Nghiên cứu phương pháp so sánh độ tương đồng văn bản bằng độ đo cosine

Phạm Thị Hải Vân*, Phạm Hữu Lợi

Trường Đại học Mỏ - Địa chất

Ngày nhận bài 12/9/2016, ngày chuyển phản biện 15/9/2016, ngày nhận phản biện 15/11/2016, ngày chấp nhận đăng 25/11/2016

Hiện tại đã có một số giải pháp cho việc phát hiện sao chép và một vài công cụ phần mềm cho phép phát hiện một tài liệu (gọi là văn bản kiểm tra) có sao chép từ một tập hợp các tài liệu nguồn hay không. Các phương pháp này chủ yếu dựa trên tìm kiếm và so khớp chuỗi, chỉ thực sự có hiệu quả nếu việc sao chép là "nguyên văn". Bài báo đề xuất phương pháp so sánh độ tương đồng văn bản bằng độ đo cosine, là một phương pháp hiệu quả để phát hiện việc sao chép văn bản, góp phần hạn chế tình trạng xâm phạm quyền sở hữu trí tuệ về công bố khoa học hiện nay.

***Từ khóa:* độ đo cosine, tương đồng, văn bản.**

***Chỉ số phân loại* 1.2**

A research on the text comparison method using cosine similarity

Summary

Currently, there are a number of solutions for detection of copy and a few software tools that allow detecting whether a document (called writing checks) has copied from a set of source materials or not. These methods are mainly based on the string search and matching, only really effective if the copy is "verbatim". This paper proposes the method of text similarity comparison with cosine measure, an effective method to detect copying document, contributing to limit the infringement of intellectual property rights on scientific publications at present.

***Keywords:* cosine measure, similarities, text.**

***Classification number* 1.2**

Mở đầu

Trong thời đại công nghệ số như hiện nay, các nguồn tài liệu là vô cùng phong phú. Việc "sao chép tài liệu" theo nghĩa tiêu cực như đạo văn, sao chép các luận án, luận văn, đồ án trở nên phổ biến và là một vấn nạn. Làm thế nào để phát hiện sự sao chép tài liệu theo nghĩa tiêu cực? Làm thế nào ngăn chặn việc sao chép trái phép, đạo văn, đạo nhạc, đạo luận văn, đồ án? Chủ đề này đã được nghiên cứu từ hơn 10 năm qua. Hiện tại, đã có một số giải pháp cho việc phát hiện sao chép và một vài công cụ phần mềm cho phép phát hiện một tài liệu (gọi là văn bản kiểm tra) có sao chép từ một tập hợp các tài liệu nguồn hay không [1, 2]. Đã có một số nghiên cứu đề xuất các phương pháp khác nhau để xác định xem một đoạn văn bản của một số tài liệu có nằm trong một tài liệu nào đó hay không. Các phương pháp này chủ yếu dựa trên tìm kiếm và so khớp chuỗi (string matching). Tuy nhiên, các phương pháp so khớp chuỗi chỉ có hiệu quả nếu việc sao chép là "nguyên văn". Nó không thể phát hiện các sao chép có sửa đổi đôi chút như thay thế một số từ bằng từ đồng nghĩa hay thay đổi thứ tự các câu trong văn bản. Phương pháp độ đo cosine được sử dụng trong bài toán so sánh độ tương đồng văn bản, trong lĩnh vực xử lý ngôn ngữ tự nhiên và có nhiều kết quả khả quan. Với các văn bản tiếng Việt đặt dấu cách giữa các âm tiết chứ không phải giữa các từ. Một từ có thể có 1, 2 hoặc nhiều âm tiết nên có nhiều cách phân chia các âm tiết thành các từ, gây ra nhập nhằng. Vì vậy để so sánh độ tương đồng giữa 2 văn bản tiếng Việt sẽ khó khăn hơn so với các văn bản bằng tiếng ngước ngoài. Tiếng Việt là một sinh ngữ, có sự biến đổi, các từ mới thuần Việt cũng như từ vay mượn được tạo ra thường xuyên. Nếu một ứng dụng không xử lý được vấn đề này thì hiệu năng sẽ bị giảm dần theo thời gian. Do đó, vấn đề nghiên cứu trong

*Tác giả liên hệ: Email: phamhaivan1979@gmail.com

bài báo chỉ tập trung vào so sánh độ tương đồng giữa 2 văn bản bằng tiếng Việt.

