PHƯƠNG PHÁP TRÍCH RÚT TỪ KHÓA TÌM TẬP ỨNG CỬ TRONG BÀO TOÁN PHÁT HIỆN ĐẠO VĂN

Nguyễn Văn Sơn[1*], Lê Thanh Hương[2], Nguyễn Chí Thành[1]

Tóm tắt: *Trong bài toán phát hiện đạo văn, hai vấn đề quan trọng cần thực hiện là tìm tập tài liệu nghi ngờ bị sao chép và kiểm trùng văn bản. Để tìm tập tài liệu nghi ngờ bị sao chép, vấn đề cốt yếu là phải đưa ra được tập từ khóa đại diện cho tài liệu đầu vào và cho các đoạn trong tài liệu đó. Tập từ khóa này được dùng để sinh câu truy vấn tìm kiếm các tài liệu nghi ngờ bị sao chép. Bài báo này đề xuất một phương pháp trích rút tập từ khóa đại diện cho tài liệu đầu vào dựa trên các độ đo tf.idf mức tài liệu và mức đoạn, có xem xét yếu tố từ loại với thứ tự ưu tiên lần lượt là danh từ, tính từ, động từ. Để đánh giá phương pháp đề xuất, chúng tôi tiến hành xây dựng tập dữ liệu thử nghiệm tiếng Việt gồm 10 tài liệu cần kiểm tra với mỗi tài liệu có 10 tài liệu liên quan. Kết quả thử nghiệm cho thấy với các truy vấn tìm kiếm do hệ thống sinh ra có thể trả về tập tài liệu nghi ngờ với độ chính xác 67,77%. Điều này cho thấy cách tiếp cận đề xuất là có triển vọng.*

Từ khóa: Đạo văn, Trích rút từ khóa, Tập ứng cử, Tf.idf, Từ loại.

