(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **1-0026606**

(51)[2020.01] G06F 40/194

(13) **B**

(21) 1-2018-05587

(22) 11/12/2018

(45) 25/12/2020 393

(43) 25/02/2019 371A

(73) Trường Đại học Công nghệ (VN)
Nhà E3, 144 Xuân Thủy, quận Cầu Giấy, thành phố Hà Nội

(72) Võ Đình Hiếu (VN); Nguyễn Ngọc Sơn (VN); Trần Minh Tuấn (VN); Nguyễn Văn Sơn (VN).

(54) QUY TRÌNH KIỂM TRA TRÙNG LẶP TRONG NHÓM VĂN BẢN

(57) Sáng chế đề cập đến quy trình kiểm tra trùng lặp trong nhóm văn bản để chỉ ra các nội dung trùng lặp hoặc tương đồng của văn bản cần kiểm tra khi so sánh với các văn bản khác trong nhóm văn bản, quy trình này bao gồm các bước: trích xuất nội dung tài liệu của từng văn bản trong nhóm các văn bản và ghi vào kho dữ liệu có thể tìm kiếm được, trong đó nội dung tài liệu được trích xuất là các câu văn có ý nghĩa và kho dữ liệu có thể tìm kiếm được sử dụng kỹ thuật đánh chỉ mục ngược; truy vấn nội dung của văn bản cần kiểm tra và so sánh tính trùng lặp của từng câu văn có ý nghĩa của văn bản cần kiểm tra này lần lượt với các câu văn có ý nghĩa của các văn bản khác trong nhóm văn bản, trong đó việc so sánh tính trùng lặp dựa trên giá trị độ tương đồng được tính toán dựa trên n-gram (một chuỗi liên tục các đối tượng trong một câu văn); xác định các câu văn trong các văn bản là câu tương đồng khi kết quả giá trị độ tương đồng tính được lớn hơn giá trị ngưỡng tương đồng xác định trước; đánh dấu và hiển thị các nội dung tương đồng trên nền web, máy tính hoặc thiết bị di động.

BẰNG ĐỘC QUYỀN SÁNG CHẾ SỐ: 26606

Danh sách các Chủ bằng độc quyền tiếp theo:


Nhóm tài liệu
Tài liệu 1
Tài liệu 2
Tài liệu N
201.1
Chuyển sang PDF
201.2
Chuyển sang PDF
201.n
Chuyển sang PDF
202.1
Trích xuất nội dung
202.2
Trích xuất nội dung
202.n
Trích xuất nội dung
203.1
Ghi dữ liệu vào kho
203.2
Ghi dữ liệu vào kho
203.n
Ghi dữ liệu vào kho
Đồng bộ
204.1
Tìm kiếm trên kho
204.2
Tìm kiếm trên kho
204.n
Tìm kiếm trên kho
205.1
Đánh dấu, hiển thị
205.2
Đánh dấu, hiển thị
205.n
Đánh dấu, hiển thị
Kết quả 1
Kết quả 2
Kết quả N

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến quy trình kiểm tra trùng lặp trong nhóm văn bản, cụ thể hơn là quy trình kiểm tra có khả năng kiểm tra tính trùng lặp trong một nhóm văn bản tiếng Việt theo câu và theo đoạn văn.

Tình trạng kỹ thuật của sáng chế

Hiện nay ở nước ta chưa thấy có phương pháp/quy trình kiểm tra trùng lặp cho nhóm văn bản tiếng Việt được công bố. Tuy vậy, trên thế giới đã có rất nhiều phương pháp kiểm tra trùng lặp và phát hiện đạo văn. Đa số các phương pháp này vẫn còn có các nhược điểm như sẽ được phân tích và chỉ ra dựa trên một số phương pháp đã biết làm ví dụ dưới đây.

Phương pháp đã biết thứ nhất là phương pháp phát hiện đạo văn được đề cập trong tài liệu US 6,976,170 B1, có tên là "Method for detecting plagiarism" của tác giả Adam W. Kelly, đặc điểm của phương pháp này là tách mỗi văn bản thành các câu văn, sau đó sử dụng công thức tính để tính giá trị số của mỗi câu văn, sau đó so sánh giá trị số của mỗi câu với các câu văn khác, nếu hai câu trùng giá trị hoặc ở trong một khoảng định trước thì đánh dấu lại và hiển thị cho giảng viên.

