần mềm Protégé

[1] Chương trình nghiên cứu "DARPA Agent Markup Language (DAML)" do Michael Pagels làm giám đốc. Mục tiêu của chương trình DAML là phát triển ngôn ngữ và công cụ hỗ trợ đắc lực trong vấn đề Semantic Web.

c. Chuẩn bị dữ liệu huấn luyện

Sau khi xây dựng ontology với hệ thống các khái niệm trong phần 2b, chúng tôi bắt đầu thực hiện phương pháp bán thủ công để xây dựng tập dữ liệu huấn luyện dựa trên ontology. Cụ thể, tập dữ liệu huấn luyện là tập văn bản để rút ra từ đặc trưng cho từng khái niệm thể hiện trên ontology. Phương pháp thực hiện bao gồm: tự động rút trích tập từ đặc trưng từ 300 văn bản được thu thập (báo cáo, thông báo, quy định trong trường đại học, các thảo luận trao đổi trên diễn đàn và mạng xã hội) có nội dung đề cập đến các hoạt động trong trường đại học. Việc tự động rút trích từ đặc trưng và gom nhóm các từ đặc trưng về cùng chủ đề chúng tôi áp dụng mô hình LDA. Tuỳ theo từng chủ đề mà có số từ đặc trưng khác nhau. Sau đó, chúng tôi gán nhãn chủ đề bằng phương pháp chuyên gia dựa trên tập khái niệm thể hiện trên ontology được xây dựng trong phần 2b.



Hình 4. Phân cấp chủ đề và tập từ đặc trưng thuộc chủ đề

Hình 4 trình bày chi tiết về kiến trúc phân cấp các chủ đề và tập từ đặc trưng chủ đề. Chẳng hạn, chủ đề "hoạt động đoàn hội" gồm có 4 chủ đề con (2) là "tình nguyện", "câu lạc bộ", "đoàn viên" và "hội sinh viên". Trong 4 chủ đề con này sẽ có những từ đặc trưng (4) mô tả cho từng chủ đề. Tuy nhiên, riêng đối với chủ đề cha "hoạt động đoàn hội" có tập từ đặc trưng trực tiếp là (2). Như vậy, xét tổng thể tập từ đặc trưng dùng cho tập dữ liệu huấn luyện cho chủ đề (1) "hoạt động đoàn hội" thì bao gồm cả (2), (3) và (4). Trong đó, (2) là danh sách chủ đề con mà cũng chính là đặc trưng của chủ đề cha (1).

d. Các bước thực hiện mô hình tổng quát (Hình 1)

Bước 1 (B1): Xử lý thu thập và tích hợp dữ liệu, thu thập nguồn dữ liệu từ mạng xã hội, diễn đàn. Trong giai đoạn này dữ liệu cần được tiền xử lý, phân tách từ loại và gán nhãn cho từ loại, vì các thông tin trao đổi trên mạng xã hội luôn có thông tin nhiễu như: các hư từ (stopwords), ký hiệu đặc biệt,… Chính vì thế, việc xử lý tiền dữ liệu là quan trọng, hệ thống sẽ lọc các thông tin nhiễu hay nội dung rác ra khỏi văn bản, sau đó dùng các phương pháp tách từ như vnTokenizer [16] để xác định từ loại là từ đơn, từ ghép trong tiếng Việt. Tiếp đến công cụ vnTagger [16] được sử dụng để xác định từ loại của mỗi từ đã được phân tách là danh từ, tính từ, động từ hoặc trạng từ… Việc xác định được thể loại từ sẽ giúp cho chúng ta tóm tắt lại nội dung và chỉ lấy những từ thực sự có nghĩa để sử dụng cho các bước sau. Tuy nhiên, trong bước tách từ và gán nhãn từ loại cho kết quả vẫn còn xảy ra lỗi từ. Chẳng hạn, dữ liệu sau xử lý chưa loại bỏ hết những lỗi từ. Điều này dẫn đến vẫn còn lỗi trong tập từ đặc trưng được rút trích. Mục đích là làm sạch dữ liệu để đạt độ chính xác trong khám phá chủ đề, chính vì vậy chúng tôi đã cải tiến tại bước này bằng phương pháp tiền xử lý dữ liệu nhằm giảm thiểu lỗi xảy ra. Bảng 1 dưới đây mô tả kết quả trước và sau khi thực hiện bước cải tiến tiền xử lý dữ liệu khi thực hiện tách từ và gán nhãn từ loại bằng công cụ vnTokenizer và vnTagger:

Bảng 1.

