BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ QUỐC PHÒNG
VIỆN KHOA HỌC VÀ CÔNG NGHỆ QUÂN SỰ

NGUYỄN NHẬT AN

NGHIÊN CỨU, PHÁT TRIỂN CÁC KỸ THUẬT TỰ ĐỘNG TÓM TẮT VĂN BẢN TIẾNG VIỆT

LUẬN ÁN TIẾN SĨ TOÁN HỌC

HÀ NỘI – 2015

BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ QUỐC PHÒNG
VIỆN KHOA HỌC VÀ CÔNG NGHỆ QUÂN SỰ

NGUYỄN NHẬT AN

NGHIÊN CỨU, PHÁT TRIỂN CÁC KỸ THUẬT TỰ ĐỘNG TÓM TẮT VĂN BẢN TIẾNG VIỆT

Chuyên ngành : Cơ sở toán học cho tin học

Mã số : 62 46 01 10

LUẬN ÁN TIẾN SĨ TOÁN HỌC

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. TSKH NGUYỄN QUANG BẮC
2. PGS.TS NGUYỄN ĐỨC HIẾU

HÀ NỘI - 2015

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả trong luận án là trung thực và chưa từng công bố trong bất kỳ công trình nào khác.

Tác giả

Nguyễn Nhật An

LỜI CẢM ƠN

Luận án được thực hiện tại Viện Công nghệ thông tin - Viện Khoa học Công nghệ quân sự - Bộ Quốc phòng, dưới sự hướng dẫn khoa học của Thiếu tướng, TSKH Nguyễn Quang Bắc và Đại tá PGS.TS Nguyễn Đức Hiếu.

Trước tiên tôi xin bày tỏ lòng biết ơn sâu sắc tới tập thể giáo viên hướng dẫn, những người đã đưa tôi đến với lĩnh vực nghiên cứu này. Các thầy đã tận tình giảng dạy, hướng dẫn giúp tôi tiếp cận và đạt được thành công trong các nghiên cứu của mình; luôn tận tâm động viên, khuyến khích và chỉ dẫn giúp tôi hoàn thành được bản luận án này.

Tôi xin bày tỏ lòng biết ơn tới Đảng uỷ, ban lãnh đạo, các cán bộ Phòng Quản trị Cơ sở dữ liệu - Viện Công nghệ thông tin và Phòng Đào tạo - Viện Khoa học Công nghệ quân sự, đã tạo mọi điều kiện thuận lợi giúp đỡ tôi trong quá trình học tập và nghiên cứu tại đơn vị.

Tôi xin cảm ơn PGS.TS Đào Thanh Tĩnh, TS Nguyễn Phương Thái, TS Nguyễn Thị Thu Hà, TS. Đỗ Đức Đông và TS Ngôn ngữ học Phan Thị Nguyệt Hoa đã chia sẻ những tài liệu và kinh nghiệm nghiên cứu.

Cuối cùng, tác giả xin chân thành cảm ơn các thành viên trong Gia đình, những người luôn dành cho tác giả những tình cảm nồng ấm và sẻ chia những lúc khó khăn trong cuộc sống, luôn động viên giúp đỡ tác giả trong quá trình nghiên cứu. Luận án cũng là món quà tinh thần mà tác giả trân trọng gửi tặng đến các thành viên trong Gia đình.

MỤC LỤC

Trang

DANH MỤC CÁC KÍ HIỆU, CÁC CHỮ VIẾT TẮT

d	văn bản
D	tập văn bản huấn luyện (gốc)
m	số văn bản huấn luyện
SH	tập các văn bản tóm tắt hệ thống
sh	văn bản do hệ thống tóm tắt
s	câu văn bản
a	tỷ lệ tóm tắt
f	tập các đặc trưng văn bản
p	tập các phương pháp tóm tắt văn bản
k	tập hệ số đặc trưng hoặc phương pháp
$Score(s)$	giá trị trọng số của câu s
$Sim(s_1, s_2)$	Hàm tính độ tương tự giữa văn bản s_1 và s_2
$F(k)$	Hàm thích nghi (mục tiêu) theo bộ hệ số k
G^{max}	số vòng lặp (điều kiện dừng)
ACO	Tối ưu đàn kiến (Ant Colony Optimization)
AS	Tóm tắt tóm lược (Abstraction Summarization)
CRF	Miền ngẫu nhiên điều kiện (Conditional Random Field)
CSSD	Cloneproof Schwartz Sequential Dropping
EA	Giải thuật tiến hóa (Evolutionary Algorithm)
ES	Tóm tắt trích rút (Extraction Summarization)
GA	Giải thuật di truyền (Genetic Algorithm)
GP	Lập trình di truyền (Genetic Programming)
HMM	Mô hình Markov ẩn (Hidden Markov Model)
LCS	Chuỗi con chung dài nhất (Longest Common Subsequence)
LSA	Phân tích ngữ nghĩa tiềm ẩn (Latent Semantic Analysis)
MEM	Mô hình cực đại hóa Entropy (Maximum Entropy Model)

NLP	Xử lý ngôn ngữ tự nhiên (Natural Language Processing)
NMF	Phép nhân tử hóa ma trận không âm (Non-negative Matrix Factorization)
PGA	Giải thuật di truyền song song (Parallel Genetic Algorithms)
ROUGE	Độ đo đánh giá độ tương tự văn bản (Recall-Oriented Understudy for Gisting Evaluation)
RST	Lý thuyết cấu trúc tu từ (Rhetorical Structure Theory)
SDD	Khai triển ma trận nửa rời rạc (Semi-discrete Matrix Decomposition)
SSD	Schwartz Sequential Dropping
SVD	Phương pháp phân tích giá trị đơn (Singular Value Decomposition)
SVM	Máy vector hỗ trợ (Support Vector Machine)
TF	Tần suất thuật ngữ (Term Frequency)
TF.ISF	Tần suất từ - nghịch đảo tần suất câu (Term frequency- inverse sentence frequency)
TTVB	Tóm tắt văn bản
TTĐVB	Tóm tắt đơn văn bản
n-gram	Mô hình ngôn ngữ n-gram [81]
unigram	Mô hình n-gram với gram là một từ (1-gram)
Voting	Bầu chọn
Vietworknet	Mạng từ tiếng Việt
Wordnet	Mạng từ

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, THUẬT TOÁN

MỞ ĐẦU

1. Tình hình nghiên cứu trong nước và ngoài nước

Trong thời gian gần đây, với sự phát triển nhanh chóng của các dịch vụ trực tuyến và công nghệ lưu trữ hiện đại, thông tin văn bản được lưu trữ trên mạng Internet trở nên vô cùng lớn. Hằng ngày, số lượng thông tin văn bản tăng lên không ngừng. Lượng thông tin văn bản khổng lồ đó đã và đang mang lại lợi ích không nhỏ cho con người. Tuy nhiên, nó gây ra sự quá tải thông tin khiến chúng ta gặp nhiều khó khăn trong việc tìm kiếm và tổng hợp thông tin. Để cải thiện tìm kiếm cũng như tăng hiệu quả cho việc xử lý thông tin, tóm tắt văn bản tự động là giải pháp không thể thiếu để giải quyết vấn đề này.

Trên thế giới, bài toán tóm tắt văn bản xuất hiện từ rất lâu. Những kỹ thuật đầu tiên áp dụng để tóm tắt văn bản đã được đề xuất từ những năm 50 của thế kỷ trước [47],[17]. Sau đó, chúng tiếp tục được nghiên cứu và đạt nhiều kết quả ngày càng tốt hơn cho nhiều loại ngôn ngữ như tiếng Anh, tiếng Pháp, tiếng Nhật, tiếng Trung… Các nghiên cứu tập trung vào hai hướng chính: tóm tắt trích rút ES (Extraction Summarization) và tóm tắt tóm lược AS (Abstraction Summarization) [37] cho bài toán tóm tắt đơn văn bản (bản tóm tắt được tạo thành từ một văn bản) và đa văn bản (văn bản tóm tắt được tạo thành từ nhiều văn bản cùng chủ đề). Hầu hết các nghiên cứu về tóm tắt văn bản là ES vì nó dễ thực hiện và có tốc độ nhanh hơn so với AS. Hướng tiếp cận ES chủ yếu là dựa vào các đặc trưng quan trọng của văn bản để tính trọng số câu để trích rút. Trong khi đó, AS là dựa vào các kỹ thuật xử lý ngôn ngữ tự nhiên kết hợp với thông tin về ngôn ngữ để tạo ra các tóm tắt cuối cùng.

Đối với tiếng Việt, do tính phức tạp và đặc thù riêng của nó, số lượng những nghiên cứu về tóm tắt văn bản tiếng Việt so với tiếng Anh vẫn còn ít. Phần lớn các nghiên cứu mới chỉ là các nghiên cứu ở mức đề tài tốt nghiệp đại học, luận văn thạc sĩ, tiến sĩ và đề tài KHCN cấp bộ [5],[9],[13],[55],[57],[76].

Các bài báo công bố kết quả nghiên cứu về tóm tắt văn bản phần lớn dựa trên hướng trích rút cho bài toán tóm tắt đơn văn bản. Tuy nhiên vẫn có hai hướng là tóm tắt trích rút và tóm tắt theo tóm lược. Mặt khác, do chưa có kho ngữ liệu chuẩn phục vụ cho tóm tắt văn bản tiếng Việt nên hầu hết thử nghiệm của các nghiên cứu đều dựa trên các kho ngữ liệu tự xây dựng. Do vậy, việc đánh giá hiệu quả của từng phương pháp chưa được khách quan và cần phải xem xét một cách kỹ lưỡng.

2. Tính cấp thiết

Với sự bùng nổ thông tin lưu trữ trên các hệ thống máy tính và trên Internet, một lượng thông tin khổng lồ được lưu trữ trên đó. Để khai thác hiệu quả lượng thông tin khổng lồ này cần phải có các hệ thống xử lý ngôn ngữ tự nhiên đủ mạnh. Tóm tắt văn bản là một trong những bài toán quan trọng đó.

Bài toán tóm tắt văn bản tiếng Việt đóng một vai trò quan trọng trong việc khai thác hiệu quả thông tin trong kho ngữ liệu văn bản tiếng Việt lớn. Nó có ứng dụng rất lớn trong các hệ thống như: tìm kiếm thông minh, đa ngôn ngữ, tổng hợp thông tin... Đối với lĩnh vực an ninh quốc phòng, tóm tắt tin tức có thể giúp cho cán bộ nghiệp vụ thu thập đủ các thông tin cần thiết và kịp thời theo dõi, đánh giá, xử lý nguồn thông tin một cách nhanh chóng [CT1].

Do tính chất quan trọng như vậy, hiện nay bài toán tóm tắt văn bản tiếng Việt đã được các nhà nghiên cứu xử lý ngôn ngữ trong nước quan tâm. Tuy nhiên, số lượng cũng như chất lượng các nghiên cứu còn khá khiêm tốn. Nguyên nhân của những vấn đề này có thể xuất phát từ những lý do sau:

- Nghiên cứu xử lý ngôn ngữ tự nhiên tiếng Việt đang tập trung vào những vấn đề cơ bản của tiếng Việt như:
  - Giải quyết bài toán tách từ, gán nhãn từ loại, cây cú pháp.
  - Xây dựng kho ngữ liệu: tách từ, gán nhãn từ loại.
  - Xây dựng wordnet tiếng Việt…

đây là những bước tiền xử lý cho bài toán Tóm tắt văn bản tiếng Việt.

- Chưa xác định được đầy đủ các đặc trưng quan trọng của văn bản tiếng Việt và xác định ảnh hưởng của từng đặc trưng trong bài toán tóm tắt văn bản tiếng Việt.
- Chưa xây dựng được kho ngữ liệu tiếng Việt chuẩn và lớn dùng cho huấn luyện và đánh giá trong bài toán tóm tắt văn bản tiếng Việt.
- Chưa có một hệ thống tóm tắt văn bản tiếng Việt hoàn chỉnh nào được công bố rộng rãi cho cộng đồng sử dụng, nghiên cứu.

Vì thế, đề tài luận án **"Nghiên cứu, phát triển các kỹ thuật tự động tóm tắt văn bản tiếng Việt"** có tính cấp thiết và tính ứng dụng thực tiễn cao, nhất là trong lĩnh vực an ninh quốc phòng.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của Luận án:

- Các phương pháp tóm tắt văn bản trên thế giới.
- Các phương pháp đánh giá tóm tắt văn bản.
- Các phương pháp tóm tắt văn bản tiếng Việt.
- Các đặc trưng quan trọng của văn bản tiếng Việt.
- Các giải thuật tối ưu phỏng sinh học.
- Kho ngữ liệu huấn luyện tóm tắt văn bản.
- Kho ngữ liệu đánh giá tóm tắt văn bản.

Phạm vi nghiên cứu của Luận án:

- Luận án tập trung nghiên cứu, đề xuất phương pháp mới nâng cao độ chính xác trong bài toán tóm tắt đơn văn bản tiếng Việt theo hướng trích rút.

4. Mục tiêu nghiên cứu

Mục tiêu của luận án là nghiên cứu các đặc trưng quan trọng của văn bản cho bài toán tóm tắt đơn văn bản tiếng Việt. Qua đó đề xuất hai phương pháp tóm tắt văn bản tiếng Việt: một là, phương pháp tóm tắt văn bản tiếng Việt dựa

trên bộ hệ số đặc trưng văn bản, bộ hệ số này được xác định bằng phương pháp học máy sử dụng giải thuật tối ưu phỏng sinh học. Hai là, phương pháp tóm tắt văn bản tiếng Việt bằng kỹ thuật Voting (bầu chọn) có hệ số phương pháp trên cơ sở kế thừa kết quả của các phương pháp tóm tắt văn bản trước đây.

Mục tiêu cụ thể:

- Nghiên cứu các đặc trưng quan trọng của văn bản tiếng Việt, qua đó đề xuất lựa chọn tập đặc trưng để đưa vào mô hình.
- Đề xuất phương pháp tóm tắt văn bản tiếng Việt dựa trên bộ hệ số đặc trưng văn bản, bộ hệ số này được xác định bằng phương pháp học máy sử dụng giải thuật tối ưu phỏng sinh học.
- Đề xuất mô hình tóm tắt văn bản tiếng Việt sử dụng kỹ thuật Voting (bầu chọn) có hệ số phương pháp được xác định thông qua quá trình học văn bản tóm tắt mẫu bằng phương pháp học máy.

5. Phương pháp nghiên cứu

- Dựa trên các phương pháp tóm tắt văn bản của thế giới và trong nước.
- Dựa trên phân tích các hạn chế của các phương pháp tóm tắt văn bản tiếng Việt.
- Đề xuất các phương pháp tóm tắt văn bản tiếng Việt mới dựa trên một số mô hình toán học phù hợp (phỏng sinh học, voting…).
- Kiểm chứng kết quả các phương pháp đề xuất bằng thực nghiệm.

6. Nội dung nghiên cứu

- Nghiên cứu và đề xuất lựa chọn 8 đặc trưng quan trọng cho bài toán tóm tắt văn bản tiếng Việt bằng phương pháp khảo sát trên kho ngữ liệu văn bản tiếng Việt:

  - Vị trí câu.
  - Độ dài câu.
  - Tần suất từ - nghịch đảo tần suất câu (TFxISF).

- Xác suất thực từ.
- Thực thể tên.
- Dữ liệu số.
- Tương tự với tiêu đề.
- Câu trung tâm.

- Nghiên cứu và đề xuất hai phương pháp tóm tắt văn bản tiếng Việt mới:
  - Phương pháp tóm tắt văn bản tiếng Việt dựa vào bộ hệ số đặc trưng: Xác định bộ hệ số đặc trưng văn bản nêu trên bằng phương pháp học máy trên kho ngữ liệu tóm tắt mẫu của nhiều lĩnh vực khác nhau. Sau khi xác định các hệ số đặc trưng, thực hiện tóm tắt văn bản thông qua sự kết hợp tuyến tính của 8 đặc trưng đó.
  - Phương pháp tóm tắt văn bản tiếng Việt sử dụng kỹ thuật Voting: Ý tưởng của phương pháp này là xem kết quả của mỗi phương pháp tóm tắt văn bản khác nhau là lá phiếu đã được sắp xếp thứ tự ưu tiên theo trọng số của các câu (số lá phiếu giống nhau được định nghĩa là hệ số phương pháp được xác định thông quá trình học kho ngữ liệu tóm tắt mẫu), sử dụng kỹ thuật Voting để lựa chọn các câu có trọng số voting cao dựa trên các lá phiếu.

7. Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học: Nghiên cứu chuyên sâu và có hệ thống về văn bản tiếng Việt và bài toán tóm tắt văn bản tiếng Việt. Làm rõ cơ sở toán học của các đặc trưng văn bản tiếng Việt và phương pháp tiếp cận mới, góp phần giải quyết các bài toán tóm tắt văn bản tiếng Việt sau này.

Ý nghĩa thực tiễn: Nghiên cứu xây dựng tập đặc trưng văn bản quan trọng của tiếng Việt và phương pháp xác định các hệ số đặc trưng trong bài toán tóm tắt văn bản tiếng Việt. Nghiên cứu kỹ thuật Voting và ứng dụng trong bài toán tóm tắt văn bản tiếng Việt. Kết quả của hai phương pháp mới này cho kết quả

khả quan và có thể áp dụng xây dựng các phần mềm tóm tắt văn bản tiếng Việt chất lượng cao phục vụ trong nhiều lĩnh vực, nhất là lĩnh vực an ninh quốc phòng. Ngoài ra, kho ngữ liệu tiêu chuẩn có chú giải do tác giả xây dựng có thể đóng góp vào cộng đồng nghiên cứu xử lý ngôn ngữ tự nhiên tiếng Việt.

8. Bố cục của luận án

Luận án gồm 03 chương cùng với các phần mở đầu, kết luận, phụ lục, tài liệu tham khảo và danh mục các công trình nghiên cứu đã công bố của tác giả.

Chương 1: Tổng quan về tóm tắt văn bản và tóm tắt văn bản tiếng Việt.

Trong chương này, luận án trình bày tổng quan về bài toán tóm tắt văn bản, các phương pháp giải quyết, các phương pháp đánh giá tóm tắt văn bản; Hiện trạng các nghiên cứu về tóm tắt văn bản tiếng Việt. Ngoài ra luận án còn đề cập những kiến thức cơ sở liên quan là giải thuật di truyền và phương pháp voting Schulze. Các nghiên cứu trên là tiền đề để phát triển các phương pháp tóm tắt văn bản tiếng Việt được trình bày trong chương 2 và chương 3.

Chương 2: Tóm tắt văn bản tiếng Việt dựa trên bộ hệ số đặc trưng.

Trong chương này, luận án trình bày các kết quả nghiên cứu mới về phương pháp tóm tắt văn bản tiếng Việt dựa trên bộ hệ số đặc trưng, bao gồm: Lựa chọn 8 đặc trưng quan trọng của văn bản tiếng Việt; Xác định các hệ số đặc trưng quan trọng của văn bản tiếng Việt bằng phương pháp học máy sử dụng giải thuật di truyền GA và giải thuật tối ưu đàn kiến ACO thông qua kho ngữ liệu tóm tắt mẫu; Các thử nghiệm.

Chương 3: Tóm tắt văn bản tiếng Việt sử dụng kỹ thuật Voting

Trong chương này, luận án trình bày các kết quả nghiên cứu mới về phương pháp tóm tắt văn bản tiếng Việt dựa trên kỹ thuật Voting và các thử nghiệm.

Chương 4: Quy trình xây dựng kho ngữ liệu có chú giải cho bài toán tóm tắt văn bản tiếng Việt

Trong chương này, luận án trình bày đề xuất về quy trình xây dựng kho ngữ liệu có chú giải dùng cho huấn luyện và đánh giá trong bài toán tóm tắt Văn bản tiếng Việt bao gồm các giai đoạn thu thập, xây dựng bản tóm tắt con người, chú giải cấu trúc hóa và lưu trữ. Ngoài ra luận án còn trình bày các phương pháp đánh giá kho ngữ liệu xây dựng.

Phụ lục.

Trong phần này, luận án trình bày kho ngữ liệu tiêu chuẩn có chú giải ViEvTEXTSUM do tác giả xây dựng, kho ngữ liệu Corpus_LTH của đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt" và phần thử nghiệm.

CHƯƠNG 1. TỔNG QUAN VỀ TÓM TẮT VĂN BẢN VÀ TÓM TẮT VĂN BẢN TIẾNG VIỆT

Trong chương này, luận án giới thiệu tổng quan về tóm tắt văn bản và tóm tắt văn bản tiếng Việt bao gồm các khái niệm cơ bản, các phương pháp tiếp cận tóm tắt văn bản và các phương pháp đánh giá. Bên cạnh đó, luận án cũng trình bày đặc điểm của tiếng Việt, hiện trạng về nghiên cứu xử lý ngôn ngữ tự nhiên tiếng Việt và các phương pháp tóm tắt văn bản tiếng Việt đã công bố. Ngoài ra luận án còn đề cập những nội dung cơ bản về giải thuật di truyền, giải thuật tối ưu đàn kiến và phương pháp voting Schulze, đây là kiến thức cơ sở liên quan được sử dụng trong Chương 2 và Chương 3.

1.1 Giới thiệu về tóm tắt văn bản

Như trên đã nêu, các nghiên cứu về phương pháp tóm tắt văn bản tập trung vào hai hướng chính: tóm tắt trích rút và tóm tắt tóm lược. Tóm tắt văn bản theo hướng trích rút dễ thực hiện và có tốc độ nhanh hơn so với tóm tắt tóm lược. Hướng tiếp cận tóm tắt trích rút chủ yếu là dựa vào các đặc trưng quan trọng của văn bản để tính trọng số câu để trích rút. Trong khi đó, tóm tắt tóm lược là dựa vào các kỹ thuật xử lý ngôn ngữ tự nhiên kết hợp với thông tin về ngôn ngữ để tạo ra các tóm tắt cuối cùng.

Bài toán tóm tắt văn bản được nêu như sau:

"Tóm tắt văn bản là quá trình trích rút những thông tin quan trọng nhất từ một hoặc nhiều nguồn để tạo ra phiên bản cô đọng, ngắn gọn phục vụ cho một hoặc nhiều người dùng cụ thể, hay một hoặc nhiều nhiệm vụ cụ thể" [48]

Ví dụ minh hoạ về tóm tắt văn bản với 120 từ:

> *Ngày 11/4, Đại sứ Liên bang Nga tại Việt Nam Andrey Kovtun cùng đoàn công tác đã thăm và làm việc với tỉnh Ninh Thuận về tình hình triển khai xây dựng nhà máy điện hạt nhân Ninh Thuận.*
>
> *Tại buổi làm việc, Chủ tịch Ủy ban Nhân dân tỉnh Ninh Thuận Nguyễn Đức Thanh cho biết tỉnh đã hoàn chỉnh chính sách, cơ chế đặc thù và đã trình Thủ tướng Chính phủ phê duyệt. Tỉnh cũng đã hoàn thành công tác đo đạc lập bản đồ*

thu hồi đất và quy chủ sử dụng đất tại các khu vực triển khai dự án gồm khu vực thu hồi xây dựng nhà máy, khu tái định cư, khu nghĩa trang và hệ thống cấp nước phục vụ khu tái định cư nhà máy điện hạt nhân Ninh Thuận 1; đồng thời hoàn thành công tác kiểm kê khu vực vùng lõi nhà máy.

Hiện nay tỉnh đã hoàn thành việc khảo sát đo đạc địa hình, địa chất phục vụ công tác lập quy hoạch và dự án đầu tư; hoàn thành công tác lập quy hoạch chi tiết khu tái định cư nhà máy 1 với diện tích 86,9 ha và khu nghĩa trang với diện tích hơn 10,8 ha.

Tỉnh cũng đã thành lập Ban Quản lý dự án điện hạt nhân để thực hiện dự án di dân, tái định cư do Ủy ban Nhân dân tỉnh làm chủ đầu tư. Bên cạnh đó, tỉnh phấn đấu hoàn thành công tác bồi thường, giải phóng mặt bằng, đồng thời tổ chức thi công xây dựng các công trình hạ tầng phục vụ di dân tái định cư gồm khu tái định cư tập trung, nghĩa trang và hệ thống cấp nước phục vụ khu tái định cư nhà máy điện hạt nhân Ninh Thuận 1.

Theo quy hoạch được duyệt, khu tái định cư tập trung là khu nằm trong vành đai du lịch, do đó sẽ đầu tư đồng bộ hệ thống hạ tầng kỹ thuật, hạ tầng xã hội theo tiêu chuẩn khu đô thị. Ngoài ra khi được bàn giao mốc ranh giới, mốc hàng rào nhà máy điện hạt nhân, tỉnh sẽ xác định cụ thể vị trí, quy mô xây dựng khu tái định canh, đảm bảo ổn định và phát triển sản xuất lâu dài cho người dân.

Đại sứ Andrey Kovtun đánh giá cao công tác chuẩn bị cho việc xây dựng nhà máy điện hạt nhân Ninh Thuận 1. Phía Nga luôn ưu tiên cao nhất cho Việt Nam trong công tác xây dựng nhà máy điện hạt nhân, dự kiến cuối năm 2013, Nga sẽ hoàn thành hồ sơ triển khai xây dựng nhà máy điện hạt nhân số 1 tại Ninh Thuận, đồng thời sẽ tổ chức hội thảo tại Ninh Thuận để các công ty, các doanh nghiệp của tỉnh và cả nước tham gia đầu tư vào các ngành công nghiệp phụ trợ cho xây dựng nhà máy điện hạt nhân.

Tỉnh Ninh Thuận mong muốn nhận được sự quan tâm, hỗ trợ của Chính phủ Liên bang Nga trong việc đào tạo nguồn nhân lực cho các lĩnh vực khác tỉnh đang có nhu cầu (ngoài chương trình đào tạo của Chính phủ hai nước đã hợp tác), đồng thời hỗ trợ tỉnh trong việc xúc tiến đầu tư, vận động các doanh nghiệp Nga đầu tư vào tỉnh trong các lĩnh vực sản xuất, chuyển giao công nghệ phục vụ cho việc xây dựng nhà máy điện hạt nhân và các ngành công nghiệp phụ trợ.

Hình 1-1 Văn bản gốc.

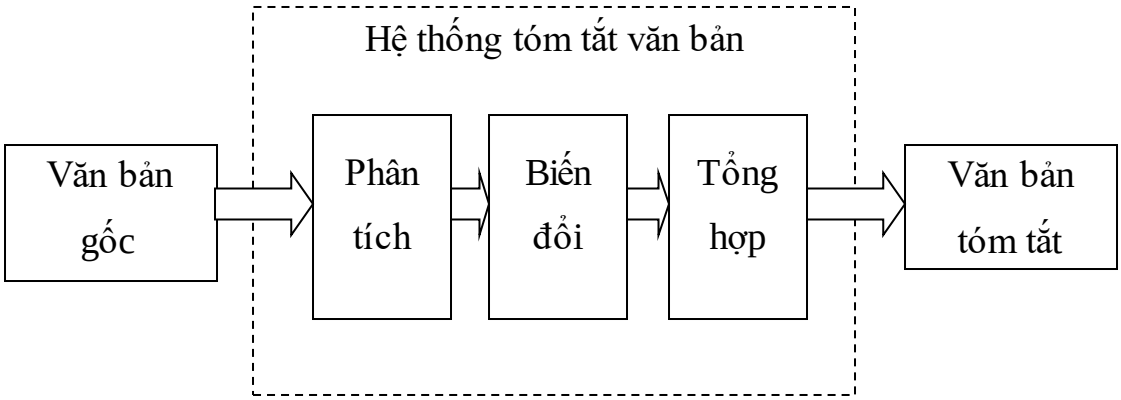
Ngày 11/4, Đại sứ Liên bang Nga tại Việt Nam Andrey Kovtun cùng đoàn công tác đã thăm và làm việc với tỉnh Ninh Thuận về tình hình triển khai xây dựng nhà máy điện hạt nhân Ninh Thuận. Phía Nga luôn ưu tiên cao nhất cho Việt Nam trong công tác xây dựng nhà máy điện hạt nhân, dự kiến cuối năm 2013, Nga sẽ hoàn thành hồ sơ triển khai xây dựng nhà máy điện hạt nhân số 1 tại Ninh Thuận, đồng thời sẽ tổ chức hội thảo tại Ninh Thuận để các công ty, các doanh nghiệp của tỉnh và cả nước tham gia đầu tư vào các ngành công nghiệp phụ trợ cho xây dựng nhà máy điện hạt nhân.

Hình 1-2 Văn bản tóm tắt với 120 từ.

1.1.1 Các giai đoạn và các tham số của hệ thống tóm tắt văn bản

Theo quan điểm của các nhà nghiên cứu TTVB thì bản tóm tắt là một bản rút gọn của văn bản gốc thông qua việc lựa chọn và tổng quát hóa các khái niệm quan trọng [34],[48],[35]. Hệ thống tóm tắt văn bản tự động được chia thành 3 giai đoạn chính:

- Phân tích (Analysis or Interpretation): Phân tích văn bản đầu vào để đưa ra những mô tả bao gồm các thông tin dùng để tìm kiếm, đánh giá các đơn vị ngữ liệu quan trọng cũng như các tham số đầu vào cho việc tóm tắt.
- Biến đổi (Transformation): Lựa chọn các thông tin trích chọn được, biến đổi để giản lược và thống nhất, kết quả là các đơn vị ngữ liệu đã được tóm tắt.
- Tổng hợp (Synthesis or Realization): Từ các đơn vị ngữ liệu đã tóm tắt, tạo văn bản mới chứa những điểm chính, quan trọng của văn bản gốc.



Hình 1-3 Các giai đoạn của hệ thống tóm tắt

Các giai đoạn của quá trình tóm tắt văn bản chịu ảnh hưởng bởi các tham số khác nhau như các tham số đầu vào, đầu ra và các tham số mục đích [37],[35].

Các tham số đầu vào: Các đặc trưng của văn bản đầu vào có thể ảnh hưởng tới kết quả tóm tắt theo các yếu tố sau:

- Cấu trúc của văn bản: Cấu trúc là tổ chức của một văn bản cho trước như tiêu đề, nội dung, đoạn (paragraph),… Cấu trúc của một văn bản có thể cung cấp rất nhiều thông tin khi tạo bản tóm tắt.

- Kích thước: Kích thước là độ dài của văn bản cho trước tính theo đơn vị thuật ngữ, ví dụ như tài liệu nghiên cứu dài thường đề cập nhiều chủ đề ít thuật ngữ lặp lại trong khi văn bản ngắn chỉ trình bày một chủ đề nhưng chứa nhiều thuật ngữ lặp lại hơn.
- Ngôn ngữ: Ngôn ngữ được sử dụng trong văn bản đầu vào có thể ảnh hưởng tới kết quả tóm tắt. Các thuật toán tóm tắt có thể có sử dụng hoặc không sử dụng thông tin ngôn ngữ.
- Lĩnh vực: Văn bản đầu vào thường liên quan tới một lĩnh vực cụ thể nào đó. Do đó, người ta có thể sử dụng các tri thức (như kho ngữ liệu) liên quan đến lĩnh vực đó để tạo ra bản tóm tắt tốt hơn.
- Đơn vị: Nếu một bản tóm tắt được tạo thành từ một văn bản riêng lẻ thì hệ thống tóm tắt đó được gọi là hệ thống tóm tắt đơn văn bản (single-document). Nếu một bản tóm tắt được tạo thành từ nhiều văn bản liên quan tới một chủ đề riêng lẻ thì hệ thống tóm tắt đó gọi là hệ thống tóm tắt đa văn bản (multi-document).

Các tham số mục đích: Các hệ thống tóm tắt tự động có thể tạo ra các bản tóm tắt tổng quát của một văn bản cho trước, hay có thể tạo ra các bản tóm tắt cho một tác vụ được định nghĩa trước. Các yếu tố sau đây có liên quan tới các tham số mục đích của các hệ thống tóm tắt.

- Tình huống: Tình huống liên quan tới ngữ cảnh của bản tóm tắt. Môi trường mà ta sẽ sử dụng bản tóm tắt, giả sử như người ta sử dụng bản tóm tắt khi nào và nhằm mục đích gì, có thể biết trước hoặc không.
- Chủ đề: Nếu ta biết trước mối quan tâm của người đọc thì ta có thể tạo ra các bản tóm tắt có liên quan tới chủ đề đó.
- Mục đích sử dụng: Tham số này quan tâm tới mục đích tạo ra bản tóm tắt như để xem qua trước khi đọc toàn bộ văn bản,...

Các tham số đầu ra: Bản tóm tắt có thể ảnh hưởng bởi các tham số đầu

ra như sau:

- Tài nguyên: Bản tóm tắt của một văn bản có thể liên quan tới tất cả các khái niệm xuất hiện trong văn bản, hoặc có thể liên quan tới các khái niệm đã chọn trước. Thường thì các hệ thống tóm tắt tổng quát có thể nắm bắt tất cả các khái niệm trong văn bản. Trong các hệ thống tóm tắt hướng người dùng như các hệ thống tóm tắt dựa trên truy vấn chẳng hạn, thì bản tóm tắt có thể chứa các khái niệm liên quan tới nhu cầu của người dùng.
- Định dạng: Bản tóm tắt khi tạo ra có thể được tổ chức thành các trường (như sử dụng các heading chẳng hạn) hoặc có thể được tổ chức như một văn bản không cấu trúc (như phần tóm tắt của một bài báo).
- Văn phong (style): Một bản tóm tắt có thể chứa nhiều thông tin (informative), mang tính ngụ ý (indicative), kết tụ (aggregative) hoặc mang tính chất bình phẩm (critical). Các bản tóm tắt chứa nhiều thông tin cho ta thông tin về các khái niệm được nhắc đến trong văn bản đầu vào. Các bản tóm tắt mang tính ngụ ý chỉ rõ văn bản đầu vào nói về cái gì. Các bản tóm tắt kết tụ cho ta thông tin bổ sung không có trong văn bản đầu vào. Các bản tóm tắt mang tính bình phẩm xem xét lại tính đúng và sai của văn bản đầu vào.

1.1.2 Phân loại các hệ thống tóm tắt văn bản

Như đã trình bày ở phần trên, các tham số khác nhau đều ảnh hưởng đến kết quả tóm tắt văn bản. Do vậy chúng ta có thể phân loại các hệ thống tóm tắt văn bản theo các hướng sau:

Theo kết quả (output):

- Tóm tắt trích rút (Extract): là một bản tóm tắt bao gồm các đơn vị văn bản quan trọng như câu, đoạn... được trích rút từ văn bản gốc [32].

- Tóm tắt tóm lược (Abstract): tương tự như cách con người thực hiện tóm tắt, nghĩa là đầu tiên phải hiểu các khái niệm chính của một văn bản, sau đó tạo

ra bản tóm tắt có chứa các nội dung không được thể hiện trong văn bản [23].

Theo mục đích hay chức năng tóm tắt (Function):

- Tóm tắt chỉ thị (Indicative): tóm tắt nhằm cung cấp một chức năng tham khảo để chọn tài liệu đọc chi tiết hơn (ứng dụng trong tóm tắt kết quả tìm kiếm).

Ví dụ: Trong tóm tắt tin tức, tóm tắt đưa ra chi tiết chính của từng sự kiện.

- Tóm tắt thông tin (Information): tóm tắt bao gồm tất cả các thông tin nổi bật của văn bản gốc ở nhiều mức độ chi tiết khác nhau.

- Tóm tắt đánh giá (Evaluation): tóm tắt nhằm mục đích đánh giá vấn đề chính của văn bản gốc theo quan điểm của người đánh giá.

Theo nội dung:

- Tóm tắt chung (Generalized): tóm tắt nhằm mục đích đưa ra các nội dung quan trọng phản ánh toàn bộ nội dung văn bản gốc.

- Tóm tắt hướng truy vấn (Query-based): tóm tắt nhằm mục đích đưa ra kết quả dựa vào câu truy vấn của người. Tóm tắt này thường được sử dụng trong quá trình tìm kiếm thông tin (information retreival).

Theo miền dữ liệu:

- Tóm tắt trên một miền dữ liệu (Domain): tóm tắt nhắm vào một miền nội dung nào đó, như tin tức khủng bố, tin tức tài chính…

- Tóm tắt trên một thể loại (Genre): tóm tắt nhắm vào một thể loại văn bản nào đó, như báo chí, email, web, bài báo…

- Tóm tắt độc lập (Independent): tóm tắt cho nhiều thể loại và nhiều miền dữ liệu.

Theo mức độ chi tiết:

- Tóm tắt tổng quan (overview): tóm tắt miêu tả tổng quan tất cả các nội dung nổi bật trong văn bản nguồn.

- Tóm tắt tập trung sự kiện (event): tóm tắt miêu tả một sự kiện cụ thể nào đó trong văn bản nguồn.

Theo số lượng:

- Tóm tắt đơn văn bản: Nếu một bản tóm tắt được tạo thành từ một văn bản riêng lẻ thì hệ thống tóm tắt đó được gọi là hệ thống tóm tắt đơn văn bản.

- Tóm tắt đa văn bản: Nếu một bản tóm tắt được tạo thành từ nhiều văn bản liên quan tới một chủ đề riêng lẻ thì hệ thống tóm tắt đó gọi là hệ thống tóm tắt đa văn bản.

Theo ngôn ngữ:

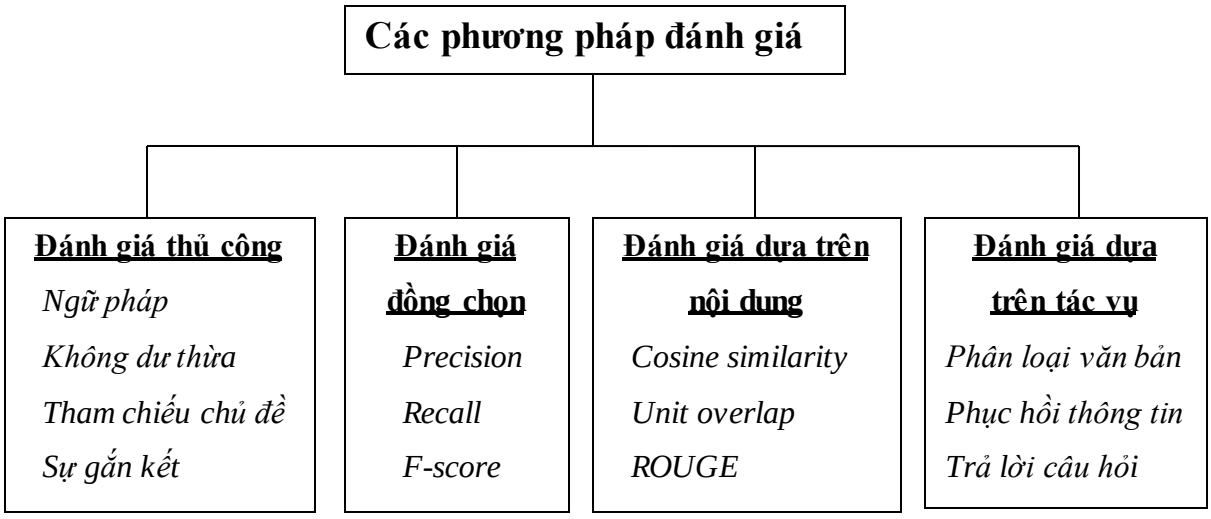
- Tóm tắt đơn ngôn ngữ: Văn bản nguồn chỉ có một loại ngôn ngữ. Kết quả ra là văn bản ngôn ngữ đó.

- Tóm tắt đa ngôn ngữ: Mỗi văn bản nguồn chỉ có một loại ngôn ngữ. Nhưng ứng dụng có khả năng tóm tắt trên nhiều loại ngôn ngữ. Tùy vào văn bản nguồn hoặc tham số mà hệ thống tóm tắt trên một ngôn ngữ được chọn.

- Tóm tắt xuyên ngôn ngữ (cross-language): Trong văn bản nguồn chứa hai hay nhiều ngôn ngữ khác nhau, hệ thống có thể tùy vào từng đơn vị ngữ liệu mà nhận dạng và tóm tắt cho phù hợp. Đây là loại tóm tắt phức tạp nhất trong ba loại phân chia theo số lượng ngôn ngữ.

1.2 Các phương pháp đánh giá tóm tắt văn bản

Các phương pháp đánh giá được phân thành 4 loại [65],[73], được trình bày như trong hình 1-4.



Hình 1-4 Phân loại các phương pháp đánh giá tóm tắt văn bản.

1.2.1 Đánh giá thủ công

Nhà ngôn ngữ học trực tiếp đánh giá bản tóm tắt dựa vào chất lượng bản văn, nghĩa là sử dụng các tham số ngữ pháp, không dư thừa, phân lớp tham chiếu và sự gắn kết để cho điểm bản tóm tắt do hệ thống tạo ra. Cách đánh giá là xem xét lỗi ngữ pháp trong bản văn như sai từ, lỗi dấu câu. Bản tóm tắt khi hệ thống tạo ra không được chứa thông tin dư thừa và các tham chiếu trong bản tóm tắt phải được liên kết rõ ràng với chủ đề của văn bản gốc. Độ gắn kết của văn bản cũng là một tiêu chí quan trọng để đánh giá bản tóm tắt hệ thống. Tuy nhiên, phương pháp này có một số hạn chế như việc chấm điểm do con người thực hiện không ổn định và là phương pháp đánh giá tiêu tốn thời gian và tiền bạc [23].

1.2.2 Đánh giá đồng chọn

Phương pháp đánh giá dựa trên đồng chọn chỉ có thể sử dụng với các bản tóm tắt theo hướng trích rút câu. Các câu được trích chọn kết nối với nhau, tạo nên văn bản tóm tắt, không cần hiệu chỉnh thêm. Phương pháp này đánh giá giữa bản tóm tắt do hệ thống trích rút với bản tóm tắt do con người trích rút sử dụng độ đo chính xác (precision), triệu hồi (recall), các giá trị f- measure.

Độ đo chính xác (precision) [15]**:** là tỉ số giữa số lượng các câu được cả hệ thống và con người trích rút trên số các câu được hệ thống trích rút.

$$Precision = \frac{|SH \cap SM|}{|SM|} \quad (1.1)$$

trong đó: $|SM|$ là số lượng câu của bản tóm tắt do hệ thống trích rút;

$|SH|$ là số lượng câu của bản tóm tắt do con người trích rút;

$|SH \cap SM|$ là số lượng những câu được cả hệ thống và con người trích rút.

Độ đo triệu hồi (recall)[15]**:** là tỉ số giữa số lượng các câu được trích rút bởi hệ thống trùng với số các câu mà con người trích rút trên số các câu chỉ được lựa chọn bởi con người.

$$Recall = \frac{|SH \cap SM|}{|SH|} \quad (1.2)$$

trong đó: $|SM|$ là số lượng câu của bản tóm tắt do hệ thống trích rút;

$|SH|$ là số lượng câu của bản tóm tắt do con người trích rút;

$|SH \cap SM|$ là số lượng những câu được cả hệ thống và con người trích rút.

Độ đo f-score: là một độ đo kết hợp hai đại lượng precision và recall. Theo truyền thống thì f-score được định nghĩa là trung bình hàm điều hòa của recision và recall. Các giá trị f-score nhận giá trị trong đoạn [0, 1], trong đó giá trị tốt nhất là 1.

$$f - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (1.3)$$

Trong tóm tắt văn bản, người ta cũng thường dùng các trọng số khác nhau cho precision và recall trong khi tính f-score. Giá trị trọng số β là một số thực không âm. Trọng số lớn hơn 1 nghĩa là precision quan trọng hơn, còn trọng số nhỏ hơn 1 nghĩa là recall quan trọng hơn.

$$F - score = \frac{(\beta^2 + 1) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \quad (1.4)$$

Các giá trị precision và recall có thể không phù hợp trong một số trường hợp của tóm tắt văn bản. Ví dụ, từ một văn bản có 5 câu (1, 2, 3, 4, 5), ta tạo ra hai bản tóm tắt khác nhau. Bản tóm tắt thứ nhất chứa các câu (1, 2, 5) và bản kia chứa các câu (1, 4, 5). Bản tóm tắt lý tưởng chứa các câu (1, 2, 5). Khi đánh giá bằng precision và recall, ta có thể quyết định bản tóm tắt đầu tiên tốt hơn bản thứ hai. Nhưng quá trình tóm tắt cũng có tính chủ quan, nên có thể bản tóm tắt thứ hai tốt như bản tóm tắt đầu.