Phương pháp này được thực hiện nhờ hệ thống được chia làm ba thành phần chính bao gồm: chương trình nộp bài cho sinh viên (Hand in program), chương trình xem và kiểm tra bài tập cho giảng viên (validation program) và chương trình phát hiện đạo văn (plagiarism prevention program). Trong đó, khi sinh viên nộp bài, chương trình sẽ tách văn bản thành các câu và tính giá trị số của câu đó bằng công thức dựa trên mã ASCII của các ký tự. Sau đó, khi kiểm tra trùng lặp, chương trình sẽ sắp xếp câu theo giá trị số và tìm kiếm câu văn

tương đồng bằng cách duyệt danh sách từ trên xuống. Cuối cùng hiển thị kết quả cho người dùng.

Ưu điểm

- Tốc độ kiểm tra nhanh do các câu được chuyển thành giá trị số ngay khi nộp bài và việc tính toán trùng lặp chỉ thao tác trên các giá trị số đó.

Nhược điểm

- Không đảm bảo chắc chắn phát hiện ra các trường hợp đã chỉnh sửa thêm một số từ trong câu, nhất là với các trường hợp hai câu có độ dài chênh lệch nhau nhiều hay có thêm bớt nhiều từ (nhất là các từ có mã ASCII lớn).

- Chỉ xử lý cho tài liệu chữ thuần không định dạng (tệp .txt), chưa xử lý các tài liệu PDF, DOC và việc hiển thị kết quả cho các tài liệu đó.

Phương pháp đã biết thứ hai là phương pháp phát hiện đạo văn được đề cập trong tài liệu US 2016/0196342 A1, có tên là "System, Method, and Computer-Readable Medium for Plagiarism Detection" của các tác giả Susan M. LOTTRIDGE, Monterey, CA (US); Howard C.MITZEL, Seaside, CA (US); John A. PENNINGTON, Monterey, CA (US), đặc điểm của phương pháp này là phát hiện đạo văn trong một nhóm bài trả lời bằng cách truy cập và tiền xử lý, sau đó ghép cặp và so sánh các bài trả lời với nhau, và kết quả độ tương đồng khi so sánh được coi là một tiêu chí để đánh giá khả năng đạo văn.

Phương pháp này được thực hiện thông qua các bước bao gồm: đầu tiên là truy cập vào nhóm bài tập; sau đó tiền xử lý dữ liệu đó, có thể bao gồm loại bỏ các cụm từ phổ thông và so sánh độ dài của bài tập với một ngưỡng độ dài tối thiểu được xác định trước; sau đó ghép cặp các bài trả lời với nhau và tính toán giá trị tương đồng của mỗi cặp; sau khi tính toán, kết quả độ tương đồng được so sánh với ngưỡng tương đồng đã xác định trước; và nếu kết quả tương đồng lớn hơn ngưỡng đó thì kết quả đó được xác định có thể là đạo văn.

Ưu điểm

- Phương pháp hoàn chỉnh, được mô tả rõ ràng, chi tiết, có thể cài đặt và triển khai hiệu quả trên thực tế.

- Có thể tìm kiếm tương đối cho các trường hợp chỉ chỉnh sửa một vài từ trong câu.

Nhược điểm

- Việc kiểm tra theo từng cặp một sẽ tốn nhiều tài nguyên xử lý, nhất là khi số lượng bài tập trong nhóm lớn.

- Chưa chi tiết về cách hiển thị kết quả tương đồng.

- Sử dụng độ đo đối xứng khiến việc kiểm tra giữa một câu văn dài và ngắn sẽ không chính xác, ví dụ khi câu văn thứ nhất chỉ sao chép một nửa câu thứ hai thì độ tương đồng của câu thứ nhất so với câu thứ hai là 100% (vì toàn bộ câu 1 giống câu 2) trong khi ngược lại nên là khoảng 50% (vì câu 2 chỉ có một nửa nội dung giống câu 1).

Do vậy, có nhu cầu về phương pháp kiểm tra tính trùng lặp trong văn bản tiếng Việt có khả năng khắc phục được các nhược điểm nêu trên, chẳng hạn như các vấn đề được nêu dưới đây.

- Đảm bảo được tìm kiếm tương đối, sử dụng độ đo bất đối xứng có thể phản ánh mức độ tương đồng của một câu văn so với một câu khác.

- Tập trung vào văn bản tiếng Việt, trong đó bỏ qua việc xử lý từ biến đổi hình thái trong tiếng Anh và tập trung vào việc xử lý các từ nhiều âm, cụm từ.

