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.2017.00021

ĐỘ TƯƠNG ĐỒNG NGỮ NGHĨA CỦA CÁC BÀI VIẾT TRÊN MẠNG XÃ HỘI DỰA THEO WIKIPEDIA

Nguyễn Thị Hội[1], Đàm Gia Mạnh[1], Trần Đình Quế[2]

[1] Trường Đại học Thương mại

[2] Học viện Công nghệ Bưu chính Viễn thông, Hà Nội

hoint2002@gmail.com, damgiamanh@gmail.com,tdque@yahoo.com

TÓM TẮT: *Các bài viết ở các dạng như Blog, tags, các trang chia sẻ nội dung,... được đăng trên các mạng xã hội là nguồn tài nguyên vô giá thu hút nhiều nghiên cứu và khám phá quan tâm, sở thích của người dùng cho phát triển các ứng dụng trong sản xuất kinh doanh, hoạt động chính trị, giáo dục, thương mại điện tử, tư vấn bạn đọc, tư vấn dịch vụ,..... Các nghiên cứu đa phần tập trung vào việc phân loại các bài viết, tìm kiếm hoặc trích chọn đặc trưng các bài viết dựa trên các đoạn văn bản, các mô tả ngắn nào đó để từ đó có thể phân loại người sử dụng. Một trong những cơ sở cho việc phân loại như vậy, là vấn đề ước lượng độ tương đồng của những bài viết này. Hầu hết các nghiên cứu hiện nay chú trọng tính toán độ tương đồng chỉ dựa vào một đặc trưng nào đó như nội dung hay tags,... hơn là xem xét nhiều khía cạnh liên quan. Mục đích của bài báo này trước hết đề xuất mô hình "bài viết" trên mạng xã hội dựa trên một số đặc trưng như tiêu đề, chủ đề, các đánh dấu và nội dung của bài viết. Sau đó, chúng tôi trình bày một độ đo tích hợp để ước lượng độ tương đồng giữa các bài viết theo ngữ nghĩa dựa trên thư viện bách khoa toàn thư Wikipedia. Kết quả thử nghiệm của chúng tôi đã chỉ ra rằng, việc ước lượng tương đồng tích hợp trên nhiều thuộc tính được đánh giá là tốt hơn so với ước lượng cho từng đặc trưng riêng của các bài viết trên mạng xã hội.*

Từ khóa: *Mạng xã hội, bài viết, mô hình, độ tương đồng, ngữ nghĩa.*

I. GIỚI THIỆU

Theo Zafarani cùng cộng sự [16] thì mạng xã hội là những ứng dụng dựa trên nền tảng internet nhằm thiết lập và chuyển đổi nội dung được tạo ra bởi các người dùng cho những mục đích sử dụng khác nhau. Đặc trưng quan trọng nhất của các mạng xã hội là khả năng chia sẻ nội dung số từ văn bản số, hình ảnh số, các đoạn video,… nhằm giúp cho các người dùng có thể tương tác với nhau và thể hiện quan điểm thông qua các chức năng được cung cấp bởi các mạng xã hội này. Theo thống kê của trang Statistic.com thì đến tháng 4.2017 mạng xã hội Facebook.com đã có 1968 triệu người dùng, Whatsapp đã có 1200 triệu người dùng, YouTube.com đã có 1000 triệu người dùng, Instagram đã có 600 triệu người dùng,... Với thông tin từ lượng người dùng khổng lồ như vậy, các mạng xã hội đã đem lại nguồn tài nguyên vô giá cho việc nghiên cứu và phát triển ứng dụng trong nhiều lĩnh vực khác nhau như sản xuất, kinh doanh, thương mại, giáo dục, y tế,... đặc biệt trong lĩnh vực quảng cáo trực tuyến, các hệ thống tư vấn sản phẩm, các hệ thống khuyến nghị người dùng.

Một video clip, một hình ảnh, một văn bản, hoặc là kết hợp của một số nội dung đó khi được đăng có thể gọi là một "bài viết" trên mạng xã hội. Trong hạn chế của bài báo này, chúng tôi chỉ xem xét các bài viết có chứa văn bản trong các thuộc tính của chúng, nếu các bài viết chỉ có video clip, hoặc chỉ có hình ảnh,… và không chứa văn bản sẽ không được đề cập đến. Như vậy, bài toán xem xét và ước lượng độ tương đồng về ngữ nghĩa giữa các bài viết trên mạng xã hội chủ yếu tập trung vào xem xét, ước lượng độ tương đồng về ngữ nghĩa giữa các văn bản và sử dụng cơ sở tri thức là thư viện Wikipedia trực tuyến để ước lượng so sánh độ tương đồng về mặt ngữ nghĩa của các thuộc tính của các bài viết trên các mạng xã hội.