Cơ sở lý thuyết và phương pháp nghiên cứu

Độ tương đồng

Trong toán học, một độ đo là một hàm số cho tương ứng với một "chiều dài", một "thể tích" hoặc một "xác suất" với một phần nào đó của một tập hợp cho sẵn, là một khái niệm quan trọng trong giải tích và trong lý thuyết xác suất. Ví dụ, độ đo đếm được định nghĩa bởi μ(S) = số phần tử của S. Tuy nhiên, rất khó để đo sự giống nhau, sự tương đồng. Sự tương đồng là một đại lượng (con số) phản ánh cường độ của mối quan hệ giữa 2 đối tượng hoặc 2 đặc trưng. Đại lượng này thường ở trong phạm vi từ -1 đến 1 hoặc từ 0 đến 1. Như vậy, độ đo tương đồng có thể coi là một loại scoring function (hàm tính điểm). Ví dụ, trong mô hình không gian véc-tơ, sử dụng độ đo cosine để tính độ tương đồng giữa 2 văn bản, mỗi văn bản được biểu diễn bởi một véc-tơ.

Độ tương đồng văn bản

Phát biểu bài toán tính độ tương đồng văn bản như sau: xét một tài liệu d gồm có n câu: d = s_1, s_2,… s_n. Mục tiêu của bài toán là tìm ra một giá trị của hàm S(s_i, s_j) với S (0,1), và i, j = 1,… n. Hàm S(s_i, s_j) được gọi là độ đo tương đồng giữa 2 văn bản s_i và s_j. Giá trị càng cao thì sự giống nhau về nghĩa của 2 văn bản càng nhiều [3, 4].

Ví dụ, xét 2 câu "Tôi là nam" và "Tôi là nữ", bằng trực giác có thể thấy rằng 2 câu trên có sự tương đồng khá cao.

Độ tương đồng ngữ nghĩa là một giá trị tin cậy phản ánh mối quan hệ ngữ nghĩa giữa 2 câu. Trên thực tế, khó có thể lấy một giá trị chính xác bởi vì ngữ nghĩa chỉ được hiểu đầy đủ trong một ngữ cảnh cụ thể [5].

Thuật toán tách từ (Tokenizer) [2]

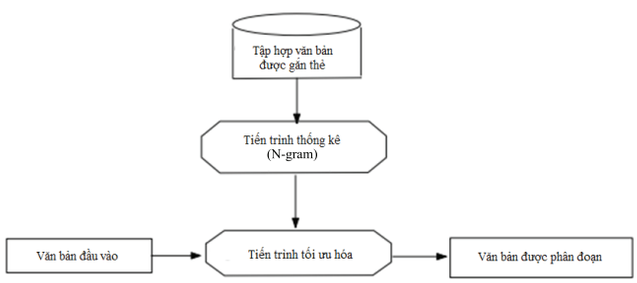
Tiếng Việt không sử dụng các hình thái (morpheme) để tạo ra các ý nghĩa của từ (trong tiếng Anh, cup => cups thì 's' là hình thái số nhiều). Vì thế trong tiếng Việt, các từ không bị thay đổi. Để tạo ra các sắc thái ý nghĩa khác nhau, tiếng Việt phụ thuộc vào trật tự của từ.

Ví dụ, với 5 âm tiết (năm từ đơn): "sao, Tuấn, bảo, không, đến" khi sắp xếp theo các trật tự khác nhau sẽ cho ra các nghĩa khác nhau. Cụ thể như: "Sao Tuấn bảo đến không?" "Sao Tuấn không đến bảo?", "Sao? Tuấn bảo đến không?", "Sao Tuấn đến không bảo?", "Sao bảo Tuấn không đến?", "Sao? bảo Tuấn đến không?", "Sao bảo không đến, Tuấn?", "Sao? đến bảo Tuấn không?", "Sao đến không bảo Tuấn?", "Sao đến Tuấn bảo không?", "Sao không đến bảo Tuấn?", "Sao không bảo Tuấn đến?".