Nhóm văn bản (2000 văn bản cho mỗi nhóm)	Số lượng lỗi từ xảy ra theo phương pháp cũ	Số lượng lỗi xảy ra theo phương pháp cải tiến
Nhóm 1	120	11
Nhóm 2	50	0
Nhóm 3	280	13

Quan sát trong bảng 1, thử nghiệm phương pháp cải tiến trên 3 nhóm văn bản (thông điệp) gồm 6000 trong 13.208 thông điệp, ta thấy rằng trong nhóm 1 với 2000 thông điệp, phương pháp cũ còn xảy ra lỗi từ là 120, phương pháp cải tiến chỉ còn 11 lỗi từ, nhóm 2 và nhóm 3 cũng với 2000 thông điệp, kết quả cũng đã giảm đáng kể những lỗi từ trong tập kết quả sau bước 1 này. Với kết quả cải tiến này, trong bước 2 và 3, chúng tôi thực hiện khám phá và gán nhãn chủ đề đạt độ chính sát hơn.

Bước 2 (B2): Sau khi dữ liệu đã được tách từ và gán nhãn từ loại, tại bước 2 này mô hình sẽ áp dụng giải thuật LDA [1][5] và phương pháp lấy mẫu Gibbs để khám phá chủ đề tiềm ẩn và rút trích tập từ đặc trưng cho từng chủ đề từ nội dung thông điệp được trao đổi trên mạng xã hội... Kết quả thu được của bước này là danh sách các chủ đề chưa được gán nhãn mà chỉ được đánh số thứ tự, chẳng hạn như: chủ đề 1, chủ đề 2… chủ đề n.

Bước 3 (B3): Kết quả thu được ở B2 là tập hợp danh sách các chủ đề chưa được gán nhãn cùng tập từ đặc trưng, nhiệm vụ của bước B3 này phát hiện được chủ đề nào tương ứng với nhãn nội dung nào (gán nhãn chủ đề) và phân lớp các thông điệp vào từng chủ đề được gán nhãn. Tại bước này, áp dụng phương pháp học máy SVM [6][7] để học trên tập dữ liệu huấn luyện trình bày trong phần 3c. Giai đoạn này thực hiện quá trình phân lớp và gán nhãn phù hợp cho các chủ đề [6][8][9][10][11]. Tuy nhiên, trong quá trình phân lớp thông điệp, theo quan sát chúng tôi thấy rằng có những thông điệp có độ đo chính xác không phải là cao nhất ở bất kỳ chủ đề nào. Khi đó, những thông điệp này sẽ được xếp vào chủ đề nào có độ chính xác cao hơn so với các chủ đề còn lại của thông điệp đang xét phân lớp và quá trình phân lớp văn bản theo chủ đề được hoàn thành. Bảng 2 dưới đây mô tả một số kết quả trong quá trình gán nhãn và phân lớp chủ đề.

Bảng 2. Quá trình thực hiện phân lớp và gán nhãn bằng phương pháp SVM và tập dữ liệu huấn luyện

Chủ đề huấn luyện	Chủ đề tiềm ẩn	Độ chính xác	Kết quả gán nhãn
Hoạt động đoàn hội	Chủ đề 4	**0.93**	Gán nhãn "Hoạt động đoàn hội" cho Chủ đề 4
Hợp tác quốc tế	Chủ đề 4	0.65	
Đời sống sinh hoạt	Chủ đề 4	0.78	
Quản lý đào tạo	Chủ đề 4	0.80	
Hoạt động đoàn hội	Chủ đề 12	0.75	
Hợp tác quốc tế	Chủ đề 12	0.82	
Đời sống sinh hoạt	Chủ đề 12	0.65	
Quản lý đào tạo	Chủ đề 12	**0.91**	Gán nhãn "Quản lý đào tạo" cho Chủ đề 12
….	Chủ đề n	….	….

