ISSN 2354-0575

NGHIÊN CỨU CÁC PHƯƠNG PHÁP THU THẬP VÀ TÓM TẮT THÔNG TIN TRONG CÁC MẠNG XÃ HỘI

Vũ Xuân Thắng, Trần Đỗ Thu Hà, Đặng Vân Anh, Nguyễn Vinh Quy

Trường Đại học Sư phạm Kỹ thuật Hưng Yên

Ngày tòa soạn nhận được bài báo: 14/11/2019

Ngày phản biện đánh giá và sửa chữa: 04/12/2019

Ngày bài báo được chấp nhận đăng: 24/12/2019

Tóm tắt:

Bài báo này chúng tôi trình bày quá trình nghiên cứu một số phương pháp thu thập, tóm tắt thông tin trong các mạng xã hội như Facebook, google+, các trang báo điện tử. Nhóm tác giả đã thực thực hiện giải pháp thu thập thông tin từ các mạng xã hội sau đó thực hiện đưa ra các giải pháp tóm tắt tắt văn bản từ đó tối ưu được phương pháp sử dụng để phân loại thông tin trên mạng xã hội.

Từ khoá: *Collect information, information analysis, information evaluates.*

1. Đặt vấn đề

Vào năm 1958, Luhn của IBM đã trình bày phương pháp tóm tắt tự động cho các bài báo kĩ thuật sử dụng phương pháp thống kê thông qua tần suất và phân bố của các từ trong văn bản [5]. Chúng ta đang ở thế kỷ 21, với sự phát triển của Internet, lượng thông tin bùng nổ nhanh chóng đặc biệt là trong các mạng xã hội, việc thu nhận những thông tin quan trọng cũng trở thành một vấn đề thiết yếu thì bài toán tóm tắt văn bản tự động mới được sự quan tâm thiết thực của nhiều nhà nghiên cứu.

Theo Inderjeet Mani, mục đích của tóm tắt văn bản tự động là: "*Tóm tắt văn bản tự động nhằm mục đích trích xuất nội dung từ một nguồn thông tin và trình bày các nội dung quan trọng nhất cho người sử dụng theo một khuôn dạng súc tích và gây cảm xúc đối với người sử dụng hoặc một chương trình cần đến*" [6].

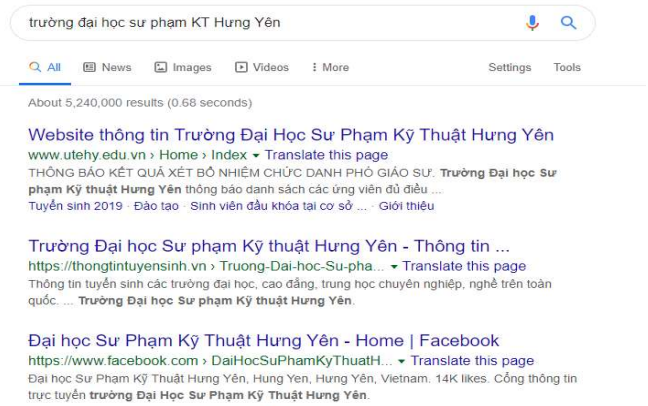
Việc đưa ra được một văn bản kết quả tóm tắt có chất lượng như là văn bản do con người làm ra mà không bị giới hạn bởi miền ứng dụng là được xác định là cực kỳ khó khăn. Vì vậy, các bài toán được giải quyết trong tóm tắt văn bản thường chỉ hướng đến một kiểu văn bản cụ thể hoặc một kiểu tóm tắt cụ thể.