Mặt khác, dấu cách (space) trong tiếng Việt không có tác dụng phân tách các từ như tiếng Anh mà chỉ có tác dụng phân tách các âm tiết. Vì thế, việc phân tách từ trong tiếng Việt là một việc không thể thiếu trong so sánh độ tương đồng văn bản. Tách từ là một quá trình xử lý nhằm mục đích xác định ranh giới của các từ trong câu văn, cũng có thể hiểu đơn giản rằng tách từ là quá trình xác định các từ đơn, từ ghép… có trong câu. Trong các mô hình được sử dụng để tách từ có mô hình N-gram là mô hình thể hiện mối quan hệ khá tốt theo ngữ cảnh của từ. Trong mô hình đó, mỗi từ thứ n được coi như phụ thuộc xác suất vào (n-1) từ trước nó.

$$P(W) = P(w_1 w_2 ... w_n) = \prod_{i=1}^{n} P(w_i | w_{i-n+1} ... w_{i-1}).$$

Mô hình N-gram được ứng dụng để phân đoạn từ trong đó với mỗi câu thì phân đoạn tốt nhất theo mô hình này là phân đoạn có xác suất P(W) được tính theo công thức là lớn nhất. Trong đó, các xác suất về sự phụ thuộc của một từ vào n từ trước đó được thống kê dựa trên một tập hợp văn bản, ngôn ngữ đã được số hóa (corpus) đủ lớn. Tùy vào giả thiết về tính phụ thuộc mà sẽ có các mô hình 2-gram hoặc 3-gram tương ứng. Phương pháp này là một trong những phương pháp thống kê chính để giải quyết bài toán phân đoạn từ khi không có thông tin từ điển và dữ liệu gán nhãn. Mô hình phân đoạn từ sử dụng N-gram được biểu diễn như hình 1.



Hình 1: mô hình phân đoạn từ sử dụng N-gram

Ví dụ, Maosong Sun và cộng sự [6] đã sử dụng độ đo mutal information và t-scores và một số kỹ thuật khác để xác định từ cho tiếng Trung và đã thu được kết quả khá cao (> 90%). Đối với tiếng Việt, tác giả Lê An

Hà [7] đơn giản sử dụng tần suất N-gram để tối ưu xác suất của mỗi chunk. Kết quả thực nghiệm tuy không cao nhưng đã chứng tỏ rằng N-gram là một phương pháp phù hợp có thể ứng dụng cho bài toán phân đoạn từ tiếng Việt nói riêng.

Phương pháp tính độ tương đồng văn bản cosine

Trong phương pháp này, các văn bản sẽ được biểu diễn theo mô hình không gian véc-tơ [8, 9]. Mỗi thành phần trong véc-tơ chỉ đến một từ tương ứng trong danh sách mục từ chính. Danh sách mục từ chính thu được từ quá trình tiền xử lý văn bản đầu vào, các bước tiền xử lý gồm: tách câu, tách từ, gán nhãn từ loại, loại bỏ những câu không hợp lệ (không phải là câu thực sự) và biểu diễn câu trên không gian véc-tơ [10, 11].

Không gian véc-tơ có kích thước bằng số mục từ trong danh sách mục từ chính. Mỗi phần tử là độ quan trọng của mục từ tương ứng trong câu. Độ quan trọng của từ thứ i được tính bằng công thức sau:

$$w_{i,j} = \frac{tf_{i,j}}{\sqrt{\sum_j tf_{i,j}^2}} \quad (1)$$

Trong đó, $tf_{i,j}$ là tần số xuất hiện của từ thứ i trong câu thứ j.

Với không gian biểu diễn tài liệu được chọn là không gian véc-tơ và trọng số TF (term frequency), độ đo tương đồng được chọn là cosine của góc giữa 2 véc-tơ tương ứng của 2 văn bản S_i và S_k. Véc-tơ biểu diễn 2 văn bản lần lượt có dạng:

$S_i = <W_1^i, \ldots, W_t^i>$, với W_t^i là trọng số của từ thứ t trong văn bản i

$S_k = <W_1^k, \ldots, W_t^k>$, với W_t^k là trọng số của từ thứ t trong văn bản k

Độ tương tự giữa chúng được tính theo công thức như sau:

$$Sim(S_i, S_k) = \frac{\sum_{k=1}^{t} W_k^i W_k^k}{\sqrt{\sum_{k=1}^{t} (W_k^i)^2 . \sum_{k=1}^{t} (W_k^k)^2}} \quad (2)$$

Trên các véc-tơ biểu diễn cho các văn bản lúc này chưa xét đến các quan hệ ngữ nghĩa giữa các mục từ, do đó các từ đồng nghĩa sẽ không được phát hiện, dẫn đến kết quả xét độ tương tự giữa các câu chưa tốt. Ví dụ như cho 2 văn bản ngắn như sau:

S_1: Cần trao đổi ý kiến trước khi lấy biểu quyết

S_2: Hội đàm đã diễn ra trong bầu không khí thân mật và hiểu biết lẫn nhau.