Độ đo Relative utility được giới thiệu bởi Radev, Jing và Budzikowska vào năm 2000 [64] để khắc phục vấn đề của phương pháp đánh giá dựa trên precision và recall đã nêu ở trên. Với phương pháp này, bản tóm tắt lý tưởng được biểu diễn với các câu gốc và các giá trị Relative utility của chúng. Các giá

trị Relative utility do con người phán đoán và được dùng để cung cấp thông tin về tầm quan trọng của một câu nào đó trong văn bản đã cho. Ví dụ, một bản tóm tắt lý tưởng cho một văn bản gồm 5 câu được cho trước là (1/5, 2/3, 3/2, 4/3, 5/4). Các giá trị Relative utility bao gồm: câu đầu tiên là quan trọng nhất, câu thứ 3 ít quan trọng nhất, và tầm quan trọng của câu thứ 2 và thứ 4 là như nhau. Do vậy khi hai bản tóm tắt khác nhau cùng chọn (1, 2, 5) và (1, 4, 5) thì thật ra sẽ có chỉ số đánh giá bằng nhau. Cũng như vậy cả hai đều có các chỉ số cao nhất có thể nhận được, thì nghĩa là hai bản tóm tắt đều là tối ưu.

1.2.3 Đánh giá dựa trên nội dung

Trong phương pháp đánh giá dựa trên nội dung, bản tóm tắt của hệ thống được so sánh với bản tóm tắt lý tưởng bằng cách sử dụng đơn vị so sánh là từ vựng. Nếu dùng phương pháp này, ta có thể so sánh các bản tóm tắt được trích rút với các bản tóm tắt lý tưởng ngay cả khi chúng không trùng nhau câu nào. Với các cách đánh giá dựa trên nội dung, ta sử dụng các độ đo như tính tương tự cosine, chuỗi con chung dài nhất LCS và các chỉ số ROUGE. Phương pháp dựa trên nội dung được đánh giá là tốt hơn phương pháp dựa trên đồng chọn vì nó có thể đánh giá 2 câu khác nhau nhưng có cùng nội dung thông tin.

Độ tương tự cosine [45]**:** Trong xử lý ngôn ngữ tự nhiên, công thức tính toán cosine được sử dụng để đo mức độ tương tự giữa hai câu hoặc hai văn bản. Công thứ tính độ tương tự cosine được mô tả như sau:

$$Cosine(A,B) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2} \times \sqrt{\sum_{i=1}^{n} {B_i}^2}} \quad (1.5)$$

trong đó:

$A = \{w_1^A, \dots, w_n^A\}$ là vector thuộc tính của bản tóm tắt hệ thống với w_i^A là trọng số của từ thứ i trong bản tóm tắt hệ thống;

$B = \{w_1^B, \dots, w_n^B\}$ là vector thuộc tính của bản tóm tắt lý tưởng với w_i^B là trọng số của từ thứ i trong bản tóm tắt lý tưởng.

Phương pháp đánh giá dựa trên LCS [65]**:** LCS tìm ra độ dài của chuỗi con chung dài nhất giữa văn bản *X* và *Y*, độ dài của chuỗi con chung dài nhất càng lớn thì 2 văn bản *X, Y* càng giống nhau.

$$lcs(X,Y) = \frac{length(X) + length(Y) - edit_{di}(X,Y)}{2} \quad (1.6)$$

trong đó: $length(X)$ là độ dài của chuỗi *X*; $length(Y)$ là độ dài của chuỗi *Y*; $edit_{di}(X,Y)$ là khoảng cách biên tập giữa *X* và *Y* (là số lượng tối thiểu của việc xóa và chèn thêm cần thiết để biến đổi *X* thành *Y*).

Phương pháp đánh giá BLEU [38]**:** Ý tưởng chính của BLEU là đánh giá độ tương tự giữa một bản tóm tắt hệ thống và tập các bản tóm tắt lý tưởng dựa vào trung bình có trọng số của các n-gram (một n-gram là một dãy gồm *n* ký tự (hoặc âm tiết, từ) liên tiếp nhau trong văn bản) trong bản tóm tắt hệ thống và trong tập các bản tóm tắt lý tưởng. Độ đo được tính theo công thức (1.7):

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count\ (n\text{-}gram)} \quad (1.7)$$

trong đó: $Count_{clip}(n\text{-}gram)$ là số n-gram xuất hiện lớn nhất trong bản tóm tắt hệ thống và bản tóm tắt lý tưởng; $Count(n\text{-}gram)$ là số n-gram trong bản tóm tắt hệ thống.

Phương pháp đánh giá ROUGE:

Các phương pháp đánh giá tóm tắt truyền thống thường gắn với đánh giá thủ công do chuyên gia con người thực hiện thông qua một số độ đo khác nhau, chẳng hạn: mức độ súc tích, mức độ liền mạch, ngữ pháp, mức độ dễ đọc và nội dung. Tuy nhiên, phương pháp đánh giá kết quả tóm tắt thủ công mất quá nhiều công sức và chi phí. Vì thế, đánh giá tóm tắt tự động là một yêu cầu cấp thiết. Lin và Hovy đề xuất một phương pháp đánh giá mới gọi là ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[44],[73]. Hiện nay phương pháp đo này được sử dụng như một phương pháp chuẩn đánh giá kết

quả tóm tắt tự động cho văn bản tiếng Anh. Một cách hình thức, ROUGE-N là một độ đo đối với các n-gram trong văn bản tóm tắt hệ thống và trong tập các văn bản tóm tắt lý tưởng, được tính theo công thức (1.8):

$$ROUGE - N = \frac{\sum_{S \in RSS} \sum_{n-gram \in S} Count_{match}(n - gram)}{\sum_{S \in RSS} \sum_{n-gram \in S} Count\ (n - gram)} \quad (1.8)$$

trong đó: S là bản tóm tắt hệ thống; RSS là tập văn bản tóm tắt lý tưởng; $Count_{match}(n\text{-}gram)$ là số lượng n-gram đồng xuất hiện lớn nhất giữa văn bản tóm tắt hệ thống và tập văn bản tóm tắt lý tưởng; $Count(n\text{-}gram)$ là số lượng n-gram trong văn bản tóm tắt lý tưởng.

Đối với bài toán tóm tắt đơn văn bản tiếng Việt, luận án sử dụng độ đo ROUGE-N dựa trên số n-gram từ vựng để đánh giá (mô tả chi tiết ở phần thử nghiệm của từng chương).

1.2.4 Đánh giá dựa trên tác vụ

Phương pháp cuối cùng là đánh giá dựa trên tác vụ. Trong phương pháp đánh giá này, các bản tóm tắt được tạo ra với mục đích là so sánh dựa trên hiệu năng của tác vụ đã cho của chúng. Đánh giá dựa trên tác vụ có thể dùng các phương pháp khác nhau để đánh giá hiệu năng của hệ thống tóm tắt. Một số phương pháp trong các phương pháp này là phục hồi thông tin, trả lời câu hỏi và các phương pháp phân cụm văn bản.

Hiệu năng của hệ thống tóm tắt có thể được đo bằng cách sử dụng các phương pháp phục hồi thông tin. Ta so sánh hiệu năng của phương pháp phục hồi thông tin sử dụng toàn bộ văn bản và hiệu năng của phương pháp dùng bản tóm tắt được trích rút. Nếu hiệu năng của phương pháp phục hồi thông tin không thay đổi nhiều, ta kết luận hệ thống tóm tắt đã thành công [65].

Tương tự với phương pháp phục hồi thông tin, các phương pháp trả lời câu hỏi có thể sử dụng cho đánh giá tóm tắt. Ở đây, nếu chỉ đọc bản văn đầu vào hay chỉ đọc bản tóm tắt, óc phán đoán của con người sẽ trả lời một số câu

hỏi lựa chọn. Các kết quả đúng được sử dụng để đánh giá hệ thống tóm tắt [52].

Phân loại văn bản cũng được sử dụng để đánh giá tóm tắt. Với mục đích này, ta sử dụng các kho ngữ liệu văn bản đã được gán nhãn. Phân loại do con người làm hoặc phân loại tự động được thực hiện bằng cách sử dụng văn bản gốc, các bản tóm tắt trích rút và các bản tóm tắt được tạo ngẫu nhiên. Trong khi các kết quả có các văn bản gốc đặt được cận trên, thì các bản tóm tắt tạo bởi cách chọn các câu ngẫu nhiên đặt cận dưới. Sử dụng các giá trị precision và recall, các bản tóm tắt trích rút có thể so sánh với các kết quả của phương pháp sử dụng các văn bản gốc hoặc các bản tóm tắt được tạo ngẫu nhiên.

1.3 Các hướng tiếp cận tóm tắt văn bản ngoài nước

1.3.1 Các phương pháp tóm tắt trích rút

Các phương pháp tóm tắt trích rút cố gắng tìm ra các đơn vị quan trọng nhất của một văn bản đầu vào và chọn các câu có liên quan tới các đơn vị quan trọng này để tạo ra bản tóm tắt.

a. Các phương pháp tiên phong

Nghiên cứu đầu tiên về tóm tắt văn bản vào những năm 50 của thế kỷ 20 là của Luhn [47] được dựa trên tần suất các từ trong văn bản với quan điểm từ xuất hiện thường xuyên là từ quan trọng nhất. Câu chứa nhiều từ thường xuyên quan trọng hơn các câu khác và được chọn trong bản tóm tắt.

Sau nghiên cứu của Luhn, các nhà nghiên cứu đề xuất rất nhiều phương pháp khác dựa trên các đặc trưng đơn giản khác như các từ khóa/cụm từ khóa [75],[29]; vị trí câu [17],[29],[19].

b. Các phương pháp thống kê

Các phương pháp tóm tắt nổi tiếng nhất dùng thống kê là dựa trên khái niệm tương quan và phân loại Bayes.

Dự án SUMMARIST [34] là một dự án tóm tắt văn bản nổi tiếng dùng phương pháp thống kê. Trong dự án này thông tin về khái niệm tương quan trích rút từ các từ điển và WordNet được dùng cùng với các phương pháp xử lý

ngôn ngữ tự nhiên. Trong phương pháp này, một từ được cho là có xuất hiện khi các từ khác có liên quan cũng xuất hiện. Ví dụ số các lần xuất hiện của từ "automobile" được tăng lên nếu ta đã thấy từ "car".

Một ứng dụng tóm tắt khác dựa trên thống kê là của Kupiec [39], trong đó phân loại Bayes được dùng để trích rút câu. Trong phương pháp này tác giả dùng một kho ngữ liệu các bản văn và các bản tóm tắt để huấn luyện hệ thống. Các đặc trưng được sử dụng trong hệ thống này là tần suất xuất hiện các từ, các từ viết hoa, độ dài câu, vị trí trong các đoạn và cấu trúc cụm từ.

c. Các phương pháp dựa trên kết nối bản văn

Phương pháp này liên quan tới các bài toán tham chiếu tới các phần đã được đề cập của một văn bản. Các phương pháp sử dụng chuỗi từ vựng và Lý thuyết cấu trúc tu từ RST (Rhetorical Structure Theory).

Phương pháp chuỗi từ vựng là một thuật toán nổi tiếng sử dụng kết nối bản văn. Trong phương pháp này, mối tương quan ngữ nghĩa của các từ (tính đồng nghĩa, tính trái nghĩa,…) được thực hiện bằng cách sử dụng các từ điển và WordNet. Các chuỗi từ vựng có mối tương quan ngữ nghĩa được xây dựng được sử dụng để trích rút các câu quan trọng trong một văn bản [18],[30].

Các phương pháp dựa trên RST để tổ chức các đơn vị bản văn thành cấu trúc dạng cây. Sau đó cấu trúc này được sử dụng để thực hiện tóm tắt [59],[50].

d. Các phương pháp dựa trên đồ thị

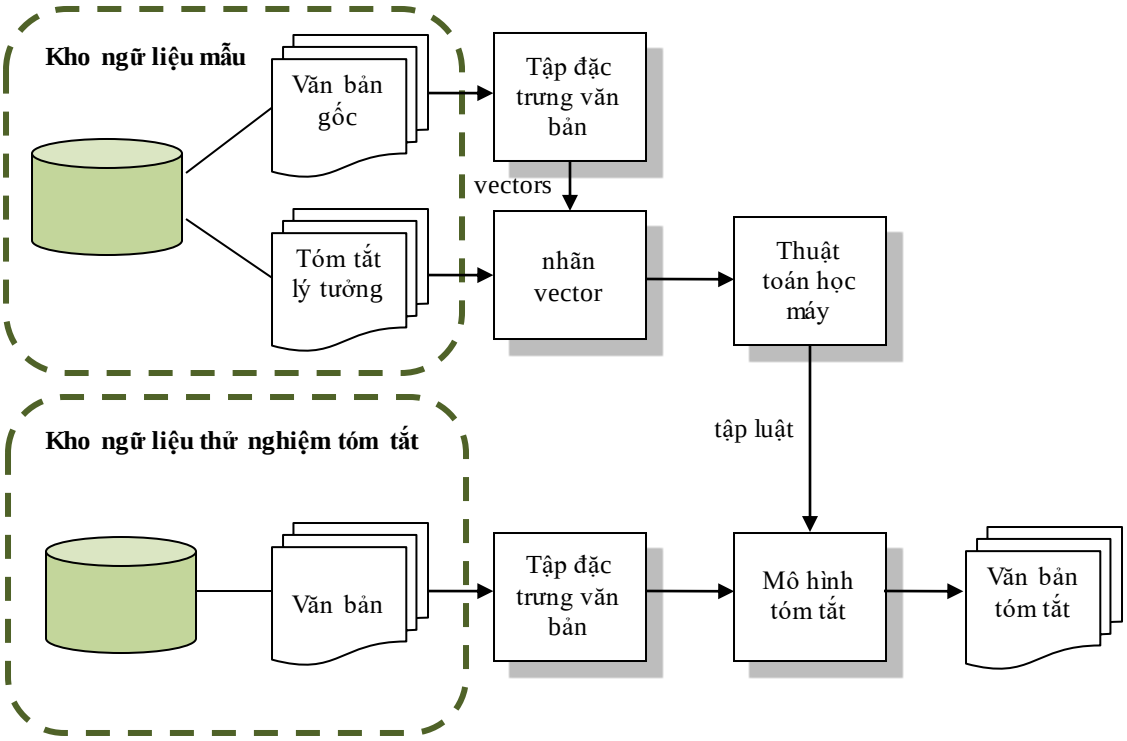
Phương pháp đồ thị được xây dựng dựa trên các thuật toán HITS [40] và Google's PageRank [20]. Các thuật toán này sau đó được dùng trong tóm tắt văn bản [36].

Trong bài toán tóm tắt văn bản dựa vào đồ thị, các đỉnh biểu diễn các câu, còn các cạnh biểu diễn độ tương tự giữa các câu. Các giá trị đo độ tương tự được tính toán bằng cách sử dụng độ tương tự giữa các từ hoặc các cụm từ. Các câu có độ tương tự cao nhất với các câu khác được chọn ra cho bản tóm tắt đầu ra theo tỷ lệ tóm tắt. Điển hình cho hướng tiếp cận tóm tắt văn bản dựa trên đồ

thị là hai phương pháp TextRank [54] và Cluster LexRank [62].

e. Các phương pháp dựa vào học máy

Các phương pháp dựa vào học máy cũng được sử dụng cho tóm tắt văn bản với sự hỗ trợ của các tiến bộ trong học máy và xử lý ngôn ngữ tự nhiên. Các phương pháp đầu tiên sử dụng giả thiết các đặc trưng độc lập với nhau. Các phương pháp phát triển sau đó lại sử dụng giả thiết các đặc trưng phụ thuộc lẫn nhau.



Hình 1-5 Framework chung cho hệ thống TTVB bằng phương pháp học máy.

Các thuật toán tóm tắt dựa trên học máy sử dụng các kỹ thuật như Naïve-Bayes [39],[21], mô hình Markov ẩn HMM [22], các mô hình logarit tuyến tính (Log-linear Models) [60], mạng nơ-ron [71] và giải thuật phỏng sinh học như [25],[31],[42],[51],[67],[72].

f. Các phương pháp đại số

Trong những năm gần đây, các phương pháp đại số như phân tích ngữ nghĩa tiềm ẩn LSA (Latent Semantic Analysis) [43], phép nhân tử hóa ma trận

không âm NMF (Non-negative Matrix Factorization) [46] và khai triển ma trận nửa rời rạc SDD (Semi-discrete Matrix Decomposition) được sử dụng cho tóm tắt văn bản. Trong đó, thuật toán LSA nổi tiếng nhất, thuật toán này dựa trên phương pháp phân tích giá trị đơn SVD (Singular Value Decomposition) [16]. Trong thuật toán LSA, độ tương tự giữa các câu và độ tương tự giữa các từ đều được trích rút. Không những ứng dụng trong tóm tắt văn bản, thuật toán LSA còn được dùng cho phân cụm văn bản và lọc thông tin.

1.3.2 Các phương pháp tóm tắt theo hướng tóm lược

Các phương pháp tóm tắt tóm lược cố gắng để hiểu đầy đủ các văn bản cần tóm tắt, ngay cả các văn bản có chủ đề không rõ ràng. Sau đó, tạo ra các câu mới cho bản tóm tắt theo tỉ lệ của người dùng yêu cầu. Phương pháp này rất giống với cách tóm tắt của con người. Nhưng về mặt thực tế, để đạt được biểu diễn của con người rất khó. Do đó, các nghiên cứu đã dựa vào các đơn vị đặc trưng như từ, cụm từ, thành phần câu quan trọng để sinh ra các câu mới cho tóm tắt văn bản.

Theo hướng này có: phương pháp dựa vào các từ hay cụm từ quan trọng để tạo ra các câu cho bản tóm tắt [24],[66]; phương pháp dựa trên kỹ thuật cô đọng văn bản [78]; phương pháp dựa trên kỹ thuật rút gọn văn bản, nối hai hay nhiều câu thành một câu [63]; phương pháp dựa trên kỹ thuật rút gọn câu để tạo ra bản tóm tắt [41].

1.4 Kho ngữ liệu tiêu chuẩn cho bài toán tóm tắt văn bản tiếng Anh

Vấn đề của lĩnh vực tóm tắt văn bản tự động là làm sao để đánh giá chính xác tính chính xác và khách quan các phương pháp tóm tắt văn bản được đề xuất. Để đánh giá chính xác đòi hỏi phải có một kho ngữ liệu tóm tắt tiêu chuẩn phù hợp. Đối với tiếng Anh, người ta đã xây dựng được một số kho ngữ liệu tóm tắt tiêu chuẩn lớn như BBC, CNN, TREC, CAST, DUC [74]. Trong các kho ngữ liệu đó, DUC được đánh giá là kho ngữ liệu lớn, luôn được cập nhật và đã được sử dụng rộng rãi.

Từ năm 2001, Viện tiêu chuẩn và công nghệ NIST đã giới thiệu 7 bộ dữ liệu liên quan đến tổng kết văn bản tự động (DUC2001-DUC2007). Các bộ số liệu này được giới thiệu với mục đích đánh giá các phương pháp tóm tắt văn bản tự động. Mỗi bộ số liệu giới thiệu được phục vụ cho một mục đích cụ thể khác nhau. DUC2001 đến DUC2004 phục vụ cho đánh giá bài toán tóm tắt đơn văn bản. DUC2005 đến DUC2007 phục vụ cho đánh giá bài toán tóm tắt đa văn bản.

DUC2007 chứa 45 chủ đề, mỗi chủ đề 25 văn bản. Mỗi văn bản được 10 thành viên của NIST tóm tắt tóm lược bằng tay và kết quả tóm tắt sẽ được lựa chọn ngẫu nhiên. Hiện nay đã có 32 hệ thống tóm tắt tham gia tóm tắt văn bản tự động cho mỗi chủ đề và sử dụng độ đo ROUGE (phép đo giữa bản tóm tắt của hệ thống với bản tóm tắt con người) để đánh giá, xếp hạng hiệu quả từng phương pháp.

1.5 Hiện trạng nghiên cứu tóm tắt văn bản tiếng Việt

1.5.1 Đặc điểm tiếng Việt

Tiếng Việt là ngôn ngữ không biến hình từ và âm tiết tính, tức là mỗi một tiếng (âm tiết) được phát âm tách rời nhau và được thể hiện bằng một chữ viết [1]. Hai đặc trưng này chi phối toàn bộ tổ chức bên trong của hệ thống ngôn ngữ Việt, do vậy trong lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt nói chung và bài toán tóm tắt tiếng Việt nói riêng chúng ta cần chú ý tới khi xử lý trên máy tính. Tiếng Việt có những đặc điểm cơ bản như sau:

Đặc điểm cấu tạo:

Đơn vị cơ sở để cấu tạo từ tiếng Việt là các tiếng hay theo ngữ âm học là các âm tiết. Từ âm tiết, người ta tạo ra các đơn vị từ vựng khác như ***từ, cụm từ, câu*** để định danh sự vật, hiện tượng,… chủ yếu nhờ phương thức ghép và phương thức láy [1]. Theo thống kê, trong tiếng Việt có khoảng hơn 6700 âm tiết [4] và trong vốn từ tiếng Việt 80% là các từ gồm 2 âm tiết trở lên.

Ví dụ: Từ *"tin"* là một từ gồm một âm tiết.

Từ *"thông tin"* là một từ gồm hai âm tiết.

Cụm từ *"công nghệ thông tin"* gồm 2 từ hay 4 âm tiết.

Do đặc điểm như vậy, khoảng trắng (space) không được sử dụng để phân biệt ranh giới từ như các ngôn ngữ khác (Anh, Pháp, Nga,…). Vì vậy, đối với tiếng Việt việc xác định ranh giới từ là một thách thức, đặc biệt là xử lý nhập nhằng và từ mới.

Ví dụ: *Hôm nay, chúng tôi đón tiếp tân giám đốc*

nhập nhằng tách từ có thể xảy ra ở 'đón tiếp' và 'tiếp tân'. Đây là một trong những nhập nhằng thường gặp trong bài toán tách từ tiếng Việt.

Ví dụ: *Ông già đi nhanh quá*

nhập nhằng về mặt danh từ 'ông già' hay động từ 'già', như vậy cần phải xét mặt ngữ cảnh trong văn bản để tách từ cho đúng.

Phân loại từ:

Theo quan điểm truyền thống, từ tiếng Việt được chia ra làm hai loại thực từ và hư từ. Trong đó, thực từ có ý nghĩa chân thực, còn hư từ thì không có ý nghĩa từ vựng chân thật mà chỉ làm công cụ ngữ pháp để biểu hiện các quan hệ ngữ pháp khác nhau. Tuy nhiên, trong nhiều trường hợp nhiều hư từ vốn bắt nguồn từ thực từ và cùng tồn tại song hành với thực từ ấy [1]. Điều này gây khó khăn trong việc nhận diện hư từ. Xem hai câu ví dụ sau:

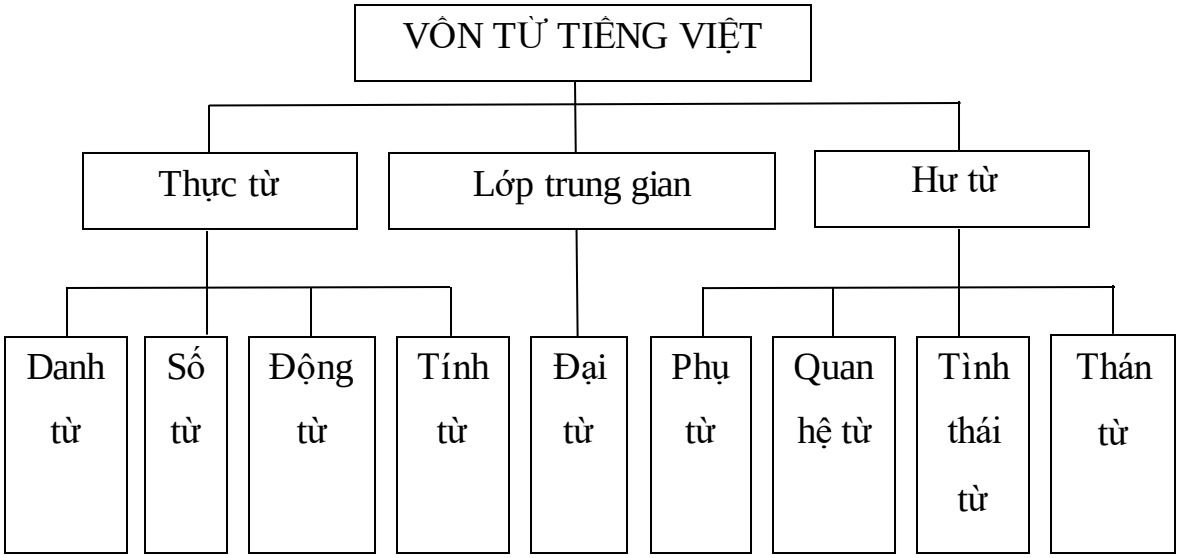
Lấy cho tôi cuốn sách ấy

và

Anh cho nó cuốn sách.

Từ *"cho"* trong câu thứ nhất là hư từ, trong câu thứ 2 là thực từ.

Trong bài toán tóm tắt văn bản tiếng Việt, việc nhận biết thực từ và hư từ là bước rất quan trọng bởi vì các phương pháp tóm tắt đều chỉ thực hiện tính toán dựa trên thực từ còn các hư từ bị loại bỏ.



Hình 1-6. Sơ đồ từ loại tiếng Việt

Từ đồng nghĩa:

"Những từ đồng nghĩa là những từ có nghĩa giống nhau. Đó là nhiều từ khác nhau cùng chỉ một sự vật, một đặc tính, một hành động nào đó. Đó là những tên khác nhau của một hiện tượng" [11].

Ví dụ: *dễ, dễ dàng, dễ dãi* là những nhóm từ đồng nghĩa.

Với bài toán tóm tắt văn bản thì từ đồng nghĩa cũng có một ý nghĩa khá quan trọng bởi trong các câu, đoạn văn trong văn bản có các từ đồng nghĩa hoặc gần nghĩa nhau và việc sử dụng từ đồng nghĩa sẽ làm nâng cao tính chính xác khi so sánh về độ tương đồng ngữ nghĩa giữa các đơn vị văn bản.

Đặc điểm chính tả:

Trong tiếng Việt, một số đặc điểm chính tả chính cần lưu ý như sau [8]:

- Các tiếng đồng âm: như kĩ/kỹ, lí, lý… thường bị sử dụng lẫn nhau như: lý luận, lí luận, kĩ thuật, kỹ thuật…

- Vị trí dấu thanh: theo quy định đánh dấu tiếng Việt, dấu được đặt trên nguyên âm có ưu tiên cao nhất. Tuy nhiên, khi viết văn bản nhiều bộ gõ văn bản không tuân thủ theo đúng nguyên tắc trên nên xảy ra hiện tượng dấu được đặt ở các vị trí khác nhau, chẳng hạn: toán, tóan, thuý, thúy…

- Phiên âm tiếng nước ngoài: hiện nay, vẫn còn nhiều tranh cãi giữa việc phiên âm tiếng nước ngoài thành tiếng Việt (Việt hoá), nên tồn tại nhiều cách viết (giữ nguyên gốc tiếng nước ngoài, phiên âm ra tiếng Việt), ví dụ: Singapore/Xin−ga−po.

- Từ gạch nối: do cách viết dấu gạch nối tuỳ tiện, không phân biệt được giữa nối tên riêng hay chú thích.

- Kí tự ngắt câu: các kí tự đặc biệt như "", ";", "!", "?", "…" ngăn cách giữa các câu hoặc các vế câu trong câu ghép.

Bảng mã tiếng Việt trên máy tính:

Hiện nay có nhiều cách mã hoá các kí tự tiếng Việt khác nhau, dẫn tới có nhiều bảng mã khác nhau được sử dụng. Theo thống kê, có tới trên 40 bảng mã tiếng Việt khác nhau được sử dụng như loại mã 1 byte TCVN, VNI… và loại mã 2byte Unicode. Do đó, việc khai thác tài liệu cũng như xử lý dữ liệu rất phức tạp. Do vậy, trong các bài toán xử lý ngôn ngữ tiếng Việt, các văn bản cần phải thống nhất về một bảng mã chuẩn Unicode.

1.5.2 Hiện trạng nghiên cứu xử lý ngôn ngữ tự nhiên tiếng Việt

Hiện nay, lĩnh vực xử lý ngôn ngữ tiếng Việt đã nhận được nhiều sự quan tâm của các nhà nghiên cứu. Tuy nhiên, các nghiên cứu chủ yếu đang tập trung vào những vấn đề cơ bản của tiếng Việt như: Xây dựng kho ngữ liệu và công cụ tách từ tiếng Việt, xây dựng kho ngữ liệu và công cụ gán nhãn tiếng Việt,… Bắt đầu từ năm 2006, nhánh đề tài "Xử lí văn bản" là một phần của đề tài KC01.01/06-10 *"Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt"* giai đoạn 1 đã được triển khai [79]. Cho đến nay, nhánh đề tài này đã thu được một số kết quả bao gồm kho ngữ liệu: từ điển, kho ngữ liệu tách từ, kho ngữ liệu gán nhãn, song ngữ Anh – Việt; và các bộ công cụ phục vụ cho xử lý văn bản: công cụ tách từ, gán nhãn từ loại, phân tích cú pháp…

Trong giai đoạn 2, đề tài "*Nghiên cứu, xây dựng và phát triển một số tài*

nguyên và công cụ thiết yếu cho xử lí văn bản tiếng Việt" mã số "*KC.01.20/11-15*" đã được triển khai và tập trung xây dựng Wordnet tiếng Việt. Tuy nhiên, đến hiện nay các công bố về Wordnet tiếng Việt mới chỉ ở mức thử nghiệm.

Ngoài ra, còn có các nghiên cứu của các tác giả khác về tách từ, gán nhãn từ loại, trích rút thông tin, tóm tắt văn bản tiếng Việt đã được công bố và thử nghiệm trên kho ngữ liệu do cá nhân xây dựng. Tuy nhiên, rất ít các công cụ được công bố cho cộng đồng thử nghiệm, đánh giá.

1.5.3 Một số hướng tiếp cận tóm tắt văn bản tiếng Việt

Do tính phức tạp và đặc thù riêng của tiếng Việt, số lượng những nghiên cứu về tóm tắt văn bản tiếng Việt so với tiếng Anh vẫn còn ít. Phần lớn các nghiên cứu đó mới chỉ là các nghiên cứu ở mức đề tài tốt nghiệp đại học, luận văn thạc sĩ và tiến sĩ, đề tài nghiên cứu. Tuy nhiên, các phương pháp hầu hết chỉ dừng ở mức thử nghiệm mà chưa xây dựng một ứng dụng hoàn chỉnh để công bố cho cộng đồng thử nghiệm. Mặt khác, do chưa có kho ngữ liệu chuẩn phục vụ cho tóm tắt nên hầu hết thử nghiệm của các nghiên cứu đều thực hiện trên các kho ngữ liệu tự xây dựng. Do vậy, việc đánh giá từng phương pháp cần phải xem xét một cách kỹ lưỡng.

Hiện nay, hầu hết các nghiên cứu tóm tắt văn bản tiếng Việt đã được công bố thực hiện theo hướng trích rút, chỉ có một vài nghiên cứu thực hiện theo hướng tóm tắt tóm lược. Có thể liệt kê một số công trình tiêu biểu theo các hướng cụ thể sau:

Hướng tóm tắt trích rút:

Nghiên cứu của Lê Hà Thanh, Huỳnh Quyết Thắng, Lương Chi Mai (2005) [76]: dựa vào sự kết hợp tuyến tính của 5 đặc trưng: Từ tiêu đề, vị trí câu trong đoạn, danh từ, độ tương đồng giữa hai đoạn, TFxIPF (Term Frequency times InverParagraph Frequency) để tính trọng số câu. Nghiên cứu này đã đề cập đến hệ số đặc trưng và cách tìm qua quá trình thực nghiệm.

Nghiên cứu của Đỗ Phúc, Hoàng Kiếm (2006) [2]: trích rút các ý chính từ văn bản hỗ trợ tạo tóm tắt văn bản tiếng Việt dựa trên việc sử dụng cây hậu tố để phát hiện các dãy từ phổ biến trong các câu của văn bản, dùng từ điển để tìm các dãy từ có nghĩa, dùng WordNet tiếng Việt hoặc từ điển để giải quyết vấn đề ngữ nghĩa của các từ. Cuối cùng dùng kỹ thuật gom cụm để gom các các câu trong văn bản (vector đặc trưng cho câu) và hình thành các vector đặc trưng cụm, sau đó rút ra các câu chứa nhiều thành phần của các vector đặc trưng cụm.

Nghiên cứu của Nguyễn Lê Minh, Akira Shimazu, Xuân Hiếu Phan, Hồ Tú Bảo và Susumu Horiguchi [55]: sử dụng phương pháp học máy SVM (Support Vector Machine) dựa trên tập đặc trưng vị trí câu (câu đầu và cuối trong văn bản là quan trọng), chiều dài câu (ưu tiên câu ngắn), từ liên quan tiêu đề, cụm từ gợi ý, từ xuất hiện nhiều để chọn ra câu quan trọng.

Nghiên cứu của Nguyễn Hoàng Tú Anh [7]: biểu diễn văn bản bằng đồ thị với mỗi đỉnh là một câu, trọng số cạnh là độ tương tự ngữ nghĩa giữa 2 câu bằng độ đo Cosin. Sử dụng thuật toán PageRank cải tiến cho đồ thị vô hướng để chọn ra những câu quan trọng.

Nghiên cứu của Trương Quốc Định, Nguyễn Quang Dũng [13]: biểu diễn văn bản bằng đồ thị với mỗi đỉnh là một câu, sử dụng thuật toán PageRank cải tiến cho đồ thị vô hướng với trọng số cạnh là độ tương tự giữa hai câu được thử nghiệm bằng 3 độ đo: khoảng cách Jaro, hệ số Jaccard và Cosin. Sau khi thử nghiệm, tác giả chỉ ra rằng sử dụng hệ số Jaccard là hiệu quả hơn cả.

Nghiên cứu của nhóm Nguyễn Quang Uy [57]: Sử dụng lập trình di truyền qua tập đặc trưng: vị trí đoạn, vị trí câu trong đoạn, độ dài câu, tần suất từ (Content-word Frequencies) để xác định những câu quan trọng nhất của văn bản qua quá trình học văn bản mẫu được tóm tắt bằng con người với tỉ lệ 30%.

Đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt" [5]: sử dụng thuật toán PageRank cải tiến để trích rút ra những câu quan trọng dựa trên đặc trưng TFxISF và hệ số nhân cho các

từ xuất hiện trong tiêu đề của văn bản. Kết quả tóm tắt trên kho ngữ liệu được tác giả công bố theo độ đo ROUGE-N với các giá trị 1-gram, 2-gram, 3-gram, 4-gram được trình bày trong bảng 1-1:

Bảng 1-1. Kết quả thử nghiệm của đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt"

n-gram	1	2	3	4
Tóm tắt trích rút	0.539	0.389	0.337	0.311

Hướng tóm tắt tóm lược:

Nghiên cứu của Nguyễn Lê Minh, Akira Shimazu, Xuân Hiếu Phan, Hồ Tú Bảo và Susumu Horiguchi [55]: sử dụng cây cú pháp nhằm rút gọn câu tiếng Việt. Tuy nhiên, các hệ thống phân tích cú pháp tiếng Việt hiện nay có độ chính xác chưa cao nên cách tiếp cận này vẫn chưa thực sự khả thi.

Nghiên cứu của Nguyễn Trọng Phúc và Lê Thanh Hương [10]: sử dụng cấu trúc diễn ngôn trong tóm tắt văn bản tiếng Việt. Cấu trúc diễn ngôn là một phương tiện cho phép biểu diễn mối quan hệ diễn ngôn giữa các đoạn văn bản (như quan hệ nguyên nhân – kết quả). Cây cấu trúc diễn ngôn cho phép đánh giá được tầm quan trọng của các mệnh đề trong câu, các câu trong văn bản. Trên cơ sở đó có thể trích ra được các mệnh đề và các câu quan trọng trong văn bản để đưa vào tóm tắt.

Nghiên cứu của Nguyễn Thị Thu Hà [9] đề xuất xây dựng hệ thống tóm tắt văn bản tiếng Việt dựa trên việc trích rút câu và rút gọn câu với bốn phương pháp khác nhau. Việc trích rút câu được thực hiện theo hai phương pháp: (i) dựa trên lý thuyết tập mờ và mô hình chủ đề; và (ii) dựa trên lượng thông tin và độ ngôn ngữ. Việc rút gọn câu được thực hiện theo hai cách: (i) xác định chuỗi phù hợp và (ii) kết nối các chuỗi con phù hợp nhất.

Đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt" [5]: sử dụng thuật toán PageRank cải tiến để trích

rút ra những câu quan trọng dựa trên đặc trưng TFxISF và hệ số nhân cho các từ xuất hiện trong tiêu đề của văn bản. Sau đó sử dụng các luật diễn ngôn để rút gọn câu đã trích rút tạo ra bản tóm tắt tóm lược cuối cùng.

Hướng tóm tắt đa văn bản:

Trần Mai Vũ [12]: xây dựng hệ thống tóm tắt đa văn bản dựa trên trích rút câu. Để tính độ tương đồng câu, tác giả dựa vào chủ đề ẩn (Latent Dirichlet Allocation), bách khoa toàn thư Wikipedia, và đồ thị quan hệ thực thể.

Đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt" [5]: đề xuất 2 phương pháp tóm tắt đa văn bản ở mức khái quát và ở mức tài liệu. Ở mức khái quát, từng văn bản thuộc cùng một cụm (cluster) sẽ được đưa qua các bộ tóm tắt đơn văn bản để sinh ra văn bản tóm tắt tương ứng. Các văn bản tóm tắt sau đó sẽ được kết hợp lại thành một văn bản tóm tắt tổng hợp. Văn bản này cũng sẽ được đưa qua thành phần tóm tắt đơn văn bản để sinh ra văn bản tóm tắt của toàn bộ cụm. Ở mức tài liệu, nhóm tác giả đề xuất phương pháp tiếp cận khai phá quan điểm dựa trên học máy (cụ thể là xây dựng các bộ phân lớp). Hệ thống bao gồm năm khối: (i) Thu thập và tiền xử lý dữ liệu; (ii) Học bộ phân lớp văn bản chủ quan/khách quan; (iii) Học bộ phân lớp tích cực/tiêu cực; (iv) Áp dụng các bộ phân lớp đã có; (v) Tổng hợp quan điểm. Phương pháp tiếp cận này dựa vào phần mềm dự báo tăng/giảm chứng khoán từ Twitter.

1.5.4 Hiện trạng kho ngữ liệu huấn luyện và đánh giá cho bài toán tóm tắt văn bản tiếng Việt

Cho đến nay, chưa có một kho ngữ liệu huấn luyện và đánh giá phục vụ cho bài toán tóm tắt văn bản tiếng Việt được công bố. Lý do có thể là do để xây dựng kho ngữ liệu lớn cần một số lượng chuyên gia ngôn ngữ và kinh phí đủ lớn. Việc thiếu kho ngữ liệu huấn luyện và đánh giá cho bài toán tóm tắt văn bản tiếng Việt là một lý do quan trọng để giải thích việc tại sao đến nay các nghiên cứu tóm tắt văn bản tiếng Việt còn ít. Mặt khác, do thiếu kho ngữ liệu

huấn luyện và đánh giá mà các phương pháp tóm tắt đã đề xuất cũng chưa được đánh giá so sánh với nhau.

1.5.5 Đặc điểm của các phương pháp tóm tắt văn bản tiếng Việt

Với đối tượng nghiên cứu của đề tài là tập trung vào hướng tóm tắt văn bản theo hướng trích rút. Do vậy, các phương pháp tóm tắt trích rút đã trình bày ở mục 1.1.4.3 có những đặc điểm chung như sau:

- Các đặc trưng văn bản sử dụng trong các phương pháp hầu hết dựa trên các đặc trưng văn bản tiếng Anh mà chưa có khảo sát kỹ việc sử dụng các đặc trưng đó trong văn bản tiếng Việt có phù hợp hay không. Mặt khác, số lượng đặc trưng được sử dụng trong hầu hết các phương pháp còn chưa nhiều (<= 5 đặc trưng) cho nên kết quả tóm tắt còn chưa được cao.

- Chưa có phương pháp xác định ảnh hưởng của từng đặc trưng văn bản trên từng lĩnh vực văn bản trong bài toán tóm tắt văn bản tiếng Việt.

- Chưa có kho ngữ liệu tiêu chuẩn có chú giải dùng cho việc huấn luyện trong bài toán tóm tắt văn bản tiếng Việt. Do vậy, việc so sánh đánh giá chất lượng tóm tắt của từng hệ thống chưa được khách quan và chính xác.

- Hầu hết các phương pháp tóm tắt văn bản mới dừng lại ở mức thử nghiệm, chưa được xây dựng thành các hệ thống ứng dụng trong thực tế

1.6 Các kiến thức cơ sở liên quan

1.6.1 Giải thuật di truyền

Giải thuật di truyền (GA – Genetic Algorithm) là một trong những công cụ chính trong hệ thống tính toán mềm hay còn gọi là trí tuệ tính toán. GA được John Holland đề xuất từ khoảng những năm 70 của thế kỷ trước dựa trên sự mô phỏng quá trình tiến hoá tự nhiên [53]. GA chủ yếu giải quyết vấn đề tìm nghiệm trong lớp các bài toán tối ưu có độ phức tạp tính toán lớn. GA tìm kiếm lời giải của bài toán dựa trên một quần thể được hiểu như một tập những lời giải và tiến hoá quần thể đó dựa trên các toán tử di truyền như chọn lọc, lai ghép, đột biến. Sau khi được giới thiệu, GA đã được các nhà toán học và tin

học nghiên cứu và phát triển rất nhanh, nhiều dạng biến thể cũng như vấn đề cải tiến các toán tử được đề xuất và kết quả thử nghiệm cho thấy tính hiệu quả rõ rệt của giải thuật này.

Giải thuật di truyền đơn giản gồm các bước sau:

- Biểu diễn giải pháp: Đây là một trong những công việc quan trọng trong thiết kế giải thuật di truyền, quyết định việc áp dụng các toán tử tiến hoá. Một trong những biểu diễn truyền thống của GA là biểu diễn nhị phân. Với phép biểu diễn này, giải pháp cho một bài toán được biểu diễn như là một vector bit, còn gọi là nhiễm sắc thể. Mỗi nhiễm sắc thể bao gồm nhiều gen, trong đó một gen đại diện cho một tham số thành phần của giải pháp.

- Lựa chọn: Việc lựa chọn các cá thể được thực hiện khi cần một số cá thể để thực hiện sinh sản ra thế hệ sau. Mỗi cá thể có một giá trị thích nghi (fitness). Giá trị này được dùng để quyết định xem lựa chọn cá thể nào. Một số phương pháp lựa chọn thường dùng bao gồm:

+ Roulette wheel: Dựa trên xác suất (tỷ lệ thuận với giá trị hàm thích nghi) để lựa chọn cá thể.

+ Giao đấu (nhị phân): Chỉ định ngẫu nhiên 2 cá thể, sau đó chọn cá thể tốt hơn trong hai cá thể đó.

- Lai ghép: Toán tử lai ghép được áp dụng nhằm sinh ra các cá thể con mới từ các cá thể cha mẹ, thừa hưởng các đặc tính tốt từ cha mẹ. Trong ngữ cảnh tìm kiếm thì toán tử lai ghép thực hiện tìm kiếm xung quanh khu vực của các giải pháp biểu diễn bởi các cá thể cha mẹ.

- Đột biến: Tương tự như lai ghép, đột biến cũng là toán tử mô phỏng hiện tượng đột biến trong sinh học. Kết quả của đột biến thường sinh ra các cá thể mới khác biệt so với cá thể cha mẹ. Trong ngữ cảnh tìm kiếm, toán tử đột biến nhằm đưa quá trình tìm kiếm ra khỏi khu vực cục bộ địa phương.

1.6.2 Giải thuật tối ưu đàn kiến

Tối ưu đàn kiến ACO là một phương pháp nghiên cứu lấy cảm hứng từ việc mô phỏng hành vi của đàn kiến trong tự nhiên nhằm mục tiêu giải quyết các bài toán tối ưu phức tạp.

