Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00043

MÔ HÌNH ƯỚC LƯỢNG ĐỘ TƯƠNG TỰ GIỮA CÁC BÀI VIẾT TRÊN CÁC CỔNG THÔNG TIN GIẢI TRÍ

Nguyễn Thị Hội [1], Trần Đình Quế [2], Đàm Gia Mạnh[1], Nguyễn Mạnh Hùng[2,3]
[1] Trường Đại học Thương mại, Hà Nội, Việt Nam
[2] Học viện Công nghệ Bưu chính Viễn thông, Hà Nội, Việt Nam
[3] UMI UMMISCO 209 (IRD/UPMC), Hanoi, Vietnam
hoint2002@gmail.com, tdque@yahoo.com, damgiamanh@gmail.com, nmh.nguyenmanhhung@gmail.com

TÓM TẮT— *Ngày nay, với sự bùng nổ của các cổng thông tin cũng như các phương tiện giải trí và các mạng xã hội, mỗi giây, mỗi phút có rất rất nhiều các bài viết được đăng trên các phương tiện này. Nhiều nhà nghiên cứu và quan tâm đến các phương tiện truyền thông xã hội đã đưa ra một số cách thức để lọc, phân loại, tìm kiếm hoặc đưa ra các bài viết tương tự nhau dựa trên các đoạn văn bản, các mô tả ngắn hoặc một thuộc tính nào đó của bài viết,... Vấn đề đặt ra là làm thế nào để ước lượng được độ tương tự giữa các bài viết được đăng trên các cổng thông tin giải trí đó? Hay làm thế nào để phát hiện được bài viết đang xem xét có độ tương tự cao nhất với một hoặc một số bài viết đã đăng trên các cổng thông tin giải trí đó. Để giải quyết vấn đề này, trong bài viết này chúng tôi đề xuất hai vấn đề: Thứ nhất là mô hình hóa các bài viết được đăng trên một số cổng thông tin giải trí phổ biến hiện nay dựa trên một số thuộc tính của chúng như: tiêu đề bài viết, chủ đề bài viết, các đánh dấu của bài viết, và nội dung của bài viết...; Thứ hai là đề xuất một mô hình ước lượng độ tương tự giữa các bài viết trên các cổng thông tin giải trí dựa trên các thuộc tính đã được mô hình hóa ở theo mô hình đã đề xuất. Cuối cùng chúng tôi thảo luận một số giới hạn của mô hình và các hướng nghiên cứu tiếp theo.*

Từ khóa— *Độ tương tự, bài viết tương tự, phương tiện truyền thông, độ đo tương tự, cổng thông tin giải trí*

I. GIỚI THIỆU

Ngày nay, với sự bùng nổ của các cổng thông tin, các phương tiện giải trí cũng như các mạng xã hội, hàng ngày, hàng giờ có rất nhiều các bài viết được đăng lên các phương tiện này. Với nguồn thông tin khổng lồ và vô cùng phong phú từ các bài viết trên các cổng thông tin giải trí, đây cũng là mảnh đất màu mỡ cho các nhà nghiên cứu, những người quan tâm tìm kiếm các phương pháp, cách thức dùng để lọc, phân loại và tìm kiếm các bài viết trên các cổng thông tin hoặc phương tiện giải trí dựa trên các đoạn văn bản, các mô tả ngắn hoặc tập các đoạn trích chọn từ bài viết, … Vấn đề căn bản là làm thế nào để ước lượng được độ tương tự giữa các bài viết đã đăng trên các phương tiện giải trí này? Nói cách khác là làm thế nào để phát hiện được một bài viết vừa đăng có độ tương tự cao nhất với một hay một số bài viết trong một tập hợp các bài đã đăng trước đó hay không?

Về bài toán phát hiện độ tương tự giữa hai đối tượng đã được khá nhiều nhà nghiên cứu quan tâm và đề xuất phương pháp giải quyết như D. Lin [9] đề xuất một mô hình ước lượng tương tự giữa hai đối tượng dựa trên hướng tiếp cận của lý thuyết thông tin, Say và Kumar [18] lại đề xuất một mô hình phân nhóm dựa trên các tập dữ liệu quan hệ sử dụng các tính chất của phụ thuộc hàm như là các tham số để ước lượng độ tương tự. Reddy và Krishnaiah [17] đề xuất một độ đo tương tự được gọi là độ đo tương tự đa điểm (multi – viewpoint) để phân cụm dựa trên tất cả các mối quan hệ giữa các đối tượng. Nguyen và Nguyen [12] giới thiệu một mô hình tổng quát để ước lượng độ tương tự giữa hai đối tượng dựa trên các thuộc tính của chúng. Trong mô hình trên, độ tương tự trên mỗi thuộc tính được định nghĩa trên các đặc trưng và tính chất khác nhau của các đối tượng.