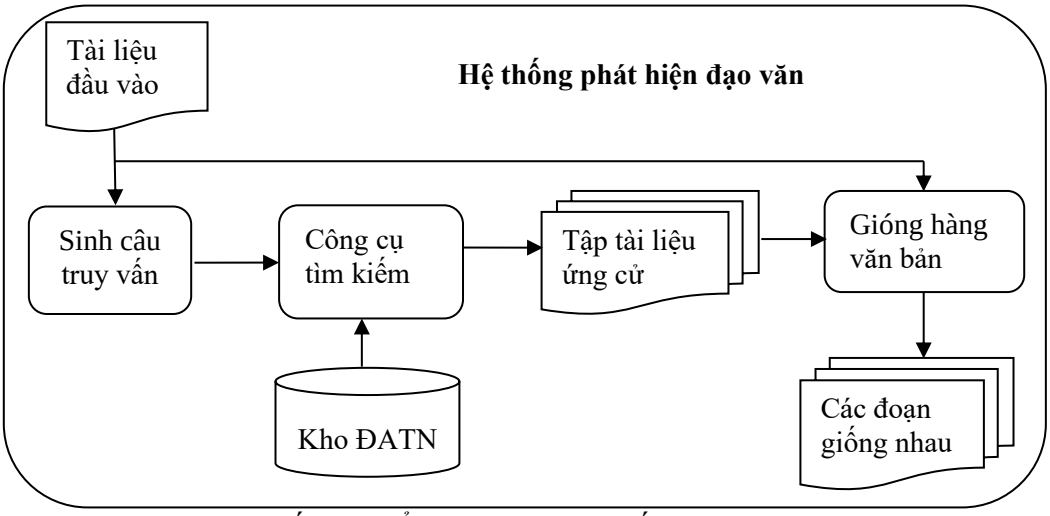
1. ĐẶT VẤN ĐỀ

Sự phát triển của Internet đem lại cho chúng ta nhiều tiện nghi như có thể dễ dàng tìm thấy thông tin, tài liệu mình quan tâm, nhưng nó cũng đặt ra nhiều vấn đề như hiện tượng sao chép nội dung của các tài liệu. Đặc biệt là với các báo cáo bài tập lớn, tiểu luận, đồ án tốt nghiệp (ĐATN) và luận văn thạc sĩ thì vấn nạn đó xảy ra rất nhiều. Theo Báo Tuổi trẻ Online số tháng 5/2015, tỉ lệ sinh viên đại học "đạo văn" ở một số trường đại học Việt Nam chiếm tỉ lệ cao so với thế giới. Số liệu khảo sát sinh viên tại Trường Đại học Duy Tân cho thấy trên 70% sinh viên "đạo văn". Tuy nhiên, việc phát hiện đạo văn không đơn giản. Do hiện nay việc tổ chức lưu trữ, quản lý và khai thác nguồn tri thức đó còn chưa được quan tâm đúng mức, các tài liệu đó xuất hiện tản mát ở một số nơi dẫn đến tình trạng các tài liệu sao chép bất hợp pháp xảy ra mà các giáo viên hoặc những người làm công tác phản biện rất khó kiểm soát.

Đạo văn là hình thức sao chép, cắt dán, gõ lại, viết lại, sử dụng lại ý tưởng, kết quả mà không có trích dẫn đến tác giả hoặc nguồn thông tin. Đạo văn thường xuất hiện dưới hai hình thức: sao chép nguyên văn và sao chép ý tưởng. Để thực hiện việc đạo văn, người sao chép thực hiện thu thập các đoạn văn bản từ nhiều nguồn khác nhau để tạo nên văn bản của mình.

Hai công việc chính để giải quyết bài toán phát hiện đạo văn là: tìm tập tài liệu ứng cử và tìm các đoạn văn bản giống nhau giữa hai văn bản. Để kiểm tra một tài liệu đầu vào có sao chép từ các tài liệu khác lưu trong hệ thống hay không, trước tiên hệ thống cần xác định các từ khóa là cụm từ đại diện cho tài liệu đầu vào, và sử dụng một công cụ tìm kiếm để tìm các tài liệu chứa các từ đó. Sau đó, từng tài liệu trong tập tài liệu trả về (tập tài liệu ứng cử) sẽ được đối sánh (giống hàng) với tài liệu đầu vào để tìm ra các đoạn trùng nhau giữa các tài liệu đó. Việc tài liệu đầu vào có bị coi là đạo văn hay không là do con người quyết định.

Nội dung thực hiện trong bài báo này nằm trong công việc thứ nhất – tìm tập tài liệu ứng cử. Kiến trúc tổng quát của hệ thống phát hiện đạo văn được mô tả trong hình 1 dưới đây.



Hình 1. *Kiến trúc tổng quát của hệ thống phát hiện đạo văn*

Trong bài báo này chúng tôi xây dựng phương pháp trích rút từ khóa của một tài liệu được sử dụng trong câu truy vấn tìm tài liệu ứng cử. Nội dung bài báo gồm bốn phần. Phần 2 giới thiệu phương pháp trích rút từ khóa. Phần 3 trình bày kết quả thử nghiệm và đánh giá. Phần 4 gồm kết luận và hướng phát triển tiếp theo.