Trong quá trình phân lớp, vấn đề so sánh độ chính xác khi phân lớp giữa chủ đề tiềm ẩn và chủ đề huấn luyện là rất quan trọng, cặp chủ đề (chủ đề huấn luyện - chủ đề tiềm ẩn được khám phá từ bước 2 trên hình 1) được so sánh với độ chính xác cao nhất thì được hệ thống tự động gán nhãn và phân lớp theo đúng chủ đề được phương pháp SVM xác định (xem bảng 2). Bước cuối cùng chúng tôi áp dụng phương pháp thủ công để kiểm tra nhãn chủ đề đã được gán chính xác.

Trên bảng 2, độ chính xác (accuracy) là độ đo quá trình phân lớp và gán nhãn giữa chủ đề huấn luyện và chủ đề dựa trên phương pháp SVM. Dựa trên độ đo này, hệ thống xác định chính xác được chủ đề để gán nhãn. Chẳng hạn, đối với Chủ đề 4 được phân lớp theo 4 chủ đề huấn luyện là "Hoạt động đoàn hội", "Đời sống sinh hoạt", "Quản lý đào tạo" và chủ đề "Hợp tác quốc tế". Tuy nhiên, kết quả so với chủ đề huấn luyện "Hoạt động đoàn hội" có độ chính xác cao nhất là 0.93 hệ thống sẽ chọn chủ đề "Hoạt động đoàn hội" gán nhãn cho Chủ đề 4 và các chủ đề khác cũng được thực hiện tương tự. Sau khi kết thúc bước huấn luyện phân lớp và gán nhãn chủ đề, kết quả thu thập được tập chủ đề được gán nhãn. Ứng với mỗi chủ đề được gán nhãn sẽ có tập từ đặc trưng kèm theo xác suất.

4. Kết quả thử nghiệm và thảo luận

Thông tin trao đổi trên diễn đàn hay mạng xã hội là nguồn dữ liệu rất quan trọng để chúng ta phân tích và rút trích thông tin. Về dữ liệu thử nghiệm cho mô hình đề xuất có thể được thu thập từ nhiều nguồn khác nhau để phân tích như: lĩnh vực chính trị, xã hội, tư vấn sức khỏe, diễn đàn kinh tế, giáo dục… ngoài ra với sự phát triển vượt bậc về mạng xã hội ngày nay như các trang: Facebook, Twitter, Zing me… với các hàm API được cung cấp sẵn chúng ta dễ dàng thu thập nguồn dữ liệu thử nghiệm. Trong phạm vi nghiên cứu này, bài báo tập trung phân tích dữ liệu trong lĩnh vực giáo dục đại học. Thử nghiệm mô hình trên cơ sở dữ liệu diễn đàn sinh viên được chúng tôi đã tổ chức lại như một cấu trúc của một mạng xã hội giáo dục để khảo sát và phân tích từ đó khám phá, phân lớp và gán nhãn chủ đề. Dữ liệu này được xem như là một mạng xã hội, đây là nơi người dùng được phép đăng ký làm thành viên, trao đổi về các nội dung học tập, đời sống, văn hóa, xã hội, đoàn hội, đào tạp… Cơ sở dữ liệu bao gồm: kích thước dữ liệu: 350 MB, số lượng người dùng: 2.494 (người dùng) có trao đổi thảo luận, số lượng thông điệp trao đổi giữa người dùng: 13.208 (bài viết), số lượng bài viết này được chọn lọc hình thức bán tự động và đã loại bỏ đi nhiều bài viết ngoài tiếng Việt hoặc những bài viết bằng ký tự đặc biệt, bài viết không có ý nghĩa. Chúng tôi gọi tên tập dữ liệu là FTSN.

Bảng 3. Kết quả tốc độ xử lý trung bình của chương trình trên 2000 bài viết chạy trên hệ điều hành Windows 2007, CPU 2 GB, RAM 4GB.