Một trong những vấn đề quan trọng trong nghiên cứu tính toán trên mạng xã hội là phát hiện sự tương đồng giữa hai đối tượng như người sử dụng hay các bài viết và đã thu hút rất nhiều quan tâm nghiên cứu. D. Lin [4] đã đề xuất một mô hình ước lượng độ tương đồng giữa hai đối tượng dựa trên cách tiếp cận của lý thuyết thông tin. Say và Kumar [13] lại đề xuất một mô hình phân nhóm dựa trên các tập dữ liệu quan hệ bằng cách sử dụng các tính chất phụ thuộc hàm như là các tham số để ước lượng độ tương đồng. Nguyen et al [10] đã đưa ra một mô hình tổng quát để ước lượng độ tương đồng giữa hai đối tượng, bằng cách ước lượng độ tương đồng trên các thuộc tính của chúng, sau đó tích hợp dựa trên trọng số để đưa ra độ tương đồng của hai đối tượng cần xem xét.

Đặc biệt về khía cạnh văn bản, các nghiên cứu tính toán độ tương đồng giữa các văn bản có thể nhóm vào hai hướng tiếp cận chính: Hướng thứ nhất dựa trên thống kê: Theo hướng tiếp cận này, các văn bản được so sánh dựa trên việc thống kê các từ, các nhóm từ, các cấu trúc của từ,… hoặc dựa trên thống kê số lượng các từ xuất hiện trong các văn bản. Điển hình như các nghiên cứu của Bollegala et al [2], Buscaldi et al [3], Sultan et al [11],… Hướng tiếp cận thứ hai là xem xét độ tương đồng của các văn bản dựa trên ngữ nghĩa theo môt cơ sở tri thức nào đó, các nghiên cứu theo hướng tiếp cận này chủ yếu xem xét ngữ nghĩa dựa trên các hệ thống tri thức thông dụng như WordNet, các ontology hoặc các hệ thống từ điển tự xây dựng. Điển hình như các nghiên cứu của Buscaldi et al [4], Han et al [8], Nguyen và Tran [5], J.Xu và Qin Lu [7]. Ở Việt Nam, phân tích Tiếng Việt đã có rất nhiều nghiên cứu của nhiều nhóm nghiên cứu và các tác giả của nhiều trường Đại học lớn, điển hình nhất là dự án do GS. Hồ Tú Bảo chủ trì, về phân tích và xử lý Tiếng Việt, các nghiên cứu chủ yếu chia thành các nhóm chính gồm: xử lý tiếng nói, nhận dạng tiếng nói, nhận dạng chữ viết, dịch tự động, tóm tắt văn bản, tìm kiếm thông tin, trích chọn thông tin, phát hiện tri thức và khai

phá dữ liệu văn bản,… Tuy vậy, bài toán trích chọn thông tin và phát hiện thông tin từu văn bản luôn là một bài toán mở, đến nay vẫn có nhiều hướng tiếp cận và ứng dụng thực tiễn, đặc biệt là các nghiên cứu về so sánh ngữ nghĩa của các văn bản tiếng Việt, đã có nhiều hướng nghiên cứu cũng như nhiều công nghệ được ứng dụng trong xử lý ngôn ngữ như giao diện người máy bằng ngôn ngữ tự nhiên, các hệ hỏi đáp, các hệ sinh ra ngôn ngữ,… các nghiên cứu này đa phần tập trung xem xét ước lượng độ tương đồng trên một hoặc vài thuộc tính của văn bản, các văn bản chủ yếu là các văn bản có cấu trúc rõ ràng, hoặc các ngữ cảnh cụ thể, vì vậy, chúng tôi đề xuất một hướng tiếp cận khác trong phân tích và so sánh ngữ nghĩa Tiếng Việt sử dụng từ điển Wikipedia. Do giới hạn về các công bố khoa học ở Việt Nam, chúng tôi chưa thấy các kết quả nghiên cứu so sánh ngữ nghĩa Tiếng Việt theo Wikipedia ở Việt Nam trong thời gian gần đây, do đó, chúng tôi lựa chọn thư viện Wikipedia trực tuyến bởi có hai lý do chính: Thứ nhất, Wikipedia hiện nay là thư viện có khối lượng bài viết rất phong phú, đa dạng: theo thống kê mới nhất vào tháng 5 năm 2017 thì hiện nay trên Wikipedia có 1,155 đề mục và bài viết Tiếng Việt; có 5,392 triệu đề mục và bài viết Tiếng Anh; 1,864 triệu đề mục và bài viết Tiếng Pháp... Các đề mục được định nghĩa rất chi tiết bao gồm cả văn bản, hình ảnh, sơ đồ, minh họa,... rất dễ hiểu và dễ dùng. Thứ hai, đây là một thư viện bách khoa toàn thư động, trực tuyến, được bổ sung, chỉnh sửa thường xuyên bởi hàng triệu người dùng trên thế giới và nó cho phép người dùng sử dụng miễn phí và các nghiên cứu so sánh ngữ nghĩa văn bản bằng Wikipedia còn rất ít ở Việt Nam.