Được giới thiệu lần đầu tiên vào năm 1991 bởi A. Colorni và M. Dorigo [27], Giải thuật tối ưu đàn kiến đã nhận được sự chú ý rộng lớn nhờ vào khả năng tối ưu của nó trong nhiều lĩnh vực khác nhau. Khái niệm ACO lấy cảm hứng từ việc quan sát hành vi của đàn kiến trong quá trình chúng tìm kiếm nguồn thức ăn. Người ta đã khám phá ra rằng, đàn kiến luôn tìm được đường đi ngắn nhất từ tổ của chúng đến nguồn thức ăn. Phương tiện truyền đạt tín hiệu được kiến sử dụng đề thông báo cho những con khác trong việc tìm đường đi hiệu quả nhất chính là mùi của chúng (*pheromone*). Kiến để lại vệt mùi trên mặt đất khi chúng di chuyên với mục đích đánh dấu đường đi cho các con theo sau. Vệt mùi này sẽ bay hơi dần và mất đi theo thời gian, nhưng nó cũng có thể được củng cố nếu những con kiến khác tiếp tục đi trên con đường đó lần nữa. Dần dần, các con kiến theo sau sẽ lựa chọn đường đi với lượng mùi dày đặc hơn, và chúng sẽ làm gia tăng hơn nữa nồng độ mùi trên những đường đi được yêu thích hơn. Các đường đi với nồng độ mùi ít hơn rốt cuộc sẽ bị loại bỏ và cuối cùng, tất cả đàn kiến sẽ cùng kéo về một đường đi mà có khuynh hướng trở thành đường đi ngắn nhất từ tổ đến nguồn thức ăn của chúng.

Để bắt chước hành vi của các con kiến thực, Dorigo xây dựng các con kiến nhân tạo *(artificial ants)* cũng có đặc trưng sản sinh ra vết mùi để lại trên đường đi và khả năng lần vết theo nồng độ mùi để lựa chọn con đường có nồng độ mùi cao hơn để đi. Gắn với mỗi cạnh *(i,j)* nồng độ vết mùi τ_{ij} và thông số heuristic η_{ij} trên cạnh đó.

Ban đầu, nồng độ mùi trên mỗi cạnh (i,j) được khởi tạo bằng một hằng số *c*, hoặc được xác định theo công thức:

$$\tau_{ij} = \tau_{ij} = \frac{m_{ant}}{C^{nn}}, \forall(i,j) \quad (1.9)$$

trong đó: τ_{ij} là nồng độ vết mùi trên cạnh *i,j*;

m_{ant} là số lượng kiến ;

C^{nn} là chiều dài hành trình cho bởi phương pháp tìm kiếm gần nhất.

Tại đỉnh *i*, một con kiến *k* sẽ chọn đỉnh *j* chưa được đi qua trong tập láng giềng của *i* theo một quy luật phân bố xác suất được xác định theo công thức:

$$p_{ij}^{k} = \frac{[\tau_{ij}]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{u \in N_i^k}[\tau_{iu}]^{\alpha}[\eta_{iu}]^{\beta}}, j \in N_i^k \quad (1.10)$$

trong đó: p_{ij}^{k} là xác suất con kiến *k* lựa chọn cạnh *i,j ;*

α là hệ số điều chỉnh ảnh hưởng của τ_{ij};

η_{ij} là thông tin heuristic giúp đánh giá chính xác sự lựa chọn của con kiến khi quyết định đi từ đỉnh *i* qua đỉnh *j*; được xác định theo công thức:

$$\eta_{ij} = \frac{1}{d_{ij}} \quad (1.11)$$

d_{ij}: khoảng cách giữa đỉnh *i* và đỉnh *j*;

β: hệ số điều chỉnh ảnh hưởng của η_{ij};

N_i^k: tập các đỉnh láng giềng của *i* mà con kiến *k* chưa đi qua.

Quy luật này mô phỏng hoạt động của một vòng quay xổ số nên được gọi là kỹ thuật bánh xe xổ số.

Con kiến *k* ở đỉnh *i* sẽ lựa chọn đỉnh *j* kế tiếp để đi theo một quy tắc lựa chọn được mô tả bởi công thức sau:

$$j = \begin{cases} \arg_{l \in N_i^k} max\left[(\tau_{il})^{\alpha} \times (\eta_{il})^{\beta}\right] & nếu\ q \le q_0 \\ J & ngược\ lại \end{cases} \quad (1.12)$$

q: giá trị được lựa chọn một cách ngẫu nhiên với một xác suất không thay đổi trong khoảng [0,l]

q_0: là một hằng số cho trước trong khoảng [0,1]

J: là một biến số ngẫu nhiên được lựa chọn theo sự phân bố xác suất cho bởi quy luật phân bố xác suất theo công thức (1.10)

Sau khi cũng như trong quá trình các con kiến tìm đường đi, các vết mùi (τ_{ij}) trên mỗi cạnh sẽ được cập nhật lại, vì chúng bị biến đổi do quá trình bay hơi cũng như quá trình tích lũy mùi khi các con kiến đi trên cạnh đó.

Sau mỗi vòng lặp, vệt mùi trên mỗi cạnh được cập nhật lại:

$$\tau_{ij}(t+1) = (1-\rho) \times \tau_{ij}(t) + \sum_{k=1}^{m_{ant}} \emptyset\, \tau_{ij}^{k}(t) \;\; \forall (i,j) \tag{1.13}$$

trong đó: $0 \leq \rho \leq 1$: tỷ lệ bay hơi của vệt mùi;

$\emptyset\tau_{ij}^{k}(t)$: lượng mùi mà con kiến *k* để lại trên cạnh *ij*, xác định như sau:

$$\emptyset\tau_{ij}^{k} = \begin{cases} \dfrac{Q}{f(k)} & nếu\ con\ kiến\ k\ đi\ qua\ cạnh\ (i,j) \\ 0 & ngược\ lại \end{cases} \tag{1.174}$$

Q: là một hằng số;

f(k): giá trị mục tiêu trong mỗi vòng lặp.

1.6.3 Phương pháp Voting Schulze

Phương pháp Schulze là một phương pháp bầu cử (voting) được Markus Schulze phát triển [80]. Phương pháp này lựa chọn ra một người chiến thắng sử dụng các phiếu bầu có thứ tự (các ứng cử viên được sắp xếp trên phiếu bầu theo thứ tự ưu tiên do người bầu cử quyết định). Phương pháp này cũng có thể đưa ra danh sách theo thứ tự những người chiến thắng. Phương pháp Schulze còn được gọi bằng một số tên khác như Schwartz Sequential Dropping (SSD), Cloneproof Schwartz Sequential Dropping (CSSD), the Beatpath Method, Beatpath Winner, Path Voting, và Path Winner.

Kết quả đầu ra của phương pháp Schulze cho chúng ta một danh sách thứ tự các ứng cử viên. Do đó, nếu cần bầu cử lấy *k* vị trí thì có thể sử dụng ngay

phương pháp này không cần sửa đổi gì bằng cách lấy *k* ứng cử viên có thứ hạng cao nhất là những người được chọn vào *k* vị trí.

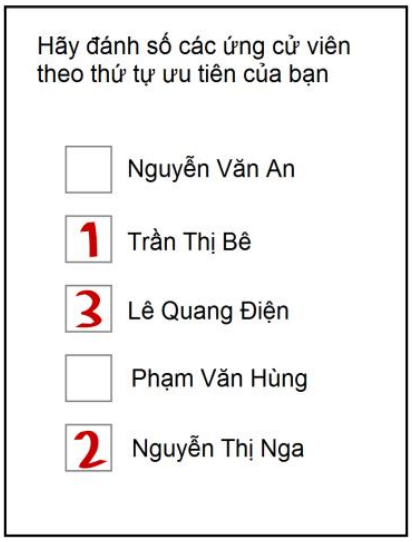
Phương pháp Schulze được sử dụng bởi nhiều tổ chức như Debian, Ubuntu, Gentoo, Software in the Public Interest, Free Software Foundation Europe, Pirate Party associations, …

1.6.3.1 Lá phiếu

Đầu vào cho phương pháp Schulze giống như đầu vào cho các phương pháp bầu cử phiếu bầu có thứ tự ưu tiên khác: mỗi người đi bầu phải sắp xếp danh sách các ứng cử viên theo thứ tự ưu tiên, trong đó có thể cho phép hai ứng cử viên có thứ tự ưu tiên bằng nhau.

Hình 1-8 minh họa một lá phiếu bầu chọn của mô hình chọn nhiều ứng viên. Người đi bầu đánh số đánh số thứ tự ưu tiên của họ trên lá phiếu. Ghi số 1 bên cạnh ứng cử viên ưu tiên cao nhất, ghi số 2 bên cạnh ứng cử viên ưu tiên thứ hai, .v.v. Mỗi người đi bầu có thể:

- Đánh cùng một số thứ tự ưu tiên cho nhiều hơn một ứng cử viên, có nghĩa là đối với người đi bầu thì các ứng cử viên này là tương đương nhau.



Hình 1-7 Ví dụ một lá phiếu cho phương pháp Schulze

- Sử dụng các số không liên tiếp khi đánh thứ tự. Việc này không ảnh

hưởng đến kết quả của cuộc bầu chọn vì chúng ta chỉ quan tâm tới thứ tự của các ứng cử viên mà người đi bầu sắp xếp chứ không phải là con số tuyệt đối do người đi bầu chọn.

- Không đánh số thứ tự một số ứng cử viên. Khi một người đi bầu không đánh số một số ứng cử viên thì có thể hiểu là (i) người đi bầu này ưu tiên tất cả những ứng cử viên được đánh số hơn nhiều những ứng cử viên không được đánh số, và (ii) đối với người đi bầu này thì tất cả những ứng cử viên không được đánh số là tương đương nhau.

1.6.3.2 Phương pháp tính toán

Gọi $d[V,W]$ là số lượng người đi bầu ưu tiên ứng cử viên V hơn ứng cử viên W. Một *đường đi* từ ứng cử viên X đến ứng cử viên Y với độ mạnh p là một chuỗi các ứng cử viên $C(1),\dots,C(n)$ thỏa mãn các tính chất sau:

- $C(1)=X$ *và* $C(n)=Y$.

- *Với mọi* $i=1,\dots,(n-1)$: $d[C(i),C(i+1)]>d[C(i+1),C(i)]$.

- *Với mọi* $i=1,\dots,(n-1)$: $d[C(i),C(i+1)]\geq p$.

Độ mạnh của đường đi mạnh nhất từ ứng cử viên A đến ứng cử viên B, kí hiệu là $p[A,B]$, là giá trị lớn nhất sao cho tồn tại một đường đi từ ứng cử viên A đến ứng cử viên B có độ mạnh bằng giá trị đó. Nếu không tồn tại một đường đi nào từ ứng cử viên A đến ứng cử viên B thì $p[A,B]=0$.

Ứng cử viên D được định nghĩa là được bầu cao hơn ứng cử viên E khi và chỉ khi $p[D,E]>p[E,D]$.

Ứng cử viên D là một người chiến thắng tiềm năng khi và chỉ khi $p[D,E]\geq p[E,D]$ với mọi ứng cử viên E khác D.

Mối quan hệ "được bầu cao hơn" $\mathcal{O}$ được định nghĩa như sau:

$$XY\in\mathcal{O}\Leftrightarrow p[X,Y]>p[Y,X].$$

Tập hợp $S=\{X|\forall Y\neq X: XY\notin\mathcal{O}\}$ là tập hợp những người chiến thắng.

Định lý 1.1 [49]**:** Mối quan hệ $\mathcal{O}$ có tính chất bắc cầu.

Định lý 1.2 [49]**:** Trong mọi trường hợp, phương pháp Schuzle luôn luôn tìm được người chiến thắng.

Bước khó nhất khi cài đặt thuật toán cho phương pháp Schulze là bước tính toán độ mạnh của các đường đi mạnh nhất. Có thể sử dụng thuật toán Floyd [68] để giải quyết vấn đề này. Các bước của thuật toán được mô tả cụ thể trong [49].

Để hiểu rõ hơn về phương pháp Schulze, chúng ta có thể xem ví dụ minh họa phương pháp trong [80].

1.7 Kết luận Chương 1

Các kết quả Chương 1 đạt được bao gồm:

(1). Đã nghiên cứu, trình bày tổng quan các giai đoạn và tham số của hệ thống tóm tắt văn bản. Các phương pháp tiếp cận tóm tắt văn bản trên thế giới theo hai hướng: Tóm tắt trích rút (ES) và tóm tắt tóm lược (AS).

(2). Đã nghiên cứu, trình bày tổng quan các phương pháp tiếp cận tóm tắt văn bản tiếng Việt trong những năm gần đây. Qua đó phân tích, đánh giá hiện trạng nghiên cứu tóm tắt văn bản tiếng Việt.

(3). Đã nghiên cứu, trình bày tổng quan về giải thuật di truyền, giải thuật tối ưu đàn kiến và phương pháp Voting Schulze.

Việc nghiên cứu các phương pháp tiếp cận tóm tắt văn bản, các phương pháp đánh giá tóm tắt văn bản và kiến thức cơ sở liên quan là tiền đề để nghiên cứu, xây dựng phát triển các kỹ thuật tóm tắt văn bản tiếng Việt được trình bày trong chương 2 và chương 3.

CHƯƠNG 2. TÓM TẮT VĂN BẢN TIẾNG VIỆT DỰA TRÊN BỘ HỆ SỐ ĐẶC TRƯNG

Trong chương này, luận án trình bày việc lựa chọn tập đặc trưng quan trọng cho văn bản tiếng Việt thông qua khảo sát kho ngữ liệu mẫu, qua đó đề xuất cải tiến một số đặc trưng cho phù hợp với văn bản tiếng Việt. Trên cơ sở các đặc trưng này, luận án đề xuất phương pháp tóm tắt văn bản tiếng Việt dựa trên bộ hệ số đặc trưng được xác định bằng phương pháp học máy sử dụng giải thuật di truyền và giải thuật tối ưu đàn kiến. Cuối cùng, luận án trình bày các kết quả thử nghiệm và đánh giá.

2.1 Mô hình tóm tắt văn bản tiếng Việt dựa trên bộ hệ số đặc trưng

Như đã trình bày tổng quan về tóm tắt văn bản và tóm tắt văn bản tiếng Việt ở mục 1.1, hiện nay, tiếp cận tóm tắt văn bản dựa trên trích rút câu được dùng phổ biến nhất. Mục đích của cách tiếp cận này là trích rút ra những câu quan trọng trong văn bản, phản ánh được nhiều thông tin từ văn bản gốc.

2.1.1 Quy trình tóm tắt văn bản theo hướng trích rút

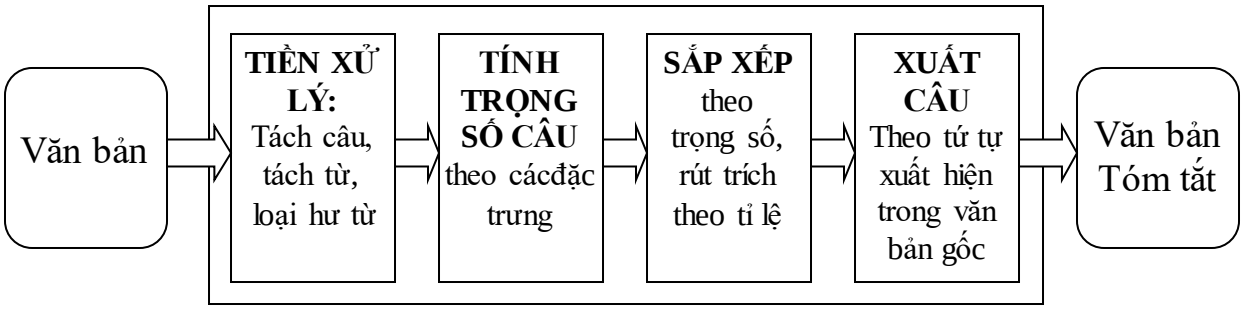
Quy trình tóm tắt văn bản theo hướng trích rút câu được mô tả như sau:

Bước 1. Tiền xử lý văn bản đầu vào: tách câu, tách từ, gán nhãn từ loại, lọc bỏ các hư từ.

Bước 2. Tính trọng số các câu theo các đặc trưng văn bản.

Bước 3. Chọn các câu có trọng số tốt nhất từ trên xuống theo tỉ lệ.

Bước 4. Xuất các câu đã trích rút theo thứ tự xuất hiện trong văn bản gốc.



Hình 2-1 Quy trình cách tiếp cận TTVB dựa trên trích rút câu.

Để xác định được trọng số của câu người ta thường dựa trên các đặc trưng quan trọng như: vị trí của câu trong văn bản, các từ quan trọng xuất hiện trong câu, độ tương tự tiêu đề, ... [17],[76]. Công thức tổng quát để tính trọng số câu thông qua tập đặc trưng quan trọng:

$$Score(s) = \sum_{i=1}^{n} k_i \times Score_{f_i}(s) \quad (2.1)$$

trong đó: s là một câu trong văn bản;

n là số đặc trưng;

k_i là hệ số đặc trưng thứ i của văn bản;

$Score_{f_i}(s)$ là trọng số của đặc trưng thứ i trong câu s;

Ta cho thể biểu diễn bài toán tóm tắt đơn văn bản tiếng Việt theo hướng trích rút như sau:

Bài toán tóm tắt văn bản theo hướng trích rút được xác định bởi các dữ liệu:

$$(a, d = (s_1, s_2, \dots, s_{Ns}), f = (f_1, f_2, \dots, f_n), k = (k_1, k_2, \dots, k_n))$$

trong đó:

a: tỷ lệ tóm tắt ($a < 1$);

$d = (s_1, s_2, \dots, s_{Ns})$: văn bản cần tóm tắt;

$s_j, (j = 1, \dots, Ns)$: câu thứ j của văn bản cần tóm tắt d;

$f = (f_1, f_2, \dots, f_n)$: tập đặc trưng văn bản;

$f_i, (i = 1, \dots, n)$: đặc trưng thứ i;

$k = (k_1, k_2, \dots, k_n)$: tập hệ số đặc trưng;

$k_i, (i = 1, \dots, n)$: hệ số đặc trưng thứ i .

Định nghĩa 2.1: *Bài toán tóm tắt văn bản theo hướng trích rút số câu gốc của văn bản d theo tỉ lệ tóm tắt a<1 được biểu diễn như sau:*

$$Sum(a, d, f, k) = (s_{y1,} s_{y2}, \dots, s_{yz,}); \ 1 \le y1 < y2 \dots \le l \text{ và } z = [l * a]$$

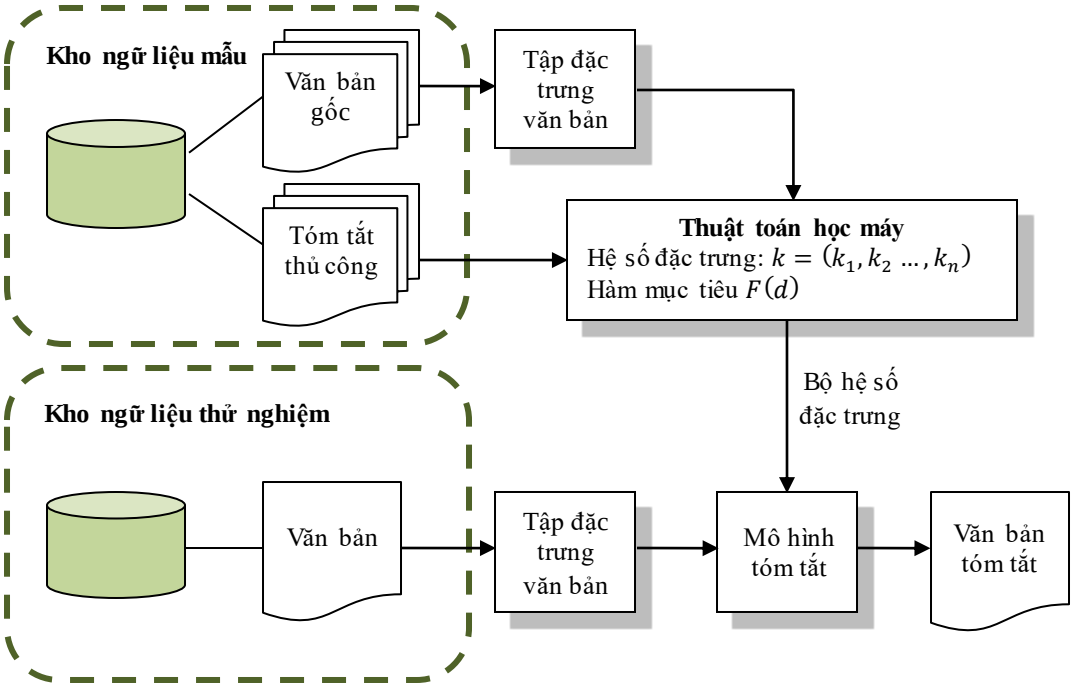
trong đó: m là số câu trong văn bản d; s_i là câu thứ i trong văn bản d.

2.1.2 Mô hình tóm tắt văn bản dựa trên bộ hệ số đặc trưng

Qua công thức (2.1), ta có thể nhận xét rằng, bài toán tóm tắt văn bản tiếng Việt cần xác định được 2 yếu tố quan trọng là:

- Cần phải xác định tập đặc trưng quan trọng của văn bản tiếng Việt.
- Phương pháp xác định bộ hệ số đặc trưng (vai trò của từng đặc trưng).

Đây chính là hạn chế của phương pháp tóm tắt văn bản tiếng Việt theo hướng trích rút trước đây [76]. Chính vì vậy luận án đề xuất mô hình tóm tắt đơn văn bản tiếng Việt theo hướng trích rút dựa trên bộ hệ số đặc trưng (sau đây gọi tắt là ***VTS_FC***) được mô tả như hình 2-2:



*Hình 2-2 Mô hình tóm tắt văn bản tiếng Việt **VTS_FC***

Mô hình được thực hiện theo 2 bước:

Bước 1: Sử dụng phương pháp học máy có giám sát để xác định bộ hệ số đặc trưng quan trọng của văn bản tiếng Việt thông qua việc học kho ngữ liệu tóm tắt mẫu.

Bước 2: Sử dụng bộ hệ số đặc trưng để tính toán trọng số câu theo công thức (2.1). Sau đó, sắp xếp lại câu theo trọng số và trích rút ra theo tỉ lệ tóm tắt.

Nhận xét:

Mô hình tóm tắt văn bản tiếng Việt ***VTS_FC*** được đề xuất dựa trên ý tưởng: Hệ thống tóm tắt học được "cách tóm tắt của con người" thông qua việc đánh giá vai trò của từng đặc trưng trong bản tóm tắt do con người thực hiện. Đây là mô hình phù hợp để tóm tắt văn bản theo từng lĩnh vực cụ thể vì thực tế mỗi lĩnh vực có quan điểm tóm tắt văn bản khác nhau. Do vậy, dựa vào tập ngữ liệu tóm tắt mẫu của từng lĩnh vực, mô hình *VTS_FC* sẽ xác định được bộ hệ số đặc trưng thích hợp cho từng quan điểm tóm tắt. Thông qua bộ hệ số này, văn bản gốc thuộc lĩnh vực nào sẽ được hệ thống tóm tắt với độ chính xác thích hợp nhất dựa vào bộ hệ số đặc trưng của lĩnh vực đó.

2.2 Lựa chọn tập đặc trưng cho văn bản tiếng Việt

Qua quá trình nghiên cứu về tiếng Việt và các phương pháp tóm tắt văn bản tiếng Việt đã công bố đã được trình bày ở mục 1.1.4.3. Có thể nhận thấy rằng, hầu hết các đặc trưng được các tác giả lựa chọn đều dựa vào đặc trưng cho văn bản tiếng Anh, tuy nhiên các tác giả chưa có một sự khảo sát kỹ về việc sử dụng các đặc trưng đó trong tiếng Việt có phù hợp hay không. Do vậy, để xây dựng tập đặc trưng sử dụng cho phương pháp này, luận án tập trung khảo sát từng đặc trưng một cách khoa học dựa trên bộ kho ngữ liệu văn bản tiếng Việt khá lớn được trình bày trong phần phụ lục. Qua quá trình khảo sát, luận án đề xuất cải tiến một số đặc trưng phù hợp với văn bản tiếng Việt.

Trong nghiên cứu, luận án sử dụng quan điểm phân loại từ vựng tiếng Việt thành 2 lớp là thực từ và hư từ của Diệp Quang Ban [1] được mô tả trong hình 1-6. Thực từ là những từ mang thông tin còn hư từ là những từ chỉ có chức năng ngữ pháp (không mang thông tin). Do vậy, luận án chỉ lựa chọn và thực hiện tính toán các đặc trưng dựa trên thực từ, còn hư từ bị loại bỏ. Ngoài ra, để nâng cao độ chính xác, trong quá trình tính giá trị các đặc trưng thì các thực từ đồng nghĩa trong tiêu đề, nội dung được thay thế bằng một từ duy nhất bằng cách sử dụng từ điển đồng nghĩa của tác giả Nguyễn Văn Tu [11].

2.2.1 Ví trí câu

Định nghĩa 2.2: *Độ quan trọng của câu của văn bản dựa theo đặc trưng vị trí được xác định là giá trị vị trí của câu trong một đoạn văn bản.*

Với các nghiên cứu trước đây về tóm tắt văn bản, vị trí câu đóng vai trò khá quan trọng. Có phương pháp sử dụng câu đầu tiên trong đoạn (hoặc toàn bộ văn bản) là quan trọng hơn các câu khác trong đoạn (hoặc toàn bộ văn bản) [29],[19], có phương pháp sử dụng cả câu đầu tiên và câu cuối trong đoạn (hoặc toàn bộ văn bản) là câu quan trọng hơn các câu khác trong đoạn (hoặc toàn bộ văn bản) [76],[55]. Để xác định vai trò của đặc trưng vị trí câu trong văn bản tiếng Việt, chúng ta dựa vào khảo sát phân bố vị trí câu quan trọng trong kho ngữ liệu mẫu văn bản tiếng Việt là Corpus_LTH và ViEvTextSum (trình bày trong phần phụ lục). Qua đó xây dựng công thức tính giá trị vị trí câu phù hợp với văn bản tiếng Việt.

Thực hiện khảo sát phân bố của vị trí câu quan trọng trong đoạn văn theo các bước như sau:

Bước 1: Các câu trong văn bản của kho ngữ liệu mẫu được gán nhãn: D: câu đầu, G: Câu giữa, C: câu cuối. Các câu giữa được gán nhãn G_d: đoạn đầu của các câu giữa, G_g: đoạn giữa của các câu giữa, G_c: đoạn cuối của các câu giữa theo quy tắc:

- Nếu đoạn có 1 câu: gán câu = "DC";
- Nếu đoạn có 2 câu: gán câu đầu = "D" câu cuối = "C";
- Nếu đoạn có 3 câu: câu đầu ="D"; câu giữa= "GdGgGc"; câu cuối = "C";
- Nếu đoạn có 4 câu: câu đầu ="D"; câu thứ 2= "Gd"; câu thứ 3= "Gc"; câu cuối = "C";
- Nếu đoạn có nhiều hơn 4 câu: câu đầu ="D"; câu thứ 2= "Gd"; câu thứ 3 đến câu gần câu gần cuối= "Gg"; câu gần cuối = "Gc"; câu cuối = "C".

Bước 2: Trọng số câu được xác định bằng độ đo đồng xuất hiện của các thực từ trong câu với đoạn văn bản tóm tắt do con người thực hiện. Độ đo đồng

xuất hiện được tính theo công thức (2.2):

$$Sim(S,SH) = \frac{|S \cap SH|}{|SH|} \quad (2.2)$$

trong đó: $S = \{s_1, s_2, \dots, s_N\}$: vector thực từ khác nhau của câu;

$SH = \{sh_1, sh_2, \dots, sh_M\}$: vector thực từ khác nhau của đoạn văn bản tóm tắt con người;

$|S \cap SH|$: là số thực từ đồng xuất hiện trong S và SH .

Bước 3: Thực hiện tóm tắt toàn bộ văn bản gốc trong kho ngữ liệu mẫu dựa vào giá trị câu tính theo công thức (2.2).

Bước 4: Thống kê phân bố vị trí của các câu quan trọng theo tập nhãn trong kết quả tóm tắt văn bản vừa thực hiện.

Kết quả phân bố xác suất câu quan trọng trong kết quả tóm tắt được mô tả dưới bảng 2-1:

Bảng 2-1. Kết quả khảo sát vị trí câu quan trọng kho ngữ liệu tiếng Việt

Vị trí câu	Câu đầu (D)	Câu giữa (G)			Câu cuối
		G_d	G_g	G_c	
Phân bố $F_{vt}(s)$	0,60	0,08	0,06	0,07	0,19

Qua bảng kết quả khảo sát, luận án sử dụng giá trị phân bố vị trí câu làm cơ sở để tính độ quan trọng của câu theo đặc trưng vị trí câu.

$$Score_{f1}(s) = F_{vt}(s) \quad (2.3)$$

trong đó: $F_{vt}(s)$ là giá trị phân bố vị trí câu được tính theo bảng 2-1.

2.2.2 Trọng số TF.ISF

Định nghĩa 2.3: *Độ quan trọng của câu trong văn bản dựa theo đặc trưng trọng số TF.ISF được tính bằng giá trị trung bình cộng các trọng số TF.ISF của các thực từ trong câu (được chuẩn hóa về đoạn [0,1]).*

Phương pháp này bắt nguồn từ công thức nổi tiếng *TF.IDF* (term frequency – inverse document frequency)[70] được sử dụng để xác định mức

độ quan trọng của từ trong một văn bản, mà văn bản đó nằm trong một tập hợp các văn bản. Ở đây, luận án tiếp cận bài toán đơn văn bản nên sử dụng trọng số *TF.ISF* (Term frequency- inverse sentence frequency) làm đặc trưng để tính độ quan trọng của câu.

$$Score_{TF-ISF}(s) = \frac{1}{N_w}\sum_{k=1}^{N_w} TF(w_k, s) \times ISF(w_k) \quad (2.4)$$

trong đó: N_w là số các thực từ có trong câu *s*;

w_k là thực từ thứ *k* trong câu *s*;

$TF(w_k, s)$ là số lần xuất hiện của thực từ w_k trong câu *s*;

$ISF(w_k) = log\left(\frac{N_s}{SF(w_k)}\right)$ là tần số nghịch của từ w_k trong tập câu thuộc văn bản (N_S là tổng số câu có trong văn bản; $SF(w_k)$ là tổng số câu trong văn bản có chứa thực từ w_k).

Do giá trị của công thức (2.4) tương đối lớn, do vậy giá trị đặc trưng này được chuẩn hóa về khoảng [0,1]. Công thức tính giá trị câu theo đặc trưng TF-ISF được tính theo công thức (2.5):

$$Score_{f2}(s) = \frac{Score_{TF-ISF}(s)}{Max(Score_{TF-ISF}(s), s \in d)} \quad (2.5)$$

trong đó: *d* là văn bản gốc.

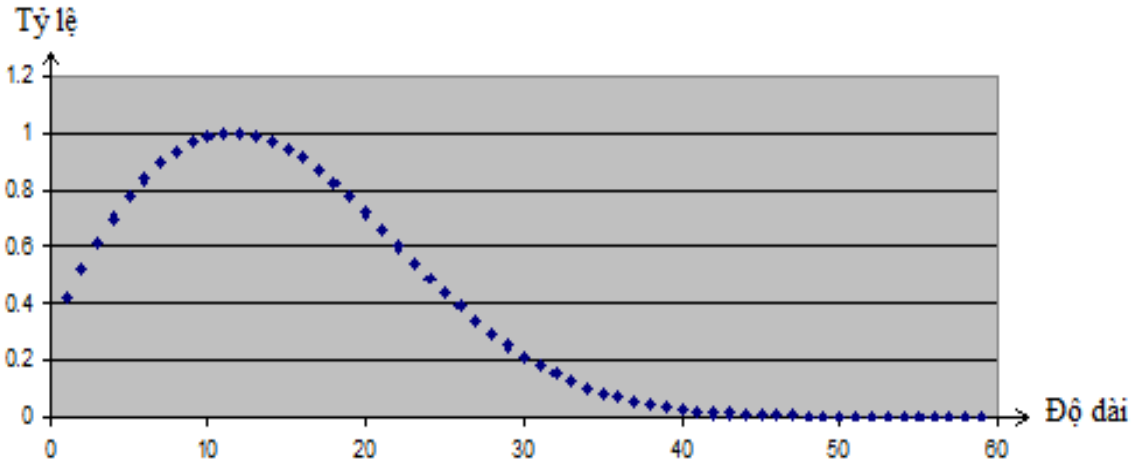
2.2.3 Độ dài câu

Định nghĩa 2.4: *Độ quan trọng của câu trong văn bản dựa theo đặc trưng độ dài câu được tính bằng giá trị phân bố độ dài câu tính theo thực từ trong kho ngữ liệu lớn.*

Khác với các quan điểm trước đây về độ dài câu của các nghiên cứu tóm tắt văn bản là câu quá ngắn hoặc quá dài đều không xuất hiện trong bản tóm tắt [75],[76]. Ở đây, sau khi khảo sát kho ngữ liệu tiếng Việt, kết quả cho thấy giá trị độ dài câu đều có vai trò trong việc xác định độ quan trọng của từng câu. Do vậy, giá trị đặc trưng độ dài câu được xác định thông qua sự phân bố độ dài câu

trong toàn bộ kho ngữ liệu tiếng Việt đã được thu thập.

Sơ đồ phân bố độ dài câu theo thực từ và chuẩn hoá về đoạn [0,1] của kho ngữ liệu tiếng Việt hơn 20.000 văn bản tiếng Việt với 202.785 câu được thu thập được mô tả trong hình 2-3.



Hình 2-3 Sơ đồ phân bố độ dài câu tính theo thực từ.

Công thức độ dài câu được xây dựng dựa theo đồ thị phân bố trong hình 2-3, mô tả như sau:

$$Score_{f3}(s) = \begin{cases} ax^2 + bx + c, 0 < x < 12 \\ \dfrac{\alpha}{\sigma\sqrt{2\pi}} exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right), x > 12 \end{cases} \quad (2.6)$$

trong đó: x là độ dài câu s tính theo thực từ; a = - 0.00529; b = 0.12174; c = 0.3; α = 26.3; μ = 11.5; σ = 10.5.

2.2.4 Xác suất thực từ

Định nghĩa 2.5: *Độ quan trọng của câu trong văn bản dựa theo đặc trưng xác xuất thực từ được tính bằng giá trị trung bình cộng xác suất của các thực từ trong câu.*

Đặc trưng này sử dụng xác suất thực từ để làm nền tảng tính toán trọng số câu. Câu có chứa nhiều thực từ có tần suất xuất hiện cao trong toàn bộ văn bản thì câu đó càng quan trọng [69]. Công thức tính độ quan trọng của câu tính theo xác suất thực từ được mô tả như sau:

$$Score_{f4}(s) = \frac{1}{N_w}\sum_{k=1}^{N_w}\frac{C(w_k)}{N} \quad (2.7)$$

trong đó: N_w : số các thực từ có trong câu *s*;

$C(w_k)$: số lần xuất hiện của thực từ w_k của câu *s* trong toàn bộ văn bản;

N: số các thực từ có trong văn bản.

2.2.5 Thực thể tên

Đặc trưng thực thể tên được đưa ra theo quan điểm các thuật ngữ tên riêng (tên người, tên địa danh, tên tổ chức, tên quốc gia…) thường truyền đạt những thông tin quan trọng trong các loại văn bản tin tức [69]. Do vậy, luận án đã thực hiện khảo sát sự phân bố thực thể tên trong toàn bộ kho ngữ liệu tóm tắt tiếng Việt (Corpus_LTH và ViExTextSum) trên văn bản gốc và bản tóm tắt thủ công để xác định đặc trưng này đóng vai trò như thế nào trong văn bản tiếng Việt. Thực thể tên được xác định thông qua quá trình gán nhãn cho kho ngữ liệu huấn luyện bằng nhãn Np, Ny [79].

Qua kết quả khảo sát 2 kho ngữ liệu được trình bày trong bảng 2.2 và 2.3, có thể nhận thấy rằng thực thể tên đóng vai trò quan trọng trong văn bản tiếng Việt thuộc thể loại tin tức. Do vậy, việc sử dụng đặc trưng này trong bài toán tóm tắt văn bản tiếng Việt là hợp lý.

Định nghĩa 2.6: *Độ quan trọng của câu trong văn bản dựa theo đặc trưng thực thể tên được tính bằng thương của số thực thể tên xuất hiện trong câu và số thực từ có trong câu.*

$$Score_{f5}(s) = \frac{N_{name}(s)}{N_w(s)} \quad (2.8)$$

trong đó: $N_{name}(s)$ là số thực thể tên xuất hiện trong câu s;

$N_w(s)$ số các thực từ có trong câu s.

Bảng 2-2. Kết quả phân bố thực thể tên trên văn bản tóm tắt mẫu

	Corpus_LTH (văn bản)	ViEvTextSum (văn bản)
Không chứa thực thể tên	*18*	*1033*
Chứa 1 thực thể tên	24	1420
Chứa 2 thực thể tên	17	1431
Chứa 3 thực thể tên	31	1261
Chứa 4 thực thể tên	22	992
Chứa 5 thực thể tên	14	673
Chứa 6 thực thể tên	10	441
Chứa 7 thực thể tên	16	311
Chứa 8 thực thể tên	13	191
Chứa 9 thực thể tên	9	115
Chứa 10 thực thể tên	7	75
Chứa hơn 10 thực thể tên	19	118
Tổng số văn bản	**200**	**8061**

Bảng 2-3. Kết quả phân bố thực thể tên trên các câu của văn bản gốc

	Corpus_LTH (câu)	ViEvTextSum (câu)
Số câu không chứa thực thể tên	1.651	77.456
Số câu chứa thực thể tên	2.212	82.933
Tổng số câu	**3.863**	**160.389**

2.2.6 Dữ liệu số

Đặc trưng này được đưa ra dựa theo quan điểm của một số nhà nghiên cứu tóm tắt văn bản trên thế giới xem rằng các thuật ngữ được viết dưới hình thức số (số, số bằng chữ, ngày tháng năm, …) đôi khi truyền đạt thông tin quan trọng [21],[69]. Để xác định đặc trưng này đóng vai trò như thế nào trong văn bản tiếng Việt, thực hiện khảo sát phân bố thực từ là dữ liệu số trên 2 kho ngữ liệu văn bản Corpus_LTH và ViExTextSum trên cả bản tóm tắt mẫu và văn bản

gốc. Các thực từ là dữ liệu số được nhận biết bằng nhãn M được định nghĩa trong thông qua quá trình gán nhãn [79].

Qua kết quả khảo sát 2 kho ngữ liệu được trình bày trong bảng 2.4 và 2.5, có thể nhận thấy rằng đặc trưng dữ liệu số cũng có vai trò trong văn bản tiếng Việt thuộc thể loại tin tức. Do vậy, việc sử dụng đặc trưng này trong bài toán tóm tắt văn bản tiếng Việt là hợp lý.

Bảng 2-4. Kết quả phân bố dữ liệu số trên văn bản tóm tắt mẫu

	Corpus_LTH (văn bản)	ViEvTextSum (văn bản)
Không chứa dữ liệu số	20	1468
Chứa 1 dữ liệu số	32	2186
Chứa 2 dữ liệu số	32	1699
Chứa 3 dữ liệu số	33	1220
Chứa 4 dữ liệu số	25	709
Chứa 5 dữ liệu số	7	395
Chứa 6 dữ liệu số	17	201
Chứa 7 dữ liệu số	10	89
Chứa 8 dữ liệu số	6	34
Chứa 9 dữ liệu số	7	33
Chứa 10 dữ liệu số	1	12
Chứa hơn 10 dữ liệu số	8	15
Tổng số văn bản	**200**	**8061**

Bảng 2-5. Kết quả phân bố dữ liệu số trên các câu của văn bản gốc

	Corpus_LTH (câu)	ViEvTextSum (câu)
Số câu không chứa dữ liệu số	1.923	84.971
Số câu chứa dữ liệu số	1.940	75.418
Tổng số câu	**3.863**	**160.389**

Định nghĩa 2.7: *Độ quan trọng của câu trong văn bản dựa theo đặc trưng dữ liệu số được tính bằng thương của số thực từ là dữ liệu số xuất hiện trong câu và số thực từ có trong câu.*

$$Score_{f6}(s) = \frac{N_{num}(s)}{N_w(s)} \tag{2.9}$$

trong đó: $N_{num}(s)$ là số thực từ dữ liệu số xuất hiện trong câu s;

$N_w(s)$ số các thực từ có trong câu s.

2.2.7 Tương tự với tiêu đề

Định nghĩa 2.8: *Độ quan trọng của câu trong văn bản dựa theo đặc trưng tương tự với tiêu đề được tính bằng phép đo đồng xuất hiện thực từ giữa câu và câu tiêu đề.*

Đặc trưng này xem xét độ đồng xuất hiện thực từ giữa câu và câu tiêu đề của văn bản. Được tính dựa theo phép đo đồng xuất hiện Dice [26]:

$$Score_{f7}(s) = Sim_{Dice}(s,T) = 2 \times \frac{|s \cap T|}{|s| + |T|} \tag{2.10}$$

trong đó: $s = \{s_1, s_2, \dots, s_N\}$: vector thực từ khác nhau của câu;

$T = \{t_1, t_2, \dots, t_M\}$: vector thực từ khác nhau của câu tiêu đề;

$|S \cap T|$: là số thực từ đồng xuất hiện trong S và T.

2.2.8 Câu trung tâm

Định nghĩa 2.9: *Độ quan trọng của câu trong văn bản dựa theo đặc trưng câu trung tâm được tính bằng giá trị trung bình cộng độ tương tự giữa một câu và các câu khác trong văn bản.*

Đặc trưng này xem xét độ đồng xuất hiện của các thực từ giữa một câu và các câu khác trong văn bản. Giá trị đặc trưng này được tính toán dựa vào phương pháp Aggregation Similarity [51], được mô tả bằng công thức (2.11):

$$Score_{f8}(s_i) = \frac{1}{N_s} \sum_{j=1, j \neq i}^{N_s} Sim_{Dice}(s_i, s_j) \tag{2.11}$$

trong đó: N_s là tổng số câu có trong văn bản;

$Sim_{Dice}(s_i, s_j)$ là phép đo đồng xuất hiện thực từ giữa câu thứ *i* với câu thứ *j* được tính bằng công thức (2.10).

2.3 Xác định hệ số đặc trưng bằng phương pháp học máy

2.3.1 Đặt bài toán

Theo phương pháp tóm tắt văn bản theo hướng trích rút đã được trình bày trong mục 2.1, có hai vấn đề cần được xem xét: Thứ nhất, phải xem xét sự phù hợp của từng đặc trưng trong bài toán tóm tắt văn bản tiếng Việt và lựa chọn được tập đặc trưng quan trọng của tiếng Việt. Thứ hai, mỗi giá trị đặc trưng sử dụng phải được xác định hệ số sao cho thích hợp nhất đối với bài toán. Trong phương pháp tóm tắt văn bản tiếng Việt theo hướng trích rút [76], tác giả đã dựa vào 5 đặc trưng văn bản: Từ tiêu đề, vị trí câu trong đoạn, danh từ, độ tương đồng giữa hai đoạn, TFxIPF, sau đó thực hiện điều chỉnh các hệ số đặc trưng thông qua quá trình thử nghiệm mà chưa có một phương pháp hiệu quả để thực hiện việc xác định hệ số đặc trưng này.

Trong phần này của luận án sẽ đề cập phương pháp xác định bộ hệ số của các đặc trưng trong mô hình ***VTS_FC*** dựa trên phương pháp tối ưu. Như vậy bài toán đặt ra là tìm kiếm bộ hệ số của các đặc trưng sao cho bản tóm tắt thu được dựa vào công thức (2.1) là "tốt nhất".

Tuy nhiên với số lượng đặc trưng được sử dụng nhiều thì sẽ tạo ra tổ hợp số lượng các bộ hệ số *k* lớn. Do đó để xác định được bộ hệ số *k* tối ưu khó thực hiện theo cách trực quan của người dùng do độ phức tạp được tăng theo hàm mũ. Do vậy, chúng ta sẽ đưa việc xác định bộ hệ số *k* vào bài toán tìm kiếm tối ưu sử dụng các giải thuật phỏng quá trình tự nhiên.