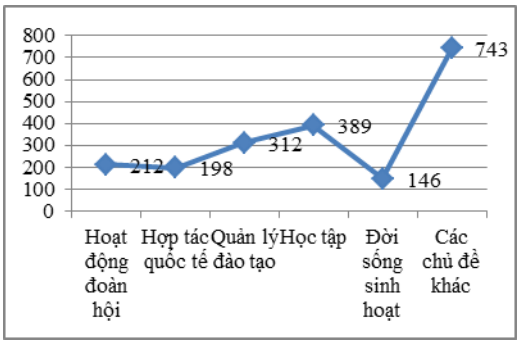
Quá trình xử lý		Tốc độ (giây)
Tiền xử lý dữ liệu	Lọc thẻ html, khoảng trắng, các ký hiệu đặc biệt, hư từ,…	77
	vnTokenizer	383.4
	vnTagger	207.3
Xử lý LDA	Khám phá chủ đề	125
Xử lý SVM	Phân lớp và gán nhãn chủ đề	36

Áp dụng mô hình kết hợp khám phá chủ đề LDA và phân lớp, gán nhãn chủ đề SVM, trong quá trình thử nghiệm thực tế, kết quả thu được trong Bảng 3 và 4.

Bảng 4. Kết quả phân loại vă bản theo chủ đề khảo sát trên 2000 bài viết đầu tiên

Chủ đề	Số văn bản phân loại
Hoạt động đoàn hội	212
Hợp tác quốc tế	198
Quản lý đào tạo	312
Học tập	389
Đời sống sinh hoạt	146
Các chủ đề khác	743

Trên bảng 4, qua quá trình thử nghiệm thực tế trên 2000 dữ liệu trao đổi đầu tiên trong tập ngữ liệu và tập chủ đề huấn luyện, hệ thống rút trích ra được 212 thông điệp nói về chủ đề "Hoạt động đoàn hội" và 198 thông điệp nói về chủ đề "Hợp tác quốc tế", 312 thông điệp nói về chủ đề "Quản lý đào tạo", 156 thông điệp nói về chủ đề "Đời sống sinh hoạt",....



Hình 2. Biểu đồ thể hiện kết quả phân loại chủ đề trên 2000 bài viết đầu tiên

Với hệ thống rút trích phân loại phát hiện thông tin trao đổi theo chủ đề, chúng ta đã có phân loại, gán nhãn chủ đề bằng phương pháp máy học trên tập dữ liệu huấn luyện và phương pháp chuyên gia. Trong Bảng 5, kết quả trình bày 3 chủ đề trong 20 chủ đề được hệ thống đề xuất từ dữ liệu FTSN. Việc xác định 20 chủ đề là xác định giá trị tham số đầu vào khi chạy mô hình, chính vì vậy, việc số lượng bao nhiêu chủ đề sẽ phụ thuộc vào lĩnh vực được khảo sát hay yêu cầu cần khảo sát.

Bảng 5. Danh sách các chủ đề đã được gán nhãn cùng tập từ đặc trưng (kèm xác suất) đại diện cho mỗi chủ đề

Chủ đề 4 "Hoạt động đoàn hội"		Chủ đề 12 "Quản lý đào tạo"		Chủ đề 17 "Tuyển dụng việc làm"	
Đoàn	0.03189	Phòng	0.03925	công_ty	0.03245
hội	0.01957	quản_lý	0.02322	việc_làm	0.01767
Bác	0.01769	tín_chỉ	0.01933	kiến_tập	0.01485
hoạt_động	0.01689	học_phí	0.01402	nhân_viên	0.01093
đoàn_viên	0.01433	giảng_dạy	0.01021	tham_quan	0.00952
nội_dung	0.01277	thi	0.00967	lương	0.00711
công_tác	0.00677	học_lại	0.00833	người	0.00429
năm_học	0.00597	thời_khoá_biểu	0.00790	môi_trường	0.00641
qui_định	0.00535	phòng_học	0.00718	hoạt_động	0.00500
phong_trào	0.00499	máy_chiếu	0.00554	học_hỏi	0.00359

Trong Bảng 5, Chủ đề 4 được gán nhãn "Hoạt động đoàn hội, Chủ đề 12 được gán nhãn "Quản lý đào tạo" và Chủ đề 17 được gán nhãn "Tuyển dụng việc làm". Mỗi chủ đề được trình bày 10 từ đặc trưng với xác suất cao nhất tương ứng. Những chủ đề chỉ ra rằng sinh viên thường quan tâm và trao đổi trên mạng những nội dung liên quan. Sinh viên trao đổi về các "Hoạt động đoàn hội" như phong trào, hoạt động, qui định, sinh viên nói về Bác và những hoạt động hướng về Bác,... Sinh viên trao đổi về "Quản lý đào tạo" thường trao đổi về phòng, quản lý, tín chỉ, học phí,...Sinh viên trao đổi về "Tuyển dụng việc làm" thường đề cập về công ty, việc làm, môi trường, lương,...