Nếu không xét đến quan hệ ngữ nghĩa giữa các từ thì 2 văn bản trên không có mối liên hệ gì cả và độ tương đồng bằng 0. Nhưng thực chất, 2 câu trên đều nói về hoạt động hội họp, trao đổi ý kiến, do đó giữa 2 văn bản có một sự liên quan nhất định và với công thức tính độ tương tự như trên thì độ tương tự giữa 2 câu này phải khác 0.

Công thức tính tương đồng cosine (cosine similarity):

$$cos(\vec{q}, \vec{d}) = \frac{\vec{q}.\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|} . \frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|v|} q_i d_i}{\sqrt{\sum_{i=1}^{|v|} q_i^2} \sqrt{\sum_{i=1}^{|v|} d_i^2}} \quad (3)$$

Trong đó, q_i là trọng số tf-idf (term frequency - inverse document frequency) của từ i trong văn bản truy vấn; d_i là trọng số tf-idf của văn bản trong tập tài liệu; cos $(\vec{q},\vec{d})$ là sự tương đồng cosine giữa $\vec{q}$ và $\vec{d}$ hay là cosine giữa $\vec{q}$ và $\vec{d}$.

Đối với những véc-tơ đã được chuẩn hóa về độ dài, sự tương đồng cosine chỉ đơn giản là tích vô hướng của 2 véc-tơ (scalar product).

$$cos(\vec{q}, \vec{d}) = \vec{q}.\vec{d} = \sum_{i=1}^{|v|} q_i d_i \quad (4)$$

Mục đích là cung cấp định nghĩa chính thức về khái niệm độ tương tự, đầu tiên là trực giác về độ tương tự:

- Trực giác 1: độ tương tự giữa A và B có liên quan đến sự tương đồng của chúng. Sự tương đồng càng nhiều, độ tương tự càng lớn.

- Trực giác 2: độ tương tự giữa A và B liên quan đến những sự khác biệt giữa chúng. Càng nhiều sự khác biệt, độ tương tự càng thấp.

- Trực giác 3: độ tương tự lớn nhất giữa A và B đạt được khi A và B giống hệt nhau.

Xuất phát từ trực giác về độ tương tự, kết hợp với sự logic của toán học, chương trình áp dụng lý thuyết

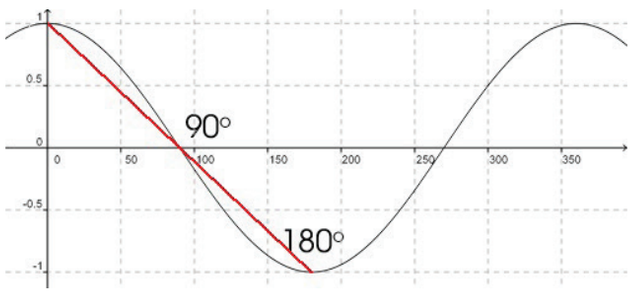
đó để giải quyết vấn đề của bài toán trùng lặp văn bản. Để phát hiện một tài liệu có phải là tài liệu đi sao chép hay không, đơn giản nhất là so sánh tài liệu cần kiểm tra đó với các tài liệu khác trong cơ sở dữ liệu. Ý tưởng là coi tài liệu kiểm tra như một véc-tơ và xếp hạng các tài liệu dựa vào sự tương đồng với tài liệu cần kiểm tra. Để so sánh 2 véc-tơ có thể tính khoảng cách giữa 2 véc-tơ hoặc tính góc tạo ra bởi 2 véc-tơ. Tuy nhiên, cách tính khoảng cách có nhược điểm là không chính xác, bởi khoảng cách lớn với các véc-tơ có chiều dài khác nhau. Do vậy, thường dựa vào góc trong không gian véc-tơ hơn là so sánh về khoảng cách.

Từ những phân tích trên, chương trình áp dụng độ đo cosine làm độ đo sự tương đồng giữa các văn bản. Có thể thấy, góc giữa 2 véc-tơ càng lớn thì cosine hay điểm xếp hạng sẽ càng thấp, khi góc giữa 2 véc-tơ bằng 0 thì điểm sẽ lớn nhất (= 1).