- Phương pháp hiển thị kết quả kiểm tra qua giao diện thân thiện với người dùng, có thể hiển thị ở dạng web và có thể tùy biến được.

- Có thể xử lý các tiêu chuẩn tài liệu TXT, DOC, PDF.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất quy trình kiểm tra trùng lặp trong nhóm văn bản có khả năng chỉ ra các nội dung trùng lặp hoặc tương đồng của văn bản cần kiểm tra khi so sánh với các văn bản khác trong nhóm văn bản, và khắc phục được các vấn đề được nêu ra trong tình trạng kỹ thuật của sáng chế.

Để đạt được mục đích nêu trên, theo một khía cạnh, sáng chế đề xuất quy trình kiểm tra trùng lặp trong nhóm văn bản để chỉ ra các nội dung trùng lặp hoặc tương đồng của văn bản cần kiểm tra khi so sánh với các văn bản khác trong nhóm văn bản, sử dụng các kỹ thuật trích xuất nội dung tài liệu, kho dữ liệu tìm kiếm và công thức tính điểm, quy trình này bao gồm các bước:

trích xuất nội dung tài liệu của từng văn bản trong nhóm các văn bản và ghi vào kho dữ liệu có thể tìm kiếm được, trong đó nội dung tài liệu được trích xuất là các câu văn có ý nghĩa và kho dữ liệu có thể tìm kiếm được sử dụng kỹ thuật đánh chỉ mục ngược;

kiểm tra và so sánh tính trùng lặp của từng câu văn của văn bản cần kiểm tra với kho dữ liệu có thể tìm kiếm được và tính giá trị độ tương đồng, trong đó việc so sánh tính trùng lặp dựa trên giá trị độ tương đồng được tính toán theo công thức:

độ tương đồng_AB = số n-gram_AB/số n-gram_A

trong đó:

n bằng 2 hoặc 3,

độ tương đồng_AB là giá trị độ tương đồng của câu A so với câu B,

số n-gram_AB là số n-gram của câu A trùng với số n-gram của câu B,

số n-gram_A là tổng số n-gram của câu A,

n-gram là một chuỗi liên tục các đối tượng trong một câu văn; và

xác định các câu văn trong các văn bản là câu tương đồng khi kết quả giá trị độ tương đồng tính được lớn hơn giá trị ngưỡng tương đồng xác định trước.

Theo một khía cạnh khác nữa, sáng chế đề xuất quy trình kiểm tra trùng lặp trong nhóm văn bản nêu trên, trong đó quy trình này còn bao gồm các bước:

chuẩn hóa các văn bản trong nhóm văn bản trước khi trích xuất nội dung tài liệu của từng văn bản trong nhóm các văn bản này, trong đó các văn bản này được chuẩn hóa theo định dạng tệp tin pdf;

trích xuất nội dung của văn bản, trong đó kết quả là danh sách các câu văn, vị trí và định dạng của từng từ trong văn bản và ảnh quét của từng trang trong văn bản;

ghi tất cả các câu văn đã được trích xuất vào kho dữ liệu có thể tìm kiếm được, bổ sung thông tin cần thiết và gán định danh để phân biệt, kết thúc quá trình này khi tất cả các văn bản đều đã được ghi vào kho;

tìm kiếm tất cả câu văn của từng văn bản với kho dữ liệu, tính giá trị điểm tương đồng, loại bỏ câu có điểm tương đồng thấp hoặc câu văn phổ thông, sau đó gộp các câu văn liền kề nhau và có chung nguồn tương đồng;

đánh dấu và hiển thị kết quả sử dụng kết quả kiểm tra tương đồng, vị trí và định dạng các từ, ảnh quét đã được trích xuất ở bước trên.

Theo một khía cạnh khác nữa, sáng chế đề xuất quy trình kiểm tra trùng lặp trong nhóm văn bản nêu trên, trong đó quy trình này còn bao gồm các bước: đánh dấu và hiển thị các nội dung tương đồng gồm (các) đoạn tương đồng và/hoặc (các) câu tương đồng, trong đó văn bản có các nội dung tương đồng này được chuyển từ định dạng tệp tin pdf thành định dạng ảnh, tô màu lên hình ảnh tại vị trí có các nội dung tương đồng dựa trên các thông tin vị trí, hình dạng của các từ trong văn bản, bổ sung thêm thông tin và hiển thị.