2. Các phương pháp thu thập và tóm tắt thông tin trên mạng xã hội

2.1. Các phương pháp thu thập thông tin trên mạng xã hội

2.1.1. Tìm kiếm thông tin

Tìm trên các thư mục chủ đề: là tìm tin theo hệ thống phân loại như mục chủ đề: nhắp chuột trên hạng mục chủ đề mong muốn. Một trang vừa trình bày cho mục vừa chọn gồm 3 phần: đầu trang là những dịch vụ đặc biệt của Google; kế tiếp là các liên kết đến các hạng mục con và cuối cùng là các liên kết đến các Site liên quan đến hạng mục đó. Các hạng mục con có chữ số nằm trong ngoặc đơn biểu thị số Site nằm trong hạng mục con đó. Một số hạng mục có dấu @ ở bên phải cho biết nó cũng nằm trong một phần khác trong hệ thống của Google.



Ngoài ra còn có thể sử dụng các toán tử logic để tìm tin nâng cao như: toán tử AND (&); toán tử

OR; toán tử NOT; toán tử NEAR.

Tìm tin theo từ khóa: trên các máy tìm kiếm Search Engine nói chung và tìm trên Google (hay trên một website). Tìm tin theo từ khóa có hai cách: tìm tin thông thường và tìm tin nâng cao (Advanced Search).

2.1.2. Lấy thông tin từ website về máy tính

Ở phép tìm kiếm thông tin trên ta có thể lấy được thông tin các bản tin liên quan đến từ khóa tìm kiếm. Việc tiếp theo phải làm là sử dụng phương pháp đặc biệt để lấy về tự động các bản tin đã được tìm kiếm ở trên. Ta lấy về nội dung các bản tin như sau:

Bước 1. Lấy về các đường link trên website sau khi thực hiện lệnh tìm kiếm.

Bước 2. Thực hiện lấy về nội dung của đường link sau đó ghi vào 1 file văn bản như đoạn lệnh dưới đây.

```
using System;
using System.Net.Http;
using System.Threading.Tasks;
using System.IO;
namespace HttpClientEx
{
    class Program
    {
        static async Task Main(string[] args)
        {
            using var client = new
HttpClient();
            var content = await
client.GetStringAsync("http://utehy.edu.vn/Newsl
etters/NewsDetail/9375");
    StreamWrite inp=new
StreamWrite(“baidang.txt”);
    inp.Write(content);
    inp.close();
    }
  }
}
```

Bước 3. Thực hiện lặp lại cho đến khi hoàn thành.

2.2. Một số vấn đề trong tóm tắt thông tin

2.2.1. Một số khái niệm về bài toán tóm tắt văn bản

- Tỷ lệ nén (Compression Rate): là độ đo thể hiện bao nhiêu thông tin được cô đọng trong văn bản tóm tắt được tính bằng công thức:

$$\text{CompressionRate} = \frac{\text{SummaryLength}}{\text{SourceLength}} \quad (1)$$

SummaryLength: Độ dài văn bản tóm tắt

SourceLength: Độ dài văn bản nguồn

- Độ nổi bật hay liên quan (Salience or Relevance): Là trọng số được gán cho thông tin trong văn bản thể hiện độ quan trọng của thông tin đó đối với toàn văn bản hay để chỉ sự liên quan của thông tin đó đối với chương trình của người sử dụng.

- Sự mạch lạc (coherence): Một văn bản tóm tắt gọi là mạch lạc nếu tất cả các thành phần nằm trong nó tuân theo một thể thống nhất về mặt nội dung và không có sự trùng lặp giữa các thành phần.

2.2.2. Các loại hình tóm tắt văn bản

Tóm tắt đơn văn bản: Bài toán tóm tắt văn bản đơn cũng giống như các bài toán tóm tắt khác, là một quá trình tóm tắt tự động với đầu vào là một văn bản, đầu ra là một đoạn mô tả ngắn gọn nội dung chính của văn bản đầu vào đó. Văn bản đơn có thể là một trang Web, một bài báo, hoặc một tài liệu với định dạng xác định (ví dụ: .doc, .txt) … Tóm tắt văn bản đơn là bước đệm cho việc xử lý tóm tắt đa văn bản và các bài toán tóm tắt phức tạp hơn. Chính vì thế những phương pháp tóm tắt văn bản ra đời đầu tiên đều là các phương pháp tóm tắt cho văn bản đơn.