Có 2 cách ghi tương đồng nhau:

- Xếp hạng văn bản theo tứ tự giảm dần dựa trên góc giữa văn bản cần kiểm tra với tập văn bản có sẵn (hình 2).

- Xếp hạng tài liệu theo thứ tự tăng dần dựa trên cosine giữa văn bản cần kiểm tra với tập văn bản có sẵn.



Hình 2: tài liệu được xếp hạng bởi giá trị cosine giảm dần

Theo hình 2, cos(d, q) = 1 khi d = q; cos(d, q) = 0 khi tài liệu d và tài liệu kiểm tra q không có bất kỳ từ chung nào.

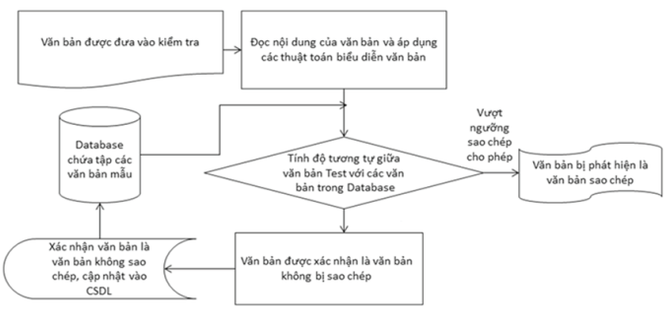
Sau khi biểu diễn véc-tơ đặc trưng của từng văn bản, áp dụng thuật toán tính độ đo cosine để tính độ tương tự của 2 văn bản. Theo như thuật toán này thì coi 2 văn bản là 2 véc-tơ, độ tương tự của 2 văn bản sẽ là góc tạo bởi 2 véc-tơ đó. Khi góc tạo bởi 2 véc-tơ càng lớn tức là 2 văn bản càng khác nhau và cosine giữa véc-tơ 1 và véc-tơ 2 càng nhỏ, ngược lại, nếu cosine càng lớn, càng tiến gần về 1 thì góc tạo bởi giữa 2 véc-tơ đặc trưng của văn bản càng nhỏ, tức là 2 văn bản càng giống nhau.

Module này có mã như sau:

```
Input vector1, vector2
Set tu, mau, sqrt1 and sqrt2 to zero
While grade counter is less than the size of Vector1
Tu is set to sum of its value and product of vector1 at this offset and vector2 at this offset.
Sqrt1 is set to sum of its value and product of vector1 at this offset and vector2 at this offset.
Sqrt2 is set to sum of its value and product of vector1 at this offset and vector2 at this offset.
Set mau to the product of square root of sqrt1 and square root of sqrt2.
Output the division of tu and mau;
```

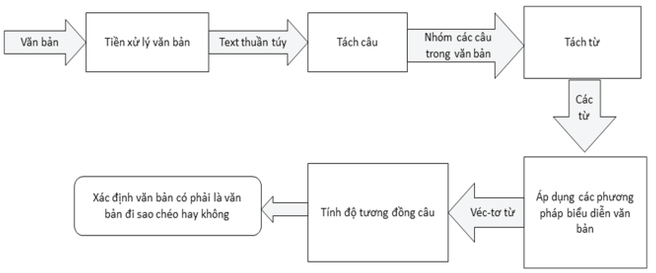
Cài đặt và thử nghiệm chương trình tính độ tương đồng văn bản sử dụng độ đo cosine

Yêu cầu chính của việc phát hiện trùng lặp văn bản đó là việc xác định một văn bản sau khi xử lý sẽ xác định được văn bản đó có trùng với văn bản nào đó đã được xác định trước (hình 3). Đối với các văn bản có sao chép nhưng áp dụng việc sao chép có sử dụng tính "nhập nhằng" của ngôn ngữ thì chương trình không thể vét cạn các trường hợp và đòi hỏi người sử dụng phải xác định thủ công xem văn bản đó được sao chép từ văn bản nào, bao nhiêu phần trăm. Sau khi xác định được sự trùng lặp này, nếu sự trùng lặp, sao chép ở ngưỡng cho phép thì văn bản đó phải được cập nhật vào hệ thống lưu trữ văn bản để tăng quy mô dữ liệu mẫu cho chương trình và áp dụng việc kiểm tra trùng lặp cho văn bản trong lần tiếp theo.