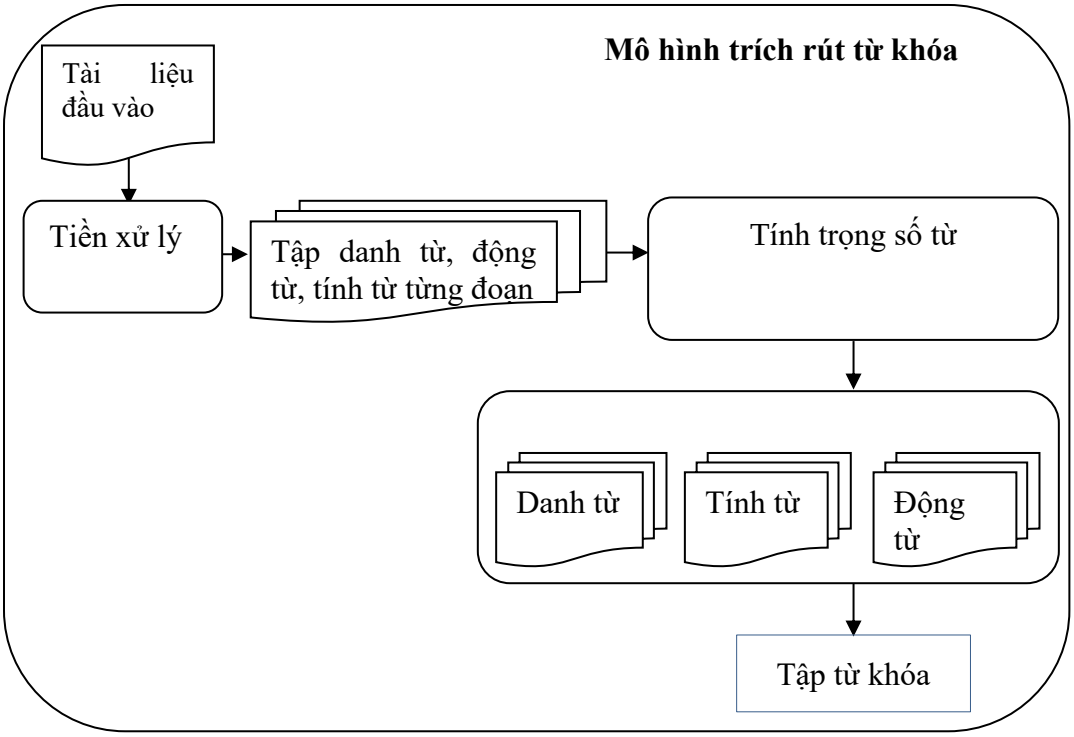
2. PHƯƠNG PHÁP TRÍCH RÚT TỪ KHÓA

2.1. Phát biểu bài toán và đề xuất phương pháp

Cho một tập tài liệu $D=\{d_1,d_2,...d_N\}$ và tài liệu cần kiểm tra d. Tìm tập tài liệu ứng cử $C=\{c_1,c_2,...,c_k\}$ với $c_i \in D$ là tài liệu nghi ngờ bị tài liệu d sao chép.

Để tìm tập tài liệu ứng cử C thông qua công cụ tìm kiếm chúng tôi thực hiện truy vấn từ kho tài liệu D mà đầu vào của câu truy vấn là tập từ khóa. Trích rút từ khóa từ một văn bản là tự động xác định tập các từ đại diện biểu diễn chủ đề chính của văn bản [1]. Có nhiều phương pháp trích rút từ khóa, tuy nhiên chất lượng của tập từ khóa thu được phụ thuộc vào nhiều yếu tố như chất lượng của tài liệu và độ dài của tài liệu. Với những đoạn văn bản ngắn, việc sinh ra tập từ khóa trở lên khá khó khăn và không hiệu quả, đặc biệt với các đoạn văn bản ngắn chứa từ viết tắt hoặc các câu không đúng ngữ pháp (như các đoạn tin nhắn). Với các văn bản dài, việc trích rút từ khóa dựa trên các phương pháp chính như sử dụng độ đo tf.idf, phương pháp TextRank [2] hay phương pháp RAKE (Rapid Automatic Keyword Extraction) [3]. Mihalcea và Tarau[2] chỉ ra rằng phương pháp TextRank đạt hiệu quả tốt nhất khi chọn từ khóa là danh từ và tính từ. Bên cạnh đó, phân tích [6] chỉ ra rằng các câu quá ngắn thường ít mang thông tin quan trọng.

Trong bài báo này chúng tôi thực hiện trích rút từ khóa dựa trên độ đo tf.idf [4] có xem xét đến yếu tố từ loại theo mô hình như hình 2 dưới đây.