Một cách tổng quát, một bài viết trên các cổng thông tin giải trí hay mạng xã hội có thể là một video clip, một hình ảnh, một văn bản, hoặc một sự kết hợp của tất cả các nội dung đó. Tuy nhiên, trong bài báo này, chúng tôi chỉ xem xét các bài viết có chứa văn bản còn các bài viết như video, hình ảnh, … không chứa văn bản được bỏ qua trong bài báo. Do đó, bài toán xem xét và ước lượng độ tương tự giữa các bài viết chủ yếu tập trung vào xem xét và ước lượng độ tương tự giữa các văn bản.

Hiện nay trên thế giới cũng như ở Việt Nam đã có rất nhiều nghiên cứu về chủ đề tương tự giữa các văn bản, các nghiên cứu này có thể gom lại vào hai nhóm chính: Nhóm thứ nhất nghiên cứu về độ tương tự dựa trên ngữ nghĩa của văn bản. Trong nhóm này, độ tương tự được so sánh dựa trên độ tương tự về ngữ nghĩa của các văn bản. Các phương pháp so sánh dựa trên hệ thống từ ngữ (WordNet) điển hình như nghiên cứu của Buscaldi et al [4], Han et al [7], Lee et al [8], Marsi et al [11], Oliva et al [15] hoặc so sánh độ tương tự trên các ontology như Agirre et al [1], Nguyen và Tran [13, 21], Novelli và Oliveira [14]. Nhóm thứ hai nghiên cứu độ tương tự của các văn bản dựa trên thống kê. Với hướng này, các văn bản được so sánh dựa trên việc thống kê các từ, các ngữ danh từ, các cấu trúc của các từ, của câu …và/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 [10], Croce et al [5], Finkel et al [6], Lintean và Rus [10], Proisl et al [16], Sarie et al [22], Severyn et al [19], Sultan et al [20], Xu và Lu [23].

Hầu hết các mô hình chỉ xem xét và ước lượng độ tương tự dựa trên nội dung bản thân văn bản của bài viết, cũng đã có một số mô hình xem xét thêm tiêu đề của bài viết. Tuy nhiên, nếu chỉ xem xét bản thân văn bản của bài viết có thể dẫn đến các mô hình có thể bỏ qua các thông tin, các đặc trưng của bài viết như các đánh dấu (tags), các nhóm (category), các tiêu đề (title), các từ khóa (key words) … của bài viết. Một số nhà nghiên cứu đồng ý rằng các loại

thông tin đó có thể được trích chọn từ bản thân văn bản của bài viết, vì vậy, chúng không cần thiết phải đưa vào mô hình ước lượng hoặc cần thống kê chúng khi xem xét. Các kết quả thực nghiệm đã chỉ ra trong bài báo này lại cho thấy rằng, việc lấy các loại thông tin khác nhau trực tiếp trong mô hình được đề xuất có thể làm tăng khả năng tính toán của mô hình trong việc ước lượng độ tương tự giữa các bài viết trên các cổng thông tin giải trí.

Chính vì vậy, trong bài báo này chúng tôi đề xuất hai nội dung chính, thứ nhất là mô hình hóa các bài viết được đăng trên các cổng thông tin giải trí với các thuộc tính của chúng như tiêu đề (title), nhóm (category), đánh dấu (tags), nội dung (content), …, thứ hai là đề xuất mô hình ước lượng độ tương tự giữa các bài viết dựa trên các thuộc tính đã được mô hình hóa. Trong mô hình hóa các bài viết đã được đăng trên các cổng thông tin giải trí, không chỉ bản thân văn bản của bài viết được xem xét và ước lượng mà các thuộc tính khác của bài viết cũng được đưa vào trong mô hình để ước lượng độ tương tự giữa các bài viết. Nói cách khác, một bài viết được đăng được biểu diễn bởi một tập hợp các đặc tính, các thuộc tính này sẽ được dùng để xem xét khi ước lượng độ tương tự của bài viết. Trong mô hình ước lượng độ tương tự giữa các bài viết thì các bài viết sẽ được so sánh độ tương tự trên các thuộc tính của chúng, sau đó sẽ tích hợp các độ tương tự trên các thuộc tính riêng thành độ tương tự tổng quát giữa các bài viết.

Bài báo có cấu trúc như sau: Phần II trình bày mô hình ước lượng độ tương tự giữa các bài viết, phần III trình bày một số kết quả thực nghiệm và thảo luận về kết quả, phần IV kết luận của bài viết và những kế hoạch nghiên cứu tiếp theo.