Theo một khía cạnh khác nữa, sáng chế đề xuất quy trình kiểm tra trùng lặp trong nhóm văn bản nêu trên, trong đó quy trình này còn bao gồm bước loại bỏ các câu tương đồng có nội dung là các câu văn phổ thông, trong đó câu văn phổ thông được tổng hợp thành kho dữ liệu phổ thông, các câu tương đồng có tính trùng lặp cao với các câu văn trong kho dữ liệu phổ thông này được loại bỏ để tránh việc hiển thị các nội dung trùng lặp không cần thiết.

Mô tả vắn tắt các hình vẽ

Hình 1 là lưu đồ thể hiện các bước chính của quy trình kiểm tra trùng lặp trong nhóm văn bản theo một phương án thực hiện sáng chế;

Hình2 là lưu đồ thể hiện các bước chính của quy trình kiểm tra trùng lặp trong nhóm văn bản được chia thành các giai đoạn có thể chạy song song, đa luồng theo một phương án thực hiện sáng chế;

Hình 3 là bảng chỉ mục minh họa một ví dụ về cơ sở dữ liệu có thể tìm kiếm được sử dụng kỹ thuật đánh chỉ mục ngược; và

Hình 4 là sơ đồ khối minh họa một ví dụ về xác định tính trùng lặp của hai câu văn dựa trên tính toán tỉ lệ n-gram trùng nhau của hai câu văn đó, trong đó n-gram là một chuỗi liên tục các đối tượng trong một câu văn và n có giá trị bằng 2 hoặc 3.

Mô tả chi tiết sáng chế

Dưới đây, các ưu điểm và nguyên lý cơ bản của sáng chế sẽ được hiểu rõ hơn thông qua phần mô tả chi tiết các phương án thực hiện có dựa vào các hình vẽ kèm theo. Cần hiểu rằng các phương án này chỉ được mô tả với mục đích làm ví dụ minh họa mà không làm giới hạn phạm vi của sáng chế.

Như được thể hiện trên Hình 1, quy trình kiểm tra trùng lặp trong nhóm văn bản theo sáng chế bao gồm các bước chính từ bước 101 đến bước 105, trong đó bước 101 thực hiện việc chuyển đổi tài liệu là các tệp tin văn bản ở định dạng bất kỳ sang định dạng pdf, nếu tài liệu là tệp tin văn bản ở định dạng pdf thì bước 101 này có thể được bỏ qua và chuyển tới bước 102 thực hiện trích xuất nội dung của văn bản từ tệp tin có định dạng pdf để lấy các câu văn có ý nghĩa và chuyển sang bước 103 thực hiện ghi các nội dung được trích xuất này vào kho, ở đây khái niệm kho được hiểu theo nghĩa rộng, kho có thể là một phần của cơ sở dữ liệu nhất định, kho cũng có thể là toàn bộ cơ sở dữ liệu hoặc thậm chí

kho có thể là tập hợp của nhiều cơ sở dữ liệu được kết nối mạng miễn là có thể trả về kết quả khi tìm kiếm hoặc truy vấn, theo đó bước 104 thực hiện việc tìm kiếm dữ liệu trong kho cũng không bị giới hạn ở một phần cơ sở dữ liệu, toàn bộ cơ sở dữ liệu hay tập hợp của nhiều cơ sở dữ liệu miễn là có thể quản lý được việc ghi và truy xuất các dữ liệu này để thực hiện chức năng tìm kiếm, bước cuối cùng là bước 105 thực hiện việc hiển thị các nội dung có sự tương đồng để người dùng có thể xác định được chính xác các nội dung nào bị trùng lặp/tương đồng và các nội dung này thuộc các tài liệu nào, v.v..

Trên Hình 2 thể hiện các bước chính của quy trình kiểm tra trùng lặp trong nhóm văn bản được chia thành các giai đoạn có khả năng chạy song song, đa luồng với nhau, với ràng buộc phải đồng bộ với nhau sau khi hoàn thành bước 203.1 đến 203.n là ghi dữ liệu của các văn bản khác nhau được xử lý đa luồng vào kho. Các bước từ 201.i đến 205.i (i=1,n) có nội dung tương tự các bước từ 101 đến 105 được mô tả trên đây sẽ được bỏ qua việc mô tả, thay vào đó chi tiết của từng giai đoạn trong các bước xử lý sẽ được mô tả rõ hơn dưới đây.