Bài toán tìm hệ số đặc trưng cho bài toán tóm tắt văn bản được xác định bởi các dữ liệu sau:

$$(m, a, D = (d_1, d_2, \dots, d_m), SH = (sh_1, sh_2, \dots, sh_m), f = \left(f_1^i, f_2^i, \dots, f_n^i\right);\ i = 1..m)$$

trong đó:

- m là số văn bản huấn luyện;
- n là số đặc trưng văn bản (n = 8);
- a là tỷ lệ tóm tắt;
- D là tập văn bản gốc;
- Đối với mỗi văn bản học thứ j trong tập văn bản mẫu D:
  + d_j là văn bản gốc thứ j (chứa tiêu đề và nội dung);
  + sh_j là bản tóm tắt do con người thực hiện của văn bản d_j;
  + $f_i^j; i = 1 \dots n$ là giá trị đặc trưng thứ i của văn bản gốc thứ j.

Bài toán đặt ra là tìm bộ hệ số đặc trưng k sao cho bản tóm tắt trích rút dựa vào các đặc trưng theo tỉ lệ tóm tắt a "gần giống" với bản tóm tắt con người.

Định nghĩa 2.10: *Một bộ hệ số là một vector* $k = (k_1, k_2 \dots, k_n), k_i \in \mathbb{R}$ *với* k_i *là hệ số của đặc trưng* t_i*. Bộ hệ số gọi là chấp nhận được nếu nó thỏa mãn điều kiện* $1 \geq k_i \geq 0$.

Một bản "tóm tắt vàng" của hệ thống sinh ra cần đạt được tiêu chí là chứa hầu hết các từ liên quan trong văn bản tóm tắt của con người. Độ đo đánh giá văn bản tóm tắt được định nghĩa như sau:

Định nghĩa 2.11: *Độ đo đánh giá văn bản tóm tắt được định nghĩa bằng độ tương tự giữa văn bản tóm tắt của hệ thống với văn bản tóm tắt con người theo độ đo đồng xuất hiện của thực từ trong văn bản tóm tắt hệ thống và văn bản tóm tắt con người:*

$$Sim(Sum(a, d_i, f, k), sh_i) = \frac{|Sum(a, d_i, f, k) \cap sh_i|}{|sh_i|}; i = 1 \dots m \qquad (2.12)$$

trong đó: $Sum(a, d_i, f, k) = \{sm_{i1}, \dots, sm_{ir}\}$ là vector thực từ khác nhau của văn bản tóm tắt của hệ thống theo bộ đặc trưng f và bộ hệ số k theo tỉ lệ tóm tắt

a của văn bản d_i; $sh_i = \{sh_{i1}, \dots, sh_{il}\}$ là vector thực từ khác nhau của văn bản sh_i.

Phát biểu bài toán (sau đây được gọi $DFC(m, a, D, SH, f)$):

Giả sử $k = (k_1, k_2 \dots, k_n)$ là bộ hệ số đặc trưng chấp nhận được. Tìm k sao cho hàm mục tiêu:

$$F(k) = \sum_{i=1}^{m} \frac{Sim(Sum(a, d_i, f, k), sh_i)}{m} \Longrightarrow \text{Giá trị cực đại} \quad (2.13)$$

với miền ràng buộc:

$$1 \geq k_i \geq 0 \quad (2.14)$$

Nhận xét:

Việc tìm bộ hệ số tối ưu cho bài toán (2.13)-(2.14) bằng các phương pháp tối ưu thông thường là hết sức khó khăn do hàm mục tiêu (2.13) phụ thuộc vào các vec tơ thực từ nên có tính rời rạc, còn ràng buộc (2.14) lại là miền liên tục.

Với số đặc trưng tăng lên, không gian tìm kiếm càng lớn yêu cầu cần phải có một giải thuật tốt để tăng tốc độ và hiệu quả của giải thuật. Sự ra đời của giải thuật Meta-Heuristic đã giải quyết các bài toán tối ưu với hiệu quả cao cho kết quả lời giải gần tối ưu như họ giải thuật đàn kiến (Ant Algorithm), giải thuật luyện SA (Simulated Annealing), thuật giải tối ưu đàn kiến ACO (Ant colony optimization), giải thuật di truyền GA (Genetic Algorithm).

Để có thể áp dụng các giải thuật nêu trên, do tính chất rời rạc của hàm mục tiêu, luận án đề xuất tìm kiếm bộ hệ số tối ưu k trong không gian rời rạc: $k_i \in \{h, 2.h, \dots, M.h = 1\}, i = 1,2, \dots, n$; với h là bước chia hay độ chính xác tìm kiếm.

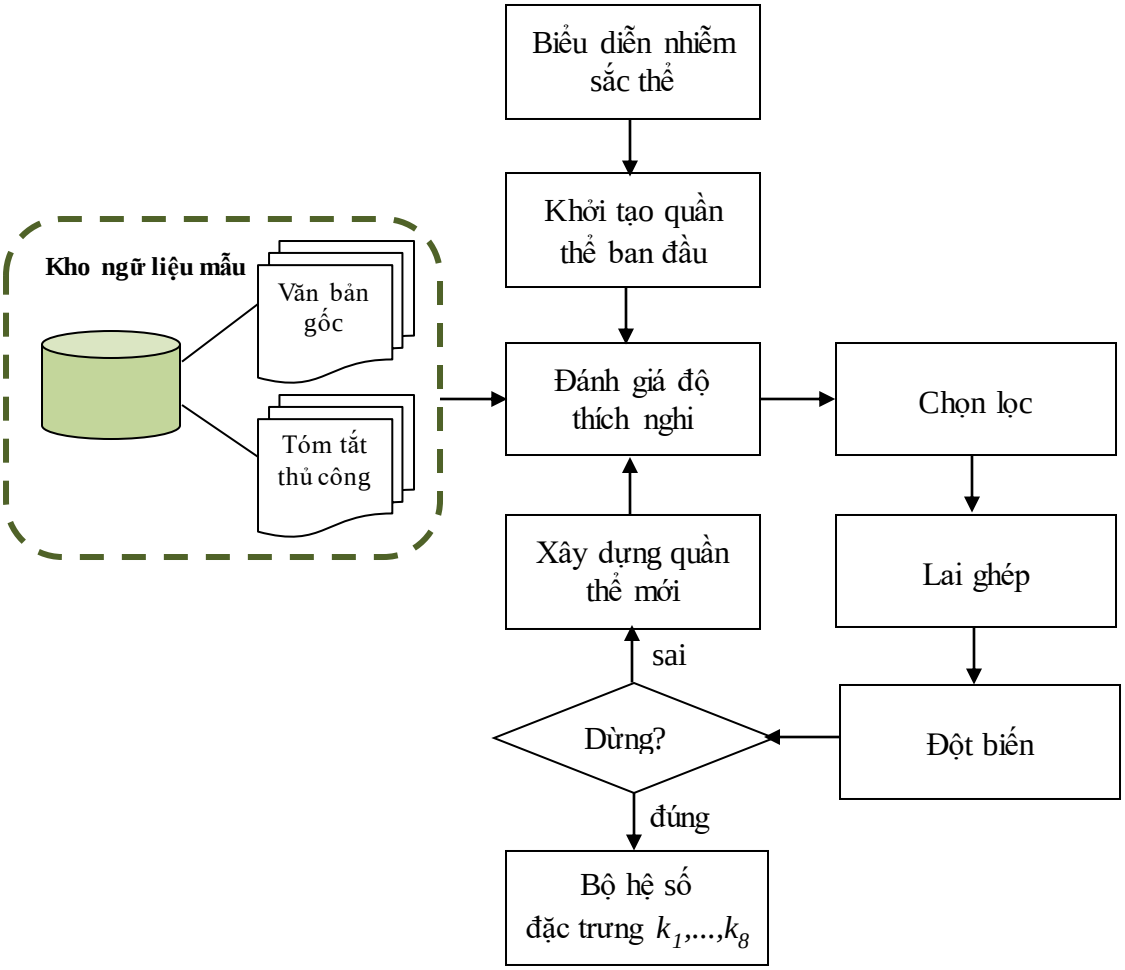
Phần tiếp theo, trình bày mô tả bài toán xác định bộ hệ số đặc trưng bằng giải thuật di truyền và Tối ưu đàn kiến.

2.3.2 Xác định hệ số bằng giải thuật di truyền

Thuật giải di truyền (GA) là thuật giải tìm kiếm ngẫu nhiên dựa trên cơ

chế chọn lọc tự nhiên. GA bao gồm 3 bước quan trọng là: chọn lọc (selection), lai ghép (crossover) và đột biến (mutation). Đây là phương pháp được nhiều nhà nghiên cứu sử dụng trong việc xác định các bộ tham số tối ưu cho nhiều lĩnh vực như hiệu chỉnh tự động các thông số trong mô hình thủy văn, tối ưu tiến độ xây dựng… Trong bài toán tóm tắt văn bản, một số nghiên cứu tóm tắt văn bản tiếng Anh sử dụng GA [72],[67] và cho kết quả khả quan.

Mô hình xác định bộ hệ số đặc trưng bằng phương pháp học máy sử dụng giải thuật di truyền được mô tả trong hình 2-4.



Hình 2-4 Mô hình xác định hệ số đặc trưng bằng thuật toán di truyền

2.3.2.1 Biểu diễn bài toán

Sau đây chúng ta sẽ lần lượt hình thức hóa bài toán xác định hệ số đặc trưng bằng giải thuật di truyền cho bài toán tóm tắt văn bản trên ngôn ngữ của

giải thuật di truyền.

Nhiễm sắc thể. Chúng ta sử dụng nhiễm sắc thể có cấu trúc mã hoá là một vector n chiều $(k_1, k_2 \dots, k_n), k_i \in \mathbb{R}^+$ để biểu diễn các cá thể (các điểm) trong không gian tìm kiếm. Mỗi quần thể là một tập bao gồm một số cố định các cá thể.

Độ đo thích nghi: Với mỗi cá thể $k = (k_1, k_2 \dots, k_n)$ ta xác định mức độ thích nghi của cá thể $F(k)$ bằng công thức sau:

$$F(k) = \sum_{i=1}^{m} \frac{Sim(Sum(a, d_i, f, k), sh_i)}{m} \quad (2.15)$$

Toán tử lai ghép: Giả sử $k^1 = (k_1^1, k_2^1 \dots, k_n^1)$ và $k^2 = (k_1^2, k_2^2 \dots, k_n^2)$ là 2 cá thể bất kỳ trong quần thể. Chúng ta đưa ra một số dạng toán tử lai ghép sau đây:

Toán tử lai ghép một điểm: Giả sử z là một số được lựa chọn ngẫu nhiên, $1 \le z \le n$. Từ hai cá thể cha mẹ là k^1 và k^2 mô tả trên, có thể tạo ra hai cá thể con $k^{1'}$ và $k^{2'}$ với các vector cột tương ứng của chúng được xác định như sau:

$$k_i^{1'} = k_i^1, i = 1, \dots, z;\ k_i^{1'} = k_i^2, i = z+1, \dots, n \quad (2.16)$$

$$k_i^{2'} = k_i^2, i = 1, \dots, z;\ k_i^{2'} = k_i^1, i = z+1, \dots, n \quad (2.17)$$

Có thể biểu diễn toán tử lai ghép một điểm có dạng biểu diễn dưới dạng nhân ma trận như sau:

$$\left(k^{1'}, k^{2'}\right) = (k^1, k^2) \times M$$

trong đó: (k^1, k^2) là vector thuộc $\mathbb{Z}^{2n}$; M là ma trận vuông cấp 2n:

$$M = \begin{bmatrix} A & I - A \\ I - A & A \end{bmatrix}$$

trong đó: A là ma trận chéo cấp n với $a_{ii} = \begin{cases} 1 & khi\ i = 1 \dots z \\ 0 & khi\ i = z+1 \dots n \end{cases}$

Ví dụ:

Cho hai nhiễm sắc thể k^1 và k^2 và điểm lai ghép z:

$k^1 = (1,2,3)$; $k^2 = (4,5,6)$; $z = 2$, ta có:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}; I - A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}; M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\left(k^{1\prime}, k^{2\prime}\right) = (1\ 2\ 3\ 4\ 5\ 6) \times \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} = (1\ 2\ 6\ 4\ 5\ 3)$$

$=> k^{1\prime} = (1,2,6); k^{2\prime} = (4,5,3)$

Toán tử đột biến phân phối đều: Với một gen *i* được chọn ngẫu nhiên để đột biến từ cá thể $k = (k_1, k_2 \ldots, k_n)$, thành phần k_i được thay thế bởi một số ngẫu nhiên α trong khoảng xác định $[0,1]$ của k_i. Cá thể s' sau khi đột biến với các vector cột tương ứng của chúng được xác định như sau:

$$k_j{}' = k_j, j \neq i;\ k_j{}' = \alpha, j = i; j = 1 \ldots n \tag{2.18}$$

Toán tử chọn lọc: Toán tử chọn lọc được xác định theo luật tỷ lệ thuận với mức độ thích nghi:

$$p_s = \frac{F(k)}{\sum_{s \in G} F(k)} \tag{2.19}$$

trong đó: s là cá thể và *G* là quần thể đang xem xét có chứa s.

2.3.2.2 Xây dựng thuật toán

THUẬT TOÁN DI TRUYỀN XÁC ĐỊNH HỆ SỐ ĐẶC TRƯNG

Input: m, a, D, SH, f, G^{max}

trong đó: m là số văn bản huấn luyện; *n* là số đặc trưng; *a* là tỷ lệ tóm tắt; *D* là tập văn bản gốc; SH là tập bản tóm tắt thủ công tương ứng của tập văn bản *D*; f là tập đặc trưng; G^{max} là số thế hệ.

Output: Nghiệm tối ưu của bài toán $DFC(m, a, D, SH, f)$ là vector hệ số đặc trưng $k = (k_1, k_2 \ldots, k_n)$.

Bước 0. Khởi tạo quần thể gồm y cá thể $G_0 = \left(k^{1\,0}, \dots, k^{y\,0}\right)$, trong đó: $k^{i\,0} = \left(k_1^{i\,0}, \dots, k_n^{i\,0}\right); i = 1 \dots y$

Bước 1. Giải các bài toán $Sum\left(a, d_i, f, k^{j\,t}\right), i = 1, \dots, m, j = 1, \dots, y$, t là số thế hệ thứ t của quần thể. Tính mức độ thích nghi $F\left(k^{j\,t}\right), i = 1, \dots, y$ cho từng cá thể của G_t theo (2.15). Áp dụng toán tử chọn lọc theo công thức (2.19) lên G_t để chọn ra y cá thể có mức độ thích nghi lớn nhất.

Bước 2. Nếu $t < G^{max}$ thì chạy tiếp đến Bước 3. Ngược lại thuật toán dừng và cho nghiệm tối ưu là bộ hệ số đặc trưng tối ưu k có mức độ thích nghi lớn nhất trong y cá thể, nghĩa là thỏa mãn $F(\mathrm{k}) = max(F(k^{j\,t}), j = 1, \dots, y)$.

Bước 3. Lựa chọn các cha-mẹ trong G_t theo mức độ thích nghi để ghép cặp theo toán tử lai ghép một điểm (2.16) (2.17) để tạo nên tập các hậu thế G_t^{lg} với y_1 phần tử.

Bước 4. Tác động toán tử đột biến phân phối đều (2.18) vào $G_t \cup G_t^{\mathrm{lg}}$ để nhận được G_{t+1} đặt $t = t + 1$ và quay lại bước 1.

Thuật toán di truyền được biểu diễn dưới dạng giả mã (pseudocode):

Algorithm: $GA(y, \chi, \mu, G^{max})$
//0. Khởi tạo quần thể ban đầu:
$t \leftarrow 0$;
$G_\mathrm{t} \leftarrow \left(k^{1\,t}, \dots, k^{y\,t}\right)$, trong đó: $k^{i\,t} = \left(k_1^{i\,t}, \dots, k_n^{i\,t}\right); i = 1 \dots y$;
While ($t < G^{max}$) **do**
{**//Tạo thế hệ t+1:**
// 1. Tính độ thích nghi:
$fitness(k^{i\,t}), i \in G_\mathrm{t}$
// Lựa chọn y cá thể có độ thích nghi cao nhất:
$G_\mathrm{t} \leftarrow select(G_\mathrm{t})$;
// 2. Lai ghép:

$\boldsymbol{G_t^{lg} \leftarrow crossover(G_t, \chi);}$

//3. Đột biến:

$\boldsymbol{G_{t+1} \leftarrow mutate(G_t \cup G_t^{lg}, \mu);}$

// Tăng số thế hệ:

$t = t + 1;$

}

Return bộ hệ số đặc trưng $\boldsymbol{k}$; $\boldsymbol{fitness}(\mathbf{k}) = \boldsymbol{max}(\boldsymbol{fitness}(\boldsymbol{k^{j^t}}), \boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{y})$

Hình 2-5 Thuật toán xác định hệ số đặc trưng bằng thuật toán di truyền

trong đó: y là số cá thể trong quần thể; χ là xác suất lai ghép; μ là xác suất đột biến; G^{max} số thế hệ (điều kiện dừng).

Thuật toán tính độ thích nghi của cá thể k (bộ hệ số đặc trưng) theo tập văn bản huấn luyện được mô tả như sau:

Algorithm: $fitness(k)$

//Tập văn bản huấn luyện

$D \leftarrow$; //tập văn bản gốc

$SH \leftarrow$; // tập tóm tắt thủ công

$i \leftarrow 0;$

$gt \leftarrow 0;$

$a \leftarrow 0;$ //tỷ lệ tóm tắt

While ($d_i \in D$ **) do**

{// Tính độ tương tự:

$gt \leftarrow gt + Sim(Sum(a, d_i, k), sh_i);$

$i \leftarrow i + 1;$

}

Return $gt \leftarrow {}^{gt}/_{i}$

Hình 2-6 Thuật toán tính độ thích nghi của cá thể

Thuật toán tóm tắt văn bản theo hệ số đặc trưng k được mô tả như sau:

Algorithm: $Sum(a, d, k)$
$S \leftarrow D$; //tập câu văn bản
$score \leftarrow \emptyset$; tập giá trị của câu văn bản
$S_{SUM} \leftarrow \emptyset$; //tập câu văn bản tóm tắt
$f \leftarrow$; // tập giá trị đặc trưng của từng câu văn bản tương ứng
$i \leftarrow 0$;
While ($s_i \in S$ **) do**
{// **Tính giá trị của từng câu theo hệ số k:**
$score[i] \leftarrow$ tính giá trị của câu theo hệ số k và giá trị đặc trưng f tương ứng;
$i \leftarrow i + 1$;
}
$score[].Sort$ // sắp xếp giá trị trọng số câu từ cao xuống thấp
$S_{SUM} \leftarrow$ lấy số câu tóm tắt có trọng số cao theo tỉ lệ tóm tắt a;
Return S_{SUM}

Hình 2-7 Thuật toán tóm tắt văn bản theo hệ số đặc trưng

trong đó: văn bản đã được tiền xử lý tách câu, tách từ, loại bỏ hư từ và tính toán giá trị tập đặc trưng f cho mỗi câu trong văn bản.

Thuật toán tính độ tương đồng giữa bản tóm tắt hệ thống và bản tóm tắt thủ công được mô tả như sau:

Algorithm: $Sim(S_{SUM}, sh)$
$tuloaiTTHT \leftarrow S_{SUM}$; //tập từ loại tóm tắt hệ thống
$tuloaiTTTC \leftarrow sh$; //tập từ loại tóm tắt thủ công
$dongxuathien \leftarrow 0$;
for $i = 0$ **to** $tuloaiTTTC.Count$
{//**Tính số từ đồng xuất hiện giữa 2 văn bản**
$timthay = tuloaiTTHT.IndexOf(tuloaiTTTC[i])$

```
            if (timthay >= 0)
            {
               dongxuathien ← dongxuathien + 1;
            }
         }
Return dongxuathien / tuloaiTTTC.Count
```

Hình 2-8 Thuật toán tính độ tương đồng giữa bản tóm tắt hệ thống và bản tóm tắt thủ công

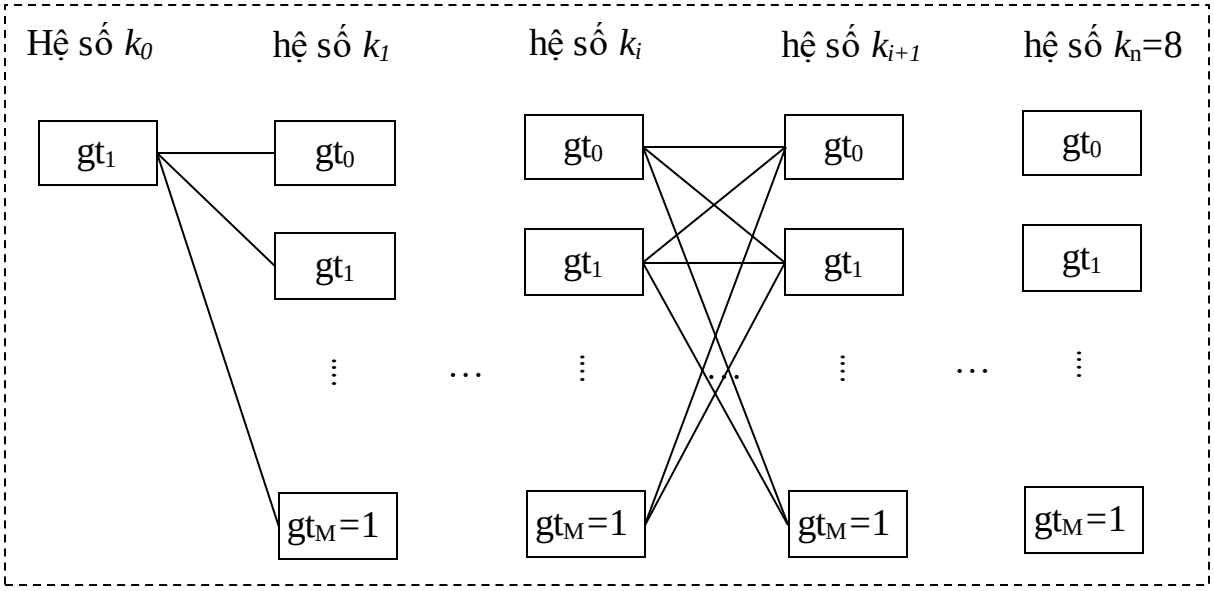
2.3.2.3 Đánh giá độ phức tạp thuật toán

Độ phức tạp của thuật toán di truyền xác định hệ số đặc trưng trình bày ở trên là $O(G^{max} \times y \times n \times m^2)$. Trong đó G^{max} là số thế hệ tiến hoá, độ lớn của G^{max} tuỳ thuộc vào từng bài toán cụ thể, thường là G^{max} có thể lớn đến hàng nghìn; y là kích thước quần thể - số cá thể trong quần thể (thông thường chỉ đến vài chục cá thể); m là số văn bản huấn luyện, m^2 là thời gian thực hiện tính hàm thích nghi; n là chiều dài nhiễm sắc thể (được tính bằng số đặc trưng), n và y thường rất nhỏ (coi như hằng số), do đó độ phức tạp của thuật giải chỉ là $O(G^{max} \times m^2)$ cho một lần tìm kiếm, chỉ tương đương hoặc nhỏ hơn độ phức tạp $O(N \times m)$ với N là số rất lớn. Trong nghiên cứu này ta dùng giải thuật di truyền để giải bài toán xác định hệ số đặc trưng sẽ đáp ứng được tốt yêu cầu về thời gian.

2.3.3 Xác định hệ số bằng giải thuật tối ưu đàn kiến

Như đã trình bày ở mục 1.3, giải thuật tối ưu đàn kiến ACO là một phương pháp nghiên cứu lấy cảm hứng từ việc mô phỏng hành vi của đàn kiến trong tự nhiên nhằm mục tiêu giải quyết các bài toán tối ưu phức tạp. Qua thử nghiệm giải thuật tối ưu đang kiến cho bài toán tách từ tiếng Việt và nhận thấy có hiệu quả [CT5]. Do vậy, trong phần này luận án trình bày nghiên cứu phương pháp xác định hệ số đặc trưng bằng giải thuật tối ưu đàn kiến.

Thực hiện chuyển đổi bài toán tìm hệ số đặc trưng tối ưu đã được mô tả ở mục 2.3.1 thành bài toán tối ưu tổ hợp và tìm lời giải tối ưu dựa trên thuật toán tối ưu đàn kiến. Hình 2-5 biểu diễn bài toán tối ưu tổ hợp tìm kiếm bộ hệ số đặc trưng tối ưu với bước chia $h = {}^{1}/_{M}$:



Hình 2-9 Biểu diễn bài toán xác định hệ số đặc trưng dưới dạng bài toán tối ưu tổ hợp với bước chia h=1/M

Mỗi nút trong hình 2-5 biểu thị một giá trị hệ số gt_i được lựa chọn cho hệ số đặc trưng k_i. Ví dụ, nút thứ j trên cột i $(i = 1,2,\dots,n)$ cho biết rằng hệ số k_i của đặc trưng thứ i được chọn giá trị bằng gt_j (giá trị ở ô thứ j). Cột 0 là một hệ số ảo đại diện cho điểm bắt đầu. Các cạnh trên hình 2-5 được mô tả bởi một ma trận với 3 yếu tố, ví dụ (i, j_1, j_2) miêu tả hệ số đặc trưng thứ i được lựa chọn giá trị bằng giá trị ở ô thứ j_1, trong khi hệ số đặc trưng thứ $j + 1$ được lựa chọn giá trị bằng giá trị ở ô thứ j_2. Mỗi đường đi là $d = (k_0, k_1 j_1, k_2 j_2, \dots, k_n j_n)$ từ cột 0 qua các đỉnh từ cột 1 đến cột n thể hiện một bộ hệ số $k = (gt_{j1}, gt_{j2} \dots, gt_{jn})$ tương ứng của bài toán. Năng lượng tiêu phí trên đường đi được tính theo công thức (2.20):

$$F(d) = F(k) = \sum_{i=1}^{m} \frac{m}{Sim(Sum(a, d_i, f, k), sh_i)} \quad (2.20)$$

Do vậy, việc giải quyết bài toán xác định hệ số đặc trưng là tập trung tìm kiếm một đường đi có thể làm cho cực tiểu hàm mục tiêu.

2.3.3.1 Phát biểu bài toán:

Bài toán xác định bộ hệ số đặc trưng bằng giải thuật tối ưu đàn kiến (sau đây được gọi $F\_ACO(m, a, D, SH, f)$ được phát biểu như sau:

Giả sử $k = (k_1, k_2 \ldots, k_n)$ là bộ hệ số đặc trưng chấp nhận được. Tìm k sao cho hàm mục tiêu:

$$F(d) = F(k) \Longrightarrow \text{Giá trị cực tiểu} \quad (2.21)$$

trong đó: d là đường đi từ cột k_0 đến k_n.

2.3.3.2 Xây dựng thuật toán

Quan trọng nhất của giải thuật đàn kiến là phương pháp cập nhật mùi trên đường đi của kiến sau mỗi vòng lặp và hệ thông tin heuristic (tầm nhìn) của kiến đến đỉnh tiếp theo. Ở đây để đạt sự hội tụ và kết quả tốt, chúng ta cải tiến thuật toán ACO gốc bằng cách sử dụng công thức cập nhật mùi Mi-max trơn (Smoothed Max Min Ant System – SMMAS) đã được đề xuất và chứng minh tính hội tụ tốt trong trong [3] và xây dựng công thức heuristic (2.28) phù hợp với bài toán xác định hệ số đặc trưng.

THUẬT TOÁN TỐI ƯU ĐÀN KIẾN XÁC ĐỊNH HỆ SỐ ĐẶC TRƯNG

Input: m, a, D, SH, f, G^{max}

trong đó: m là số văn bản huấn luyện; n là số đặc trưng; a là tỷ lệ tóm tắt; D là tập văn bản gốc; SH là tập bản tóm tắt thủ công tương ứng của tập văn bản D; f là tập đặc trưng; G^{max} là số thế hệ.

Output: Nghiệm tối ưu của bài toán $F\_ACO(m, a, D, SH, f)$ là vector hệ số đặc trưng $k = (k_1, k_2 \ldots, k_n)$

Bước 1: Khởi tạo các đáp án ban đầu:

Trước tiên, tất cả các con kiến nhân tạo được đặt ở nút khởi đầu. Tiếp theo, tạo ra một cách ngẫu nhiên một đường đi từ nút khởi đầu đến nút kết thúc cho mỗi con kiến. Điều này có nghĩa là mỗi con kiến sẽ chọn lựa một cách ngẫu nhiên một giá trị hệ số cho mỗi đặc trưng để tạo ra một đáp án khả thi cho bài toán.

Bước 2: Tính toán hàm mục tiêu theo công thức (2.20). Giá trị này được sử dụng để chọn ra phương án tối ưu là bộ hệ số đặc trưng *k* trong mỗi lần thử.

Bước 3: Thiết lập vùng đáp án (solution pool) đặt tên là E:

Mục đích của việc thiết lập vùng đáp án là làm giảm việc tính toán lặp lại một cách không cần thiết trong suốt quá trình chạy thuật toán. Khi tạo ra một đáp án mới, trước tiên sẽ tìm kiếm trong vùng đáp án. Nếu đáp án này đã xuất hiện trong vùng đáp án, thì loại bỏ nó, nếu không thì tính toán giá trị hàm mục tiêu theo công thức (2.20).

Bước 4: Tính toán giá trị cập nhật của vệt mùi trên mỗi đường đi sau một vòng lặp:

Sử dụng phương pháp Max-Min trơn [3] cập nhật mùi sau mỗi vòng lặp, phương pháp này đảm bảo vết mùi ở các cạnh không bị giảm quá nhanh, dẫn đến các cạnh tốt trong một vài trường hợp bị loại bỏ sớm. Do vậy, kết quả tìm kiếm không hội tụ về phương án tối ưu. Quy tắc SMMAS tính giá trị cập nhật của vệt mùi trên mỗi cạnh (i, j_1, j_2) sau mỗi vòng lặp theo công thức sau:

$$\Delta\tau_{i,j_1,j_2} = \begin{cases} \rho\tau_{max} \; nếu \; (i, j_1, j_2) \in w(t) \\ \rho\tau_{min} \; nếu \; (i, j_1, j_2) \notin w(t) \end{cases} \tag{2.22}$$

trong đó:

$\Delta\tau_{i,j_1,j_2}$: giá trị cập nhật của vệt mùi trên cạnh (i, j_1, j_2) sau một vòng lặp;

ρ: tham số đặc trưng cho việc bay hơi (trong thực nghiệm $\rho = 0.05$);

τ_{max}, τ_{min}: Các tham số đặc trưng cho hành vi của kiến. Khi τ_{min} nhỏ hơn nhiều so với τ_{max}, tính khám phá sẽ kém, còn nếu chọn τ_{min} gần với

τ_{max} thì thuật toán chủ yếu là tìm kiếm ngẫu nhiên dựa theo thông tin heuristic. Trong thực nghiệm, chọn $\tau_{max} = 1.0$ và $\tau_{min} = 0{,}01$;

$w(t)$: hành trình tối ưu của đàn kiến trong mỗi lần thử.

Bước 5: Cập nhật vệt mùi trên mỗi cạnh:

Cuối mỗi vòng lặp, cường độ của vệt mùi trên mỗi cạnh được cập nhật lại theo quy tắc sau:

$$\emptyset\tau_{i,j1,j2}(nc+1) = (1-\rho) \times \emptyset\tau_{i,j1,j2}(nc) + \Delta\tau_{i,j_1,j_2} \quad (2.23)$$

trong đó:

$\emptyset\tau_{i,j1,j2}(nc)$: vệt mùi trên cạnh (i, j_1, j_2) sau vòng lặp nc

$\emptyset\tau_{i,j1,j2}(nc+1)$: vệt mùi trên cạnh (i, j_1, j_2) sau vòng lặp $nc+1$

$\rho \in [0,1]$: là hằng số, đặc trưng cho tỷ lệ tồn tại của vệt mùi trước đó.

$\Delta\tau_{i,j_1,j_2}$: giá trị cập nhật vệt mùi theo công thức (2.22)

Bước 6: Tính toán xác suất lựa chọn đường đi trên mỗi cạnh của các con kiến:

Kiến lựa chọn đường đi dựa trên cường độ mùi và tầm nhìn của mỗi cạnh. Do đó, xác suất lựa chọn cho mỗi cạnh được tính theo công thức (2.24):

$$\begin{cases} p^z_{i,j1,j2} = \dfrac{\left[\tau_{i,j1,j2}\right]^{\alpha}\left[\eta_{i,j1,j2}\right]^{\beta}}{\sum_{u \in J_z(i)}\left[\tau_{i,j1,u}\right]^{\alpha}\left[\eta_{i,j1,u}\right]^{\beta}}, nếu\ j \in J_z(i) \\ Ngược\ lại\ p^k_{i,j1,j2} = 0 \end{cases} \quad (2.24)$$

trong đó:

$p^z_{i,j1,j2}$: xác suất để con kiến z lựa chọn cạnh (i, j_1, j_2) để đi;

α: thông số điều chỉnh ảnh hưởng của vệt mùi $\emptyset\tau_{i,j1,j2}$;

β: thông số điều chỉnh ảnh hưởng của $\eta_{i,j1,j2}$;

$J_z(i)$: tập hợp các nút mà con kiến z ở nút i chưa đi qua;

$\tau_{i,j1,j2}$: nồng độ của vệt mùi trên cạnh $\emptyset\tau_{i,j1,j2}$;

$\eta_{i,j1,j2}$: thông tin heuristic (hay gọi là tầm nhìn) giúp đánh giá chính xác sự lựa chọn của con kiến khi quyết định đi trên cạnh (i, j_1, j_2), tượng trưng cho thông tin cục bộ xem xét trong quá trình; được xác định theo công thức:

$$\eta_{ij} = \frac{dc_{i+1}^{max} - dc_{i+1}^{(z)} + \gamma}{dc_{i+1}^{max} - dc_{i+1}^{min} + \gamma} \quad (2.25)$$

trong đó:

dc_{i+1}^{max}: giá trị hàm mục tiêu cực đại được tính với bộ hệ số đặc trưng có giá trị hệ số đặc trưng i+1 theo những lựa chọn khác nhau;

dc_{i+1}^{min}: giá trị hàm mục tiêu cực tiểu được tính với bộ hệ số đặc trưng có giá trị hệ số đặc trưng i+1 theo những lựa chọn khác nhau;

$dc_{i+1}^{(z)}$: giá trị hàm mục tiêu được tính tính với bộ hệ số đặc trưng có giá trị hệ số đặc trưng i+1 theo lựa chọn thứ z;

γ: là một hằng số cho trước trong đoạn (0,1)

Bước 7: Lựa chọn đường đi cho mỗi con kiến:

Đề lựa chọn một giá trị hệ số đặc trưng, con kiến sẽ sử dụng thông tin heuristic biểu thị bởi $\eta_{i,j1,j2}$ cũng như là thông tin về vệt mùi biểu thị bởi $\emptyset\tau_{i,j1,j2}$. Quy tắc lựa chọn đươc mô tả bởi công thức sau đây:

$$j = \begin{cases} \arg_{u \in J_z(i)} \underset{J}{max} \left[\left(\tau_{i,j1,u}\right)^{\alpha} \times \left(\eta_{,j1,u}\right)^{\beta} \right] & nếu\ q \le q_0 \\ ngược\ lại & \end{cases} \quad (2.26)$$

q: giá trị được lựa chọn một cách ngẫu nhiên với một xác suất không thay đổi trong khoảng [0,l];

q_0: là một hằng số cho trước trong khoảng [0,1];

J: là một biến số ngẫu nhiên được lựa chọn theo sự phân bố xác suất cho bởi quy luật phân bố xác suất theo công thức (2.24).

Bước 8: Thêm đáp án mới từ quá trình vào vùng đáp án E. Lặp lại quá trình từ Bước 4 đến Bước 8 cho đến khi điều kiện kết thúc được thỏa mãn. Ở

đây điều kiện dừng là đạt đến số bước lặp cho trước G^{max}. Khi đó nghiệm tối ưu của bài toán chính là bộ hệ số k.

Thuật toán tối ưu đàn kiến được biểu diễn dưới dạng giả mã (pseudocode):

Algorithm: $\boldsymbol{ACO}(nAnts, \alpha, \beta, G^{max})$
// 1. Khởi tạo tập đường đi của kiến ban đầu:
$t \leftarrow 0$;
$G_t \leftarrow \left(k^{1^t}, \dots, k^{nAnts^t}\right)$, trong đó: $k^{i^t} = \left(k_1^{i^t}, \dots, k_n^{i^t}\right); i = 1 \dots nAnts$ là bộ hệ số đặc trưng tương ứng với đường đi ngẫu nhiên đầu tiên của kiến từ cột 0 đến cột n ($duongdi = (k_0^t, k_0^t, \dots., k_n^t)$)
//2. Tính hàm mục tiêu
$fitness(k^{i^t}), i \in G_t$
// 3. Lập vùng đáp án
$E \leftarrow G_t$;
While ($t < G^{max}$ **) do**
{
// 4. Tính toán giá trị cập nhật mùi:
$Max - min(k^{i^t}), i \in E$
// 5. Cập nhật mùi:
$UpdateTrails()$;
// 6. Tính toán xác suất lựa chọn đường đi:
$SelectPro(\alpha, \beta)$;
//7. Lựa chọn đường đi:
$bestlength \leftarrow max(fitness(k^{i^t})), i \in E$;
// Thêm đáp án mới:
If $bestlength \notin E$ **then** $E \leftarrow E + bestlength$;
$t = t + 1$;
}
Return bộ hệ số đặc trưng $\boldsymbol{k}$; $\boldsymbol{fitness}(k) = \min(fitness(k^i)), k^i \in E$

Hình 2-10 Thuật toán xác định hệ số đặc trưng bằng giải thuật ACO

trong đó: $nAnts$ là số kiến được thả; α là thông số điều chỉnh ảnh hưởng của vệt mùi; μ là thông số điều chỉnh ảnh hưởng của heuristic (hay gọi là tầm nhìn); G^{max} số vòng lặp (điều kiện dừng)

2.3.3.3 Đánh giá độ phức tạp thuật toán

Độ phức tạp của thuật toán tối ưu đàn kiến đã được trình bày chi tiết trong tài liệu [28]. Với bài toán này độ phức tạp được xác định là $O(G^{max} \times n^3)$, trong đó n là số đặc trưng (trong bài toán này n = 8).

2.4 Các kết quả thử nghiệm

2.4.1 Kho ngữ liệu thử nghiệm

Sử dụng 2 kho ngữ liệu Corpus_LTH và ViEvTextSum (được trình bày trong phần phụ lục). Đặc điểm của 2 kho ngữ liệu này như sau:

- Bản tóm tắt thủ công của kho ngữ liệu Corpus_LTH được xây dựng trên quan điểm trích chọn những câu quan trọng trong văn bản, sau đó rút gọn câu bằng cách bỏ những phần không quan trọng trong câu. Tạo bản tóm tắt thủ công cuối cùng với độ dài khoảng 120 từ.

- Bản tóm tắt thủ công của kho ngữ liệu ViEvTextSum được xây dựng trên quan điểm: tác giả đọc hiểu toàn bộ văn bản và viết lại bản tóm tắt theo quan điểm của tác giả với độ dài xấp xỉ 120 từ.

Để làm chính xác kết quả ở mỗi bước thử nghiệm, thực hiện 5 lần lấy 80% văn bản mẫu ngẫu nhiên để làm văn bản huấn luyện. Bộ hệ số thu được chính là bộ hệ số trung bình của 5 lần thực hiện đó. Sau khi có bộ hệ số, thực hiện tóm tắt 5 lần trên 20% văn bản ngẫu nhiên còn lại và thu được độ đo ROUGE-N trung bình của 5 lần tóm tắt.

2.4.2 Phương pháp đánh giá kết quả tóm tắt

Sử dụng phương pháp đánh giá ROUGE-N đã được đề cập trong mục [1.1.2.3]. Phương pháp này đánh giá chất lượng của một bản tóm tắt dựa trên độ đo đồng xuất hiện n-gram từ vựng giữa văn bản tóm tắt do hệ thống tạo ra và và văn bản tóm tắt do con người thực hiện. Độ đo ROUGE-N được tính theo

công thức (2.27):

$$ROUGE - N = \frac{|SH_{n-gram} \cap SM_{n-gram}|}{|SH_{n-gram}|} \quad (2.27)$$

trong đó: $SM_{n-gram} = \{sm_1, \dots, sm_r\}$ là vector n-gram từ vựng khác nhau của văn bản tóm tắt của hệ thống; $SH_{n-gram} = \{sh_1, \dots, sh_l\}$ là vector từ vựng khác nhau của văn bản tóm tắt do con người thực hiện.

Độ đo trung bình của toàn bộ kho ngữ liệu tóm tắt bằng độ đo ROUGE-N được tính theo công thức (2.28):

$$ROUGE - N_{Avg}(D) = \frac{1}{m}\sum_{i=1}^{m} ROUGE - N(d_i) \quad (2.28)$$

trong đó: *D* là tập văn bản tóm tắt; d_i văn bản tóm tắt thứ *i* của tập văn bản tóm tắt *D; m* là số văn bản của tập văn bản tóm tắt *D*.

Nhận xét: Với tập văn bản tóm tắt lớn, giá trị $ROUGE - N_{Avg}(D)$ của các phương pháp chênh lệch nhau 0.01 (1%) thì có thể xem là kết quả chênh lệch đáng kể để đánh giá độ chính xác của từng phương pháp.

2.4.3 Các kết quả thử nghiệm

2.4.3.1 Thử nghiệm đánh giá vai trò của từng đặc trưng

Trước hết, cần phải nghiên cứu ảnh hưởng của mỗi đặc trưng văn bản được sử dụng trong mô hình tóm tắt đã trình bày ở trên. Chúng ta thực hiện tính toán những ảnh hưởng này bằng cách sử dụng công thức tính trọng số câu (2.1) với chỉ một hệ số k_i bằng 1, các hệ số còn lại bằng 0. Công thức tính trọng số câu của đặc trưng thứ *i* được viết lại như sau:

$$Score(s) = Score_{f_i}(s) \quad (2.29)$$

Mục đích của việc nghiên cứu này là xem ảnh hưởng của từng đặc trưng trong 8 đặc trưng đã chọn ở trên đến hệ thống tóm tắt văn bản như thế nào, qua đó có thể đánh giá các đặc trưng được cải tiến (vị trí câu, độ dài câu) ảnh hưởng đến hiệu quả tóm tắt như thế nào.

Bảng 2-6; 2-7 cho thấy độ chính xác trung bình thu được bằng cách sử dụng từng đặc trưng văn bản để tóm tắt các tài liệu trong kho ngữ liệu Corpus_LTH và ViExTextSum.

Bảng 2-6. Kết quả tóm tắt từng đặc trưng trên kho ngữ liệu Corpus_LTH

Đặc trưng	ROUGE-N			
	N=1	N=2	N=3	N=4
F1 - Vị trí câu	0.584	0.402	0.341	0.317
F1a - Câu đầu	0.527	0.312	0.245	0.210
F1b - Câu đầu và câu cuối	0.564	0.371	0.333	0.302
F2 - Trọng số TF.ISF	0.512	0.284	0.227	0.208
F3 - Độ dài câu	0.365	0.188	0.141	0.126
F4 - Xác suất thực từ	0.501	0.347	0.298	0.271
F5 - Danh từ riêng	0.513	0.321	0.272	0.248
F6 - Dữ liệu số	0.492	0.301	0.257	0.233
F7 - Độ tương đồng giữa câu với tiêu đề	0.564	0.412	0.361	0.336
F8 - Câu trung tâm	0.592	0.435	0.391	0.354

Qua 2 bảng kết quả 2-6 và 2-7, chúng ta có thể nhận thấy đặc trưng vị trí câu, độ tương tự tiêu đề và câu trung tâm cho kết quả tốt nhất. Riêng đặc trưng vị trí câu, do đã khảo sát kỹ kho ngữ liệu tóm tắt mẫu tiếng Việt, do vậy sự cải tiến công thức tính vị trí câu theo phân bố đã phát huy hiệu quả hơn là những phương pháp tính vị trí câu trước đây là định nghĩa câu đầu, hoặc câu đầu và câu cuối là quan trọng nhất. Đặc trưng tương tự tiêu đề cho kết quả cao, nghĩa là đánh giá cao những câu sát với chủ đề của câu tiêu đề đưa ra. Đối với đặc trưng câu trung tâm, chúng ta cũng có thể dự đoán được vì đặc trưng này đánh giá cao những câu đề cập đến nhiều chủ đề xuất hiện trên khắp văn bản hơn những câu chỉ đề cập đến một chủ đề. Ngược lại, đặc trưng độ dài câu cho kết quả kém nhất chứng tỏ đặc trưng này không ảnh hưởng nhiều đến kết quả tóm

tắt. Điều đó cũng dễ hiểu bởi vì độ dài câu không phản ánh mức độ ngữ nghĩa hoặc bố cục của văn bản.