II. MÔ HÌNH ƯỚC LƯỢNG ĐỘ TƯƠNG TỰ GIỮA CÁC BÀI VIẾT

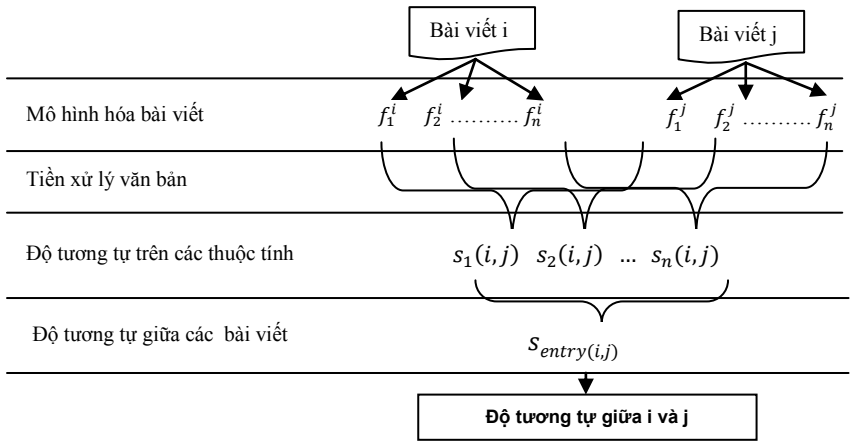
A. Tổng quan về mô hình

1. Giới thiệu mô hình

Đầu vào là hai bài viết i và j, đầu ra là kết quả ước lượng độ tương tự giữa hai bài viết i và j. Mô hình có 4 bước xử lý cơ bản như sau:

- Mô hình hóa các bài viết
- Tiền xử lý các thuộc tính văn bản
- Ước lượng độ tương tự trên các thuộc tính
- Tổng hợp độ tương tự của bài viết dựa trên độ tượng tự của các thuộc tính

Mô hình tổng quát được minh họa như hình 1 sau:



Hình 1. Mô hình ước lượng độ tương tự giữa bài viết i và bài viết j

2. Mô hình hóa các bài viết

Không mất tính tổng quát, chúng ta giả sử rằng:

- Một cổng thông tin giải trí có chứa một tập các bài viết $\{1,2,\dots,m\}$
- Một bài viết được đặc trưng bởi các thuộc tính của chúng như: tiêu đề bài viết (title), nhóm các bài viết (category), các đánh dấu của bài viết (tags) và nội dung của bài viết (content), …

Trong mô hình này, chúng tôi xem xét mỗi bài viết i trong tập các bài viết trên một cổng thông tin giải trí có n thuộc tính, được ký hiệu là $(f_1^i, f_2^i, \dots, f_n^i)$. Trong thực nghiệm, chúng tôi xem xét và ước lượng các thuộc tính của bài viết bao gồm:

- Title hay tiêu đề của bài viết i ký hiệu là f_{tit}^i. Nó có thể là một câu ngắn, chú ý rằng nếu trường hợp bài viết là hình ảnh thì tiêu đề của bài viết được xem xét chính là chú thích của hình ảnh hay caption của hình ảnh nếu bài viết đó không có tiêu đề nào khác
- Content hay nội dung của bài viết i, ký hiệu là f_{con}^i. Một bài viết có thể là một video clip, một hình ảnh, một văn bản hoặc là một sự kết hợp giữa chúng. Tuy nhiên, trong mô hình này chúng tôi chỉ xem xét các thuộc

tính của bài viết là văn bản, các phần khác của bài viết như hình ảnh, clip, … sẽ không được xem xét trong bài báo này. Trong trường hợp nội dung của bài viết không có văn bản chúng tôi coi như không có dữ liệu cho thuộc tính này

- Tags hay các đánh dấu của bài viết, ký hiệu là f_{tag}^{i}. Trên các cổng thông tin giải trí, mỗi bài viết có thể được đánh dấu bởi một tập các đánh dấu. Mỗi đánh dấu là một từ, một ngữ danh từ hay một biểu diễn độc lập
- Category hay nhóm các bài viết, ký hiệu là là f_{cat}^{i}. Trên các cổng thông tin giải trí, mỗi bài viết thường được sắp xếp vào ít nhất một chủ đề hay nhóm cùng loại. Mỗi chủ đề hay nhóm được biểu diễn bởi một từ, một ngữ danh từ độc lập

Như vậy, sau khi được mô hình hóa, một bài viết được đặc trưng bởi một tập các thuộc tính. Trong các thuộc tính của bài viết, bài báo này chỉ xem xét và ước lượng các thuộc tính có chứa văn bản. Dó đó, bài toán ước lượng độ tương tự giữa các bài viết (dựa trên các thuộc tính của chúng) được chuyển thành bài toán ước lượng độ tương tự giữa các văn bản hay tập các biểu diễn bằng văn bản của các bài viết với nhau.