Tóm tắt đa văn bản: Tóm tắt đa văn bản có thể được coi như là một mở rộng của tóm tắt đơn văn bản. Mục đích của tóm tắt đa văn bản: Là quá trình trích xuất nội dung từ một tập các văn bản có liên quan đến nhau, trong quá trình đó các thông tin dư thừa sẽ được loại bỏ và những thông tin quan trọng sẽ được biểu diễn dưới hình thức cô đọng, súc tích và giàu cảm súc đến người sử dụng hoặc chương trình cần dùng.

2.2.3. Các thách thức của quá trình tóm tắt đa văn bản

Trùng lặp đại từ và đồng tham chiếu:

Thông thường, chúng ta đề cập đến một tên thực thể chính là nói đến tên ban đầu của thực thể đấy và sau đó thường hay sử dụng một đại từ thay thế nói về thực thể trên. Xác định chính xác được thực thể mà đại từ chỉ đến được gọi là việc xác định trùng lặp đại từ (Pronominal Anaphora resolution).

Nhập nhằng mặt thời gian: Các văn bản trong cụm tài liệu có thể được chỉ đến bởi nhiều từ hay cụm từ chỉ thời gian ví dụ: hôm qua, hôm nay… Việc xác định rõ ràng các mốc thời gian tương ứng là một điều kiện cần để sắp xếp các câu hay các văn bản theo đúng trình tự hợp lý. Một số hệ thống có khả năng xác định được mốc thời gian và thay thế các mốc thời gian tương đối thành các mốc thời gian tuyệt đối bằng việc phân tích nội dung của văn bản.

Sự chồng chéo nội dung giữa các tài liệu: Một câu hỏi mà nhiều người đặt ra đối với tóm tắt đa văn bản đó là:

- Liệu có thể ghép các văn bản lại với nhau rồi sử dụng tóm tắt đơn văn: Câu trả lời ở đây là **không**!

Bằng cách đó chúng ta sẽ không tạo ra được một văn bản tóm tắt tốt do không loại bỏ được sự chống chéo về mặt nội dung cũng như xác định được mối quan hệ giữa các văn bản.

2.3. Các phương tóm tắt thông tin

Để thực hiện được việc tóm tắt thông tin ta cần thực hiện một số phương pháp sau:

Phương pháp loại bỏ chồng chéo và sắp xếp độ quan trọng giữa các văn bản trong cụm văn bản: Là một trong những vấn đề quan trọng nhất của bài toán tóm tắt đa văn bản. Một trong các phương pháp phổ biến để tính được độ quan trọng này là phương pháp MMR (Maximal Maginal Relevance) do Jaime Carbonell và Jade Goldstein đề xuất năm1998 [3]. Đầu vào của phương pháp này là một cụm văn bản đã được sắp xếp sẵn và đầu ra là cụm văn bản đã được sắp xếp lại theo thứ tự về ngữ nghĩa. Phương pháp này sắp xếp các văn bản dựa vào việc xác định một độ đo làm rõ ranh giới về ngữ nghĩa giữa các văn bản trong cụm. Mỗi một văn bản có độ đo này cực đại nếu độ đo về sự tương đồng giữa văn bản với câu truy vấn cao và cực tiểu được sự tương đồng giữa văn bản này và các văn bản khác đã được chọn trước đấy. Công thức để tính độ đo này như công thức (2) dưới đây:

$$MMT \stackrel{\text{def}}{=} Arg \max_{Di \in R/S} [\mathrm{Sim}_i(\mathrm{D}_i, Q) - (1-\lambda) \max_{\mathrm{D}_j \in S} [\lambda * (\mathrm{Sim}_2(\mathrm{D}_i, \mathrm{D}_j) \quad ((2)$$

Trong đó:

λ: là tham số nằm trong ngưỡng [0,1] để quyết định việc đóng góp giữa 2 độ đo. Nếu λ=1 thì độ quan trọng của văn bản chỉ phụ thuộc vào độ đo tương đồng giữa văn bản và câu truy vấn, còn nếu λ=0 thì độ đo sự tương đồng giữa văn bản này và văn bản khác sẽ đạt giá trị cực đại trong biểu thức trên.