***Hình 2**. Mô hình trích rút từ khóa*

Tài liệu đầu vào bao gồm các tệp văn bản như word hoặc pdf. Quá trình trích rút từ khóa từ văn bản đầu vào gồm các bước sau:

1. Tiền xử lý`
2. Tính các trọng số cho các từ trong đoạn
3. Lựa trọn từ khóa.

2.2. Tiền xử lý

2.2.1. Tách từ, tách câu và gán nhãn từ loại

Tiền xử lý là bước quan trọng đối với các hệ thống tìm kiếm. Tệp tin đầu vào có dạng .pdf, .doc hoặc .docx, đọc nội dung và loại bỏ các ký tự đặc biệt (như các ký tự điều khiển, ký tự xuống dòng) và thực hiện tách câu, tách từ và gán nhãn từ loại. Sau khi gán nhãn chúng tôi lựa chọn tất cả các từ là danh từ, động từ và tính từ [17] để thực hiện các bước tiếp theo. Trong bài báo này, chúng tôi sử dụng công cụ tách từ vnTagger [16] phiên bản 4.1.1, được phát triển bởi nhóm tác giả Lê Hồng Phương để tách nội dung của văn bản thành các câu, các đơn vị từ và gán nhãn từ loại. Với chuỗi đầu vào *"Hỗ trợ phân tích các chuẩn Log phổ biến hiện nay, tập trung vào vấn đề giám sát an ninh, hỗ trợ cảnh báo qua Email và SMS"* sau khi chạy chương trình vnTagger chúng ta thu được kết quả:

```
<doc>
    <s>
            <w pos="V">Hỗ trợ</w>
            <w pos="V">phân tích</w>
```

```
            <w pos="L">các</w>
            <w pos="N">chuẩn</w>
            <w pos="Np">Log</w>
            <w pos="V">phổ biến</w>
            <w pos="N">hiện nay</w>
            <w pos=",">,</w>
            <w pos="V">tập trung</w>
            <w pos="E">vào</w>
            <w pos="N">vấn đề</w>
            <w pos="V">giám sát</w>
            <w pos="N">an ninh</w>
            <w pos=",">,</w>
            <w pos="V">hỗ trợ</w>
            <w pos="V">cảnh báo</w>
            <w pos="E">qua</w>
            <w pos="Np">Email</w>
            <w pos="CC">và</w>
            <w pos="Np">SMS</w>
        </s>
    </doc>
```

Trong đó ký hiệu các nhãn từ loại chính [16] gồm:

N: Danh từ; V: Động từ; A: Tính từ; Np: Danh từ riêng; P: Đại từ; L: Định từ; M: Số từ; R: Phó từ; E: Giới từ

2.1.2. Chia đoạn văn bản

Sau bước tiền xử lý dữ liệu, mỗi tài liệu được chia thành các đoạn sao cho mỗi mỗi câu không thuộc hai đoạn. Bằng phương pháp thống kê các tài liệu trong kho ĐATN có khoảng 90% số đồ án có độ dài 70-80 trang A4, mỗi trang có từ 30 đến 35 dòng, mỗi dòng khoảng 15 tiếng. Có nhiều phương án chia văn bản thành các đoạn như coi văn bản là một đoạn [12], mỗi đoạn 50 dòng [14], mỗi đoạn được lựa chọn dựa trên tiêu đề đoạn [12], mỗi đoạn gồm 100 từ [13], hay mỗi đoạn 5 câu [15]. Phân tích trên các văn bản đầu vào, số tiếng trong mỗi văn bản trong xấp xỉ 35.000 tiếng, các đoạn dựa theo tiêu đề có độ dài không đồng đều do vậy bài báo lựa chọn độ dài mỗi đoạn khoảng 500 tiếng tương đương với khoảng xấp xỉ 70 đoạn trong một văn bản.

2.3. Tính trọng số và xác định từ khóa đoạn

Ở bước này, văn bản đã được chia thành các đoạn. Với mỗi đoạn ta cần tìm các từ khóa đại diện cho đoạn đó. Có những từ khóa đại diện cho văn bản nhưng trong một số đoạn, có thể từ khóa lại ít xuất hiện. Vì vậy, bên cạnh các từ khóa của văn bản, chúng tôi còn sử dụng cả những từ khóa của đoạn văn bản.

2.2.1. Tính trọng số của từ

Trọng số của một từ được được xác định thông qua giá trị trọng số tf.idf [4] của nó. Từ có trọng số cao sẽ được chọn làm từ khóa của văn bản. Hai loại trọng số được sử dụng là:

1. tf.idf1: với tf là số lần xuất hiện của từ trong đoạn, idf là nghịch đảo số lần xuất hiện của từ trong tài liệu đầu vào.