B. Độ tương tự giữa các bài viết

1. Độ tương tự trên mỗi thuộc tính của bài viết

Khi ước lượng độ tương tự giữa các bài viết, chúng tôi chỉ xem xét các thuộc tính có chứa văn bản, vì vậy, cần một số bước để tinh chỉnh và xử lý trước khi so sánh các văn bản với nhau. Để làm được điều đó, chúng tôi phân biệt 2 loại thuộc tính có chứa văn bản như sau:

- Thứ nhất nếu giá trị thuộc tính đã được chuẩn hóa là một tập các biểu diễn như các đánh dấu, các chủ đề hay nhóm bài viết, khi đó, độ tương tự của các thuộc tính này là độ tương tự của tập các biểu diễn này
- Thứ hai nếu giá trị của thuộc tính được thể hiện là các văn bản nói chung như giá trị của thuộc tính nội dung (content) thì độ tương tự của chúng chính là độ tương tự của các văn bản
- Trong trường hợp thuộc tính là tiêu đề (title), thì thông thường các tiêu đề là một câu, bỏ qua sự khác nhau về độ dài ngắn, chúng tôi xem xét thuộc tính này của bài viết như một văn bản

- Trường hợp giá trị thuộc tính là một tập các biểu diễn :

Khi giá trị thuộc tính là một biểu diễn, độ tương tự của chúng là độ tương tự của hai tập các biểu diễn. Chúng tôi định nghĩa độ tương tự giữa hai tập các biểu diễn như sau:

Giả sử rằng $A_1 = (a_1^1, a_1^2, \dots, a_1^m), A_2 = (a_2^1, a_2^2, \dots, a_2^n)$ là hai tập hợp các biểu diễn. Trong đó, m và n là kích thước hay độ dài của A_1 và A_2

Gọi v là kích thước của tập giao của A_1 và A_2, khi đó, độ tương tự giữa A_1 và A_2 được định nghĩa như sau:

$$s_{exp}(A_1, A_2) = \frac{2*|A_1 \cap A_2|}{|A_1|+|A_2|} = \frac{2*v}{m+n} \quad (1)$$

Dễ dàng thấy rằng các giá trị của $s_{exp}(A_1, A_2)$ nằm trong khoảng đơn vị [0,1]. Nghĩa là sau bước này thì tất cả các độ tương tự của hai tập biểu diễn có thể được chuẩn hóa vào khoảng đơn vị. Việc chuẩn hóa này cho phép chúng tôi tránh được các trường hợp ngoại lệ xảy ra như miền giá trị của một số thuộc tính có thể quá lớn hoặc quá bé. Việc chuẩn hóa này cũng được áp dụng cho tất cả các thuộc tính chúng tôi xem xét đối với mỗi bài viết nếu thuộc tính này là một biểu diễn.

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, chúng ta xem xét thuộc tính thứ k của tập các biểu diễn, khi đó độ tượng tự giữa hai bài viết i và j trên thuộc tính thứ k được định nghĩa như sau:

$$s_k(i,j) = s_{exp}(f_k^i, f_k^j), \quad (2)$$

Trong đó f_k^i và f_k^j là giá trị thuộc tính thứ k của hai bài viết tương ứng i và j.

Trong thực nghiệm chúng tôi xem xét 4 thuộc tính của bài viết để ước lượng độ tương tự của các bài viết trong đó có hai thuộc tính là biểu diễn là đánh dấu và nhóm của bài viết. Khi đó độ tương tự của chúng được tính bằng:

$$s_{tag}(i,j) = s_{exp}(f_{tag}^i, f_{tag}^j), \quad (3)$$

$$s_{cat}(i,j) = s_{exp}(f_{cat}^i, f_{cat}^j), \quad (4)$$

- Trường hợp giá trị thuộc tính là một văn bản

Trong trường hợp giá trị thuộc tính của bài viết là một văn bản thì bài toán ước lượng độ tương tự giữa hai thuộc tính được chuyển thành bài toán ước lượng độ tương tự giữa hai văn bản. Với bài toán này có thể áp dụng TF-IDF để phân loại văn bản, có thể sử dụng nhiều mô hình dựa trên phương pháp thống kê cho bài toán này như Bollegala et al[2], Buscaldi et al [3], Croce et al [5], Finkel et al [6], …Trong mô hình này TF-IDF cũng được dùng để tối ưu hóa độ tương tự giữa hai thuộc tính của văn bản như sau:

- Trích chọn giá trị thuộc tính (là văn bản) vào một tập các

$$n-gram, t^1 = (g_1^1, g_2^1, \dots, g_n^1) \text{ và } t^2 = (g_1^2, g_2^2, \dots, g_n^2)$$