C: cụm văn bản.

D_i: văn bản thuộc cụm C.

Q: là câu truy vấn (hay câu hỏi người dùng đưa vào).

R=IR (C, Q, θ): là tập các văn bản của C đã được sắp xếp thứ tự theo sự liên quan với câu truy vấn Q dựa vào một ngưỡng xác định θ.

S: là tập các văn bản của R đã được chọn.

R/S: là tập các văn bản chưa được chọn của R.

Sim_1, Sim_2: là độ đo về sự tương đồng giữa hai văn bản.

Phương pháp sắp xếp câu: Xác định độ quan trọng câu là bước xuất hiện hầu hết trong các phương pháp tóm tắt đơn văn bản cũng như tóm tắt đa văn bản hiện nay. Độ đo quan trọng này có thể được xây dựng bằng cách kết hợp nhiều độ đo độ tương đồng câu khác nhau với các phương pháp cải tiến từ phương pháp MMR để làm tăng độ quan trọng đối với mức ngữ nghĩa câu [7, 4, 3]. Công thức của phương pháp MMR được cải tiến cho mức ngữ nghĩa câu:

$$Score(S_i) = \arg\max_{S_i} [\lambda * \mathrm{sim}(s, q) - (1-\lambda) * \max sim(S_i, S_j)] \quad (3)$$

Trong đó:

λ: là tham số nằm trong ngưỡng [0,1] để quyết định việc đóng góp giữa 2 độ đo.

q: là câu truy vấn (hay câu hỏi người dùng đưa vào).

s_i: là một câu trong cụm văn bản.

s_j: các câu khác nằm trong cụm văn bản

sim: độ đo về sự tương đồng giữa hai câu.

2.4. Đề xuất giải pháp tăng cường tính ngữ nghĩa cho việc tóm tắt thông tin trên mạng xã hội

Thông thường, để xây dựng các độ đo tương đồng ngữ nghĩa tốt, phương pháp phổ biến là sử dụng việc kết hợp nhiều độ đo lại với nhau thông qua một hàm tính hạng tuyến tính. Công thức biểu diễn việc kết hợp các độ đo như sau (4):

$$SimTotal(S_1, S_2) = \sum_i \alpha_i \, sim_i(S_1, S_2)$$

Với điều kiện: (4)

$$\sum_i \alpha_i = 1$$

Trong đó:

- s_1, s_2: là hai câu cần tính độ tương đồng
- i: là số lượng các độ đo tương đồng kết hợp lại
- sim_i: là các độ đo tương đồng thành phần
- α_i: là các hằng số trộn nằm trong ngưỡng [0,1] thể hiện sự đóng góp của các độ đo tương đồng thành phần với độ đo SimTotal. Các tham số này phải thỏa mãn điều kiện, tổng tất cả các hằng số trong công thức bằng 1 (Các hằng số này sẽ được ước lượng trong quá trình thực nghiệm).

Từ những nghiên cứu được nêu ở các mục trên, tác giả đã đưa ra một mô hình tóm tắt đa văn bản cho các cụm dữ liệu trên trang mạng xã hội trả về từ máy tìm kiếm như sau:

Pha tiền xử lý dữ liệu

Pha xử lý này nhận đầu vào tập các trang web thuộc trên mạng xã hội một cụm dữ liệu. Các quá trình được thực hiện theo các bước sau:

Bước 1. Loại bỏ các trang web có nội dung trùng lặp.

Bước 2. Lọc nhiễu, loại bỏ các thẻ HTML, lấy nội dung chính của trang Web.