Hình 3: các yêu cầu đối với việc sao chép văn bản

Dựa vào phân tích yêu cầu và bài toán, có thể thấy rằng việc nhận dạng một văn bản là văn bản sao chép hay không, cần phải thực hiện các bước như trong hình 4.



Hình 4: cấu trúc chương trình

Chương trình sử dụng công nghệ .NET của Microsoft do vậy hỗ trợ hoàn toàn Unicode tiếng Việt và được viết trên nền Window (Winform) với ngôn ngữ lập trình là C#. Chương trình sử dụng công cụ tách từ của dự án "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt", do đó yêu cầu máy tính phải hỗ trợ môi trường để chạy file Java.

Các cơ sở dữ liệu sử dụng trong chương trình bao gồm: bộ từ điển cho chức năng tách từ, sử dụng bộ từ điển trong dự án bộ dữ liệu huấn luyện được dùng là các file văn bản được lưu trữ trên SQL Server.

Để chương trình có thể thực hiện được thì yêu cầu tối thiểu để chương trình thực hiện là: máy tính phải cài đặt bộ .NET framework, SQL Server và môi trường JRE, hỗ trợ chạy Java. Các dữ liệu kiểm thử cần phải tạo thành tệp tin trên máy sử dụng phần mở rộng ".txt". Để kiểm tra thuật toán, tiến hành so sánh độ tương đồng văn bản với một số mẫu sau:

Ví dụ 1:

Văn bản 1 (Text1): Tuấn bảo sao cậu không đến?

Văn bản 2 (Text2): Tuấn đến sao cậu không bảo?

Kết quả Cosine(Text1, Text2) = 0,9137256 cho thấy, 2 văn bản trên dù giống nhau hoàn toàn về từ, nhưng cách sắp xếp khác nhau, nên ngữ nghĩa khác nhau, do đó độ tương đồng < 1.

Ví dụ 2:

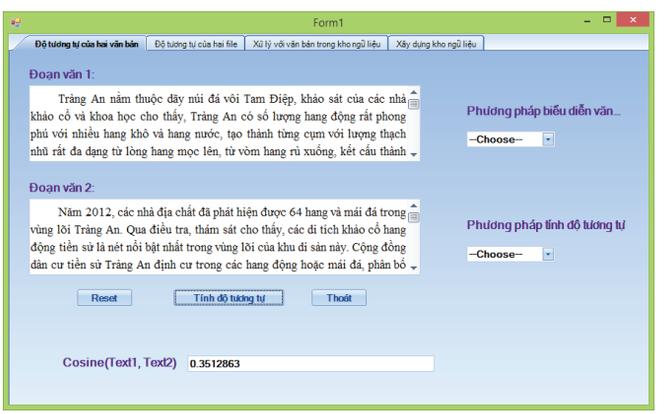
Văn bản 1 (Text1):

Tràng An nằm thuộc dãy núi đá vôi Tam Điệp, khảo sát của các nhà khảo cổ và khoa học cho thấy, Tràng An có số lượng hang động rất phong phú với nhiều hang khô và hang nước, tạo thành từng cụm với lượng thạch nhũ rất đa dạng từ lòng hang mọc lên, từ vòm hang rủ xuống, kết cấu thành nhiều hình thù kỳ lạ tuỳ theo sức tưởng tượng của mỗi du khách. Tại đây còn có khoảng 30 thung, mỗi thung là một bức tranh thuỷ mặc. Sự nguyên sơ của vùng đất này cùng với kiến tạo của địa chất đã tạo nên một thảm thực vật đa dạng về chủng loại là nơi trú ngụ, cư trú của các loài chim thú, là "mái nhà" chung cho các loài động vật sinh sống và phát triển.

Văn bản 2 (Text2):

Năm 2012, các nhà địa chất đã phát hiện được 64 hang và mái đá trong vùng lõi Tràng An. Qua điều tra, thám sát cho thấy, các di tích khảo cổ hang động tiền sử là nét nổi bật nhất trong vùng lõi của khu di sản này. Cộng đồng dân cư tiền sử Tràng An định cư trong các hang động hoặc mái đá, phân bố tập trung trong thung lũng đầm lầy núi đá vôi, chịu sự tác động to lớn của biến đổi cảnh quan môi trường do các đợt biển tiến, biển thoái. Cư dân tiền sử nơi đây là những người tiếp cận và khai thác biển đầu tiên ở Việt Nam, sáng tạo ra tổ hợp công cụ lao động bằng đá vôi, duy trì lâu dài kỹ nghệ ghè đẽo, sớm nảy sinh kỹ thuật cưa, mài; chế tạo và sử dụng phổ biến đồ gốm.