- Tính toán TF-IDF của mỗi $n-gram$ trong văn bản. Sau đó biểu diễn giá trị thuộc tính bằng một véc tơ với mỗi thành phần là một cặp:

$$< n-gram, td-id >: v^1 = (< g_1^1, v_1^1 >, (< g_2^1, v_2^1 >, \dots, < g_n^1, v_n^1 >) \text{ và}$$
$$v^2 = (< g_1^2, v_1^2 >, (< g_2^2, v_2^2 >, \dots, < g_m^2, v_m^2 >)$$

- Tính toán khoảng cách giữa hai véc tơ: $D(v^1, v^2) = \frac{1}{N}\sum_1^N d_k$, (5) trong đó, N là số lượng các $n-gram$ khác nhau được xem xét trong $t^1 \cup t^2$, d_k là khoảng cách đến mỗi $< g_i^1, v_i^1 >$ của v^1 hoặc $< g_j^2, v_j^2 >$ của v^2

- Nếu có thành phần $< g_l^2, v_l^2 >$ của v^2 hoặc thành phần $< g_l^1, v_l^1 >$ của v^1 mà có $g_l^2 = g_l^1$ thì khi đó

$$d_k = \frac{|v_l^1 - v_l^2|}{\max(v_l^1, v_l^2)}, \quad (6)$$

- Các trường hợp khác thì $d_k = 1$

Dễ dàng thấy rằng giá trị của $D(v^1, v^2)$ nằm trong khoảng [0,1]. Độ tương tự giữa hai thuộc tính là:

$$s_{txt}(t^1, t^2) = 1 - D(v^1, v^2) \quad (7)$$

Trong thực nghiệm của chúng tôi, độ tương tự giữa hai thuộc tính tiêu đề và nội dung của bài viết i và bài viết j tương ứng là:

$$s_{tit}(i,j) = s_{txt}(f_{tit}^i, f_{tit}^j) \quad (8)$$

$$s_{con}(i,j) = s_{txt}(f_{con}^i, f_{con}^j) \quad (9)$$

2. Độ tương tự giữa hai bài viết

Để ước lượng độ tương tự giữa hai bài viết dựa trên độ tương tự của các thuộc tính của các bài viết đã được tính toán ở phần II.B.1. Việc ước lượng độ tương tự của hai bài viết i và bài viết j được định nghĩa 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 tự của hai bài viết i và j được tính toán theo công thức sau:

$$s_{entry}(i,j) = \sum_{k=1}^n w_k * s_k(i,j) \quad (10)$$

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

$$\sum_{k=1}^n w_k = 1 \quad (11)$$

Độ tương tự càng gần đến 1 thì hai bài viết càng giống nhau. Ngược lại, nếu độ tương tự càng gần đến 0 thì hai bài viết càng khác nhau.

III. THỰC NGHIỆM VÀ ĐÁNH GIÁ

A. Phương pháp thực hiện

Bước 1: Xây dựng tập dữ liệu mẫu. 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 như Youtube, CNN, News, … Các bài viết này được gọi lần lượt là A, B và C
- Chúng tôi hỏi một số người được lựa chọn để trả lời cho câu hỏi: Giữa bài viết B và C thì bài viết nào tương tự nhiều hơn với bài viết A?
- Sau đó chúng tôi so sánh số lượng người chọn B và số lượng người chọn 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 1. Dữ liệu về 3 bài viết được trích chọn trên Youtube

Bài viết	Tiêu đề (title)	Nhóm (category)	Đánh dấu (tag)	Nội dung (content)
A	Top 30 Goals World Cup 2014	Sports	Worldcup, Football	no text
B	Top 10 Goals: 2014 FIFA World Cup Brazil [Official]	Sports	Worldcup, Football, Brazil, FIFA	no text
C	The Speech that Made Obama President	Education	Obama, President speech	no text

Để so sánh bài viết A với hai bài viết còn lại, chúng tôi đã hỏi một nhóm 9 người tình nguyện tham gia cuộc khảo sát của chúng tôi: Câu hỏi là: So sánh giữa hai bài viết B và C thì bài viết nào có độ tương tự nhiều hơn với bài viết A? Và kết quả được trình bày trong bảng 2

Bảng 2. Dữ liệu được chọn của người dùng từ các bài viết trên Youtube

Câu hỏi	Đa số chọn	Thiểu số chọn
1	9 (cho bài viết B)	0 (cho bài viết C)

Từ kết quả này có thể thấy rằng bài viết B và bài viết A có độ tương tự cao hơn so với bài viết C và bài viết A. Do đó, giá trị của mẫu này được đặt là 1

Sau bước này chúng tôi có một tập các mẫu. Chúng tôi cũng dùng một số nguồn của các mẫu khác và lưu chúng trong một số tập mẫu . Trong quá trình thực nghiệm, chúng tôi lấy mẫu từ 3 nguồn, và các tập mẫu được mô tả trong bảng 3.