Bước 3. Tách từ, tách câu các văn

Pha sắp xếp văn bản và câu theo độ quan trọng

Pha này nhận dữ liệu đầu vào là các văn bản và nhãn cụm đã qua tiền xử lý, đầu ra là danh sách các câu, các văn bản đã được sắp xếp lại theo độ quan trọng về mặt ngữ nghĩa.

Vì máy nhận dữ liệu đầu vào là các văn bản và nhãn cụm đã qua tiền xử lý, đầu ra là cụm đã qua tiền xử lý, đầu ra là danh sách các câu, các văn bản đã được sắp

Pha sinh văn bản ban đan

Trong pha sinh văn bản tóm tắt, các câu được sắp xếp đã được sắp xếp ở pha trên sẽ được sắp xếp lại. Trọng số độ quan trọng của câu sẽ được bổ sung thêm trọng số của văn bản chưa câu đấy, việc này sẽ giúp văn bản tóm tắt không có sự chồng chéo về mặt nội dung.

2.5. Đánh giá

Cả hai vấn đề cần giải quyết trong bài toán tóm tắt đa văn bản (dựa vào trích xuất câu đều tập trung vào việc xác định được sự tương đồng giữa hai văn bản nói chung và giữa hai câu nói riêng. Trên thực tế, các phương pháp áp dụng và cải tiến cho tóm tắt đa văn bản dựa vào đều tập trung vào vấn đề là tăng cường tính ngữ nghĩa cho độ đo tương đồng giữa hai câu hay hai văn bản [4, 7, 6]. Chính vì vậy nhóm tác giả đã thông qua việc khắc phục các hạn chế của việc tóm tắt văn bản để đưa ra giải pháp tổng thể nhằm áp dụng triệt để cho bài toán phân tích thông tin trên mạng xã hội.

3. Kết luận

Trong thế giới hiện tại, việc thu thập và tóm tắt thông tin giúp cho việc xử lý thông tin được nhanh hơn. Các phương pháp thu thập và tóm tắt thông tin trên mạng xã hội là một giải pháp giúp các doanh nghiệp, các đơn vị sự nghiệp có thể sử dụng để nâng cao khả năng tiếp cận thông tin, giảm công sức thực hiện từ đó giảm chi phí cho vấn đề lên kế hoạch thực hiện thu thập thông tin.

Tài liệu tham khảo

[1]. Blake,C., Kampov, J., Orphanides, A., West,D., & Lown, C. (2007). UNC- CH at DUC 2007: Query Expansion, Lexical Simplification, and Sentence Selection Strategies for Multi-Document Summarization, In DUC07.

[2]. D. Bollegara, Y. Matsuo, and M. Ishizuka (2006). Extracting key phrases to disambiguate personal names on the web, In CICLing 2006.

[3]. Jaime Carbonell, Jade Goldstein (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, In SIGIR-98, Melbourne, Australia, Aug. 1998.

[4]. K. Filippova, M. Mieskes, V. Nastase, S. Paolo Ponzetto, M. Strube (2007). Cascaded Filtering for Topic-Driven Multi-Document Summarization, In EML Research gGmbH, 2007.

[5]. H. Luhn (1958). The automatic creation of literature abstracts, IBM Journal of Research and Development, 2(2):P.159-165, 1958.

[6]. Inderjeet Mani and Mark T. Maybury (eds) (1999). Advances in Automatic Text Summarization, MIT Press, 1999, ISBN 0-262-13359-8.

[7]. B. Hachey, G. Murray, D. Reitter (2005). Query-Oriented Multi-Document Summarization With a Very Large Latent Semantic Space, In The Embra System at DUC, 2005.

STUDY ON COLLECTION METHODS AND SUMMARY OF INFORMATION IN SOCIAL NETWORKS

Abstract:

This paper presents the process of studying some methods of collecting and summarizing information in social networks such as Facebook, google + and orther news. The author has implemented a solution to collect information from social networks, then implemented solutions to summarize the text, thereby optimizing the method used to classify information on social networks.

Keyword: *Collect information, information analysis, information evaluates.*