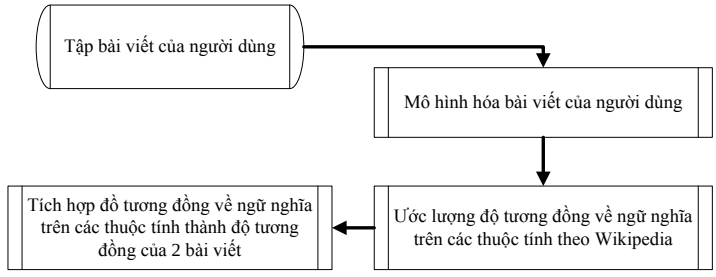
Bài báo của chúng tôi trình bày hai nội dung chính bao gồm: Đề xuất mô hình hóa các bài viết được đăng trên các mạng xã hội với các thuộc tính của chúng như tiêu đề (title), chủ đề (topic), các đánh dấu (tags), nội dung (content),… Tiếp theo là đề xuất mô hình ước lượng độ độ tương đồng về ngữ nghĩa của các bài viết của người dùng trên mạng xã hội dựa vào thư viện trực tuyến Wikipedia. Phần còn lại của bài báo được cấu trúc như sau: Phần II giới thiệu mô hình bài viết và ước lượng độ tương đồng; Phần III là thực nghiệm và đánh giá kết quả và phần IV là kết luận và các hướng nghiên cứu tiếp theo của chúng tôi.

II. MÔ HÌNH ĐỘ TƯƠNG ĐỒNG GIỮA CÁC BÀI VIẾT DỰA TRÊN WIKIPEDIA

Giới thiệu mô hình

Giả sử chúng ta có hai bài viết a_i và a_j cùng các thuộc tính của chúng, bài toán đặt ra cần ước lượng độ tương đồng về ngữ nghĩa của hai bài viết theo thư viện bách khoa toàn thư Wikipedia, mô hình có các bước chính như sau (Hình 1):

- Mô hình hóa các bài viết của người dùng trên các mạng xã hội.
- Ước lượng độ tương đồng về ngữ nghĩa trên từng đặc trưng theo Wikipedia.
- Tích hợp độ tương đồng về ngữ nghĩa của các bài viết dựa trên độ tương đồng trên các đặc trưng của bài viết.



Hình 1. Mô hình tổng quát ước lượng độ tương đồng dựa trên Wikipedia

A. Mô hình hóa các bài viết

Giả sử rằng trên một mạng xã hội có một tập các người dùng: $U = \{u_1, u_2, \dots, u_n\}$, mỗi người dùng u_i có một tập các bài viết: $A = \{a_1, a_2, \dots, a_m\}$. Mỗi bài viết có thể được đặc trưng bởi các thuộc tính như: tiêu đề (title), chủ đề (topic), các đánh dấu (tags), nội dung (content),... Để tiện trình bày, mỗi bài viết trong tập các bài viết của một người dùng u_i có p thuộc tính, được ký hiệu là: $(f_1^i, f_2^i, \dots, f_p^i)$ và trong thực nghiệm, chúng tôi xem xét và ước lượng độ tương đồng các thuộc tính của các bài viết bao gồm:

- *Tiêu đề (Title)* của bài viết ký hiệu là: f_{tit}^i. Nó có thể là một câu ngắn, một câu tóm lược nội dung của bài viết. Nếu nội dung bài viết là một hình ảnh thì tiêu đề của bài viết chính là nhãn hay chú thích (caption) của ảnh nếu bài viết không có tiêu đề nào khác.
- *Nội dung (Content)* của bài viết ký hiệu là: f_{con}^i. Nó có thể một video clip, một hình ảnh, một văn bản hoặc là một sự kết hợp giữa chúng. Tuy nhiên, trong bài báo này, chúng tôi chỉ xem xét nội dung là một câu văn ngắn, một câu cảm thán, một đoạn văn hoặc một bài văn. Nói cách khác,chúng tôi chỉ xem xét nội dung bài viết là văn bản, nếu trong trường hợp không có nội dung, chúng tôi sẽ coi thuộc tính này không có hoặc không tồn tại.