Bảng 3. Cấu trúc của 3 tập mẫu

Nguồn	Số lượng các mẫu
CNN News	100
Fox News	100
YouTube	100
Tổng	**300**

Bước 2: Cách thực thi mô hình:

- 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, và ước lượng độ tương tự giữa bài viết A và bài viết C
- Nếu bài viết B có độ tương tự nhiều hơn với bài viết A thì kết quả trả về của mẫu bằng 1. Ngược lại nếu bài viết C tương tự nhiều hơn với bài viết A thì kết quả trả về mẫu bằng 2
- Sau đó chúng tôi so sánh kết quả và giá trị của mỗi mẫu. Nếu chúng được xác định, thì chúng tôi tăng số lượng độ chính xác của mẫu lên 1

Bước 3: Phương pháp đánh giá kết quả mô hình

Độ chính xác CR (Correct Ratio) của mô hình trên các mẫu đã lấy được tính toán theo công thức sau:

$$CR = \frac{Số\ lượng\ các\ mẫu\ đúng}{Tổng\ số\ các\ mẫu} * 100\% \quad (12)$$

Độ chính xác CR càng gần đến 100% thì mô hình được đề xuất càng chính xác. Chúng tôi hi vọng kết quả của mô hình có độ chính xác CR càng cao càng tốt.

Tính toán và lựa chọn trọng số tốt nhất cho mỗi thuộc tính của bài viết

Các bài viết trước khi ước lượng độ tương tự cần được xác định trọng số tốt nhất của mỗi thuộc tính của chúng, theo mô hình đề xuất ở II.A.2, các bài viết trên các cổng thông tin giải trí có 4 thuộc tính là tiêu đề, nhóm, đánh dấu và nội dung thì ta đặt các trọng số của các thuộc tính tương ứng là: (w_1, w_2, w_3, w_4). Vì thế kịch bản để tính toán và lựa chọn trọng số của các thuộc tính của bài viết được thực hiện như sau:

- Kiểm tra tất cả các mẫu một lần và đặt các thuộc tính tiêu đề (title), nội dung (content), đánh dấu (tags) và nhóm (category) của bài viết có trọng số cho mỗi thuộc tính là 1, các thuộc tính không được xem xét thì đặt bằng 0. Tính toán độ chính xác CR
- Càng nhiều thuộc tính đơn thì độ chính xác CR ta thu được càng cao, và khi đó độ quan trọng của thuộc tính đó trong mô hình cũng cao hơn các thuộc tính khác

Kết quả của thực nghiệm được trình bày trong bảng 4. Trọng số của các thuộc tính của mỗi bài viết trên các cổng thông tin giải trí đã thu được là: $(w_1 : w_2 : w_3 : w_4) = (0.25 : 0.34 : 0.29 : 0.12)$. Vì thế chúng tôi chọn trọng số $(0.25 : 0.35 : 0.30 : 0.10)$. cho tất cả các lần thực hiện thực nghiệm mô hình để ước lượng độ tương tự của các bài viết trên các cổng thông tin giải trí

Bảng 4. Tỷ lệ chính xác CR (%) và trọng số tương ứng của các đặc tính

Nguồn	Chỉ có tiêu đề (title only)	Chỉ có nội dung (content only)	Chỉ có đánh dấu (tags only)	Chỉ có nhóm (category only)
CNN News	69	74	77	31
Fox News	32	82	62	31
YouTube	72	-	62	26
Độ CR trung bình	**57.67**	**78.00**	**67.00**	**29.33**
Trọng số chuẩn hóa	**0.25**	**0.34**	**0.29**	**0.12**

B. Thảo luận về kết quả mô hình

Trong phần này chúng tôi thảo luận về giới hạn về nguồn dữ liệu của mô hình đề xuất

1. Những giới hạn về nội tại của nguồn dữ liệu

Với kết quả thực nghiệm thu được ở mục III.A có thể dễ dàng thấy rằng thuộc tính nhóm (category) của cả ba nguồn dữ liệu không có nhiều hỗ trợ tốt trong việc phân biệt giữa các bài viết. Sau khi xem xét lại dữ liệu từ các ba nguồn, chúng tôi phát hiện ra nguyên nhân đó là: Trong tất cả ba nguồn dữ liệu, mỗi bài viết chỉ được nhóm vào chỉ duy nhất một nhóm. Và có một số bài viết được nhóm vào các nhóm không liên quan đến chủ đề nhiều như chủ ý của bài viết. Ví dụ như, từ dữ liệu trên Youtube với bài viết "*50 Most shocking moments in World Cup history*" được xếp vào mục Entertainment (Giải trí). Hoặc bài viết "*Germany Argentina 2014 World Cup Final Full Game ESPN*" lại được nhóm vào nhóm People & Blogs. Trong khi đó cả hai bài này đáng lẽ cần được nhóm vào mục Sports thì hợp lí hơn.