Bảng 2-7. Kết quả tóm tắt từng đặc trưng trên kho ngữ liệu ViEvTextSum

Đặc trưng	ROUGE-N			
	N=1	N=2	N=3	N=4
F1 - Vị trí câu	0.401	0.122	0.076	0.043
F1a - Câu đầu	0.356	0.091	0.043	0.021
F1b - Câu đầu và câu cuối	0.381	0.105	0.056	0.037
F2 - Trọng số TF.ISF	0.393	0.112	0.063	0.038
F3 - Độ dài câu	0.295	0.071	0.022	0.009
F4 - Xác suất thực từ	0.352	0.093	0.042	0.021
F5 - Danh từ riêng	0.364	0.097	0.045	0.026
F6 - Dữ liệu số	0.347	0.089	0.038	0.020
F7 - Độ tương đồng giữa câu với tiêu đề	0.406	0.124	0.079	0.049
F8 - Câu trung tâm	0.418	0.133	0.081	0.053

Tuy nhiên, để đánh giá chính xác ảnh hưởng của từng đặc trưng văn chúng ta cần đánh giá vai trò của các đặc trưng trên mô hình kết hợp sẽ được trinh bày ở phần tiếp theo.

2.4.3.2 Kết quả thử nghiệm của mô hình VTS_FC dựa trên giải thuật di truyền (VTS_FC_GA)

Trong phần này, chúng ta sẽ xem xét kết quả tóm tắt của mô hình ***VTS_FC_GA*** khi sử dụng kết hợp các đặc trưng văn bản được lựa chọn, trong đó tập trung xem xét kết quả của mô hình ***VTS_FC_GA*** kết hợp của các đặc trưng mà các nghiên cứu trước đây về tóm tắt văn bản thường sử dụng và so sánh với kết quả của mô hình ***VTS_FC_GA*** kết hợp tất cả 8 đặc trưng đã lựa chọn ở trên.

Dựa vào công thức (2.1) để tính trọng số câu và lựa chọn ra những câu có

điểm số cao tạo thành bản tóm tắt theo tỉ lệ người dùng mong muốn. Trong đó, bộ hệ số đặc trưng được xác định từ kết quả của quá trình huấn luyện. Trong quá trình huấn luyện, giải thuật di truyền sẽ được thực hiện với các thông số:

- Có 100 cá thể trong một quần thể;
- Xác suất lai ghép 0.8;
- Xác suất đột biến 0.1;
- Thuật toán dừng khi đạt được 1000 thế hệ;
- Tỷ lệ tóm tắt là 30%.

Bộ tham số này được xác định bằng phương pháp thử nghiệm. Đầu tiên chúng ta dựa vào bộ hệ số thông thường được đề xuất cho giải thuật di truyền gốc. Sau độ bộ tham số này được điều chỉnh trong quá trình thử nghiệm thông qua việc thay đổi các giá trị và đánh giá sự hội tụ của giải thuật thông qua Hàm thích nghi (công thức 2.15).

Sau khi tìm được bộ hệ số đặc trưng tối ưu, thực hiện bước chuẩn hoá bộ hệ số đặc trưng theo điều kiện:

$$\sum_{i=1}^{n} k_i = 1 \qquad (2.30)$$

Thử nghiệm 1: Đánh giá kết quả mô hình ***VTS_FC_GA*** sử dụng kết hợp 5 đặc trưng mà các nghiên cứu tóm tắt văn bản tiếng Việt trước đây đã đề xuất [76],[55] trên 2 kho ngữ liệu Corpus_LTH và ViEvTextSum. Kết quả thử nghiệm được trình bày trong bảng 2-8.

Qua kết quả thử nghiệm, chúng ta có thể thấy kết quả tóm tắt của mô hình ***VTS_FC_GA*** khi kết hợp 5 đặc trưng cao hơn hẳn kết quả tóm tắt theo từng đặc trưng riêng biệt (đã được trình bày ở mục 2.2.3). Ngoài ra, đặc trưng xác suất thực từ, độ tương đồng giữa câu với tiêu đề và vị trí câu đóng vai trò quan trọng hơn 2 đặc trưng còn lại là danh từ riêng và dữ liệu số.

*Bảng 2-8. Kết quả của mô hình **VTS_FC_GA** dựa trên 5 đặc trưng.*

Đặc trưng	**Hệ số**			
F1b - câu đầu và câu cuối	0.23			
F4 - Xác suất thực từ	0.39			
F5 - Danh từ riêng	0.10			
F6 - Dữ liệu số	0.03			
F7 - Độ tương đồng giữa câu với tiêu đề	0.26			
Kết quả tóm tắt (ROUGE-N)	**N=1**	**N=2**	**N=3**	**N=4**
Corpus_LTH	0.620	0.469	0.420	0.387
ViEvTextSum	0.437	0.152	0.082	0.051

Thử nghiệm 2: Đánh giá kết quả mô hình ***VTS_FC_GA*** sử dụng kết hợp 8 đặc trưng đã được lựa chọn ở mục 2.1. trên 2 kho ngữ liệu Corpus_LTH và ViEvTextSum. Kết quả thử nghiệm được trình bày trong bảng 2-9.

*Bảng 2-9. Kết quả của mô hình **VTS_FC_GA** dựa trên 8 đặc trưng.*

Đặc trưng	**Hệ số**			
F1 - Vị trí câu	0.36			
F2 - Trọng số TF.ISF	0.12			
F3 - Độ dài câu	0.02			
F4 - Xác suất thực từ	0.05			
F5 - Danh từ riêng	0.07			
F6 - Dữ liệu số	0.05			
F7 - Độ tương đồng giữa câu với tiêu đề	0.07			
F8 - Câu trung tâm	0.26			
Kết quả tóm tắt (ROUGE-N)	**N=1**	**N=2**	**N=3**	**N=4**
Corpus_LTH	0.654	0.480	0.436	0.401
ViEvTextSum	0.452	0.160	0.083	0.051

Với kết quả tóm tắt này, chúng ta có thể nhận xét rằng khi thêm vào 4 đặc trưng F1-vị trí câu (đã cải tiến), F2- trọng số TFxISF, F3- độ dài câu và F8-câu trung tâm vào thì kết quả tóm tắt của mô hình ***VTS_FC_GA*** cao hơn kết quả của mô hình ***VTS_FC_GA*** sử dụng 5 đặc trưng mà các phương pháp tóm tắt văn bản tiếng Việt trước đây đã đề xuất. Tuy nhiên xét độ ảnh hưởng thì 3 đặc trưng F1-vị trí câu, F2- trọng số TFxISF và F8-câu trung tâm có ảnh hưởng nhiều đến kết quả tóm tắt, còn đặc trưng độ dài câu không đóng vai trò gì nhiều. Mặt khác, xét độ ảnh hưởng cả 8 đặc trưng thì đặc trưng dữ liệu số cũng không đóng vai trò gì nhiều trong kết quả tóm tắt.

2.4.3.3 Kết quả thử nghiệm của mô hình VTS_FC dựa trên giải thuật tối ưu đàn kiến (VTS_FC_ACO)

Trong phần này, chúng ta sẽ xem xét kết quả thử nghiệm của mô hình ***VTS_FC_ACO*** theo các bước giống như thử nghiệm của mô hình ***VTS_FC_GA*** (mục 2.2.4). Trong quá trình huấn luyện, giải thuật tối ưu đàn kiến sẽ được thực hiện với các thông số như trong bảng 2-10.

Bảng 2-10. Lựa chọn các thông số cho thuật toán ACO

Thông số (Parameters)	**Giá trị (Value)**
Số lượng kiến z	40
Số vòng lặp G^{max}	100
Hệ số α	3
Hệ số β	2
Thông số bay hơi ρ	0.05
q_0	0.9
Q	2
Nồng độ mùi ban đầu τ_0	0

Bộ tham số của giải thuật tối ưu đàn kiến được xác định bằng phương pháp thử nghiệm. Đầu tiên chúng ta dựa vào bộ hệ số thông thường được đề xuất cho giải thuật tối ưu đàn kiến gốc [28]. Sau độ bộ tham số này được chọn trong quá

trình thử nghiệm thông qua việc thay đổi các giá trị và đánh giá sự hội tụ của giải thuật thông qua Hàm mục tiêu (công thức 2.20).

Thử nghiệm 1: Đánh giá kết quả thử nghiệm của mô hình ***VTS_FC_ACO*** sử dụng kết hợp 5 đặc trưng mà các nghiên cứu trước đây về tóm tắt văn bản tiếng Việt đã đề xuất [76],[55] trên 2 kho ngữ liệu Corpus_LTH và ViEvTextSum. Kết quả thử nghiệm được trình bày trong bảng 2-11.

*Bảng 2-11. Kết quả thử nghiệm của mô hình **VTS_FC_ACO** dựa trên 5 đặc trưng thường dùng*

Đặc trưng	**Hệ số**			
F1b - câu đầu và câu cuối	0.35			
F4 - Xác suất thực từ	0.26			
F5 - Danh từ riêng	0.07			
F6 - Dữ liệu số	0.02			
F7 - Độ tương đồng giữa câu với tiêu đề	0.30			
Kết quả tóm tắt (ROUGE-N)	**N=1**	**N=2**	**N=3**	**N=4**
Corpus_LTH	0.629	0.476	0.422	0.389
ViEvTextSum	0.439	0.148	0.059	0.045

Qua kết quả thử nghiệm, chúng ta có thể thấy kết quả tóm tắt của mô hình ***VTS_FC_ACO*** khi kết hợp 5 đặc trưng cao hơn kết quả tóm tắt theo từng đặc trưng riêng biệt (đã được trình bày ở mục 2.2.3). Ngoài ra, ta có thể thấy rằng đặc trưng xác suất thực từ, độ tương đồng giữa câu với tiêu đề và vị trí câu đóng vai trò quan trọng hơn 2 đặc trưng còn lại là danh từ riêng và dữ liệu số.

Thử nghiệm 2: Đánh giá kết quả mô hình ***VTS_FC_ACO*** sử dụng kết hợp 8 đặc trưng đã được lựa chọn ở mục 2.1. trên 2 kho ngữ liệu Corpus_LTH và ViEvTextSum. Kết quả thử nghiệm được trình bày trong bảng 2-12.

Với kết quả thử nghiệm này, chúng ta có thể nhận xét rằng cũng như mô hình ***VTS_FC_GA,*** khi thêm vào 4 đặc trưng F1-vị trí câu (đã cải tiến), F2-

trọng số TFxISF, F3- độ dài câu và F8-câu trung tâm vào thì kết quả tóm tắt của mô hình ***VTS_FC_ACO*** cao hơn kết quả của mô hình ***VTS_FC_ACO*** sử dụng 5 đặc trưng mà các phương pháp tóm tắt văn bản tiếng Việt trước đây đã đề xuất. Tuy nhiên xét độ ảnh hưởng thì 3 đặc trưng F1-vị trí câu, F2- trọng số TFxISF và F8-câu trung tâm có ảnh hưởng nhiều đến kết quả tóm tắt, còn đặc trưng vị trí câu không đóng vai trò gì nhiều. Mặt khác, xét độ ảnh hưởng cả 8 đặc trưng thì đặc trưng dữ liệu số cũng không đóng vai trò gì nhiều trong kết quả tóm tắt.

*Bảng 2-12. Kết quả tóm tắt của mô hình **VTS_FC_ACO** dựa trên 8 đặc trưng.*

Đặc trưng	**Hệ số**			
F1 - Vị trí câu	0.32			
F2 - Trọng số TF.ISF	0.13			
F3 - Độ dài câu	0.02			
F4 - Xác suất thực từ	0.09			
F5 - Danh từ riêng	0.06			
F6 - Dữ liệu số	0.02			
F7 - Độ tương đồng giữa câu với tiêu đề	0.11			
F8 - Câu trung tâm	0.26			
Kết quả tóm tắt (ROUGE-N)	**N=1**	**N=2**	**N=3**	**N=4**
Corpus_LTH	0.665	0.500	0.445	0.408
ViEvTextSum	0.464	0.167	0.088	0.058

Mặt khác, qua kết quả thử nghiệm, chúng ta có thể thấy kết quả tóm tắt của mô hình ***VTS_FC_ACO*** cao hơn kết quả tóm tắt của mô hình ***VTS_FC_GA*** trong cả 2 thử nghiệm. Chứng tỏ trong bài toán xác định bộ hệ số đặc trưng thì giải thuật tối ưu đàn kiến hiệu quả hơn giải thuật di truyền.

2.4.3.4 Kết quả thử nghiệm mô hình VTS_FC_ACO trên từng lĩnh vực

Trong phần này, trình bày kết quả thử nghiệm mô hình ***VTS_FC_ACO***

trên từng lĩnh vực văn bản của kho ngữ liệu ViEvTextSum để đánh giá vai trò của từng đặc trưng trong từng lĩnh vực thông qua bộ hệ số đặc trưng bằng việc học văn bản tóm tắt mẫu do con người thực hiện trên từng lĩnh vực văn. Kết quả thử nghiệm được trình bày trong bảng 2-13.

*Bảng 2-13. Kết quả tóm tắt của mô hình **VTS_FC_ACO** trên từng lĩnh vực của kho ngữ liệu ViEvTextSum.*

Đặc trưng	Hệ số			
	Chính trị	Xã hội	Kinh tế	Thể thao
F1 - Vị trí câu	0.20	0.16	0.11	0.16
F2 - Trọng số TF.ISF	0.05	0.09	0.06	0.03
F3 - Độ dài câu	0.03	0.03	0.03	0.06
F4 - Xác suất thực từ	0.16	0.11	0.09	0.21
F5 - Danh từ riêng	0.04	0.20	0.22	0.10
F6 - Dữ liệu số	0.17	0.03	0.06	0.03
F7 - Độ tương đồng giữa câu với tiêu đề	0.16	0.19	0.19	0.22
F8 - Câu trung tâm	0.20	0.20	0.23	0.18
Độ chính xác trung bình ROUGE-N (N=1)	**0.468**	**0.456**	**0.511**	**0.469**

Qua kết quả, chúng ta có thể thấy rằng, mỗi lĩnh vực sẽ có một bộ hệ số đặc trưng khác nhau, trong đó các hệ số đặc trưng nào có kết quả cao phản ảnh sự quan trọng của đặc trưng đó. Đặc trưng vị trí câu, xác suất thực từ, độ tương đồng với tiêu đề, câu trung tâm là các đặc trưng có tính chất quan trọng trong cả 4 lĩnh vực, đặc trưng độ dài câu có hệ số thấp phản ảnh đặc trưng này đóng vai trò không đáng kể trong tóm tắt văn bản. Các đặc trưng còn lại phản ảnh mức độ quan trọng tuỳ vào từng lĩnh vực cụ thể. Ví dụ như, trong lĩnh lực chính trị, đặc trưng dữ liệu số quan trọng, danh từ riêng không quan trọng nhưng trong lĩnh vực xã hội, kinh tế và thể thao thì lại ngược lại.

2.4.4 Nhận xét các kết quả thử nghiệm

Bảng 2-14. Bảng tổng kết kết quả tóm tắt của các mô hình.

Kho ngữ liệu	Kết quả tóm tắt (ROUGE-N)			
	N=1	N=2	N=3	N=4
Đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt" mã số B2012-01-24				
Corpus_LTH	0.539	0.389	0.337	0.311
Mô hình GA kết hợp 5 đặc trưng				
Corpus_LTH	0.620	0.469	0.420	0.387
ViEvTextSum	0.437	0.152	0.082	0.051
Mô hình GA kết hợp 8 đặc trưng				
Corpus_LTH	0.654	0.480	0.436	0.401
ViEvTextSum	0.452	0.160	0.083	0.051
Mô hình ACO kết hợp 5 đặc trưng				
Corpus_LTH	0.629	0.476	0.422	0.389
ViEvTextSum	0.439	0.148	0.059	0.045
Mô hình ACO kết hợp 8 đặc trưng				
Corpus_LTH	0.665	0.500	0.445	0.408
ViEvTextSum	0.464	0.167	0.088	0.058

Qua bảng 2-14 tổng hợp kết quả thử nghiệm cho thấy:

- Do đặc điểm tóm tắt của 2 kho ngữ liệu được trình bày trong mục 2.2.1 cho nên khi dùng độ đo ROUGE-N với N>1 thì kết quả của kho ngữ liệu Corpus_LTH sẽ lớn hơn kho ngữ liệu ViEvTextSum.

- Tập 8 đặc trưng được đề xuất lựa chọn đều có vai trò trong bài toán tóm tắt văn bản tiếng Việt, trong đó 3 đặc trưng đóng vai trò quan trọng nhất là F1-vị trí câu (đã cải tiến), F2-TFxISF và F8- câu trung tâm.

- Mô hình ***VTS_FC*** kết hợp 8 đặc trưng đã cho kết quả tóm tắt tốt hơn hẳn so với mô hình tóm tắt sử dụng 5 đặc trưng của các nghiên cứu tóm tắt văn bản tiếng Việt trước đây đề xuất và mô hình tóm tắt của đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt".

- Mô hình ***VTS_FC_ACO*** có kết quả tốt hơn mô hình ***VTS_FC_GA***.

- Qua kết quả thử nghiệm mô hình ***VTS_FC_ACO*** trên từng lĩnh vực, chúng ta có thể tìm ra bộ hệ số đặc trưng cho từng lĩnh vực để nâng cao hiệu quả tóm tắt cho từng lĩnh vực văn bản cụ thể.

2.5 Kết luận Chương 2

Các kết quả Chương 2 đạt được bao gồm:

(1). Đã nghiên cứu, đề xuất phương pháp tóm tắt đơn văn bản tiếng Việt theo hướng trích rút dựa trên bộ hệ số đặc trưng được xác định bằng phương pháp học máy. Cụ thể:

- Lựa chọn 8 đặc trưng quan trọng của văn bản tiếng Việt bằng phương pháp khảo sát khoa học trên kho ngữ liệu văn bản tiếng Việt.

- Xác định các hệ số đặc trưng văn bản bằng phương pháp học máy sử dụng giải thuật di truyền thông qua quá trình học kho văn bản tóm tắt mẫu.

- Xác định các hệ số đặc trưng văn bản bằng phương pháp học máy sử dụng giải thuật tối ưu đàn kiến thông qua quá trình học kho văn bản tóm tắt mẫu.

(2). Đã trình bày phương pháp thử nghiệm và kết quả thử nghiệm:

- Ảnh hưởng của từng đặc trưng trên các kho ngữ liệu.

- Kết quả thử nghiệm của hai mô hình xác định bộ hệ số đặc trưng bằng giải thuật di truyền và giải thuật tối ưu đàn kiến, cụ thể:

+ Kết quả thử nghiệm với 5 đặc trưng được đề xuất trong các nghiên cứu tóm tắt văn bản tiếng Việt trước đó.

+ Kết quả thử nghiệm với mô hình kết hợp 8 đặc trưng đã được đề xuất lựa chọn.

+ Kết quả thử nghiệm của mô hình kết hợp 8 đặc trưng đã được đề xuất lựa chọn trên từng lĩnh vực văn bản.

Nội dung của chương này đã được công bố trong công trình [CT5],[CT8],[CT9].

CHƯƠNG 3. TÓM TẮT VĂN BẢN TIẾNG VIỆT SỬ DỤNG KỸ THUẬT VOTING

Trong chương này, luận án trình bày phương pháp tóm tắt văn bản tiếng Việt mới sử dụng kỹ thuật Voting có hệ số phương pháp. Ý tưởng của phương pháp này là xem kết quả của mỗi phương pháp tóm tắt văn bản khác nhau là một lá phiếu có thứ tự ưu tiên. Trong đó mỗi lá phiếu là các câu đã được sắp xếp theo trọng số từ cao xuống thấp. Số lá phiếu có thứ tự sắp xếp câu giống nhau được gọi là hệ số phương pháp, hệ số này được tính toán thông qua học kho ngữ liệu tóm tắt mẫu bằng phương pháp học máy sử dụng giải thuật di truyền. Dựa trên kết quả từng lá phiếu và hệ số phương pháp, sử dụng kỹ thuật Voting để lựa chọn các câu có trọng số cao cho bản tóm tắt cuối cùng. Kết quả thử nghiệm cho thấy, kết quả tóm tắt của phương pháp sử dụng kỹ thuật Voting có hệ số phương pháp tốt hơn từng phương pháp đơn lẻ.

3.1 Mô hình tóm tắt văn bản sử dụng kỹ thuật Voting

Bầu chọn (voting) là một quá trình đưa ra quyết định lựa chọn một ứng viên trên lá phiếu để chọn ra ứng viên phù hợp cho một mục đích cụ thể. Người ta phân ra thành hai mô hình chính: mô hình chọn một người chiến thắng và mô hình chọn nhiều người chiến thắng. Theo 2 loại mô hình này, có nhiều phương pháp bỏ phiếu khác nhau được đề xuất như: phương pháp số phiếu đồng thuận, phương pháp đa số, phương pháp tính điểm Borda, phương pháp so sánh từng cặp Condorcet, phương pháp Schulze...[61]. Mỗi phương pháp có những điểm mạnh yếu riêng và phù hợp với các mô hình chọn ứng viên riêng. Người ta đã xây dựng các tiêu chuẩn riêng cho bài toán bầu cử. Dựa vào các tiêu chuẩn này, tuỳ vào từng mô hình chọn ứng viên (chọn một người chiến thắng hay nhiều người chiến thắng) mà người ta chọn phương pháp bỏ phiếu phù hợp.

Qua phân tích các hướng tiếp cận tóm tắt văn bản theo hướng trích rút (Hình 2-1). Chúng ta nhận thấy rằng, các phương pháp theo hướng này đều cho kết quả đầu ra là một danh sách các câu được sắp xếp theo trọng số từ cao đến thấp. Ta có thể xem đây là một lá phiếu bầu cử có thứ tự ưu tiên mà ứng viên là chính là câu. Và bài toán tóm tắt văn bản chính là lựa chọn theo kết quả Voting bằng mô hình chọn nhiều người chiến thắng.

Dựa vào quan sát này, luận án đề xuất phương pháp tóm tắt văn bản mới dựa theo kỹ thuật Voting với ý tưởng xem kết quả của mỗi phương pháp tóm tắt văn bản khác nhau là một lá phiếu đã được sắp xếp thứ tự ưu tiên các câu. Tuy nhiên, nếu ta xem mỗi phương pháp là một lá phiếu thì có khả năng xảy ra là số phương pháp yếu nhiều hơn sẽ thắng số phương pháp tốt (theo quan điểm đa số) và ngược lại. Để khắc phục điểm này, luận án đưa ra "hệ số phương pháp" (số lá phiếu của từng phương pháp). Hệ số này sẽ quyết định độ tốt của phương pháp đầu vào, những phương pháp tốt sẽ có hệ số cao, những phương pháp yếu sẽ có hệ số thấp. Hệ số này sẽ được tính toán thông qua quá trình học kho dữ liệu mẫu bằng phương pháp học máy. Sau đó, sử dụng phương pháp Voting phù hợp để lựa chọn các câu ưu tú dựa trên các lá phiếu đã nêu. Kết quả của phương pháp này sẽ nghiêng về những quan điểm có sự đồng thuận nhiều hơn nên chắc chắn sẽ có kết quả tốt hơn các phương pháp đơn lẻ.

Điểm số của câu theo kỹ thuật Voting được tính theo công thức (3.1):

$$Score_{Voting}(s) = \sum_{i=1}^{n} k_i \times Score\_Method_i(s) \qquad (3.1)$$

trong đó: $Score\_Method_i(s)$ là thứ tự sắp xếp của câu s trong văn bản theo trọng số câu từ cao xuống thấp của phương pháp tóm tắt i; k_i là hệ số phương pháp; n là số phương pháp tóm tắt đầu vào.

Để hiểu rõ hơn về công thức, ta có thể xem ví dụ: văn bản d gồm 6 câu, sử dụng phương pháp tóm tắt văn bản M cho kết quả như trong bảng 3-1.

Bảng 3-1. Ví dụ mô tả cách tính $Score\_Method(s)$

Thứ tự câu	Trọng số câu được tính theo phương pháp tóm tắt M	$Score\_Method(s)$
s_1	0.45	3
s_2	0.32	4
s_3	0.56	2
s_4	0.73	1
s_5	0.21	5
s_6	0.11	6

Như vậy, ta có thể hiểu $Score\_Method(s)$ chính là thứ tự được sắp xếp cao xuống thấp của trọng số câu được tính theo phương pháp tóm tắt đầu vào.

Thuật toán $Score\_Method(s)$ được biểu diễn dưới dạng giả mã (pseudocode):

Algorithm: $Score\_Method(s)$
$S \leftarrow D$; //tập câu văn bản
$score \leftarrow$; tập giá trị của câu văn bản theo phương pháp tóm tắt
$gt \leftarrow 0$;
$f \leftarrow$; // tập giá trị đặc trưng của từng câu văn bản tương ứng
$i \leftarrow 0$;
While ($s_i \in D$ **) do**
{//
$s\_index \leftarrow i$;
$i \leftarrow i + 1$;
If $(s_i = s)$ **then** $index \leftarrow i$;
}
$Sort(score, s\_index)$; // sắp xếp giá trị trọng số câu từ cao xuống thấp
For ($j = 1$ **) to** i **then**
{//

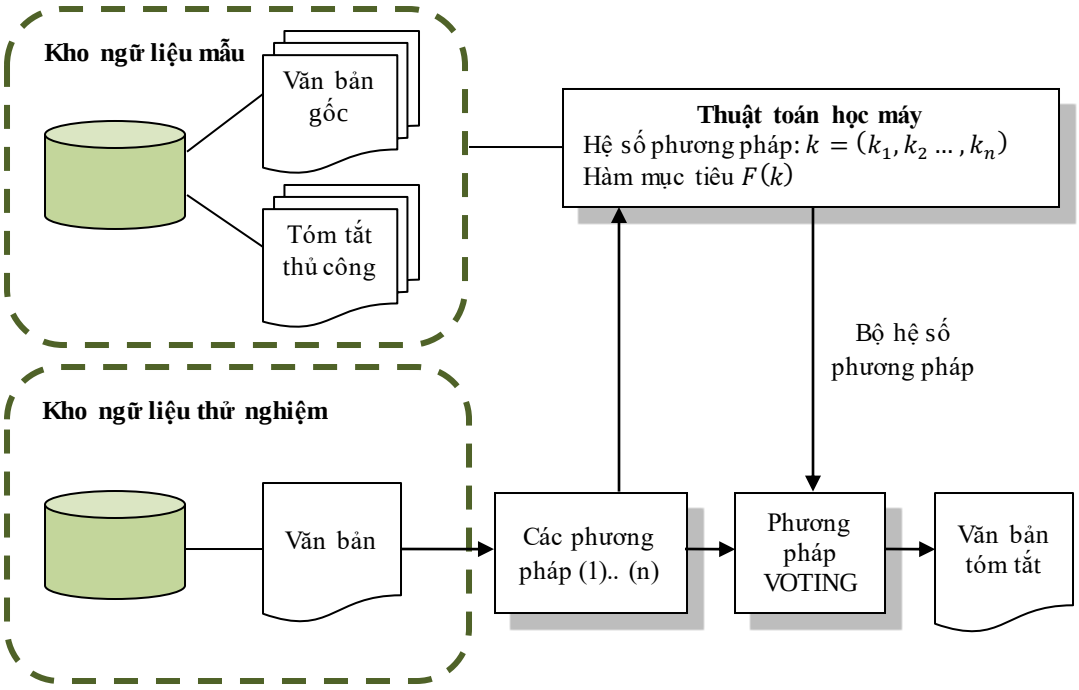
If $(j = s\_index[index])$ **then** $gt \leftarrow j;$

}

Return gt

Hình 3-1 Thuật toán gán trọng số $Score\_Method(s)$

Mô hình tóm tắt văn bản tiếng Việt được mô tả như hình 3-1.



Hình 3-2 Mô hình TTĐVB dựa theo kỹ thuật Voting

Mô hình tóm tắt văn bản dựa theo kỹ thuật Voting gồm 2 bài toán chính:

Bài toán 1: Xác định bộ hệ số phương pháp.

Bộ hệ số phương pháp được xác định thông qua quá học văn bản tóm tắt mẫu bằng phương pháp học máy.

Ở bài toán này, để tổng quát hoá luận án lựa chọn phương pháp học máy là giải thuật di truyền đã được trình bày trong chương 1 (mục 1.2). Tuy nhiên, nếu số phương pháp đầu vào cho mô hình tóm tắt văn bản tiếng Việt sử dụng kỹ thuật Voting là không nhiều, thì một số giải thuật tuần tự khác sẽ phát huy hiệu quả hơn vì nó cho nghiệm chính xác chứ không phải xấp xỉ như GA.

Bài toán 2: Tóm tắt văn bản dựa vào kỹ thuật Voting.

Với các lá phiếu đầu vào là các phương pháp tóm tắt văn bản đơn lẻ và bộ hệ số phương pháp được xác định qua bài toán 1, sử dụng phương pháp Voting chọn ra danh sách các ứng viên được bầu cao nhất (chính là các câu). theo tỷ lệ tóm tắt.

Ở bài toán này, luận án lựa chọn phương pháp Voting là phương pháp Schulze, phương pháp này hiện nay được ứng dụng nhiều trong mô hình bỏ phiếu chọn nhiều người chiến thắng (trình bày trong mục 1.4)

3.1.1 Xác định hệ số phương pháp bằng phương pháp học máy

3.1.1.1 Đặt bài toán

Trong phần này của luận án sẽ đề cập phương pháp xác định bộ hệ số phương pháp trong bài toán tóm tắt văn bản tiếng Việt sử dụng kỹ thuật Voting bằng phương pháp tối ưu. Như vậy bài toán đặt ra là tìm kiếm bộ hệ số của các phương pháp sao cho bản tóm tắt thu được dựa vào công thức (3.1) là "tốt nhất".

Tuy nhiên với số lượng các phương pháp đầu vào nhiều thì sẽ tạo ra tổ hợp số lượng các bộ hệ số k lớn. Do đó để xác định được bộ hệ số k tối ưu khó thực hiện theo các phương pháp tuần tự. Do vậy, một cách tự nhiên chúng ta sẽ đưa việc xác định bộ hệ số k vào bài toán tìm kiếm tối ưu.

Bài toán tìm hệ số phương pháp cho bài toán tóm tắt văn bản sử dụng kỹ thuật Voting được xác định bởi các dữ liệu sau:

$$\left(n, m, a, D = (d_1, d_2, \dots, d_m), SH = (sh_1, sh_2, \dots, sh_m), p = \left(p_1^i, p_2^i, \dots, p_n^i\right); \underset{i = 1 \dots m}{}\right)$$

trong đó:

- n là số phương pháp tóm tắt;
- m là số văn bản đầu vào để học;
- a là tỷ lệ tóm tắt;
- Đối với mỗi văn bản học thứ j trong tập văn bản mẫu D:

    + d_j là văn bản gốc thứ j (chứa tiêu đề và nội dung);

+ sh_j là bản tóm tắt do con người thực hiện của văn bản d_j;

+ p_i^j là các danh sách câu được sắp xếp theo trọng số của phương pháp tóm tắt văn bản i trên văn bản gốc thứ j.

Bài toán đặt ra là tìm các hệ số phương pháp k sao cho bản tóm tắt dựa vào kỹ thuật Voting $Sum_{voting}(a, d, p, k)$ theo tỉ lệ tóm tắt a "gần giống" với bản tóm tắt con người nhất.

Định nghĩa 3.1: *Một bộ hệ số là một vector* $k = (k_1, k_2 \dots, k_n), k_i \in \mathbb{R}$ *với* k_i *là hệ số phương pháp* p_i. *Bộ hệ số gọi là chấp nhận được nếu nó thỏa mãn điều kiện* $1 \geq k_i \geq 0$.

Một bản "tóm tắt vàng" của hệ thống sinh ra cần đạt được tiêu chí là chứa hầu hết các từ liên quan trong văn bản tóm tắt của con người. Độ đo đánh giá văn bản tóm tắt được định nghĩa như sau:

Định nghĩa 3.2: *Độ đo đánh giá văn bản tóm tắt được định nghĩa bằng độ tương tự giữa văn bản tóm tắt của hệ thống với văn bản tóm tắt con người theo độ đo độ đo đồng xuất hiện của thực từ trong văn bản tóm tắt hệ thống và văn bản tóm tắt con người:*

$$Sim\big(Sum_{voting}(a, d_i, p, k), sh_i\big) = \frac{\big|Sum_{voting}(a, d_i, p, k) \cap sh_i\big|}{|sh_i|}; \quad i = 1 \dots m \tag{3.2}$$

trong đó:

$Sum_{voting}(a, d_i, p, k) = \{sm_{i1}, \dots, sm_{ir}\}$ là vector thực từ khác nhau của văn bản tóm tắt theo kỹ thuật voting với bộ hệ số k theo tỉ lệ tóm tắt a của văn bản d_i;

$sh_i = \{sh_{i1}, \dots, sh_{iv}\}$ là vector thực từ khác nhau của văn bản sh_i.

Phát biểu bài toán: ($DMC(m, a, D, SH, p)$

Giả sử $k = (k_1, k_2 \dots, k_n)$ là bộ hệ số phương pháp chấp nhận được. Tìm k sao cho hàm mục tiêu:

$$F(D) = \sum_{i=1}^{m} \frac{Sim\left(Sum_{voting}(a, d_i, p, k), sh_i\right)}{m} \Longrightarrow \text{Giá trị cực đại} \quad (3.3)$$

với miền ràng buộc: $1 \geq k_i \geq 0$ (3.4)

3.1.1.2 Xác định hệ số phương pháp bằng giải thuật di truyền

Giống như chương 2 đã trình bày về phương pháp xác định hệ số đặc trưng bằng giải thuật di truyền, ở phần này chúng ta cũng thực hiện các bước tương tự để xác định hệ số phương pháp cho bài toán tóm tắt văn bản sử dụng kỹ thuật Voting. Mô hình tìm bộ hệ số phương pháp bằng giải thuật di truyền được mô tả trong hình 3-2.

Sau đây chúng ta sẽ lần lượt hình thức hóa bài toán xác định hệ số phương pháp bằng giải thuật di truyền cho bài toán tóm tắt văn bản trên ngôn ngữ của giải thuật di truyền.

Biểu diễn nhiễm sắc thể: Chúng ta sử dụng nhiễm sắc thể có cấu trúc mã hoá là một vector n chiều $(k_1, k_2 \ldots, k_n), \mathrm{k_i} \in \mathbb{Z}^+$ để biểu diễn các cá thể (các điểm) trong không gian tìm kiếm.

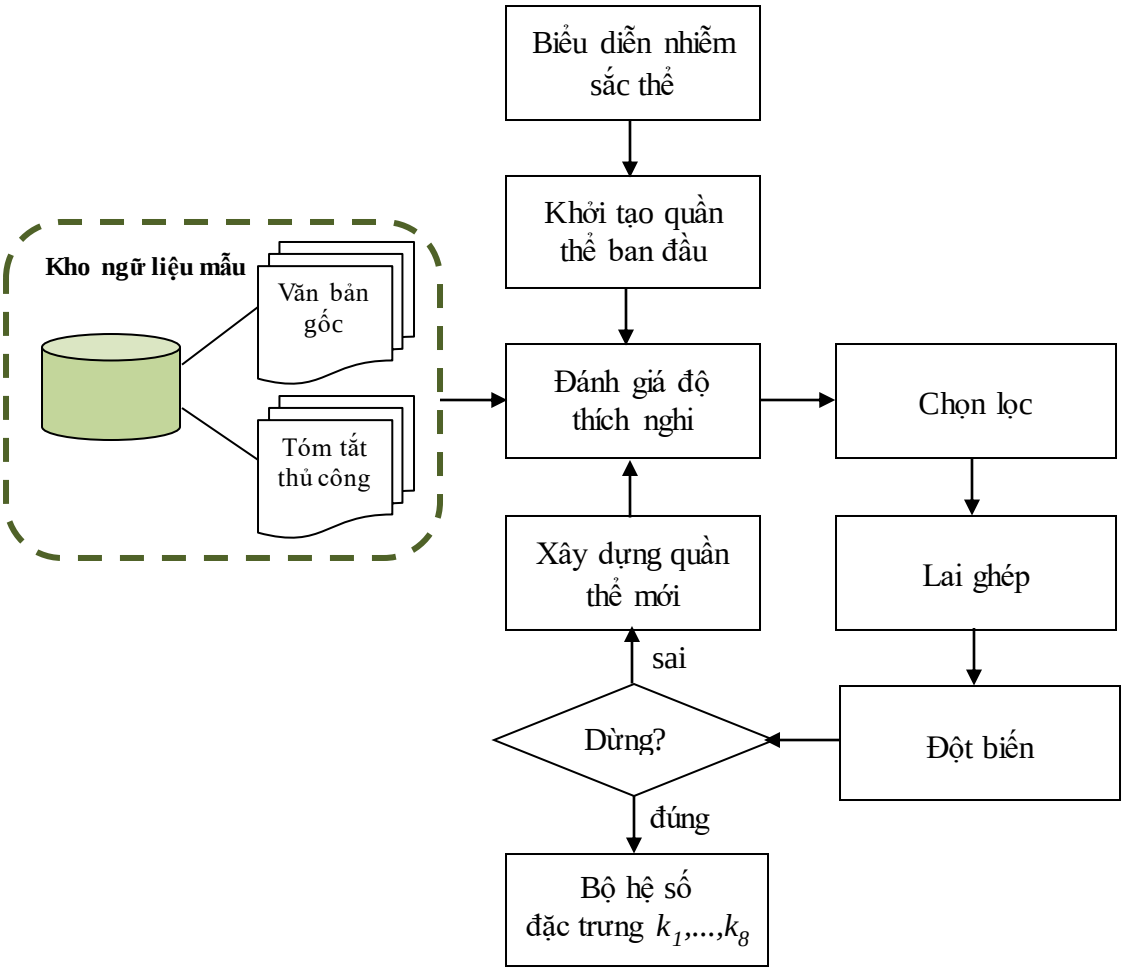
Độ đo thích nghi: Với mỗi cá thể $k = (k_1, k_2 \ldots, k_n)$ ta xác định mức độ thích nghi của cá thể $F(k)$ bằng công thức sau:

$$F(k) = \sum_{i=1}^{m} \frac{Sim(Sum_{voting}(a, d_i, p, k)_i, sh_i)}{m} \quad (3.5)$$

Toán tử lai ghép một điểm: Tương tự như (2.16) (2.17)

Toán tử đột biến phân phối đều: Tương tự như (2.18)

Toán tử chọn lọc: Tương tự như (2.19)



Hình 3-3 Mô hình học hệ số phương pháp bằng giải thuật toán truyền.

3.1.1.3 Xây dựng thuật toán

THUẬT TOÁN DI TRUYỀN XÁC ĐỊNH HỆ SỐ PHƯƠNG PHÁP

Input: $m, n, a, D, SH, p, G^{max}$

trong đó: m là số văn bản huấn luyện; n là số phương pháp; a là tỷ lệ tóm tắt; D là tập văn bản gốc; SH là tập bản tóm tắt thủ công tương ứng của tập văn bản D; p là tập các danh sách câu được sắp xếp theo trọng số của các phương pháp tóm tắt văn bản trên tập văn bản gốc D; G^{max} là số thế hệ.

Output: Nghiệm tối ưu của bài toán $DMC(m, a, D, SH, p)$ là vector hệ số phương pháp $k = (k_1, k_2 \ldots, k_n)$.

Bước 0. Khởi tạo quần thể gồm y cá thể $G_0 = \left(k^{1\,0}, \dots, k^{y\,0}\right)$, trong đó: $k^{i\,0} = \left(k_1^{i\,0}, \dots, k_n^{i\,0}\right); i = 1 \dots y$

Bước 1. Giải các bài toán $Sum_{voting}\left(a, d_i, p, k^{j\,t}\right), i = 1, \dots, m, j = 1, \dots, y$, t là số thế hệ thứ t của quần thể. Tính mức độ thích nghi $F\left(k^{j\,t}\right), i = 1, \dots, y$ cho từng cá thể của G_t theo (3.5). Áp dụng toán tử chọn lọc theo công thức (2.19) lên G_t để chọn ra y cá thể có mức độ thích nghi lớn nhất.

Bước 2. Nếu $t < G^{max}$ thì chạy tiếp đến Bước 3. Ngược lại thuật toán dừng và cho nghiệm tối ưu là bộ hệ số phương pháp tối ưu k có mức độ thích nghi lớn nhất trong y cá thể, nghĩa là thỏa mãn $F(\mathrm{k}) = max(F(k^{j\,t}), j = 1, \dots, y)$.

Bước 3. Lựa chọn các cha-mẹ trong G_t theo mức độ thích nghi để ghép cặp theo toán tử lai ghép một điểm (2.16) (2.17) để tạo nên tập các hậu thế G_t^{lg} với y_1 phần tử.

Bước 4. Tác động toán tử đột biến phân phối đều (2.18) vào $G_t \cup G_t^{\mathrm{lg}}$ để nhận được G_{t+1} đặt $t = t + 1$ và quay lại bước 1.

Các thuật toán được biểu diễn dưới dạng giả mã (pseudocode) được trình bày tương tự như mục 2.3.2.2.

3.1.1.4 Đánh giá độ phức tạp thuật toán

Thuật toán di truyền xác định hệ số phương pháp trình bày ở trên giống với thuật toán di truyền xác định hệ số đặc trưng đã được trình bày ở mục 2.3.2.2 trong chương 2. Do vậy, độ phức tạp cuả thuật toán di truyền xác định hệ số phương pháp được xác định bằng $O(N \times m)$ với N là số rất lớn.

3.1.2 Mô hình tóm tắt văn bản tiếng Việt sử dụng kỹ thuật Voting

Như đã trình bày ở trên, sau khi xác định được hệ số phương pháp chúng ta sử dụng kỹ thuật Voting trên tập kết quả của phương pháp đầu vào kết hợp

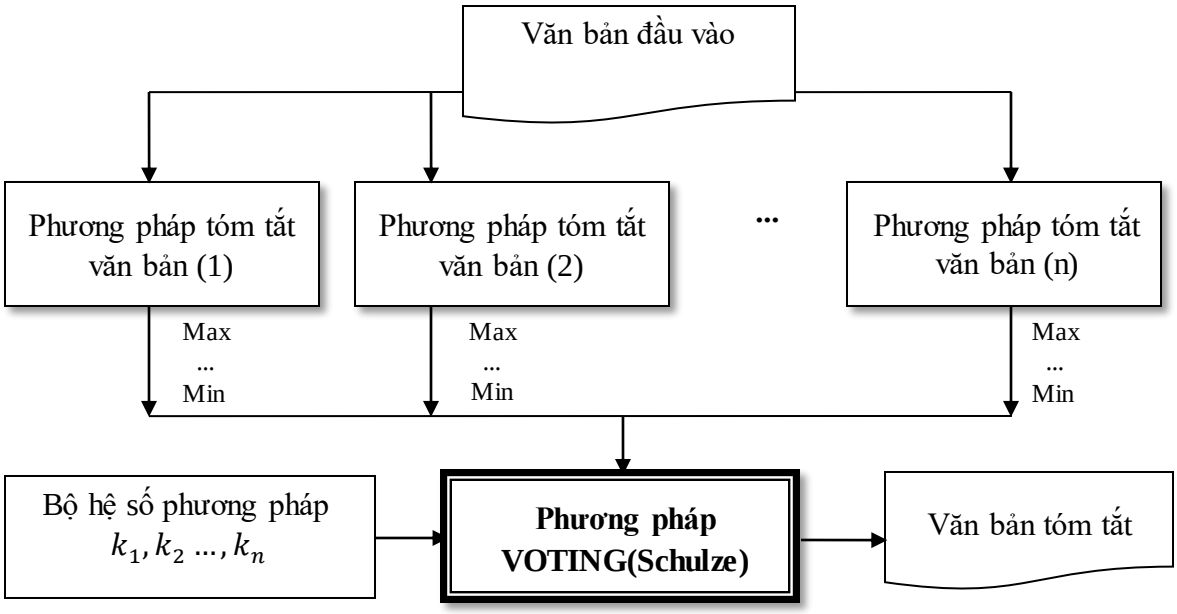
hệ số của các phương pháp đó. Kết quả Voting sẽ là tập các câu được sắp xếp theo trọng số Voting, chúng ta trích rút theo tỷ lệ để tạo ra bản tóm tắt cuối cùng. Mô hình tóm tắt văn bản tiếng Việt sử dụng kỹ thuật Voting được mô tả như hình 3-3, được mô tả các bước thực hiện như sau:

Bước 1. Tiền xử lý văn bản tiếng Việt cho đầu vào: tách câu, tách từ, gán nhãn loại bỏ hư từ...