- *Đánh dấu (Tags)* của bài viết ký hiệu là: f_{tag}^{i}. Trên thực tế, một bài viết có thể được gắn vào một đánh dấu hoặc một tập các đánh dấu. Chúng giúp người dùng có thể truy nhập các chủ đề nhanh hơn hoặc dễ gợi nhớ đến nội dung bài viết, các đánh dấu còn được gọi là hashtag và được để trong cặp dấu # #, một đánh dấu có thể là một từ, một cụm từ hay một câu.
- *Chủ đề (Topic)* của bài viết ký hiệu là: f_{top}^{i}. Trên các mạng xã hội, mỗi bài viết có thể được sắp xếp vào một hoặc nhiều chủ đề, tùy theo nội dung của bài viết hoặc sự phân loại của người dùng. Mỗi chủ đề hay nhóm bài viết được đặc trưng bởi một từ hay một nhóm từ.

Như vậy, sau khi được mô hình hóa theo cách thể hiện trên thì một bài viết sẽ được đặc trưng bởi một tập các thuộc tính, trong bài báo này mô hình chúng tôi hạn chế chỉ xem xét ước lượng độ tương đồng về ngữ nghĩa của các thuộc tính có chứa văn bản. Do đó, bài toán ước lượng độ tương đồng về ngữ nghĩa của các bài viết tương đương với bài toán ước lượng độ tương đồng về ngữ nghĩa của hai tập văn bản, việc ước lượng độ tương đồng giữa hai văn bản theo ngữ nghĩa trong bài này được tính toán theo ngữ nghĩa của từ điển bách khoa toàn thư Wikipedia.

B. Độ tương đồng về ngữ nghĩa của các bài viết

Để xem xét và ước lượng độ tương đồng về ngữ nghĩa của các bài viết, chúng tôi lần lượt ước lượng độ tương đồng trên các thuộc tính tương ứng của chúng, sau đó, kết hợp các độ tương đồng theo trọng số trên các thuộc tính thành độ tương đồng của các bài viết. Ví dụ, chúng tôi có nội dung bài viết: "*Giá máy tính đã thay đổi rất nhiều trong những năm qua!",* đầu tiên chúng tôi tách các bài viết theo n-gram, trong thực nghiệm chúng thôi chọn N= 2, 3 và 4 do khi thực hiện 1-gram Tiếng Việt, hầu như các từ không có định nghĩa trên Wikipedia, nếu loại bỏ từ dừng thì số khóa thu được không đáng kể, với 2-gram chúng tôi có các cụm từ {*giá máy, máy tính, tính đã, đã thay, thay đổi, đổi rất, rất nhiều, nhiều trong, trong những, những năm, năm qua*}, tương tư với 3-gram và 4-gram sau đó loại bỏ từ dừng chúng tôi có *{giá máy tính, máy tính, thay đổi, năm qua}*. Sau đó với các từ khóa này chúng tôi trích chọn định nghĩa của chúng từ Wikipedia bằng cách đọc nội dung từ trang của Wikipedia.com, sau đó lưu lên CSDL và thực hiện tách n-gram trên CSDL. Cuối cùng lưu các tập từ khóa thu được vào CSDL và ước lượng giữa các tập từ khóa thu được bằng độ đo Jacard. Độ đo Jacard trên hai tập từ ngữ được tính theo công thức (5).

1. Độ tương đồng về ngữ nghĩa trên các thuộc tính

Bước 1: Xây dựng tập từ khóa của thuộc tính

Bảng 1. Xây dựng tập từ khóa của các thuộc tính.

Thuật giải 1: Xác định tập từ khóa, GetKeyword (x,y)

Input: Thuộc tính i của bài viết a, a[i]

Ouput: Danh sách các từ khóa của a[i] thu được từ Wikipedia, Ka[i]

```
1:  x:= a[i]; y=∅;  T1:= ∅; T11:= ∅;
3: T2:= ∅; T22:=∅; W:= ∅; T3:= ∅;
4:  For i:=2 to n do
            T1:= T1∪separateN-gram (x, i) ;      //n là số các gram (trong mô hình chúng tôi lựa chọn n=2,3,4)
   End For
5: T2:= T2 ∪removeStopword (T1);                 //Loại bỏ từ dừng và từ chưa xác định trên Wikipedia
6: For i:=1 to count (T2) do
      W[i]:= W[i] ∪getDefWiki (T2[i]) ;          // Trích xuất định nghĩa của các từ khóa từ Wikipedia
   End For
7: For i:=1 to count(T2) do
       For j:=2 to n do
          T11:=T11∪ separateN-gram(W[i], n);  //N-gram các mẩu tin thu được, n=2,3,4
       End For
   End For
8: T22:=T22 ∪removeStopWord (T11);             //Loại bỏ từ dừng của mẩu tin
9: T3:=T3 ∪ T22;                               // Lưu lại danh sách từ khóa của thuộc tính
10: Ka[i]:=T3;
11: Return Ka[i];
```

Để ước lượng độ tương đồng của hai thuộc tính tương ứng của hai bài viết chúng tôi ước lượng độ tương đồng của hai tập từ khóa vừa tìm được dựa trên một số độ đo tương đồng phổ biến. Việc ước lượng này được chúng tôi kế thừa và cải tiến từ thuật toán mà Takale et al [14] đề xuất nhằm ước lượng độ tương đồng trên hai từ tiếng Anh.