2. tf.idf2: với tf là số lần xuất hiện của từ trong đoạn, idf là nghịch đảo số lần xuất hiện của từ trong kho tài liệu ĐATN

Cụ thể như sau. Xét từ w_{ij} (từ thứ i trong đoạn j)

$$tf.idf1 = tf_{ij} * idf1_i \quad (1)$$

tf_{ij} là tần số xuất hiện của từ thứ i trong đoạn j.

$$tf_{i,j} = \frac{freq_{i,j}}{max_l freq_{l,j}} \quad (2)$$

$idf1_i$: tần suất xuất hiện nghịch đảo của từ w_{ij} trong đoạn

$$idf1_i = \log\frac{N}{n_i} \quad (3)$$

với N là số các đoạn của văn bản đang xét; n_i là số đoạn của văn bản đang xét chứa từ w_{ij}

$$tf.idf2 = tf_{ij} * idf2_i \quad (4)$$

tf_{ij} là tần số xuất hiện của từ thứ i trong đoạn j.

$$tf_{i,j} = \frac{freq_{i,j}}{max_l freq_{l,j}} \quad (5)$$

$idf2_i$: tần suất xuất hiện nghịch đảo của từ thứ i trong kho dữ liệu văn bản.

$$idf2_i = \log\frac{M}{m_i} \quad (6)$$

với M là số lượng văn bản trong kho dữ liệu; m_i là số văn bản chứa từ w_{ij}

2.2.2. Trích rút từ khóa

Để đảm bảo tốc độ tìm kiếm các công cụ tìm kiếm luôn cấu hình để giới hạn số từ khóa đầu vào (như ChatNoir [10] cho phép 10 từ khóa, Apache Nucene [11] cho phép 1024 từ khóa).

Một từ được xác định là từ khóa của một đoạn nếu nó quan trọng trong đoạn và trong văn bản. Qua thử nghiệm chúng tôi lựa chọn 10 từ khóa có giá trị tf.idf cao nhất, 3 câu có giá trị tf.idf cao nhất và tổng số từ khóa cần trích rút k=30 đảm bảo tốc độ và kết quả tìm kiếm. Thuật toán trích rút từ khóa cho một đoạn trong văn bản sau khi tính tf.idf1 và tf.idf2 cho tất cả các từ trong đoạn như sau:

1. Chọn 10 từ có tf.idf1 và 10 từ có tf.idf2 cao nhất
2. Xác định các câu quan trọng: câu được xác định là quan trọng nếu nó chứa cả từ có tf.idf1 và tf.idf2 lựa chọn ở bước trên
3. Lấy 3 câu có tf.idf1 và tf.idf2 cao nhất từ các câu trên.
4. Từ khóa được trích rút từ các câu trên theo trình tự sau đến khi số từ khóa thu được bằng k (k cho trước):

- Các danh từ có giá trị tf.idf cao
- Các danh từ khác trong câu
- Tính từ và động từ có tf.idf1 cao

Đầu ra của thuật toán là tập từ khóa sẽ sử dụng để sinh ra câu truy vấn. Các từ này được xếp cạnh nhau theo trật tự xuất hiện trong tài liệu gốc để tạo thành câu

truy vấn. Câu truy vấn này sẽ được đưa vào các công cụ tìm kiếm để tìm các tài liệu có thể bị sao chép.

3. THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

3.1. Chuẩn bị

Tiền xử lý kho dữ liệu: Chúng tôi thực hiện chuẩn hóa tên các tệp ĐATN từ 1 đến 350 và không thay đổi nội dung cũng như định dạng tệp. Để tăng tốc độ khi tính tần suất xuất hiện tf.idf2 chúng tôi thực hiện tính toán idf2 dưới dạng từ điển với khóa là từ và giá trị là tần suất xuất hiện của từ trong toàn bộ ĐATN. Từ điển này được lưu trữ trên ổ đĩa và được nạp khi chạy chương trình.