Cuối cùng, một lần nữa chúng tôi thực hiện đánh giá kết quả gán nhãn chủ đề bằng phương pháp chuyên gia. Chúng tôi chọn năm người bao gồm chuyên gia về ngôn ngữ, cán bộ quản lý, giảng viên, sinh viên cùng kiểm tra bằng cách chọn từ đặc trưng cho chủ đề được thực hiện bằng tay và độc lập nhau. Kết quả cho thấy, độ chính xác gán nhãn của năm người chỉ chênh lệch từ 2% - 5%. Tuy nhiên, so sánh kết quả gán nhãn bằng tay trên với kết quả gán nhãn theo phương pháp gán nhãn tự động chúng tôi đề xuất đạt độ chính xác 88%. Tỉ lệ này chưa tuyệt đối có thể do việc lựa chọn các tham số đầu vào trong quá trình thực hiện lấy mẫu Gibbs cho mô hình LDA và quá trình làm sạch dữ liệu, tách từ và gán nhãn từ loại tương đối phức tạp khi thử nghiệm trên ngữ liệu tiếng Việt. Chúng tôi sẽ tiếp tục nghiên cứu để có thể thử nghiệm trên nhiều tham số khác nhau trong quá trình lấy mẫu Gibbs cho mô hình LDA để có thể cải tiến nhằm làm tăng độ chính xác cho phương pháp khám phá và gán nhãn chủ đề được đề xuất trên.

5. Kết luận và hướng phát triển

Bài báo đã đề xuất được mô hình khám phá, phân lớp và gán nhãn chủ đề trong lĩnh vực phân tích mạng xã hội và rút trích thông tin dựa theo mô hình chủ đề và thử nghiệm trên tập ngữ liệu văn bản tiếng Việt được thu thập từ diễn đàn, mạng xã hội trong giáo dục. Quá trình thử nghiệm trên tập ngữ liệu tiếng Việt, chúng tôi đã cải tiến bước xử lý và làm sạch dữ liệu đầu vào mô hình nhằm cải tiến quá trình tách từ và gán nhãn từ loại tiếng Việt. Phương pháp nghiên cứu chính của bài báo là: (1) xây dựng ontology gồm tập khái niệm trong lĩnh vực giáo dục đại học và xây dựng tập dữ liệu huấn luyện là tập từ đặc trưng theo từng khái niệm trên ontology, (2) khám phá chủ đề tiềm ẩn (chưa được gán nhãn) dựa trên tập từ đặc trưng được rút trích từ tập văn bản bằng mô hình LDA kết hợp phương pháp lấy mẫu Gibbs dựa trên xích Markov và mô hình xác suất Bayes, (3) từ tập chủ đề chưa được gán nhãn và tập văn bản chưa được phân lớp, áp dụng phương pháp học máy SVM với tập dữ liệu huấn luyện để phân lớp văn bản và gán nhãn chủ đề.

Mô hình đề xuất đã cho kết quả đáng kể, hệ thống có thể tự động làm sạch dữ liệu tương đối tốt, quá trình phát hiện chủ đề và phân loại gán nhãn chủ đề cũng đã được cải tiến với độ chính xác cao hơn, thời gian xử lý được giảm đi đáng kể. Tuy nhiên quá trình thực hiện cũng gặp phải một số hạn chế về các dữ liệu bị nhiễu, do chưa hoàn toàn làm sạch tuyệt đối dữ liệu đầu vào nên quá trình phân lớp và gán nhãn sau đó gặp một số khó khăn về mặt xử lý vì thế chưa đạt được độ chính xác tuyệt đối.

Chúng tôi sẽ tiếp tục cải tiến quá trình lọc dữ liệu đầu vào nhằm tối đa hoá độ chính xác quá trình phân lớp và gán nhãn chủ đề. Bên cạnh đó, nghiên cứu tiếp theo, chúng tôi sẽ kết hợp mô

hình đã đề xuất trong nghiên cứu này và phương pháp phân tích thông tin của người dùng mạng (profile) để từ đó tiến tới khảo sát được sự thay đổi mối quan tâm của người dùng theo từng giai đoạn thời gian.