Bước 2: Ước lượng độ tương đồng của các thuộc tính

Sử dụng các độ đo tương đồng phổ biến để ước lượng độ tương đồng của các tập từ khóa thu được, tính độ tương đồng trên thuộc tính của hai bài viết.

Bảng 2. Ước lượng độ tương đồng của thuộc tính

Thuật giải 2: Ước lượng độ tương đồng của thuộc tính, SimFeatures(x,y)
Input: Thuộc tính văn bản thứ i của bài viết a và bài viết b **Ouput**: Độ tương đồng của a[i] và b[i]
1: x:= a[i]; y=b[i]; *2: GetKeyWord (x, z1);* *// Lấy từ khóa theo Wikipedia của a[i]* *3: GetKeyWord (y, z2)* *// Lấy từ khóa theo Wikipedia của b[i]* *4: simFeaturexy:=simJacard (z1,z2);* *// Tính độ tương đồng theo Jarcard theo (5)* *5: Return simFeaturesxy;*

Trong phần thực nghiệm chúng tôi giới hạn xem xét các bài viết gồm có bốn thuộc tính tiêu đề (title), chủ đề (topic), nội dung (content) và các đánh dấu (tags). Do đó, độ tương đồng về ngữ nghĩa dựa trên Wikipedia cho bốn thuộc tính này thể hiện tương ứng với bốn biểu thức sau đây:

$$s_{tag}(i,j) = sim(f_{tag}^{i}, f_{tag}^{j}), \quad (1)$$

$$s_{top}(i,j) = sim(f_{top}^{i}, f_{top}^{j}), \quad (2)$$

$$s_{tit}(i,j) = sim(f_{tit}^{i}, f_{tit}^{j}), \quad (3)$$

$$s_{con}(i,j) = sim(f_{con}^{i}, f_{con}^{j}), \quad (4)$$

Độ đo tương đồng dùng để ước lượng. Trong quá trình phân tích và so sánh độ tương đồng giữa các văn bản, chúng tôi sử dụng độ đo tương đồng Jacard sau đây để ước lượng. Giả sử có hai tập từ khóa: $Kw_1 = \{w_1^1, w_2^1, \dots, w_n^1\}$ và $Kw_2 = \{w_1^2, w_2^2 \dots, w_n^2\}$. Ký hiệu |Kw| là kích thước của tập hợp các từ khóa Kw. Khi đó các độ đo phổ biến được tính như sau:

$$J(w_1, w_2) = \frac{|Kw_1 \cap Kw_2|}{|Kw_1 + Kw_2 - Kw_1 \cap Kw_2|} \quad (5)$$

2. Độ tương đồng về ngữ nghĩa của các bài viết

Để xem xét và ước lượng độ tương đồng về ngữ nghĩa của hai bài viết chúng tôi ước lượng độ tương đồng về ngữ nghĩa dựa trên từng thuộc tính của chúng, sau đó tích hợp có trọng số độ tương đồng trên các thuộc tính để đưa ra độ tương đồng của hai bài viết. Việc ước lượng độ tương đồng về ngữ nghĩa của hai bài viết i và j được định nghĩa tổng quát như sau, giả sử: $i = (f_1^i, f_2^i, \dots, f_n^i), j = (f_1^j, f_2^j, \dots, f_n^j)$ là hai bài viết được biểu diễn bởi các thuộc tính của chúng, khi đó, độ tương đồng của hai bài viết i và j được tính theo công thức sau:

$$s_{entry}(i,j) = \sum_{k=1}^{n} w_k * s_k(i,j) \quad (6)$$

Trong đó, $s_k(i,j)$là độ tương đồng trên thuộc tính k của bài viết i và j, w_k là trọng số của thuộc tính k và:

$$\sum_{k=1}^{n} w_k = 1 \quad (7)$$

Độ tương đồng càng gần đến 1 thì hai bài viết càng có nhiều điểm giống nhau, ngược lại, nếu độ tương đồng càng gần đến 0 thì hai bài viết càng khác nhau. Bộ trọng số chúng tôi sử dụng để tính toán dựa trên nghiên cứu trước của chúng tôi trong Nguyen et al [15], cách xây dựng bộ trọng số thực nghiệm trong [15] là chúng tôi chạy nhiều lần mô hình trên cùng một bộ dữ liệu và thay đổi các trọng số để xem xét số các bài viết ước lượng được so với số bài viết được lựa chọn trên thực tế, dựa trên nghiên cứu này, chúng tôi sử dụng bộ trọng số tương ứng gồm: tiêu đề, nội dung, đánh dấu và chủ đề: $(w_1: w_2: w_3: w_4) = (0.25: 0.35: 0.30: 0.10)$.