Dữ liệu thử nghiệm: Vì trên thế giới không có tập dữ liệu mẫu về sinh câu truy vấn đại diện cho văn bản nên việc đánh giá kết quả được tiến hành thủ công nhằm đánh giá các truy vấn đó có điển hình cho tài liệu đầu vào hay không. Để xây dựng một tài liệu đầu vào chúng tôi thực hiện sao chép một số đoạn trong kho dữ liệu (tài liệu trộn) đưa vào tài liệu mẫu. Chúng tôi tiến hành sinh câu truy vấn một cách thủ công trên 10 tài liệu đầu vào và sau đó so sánh với kết quả hệ thống sinh ra. Chúng tôi thực hiện đánh giá trên 10 kết quả tốt nhất thu được từ công cụ tìm kiếm.

3.2. Đánh giá kết quả

Hệ thống được cài đặt bằng ngôn ngữ Java, sử dụng công cụ vnTagger của tác giả Lê Hồng Phương. Hệ thống thử nghiệm trên bộ dữ liệu 350 ĐATN. Với mỗi đầu vào là một ĐATN, hệ thống tiến hành phân tích để xác định các câu truy vấn đại diện cho văn bản.

Kết quả được đánh giá trên các độ đo thường dùng trong học máy là Precision, Recall và F-score[7].

Kết quả thử nghiệm được cho trong bảng sau:

Bảng 1: *Kết quả thử nghiệm*

STT	Tên file	Số tệp trộn	Số kết quả thu được	Số tệp tìm được	Precision	Recall	F-Score
1	File1	5	6	4	0,8	0,6667	0,7273
2	File2	5	7	5	1	0,7143	0,8333
3	File3	5	8	4	0,8	0,5	0,6154
4	File4	5	7	5	1	0,7143	0,8333
5	File5	5	6	4	0,8	0,6667	0,7273
6	File6	5	5	4	0,8	0,8	0,8
7	File7	5	6	3	0,6	0,5	0,5455
8	File8	5	7	4	0,8	0,5714	0,6666
9	File9	5	9	3	0,6	0,3333	0,4285
10	File10	5	10	5	1	0,5	0,6667
Trung bình		**50**	**71**	**41**	**0,82**	**0,5775**	**0,6777**

Nhận xét: Giá trị trung bình độ đo Precision cho kết quả khá tốt, các điểm đánh giá trên toàn tập dữ liệu đều trên 80%. Tập dữ liệu cho kết quả tốt nhất là file 2, file 5 và file 10 đạt 100%. Tuy nhiên có kết quả thấp so với kết quả còn lại như file7 và file9.

Có một số văn bản có điểm đánh giá thấp do trong văn bản có nhiều hình vẽ và ký hiệu toán học. Do vậy, phương pháp này sẽ cho kết quả tốt nhất với các văn bản chứa ít ký tự đặc biệt và độ dài câu đủ lớn.

4. KẾT LUẬN

Với đặc thù của Tiếng Việt là ngôn ngữ đa âm tiết, trong bài báo này chúng tôi đã giới thiệu phương pháp trích rút từ khóa từ văn bản Tiếng Việt và sự thành công khi áp dụng phương pháp này trong việc tìm kiếm tập tài liệu ứng cử làm tiền đề để giải quyết bài toán phát hiện đạo văn. Đặc biệt bài báo đưa ra phương pháp trích rút từ khóa dựa trên hai độ đo tf.idf1 và tf.idf2 có xem xét yếu tố từ loại. Phương pháp đề xuất mang lại nhiều lợi ích trong việc phát hiện sự sao chép nguyên mẫu hoặc có sự biến đổi trật tự từ trong các bài báo khoa học hay đồ án tốt nghiệp tại các trường đại học. Điểm yếu của mô hình là khả năng phát hiện đạo văn cho các văn bản tương đồng về ngữ nghĩa. Điểm hạn chế này được phát triển trong thời gian tới.

TÀI LIỆU THAM KHẢO

[1] H. T. B. Lương Chi Mai, "Về xử lý tiếng Việt trong công nghệ thông tin," *Báo cáo Tổng kết đề tài KC.01.01/06-10,* 2009.