Bước 2. Thực hiện tóm tắt văn bản theo từng phương pháp đầu vào (1), (2),..., (n). Kết quả của mỗi phương pháp là tập các câu được sắp xếp từ cao xuống thấp theo trọng số.

Bước 3. Thực hiện kỹ thuật Voting trên tập các kết quả và bộ hệ số phương pháp, kết quả đầu ra của phương pháp Voting là tập các câu được sắp xếp lại theo trọng số Voting từ cao xuống thấp.

Bước 4. Thực hiện lấy các câu có trọng số Voting từ cao xuống thấp theo tỉ lệ tóm tắt, xuất nguyên các câu đã trích rút theo thứ tự xuất hiện trong văn bản gốc.



Hình 3-4 Mô hình tóm tắt văn bản dựa theo kỹ thuật Voting.

Sau khi tìm được bộ hệ số phương pháp, tóm tắt văn bản theo thuật toán $TextSumvoting$ được biểu diễn dưới dạng giả mã (pseudocode):

Algorithm: $TextSumvoting(a,d,k)$
$S \leftarrow d$; //tập câu văn bản
$score \leftarrow \emptyset$; tập giá trị của câu văn bản
$S_{SUM} \leftarrow \emptyset$; //tập câu văn bản tóm tắt
$p \leftarrow$; // tập giá trị phương pháp của từng câu văn bản tương ứng
$i \leftarrow 0$;
While ($s_i \in S$) **do**
{// Tạo danh sách voting theo trọng số phương pháp p
$Listpp \leftarrow$;
}
$Schulze\_Method(Listpp,k)$; // voting các câu theo phương pháp Schulze với hệ số phương pháp k
$S_{SUM} \leftarrow$ lấy số câu tóm tắt có trọng số voting cao theo tỉ lệ tóm tắt a;
Return S_{SUM}

Hình 3-5 Thuật toán tóm tắt văn bản dựa theo kỹ thuật Voting Schulze.

Thuật toán $Schulze\_Method$ được trình bày kỹ trong phụ lục 3.

3.2 Các kết quả thử nghiệm

3.2.1 Kho ngữ liệu thử nghiệm

Tương tự như chương 2, trong phần thử nghiệm sử dụng 2 kho ngữ liệu Corpus_LTH và ViEvTextSum. Trong đó, sử dụng 80% kho ngữ liệu dùng để huấn luyện, 20% dùng để kiểm tra, đánh giá kết quả tóm tắt.

Để làm chính xác kết quả ở mỗi bước thử nghiệm, thực hiện 5 lần lấy 80% văn bản mẫu ngẫu nhiên để làm văn bản huấn luyện. Bộ hệ số thu được chính là bộ hệ số trung bình của 5 lần thực hiện đó. Sau khi có bộ hệ số, thực hiện tóm tắt 5 lần trên 20% văn bản ngẫu nhiên còn lại và thu được độ đo ROUGE-N trung bình của 5 lần tóm tắt.

3.2.2 Phương pháp đánh giá kết quả tóm tắt

Phương pháp đánh giá ROUGE-N đã được đề cập trong mục 2.3.2.

3.2.3 Lựa chọn các phương pháp tóm tắt văn bản đầu vào

Trong phần này, để có cơ sở đánh giá hiệu quả của phương pháp Voting. Chúng ta lựa chọn 05 phương pháp tóm tắt văn bản đầu vào cho phương pháp Voting dựa trên phương pháp tóm tắt văn bản dựa trên bộ hệ số đặc trưng đã được trình bày trong Chương 2. Trong đó mỗi phương pháp lựa chọn một số đặc trưng khác nhau để đại diện cho phương pháp, cụ thể: phương pháp 1 chọn 6 đặc trưng trong đó đặc trưng vị trí câu có vai trò lớn nhất đại diện cho phương pháp 1; phương pháp 2 lại chọn đặc trưng trọng số TF.ISF; phương pháp 4 chọn đặc trưng Câu trung tâm; phương pháp 4 chọn tần suất thực từ; phương pháp 5 lựa chọn toàn bộ 8 đặc trưng. Với sự lựa chọn 5 phương pháp này, qua đó thử nghiệm sử dụng phương pháp Voting để xác định hiệu quả của phương pháp Voting có hoặc không có hệ số phương pháp.

Để hiểu rõ hơn, chúng ta xem mô tả các đặc trưng được lựa chọn của từng phương pháp tóm tắt được thể hiện trong bảng 3-2:

Bảng 3-2. Bảng thống kê đặc trưng của 5 phương pháp đầu vào.

Đặc trưng	**Phương pháp (1)**	**Phương pháp (2)**	**Phương pháp (3)**	**Phương pháp (4)**	**Phương pháp (5)**
F1 - Vị trí câu	✓			✓	✓
F2 - Trọng số TF.ISF		✓			✓
F3 - Độ dài câu	✓	✓	✓		✓
F4 - Xác suất thực từ	✓	✓	✓	✓	✓
F5 - Danh từ riêng	✓	✓	✓	✓	✓
F6 - Dữ liệu số	✓	✓	✓	✓	✓
F7 - Độ tương đồng giữa câu với tiêu đề	✓	✓	✓	✓	✓
F8 - Câu trung tâm			✓		✓

Với 5 phương pháp nêu trên, tiến hành thực hiện tóm tắt văn bản theo tỉ lệ 30% bằng mô hình tóm tắt dựa trên hệ số đặc trưng đã trình bày trong chương 2 với hệ số đặc trưng được xác định bằng giải thuật tối ưu đàn kiến. Kết quả tóm tắt của 5 phương pháp được mô tả trong bảng 3-3:

Bảng 3-3. Kết quả tóm tắt của 5 phương pháp đầu vào.

Phương pháp	**Kết quả tóm tắt (ROUGE-N)**			
	N=1	N=2	N=3	N=4
Kho ngữ liệu Corpus_LTH				
Phương pháp (1)	0.631	0.482	0.432	0.398
Phương pháp (2)	0.605	0.432	0.381	0.350
Phương pháp (3)	0.601	0.449	0.402	0.372
Phương pháp (4)	0.629	0.476	0.422	0.389
Phương pháp (5)	0.665	0.500	0.445	0.408
Kho ngữ liệu ViEvTextSum				
Phương pháp (1)	0.449	0.152	0.076	0.047
Phương pháp (2)	0.445	0.151	0.076	0.046
Phương pháp (3)	0.442	0.151	0.077	0.046
Phương pháp (4)	0.439	0.148	0.059	0.045
Phương pháp (5)	0.452	0.160	0.083	0.051

Phần tiếp theo, chúng ta tiến hành 2 thử nghiệm:

Thử nghiệm 1: bao gồm 3 phương pháp khá cạnh tranh nhau về kết quả là phương pháp (1)(2)(3), như ta quan sát trong bảng 3-2, mỗi phương pháp đều sử dụng đặc trưng có vai trò cao riêng làm chủ đạo cho phương pháp. Cụ thể: phương pháp (1) sử dụng đặc trưng vị trí câu, phương pháp (2) sử dụng trọng số TF.ISF, phương pháp (3) sử dụng đặc trưng câu trung tâm (Cả 3 đặc trưng này được xem là quan trọng hơn cả các đặc trưng khác còn lại đã được nêu trong phần thử nghiệm của chương 2). Mục đích của thử nghiệm này là xem

hiệu quả của phương pháp Voting có hoặc không có hệ số phương pháp trên các phương pháp tóm tắt đầu vào có kết quả cạnh tranh nhau.

Thử nghiệm 2: sử dụng cả 5 phương pháp, như quan sát ở bảng 3-3 thì phương pháp 5 chính là phương pháp cho kết quả tốt nhất (VTS_FC_ACO đã được trình bày trong chương 2). Mục đích của thử nghiệm này là xem phương pháp tóm tắt dựa trên kỹ thuật Voting có hệ số phương pháp có khắc phục được điểm yếu của phương pháp Voting không có hệ số phương pháp là các phương pháp yếu sẽ kéo kết quả Voting thấp hơn phương pháp tốt nhất.

3.2.4 Các kết quả thử nghiệm

Trong phần này, chúng ta sẽ xem xét kết quả tóm tắt của mô hình tóm tắt sử dụng kỹ thuật Voting không sử dụng bộ hệ số phương pháp ($k_i = 1$) và sử dụng bộ hệ số phương pháp được xác định bằng giải thuật di truyền.

3.2.4.1 Mô hình tóm tắt văn bản sử dụng kỹ thuật Voting không có hệ số phương pháp

Kết quả của mô hình tóm tắt văn bản sử dụng kỹ thuật Voting không có hệ số phương pháp được mô tả dưới bảng 3-4:

Bảng 3-4. Kết quả tóm tắt của mô hình sử dụng kỹ thuật Voting không có hệ số phương pháp.

Kho ngữ liệu	**Kết quả tóm tắt (ROUGE-N)**			
	N=1	N=2	N=3	N=4
Voting 3 phương pháp (1)(2)(3)				
Corpus_LTH	0.635	0.481	0.432	0.400
ViEvTextSum	0.460	0.161	0.077	0.049
Voting 5 phương pháp (1)(2)(3)(4)(5)				
Corpus_LTH	0.648	0.495	0.446	0.412
ViEvTextSum	0.461	0.162	0.077	0.049

Qua kết quả thử nghiệm, với thử nghiệm 1 dùng 3 phương pháp cạnh tranh, chúng ta có thể thấy mô hình tóm tắt sử dụng kỹ thuật Voting không có hệ số phương pháp đã cho kết quả tốt hơn từng phương pháp tóm tắt đơn lẻ trên cả 2 kho ngữ liệu. Tuy nhiên với thử nghiệm 2 khi sử dụng 5 phương pháp, trên kho ngữ liệu Corpus_LTH ta thấy 5 phương pháp cho kết quả khá khác biệt với phương pháp (5) là cho kết quả nổi trội hơn cả. Phương pháp Voting không có hệ số cho kết quả thấp hơn phương pháp (5) là phương pháp tốt nhất (0.648<0.665). Có nghĩa là, có nhiều phương pháp không tốt sẽ kéo kết quả Voting xuống thấp hơn phương pháp đầu vào tốt nhất. Trên kho ngữ liệu ViEvTextSum với kết quả tóm tắt của 5 phương pháp đơn lẻ khá cạnh tranh (trong bảng 3-4) thì kết quả Voting cao hơn các phương pháp đơn lẻ.

Như vậy, có thể kết luận rằng phương pháp tóm tắt văn bản dựa trên kỹ thuật Voting không sử dụng hệ số phương pháp chỉ phù hợp với các phương pháp đầu vào là những phương pháp có kết quả cạnh tranh với nhau.

Phần tiếp theo ta sẽ xem kết quả thử nghiệm của mô hình tóm tắt sử dụng kỹ thuật Voting kết hợp hệ số phương pháp có khắc phục được nhược điểm của mô hình tóm tắt sử dụng kỹ thuật Voting không có hệ số phương pháp không.

3.2.4.2 Mô hình tóm tắt văn bản sử dụng kỹ thuật voting kết hợp hệ số phương pháp

Thử nghiệm xác định hệ số phương pháp bằng phương pháp học máy sử dụng giải thuật di truyền. Trong quá trình huấn luyện, giải thuật di truyền sẽ được thực hiện với các bước như sau:

- Có 100 cá thể trong một quần thể;
- Xác suất lai ghép 0.8;
- Xác suất đột biến 0.1;
- Thuật toán dừng khi đạt được 1000 thế hệ.
- Tỷ lệ tóm tắt là 30%.

Bộ tham số này được xác định bằng phương pháp thử nghiệm. Đầu tiên chúng ta dựa vào bộ hệ số thông thường được đề xuất cho giải thuật di truyền gốc. Sau độ bộ tham số này được điều chỉnh trong quá trình thử nghiệm thông qua việc thay đổi các giá trị và đánh giá sự hội tụ của giải thuật thông qua Hàm thích nghi (công thức 3.5).

Sau khi tìm được bộ hệ số phương pháp tối ưu, thực hiện bước chuẩn hoá bộ hệ số phương pháp về đoạn [0,1].

Kết quả thử nghiệm tóm tắt văn bản sử dụng kỹ thuật voting với bộ hệ số phương pháp được xác định bằng giải thuật di truyền trên hai kho ngữ liệu Corpus_LTH và ViEvTextSum được trình bày trong bảng 3-5 và bảng 3-6.

Bảng 3-5. Kết quả tóm tắt của mô hình sử dụng kỹ thuật Voting với hệ số phương pháp trên kho ngữ liệu Corpus_LTH.

Phương pháp	**Hệ số**	**Kết quả tóm tắt mô hình sử dụng kỹ thuật Voting với hệ số phương pháp (ROUGE-N)**			
		N=1	N=2	N=3	N=4
Voting trên 3 phương pháp (1)(2)(3)					
Phương pháp (1)	0.4	0.644	0.488	0.439	0.406
Phương pháp (2)	0.5				
Phương pháp (3)	0.1				
Voting trên 5 phương pháp (1)(2)(3)(4)(5)					
Phương pháp (1)	0.02	0.667	0.505	0.450	0.414
Phương pháp (2)	0.02				
Phương pháp (3)	0.17				
Phương pháp (4)	0.03				
Phương pháp (5)	0.77				

Bảng 3-6. Kết quả tóm tắt của mô hình sử dụng kỹ thuật Voting với hệ số phương pháp trên kho ngữ liệu ViEvTextSum.

Phương pháp	**Hệ số**	**Kết quả tóm tắt mô hình sử dụng kỹ thuật Voting với hệ số phương pháp (ROUGE-N)**			
		N=1	N=2	N=3	N=4
Voting trên 3 phương pháp (1)(2)(3)					
Phương pháp (1)	0.47	0.462	0.165	0.084	0.051
Phương pháp (2)	0.09				
Phương pháp (3)	0.44				
Voting trên 5 phương pháp (1)(2)(3)(4)(5)					
Phương pháp (1)	0.05	0.470	0.173	0.094	0.061
Phương pháp (2)	0.13				
Phương pháp (3)	0.32				
Phương pháp (4)	0.14				
Phương pháp (5)	0.35				

Kết quả thử nghiệm trên cho thấy, việc đưa hệ số phương pháp vào bài toán tóm tắt văn bản dựa vào kỹ thuật Voting đã phát huy được hiệu quả của phương pháp. Kết quả của mô hình tóm tắt văn bản dựa trên kỹ thuật Voting có hệ số phương pháp cao hơn các phương pháp đơn lẻ, mô hình này đã tránh tình trạng nhiều phương pháp yếu sẽ kéo kết quả xuống thấp hơn phương pháp tốt như đã trình bày ở trên.

3.2.5 Nhận xét các kết quả thử nghiệm

Với các thử nghiệm ở trên, chúng ta có bảng 3-7 và 3-8 tổng hợp kết quả của tất cả các thử nghiệm.

Cụ thể như sau:

- Thử nghiệm trên kho ngữ liệu Corpus_LTH:

Bảng 3-7. Bảng tổng kết kết quả thử nghiệm trên kho ngữ liệu Corpus_LTH.

Phương pháp	Kết quả tóm tắt (ROUGE-N)			
	N=1	N=2	N=3	N=4
Kết quả từng phương pháp				
Phương pháp (1)	0.631	0.482	0.432	0.398
Phương pháp (2)	0.605	0.432	0.381	0.350
Phương pháp (3)	0.601	0.449	0.402	0.372
Phương pháp (4)	0.629	0.476	0.422	0.389
Phương pháp (5)	**0.665**	**0.500**	**0.445**	**0.408**
Mô hình tóm tắt sử dụng kỹ thuật Voting không có hệ số phương pháp				
Phương pháp (1)(2)(3)	0.635	0.481	0.432	0.400
Phương pháp (1)(2)(3)(4)(5)	0.648	0.495	0.446	0.412
Mô hình tóm tắt sử dụng kỹ thuật Voting có hệ số phương pháp				
Phương pháp (1)(2)(3)	0.644	0.488	0.439	0.406
Phương pháp (1)(2)(3)(4)(5)	0.667	0.505	0.450	0.414

- Thử nghiệm trên kho ngữ liệu ViEvTextSum:

Bảng 3-8. Bảng tổng kết kết quả thử nghiệm trên kho ngữ liệu ViEvTextSum.

Phương pháp	Kết quả tóm tắt (ROUGE-N)			
	N=1	N=2	N=3	N=4
Kết quả từng phương pháp				
Phương pháp (1)	0.449	0.152	0.076	0.047
Phương pháp (2)	0.445	0.151	0.076	0.046
Phương pháp (3)	0.442	0.151	0.077	0.046
Phương pháp (4)	0.439	0.148	0.059	0.045
Phương pháp (5)	**0.452**	**0.160**	**0.083**	**0.051**

Mô hình tóm tắt sử dụng kỹ thuật Voting không có hệ số phương pháp				
Phương pháp (1)(2)(3)	0.460	0.161	0.077	0.049
Phương pháp (1)(2)(3)(4)(5)	0.461	0.162	0.077	0.049
Mô hình tóm tắt sử dụng kỹ thuật Voting có hệ số phương pháp				
Phương pháp (1)(2)(3)	0.462	0.165	0.084	0.051
Phương pháp (1)(2)(3)(4)(5)	0.470	0.173	0.094	0.061

Qua hai bảng tổng hợp kết quả trên, cho thấy:

Phương pháp tóm tắt văn bản tiếng Việt theo hướng trích rút sử dụng kỹ thuật Voting kết hợp hệ số phương pháp được trình bày là một phương pháp hoàn toàn mới. Qua thử nghiệm, phương pháp này có kết quả tóm tắt tốt hơn các phương pháp tóm tắt đơn lẻ. Mặt khác, việc sử dụng bộ hệ số phương pháp đã phát huy hiệu quả và tránh được tình trạng nhiều phương pháp yếu sẽ kéo kết quả xuống thấp hơn phương pháp tốt. Do vậy kết quả thử nghiệm của phương pháp này sử dụng các kết quả của chương 2 làm đầu vào đã cho kết quả cao hơn kết quả tốt nhất của chương 2.

Kết quả nghiên cứu này có giá trị thực tiễn và ứng dụng rất cao, có thể phát triển thành một sản phẩm phần mềm ứng dụng hữu ích.

3.3 Kết luận Chương 3

Các kết quả Chương 3 đạt được bao gồm:

(1). Đã nghiên cứu, đề xuất phương pháp tóm tắt văn bản tiếng Việt mới dựa vào kỹ thuật Voting, cụ thể: nghiên cứu tập trung giải quyết hai bài toán của mô hình:

- Xác định các hệ số phương pháp của từng phương pháp đầu vào bằng giải thuật di truyền thông qua quá trình học kho văn bản tóm tắt mẫu.

- Sử dụng kỹ thuật Voting dựa trên tập kết quả của các phương pháp đầu vào và hệ số phương pháp của chúng để tạo ra bản tóm tắt theo tỉ lệ người dùng lựa chọn.

(2). Đã trình bày kết quả thử nghiệm 2 mô hình đề xuất:

- Kết quả thử nghiệm xác định bộ hệ số phương pháp bằng giải thuật di truyền thông qua quá trình học kho văn bản tóm tắt mẫu.

- Kết quả thử nghiệm tóm tắt văn bản tiếng Việt sử dụng kỹ thuật Voting không có hệ số phương pháp.

- Kết quả thử nghiệm tóm tắt văn bản tiếng Việt sử dụng kỹ thuật Voting dựa trên tập kết quả của các phương pháp đầu vào và hệ số phương pháp đã được xác định.

Nội dung của chương này đã được công bố trong các công trình [CT3],[CT4],[CT7].

CHƯƠNG 4. QUY TRÌNH XÂY DỰNG KHO NGỮ LIỆU CÓ CHÚ GIẢI CHO BÀI TOÁN TÓM TẮT VĂN BẢN TIẾNG VIỆT

Trong chương này, luận án trình bày đề xuất về quy trình xây dựng kho ngữ liệu có chú giải dùng cho huấn luyện và đánh giá trong bài toán tóm tắt Văn bản tiếng Việt bao gồm các giai đoạn thu thập, xây dựng bản tóm tắt con người, chú giải cấu trúc hóa và lưu trữ. Ngoài ra luận án còn trình bày các phương pháp đánh giá kho ngữ liệu xây dựng.

4.1 Đặt vấn đề

Trong nghiên cứu về lĩnh vực tóm tắt văn bản, để đánh giá hiệu quả của từng hệ thống tóm tắt, người ta thường so sánh bản tóm tắt của hệ thống với một bản tóm tắt được lưu trữ trong kho ngữ liệu lớn đủ tin cậy về nguồn thông tin và bản tóm tắt do con người xây dựng. Các phương pháp đánh giá tóm tắt văn bản đòi hỏi phải có một kho ngữ liệu chuẩn chứa đầy đủ các nguồn tài liệu và bản tóm tắt con người tương ứng với nó [33].

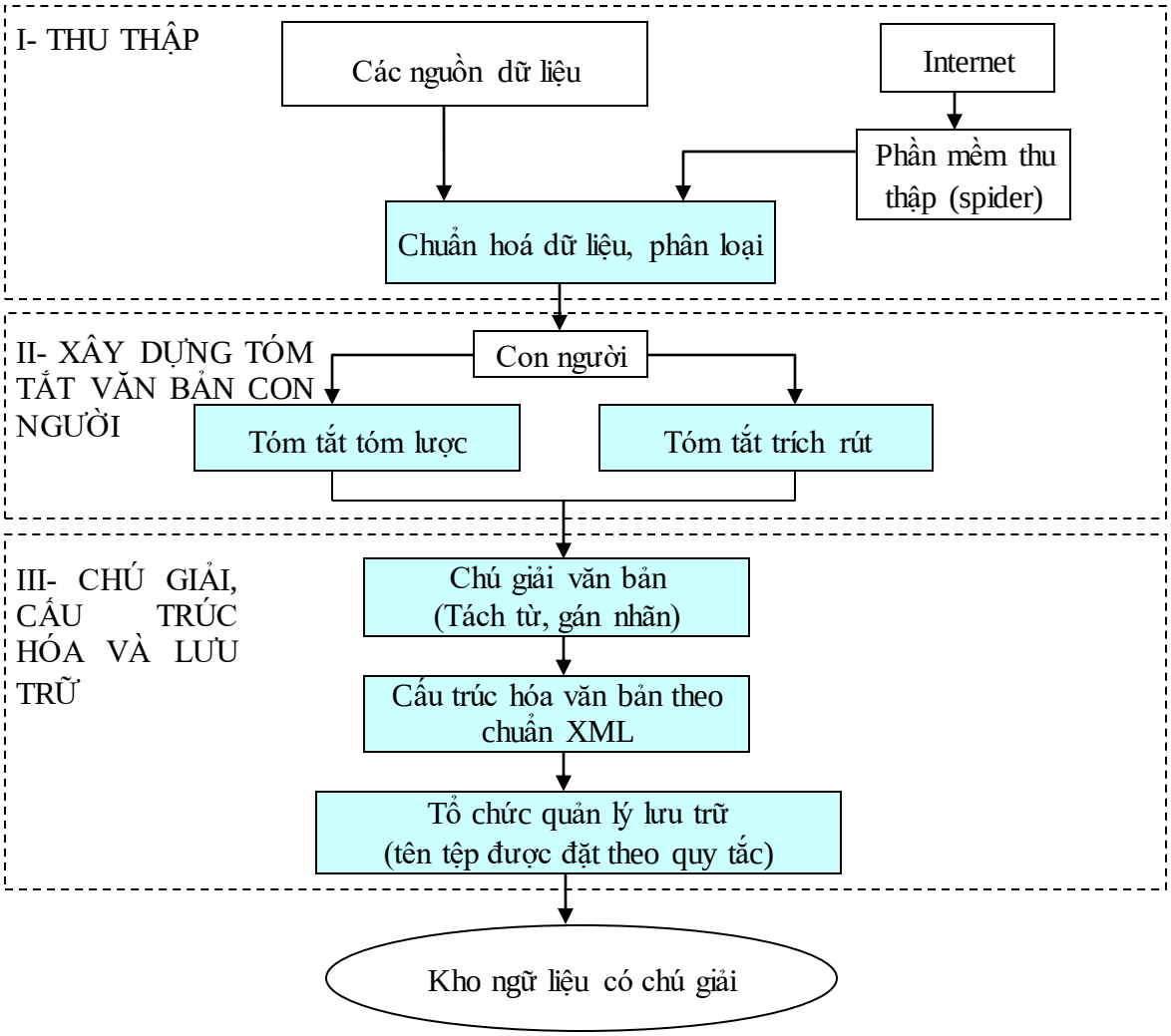
Cho đến nay, chưa có một kho ngữ liệu đầy đủ và chuẩn mực phục vụ cho bài toán tóm tắt văn bản tiếng Việt được công bố. Lý do có thể là do để xây dựng kho ngữ liệu này cần một số lượng chuyên gia ngôn ngữ và kinh phí đủ lớn. Việc thiếu kho ngữ liệu cho bài toán tóm tắt văn bản tiếng Việt là một lý do quan trọng để giải thích việc tại sao đến nay các nghiên cứu tóm tắt văn bản tiếng Việt còn ít. Mặt khác, do thiếu kho ngữ liệu chuẩn nên các phương pháp tóm tắt văn bản tiếng Việt đã đề xuất cũng chưa được đánh giá so sánh với nhau.

Chính vì vậy, trong chương này luận án trình bày quy trình xây dựng và cấu trúc kho ngữ liệu có chú giải dùng cho huấn luyện và đánh giá trong các hệ thống tóm tắt văn bản tiếng Việt.

4.2 Quy trình xây dựng kho ngữ liệu có chú giải

4.2.1 Mô hình đề xuất

Quy trình xây dựng kho ngữ liệu có chủ giải dùng cho huấn luyện và đánh giá các hệ thống tóm tắt văn bản tiếng Việt được mô tả như sơ đồ trong hình 4-1.



Hình 4-1 Quy trình xây dựng kho ngữ liệu có chú giải

Quy trình xây dựng gồm có 3 bước chính: Thu thập, xây dựng bản tóm tắt thủ công và chú giải, chuẩn hóa, lưu trữ. Phần tiếp theo sẽ mô tả chi tiết quá trình thực hiện các bước này.

4.2.2 Thu thập

Đầu vào của một hệ thống tóm tắt đơn văn bản là chỉ một văn bản duy

nhất, do vậy kho ngữ liệu cho lĩnh vực tóm tắt đơn văn bản tiếng Việt là các tài liệu đơn được thu thập từ các nguồn sau:

Nguồn dữ liệu local: Đây là những văn bản đã được thu thập, lưu trữ trên máy tính. (Ví dụ như kho ngữ liệu của các nghiên cứu tóm tắt văn bản trước đây được công bố)

Nguồn dữ liệu từ Internet: Đây được xác định là nguồn dữ liệu chính của kho ngữ liệu với số lượng văn bản dồi dào về nhiều lĩnh vực. Dữ liệu được xác định thu thập cho kho ngữ liệu là những trang thông tin (báo mạng) chính thống của nhà nước. Ưu điểm chính của nguồn dữ liệu này là thông tin đã được biên tập một cách cẩn thận về chính tả, văn phong và ngữ pháp tiếng Việt, mặt khác thông tin đã được cấu trúc và phân loại. Để thực hiện bước này một cách tự động, chúng ta có thể sử dụng các phần mềm thu thập (spider) đã được xây dựng sẵn hoặc có thể tự xây dựng phần mềm này.

Bảng 4-1 thống kê những trang báo mạng có thể thu thập để xây dựng kho ngữ liệu có chú giải cho bài toán tóm tắt văn bản tiếng Việt.

Bảng 4-1. Danh sách các trang mạng có thể lấy làm nguồn cho kho ngữ liệu

STT	Tên cơ quan	Địa chỉ web	Viết tắt
1.	Báo nhân dân điện tử	http://www.nhandan.com.vn/	BND
2.	Báo quân đội nhân dân	http://www.qdnd.vn/	BQD
3.	Báo công an nhân dân	http://www.cand.com.vn/	BCA
4.	Báo giáo dục	http://giaoduc.net.vn/	BDG
5.	Báo tiền phong điện tử	http://www.tienphong.vn/	BTP
6.	Báo tuổi trẻ	http://tuoitre.vn/	BTT
7.	Báo thanh niên	http://www.thanhnien.com.vn/	BTN
8.	Báo pháp luật	http://baophapluat.vn/	BPL
9.	Báo vietnamnet	http://vietnamnet.vn/	VNN
10.	Báo Hà tĩnh điện tử	http://baohatinh.vn	BHT

Dữ liệu sau khi thu thập về sẽ được phân loại theo các lĩnh vực. Với các nguồn dữ liệu thu thập như trên, văn bản thu thập được phân loại thành các lĩnh vực chính như trong bảng 4-2.

Bảng 4-2.Các lĩnh vực văn bản của kho ngữ liệu

STT	Lĩnh vực văn bản	Viết tắt
1.	Kinh tế	KT
2.	Văn hóa	VH
3.	Xã hội	XH
4.	Chính trị	CT
5.	Thể thao	TT
6.	Khoa học	KH

4.2.3 Xây dựng bản tóm tắt con người

Để xây dựng kho ngữ liệu dùng cho huấn luyện và đánh giá bài toán tóm tắt văn bản thì cần phải có bản tóm tắt của con người theo 2 hướng chính là tóm tắt tóm lược và tóm tắt trích rút. Phương pháp xây dựng các bản tóm tắt được mô tả như sau:

Bản tóm tắt tóm lược:

Thông thường để xây dựng tóm tắt tóm lược cho một tài liệu, người ta thường mời chuyên gia ngôn ngữ tóm tắt với số lượng từ nhất định. Tuy nhiên, phương pháp này rất tốn kém về tiền bạc và thời gian. Trong nghiên cứu này, sau khi nghiên cứu kỹ về cấu trúc bài báo tiếng Việt trên các trang báo mạng, có thể nhận thấy, một bài báo thường được cấu trúc thành 3 phần: tiêu đề, tóm tắt, nội dung. Phần tóm tắt chính là do chính tác giả tóm tắt cho bài báo của mình. Với quan sát này, chúng ta tận dụng phần tóm tắt của chính tác giả trong bài báo mạng thu thập về chứa số lượng từ đủ lớn (khoảng 120 từ trở lên) để làm phần tóm tắt tóm lược cho kho ngữ liệu.

Bản tóm tắt trích rút:

Phương pháp xây dựng bản tóm tắt trích rút là sử dụng một số chuyên gia ngôn ngữ lựa chọn các câu quan trọng bám với chủ đề văn bản làm bản tóm tắt với tỉ lệ cho trước. Sau khi có kết quả, sử dụng phương pháp voting theo đa số để chọn ra những câu được lựa chọn cao nhất.

4.2.4 Chú giải, cấu trúc hoá và lưu trữ.

4.2.4.1 Chú giải văn bản.

Do đặc thù và sự phức tạp của tiếng Việt, cho nên việc chú giải cho văn bản tiếng Việt trong kho ngữ liệu là một việc làm cần thiết giúp cho các nghiên cứu về bài toán tóm tắt văn bản tiếng Việt tiếp cận nhanh hơn trong quá trình huấn luyện và đánh giá.

Các chú giải văn bản tiếng Việt cho kho ngữ liệu bao gồm: chú giải về thông tin đoạn, câu; chú giải về từ; chú giải về từ loại.

Tách đoạn, câu:

Theo tài liệu hướng dẫn tách câu tiếng Việt của đề tài KC01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt". Câu được nhận diện qua quá trình phân tích câu đơn, câu kép. Mặt khác, thông qua nhận diện qua các đặc điểm sau:

+ Nhận diện chung: Với các kiểu câu bình thường như trên ta có thể nhận biết câu qua dấu câu: dấu chấm (câu tả, câu trần thuật, câu kể), dấu chấm than (câu cảm, câu cầu khiến), dấu chấm hỏi (câu hỏi).

+ Nhận diện câu trong hội thoại: Trong hội thoại dấu 2 chấm (:) báo hiệu cho lời nói trực tiếp, và lời nói trực tiếp này nằm trong dấu ngoặc kép ("…") hoặc bắt đầu sau dấu gạch đầu dòng (-). Trong trường hợp này, ta sẽ tách câu (nhận diện câu qua dấu hai chấm (:)). Trường hợp đoạn hội thoại có vế trích dẫn nằm ở cuối câu thì ta cũng sẽ tách câu. Vì trong lời nói trực tiếp có nhiều câu, khi ta tách chúng ra thành những câu riêng biệt, vế trích dẫn cuối cùng sẽ gắn với câu cuối cùng làm thành một câu khác có ý nghĩa khác thì câu sẽ trở

nên sai. Vì vậy ta sẽ tách vế này ra thành một câu.

+ Nhận diện câu sau dấu chấm phẩy (;) Dấu chấm phẩy (;) thường dùng để chỉ ranh giới giữa các vế trong câu ghép song song. Vì vậy ta có thể tách câu giống như câu ghép song song.

+ Nhận diện câu sau dấu ngang (-): Dấu ngang dùng để chỉ ranh giới của thành phần chú thích, đặt trước những lời đối thoại, liệt kê. Đối với câu có dấu ngang dùng để chỉ thành phần chú thích thì ta không nên tách câu.

Tách từ:

Với các ngôn ngữ biến hình như: tiếng Anh, tiếng Pháp, tiếng Đức, tiếng Nga,... việc nhận biết ranh giới từ trong các văn bản trên máy tính là khá đơn giản, chủ yếu là sử dụng khoảng trắng và các dấu câu. Bản thân các từ đã mang đầy đủ hình thái, nghĩa và ngữ pháp trong nó. Trái lại, đối với tiếng Việt, về mặt hình thức, từ được cấu tạo bởi một hay nhiều âm tiết ghép lại, nên khoảng trắng không phải dùng để phân biệt ranh giới từ.

Ví dụ. Từ đơn (có 1 âm tiết) và từ ghép (có từ 2 âm tiết trở lên)

- Từ đơn: nhà, cửa, đi, chạy, xanh, đỏ,...
- Từ ghép: gồm 3 dạng phổ biến sau

    + Từ kép: nhà trường, tổ chức, lung linh, lấp lánh, đu đưa,...
    + Từ bộ ba: phương pháp luận, bất đắc dĩ, sạch sành sanh,...
    + Từ bộ tư: xã hội chủ nghĩa, nói đi nói lại, đu đa đu đưa,...

Bài toán tách từ tiếng Việt có thể được phát biểu như sau:

Cho cụm từ gồm n âm tiết $S = s_1\ s_2\ s_3\ \ldots\ s_{i-1}\ s_i\ s_{i+1}\ \ldots\ s_{n-1}\ s_n$

Hãy tách thành dãy từ đúng $S = w_1\ w_2\ w_3 \ldots\ w_{m-1}\ w_m$

Ví dụ: *Các nghiên cứu sinh đang báo cáo.*

Được tách thành: | *Các* | *nghiên cứu sinh* | *đang* | *báo cáo* | . |

Gán nhãn từ loại:

Gắn nhãn từ loại là việc xác định các chức năng ngữ pháp của từ trong câu. Đây là bước cơ bản trước khi phân tích sâu văn phạm hay các vấn đề xử

lý ngôn ngữ phức tạp khác.

Thông thường, một từ có thể có nhiều chức năng ngữ pháp, ví dụ: trong câu "con ngựa đá đá con ngựa đá", cùng một từ "đá" nhưng từ thứ nhất và thứ ba giữ chức năng ngữ pháp là danh từ, nhưng từ thứ hai lại là động từ.

Xác định từ loại chính xác cho các từ trong văn bản là vấn đề rất quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên cụ thể là lĩnh vực tóm tắt văn bản. Việc gán nhãn từ loại tiếng Việt đúng giúp chúng ta phân biệt được những từ mang thông tin (thực từ) và những từ không mang thông tin (hư từ).

4.2.4.2 Cấu trúc tệp dữ liệu

Xây dựng cấu trúc tệp ngữ liệu với quan điểm là kho ngữ liệu phải chứa đầy đủ thông tin để phục vụ cho việc huấn luyện và đánh giá của bài toán tóm tắt đơn văn bản tiếng Việt. Do đó tài liệu cần phải được cấu trúc để chứa đầy đủ các thông tin bao gồm:

Các thông tin về tài liệu: số tài liệu, nguồn tài liệu, ngày xuất bản, lĩnh vực, tác giả.

Các thông tin về văn bản: Tiêu đề, nội dung (bao gồm thông tin về đoạn và câu)

Các thông tin về tóm tắt văn bản: tóm tắt tóm lược, tóm tắt trích rút do con người xây dựng.

Chú ý: Do một số phương pháp tóm tắt đơn văn bản theo hướng thống kê sử dụng thông tin về ví trí đoạn và vị trí câu. Vì vậy, trong phần nội dung cần phải thể hiện được thông tin về đoạn văn bản (paragraph), câu văn bản.

Cấu trúc theo chuẩn XML của tệp văn bản trong kho ngữ liệu được trình bày trong hình 4-2.

Giá trị SELECT=1 trong thẻ chứa câu <SENTENCE> có nghĩa là câu đó được chọn trong bản tóm tắt trích rút do con người tạo ra.

```
<DOC>
<INFOR> // thông tin của văn bản
        <DOCNO> Số văn bản </DOCNO>
        <SOURCE> Nguồn văn bản </SOURCE>
        <DATE-TIME> Ngày xuất bản <DATE-TIME>
        <CATEGORY> Lĩnh vực văn bản <CATEGORY>
</INFOR>
<TITLE> Tiêu đề của văn bản </TITLE>
<SUM_HUMAN> Tóm tắt tóm lược của con người </SUM_HUMAN>
<BODY>
        <PARA ID=1> //đoạn văn bản 1
                <SENTENCE ID=1 SELECT=1> Câu 1 </SENTENCE>
                <SENTENCE ID=2 > Câu 1 </SENTENCE>
                ...
        </PARA>
        <PARA ID=1> //đoạn văn bản 2
                <SENTENCE ID=1 > Câu 1 </SENTENCE>
                <SENTENCE ID=2 SELECT=1> Câu 1 </SENTENCE>
                ...
        </PARA>
        ...
</BODY>
</DOC>
```

Hình 4-2 Cấu trúc tệp ngữ liệu theo chuẩn XML.

4.2.5 Tổ chức quản lý, lưu trữ

Tài liệu khi thu thập về từ các nguồn dữ liệu được đặt tên theo quy ước:

<Nguồn văn bản> <Lĩnh vực văn bản> <Ngày xuất bản>.<Số tài liệu>.xml

Ví dụ: Tên tệp tài liệu BND.CT.20140724.068.xml có nghĩa tài liệu này được tải về từ trang "Báo nhân dân", lĩnh vực văn bản là "Chính trị", ngày xuất bản 24/07/2014 và số thứ tự lấy về là 68.

Các tệp được lưu trữ trên thư mục được lấy là tên quy ước của lĩnh vực văn bản trong bảng 4-1.

4.3 Phương pháp đánh giá kho ngữ liệu

Một bước quan trọng sau khi xây dựng kho ngữ liệu dùng cho huấn luyện và đánh giá bài toán tóm tắt văn bản tiếng Việt là đánh giá được chất lượng của

bản tóm tắt do con người tạo ra trong kho ngữ liệu. Để thực hiện điều này, luận án đề xuất 2 phương pháp đánh giá, trong đó một phương pháp đánh giá tự động dựa vào nội dung bản tóm tắt, một phương pháp đánh giá thủ công dựa vào con người chấm điểm. Chúng ta có thể sử dụng một trong hai phương pháp để đánh giá kho ngữ liệu tùy thuộc vào nhu cầu.

4.3.1 Đánh giá dựa vào độ đo đồng xuất hiện thực từ

Phương pháp đánh giá này dựa vào độ đo đồng xuất hiện thực từ giữa bản tóm tắt do con người thực hiện với văn bản gốc với quan điểm bản tóm tắt con người chứa hầu hết các từ liên quan trong văn bản gốc. Độ đo được định nghĩa như sau:

$$Sim(Sum_{human}, DOC) = \frac{|Sum_{human} \cap DOC|}{|SH_i|} \quad (4.1)$$

trong đó: $Sum_{human} = \{s_1, \dots, s_r\}$ là vector thực từ khác nhau của văn bản tóm tắt của con người; $DOC = \{d_1, \dots, d_v\}$ là vector thực từ khác nhau của văn bản gốc.

Để tăng độ chính xác cho độ đo, trong quá trình tính toán, các thực từ đồng nghĩa trong tiêu đề, nội dung được thay thế bằng một từ duy nhất bằng cách sử dụng từ điển đồng nghĩa của tác giả Nguyễn Văn Tu [11].

4.3.2 Đánh giá thủ công

Sử dụng con người đánh giá bản tóm tắt bằng phương pháp chấm điểm với thang điểm 10 cho mỗi bản tóm tắt tóm lược và tóm tắt trích rút. Để khách quan, phương pháp này nên sử dụng nhiều chuyên gia ngôn ngữ tự nhiên cùng chấm điểm. Kết quả đánh giá sẽ được tính trung bình dựa trên các bảng điểm của chuyên gia ngôn ngữ tự nhiên chấm. Các tiêu chí đưa ra cho chuyên gia ngôn ngữ chấm bao gồm:

- Bản tóm tắt bám sát chủ đề văn bản;
- Không có sự dư thừa dữ liệu;
- Văn bản có sự gắn kết giữa các câu, dễ đọc.

4.4 Kết luận Chương 4

Các kết quả mà chương 4 đạt được bao gồm:

(1). Đã nghiên cứu, đề xuất quy trình xây dựng kho ngữ liệu có chú giải dùng cho huấn luyện và đánh giá bài toán tóm tắt văn bản tiếng Việt. Bao gồm các bước sau:

- Thu thập dữ liệu.
- Xây dựng bản tóm tắt con người.
- Chú giải, cấu trúc hoá và lưu trữ.

(2). Đã nghiên cứu, trình bày các phương pháp đánh giá kho ngữ liệu có chú giải dùng cho huấn luyện và đánh giá bài toán tóm tắt văn bản tiếng Việt.

Nội dung liên quan đến chương này đã được công bố trong các công trình [CT2],[CT6].

KẾT LUẬN

Bài toán tóm tắt tiếng Việt có ý nghĩa quan trọng trong nâng cao hiệu quả khai thác thông tin từ các kho ngữ liệu văn bản tiếng Việt. Các công cụ tóm tắt tiếng Việt được ứng dụng nhiều trong các hệ thống tìm kiếm thông minh, đa ngôn ngữ, tổng hợp thông tin... Đối với lĩnh vực an ninh quốc phòng, tóm tắt tin tức có thể giúp cho cán bộ nghiệp vụ thu thập đủ các thông tin cần thiết và kịp thời theo dõi, đánh giá, xử lý nguồn thông tin một cách nhanh chóng. Nâng cao hiệu quả và độ chính xác của tóm tắt tiếng Việt là hướng nghiên cứu có ý nghĩa khoa học và thực tiễn luôn được các nhà khoa học quan tâm nghiên cứu. Chính vì vậy, mục tiêu nghiên cứu của luận án này là đề xuất các phương pháp tóm tắt văn bản mới phù hợp với văn bản tiếng Việt, có thể áp dụng xây dựng các phần mềm tóm tắt văn bản tiếng Việt chất lượng cao phục vụ trong nhiều lĩnh vực, nhất là lĩnh vực an ninh quốc phòng.