Bảng 3. Ước lượng độ tương đồng của hai bài viết

Thuật giải 3: Ước lượng độ tương đồng của hai bài viết, SimEntries(A,B)
Input: Hai bài viết A và B cùng các thuộc tính của chúng **Ouput**: Độ tương đồng của hai bài viết A và B
1: n:=số thuộc tính của bài viết A và bài viết B; *2: For i:=1 to n do* *x[i]:=a[i]; y[i]:=B[i];* *End For* *3: For i:=1 to n do* *simEntries:=w[i]* simFeatures(x[i], y[i]);* *//Trọng số w[i] được tính theo (7)* *End For* *4: Return simEntriesAB;*

III. THỰC NGHIỆM VÀ THẢO LUẬN

A. Xây dựng bộ dữ liệu

Chúng tôi thu thập dữ liệu thực tế từ mạng xã hội Facebook với hạn chế của quản lý trang này, chúng tôi chỉ trích chọn được các bài viết của những người tình nguyện cho phép ứng dụng lưu lại các bài viết trong vòng 1 tháng từ 15/01/2017 đến 15/02/2017. Chúng tôi thu thập được nội dung các bài viết của gần 40 người dùng, với khoảng 200 bài viết, sau đó thực hiện việc xử lý thô bằng cách loại bỏ các bài viết không chứa bất kỳ văn bản nào hoặc các bài viết chỉ có các hình ảnh hoặc các biểu tượng. Cuối cùng, chúng tôi thu được bộ dữ liệu thực nghiệm gồm 150 bài viết có chứa các văn bản có nghĩa. Trong 150 bài viết này chúng tôi thực hiện việc xây dựng tập dữ liệu mẫu như sau:

Bước 1: Chúng tôi thực hiện việc xây dựng dữ liệu mẫu như sau:

- Mỗi một mẫu đều chứa ba bài viết được lựa chọn từ một trong các nguồn 150 bài viết đã xử lý trước, chúng tôi gọi các bài viết tương ứng là A, B và C. Sau đó, chúng tôi hỏi một số người tình nguyện tham gia vào việc lựa chọn để trả lời cho câu hỏi: Trong hai bài viết B và C thì bài viết nào có độ tương tự nhiều với bài viết A hơn?
- Sau đó chúng tôi so sánh số lượng người chọn câu trả lời là B và số lượng người chọn câu trả lời là C. Nếu số lượng người chọn B nhiều hơn chọn C thì giá trị của mẫu này bằng 1. Ngược lại, nếu số lượng người chọn C nhiều hơn B, khi đó giá trị của mẫu được gán bằng 2. Nếu số lượng người chọn B và C ngang nhau, mẫu này sẽ bị loại ra khỏi tập mẫu.

Ví dụ với một mẫu bao gồm 3 bài viết được trích chọn như sau:

Bảng 4. Dữ liệu 3 bài viết được chọn trên Facebook

Bài viết	Tiêu đề (title)	Chủ đề (topic)	Đánh dấu (tag)	Nội dung (content)
A	Dạ Yến Thảo	Trồng hoa	#Dạ yến thảo#	Dạ Yến Thảo có 2 loại: 1 là loại bò lan để trồng vào chậu treo và 2 là loại trồng trực tiếp xuống đất hoặc trồng vào chậu đặt xuống đất.
B	0	Hoa nhà em	#Cẩm tú cầu#	Em cẩm tú cầu này lớn nhanh quá. Năm sau phải di dời mấy em ly ra chỗ khác để cho em ấy lên. Mấy em này cực dễ trồng
C	0	Linh tinh	#Ảnh các loài hoa trong vườn#	Em Azalea hồng nở sáng bừng cả góc vườn. Xem ảnh từ điện thoại thì chuẩn màu mà sao xem từ laptop màu nhạt nhòa thế. 3 em Azalea màu nghệ này mấy năm nữa to thì đẹp tuyệt vời. Em màu tím này thơm cực kỳ luôn

B. Thực thi mô hình

Chúng tôi thực hiện trên dữ liệu với mô hình đề xuất theo 2 bước ở mục 2.C.