Trong tình huống này một câu hỏi đặt ra là: *Thuộc tính nhóm (category) có nên sử dụng trong mô hình hay không*? Để trả lời cho câu hỏi này, chúng tôi làm một thực nghiệm nhỏ như sau: Lần đầu tiên, chúng tôi chạy mô hình mà không sử dụng đến thuộc tính nhóm (category) (nghĩa là chỉ chạy mô hình với ba thuộc tính là tiêu đề (title), nội dung (content), và đánh dấu (tags)) trên cả 300 mẫu dữ liệu. Lần thứ hai, chúng tôi chạy mô hình với đầy đủ các thuộc tính (nghĩa là chạy mô hình với cả 4 đặc tính). Và kết quả thu được thật đáng ngạc nhiên, kết quả của lần thứ nhất độ chính xác trung bình là 87.00% và kết quả của lần thứ hai là 92.67%. Do đó, câu trả lời ở đây là thuộc tính nhóm cũng đóng góp quan trọng trong mô hình. Đem lại độ chính xác cao hơn khi phân biệt sự tương tự giữa các bài viết.

Câu hỏi tương tự cũng được đặt ra cho thuộc tính tiêu đề (title) của nguồn dữ liệu trên Fox News. Nó cũng có vẻ như không có đóng góp tốt trong việc phân biệt sự khác nhau giữa các bài viết. Chúng tôi cũng làm một thực nghiệm nhỏ đối với mô hình. Lần đầu cũng thực hiện chạy mô hình và bỏ qua thuộc tính tiêu đề trên các nhóm dữ liệu. Bước thứ hai, chúng tôi chạy mô hình với đầy đủ các tính năng trên cả 300 mẫu dữ liệu. Kết quả là độ chính xác trong trường hợp đầu tương ứng là 92.00% trên CNN News, 96.00% trên Fox News và 71.00% trên Youtube, kết quả độ chính xác trung bình là 86.33%. Trong khi đó, khi chạy lần hai với đầy đủ các thuộc tính thì kết quả tương ứng của độ chính xác lại là 96.00% trên CNN News, 96.00% trên Fox News và 86.00% trên Youtube, kết quả độ chính xác trung bình là 92.67%. Kết quả này cho thấy rằng thuộc tính tiêu đề (title) có thể không quan trọng trên nguồn dữ liệu của Fox News nhưng trên các nhóm dữ liệu từ nguồn khác nó lại có đóng góp đáng kể trong phân biệt các bài viết. Vì vậy, câu trả lời là thuộc tính tiêu đề cũng cần được đưa vào để xem xét và ước lượng trong mô hình.

2. Những giới hạn về mô hình

Như đã xác định ở phần I. phần Giới thiệu, mô hình của chúng tôi chỉ dựa trên cú pháp của văn bản, việc ước lượng độ tương tự trong trường hợp có hai biểu diễn giống nhau về ngữ nghĩa nhưng khác nhau về cú pháp có thể gây ra kết quả không phù hợp

Ví dụ:

Bảng 5. Ba bài viết được trích chọn từ Fox News

Bài viết	Tiêu đề	Nhóm	Đánh dấu	Nội dung
1	Facebook testing digital stores within site as part of e-zommerce push	Facebook	Facebook	…
2	Twitter to lift 140-character DM limit in bid to compete with messaging apps	Twitter	Twitter	…
3	7 of the weirdest sculpture parks in the world	Extreme travel	park, sculpture	…

Với ví dụ trong bảng 5 này ta có thể thấy rằng, bài thứ nhất và bài thứ hai có thể không tương tự nhau theo mô hình của chúng tôi. Nhưng trên thực tế chúng nói về Facebook và Twitter, cả hai nhóm là hai mạng xã hội và chúng có miền giá trị chung là công nghệ, công nghệ thông tin,... Nhìn vào thì có thể thấy bài viết số 1 và bài viết số 2 có độ tương tự không lớn hơn bài thứ 1 và bài thứ 3. Trong khi trên thực tế thì bài số 1 và bài số 2 lại có độ tương tự nhiều hơn bài số 1 và bài số 3.

Tóm lại, các trường hợp ngoại lệ và những giới hạn của mô hình chúng tôi đã đưa ra hết trong mục III.B.1 và III.B.2 này. Các ngoại lệ và mô hình chạy không đúng đều xảy ra khi các biểu diễn có sự tương đồng về ngữ nghĩa. Hướng tiếp cận này chúng tôi sẽ nghiên cứu tiếp trong thời gian tới để bổ trợ cho những hạn chế còn tồn tại của mô hình dựa trên thống kê chúng tôi đã đề xuất trong bài báo này.