Lời cám ơn

"Nghiên cứu này được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (VNU-HCM) trong đề tài mã số B2013-26-02". Xin cảm ơn anh Đoàn Nguyễn Ngọc Duy đã hỗ trợ và góp ý cho chúng tôi trong quá trình thực hiện nghiên cứu này.

An integrated model for discovering, classifying and labeling topics based on topic modeling

- **Thanh Ho** [a]
- **Phuc Do** [b]

[a] University of Economics and Law, VNUHCM

[b] University of Information Technology, VNU-HCM

ABSTRACT

In this paper, we propose an integrated model for discovering, classifying and labeling topics of messages based on topic modeling to analyze and understand the topics of the messages posted by users on social networks. In which, the method of labeling is executed by machine learning on the training data and ontology. The ontology is created in the field of higher education. All parts of model are integrated on a system called social network analysis system based on topic modeling. The experiment of the model on the linguistic data of Vietnamese texts collected from a student forum is transformed into a data structure of social network, including: 13,208 messages by 2,494 users.

Keywords: *topic modeling, discovering topics, classifying topics, labeling topics.*

TÀI LIỆU THAM KHẢO

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

[2] Tom Griffiths, Gibbs Sampling in the generative model of Latent Dirichlet Allocation, Gruffydd@psych.stanford.edu (2004).

[3] Editor Charu C. Aggarwal, Social Network Data Analytics, IBM Thomas J. Watson Research Center (2011).

[4] Andrew McCallum, Andr´es Corrada, Xuerui Wang, The Author-Recipient-Topic Model for

Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email, Department of Computer Science, University of MA (2004).

[5] Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, and Quang-Thuy Ha, A hidden topic-based framework towards building applications with short Web documents, *IEEE Transactions on Knowledge and Data Engineering* (IEEE TKDE), Vol.23, No.7, pp.961-976. [PDF] (2011).

[6] T. Joachims, Transductive Inference for Text Classification using Support Vector Machines, International Conference on Machine Learning (ICML) (1999).

[7] T. Joachims, Transductive learning via spectral graph partitioning, Proceeding of The TwentiethInternational Conference on Machine Learning (ICML2003): 290-297 (2003).

[8] Davide Magatti, Automatic Labeling Of Topics, Department of Informatics, Systems and Communications, Universit´a degli Studi di Milano-Bicocca (2009).

[9] William M. Darling, A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling - School of Computer Science, University of Guelph (2011).

[10] Jey Han Lau, Best Topic Word Selection for Topic Labelling - Dept of Computer Science and Software Engineering, University of Melbourne (2010).

[11] Jey Han Lau, Automatic Labelling of Topic Models, Dept of Computer Science and Software Engineering, University of Melbourne (2011).

[12] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer (2007).

[13] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman, Indexing by latent semantic analysis, *Journal American Society for Information Science* (1990).

[14] Hofmann, Thomas, Probabilistic Latent Semantic Indexing (PDF). Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (1999).

[15] Youngchul Cha, Junghoo Cho, Social-network analysis using topic models, the 35th international ACM SIGIR conference, pp. Pages 565-574 (2012).

[16] Nhánh đề tài cấp nhà nước Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt - VLSP http://vlsp.vietlp.org:8080/demo/.

[17] http://en.wikipedia.org/wiki/Likelihood_function

[18] Hồ Trung Thành, Đỗ Phúc, Ontology tiếng Việt trong lĩnh vực giáo dục đại học, Tạp chí Khoa học Công nghệ, Viện Hàn lâm Khoa học Công nghệ Việt Nam. ISSN: 0866 – 708x (2014).

[19] Muon Nguyen, Thanh Ho, Phuc Do, Social Networks Analysis Based on Topic Modeling, The 10th IEEE RIVF International Conference on Computing and Communication Technologies (P.119-122), Hanoi (2013).

[20] Tran Quang Hoa, Vo Ho Tien Hung, Nguyen Le Hoang, Ho Trung Thanh, Do Phuc, Finding the Cluster of Actors in Social Network based on the Topic of Messages, ACIIDS 04/2014, ThaiLan (2014).

[21] Thanh Ho, Duy Doan, Phuc Do, Discovering Hot Topics On Social Network Based On Improving The Aging Theory, Advances in Computer Science : an International Journal**, ISSN : 2322-5157** (2014)**.**