- Với mỗi mẫu, chúng tôi sử dụng mô hình đã đề xuất trong bài báo này để ước lượng độ tương tự giữa bài viết B và bài viết A, ước lượng độ tương tự giữa bài viết C và bài viết A.
- Nếu bài viết B có độ tương tự với bài viết A nhiều hơn thì kết quả trả về của mẫu bằng 1. Ngược lại nếu bài viết C tương tự với bài viết A nhiều hơn thì kết quả trả về mẫu bằng 2.
- Sau đó chúng tôi so sánh kết quả và giá trị của mỗi mẫu. Nếu chúng giống với thực tế với lựa chọn ở mục 3.A, thì chúng tôi tăng số lượng độ chính xác của mẫu lên 1

Chúng tôi sử dụng độ chính xác CR (Correct Ratio) được tính toán theo công thức sau:

$$\mathrm{CR} = \frac{\text{Số lượng các mẫu đúng}}{\text{Tổng số các mẫu}} * 100\% \quad (8)$$

Để so sánh hiệu quả của mô hình với kết quả thực tế, trong thời gian ngắn chúng tôi chưa so sánh được hiệu quả của mô hình với các mô hình so sáng ngữ nghĩa khác. Chúng tôi hi vọng kết quả của mô hình có độ chính xác CR càng cao càng tốt. Vì vậy, để đánh giá mô hình, đầu tiên chúng tôi thực thi mô hình với chỉ thuộc tính nội dung, sau đó thực thi mô hình với chỉ thuộc tính đánh dấu, tương tự với các thuộc tính tiêu đề và nhóm, cuối cùng chúng tôi thực thi mô hình với toàn bộ các thuộc tính kết hợp trọng số và thu được độ chính xác như bảng 5.

Bảng 5. Tỷ lệ chính xác CR (%) và trọng số tương ứng của các đặc tính

	Chỉ có tiêu đề (title only)	Chỉ có nội dung (content only)	Chỉ có đánh dấu (tags only)	Chỉ có nhóm (category only)	Tích hợp (Bài viết)
Độ CR trên đặc tính	63%	71%	74%	45%	78%

Nhìn vào dữ liệu thể hiện ở Bảng 5, có thể rằng thấy nếu chỉ xét độ tương đồng về nội dung thì hai bài viết sẽ có độ tương đồng thấp hơn so với khi xem xét thêm các thuộc tính khác của bài viết, như vậy, khi chúng tôi xem xét và ước lượng tích hợp trên nhiều thuộc tính khác nhau của bài viết sẽ cho kết quả tốt hơn là chỉ tính toán và ước lượng trên một thuộc tính của bài viết.

C. Đánh giá mô hình và thảo luận

Có một số bài viết có độ dài hơn 250 từ thì thời gian thực thi khá chậm khi trích chọn các định nghĩa của chúng từ Wikipedia, vì vậy, chúng tôi cải tiến bằng cách lưu toàn bộ định nghĩa các từ khóa tìm thấy lần đầu tiên vào CSDL. Sau đó, khi trích chọn và tìm kiếm các định nghĩa của các từ khóa khác, đầu tiên chúng tôi tìm kiếm trên CSDL, nếu có chúng tôi lấy luôn tập từ khóa của chúng để giảm thời gian truy xuất, chỉ những từ khóa chưa có trong CSDL chúng tôi mới truy xuất trực tiếp trên Wikipedia và sau đó lại lưu chúng bổ sung vào CSDL. Với bước làm này chúng tôi giảm được thời gian truy xuất trực tuyến và mô hình thực thi nhanh hơn đáng kể, các bài viết có độ dài là một câu hoặc một đoạn văn ngắn có thời gian thực thi tốt và các từ khóa thu được nhanh chóng hơn.

Do thời gian hạn hẹp, chúng tôi chưa so sánh được mô hình đề xuất với các mô hình ước lượng theo ngữ nghĩa khác. Chúng tôi sẽ trình bày các kết quả so sánh trong các nghiên cứu tiếp theo của chúng tôi.

IV. KẾT LUẬN

Bài báo của chúng tôi đã đề xuất một mô hình để mô hình hóa các bài viết trên các mạng xã hội và ước lượng độ tương đồng về ngữ nghĩa giữa các bài viết dựa trên từ điển bách khoa toàn thư Wikipedia. Độ tương đồng của các bài viết được xem xét dựa trên kết hợp độ tương đồng của các thuộc tính của bài viết bao gồm: Tiêu đề (title), chủ đề (topic), đánh dấu (tags) và nội dung (content) của bài viết. Kết quả thực nghiệm của chúng tôi đã đánh giá độ tương đồng tích hợp này tốt hơn tính toán độ tương đồng riêng rẽ, tuy nhiên, một số vấn đề cần phải xem xét đầy đủ hơn nữa, chẳng hạn như mô hình yêu cầu trích chọn các mẫu tin trực tuyến từ hệ thống thư viện online Wikipedia nên tốc độ xử lý khá chậm; hoặc do giới hạn và tính riêng tư của dữ liệu thu thập trên các mạng xã hội nên việc tiền xử lý dữ liệu mất rất nhiều thời gian, dữ liệu thu được chưa phong phú, số lượng bài viết trong dữ liệu thực nghiệm còn khá bé, nhiều bài viết không có từ khóa, hoặc các từ không có nghĩa, hoặc chưa được đề cập đến trên Wikipedia. Những vấn đề này sẽ được nghiên cứu và công bố trong những công trình tiếp theo của chúng tôi.