A. Các kết quả đạt được của luận án

1. Đã nghiên cứu, đánh giá các phương pháp tóm tắt văn bản và tóm tắt văn bản tiếng Việt.

2. Đã nghiên cứu, đánh giá ảnh hưởng của các đặc trưng văn bản tiếng Việt trong bài toán tóm tắt văn bản tiếng Việt. Qua đó, lựa chọn ra 08 đặc trưng văn bản quan trọng sử dụng trong phương pháp tóm tắt văn bản được đề xuất.

3. Đã đề xuất phương pháp tóm tắt đơn văn bản tiếng Việt theo hướng trích rút dựa vào bộ hệ số đặc trưng, bộ hệ số đặc trưng này được xác định bằng phương pháp học máy trên kho ngữ liệu tóm tắt mẫu.

4. Đề xuất kỹ thuật tóm tắt văn bản tiếng Việt theo hướng trích rút dựa vào kỹ thuật Voting kết hợp hệ số phương pháp, bộ hệ số phương pháp này được xác định bằng phương pháp học máy trên kho ngữ liệu tóm tắt mẫu.

5. Đã nghiên cứu, đề xuất quy trình xây dựng kho ngữ liệu có chú giải dùng cho huấn luyện và đánh giá bài toán tóm tắt văn bản tiếng Việt.

B. Những đóng góp mới của luận án

1. Đề xuất phương pháp tóm tắt đơn văn bản tiếng Việt theo hướng trích rút dựa trên bộ hệ số đặc trưng:

- Lựa chọn 8 đặc trưng của văn bản tiếng Việt bằng phương pháp khảo sát, đánh giá vai trò của từng đặc trưng trong văn bản tiếng Việt, qua đó đề xuất cải tiến một số đặc trưng: vị trí câu, độ dài câu cho phù hợp với văn bản tiếng Việt;

- Xác định bộ hệ số đặc trưng bằng phương pháp học máy sử dụng giải thuật di truyền thông qua kho ngữ liệu tóm tắt mẫu.

2. Đề xuất kỹ thuật tóm tắt văn bản tiếng Việt theo hướng trích rút dựa vào kỹ thuật Voting kết hợp hệ số phương pháp.

3. Đề xuất quy trình xây dựng kho ngữ liệu có chú giải dùng cho huấn luyện và đánh giá bài toán tóm tắt văn bản tiếng việt.

Các vấn đề mà luận án đã giải quyết được công bố trong 09 bài báo trên các tạp chí chuyên ngành và hội nghị khoa học.

C. Hướng nghiên cứu tiếp theo

- Mở rộng tập đặc trưng văn bản dựa vào Wordnet tiếng Việt.

- Xây dựng kho ngữ liệu đủ lớn, nhiều lĩnh vực phục vụ cho bài toán tóm tắt văn bản tiếng Việt.

- Xây dựng bộ hệ số đặc trưng chuẩn cho từng lĩnh vực.

- Thực hiện tóm tắt đa văn bản.

DANH MỤC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ

[CT1] Phạm Việt Trung, **Nguyễn Nhật An** (2009), "Nghiên cứu, xây dựng bộ công cụ hỗ trợ xử lý văn bản tiếng Việt phục vụ công tác an ninh quốc phòng", *Tạp chí Nghiên cứu Khoa học và Công nghệ quân sự,* ISSN 1859-1043, số đặc biệt, 04/2009, tr. 67-70.

[CT2] Trần Ngọc Anh, **Nguyễn Nhật An** (2011), "Lựa chọn tập gán nhãn ranh giới từ cho mô hình Markov ẩn trong bài toán tách từ tiếng Việt", *Tạp chí Nghiên cứu Khoa học và Công nghệ quân sự,* ISSN 1859-1043, đặc san 11/2011, tr.91-99.

[CT3] **Nguyễn Nhật An**, Trần Ngọc Anh (2014), "Tóm tắt văn bản tiếng Việt dựa vào kỹ thuật Voting", *Chuyên san Công nghệ thông tin và Truyền thông (JICT) thuộc Tạp chí Khoa học và Kỹ thuật, Học viện Kỹ thuật quân sự,* ISSN 1859-0209 (160), 4/2014, tr.57-67.

[CT4] **Nguyễn Nhật An**, Trần Ngọc Anh, Phan Thị Nguyệt Hoa (2014), "Kỹ thuật Voting trong bài toán tách từ tiếng Việt", *Tạp chí Nghiên cứu Khoa học và Công nghệ quân sự,* ISSN 1859-1043, Đặc san CNTT, 04/2014, tr.54-61.

[CT5] Đặng Thanh Quyền, Trần Ngọc Anh, **Nguyễn Nhật An** (2014), "Tối ưu hoá đàn kiến trong bài toán tách từ tiếng Việt", *Tạp chí Nghiên cứu Khoa học và Công nghệ quân sự,* ISSN 1859-1043, Đặc san CNTT, 04/2014, tr.219-229.

[CT6] **Nguyễn Nhật An**, Trần Ngọc Anh, Nguyễn Đức Hiếu (2014), "Kỹ thuật voting trong bài toán gán nhãn lớp thực từ, hư từ tiếng Việt", *Tạp chí Khoa học và Công nghệ, Đại học Công nghiệp Hà nội, ISSN 1859-3585,* số 23, 08/2014, tr.15-18.

[CT7] **Nguyễn Nhật An**, Nguyễn Quang Bắc, Nguyễn Đức Hiếu, Trần Ngọc Anh (2014), "Xác định các hệ số phương pháp cho bài toán tóm tắt văn bản tiếng Việt dựa vào kỹ thuật Voting", *Tạp chí Nghiên cứu Khoa học và Công nghệ quân sự,* ISSN 1859-1043, số 32, 08/2014, tr.82-90.

[CT8] **Nguyễn Nhật An**, Nguyễn Quang Bắc, Nguyễn Đức Hiếu, Trần Ngọc Anh (2014), "Xác định các hệ số đặc trưng bằng giải thuật di truyền cho bài

toán tóm tắt văn bản tiếng Việt", *Tạp chí Nghiên cứu Khoa học và Công nghệ quân sự*, ISSN 1859-1043, số 32, 08/2014, tr.36-46.

[CT9] **Nguyễn Nhật An,** Nguyễn Quang Bắc, Nguyễn Đức Hiếu (2015), "Tóm tắt văn bản tiếng Việt dựa trên bộ hệ số đặc trưng", *Tạp chí Nghiên cứu Khoa học và Công nghệ quân sự*, ISSN 1859-1043, số 35, 02/2015, tr.59-69.

TÀI LIỆU THAM KHẢO

Tiếng Việt:

[1] Diệp Quang Ban (2008), *Ngữ Pháp Tiếng Việt,* NXB giáo dục.

[2] Đỗ Phúc, Hoàng Kiếm (2006), "Rút trích ý chính từ văn bản tiếng Việt", *Tạp chí Công nghệ Thông tin và Truyền thông*.

[3] Đỗ Đức Đông (2012), *Phương pháp tối ưu đàn kiến và ứng dụng*, Luận án Tiến sỹ, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

[4] Hoàng Phê (1998), *Từ điển tiếng Việt*, NXB giáo dục.

[5] Lê Thanh Hương (2014), *Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt,* Báo cáo tổng kết đề tài cấp KH và CN cấp bộ, Đại học Bách khoa Hà Nội.

[6] Lưu Tuấn Anh, Yamamoto Kazuhide, *Ứng dụng phương pháp Pointwise vào bài toán tách từ cho Tiếng Việt*, http://vietlex.com.

[7] Nguyễn Hoàng Tú Anh (2011), *Tiếp cận đồ thị biểu diễn, khai thác văn bản và ứng dụng*, Luận án Tiến sỹ, Trường Đại học Khoa Học Tự Nhiên, ĐHQG-HCM.

[8] Nguyễn Hồng Thái (2008), *Tóm tắt văn bản tiếng Việt theo chủ đề*, Luận án Thạc sỹ, Đại học Bách khoa Hà Nội.

[9] Nguyễn Thị Thu Hà (2012), *Phát triển một số thuật toán tóm tắt văn bản Tiếng Việt sử dụng phương pháp học bán giám sát,* Luận án Tiến sỹ, Học viện Kỹ thuật quân sự.

[10] Nguyễn Trọng Phúc, Lê Thanh Hương (2008), "Tóm tắt văn bản tiếng Việt sử dụng cấu trúc diễn ngôn", *Hội thảo ICT.rda 2008*.

[11] Nguyễn Văn Tu (2001), *Từ điển đồng nghĩa Tiếng Việt*, NXB giáo dục.

[12] Trần Mai Vũ (2010), *Tóm tắt đa văn bản dựa vào trích xuất câu,* Luận văn Thạc sỹ, Trường ĐHCN, Đại học Quốc gia Hà Nội.

[13] Trương Quốc Định, Nguyễn Quang Dũng (2012), "Một giải pháp tóm tắt văn bản tiếng Việt tự động", *Hội thảo quốc gia lần thứ XV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông- Hà Nội*.

Tiếng Anh:

[14] Aone, Chinatsu, Marry Ellen Okurowski, James Gorlinsky, and Bjornar Larsen (1999), "A trainable Summarizer with Knowledge Acquired from Robust NLP Techniques", *In Advances in Automatic Text Summarization, by Inderjeet Mani and Mark T. Maybury*, pp.71-80.

[15] Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto (1999), *Modern Information Retrieval*. Addison Wesley.

[16] Baker K. (2005), *Singular Value Decomposition Tutorial*, Available at www.cs.wits.ac.za/ michael/SVDTut.pdf.

[17] Baxendale, P B. (1958), "Machine-made index for technical literature: an experiment", *IBM Journal of Research and Development 2*, pp.354-361.

[18] Barzilay, Regina, Michael Elhadad (1997), "Using Lexical Chains for Text Summarization", *In Proceedings of the Intelligent Scalable Text Summarization Workshop*, pp.10-17.

[19] Brandow, Ronald, Karl Mitze, Lisa F Rau (1995), "Automatic condensation of electronic publications by sentence selection", *Information Processing and Management: an International Journal*, Special issue: summarizing text 31, pp.675-685.

[20] Brin, Sergey, and Lawrence Page ((1998)), "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems 30*, pp.1-7.

[21] Chin-Yew Lin, (1999), "Training a Selection Function for Extraction", *In Proceedings of the Eighteenth Annual International ACM Conference on Information and Knowledge Management (CIKM)*, Kansas City, Kansas, Nov 2-6.

[22] Conroy, John M, and Dianne P O'leary (2001), "Text summarization via hidden Markov models", *SIGIR '01 Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM New York*, pp.406-407.

[23] Das, Dispanjan, Andre F.T Martins (2007), *A Survey on Automatic Text Summarization*. Literature survey for Language and Statistics II, Carnegie Mellon University.

[24] DeJong, Gerald F (1978), *Fast Skimming of News Stories: The FRUMP System*, PhD Thesis, Computer Science Department, Yale University.

[25] Dehkordi, P. K., H. Khosravi and F. Kumarci (2009), "Text Summarization Based on Genetic Programming", *International Journal of Computing and ICT Research Volume 3, No 1*, pp. 57–64.

[26] Dice, L.R. (1945), "Measures of the amount of ecologic association between species". *Ecology 26,* pp.297–302.

[27] Dorigo, M. and Gambardella, L. (1997), "Ant colonies for the traveling salesman problem". *BioSysterns,* 43, pp. 73-81.

[28] Dorigo, M., Maniezzo, V., and Colorni, A. (1996). "The ant system: Optimization by a colony of cooperating agents", *IEEE Transactions on Systems Man and Cybernetics Part B*, pp. 26-26.

[29] Edmundson, H P (1969), "New methods in automatic extracting", *Journal of the ACM 16*, pp.264-285.

[30] Ercan, Gönenç, İlyas Çiçekli (2008), "Lexical Cohesion based Topic Modeling for Summarization", *CICLing'08 Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, pp.582-592.

[31] Fattah, M. A. and F. Ren (2009), "GA, MR, FFNN, PNN and GMM Based Models for Automatic Text Summarization", *Computer Science and Language 23*, pp. 126–144.

[32] Hahn, Udo, Inderjeet Mani (2000), "The challenges of automatic summarization", *Computer 33*, pp.29-36.

[33] H. Saggion, et aI. (2010), "Multilingual summarization evaluation without human models," *in Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 1059-1067.

[34] Hovy, Eduard, Chin-Yew Lin1999, "Automated Text Summarization in SUMMARIST", *In Advances in automaic Text Summarization, by Inderjeet Mani and Mark T Maybury*, pp.81-94.

[35] Jones, Karen (1999), "Automatic Summarising: Factors and Directions", *In Advances in Automatic Text Summarization, by Inderjeet Mani and Mark T Maybury*, pp. 1-12.

[36] Jezek, Karel, and Josef Steinberger (2008), "Automatic Text Summarization (The state of the art 2007 and new challenges)", *Znalosti, Bratislava, Slovakia*, pp. 1-12.

[37] Karel Jezek and Josef Steinberger (2008), "Automatic Text summarization", *Vaclav Snasel (Ed.)*, pp.1-12.

[38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002) "BLEU: a Method for Automatic Evaluation of Machine Translation", *Computational Linguistics (ACL)*, Philadelphia, July 2002, pp. 311-318.

[39] Kupiec, Julian, Jan Pedersen, Francine Chen (1995), "A Trainable Document Summarizer", *SIGIR '95 Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.68-73.

[40] Kleinberg, Jon M (1999), "Authoritative sources in a hyper-linked environment", *Journal of the ACM 46*, pp.604-632.

[41] Knight, Kevin, and Daniel Marcu (2000), "Statistics-based summarization-Step one: Sentence compression", *Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence (AAAI- 2000)*, pp.703-710.

[42] Kiani, A. and M. R. Akbarzadeh (2006), "Automatic Text Summarization Using: Hybrid Fuzzy GA-GP", *2006 IEEE International Conference on Fuzzy Systems*, pp. 5465–5471.

[43] Landauer, Thomas K, Pete W Foltz, and Darrell Laham (1998), "An introduction to Latent Semantic Analysis", *Discourse Processes 25*, pp.259-284.

[44] Lin, Chin-Yew. (2004), "ROUGE: a Package for Automatic Evaluation of Summaries", *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25 - 26, 2004.

[45] Salton, G. (1998): *Automatic Text Processing*, Addison-Wesley Publishing Company.

[46] Lee, Daniel D, and H Sebastian Seung (1999), "Learning the parts of objects by non- negative matrix factorization", *Nature 401*, pp.788-791.

[47] Luhn, H P. (1958), "The Automatic Creation of Literature Abstracts" *IBM Journal of Research and Development 2*, pp.159-165.

[48] Mani, I., (2001), *Automatic Summarization*, John Benjamins Publishing Company.

[49] Markus Schulze (2011), "A New Monotonic, Clone-Independent, Reversal Symmetric, and Condorcet-Consistent Single-Winner Election Method", *Social Choice and Welfare*, February 2011, Volume 36, Issue 2, pp 267-303.

[50] Marcu, Daniel (1997), "From Discourse Structures to Text Summaries", *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp 82-88.

[51] Mohamed Abdel Fattah and Fuji Ren (2008), "Automatic Text Summarization", *Proceedings of World Academy of Science, Engineering and Technology*, Vol 27, ISSN 1307-6884, pp.192-195.

[52] Morris, Andrew H, George M Kasper, and Dennis A Adams (1992), "The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance", *Information Systems Research 3*, pp.17-35.

[53] Mitchell, M (1997), *An Introduction to Genetic Algorithms (third printing)*, MIT Press, ISBN: 0-262-13316-4, London, England.

[54] Mihalcea, Rada (2004), "Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization", *ACLdemo '04 Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Association for Computational Linguistics*, pp.170-173.

[55] M.L. Nguyen, Shimazu, Akira, Xuan, Hieu Phan, Tu, Bao Ho, Horiguchi, Susumu (2005), "Sentence Extraction with Support Vector Machine Ensemble", *Proceedings of the First World Congress of the International Federation for Systems Research: The New Roles of Systems Sciences For a Knowledge-based Society*.

[56] Ngoc Anh Tran, Thanh Tinh Dao, Phuong Thai Nguyen (2013), "Identifying Coordinated Compound Words for Vietnamese Word Segmentation", *Proceedings of the Fifth International Conference of Soft Computing and Pattern Recognition (SoCPaR 2013)*.

[57] Nguyen Quang Uy, Pham Tuan Anh, Truong Cong Doan, Nguyen Xuan Hoai (2012), "A Study on the Use of Genetic Programming for Automatic Text Summarization", *KSE, 2012 Fourth International Conference on Knowledge and Systems Engineering*, pp.93-98.

[58] L. H. Phuong, N. T. M. Huyen, R. Azim, R. Mathias (2010), "An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts", *Traitement Automatique des Langues Naturelles - TALN 2010, Montreal, Canada*.

[59] Ono, Kenji, Kazuo Sumita, Seiji Miike (1994), "Abstract Generation Based on Rhetorical Structure Extraction", *COLING '94 Proceedings of the 15th conference on Computational linguistics*, pp.344-348.

[60] Osborne, Miles (2002), "Using maximum entropy for sentence extraction", *AS '02 Proceedings of the ACL-02 Workshop on Automatic Summarization*, pp.1-8.

[61] Pacuit, Eric (2012), *Voting Methods*, The Stanford Encyclopedia of Philosophy (Winter 2012 Edition),

[62] Qazvinian, Vahed, and Dragomir R Radev (2008), "Scientific paper summarization using citation summary networks", *COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics*, pp.689-696.

[63] Radev, Dragomir R, Eduard Hovy, and Kathleen McKeown (2002),

"Introduction to the special issue on summarization", *Computational Linguistics 28*, pp.399-408.

[64] Radev, Dragomir R, Hongyan Jing, and Malgorzata Budzikowska (2000), "Centroid-based summarization of multiple documents", *NAACL-ANLP-AutoSum '00 Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization. Association for Computational Linguistics Morristown*, pp. 21-30.

[65] Radev, Dragomir R, et al (2003), "Evaluation Challenges in Large-scale Document Summarization", *ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*, pp.375-382.

[66] Rau, Lisa F, and Paul S Jacobs (1991), "Creating segmented databases from free text for text retrieval", *SIGIR '91 Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval. ACM*, pp.337-346.

[67] René Arnulfo García-Hernández, Yulia Ledeneva (2013), "Single Extractive Text Summarization Based on a Genetic Algorithm", *MCPR*, pp.374-383.

[68] Robert W. Floyd (1962), "Algorithm 97": *Shortest path, Communications of the ACM Volume 5 Issue 6*, pp. 345.

[69] Rucha S. Dixit, Prof. Dr.S.S.Apte, (2012) "Improvement of Text Summarization using Fuzzy Logic Based Method", *IOSR Journal of Computer Engineering (IOSRJCE)* ISSN: 2278-0661, ISBN: 2278-8727, Volume 5, Issue 6 (Sep-Oct. 2012), pp .05-10.

[70] Salton G. and Buckley C. (1997), "Term-weighting approaches in automatic text retrieval", *Information Processing and Management 24*, 1988. 513-523. Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) Readings in I.Retrieval. Morgan Kaufmann. 323-328.

[71] Svore, Krysta, Lucy Vanderwende, Chris Burges (2007), "Enhancing

single-document summarization by combining RankNet and third-party sources", *Proceedings of EMNLP-CoNLL*, pp.448-457.

[72] Suanmali L., Salim N., Salem Binwahlan M. (2011), "Genetic Algorithm based Sentence Extraction for Text Summarization", *International Journal of Innovative Computing 1*.

[73] Steinberger, Josef (2007), *Text Summarization within the LSA Framework*, PhD Thesis.

[74] S. Ye, et aI. (2005), *"NUS at DUC 2005: Understanding documents via concept links*," in Proceedings of Document Understanding Conferences.

[75] Teufel, Simone, Marc Moens (1997), "Sentence extraction as a classification task", *ACL/EACL workshop on" Intelligent and scalable Text summarization*, pp.58-65.

[76] Thanh Le Ha, Quyet Thang Huynh, Chi Mai Luong (2005), "A Primary Study on Summarization of Documents in Vietnamese", *Proceeding of the First International Congress of the International Federation for Systems Research, Kobe, Japan, Nov 15-17*, pp.234-239.

[77] Tu Nguyen Cam, Kien Nguyen Trung, Hieu Phan Xuan, Minh Nguyen Le, Thuy Ha Quang (2008), "Vietnamese Word Segmentation with CRFs and SVMs An Investigation", *Proceedings of th 20th he PACLI Wuhan, China*, p.215-222.

[78] Witbrock, Michael J, and Vibhu O Mittal (1999), "Ultra-summarization: a statistical approach to generating highly condensed non-extractive summaries", *SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM*, pp. 315-316.

WEB:

[79] VLSP project, Vietnamese Language Processing, http://vlsp.vietlp.org

[80] http://en.wikipedia.org/wiki/Schulze_method

[81] https://en.wikipedia.org/wiki/N-gram

PHỤ LỤC 01: KHO NGỮ LIỆU VIEVTEXTSUM

1. Nguồn thu thập

Bảng 1 thống kê những trang báo mạng thu thập để xây dựng kho ngữ liệu tiêu chuẩn ViEvTextSum.

Bảng 1. Danh sách các trang báo mạng là nguồn kho ngữ liệu

STT	Tên cơ quan	Địa chỉ web	Quy ước
1.	Báo nhân dân điện tử	http://www.nhandan.com.vn/	BND
2.	Báo quân đội nhân dân	http://www.qdnd.vn/	BQD
3.	Báo công an nhân dân	http://www.cand.com.vn/	BCA
4.	Báo giáo dục	http://giaoduc.net.vn/	BDG
5.	Báo tiền phong điện tử	http://www.tienphong.vn/	BTP
6.	Báo tuổi trẻ	http://tuoitre.vn/	BTT
7.	Báo thanh niên	http://www.thanhnien.com.vn/	BTN
8.	Báo pháp luật	http://baophapluat.vn/	BPL
9.	Báo vietnamnet	http://vietnamnet.vn/	VNN
10.	Báo Hà tĩnh điện tử	http://baohatinh.vn	BHT

Dữ liệu sau khi thu thập về sẽ được phân loại theo các lĩnh vực. Với các nguồn dữ liệu thu thập như trên, văn bản thu thập được phân loại thành các lĩnh vực chính như trong bảng 2.

Bảng 2. Các lĩnh vực văn bản của kho ngữ liệu

STT	Lĩnh vực văn bản	Tên quy ước
1.	Kinh tế	KT
2.	Văn hóa	VH
3.	Xã hội	XH
4.	Chính trị	CT
5.	Thể thao	TT

2. Xây dựng bản tóm tắt con người

Do thời gian và kinh phí hạn chế, trong phần này về phần tóm tắt tóm lược, luận án sử dụng phần tóm tắt của bài báo thu thập có số lượng từ trên 120 từ để làm phần tóm tắt tóm lược cho chính văn bản thu thập đó. Phần tóm tắt tóm lược, sử dụng 5 sinh viên ngôn ngữ lựa chọn các câu quan trọng theo chủ đề văn bản để làm bản tóm tắt trích rút (tỷ lệ tóm tắt 30%).

3. Chú giải văn bản, cấu trúc và lưu trữ

Các chú giải văn bản tiếng Việt cho kho ngữ liệu ViEvTEXTSUM bao gồm: chú giải về thông tin đoạn, câu; chú giải về từ; chú giải về từ loại.

Tách đoạn, câu

Luận án sử dụng bộ công cụ vnSentDetector (một gói của vnTokenizer [79]) để thực hiện tách câu tiếng Việt.

Tách từ

Luận án sử dụng phương pháp tách từ sử dụng kỹ thuật Voting được trình bày trong [CT4] với ý tưởng kết quả của mỗi phương pháp đầu vào được gán bộ nhãn BOI [CT2] và thực hiện phương pháp Voting đa số trên từng âm tiết. Phương pháp này sử dụng các kết quả của bộ công cụ tách từ vnTokenizer [79], JvnSegmenter [77], Pointwise [6] và nghiên cứu của nhóm tác giả Trần Ngọc Anh, Đào Thanh Tĩnh và Nguyễn Phương Thái [56]. Kết quả thử nghiệm tách từ theo phương pháp sử dụng kỹ thuật Voting cao hơn các phương pháp đơn lẻ.

Gán nhãn từ loại

Luận án sử dụng phương pháp gán nhãn từ loại sử dụng kỹ thuật Voting với ý tưởng kết quả của mỗi phương pháp gán nhãn đầu vào được thống nhất lại bộ 18 nhãn và thực hiện phương pháp Voting đa số trên từng từ vựng [CT6]. Phương pháp này sử dụng các kết quả của bộ công cụ tách từ vnTagger [58], JVnTagger [79] theo mô hình MEM và CRF. Kết quả gán nhãn từ loại theo phương pháp Voting cho thấy cao hơn các phương pháp đơn lẻ.

Cấu trúc tệp dữ liệu và lưu trữ: Được thực hiện giống phần trình bày

trong chương 4.

3. Kết quả xây dựng kho ngữ liệu ViEvTextSum

Do thời gian và kinh phí hạn chế, cho nên luận án thu thập một số lượng văn bản còn khiêm tốn để phục vụ cho bài toán tóm tắt văn bản tiếng Việt.

Bảng 3. Số lượng văn bản của kho ngữ liệu ViEvTEXTSUM

STT	Lĩnh vực văn bản	Số lượng	Tóm tắt tóm lược
1.	Kinh tế	1145	1145
2.	Văn hóa	1096	1096
3.	Xã hội	2725	2725
4.	Chính trị	1580	1580
5.	Thể thao	1515	1515

Hình 2 minh họa tệp ngữ liệu có chú giải thông tin về đoạn, câu, tách từ và gán nhãn trong kho ngữ liệu ViEvTEXTSUM.

```
<SUM_HUMAN> Đường/N Trường_Sơn/Np -/CH đường_mòn/N Hồ_Chí_Minh/Np ,/CH con/Nc
đường/N huyền_thoại/N đã/R không/R ít/A sách_báo/N phim_ảnh/N giới_thiệu/V
con/Nc đường/N huyền_thoại/N này/P ./CH Tuy_nhiên/C ,/CH tuyến_đường/N
giao_liên/N chuyển/V quân/N từ/E Bắc/Np vào/V Nam/Np phải/V vượt/V qua/V nhiều/A
con/Nc sông/N lớn/A ./CH Để/E bộ_đội/N vượt/V sông/N an_toàn/A ,/CH tránh/V
tổn_thất/N do/E không_quân/N Mỹ/Np đánh_phá/V là/V yêu_cầu/N cao/A nhất/R của/E
nhiệm_vụ/N ./CH Bến/N đò/N Chợ/N Thượng/Np ,/CH một/M trong/E những/L
trọng_điểm/N của/E bộ_đội/N qua/E sông/N đã/R nói/V lên/R điều/N đó/P ./CH
</SUM_HUMAN>
<BODY>
      <PARA ID="1">
            <SENTENCE ID="1">Bến_đò/N Chợ_Thượng/Np qua/V sông/N La/Np đã/R
            có/V từ/E xa_xưa/A ,/CH thuộc/V xã/N Trường_Sơn/Np (/CH Đức_Thọ/Np
            )/CH thường_ngày/A chở/V khách/N qua/V sông/N nối/V đôi/M bờ/N
            giao_lưu/V buôn_bán/V làm_ăn/V ./CH
            </SENTENCE>
            <SENTENCE ID="2">Chiến_tranh_phá_hoại/N nổ/V ra/R ,/CH đò/N
            Chợ_Thượng/Np được/V gánh/N thêm/V nhiệm_vụ/N chở/V bộ_đội/N qua/V
            sông/N vào/V Nam/Np chiến_đấu/V ./CH
            </SENTENCE>
            <SENTENCE ID="3">Trách_nhiệm/N này/P được/V giao/V cho/E Đảng_bộ/N
            và/Cc nhân_dân/N xã/N Trường_Sơn/Np suốt/A từ/E năm/N 1965/M đến/E
            khi/N Tổ_quốc/N thống_nhất/V ./CH
            </SENTENCE>
      </PARA>
      ….
</BODY>
```

Hình 2. Minh họa đoạn dữ liệu có chú giải trong tệp ngữ liệu.

PHỤ LỤC 02: KHO NGỮ LIỆU CORPUS_LTH

Kho ngữ liệu Corpus_LTH được xây dựng dựa trên kho ngữ liệu được công bố của đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt" mã số B2012-01-24 do tiến sỹ Lê Thanh Hương làm chủ nhiệm.

Kho ngữ liệu bao gồm 200 tệp văn bản tin tức và tóm tắt con người tương ứng thuộc 6 lĩnh vực: văn hoá, xã hội, chính trị, kinh tế, khoa học công nghệ, bộ KHCN và 200 tệp văn bản tóm tắt tương ứng của nó.

Từ kho ngữ liệu của đề tài, luận án thực hiện các bước tiền xử lý, chú giải văn bản và cấu trúc lưu trữ như trình bày trong phụ lục 01 để tạo ra kho ngữ liệu Corpus_LTH.

Bảng 1. Số lượng văn bản của kho ngữ liệu ViEvTEXTSUM

STT	Lĩnh vực văn bản	Số lượng	Tóm tắt tóm lược
1.	Kinh tế	53	53
2.	Văn hóa	34	34
3.	Xã hội	35	35
4.	Chính trị	31	31
5.	Khoa học giáo dục	22	22
6.	Bộ KHCN	25	25

PHỤ LỤC 03: THỬ NGHIỆM

1. Dữ liệu thử nghiệm

Dữ liệu thử nghiệm được tiền xử lý tách câu, tách từ, gán nhãn từ loại và được lưu trữ dưới định dạng XML. Ví dụ:

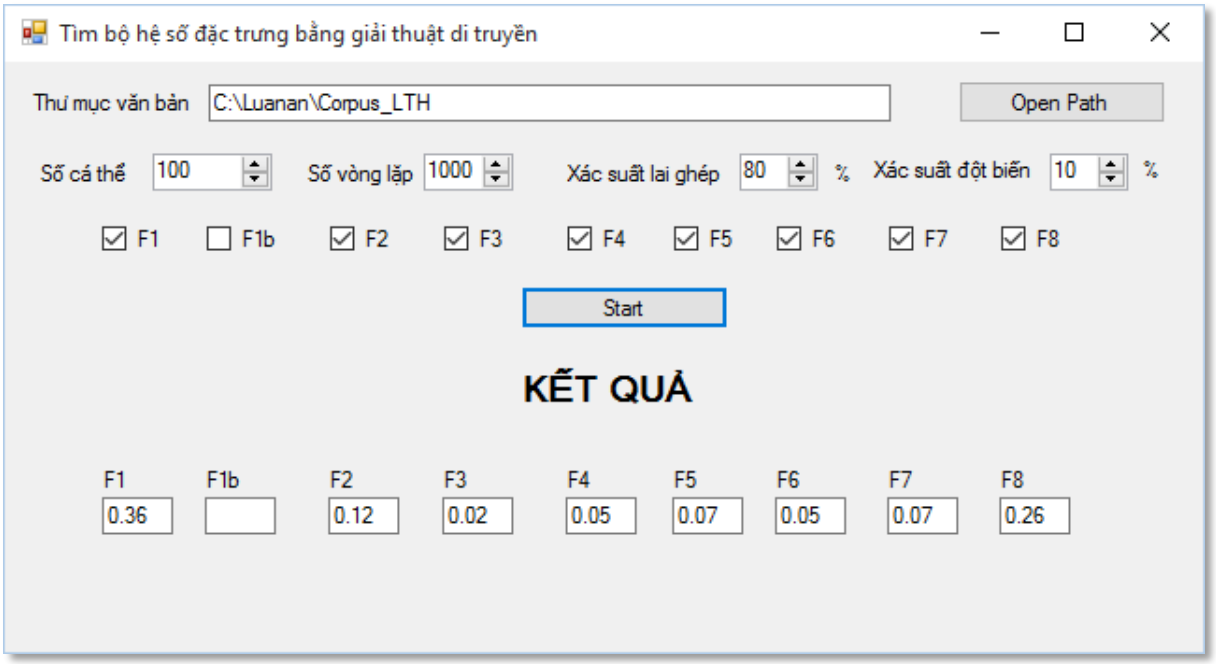
```
<DOC>
      <TITLE>Hà_Nội/Np tháo_dỡ/V hai/M cầu/N bộ_hành/N để/E xây/V cầu_vượt/Nc
./CH</TITLE>
      <INFO>Đề_tài/N của/E Lê_Thanh_Hương/Np :/CH B2012-01-24/M</INFO>
      <SUM_HUMAN>Mới/R được/V đưa/V vào/E sử_dụng/V chưa/R lâu/A ,/CH hai/M cây/N
cầu_vượt/Nc dành/V cho/E người/N đi/V bộ/N trên/E đường/N Nguyễn_Chí_Thanh/Np và/Cc
Trần_Khát_Chân/Np đã/R bị/V tháo_dỡ/V để_dành/V không_gian/N cho/E cầu/N vượt/V
dành/V cho/E xe_cơ_giới/N ./CH Đại_diện/N Sở/N Giao_thông/N vận_tải/V Hà_Nội/Np
cho/V biết/V ,/CH việc/N tháo_dỡ/V cầu/N dành/V cho/E người/N đi/V bộ/N để/E
xây_dựng/V cầu_vượt/Nc đã/R được/V tính_toán/V kỹ/A ./CH</SUM_HUMAN>
      <BODY>
          <PARA ID="1">
                <SENTENCE ID="1" F1="1,000000" F2="0,592618" F3="0,392425"
F4="0,900621" F5="0,695652" F6="0,184783" F7="0,827586" F8="1,000000" Vitri="DGC">
(/CH Dân_trí/N )/CH -/CH Mới/R được/V đưa/V vào/E sử_dụng/V chưa/R lâu/A ,/CH hai/M
cây/N cầu_vượt/Nc dành/V cho/E người/N đi/V bộ/N trên/E đường/N Nguyễn_Chí_Thanh/Np
và/Cc Trần_Khát_Chân/Np đã/R bị/V tháo_dỡ/V để_dành/V không_gian/N cho/E cầu/N
vượt/V dành/V cho/E xe_cơ_giới/N ./CH </SENTENCE>
          </PARA>
          <PARA ID="2">
                <SENTENCE ID="1" F1="1,000000" F2="1,000000" F3="0,351955"
F4="0,466270" F5="1,000000" F6="0,531250" F7="0,400000" F8="0,347178" Vitri="D">
Để/E giải_quyết/V tình_trạng/N ùn_tắc/V giao_thông/N vào/E giờ/N cao_điểm/N tại/E
nút/N giao/V Đại_Cồ_Việt/Np -/CH Trần_Khát_Chân/Np ,/CH đầu/N tháng/N 2/2013/M ,/CH
Hà_Nội/Np đã/R khởi_công/V cây/N cầu_vượt/Nc dài/A hơn/A 350/M m/Nu ,/CH rộng/A
11/M m/Nu ./CH </SENTENCE>
                <SENTENCE ID="2" F1="0,500000" F2="0,496093" F3="0,676251"
F4="0,984127" F5="0,533333" F6="0,000000" F7="0,545455" F8="0,975338" Vitri="G">
Cùng/A với/E đó/P ,/CH cây/Nc cầu/N dành/V cho/E người/N đi/V bộ/N trên/E đường/N
gần/A Trần_Khát_Chân/Np mới/R được/V đưa/V vào/E sử_dụng/V đã/R phải/V tháo_dỡ/V
./CH </SENTENCE>
                <SENTENCE ID="3" F1="0,333333" F2="0,796491" F3="0,978523"
F4="0,634921" F5="0,888889" F6="0,000000" F7="0,375000" F8="0,384687" Vitri="C">
Phần/N thân/A cầu/N được/V dùng/V lại/R ,/CH dự_kiến/V sẽ/R lắp/V trên/E đường/N
Giải_Phóng/Np ./CH </SENTENCE>
          </PARA>
          <PARA ID="3">
                <SENTENCE ID="1" F1="1,000000" F2="0,750507" F3="0,623240"
F4="0,658263" F5="0,941176" F6="1,000000" F7="0,250000" F8="0,547329" Vitri="D">
Một/M cây/N cầu_vượt/Nc dài/A 276m/M ,/CH rộng/A 17m/M ,/CH dành/V cho/E 4/M làn/Nc
xe_cơ_giới/N cũng/R mới/R được/V khởi_công/V tại/E nút/N giao/V Nguyễn_Chí_Thanh/Np
-/CH Liễu_Giai/Np ./CH </SENTENCE>
                <SENTENCE ID="2" F1="0,500000" F2="0,592232" F3="0,515148"
F4="0,857143" F5="0,400000" F6="0,212500" F7="0,666667" F8="0,950764" Vitri="C">
Cây/N cầu_vượt/Nc dành/V cho/E người/N đi/V bộ/N trên/E đường/N Nguyễn_Chí_Thanh/Np
(/CH nằm/V ngay/T đầu/N cầu/N vượt/V cho/E xe_cơ_giới/N )/CH cũng/R sẽ/R phải/V
tháo_dỡ/V ,/CH lắp_đặt/V lại/R cách/V vị_trí/N cũ/A 100m/M ./CH </SENTENCE>
          </PARA>
```

```
            <PARA ID="4">
                  <SENTENCE  ID="1"  F1="1,000000"  F2="0,620167"  F3="0,553918"
F4="0,827068" F5="0,421053" F6="0,000000" F7="0,923077" F8="0,878228" Vitri="D">
Đại_diện/N Sở/N Giao_thông/N vận_tải/V Hà_Nội/Np cho/V biết/V ,/CH việc/N tháo_dỡ/V
cầu/N dành/V cho/E người/N đi/V bộ/N để/E xây_dựng/V cầu_vượt/Nc đã/R được/V
tính_toán/V kỹ/A ./CH </SENTENCE>
                  <SENTENCE  ID="2"  F1="0,500000"  F2="0,525783"  F3="1,000000"
F4="1,000000" F5="0,800000" F6="0,000000" F7="0,705882" F8="0,749344" Vitri="G">
"/CH Cầu/Np dành/V cho/E người/N đi/V bộ/N có_thể/R tháo_dỡ/V lắp_đặt/V sang/V
vị_trí/N khác/A ./CH </SENTENCE>
                  <SENTENCE  ID="3"  F1="0,333333"  F2="0,713299"  F3="0,623240"
F4="0,714286" F5="0,000000" F6="0,000000" F7="1,000000" F8="0,644497" Vitri="C">
Do_vậy/C ,/CH việc/N tháo_dỡ/V cầu/N bộ_hành/N để/E xây_dựng/V cầu_vượt/Nc dành/V
cho/E xe_cơ_giới/N đem/V lại/R hiệu_quả/N cao/A hơn/A "/CH ,/CH đại_diện/N Sở/N
Giao_thông/N vận_tải/V nói/V ./CH </SENTENCE>
            </PARA>
      </BODY>
</DOC>
```

trong đó, các giá trị F1 đến F8 tại mỗi câu đã được tính trước theo các công thức được trình bày trong Mục 2.2.

2. Thử nghiệm tìm bộ tham số đặc trưng theo giải thuật di truyền

Màn hình chính thực hiện:



Các bước thử nghiệm:

Bước 1: Chọn thư mục dữ liệu huấn luyện.

Bước 2: Nhập các tham số như số cá thể của quần thể, số vòng lặp, xác

suất lai ghép, xác suất đột biến.

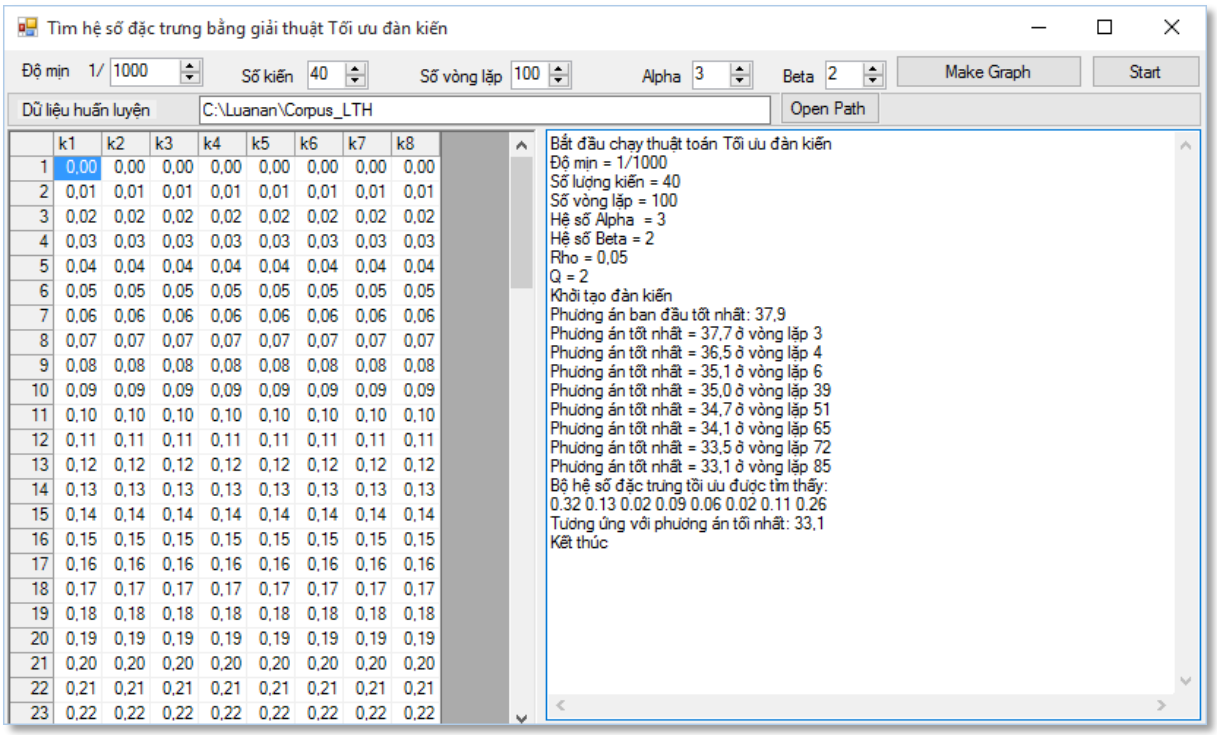
Lựa chọn các tham số cần tìm bằng cách đánh dấu tích vào các ô tham số.

Bước 3: Bấm nút Start để tìm kiếm bộ hệ số đặc trưng tối ưu nhất.

Bảng kết quả cho ta thấy bộ hệ số đặc trưng tối ưu được tìm thấy bằng giải thuật di truyền.

3. Thử nghiệm tìm bộ tham số đặc trưng theo giải thuật tối ưu đàn kiến

Màn hình chính:



Các bước thử nghiệm:

Bước 1: Nhập các tham số như độ mịn, số kiến, số vòng lặp, Hệ số alpha, beta.

Chọn thư mục dữ liệu huấn luyện.