IV. KẾT LUẬN

Bài báo này đã đề xuất một mô hình đề mô hình hóa các bài viết được đăng trên các cổng thông tin giải trí và mạng xã hội và ước lượng độ tương tự giữa các bài viết đã đăng. Độ tượng tự của các bài viết được xem xét dựa trên các thuộc tính của bài viết bao gồm: tiêu đề (title), nhóm (category), đánh dấu (tags), và nội dung (content). Mô hình có thể sử dụng để phát hiện hay phân loại một bài viết có tương tự hay khác biệt với một tập các bài đã đăng trên các cổng thông tin giải trí hoặc mạng xã hội, hoặc tìm xem bài viết nào có độ tương tự lớn nhất với bài viết đang được xem xét. Mô hình cũng có thể sử dụng để phân loại tự động các bài viết trên một số cổng thông tin giải trí và mạng xã hội phổ biến hiện nay.

Mô hình trong bài viết cũng được kiểm định lại bằng thực nghiệm và cho kết quả tốt gần giống với việc phân loại, lựa chọn của con người trên cùng một tập mẫu dữ liệu. Tuy nhiên vẫn còn một số vấn đề với mô hình hiện tại như làm thế nào để so sánh ngữ nghĩa của văn bản trong các biểu diễn của dữ liệu, làm thế nào để cải thiện được tốc độ xử lý của mô hình, … Đây sẽ là những hướng nghiên cứu tiếp theo của chúng tôi trong tương lai gần

TÀI LIỆU THAM KHẢO

[1] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. Semantic textual similarity. (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 32- 43, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[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, Paolo Rosso, Jose Manuel Gomez-Soriano, and Emilio Sanchis. Answering questions with an n-gram based passage retrieval engine. Journal of Intelligent Information Systems, 34(2):113-134, 2010.

[4] 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.

[5] Danilo Croce, Valerio Storch, and Roberto Basili. Combining text similarity and semantic Filters through sv regression. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 59-65, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[6] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 363-370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[7] 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 (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 44-52, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[8] Ming Che Lee, Jia Wei Chang, and Tung Cheng Hsieh. A grammar-based semantic similarity algorithm for natural language sentences. The Scientific World Journal, 2014:17 pages, 2014.

[9] 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.

[10] Mihai C. Lintean and Vasile Rus. Measuring semantic similarity in short texts through greedy pairing and word semantics. In G. Michael Youngblood and Philip M. McCarthy, editors, Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, Marco Island, Florida. May 23- 25, 2012. AAAI Press, 2012.

[11] Erwin Marsi, Hans Moen, Lars Bungum, Gleb Sizov, Bjorn Gamback, and Andre Lynum. Combining strong features for semantic similarity. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 66-73, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[12] 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.

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

[14] Andreia Dal Ponte Novelli and Jose Maria Parente De Oliveira. Article: A method for measuring semantic similarity of documents. International Journal of Computer Applications, 60(7):17-22, December 2012.

[15] Jess Oliva, Jos Ignacio Serrano, Mara Dolores del Castillo, and ngel Iglesias. Symss: A syntax-based measure for short-text semantic similarity. Data & Knowledge Engineering, 70(4):390-405, 2011.

[16] Thomas Proisl, Stefan Evert, Paul Greiner, and Besim Kabashi. Robust semantic similarity at multiple levels using maximum weight matching. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 532-540, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

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

[18] 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.

[19] Aliaksei Severyn, Massimo Nicosia, and Alessandro Moschitti. Tree kernel learning for textual similarity. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 53-58, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[20] 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. Association for Computational Linguistics and Dublin City University.

[21] 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.

[22] Frane Saric, Goran Glavas, Mladen Karan, Jan Snajder, and Bojana Dalbelo Basic. Takelab: Systems for measuring semantic text similarity. In Proceedings of the First Joint Conference on Lexical and Computational Semantics- Volume 1: Proceedings

of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, SemEval '12, pages 441- 448, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[23] Jian Xu and Qin Lu. Computing semantic textual similarity using overlapped senses. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 90-95, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

MULTI FEATURES-BASED SIMILARITY AMONG ENTRIES ON MEDIA PORTALS

Thi Hoi Nguyen, Dinh Que Tran, Gia Manh Dam, and Manh Hung Nguyen

ABSTRACT— *Nowadays, with the exploration of entertainment, news or media portals and the social networks, there is a huge number of entries posted on these portals. This raises several issues to filter, classify, and/or search for entries which are similar to a given text, a short description, or a selected entry, etc. The core basic problem of these issues is how to measure the similarity among the entries posted on the mentioned portals: with a given entry, and a set of entries to consider, how to detect the entry in the considered set which is the most similar to the given entry. This paper firstly models the entries on posted on media or entertainment portals based on their features such as title, category, tags, and content, etc. And secondly it presents a model for estimating the similarity among these entries.*