[2] R. a. P. T. Mihalcea, "Textrank: Bringing order into text," *Proceedings of the 2004 conference on empirical methods in natural language processing,* 2004.

[3] D. E. N. C. a. W. C. Stuart Rose, "Automatic keyword extraction from individual documents," *Text Mining: Applications and Theory,* pp. 1-20, 2010.

[4] M. Dillon, "Introduction to modern information retrieval: G. Salton and M. McGill. McGraw-Hill, New York (1983). xv+ 448 pp., $32.95 ISBN 0-07-054484-0.," pp. 402-403, 1983.

[5] R. Al-Hashemi, "Text Summarization Extraction System (TSES) Using Extracted Keywords," *International Arab Journal of e-Technology,* pp. 164-168, 2010.

[6] T. A. a. K. Y. Luu, "A pointwise approach for Vietnamese diacritics restoration," *Asian Language Processing (IALP), 2012 International Conference on. IEEE,* pp. 189-192, 2012.

[7] C. a. E. G. Goutte, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," *European Conference on Information Retrieval,* pp. 345-359, 2005.

[8] C.-T. X.-H. P. a. T.-T. N. Nguyen, "Jvntextpro: A java-based vietnamese text processing tool," *http://jvntextpro.sourceforge.net/,* 2010.

[9] Q. T. e. a. Dinh, "Word Segmentation of Vietnamese Texts: a comparison of approaches. LREC: 2008.," *Proceedings of the 10th International Conference on Information and Knowledge Management Ho Ngoc Duc, 2004: Vietnamese word list: Ho Ngoc Duc's word list–http://www. informatik. unileipzig. de/~ duc/software/misc/wordlist. html John O'Neil. 2007. Large Co.*

[10] M. e. a. Potthast, "ChatNoir: a search engine for the ClueWeb09 corpus," *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval,* pp. 1004-1004, 2012.

[11] "Apache Lucene".

[12] S. a. M. B. Suchomel, "Heterogeneous Queries for Synoptic and Phrasal Search.," *In CLEF (Working Notes),* pp. 1017-1020, 2014.

[13] A. S. S. Prakash, "Experiments on Document Chunking and Query Formation for Plagiarism Source Retrieval," *Notebook for PAN at CLEF 2014,* 2014.

[14] V. Elizalde, "Using Noun Phrases and tf-idf for Plagiarized Document Retrieval," *CLEF (Working Notes),* 2014.

[15] L. e. a. Kong, "Source Retrieval Based on Learning to Rank and Text Alignment Based on Plagiarism Type Recognition for Plagiarism Detection.," *CLEF (Working Notes),* 2014.

[16] A. R. T. M. H. N. M. R. Phuong Le-Hong, "An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts," *Traitement Automatique*

des Langues Naturelles-TALN 2010, 2010.

[17] N. T. Cẩn, "Ngữ pháp tiếng Việt," *NXB ĐHQGHN,* 2004.

ABSTRACT

KEYWORD EXTRACTION METHOD FOR CANDIDATE DOCUMENT RETRIEVAL IN VIETNAMESE PLAGIARISM DETECTION PROBLEM

Two important issues that need to be addressed in plagiarism detection are source retrieval and checking duplication. To do source retrieval, it is essential to provide a set of keywords representing for the suspected document and its paragraphs. This keyword set is used to search for relevant documents. This paper proposes a method of extracting such keyword set basing on tf.idf measures at document and paragraph levels, in companied with part-of-speech tags. To evaluate the proposed method, we generated a test set consisting of 10 suspicious documents in Vietnamese, each of which is accompanied with 10 related ones. The documents returned by the source retrieval module were compared with the above mentioned related documents to calculate the system accuracy. Experiment results gave us the accuracy of 67,77%, which proved that the proposed approach is promising in solving source retrieval task.

Keywords: Plagiarism, Keyword extraction, Candidate document, Tf.idf, Part of speech.

Địa chỉ: [1]Viện Công nghệ thông tin/Viện KH-CN quân sự;

[2]Viện Công nghệ thông tin và truyền thông/Đại học Bách khoa Hà Nội;

*Email của tác giả liên hệ : sonnv78@gmail.com