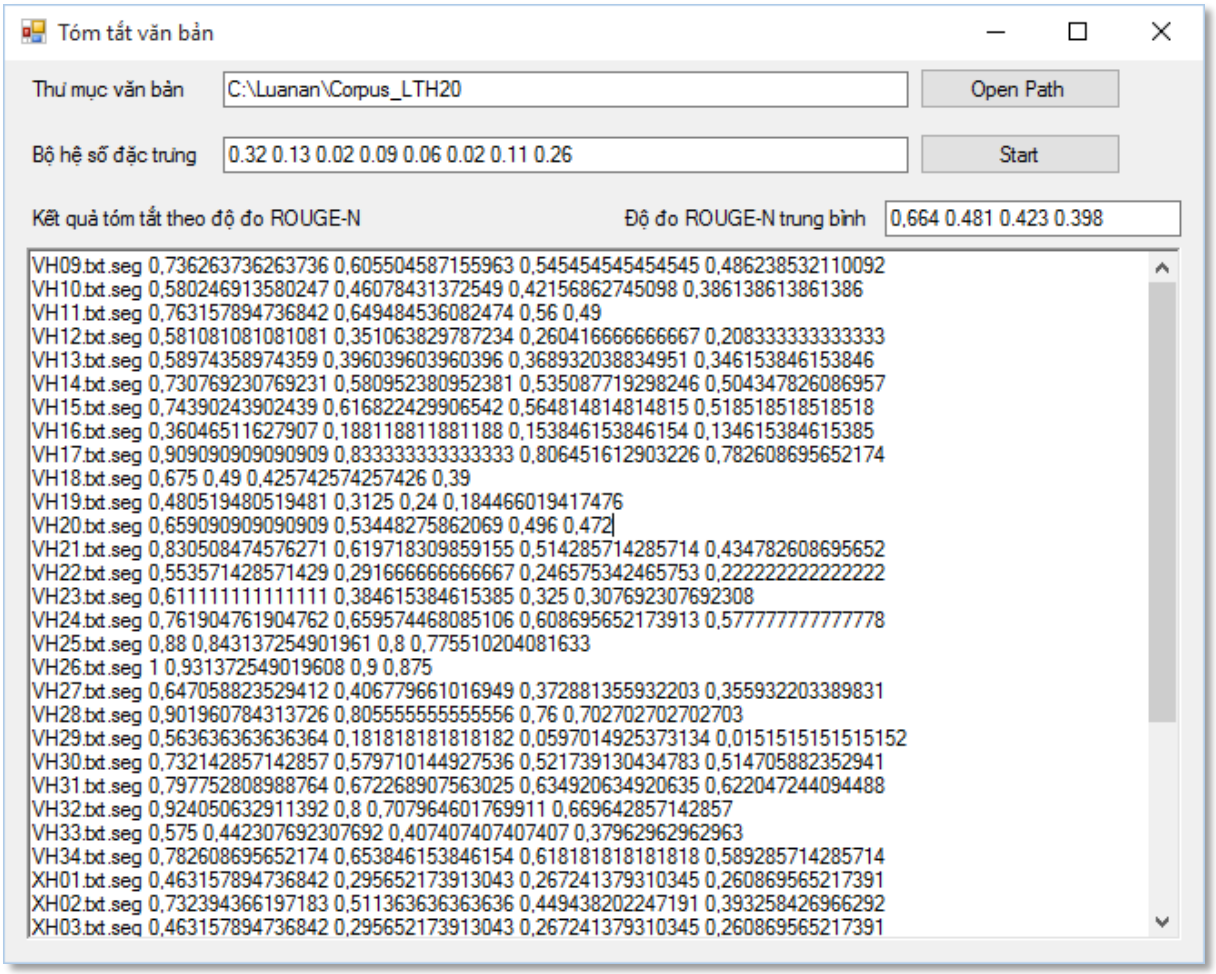
Bước 2: Tạo sơ đồ số nút bằng cách bấm Make Graph.

Bước 3: Bấm nút Start để tìm kiếm bộ hệ số đặc trưng tối ưu nhất.

Bảng kết quả cho ta thấy các bước thực hiện và kết quả của thuật toán tối ưu đàn kiến

4. Thử nghiệm tóm tắt văn bản sau khi tìm được bộ hệ số

Sau khi tìm được bộ hệ số tối ưu, tiến hành thử nghiệm tóm tắt văn bản. Màn hình chính của thử nghiệm này được trình bày như sau:



Các bước thử nghiệm:

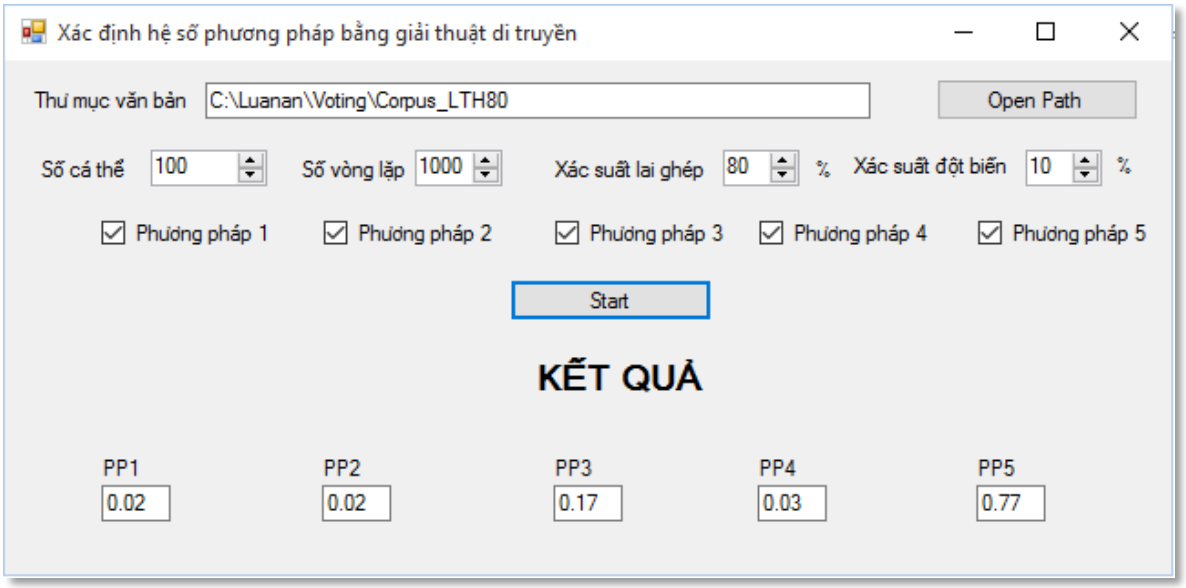
Bước 1: Chọn thư mục dữ liệu thử nghiệm; nhập hệ số tối ưu.

Bước 2: Bấm nút Start để tóm tắt toàn bộ văn bản trong thư mục theo bộ hệ số đã nhập.

Bảng kết quả cho ta thấy các độ đo ROUGE-N (1-gram, 2-gram, 3-gram, 4-gram) trung bình của toàn bộ thư mục và các văn bản trong thư mục. Kết quả văn bản tóm tắt từng văn bản gốc trong thư mục được lưu ra thư mục tomtat trong máy tính.

5. Thử nghiệm tóm tắt phương pháp Voting

Màn hình chính:



Các bước thử nghiệm:

Bước 1: Chọn thư mục dữ liệu huấn luyện.

Bước 2: Nhập các tham số như số cá thể của quần thể, số vòng lặp, xác suất lai ghép, xác suất đột biến.

Lựa chọn các phương pháp tóm tắt đầu vào bằng cách đánh dấu tích vào các ô phương pháp.

Bước 3: Bấm nút Start, chương trình sẽ thực hiện theo trình tự:

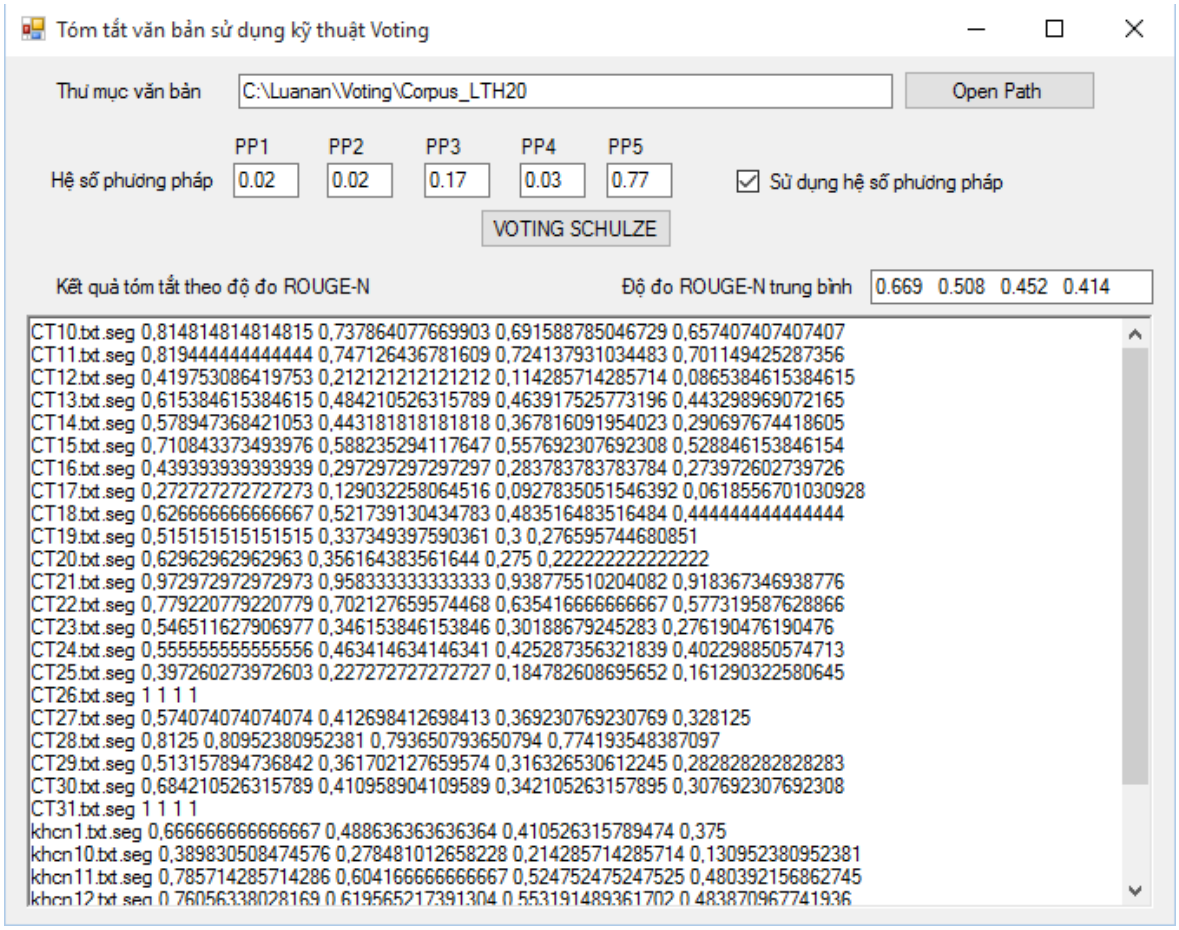
- Thực hiện tóm tắt văn bản theo từng phương pháp lựa chọn

- Gán trọng số voting của từng phương pháp lựa chọn cho các câu theo công thức (3.1).

- Thực hiện tìm hệ số phương pháp theo giải thuật di truyền.

Bảng kết quả cho ta thấy bộ hệ số phương pháp tối ưu được tìm thấy bằng giải thuật di truyền.

Sau khi tìm được bộ hệ số tối ưu, tiến hành thử nghiệm tóm tắt văn bản. Màn hình chính của thử nghiệm này được trình bày như sau:



Các bước thử nghiệm:

Bước 1: Chọn thư mục dữ liệu thử nghiệm; nhập hệ số phương pháp tối ưu được xác định bằng giải thuật di truyền.

Bước 2: Tóm tắt văn bản bằng phương pháp Voting Schulze kết hợp hệ số phương pháp.

Bảng kết quả cho ta thấy các độ đo ROUGE-N (1-gram, 2-gram, 3-gram, 4-gram) trung bình của toàn bộ thư mục và các văn bản trong thư mục. Kết quả văn bản tóm tắt từng văn bản gốc trong thư mục được lưu ra thư mục tomtat trong máy tính.

6. Kết quả tóm tắt thử nghiệm

Phần này trình bày một kết quả thử nghiệm của phương pháp tóm tắt VTS_FC_ACO

6.1. Văn bản gốc

Món ăn truyền thống của người dân tộc Mường.

Văn hoá của một tộc người nói chung và văn hóa Mường nói riêng không phải là cái gì đó quá bao la, rộng lớn hay khó nắm bắt. Đó là những nét riêng , độc đáo biểu hiện sinh động trong nội dung và hình thức của một số giá trị văn hoá tiêu biểu: Văn hoá ẩm thực, văn hoá trang phục, văn hoá nhà ở- kiến trúc, ngôn ngữ, lịch pháp, tín ngưỡng- tôn giáo, phong tục tập quán, đạo đức,văn học - nghệ thuật, y học cổ truyền,...

Như vậy, tìm hiểu một nét văn hoá cũng chính là đã tìm hiểu được tính cách, lối sống, lối sinh hoạt của dân tộc đó. Ở đây, tôi muốn đề cập đến một nét văn hoá vật chất của người Mường - mà khi soi vào đó, tâm hồn dân Mường, nếp sống, cách nghĩ, phong tục tập quán và truyền thống của họ hiện lên một cách tự nhiên, giản dị nhưng lại mang đậm nét bản sắc văn hoá riêng, không thể nhầm lẫn- Nét văn hoá ẩm thực.

Nói đến Ẩm thực Mường là nói tới nét văn hoá toát lên trong mỗi món ăn, thức uống, trong cách họ ăn như thế nào. Với cuộc sống thường nhật, người Mường sáng tạo ra những món ăn của riêng mình, và khi ta thưởng thức ẩm thực Mường, ta hiểu hơn về cuộc sống lao động, nếp sống bao đời nay của dân tộc này.

Người Mường thường sinh sống trong những thung lũng có triền núi đá vôi bao quanh, gần những con sông, con suối nhỏ. Họ trồng lúa trên những thửa ruộng bậc thang hay trong chân núi trũng nước, trồng ngô, khoai sắn trên các nương rẫy thấp, săn bắt hái lượm trên rừng và đánh bắt cá tôm ở lòng sông , khe suối. Cuộc sống chủ yếu dựa vào thiên nhiên; chính từ sự che chở của thiên nhiên đó, người Mương đã tồn tại cùng những món ăn, thức uống do họ tự sáng tạo ra, để rồi từ đó Văn hoá Ẩm thực Mường đã được khẳng định.

Người Mường rất thích ăn thức ăn có vị chua : củ kiệu, quả cà muối chua với cá, rau cải muối dưa, quả đu đủ muối dưa tép, rau sắn muối dưa cá, lá lồm nấu thịt trâu, thịt bò, lá bển, lá chau khao nấu cá đồng, muối thịt trâu, tiết bò ăn vào mùa nào cũng thích hợp. Đặc biệt, trong góc bếp của mỗi gia đình Mường không thể thiếu những hũ măng chua.Nguồn thức ăn quanh năm sẵn có nơi núi rừng. Măng chua có thể xào nấu với cá, thịt gà, vịt, nước măng chua kho thịt trâu, kho cá, chấm rau sống hay ngâm ớt tươi,...

Vị đắng cũng là vị mà người Mường rất yêu thích. Măng đắng; lá, hoa, quả đu đủ không chỉ là món ăn thường ngày mà còn là món để thờ phụng trong nhiều nghi lễ dân gian. Ngoài ra còn có rau đốm, lá kịa, vừa là thức ăn vừa là thuốc đau bụng. Đặc biệt, ruột và dạ dày con Don vừa là vị thuốc chữa dạ dày vừa là món ăn quý hiếm.

Gắn với vị cay, người Mường có món Ớt nổi tiếng . Ớt được băm lẫn với lòng cá; hay đầu, tiết luộc, ruột cắt nhỏ của con gà, vịt. Băm nhỏ cho tất cả lên màu nâu sẫm, cắt nhỏ vài loại rau thơm trộn vào là được món ớt. Vị ớt cay của người Mường

thường dùng để chế biến thành những món ăn riêng chứ không làm gia vị xào nấu như một số dân tộc khác.

Truyền thống của người Mường là thích bày cỗ trên lá chuối trong tất cả những bữa cỗ cộng đồng: Lễ hội, cưới xin, tang ma hoặc lễ cúng lớn trong năm. Trong mỗi dịp lễ tết, hội hè, món ăn và cách bày trí nó đều có những nét riêng, chứa đựng cả một tín ngưỡng. Với người Mường, phần ngọn và mép lá tượng trưng cho Mường Sáng- mường của người sống, phần gốc lá và mang lá tượng trưng cho Mường Tối- Mường ma, mường của người chết. Chính thế, khi dùng lá chuối bày cỗ, người Mường có quy tắc phân biệt: Người vào, ma ra. Tức là khi dọn cỗ cho người sống , phần ngọn lá hướng vào trong , phần gốc lá hướng ra ngoài, còn khi dọn cỗ cho người ma thì ngược lại. Đây là một quy tắc khá nghiêm ngặt, không thể vi phạm bởi người Mường tin rằng, sự vi phạm sẽ mang lại những điều dữ hoặc làm mất lòng khách.

Trong văn hoá ẩm thực Mường, tục uống rượu đúng ra thành một nét văn hoá riêng Văn hoá rượu cần. Rượu cần người Mường luôn phải uống tập thể, mỗi lần uống rượu cần là ta lại được hoà mìng vào những luật vui của các tuần rượu, được nghe hát dân ca Thường rang- Bộ mẹng, hát đối đáp của các bên tham gia. Có thể khẳng định rằng, văn hoá Ẩm thực Mường cũng văn hoá rượu Cần đã thể hiện được tính cộng đồng và tính huyết thống rất cao của dân tộc. Hoà Bình từ lâu đã được coi là tỉnh Mường , Văn hoá Mường góp phần rất lớn làm nên sự hấp dẫn đặc biệt cho mảnh đất giàu truyền thống văn hoá này. Đến với Hoà Bình, tìm hiểu văn hoá bản địa, không thể không đến Bảo tàng Không gian văn hoá Mường - nơi tái hiện và lưu giữ lại cả không gian sống, lối sinh hoạt, lao động sản xuất và những nét văn hoá đặc sắc của chủ nhân mảnh đất. Đến đây, chúng ta sẽ thực sự được hoà mình vào một xã hội Mường thu nhỏ, được thưởng thức ẩm thực dân gian trong khung cảnh nhà sàn, trong âm vang tiếng nhạc cồng chiêng, hoà cùng những lời ca tha thiết của các chàng trai, cô gái Mường. Về với Hoà Bình, về với bản sắc văn hoá Mường cũng chính là đã tìm về cội nguồn, với lịch sử của dân tộc

6.2. Văn bản con người tóm tắt

Nói đến Ẩm thực Mường là nói tới nét văn hoá toát lên trong mỗi món ăn, thức uống, trong cách họ ăn như thế nào.

Người Mường sáng tạo ra những món ăn của riêng mình.

Người Mường rất thích ăn thức ăn có vị chua .

Vị đắng cũng là vị mà người Mường rất yêu thích.

Gắn với vị cay, người Mường có món Ớt nổi tiếng.

Trong văn hoá ẩm thực Mường, tục uống rượu đúng ra thành một nét văn hoá riêng Văn hoá rượu cần.

6.3. Văn bản hệ thống tóm tắt

Văn hoá của một tộc người nói chung và văn hoá Mường nói riêng không phải là cái gì đó quá bao la , rộng lớn hay khó nắm bắt .

Như vậy , tìm hiểu một nét văn hoá cũng chính là đã tìm hiểu được tính cách , lối sống , lối sinh hoạt của dân tộc đó .

Nói đến Ẩm thực Mường là nói tới nét văn hoá toát lên trong mỗi món ăn , thức uống , trong cách họ ăn như thế nào .

Người Mường thường sinh sống trong những thung lũng có triền núi đá vôi bao quanh , gần những con sông , con suối nhỏ .

Người Mường rất thích ăn thức ăn có vị chua : củ kiệu , quả cà muối chua với cá , rau cải muối dưa , quả đu đủ muối dưa tép , rau sắn muối dưa cá , lá lồm nấu thịt trâu , thịt bò , lá bển , lá chau khao nấu cá đồng , muối thịt trâu , tiết bò ăn vào mùa nào cũng thích hợp .

Vị đắng cũng là vị mà người Mường rất yêu thích .

Gắn với vị cay , người Mường có món Ớt nổi tiếng .

Vị ớt cay của người Mường thường dùng để chế biến thành những món ăn riêng chứ không làm gia vị xào nấu như một số dân tộc khác .

Truyền thống của người Mường là thích bày cỗ trên lá chuối trong tất cả những bữa cỗ cộng đồng : Lễ hội , cưới xin , tang ma hoặc lễ cúng lớn trong năm .

Trong văn hoá ẩm thực Mường , tục uống rượu đúng ra thành một nét văn hoá riêng Văn hoá rượu cần .

6.4. Kết quả theo độ đo ROUGE-N

ROUGE-1 = 0,901960784313726	
1-gram tóm tắt lý tưởng	**1-gram tóm tắt hệ thống**
ăn	ăn
ẩm_thực	ẩm_thực
cách	bao
cay	bao_la
có	bày
của	bển
cũng	bò
chua	bữa
đắng	cà
đến	cá
đúng	cách
gắn	cải
họ	cái
là	cay

lên	có
mà	con
mình.	cỗ
món	cộng_đồng
mỗi	củ
một	của
mường	cũng
nào.	cúng
nét	cưới_xin
nói	chau
nổi_tiếng	chế_biến
người	chính
người_mường	chua
như	chuối
những	chứ
ớt	dân_tộc
ra	dùng
rất	dưa
riêng	đã
riêng	đá_vôi
rượu	đắng
rượu_cần	để
sáng_tạo	đến
toát	đó
tới	đồng
tục	đu_đủ
thành	đúng
thế	được
thích	gắn
thức	gần
thức_ăn	gì
trong	gia_vị
uống	hay
văn_hoá	họ
vị	hoặc
với	kiệu
yêu_thích	khác
	khao
	khó
	không
	là
	lá
	làm
	lễ

	lễ_hội lên lối lồm lớn ma mà món mỗi một một_số mùa muối mường nào năm nắm_bắt nấu nét nói nói_chung nói_riêng nổi_tiếng núi người nhỏ như như_thế_nào như_vậy những ớt phải quả quá quanh ra rau rất riêng rộng_lớn rượu rượu_cần sắn sinh_hoạt

	sinh_sống sông sống suối tang tất_cả tép tiết tìm_hiểu tính_cách toát tộc_người tới tục thành thích thích_hợp thịt thung_lũng thức thức_ăn thường trâu trên triền trong truyền_thống uống và vào văn_hoá vị với xào_nấu yêu_thích

ROUGE-2 = 0,805555555555556	
2-gram tóm tắt lý tưởng	**2-gram tóm tắt hệ thống**
ăn của ăn như ăn thức ăn thức_ăn ẩm_thực mường cách họ	ăn như_thế_nào ăn riêng ăn thức ăn thức_ăn ăn vào ẩm_thực mường

cay người	bao quanh
có món	bao_la rộng_lớn
có vị	bày cỗ
của riêng	bẩu lá
cũng là	bò ăn
chua vị	bò lá
đắng cũng	bữa cỗ
đến ẩm_thực	cá đồng
đúng ra	cá lá
gắn với	cà muối
họ ăn	cá rau
là nói	cách họ
là vị	cái gì
lên trong	cải muối
mà người	cay của
mình người_mường	cay người
món ăn	có món
món ớt	có triền
mỗi món	có vị
một nét	con sông
mường có	con suối
mường là	cỗ cộng_đồng
mường rất	cỗ trên
mường tục	cộng_đồng lễ_hội
nào. người_mường	củ kiệu
nét văn_hoá	của dân_tộc
nói đến	của một
nói tới	của người
nổi_tiếng trong	cũng chính
người mường	cũng là
người_mường rất	cúng lớn
người_mường sáng_tạo	cũng thích_hợp
như thế	cưới_xin tang
những món	chau khao
ớt nổi_tiếng	chế_biến thành
ra những	chính là
ra thành	chua củ
rất thích	chua với
rất yêu_thích	chuối trong
riêng mình.	chứ không
riêng văn_hoá	dân_tộc đó
rượu đúng	dân_tộc khác
sáng_tạo ra	dùng để
toát lên	dưa cá

tới nét	dưa quả
tục uống	dưa tép
thành một	đã tìm_hiểu
thế nào.	đá_vôi bao
thích ăn	đắng cũng
thức uống	để chế_biến
thức_ăn có	đến ẩm_thực
trong cách	đó nói
trong mỗi	đó quá
trong văn_hoá	đồng muối
uống rượu	đu_đủ muối
uống trong	đúng ra
văn_hoá ẩm_thực	được tính_cách
văn_hoá riêng	gắn với
văn_hoá rượu_cần	gần những
văn_hoá toát	gì đó
vị cay	gia_vị xào_nấu
vị chua	hay khó
vị đắng	họ ăn
vị mà	hoặc lễ
với vị	kiệu quả
yêu_thích gắn	khác truyền_thống
	khao nấu
	khó nắm_bắt
	không làm
	không phải
	lá bẻu
	là cái
	lá chau
	lá chuối
	là đã
	lá lồm
	là nói
	là thích
	là vị
	làm gia_vị
	lễ cúng
	lễ_hội cưới_xin
	lên trong
	lối sinh_hoạt
	lối sống
	lồm nấu
	lớn trong
	ma hoặc

	mà người món ăn món ớt mỗi món một nét một tộc_người một_số dân_tộc mùa nào muối chua muối dưa muối thịt mường có mường là mường nói_riêng mường rất mường tục mường thường nào cũng năm trong nắm_bắt như_vậy nấu cá nấu thịt nét văn_hoá nói đến nói tới nói_chung và nói_riêng không nổi_tiếng vị núi đá_vôi người mường nhỏ người như một_số như_thế_nào người như_vậy tìm_hiểu những bữa những con những món những thung_lũng ớt cay ớt nổi_tiếng phải là quá bao_la quả cà quả đu_đủ

	quanh_gần ra thành rau cải rau sắn rất thích rất yêu_thích riêng chứ riêng văn_hoá rộng_lớn hay rượu đúng sắn muối sinh_hoạt của sinh_sống trong sông con sống lối suối nhỏ tang ma tất_cả những tép rau tiết bò tìm_hiểu được tìm_hiểu một tính_cách lối toát lên tộc_người nói_chung tới nét tục uống thành một thành những thích ăn thích bày thích_hợp vị thịt bò thịt trâu thung_lũng có thức uống thức_ăn có thường dùng thường sinh_sống trâu tiết trâu thịt trên lá triền núi trong cách

	trong mỗi trong năm trong những trong tất_cả trong văn_hoá truyền_thống của uống rượu uống trong và văn_hoá vào mùa văn_hoá ẩm_thực văn_hoá của văn_hoá cũng văn_hoá mường văn_hoá riêng văn_hoá rượu_cần văn_hoá toát vị cay vị chua vị đắng vị mà vị ớt với cá với vị xào_nấu như yêu_thích gắn

ROUGE-3 = 0,76	
3-gram tóm tắt lý tưởng	**3-gram tóm tắt hệ thống**
ăn của riêng ăn như thế ăn thức uống ăn thức_ăn có ẩm_thực mường là ẩm_thực mường tục cách họ ăn cay người mường có món ớt có vị chua của riêng mình. cũng là vị chua vị đắng đắng cũng là đến ẩm_thực mường	ăn như_thế_nào người ăn riêng chứ ăn thức uống ăn thức_ăn có ăn vào mùa ẩm_thực mường là ẩm_thực mường tục bao quanh gần bao_la rộng_lớn hay bày cỗ trên bẻu lá chau bò ăn vào bò lá bẻu bữa cỗ cộng_đồng cá đồng muối

đúng ra thành	cá lá lồm
gắn với vị	cà muối chua
họ ăn như	cá rau cải
là nói tới	cách họ ăn
là vị mà	cái gì đó
lên trong mỗi	cải muối dưa
mà người mường	cay của người
mình. người_mường rất	cay người mường
món ăn của	có món ớt
món ăn thức	có triền núi
món ớt nổi_tiếng	có vị chua
mỗi món ăn	con sông con
một nét văn_hoá	con suối nhỏ
mường có món	cỗ cộng_đồng lễ_hội
mường là nói	cỗ trên lá
mường rất yêu_thích	cộng_đồng lễ_hội cưới_xin
mường tục uống	củ kiệu quả
nào. người_mường sáng_tạo	của dân_tộc đó
nét văn_hoá riêng	của một tộc_người
nét văn_hoá toát	của người mường
nói đến ẩm_thực	cũng chính là
nói tới nét	cũng là vị
nổi_tiếng trong văn_hoá	cúng lớn trong
người mường có	cũng thích_hợp vị
người mường rất	cưới_xin tang ma
người_mường rất thích	chau khao nấu
người_mường sáng_tạo ra	chế_biến thành những
như thế nào.	chính là đã
những món ăn	chua củ kiệu
ớt nổi_tiếng trong	chua với cá
ra những món	chuối trong tất_cả
ra thành một	chứ không làm
rất thích ăn	dân_tộc đó nói
rất yêu_thích gắn	dân_tộc khác truyền_thống
riêng mình. người_mường	dùng để chế_biến
riêng văn_hoá rượu_cần	dưa cá lá
rượu đúng ra	dưa quả đu_đủ
sáng_tạo ra những	dưa tép rau
toát lên trong	đã tìm_hiểu được
tới nét văn_hoá	đá_vôi bao quanh
tục uống rượu	đắng cũng là
thành một nét	để chế_biến thành
thế nào. người_mường	đến ẩm_thực mường
thích ăn thức_ăn	đó nói đến

thức uống trong thức_ăn có vị trong cách họ trong mỗi món trong văn_hoá ẩm_thực uống rượu đúng uống trong cách văn_hoá ẩm_thực mường văn_hoá riêng văn_hoá văn_hoá toát lên vị cay người vị chua vị vị đắng cũng vị mà người với vị cay yêu_thích gắn với	đó quá bao_la đồng muối thịt đu_đủ muối dưa đúng ra thành được tính_cách lối gắn với vị gần những con gì đó quá gia_vị xào_nấu như hay khó nắm_bắt họ ăn như_thế_nào hoặc lễ cúng kiệu quả cà khác truyền_thống của khao nấu cá khó nắm_bắt như_vậy không làm gia_vị không phải là lá bểu lá là cái gì lá chau khao lá chuối trong là đã tìm_hiểu lá lồm nấu là nói tới là thích bày là vị mà làm gia_vị xào_nấu lễ cúng lớn lễ_hội cưới_xin tang lên trong mỗi lối sinh_hoạt của lối sống lối lồm nấu thịt lớn trong năm ma hoặc lễ mà người mường món ăn riêng món ăn thức món ớt nổi_tiếng mỗi món ăn một nét văn_hoá một tộc_người nói_chung một_số dân_tộc khác

	mùa nào cũng muối chua với muối dưa cá muối dưa quả muối dưa tép muối thịt trâu mường có món mường là nói mường là thích mường nói_riêng không mường rất thích mường rất yêu_thích mường tục uống mường thường dùng mường thường sinh_sống nào cũng thích_hợp năm trong văn_hoá nắm_bắt như_vậy tìm_hiểu nấu cá đồng nấu thịt trâu nét văn_hoá cũng nét văn_hoá riêng nét văn_hoá toát nói đến ẩm_thực nói tới nét nói_chung và văn_hoá nói_riêng không phải nổi_tiếng vị ớt núi đá_vôi bao người mường có người mường là người mường rất người mường thường nhỏ người mường như một_số dân_tộc như_thế_nào người mường như_vậy tìm_hiểu một những bữa cỗ những con sông những món ăn những thung_lũng có ớt cay của ớt nổi_tiếng vị phải là cái

	quá bao_la rộng_lớn quả cà muối quả đu_đủ muối quanh gần những ra thành một rau cải muối rau sắn muối rất thích ăn rất yêu_thích gắn riêng chứ không riêng văn_hoá rượu_cần rộng_lớn hay khó rượu đúng ra sắn muối dưa sinh_hoạt của dân_tộc sinh_sống trong những sông con suối sống lối sinh_hoạt suối nhỏ người tang ma hoặc tất_cả những bữa tép rau sắn tiết bò ăn tìm_hiểu được tính_cách tìm_hiểu một nét tính_cách lối sống toát lên trong tộc_người nói_chung và tới nét văn_hoá tục uống rượu thành một nét thành những món thích ăn thức_ăn thích bày cỗ thích_hợp vị đắng thịt bò lá thịt trâu tiết thịt trâu thịt thung_lũng có triền thức uống trong thức_ăn có vị thường dùng để thường sinh_sống trong trâu tiết bò

	trâu thịt bò trên lá chuối triền núi đá_vôi trong cách họ trong mỗi món trong năm trong trong những thung_lũng trong tất_cả những trong văn_hoá ẩm_thực truyền_thống của người uống rượu đúng uống trong cách và văn_hoá mường vào mùa nào văn_hoá ẩm_thực mường văn_hoá của một văn_hoá cũng chính văn_hoá mường nói_riêng văn_hoá riêng văn_hoá văn_hoá toát lên vị cay người vị chua củ vị đắng cũng vị mà người vị ớt cay với cá rau với vị cay xào_nấu như một_số yêu_thích gắn với

ROUGE-4 = 0,702702702702703	
4-gram tóm tắt lý tưởng	**4-gram tóm tắt hệ thống**
ăn của riêng mình.	ăn như_thế_nào người mường
ăn như thế nào.	ăn riêng chứ không
ăn thức uống trong	ăn thức uống trong
ăn thức_ăn có vị	ăn thức_ăn có vị
ẩm_thực mường là nói	ăn vào mùa nào
ẩm_thực mường tục uống	ẩm_thực mường là nói
cách họ ăn như	ẩm_thực mường tục uống
cay người mường có	bao quanh gần những
có món ớt nổi_tiếng	bao_la rộng_lớn hay khó
có vị chua vị	bày cỗ trên lá
của riêng mình. người_mường	bểu lá chau khao
cũng là vị mà	bò ăn vào mùa

chua vị đắng cũng	bò lá bểu lá
đắng cũng là vị	bữa cỗ cộng_đồng lễ_hội
đến ẩm_thực mường là	cá đồng muối thịt
đúng ra thành một	cá lá lồm nấu
gắn với vị cay	cà muối chua với
họ ăn như thế	cá rau cải muối
là nói tới nét	cách họ ăn như_thế_nào
là vị mà người	cái gì đó quá
lên trong mỗi món	cải muối dưa quả
mà người mường rất	cay của người mường
mình. người_mường rất thích	cay người mường có
món ăn của riêng	có món ớt nổi_tiếng
món ăn thức uống	có triền núi đá_vôi
món ớt nổi_tiếng trong	có vị chua củ
mỗi món ăn thức	con sông con suối
một nét văn_hoá riêng	con suối nhỏ người
mường có món ớt	cỗ cộng_đồng lễ_hội cưới_xin
mường là nói tới	cỗ trên lá chuối
mường rất yêu_thích gắn	cộng_đồng lễ_hội cưới_xin tang
mường tục uống rượu	củ kiệu quả cà
nào. người_mường sáng_tạo ra	của dân_tộc đó nói
nét văn_hoá riêng văn_hoá	của một tộc_người nói_chung
nét văn_hoá toát lên	của người mường là
nói đến ẩm_thực mường	của người mường thường
nói tới nét văn_hoá	cũng chính là đã
nổi_tiếng trong văn_hoá ẩm_thực	cũng là vị mà
người mường có món	cúng lớn trong năm
người mường rất yêu_thích	cũng thích_hợp vị đắng
người_mường rất thích ăn	cưới_xin tang ma hoặc
người_mường sáng_tạo ra những	chau khao nấu cá
như thế nào. người_mường	chế_biến thành những món
những món ăn của	chính là đã tìm_hiểu
ớt nổi_tiếng trong văn_hoá	chua củ kiệu quả
ra những món ăn	chua với cá rau
ra thành một nét	chuối trong tất_cả những
rất thích ăn thức_ăn	chứ không làm gia_vị
rất yêu_thích gắn với	dân_tộc đó nói đến
riêng mình. người_mường rất	dân_tộc khác truyền_thống của
rượu đúng ra thành	dùng để chế_biến thành
sáng_tạo ra những món	dưa cá lá lồm
toát lên trong mỗi	dưa quả đu_đủ muối
tới nét văn_hoá toát	dưa tép rau sắn
tục uống rượu đúng	đã tìm_hiểu được tính_cách
thành một nét văn_hoá	đá_vôi bao quanh gần

thế nào. người_mường sáng_tạo thích ăn thức_ăn có thức uống trong cách thức_ăn có vị chua trong cách họ ăn trong mỗi món ăn trong văn_hoá ẩm_thực mường uống rượu đúng ra uống trong cách họ văn_hoá ẩm_thực mường tục văn_hoá riêng văn_hoá rượu_cần văn_hoá toát lên trong vị cay người mường vị chua vị đắng vị đắng cũng là vị mà người mường với vị cay người yêu_thích gắn với vị	đắng cũng là vị để chế_biến thành những đến ẩm_thực mường là đó nói đến ẩm_thực đó quá bao_la rộng_lớn đồng muối thịt trâu đu_đủ muối dưa tép đúng ra thành một được tính_cách lối sống gắn với vị cay gần những con sông gì đó quá bao_la gia_vị xào_nấu như một_số hay khó nắm_bắt như_vậy họ ăn như_thế_nào người hoặc lễ cúng lớn kiệu quả cà muối khác truyền_thống của người khao nấu cá đồng khó nắm_bắt như_vậy tìm_hiểu không làm gia_vị xào_nấu không phải là cái lá bểu lá chau là cái gì đó lá chau khao nấu lá chuối trong tất_cả là đã tìm_hiểu được lá lồm nấu thịt là nói tới nét là thích bày cỗ là vị mà người làm gia_vị xào_nấu như lễ cúng lớn trong lễ_hội cưới_xin tang ma lên trong mỗi món lối sinh_hoạt của dân_tộc lối sống lối sinh_hoạt lồm nấu thịt trâu lớn trong năm trong ma hoặc lễ cúng mà người mường rất món ăn riêng chứ món ăn thức uống món ớt nổi_tiếng vị

	mỗi món ăn thức một nét văn_hoá cũng một nét văn_hoá riêng một tộc_người nói_chung và một_số dân_tộc khác truyền_thống mùa nào cũng thích_hợp muối chua với cá muối dưa cá lá muối dưa quả đu_đủ muối dưa tép rau muối thịt trâu tiết mường có món ớt mường là nói tới mường là thích bày mường nói_riêng không phải mường rất thích ăn mường rất yêu_thích gắn mường tục uống rượu mường thường dùng để mường thường sinh_sống trong nào cũng thích_hợp vị năm trong văn_hoá ẩm_thực nắm_bắt như_vậy tìm_hiểu một nấu cá đồng muối nấu thịt trâu thịt nét văn_hoá cũng chính nét văn_hoá riêng văn_hoá nét văn_hoá toát lên nói đến ẩm_thực mường nói tới nét văn_hoá nói_chung và văn_hoá mường nói_riêng không phải là nổi_tiếng vị ớt cay núi đá_vôi bao quanh người mường có món người mường là thích người mường rất thích người mường rất yêu_thích người mường thường dùng người mường thường sinh_sống nhỏ người mường rất như một_số dân_tộc khác như_thế_nào người mường thường như_vậy tìm_hiểu một nét

	những bữa cỗ cộng_đồng những con sông con những món ăn riêng những thung_lũng có triền ớt cay của người ớt nổi_tiếng vị ớt phải là cái gì quá bao_la rộng_lớn hay quả cà muối chua quả đu_đủ muối dưa quanh gần những con ra thành một nét rau cải muối dưa rau sắn muối dưa rất thích ăn thức_ăn rất yêu_thích gắn với riêng chứ không làm rộng_lớn hay khó nắm_bắt rượu đúng ra thành sắn muối dưa cá sinh_hoạt của dân_tộc đó sinh_sống trong những thung_lũng sông con suối nhỏ sống lối sinh_hoạt của suối nhỏ người mường tang ma hoặc lễ tất_cả những bữa cỗ tép rau sắn muối tiết bò ăn vào tìm_hiểu được tính_cách lối tìm_hiểu một nét văn_hoá tính_cách lối sống lối toát lên trong mỗi tộc_người nói_chung và văn_hoá tới nét văn_hoá toát tục uống rượu đúng thành một nét văn_hoá thành những món ăn thích ăn thức_ăn có thích bày cỗ trên thích_hợp vị đắng cũng thịt bò lá bển thịt trâu tiết bò thịt trâu thịt bò

	thung_lũng có triền núi thức uống trong cách thức_ăn có vị chua thường dùng để chế_biến thường sinh_sống trong những trâu tiết bò ăn trâu thịt bò lá trên lá chuối trong triền núi đá_vôi bao trong cách họ ăn trong mỗi món ăn trong năm trong văn_hoá trong những thung_lũng có trong tất_cả những bữa trong văn_hoá ẩm_thực mường truyền_thống của người mường uống rượu đúng ra uống trong cách họ và văn_hoá mường nói_riêng vào mùa nào cũng văn_hoá ẩm_thực mường tục văn_hoá của một tộc_người văn_hoá cũng chính là văn_hoá mường nói_riêng không văn_hoá riêng văn_hoá rượu_cần văn_hoá toát lên trong vị cay người mường vị chua củ kiệu vị đắng cũng là vị mà người mường vị ớt cay của với cá rau cải với vị cay người xào_nấu như một_số dân_tộc yêu_thích gắn với vị

7. Một số thuật toán trong luận án

Thuật toán Voting Shulze_Method() được trình bày như sau:

```
private List<string> Schulze_Method(List<string> Data)
      {
         List<string> result = new List<string>();
         _enum = new List<string>(Data[0].Split('>'));
         _enum.RemoveAt(_enum.Count - 1);
```

```
        _enum.Sort();
        N = _enum.Count;
        int numEvaluators = rawDataN.Count;
        while (_enum.Count > 0)
        {
            int[,] defeats = MakeDefeatsN(Data, _enum.Count);
            int[,] strengths = MakePathStrengths(defeats, _enum.Count);
            bool[] winners = MakeWinners(strengths, _enum.Count);
            txtResult.Text += "== Best option(s) is: \r\n";
            string winner = "";
            string[] namewinner = _enum.ToArray();
            {
                for (int k = 0; k < winners.Length; k++)
                {
                    if (winners[k] == true)
                    {
                        winner = namewinner[k];
                        result.Add(winner);
                        _enum.Remove(winner);
                        Data = RemoveCandidateN(Data, winner);
                    }
                }
            }
        }
        return result;

    }
```

```
private int[,] MakeDefeatsN(List<string> rd, int N)
    {
        int[,] result = new int[N, N];

        for (int k = 0; k <= rd.Count - 1; k++)
        {
            string[] t = rd[k].Split('>');
            // one row of raw data
            for (int i = 0; i < t.Length - 1; i++)
            {
                for (int j = i + 1; j < t.Length - 1; j++)
                {
                    string winner = t[i];
                    string loser = t[j];

                    int w = _enum.IndexOf(t[i]);// Candidate.IndexOf(winner + ">") / 2;
```

```
                    int l = _enum.IndexOf(t[j]);
//Convert.ToInt32(System.Enum.Parse(typeof(options), loser));
                    result[w, l] += int.Parse(t[t.Length - 1]);
                }
            }
        }
        return result;
    }
    private int[,] MakePathStrengths(int[,] d, int N)
    {
        int[,] result = new int[N, N];
        for (int i = 0; i <= N - 1; i++)
        {
            for (int j = 0; j <= N - 1; j++)
            {
                if (d[i, j] > d[j, i])
                {
                    result[i, j] = d[i, j];
                }
                else
                {
                    result[i, j] = 0;
                }
            }
        }

        for (int k = 0; k <= N - 1; k++)
        {
            for (int i = 0; i <= N - 1; i++)
            {
                if (k == i)
                {
                    continue;
                }
                for (int j = 0; j <= N - 1; j++)
                {
                    if (k == j || i == j)
                    {
                        continue;
                    }
                    result[i, j] = Math.Max(result[i, j], Math.Min(result[i, k], result[k,
j]));
                }
                //j
```

```
            }
            //i
        }
        //k

        return result;
    }
```

```
private bool[] MakeWinners(int[,] ps, int N)
    {
        bool[] result = new bool[N];

        for (int i = 0; i <= N - 1; i++)
        {
            result[i] = true;
        }

        for (int i = 0; i <= N - 1; i++)
        {
            for (int j = 0; j <= N - 1; j++)
            {
                if (ps[i, j] < ps[j, i])
                {
                    result[i] = false;
                }
            }
        }
        return result;
    }
```

```
private List<string> RemoveCandidateN(List<string> rd, string R)
    {
        List<string> result = new List<string>();
        for (int k = 0; k <= rd.Count - 1; k++)
        {
            List<string> t = new List<string>(rd[k].Split('>'));
            t.Remove(R);
            string remo = "";
            for (int i = 0; i <= t.Count - 1; i++)
                remo += t[i] + ">";
            result.Add(remo.Substring(0, remo.Length - 1));
        }
        return result;
    }
```