TÀI LIỆU THAM KHẢO

[1] Anagnostopoulos Aris, Kumar Ravi, Mahdian Mohammad, "Influence and Correlation in Social Networks", Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pages 7-15, ACM, New York, USA.

[2] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, "A web search engineer based approach to measure semantic similarity between words", IEEE Trans, On Knowl and Data Eng, 23(7): 977-990, July 2011.

[3] Davide Buscaldi, Joseph Le Roux, Jorge J. Garca Flores, and Adrian Popescu, "Lipnecore: Semantic text similarity using n-grams, wordnet, syntactic analysis, esa and information retrieval based features", 2013.

[4] Dekang Lin, "An information-theoretic definition of similarity", In Proc. 15th International Conf. on Machine Learning, pages 296-304, Morgan Kaufmann, San Francisco, CA, 1998.

[5] Dinh Que Tran and Manh Hung Nguyen, "A mathematical model for semantic similarity measures", South-East Asian Journal of Sciences, 1(1):32-45, 2012.

[6] Gaddam Saidi Reddy and Dr. R. V. Krishnaiah, "A novel similarity measure for clustering categorical data sets", OSR Journal of Computer Engineering (IOSRJCE), 4(6):37-42, 2012.

[7] Jian Xu and Qin Lu, "Computing semantic textual similarity using overlapped senses", In Second Joint Conference on Lexical and Computational Semantics Volume 1, pages 90-95, Atlanta, Georgia, USA, June 2013.

[8] Lushan Han, Abhay L. Kashyap, Tim Finin, James May eld, and Jonathan Weese, "Semantic textual similarity systems", In Second Joint Conference on Lexical and Computational Semantics, Volume 1, pages 44-52, Atlanta, Georgia, USA, June 2013.

[9] Manh Hung Nguyen and Dinh Que Tran, "A semantic similarity measure between sentences", South-East Asian Journal of Sciences, 3(1):63-75, 2014.

[10] Manh Hung Nguyen and Thi Hoi Nguyen, "A general model for similarity measurement between objects", International Journal of Advanced Computer Science and Applications(IJACSA), 6(2):235-239, 2015.

[11] Md Arafat Sultan, Steven Bethard, and Tamara Sumner, "Sentence similarity from word alignment", In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 241-246, Dublin, Ireland, August 2014.

[12] Mihai C. Lintean and Vasile Rus, "Measuring semantic similarity in short texts through greedy pairing and word semantics", Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, Marco Island, Florida, May 23- 25, 2012, AAAI Press, 2012.

[13] Rishi Sayal and V. Vijay Kumar, "A novel similarity measure for clustering categorical data sets", International Journal of Computer Applications, 17(1):25-30, March 2011. Published by Foundation of Computer Science.

[14] Sheetal A Takale, Sushma S Nandgaonkar, "Measuring semantic similarity between words using web documents", International Journal of Advanced Computer Science and Applications (IJACSA), Volume 1, Issue 4, 2010.

[15] Thi Hoi Nguyen, Dinh Que Tran, Gia Manh Dam, and Manh Hung Nguyen, "Multifeature Based Similarity Among Entries on Media Portals", Advances in Information and Communication Technology, Proceedings of the International Conference, ICTA 12 - 2016, Advances in Intelligent Systems and Computing, ISBN 978-3-319-49072-4 (2017).

[16] Zafarani Reza, Abbasi Mohammad Ali, Liu Huan, "Social Media Mining: An Introduction", Cambridge University Press, New York, NY, USA, 2014.

SEMANTIC SIMILARITY AMONG ENTRIES ON SOCIAL NETWORKS BASED ON WIKIPEDIA

Nguyen Thi Hoi, Dam Gia Manh, Tran Dinh Que

ABSTRACT: *In this paper, we describe a model for evaluating semantic similarity among entries on social networks by comparision with information items that are extracted from Wikipedia. Our model has three stages: First, modeling entries on social networks via their features with titles, categories or topics, tags and contents. Second, estimating similarity of each feature based on comparing word by word. Finally, integrating the semantic similarity of entries based on similarities ofthese features by means of Wikipedia encyclopedia. Our experimental results show that the similarity computing based on integration is better than computing with each attribute.*