Kết quả, 2 văn bản trên có nội dung không hoàn toàn giống nhau nhưng có một số từ bị lặp lại ở cả 2 văn bản. Theo thực nghiệm trên máy độ tương tự tính theo độ đo cosine giữa 2 văn bản, nếu biểu diễn 2 văn bản trên theo phương pháp Boolean Cosine(Text1,Text2) = 0,3512863 (hình 5). Tức là, 2 bản này có 1 số điểm chung nhất định.



Hình 5: kết quả thử nghiệm tính độ đo cosine giữa 2 đoạn văn

Kết luận

Phương pháp tính độ tương đồng văn bản dựa trên độ đo cosine là phương pháp đơn giản, dễ cài đặt và thử nghiệm, với độ chính xác tương đối cao.

Sau khi tiến hành thử nghiệm chương trình trên một số ví dụ cụ thể, có thể rút ra được một số nhận xét sau:

- Trường hợp 1: độ đo cosine càng lớn (tiến gần tới 1), nghĩa là 2 văn bản có nhiều điểm giống nhau.

- Trường hợp 2: độ đo cosine càng nhỏ (tiến gần về 0), 2 văn bản khác nhau hoàn toàn.

Chương trình hiệu quả hơn với những cặp văn bản hoặc là giống nhau nhiều, hoặc là khác nhau nhiều. Độ đo cosine còn có thể sử dụng để đánh giá độ tương

đồng với các ngôn ngữ khác như tiếng Anh, Pháp… do đó hoàn toàn có thể phát triển và ứng dụng để kiểm tra độ tương đồng của văn bản có nhiều ngôn ngữ khác nhau, đây là hướng để phát triển trong các nghiên cứu tiếp theo.

Tài liệu tham khảo

[1] Trần Cao Đệ (2013), *Đo độ tương tự ngữ nghĩa tiềm ẩn để phát hiện sao chép tài liệu,* Khoa Công nghệ Thông tin và Truyền thông, Đại học Cần Thơ.

[2] Nguyễn Thanh Hùng (2006), *Hướng tiếp cận mới trong việc tách từ để phân loại văn bản tiếng Việt sử dụng giải thuật di truyền và thống kê trên Internet,* Đại học Quốc gia TP Hồ Chí Minh.

[3] Trần Mai Vũ (2009), "Tóm tắt văn bản dựa vào trích xuất câu", *Luận văn cao học*, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

[4] Trần Thị Oanh (2013), "Mô hình tách từ, gán nhãn từ loại và hướng tích hợp cho tiếng Việt", *Luận văn cao học*, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

[5] Đoàn Sơn (2002), *Phương pháp biểu diễn văn bản sử dụng tập mờ và ứng dụng khai phá dữ liệu văn bản*, Khoa Công nghệ thông tin, Đại học Quốc gia Hà Nội.

[6] Maosong Sun, D. Shen, B.K. Tsou (1998), "Chinese word segmentation without suing lexicon", *Proceedings of the Joint Conference of ACL and COLING*, Montreal, Quebec, Canada, pp.1265-1271.

[7] L.A. Ha (2003), "A method for word segmentation in Vietnamese", *Proceedings of Corpus Linguistics*, Lancaster, UK.

[8] Lê Hương Thanh, Nguyễn Thái Phương (2009), *Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt*, Viện Công nghệ thông tin.

[9] Nguyễn Kiên Trường (2005), *Tiếp xúc ngôn ngữ ở Việt Nam*, Nhà xuất bản Khoa học xã hội.

[10] D. Hand, H. Mannila, P. Smyth (2001), "Principles of Data Minning", *MIT Press*, Cambridge, MA.

[11] Phạm Thị Thu Uyên, Hoàng Minh Hiền (2008), "Độ tương đồng ngữ nghĩa giữa hai câu và ứng dụng trong tóm tắt văn bản tiếng Việt", *Hội nghị khoa học Đại học Huế*.