Chuẩn hóa định dạng tài liệu

Các tài liệu văn bản theo định dạng DOC, DOCX hoặc tương tự được chuyển thành định dạng PDF trước khi thực hiện việc trích xuất nội dung các văn bản này. Lý do là với tài liệu theo định dạng DOC, DOCX thường bị thay đổi cấu trúc khi xem ở trên nhiều máy sử dụng nhiều hệ điều hành khác nhau. Trong khi đó cấu trúc tài liệu PDF không thay đổi và luôn đồng nhất. Do đó sử dụng tài liệu định dạng PDF sẽ thuận tiện trong việc đánh dấu và hiển thị kết quả bởi việc đánh dấu các câu sao chép và lỗi chính tả yêu cầu chính xác và nhất quán tọa độ và kích thước từng từ, từng câu trong văn bản.

Việc chuyển định dạng tài liệu từ DOC, DOCX sang PDF có thể sử dụng phương pháp bất kỳ đã biết, chẳng hạn như sử dụng thư viện mã nguồn mở có

sẵn. Ngoài ra, các tài liệu định dạng khác như TXT, EPUB,... cũng có thể được xử lý và chuyển sang định dạng PDF.

Trích xuất nội dung tài liệu

Tệp tin PDF được đọc lần lượt theo thứ tự từ trên xuống từ trái qua phải tùy theo cấu trúc của văn bản. Do tệp tin PDF lưu trữ dữ liệu văn bản theo từng ký tự độc lập chứ không phải theo từng từ nên việc xử lý đầu tiên phải ghép các ký tự đơn lẻ sát nhau thành một từ bằng cách xác định vị trí, khoảng cách của các từ với nhau. Song song với quá trình gộp từ, quá trình tách câu cũng được thực hiện mỗi lúc một từ hoàn chỉnh được tạo. Dấu hiệu nhận biết từ ở đây là các kí tự ngắt câu, bao gồm dấu chấm, dấu hỏi chấm, dấu chấm than. Kết thúc quá trình, kết quả là một tệp định dạng DOM (Document Object Model) có cấu trúc dạng cây, trong đó tất cả mỗi từ đều được gán số định danh (id) duy nhất và theo thứ tự từ trên xuống, trái sang phải để phân biệt. Do đó, khi đánh dấu chỉ cần biết id của từ bắt đầu và từ kết thúc sau đó truy vấn theo id trong tệp tin DOM đã sinh ra để xác định vùng cần đánh dấu.

Tệp tin DOM được sinh ra chỉ bao gồm các thông tin về vị trí, kích thước và nội dung để đánh dấu. Tuy nhiên, trong thực tế, việc chỉ sử dụng các thông tin trên để hiển thị kết quả sẽ không đảm bảo về tính chính xác và thẩm mỹ. Hiển thị kết quả có thể coi là bôi màu trên tài liệu cũ và đính kèm thông tin về kết quả trùng lặp và lỗi chính tả. Do đó, việc hiển thị cần đảm bảo từ phông chữ, màu chữ cho đến hình ảnh, bảng biểu phải chính xác như văn bản gốc. Việc hiển thị trên web dựa trên thông tin vị trí, kích thước của từ thường mất nhiều thời gian và có thể không được chính xác tuyệt đối bởi cấu trúc tệp PDF khá phức tạp. Một cách khác là đánh dấu trực tiếp trên tệp tin PDF gốc nhưng lại khó tùy biến và bị giới hạn vì chỉ được sử dụng được các thành phần (component) trong tệp tin PDF. Do đó, việc hiển thị tốt nhất là trên trình duyệt web để có thể tùy biến và mở rộng, theo đó thì cần chuyển tệp tin PDF thành định dạng ảnh (PNG

hoặc JPG). Khi đó việc hiển thị kết quả là tô màu lên hình ảnh và thêm một số thông tin bổ sung. Kết hợp sử dụng các dữ liệu vị trí, kích thước từ tệp tin DOM và nội dung được chuyển thành ảnh thì các yếu tố cần thiết cho việc hiển thị và đánh dấu kết quả đã đủ.

Tiếp theo là lấy danh sách các câu để kiểm tra. Lý do sử dụng câu làm đơn vị để kiểm tra là trong tiếng Việt, một câu là một đơn vị có ý nghĩa hoàn chỉnh. Thông thường, khi sao chép thì thường sao chép cả câu văn hoặc một ý lớn trong đó chứ không chỉ sao chép một số từ trong đó. Đơn vị từ hoặc cụm từ thì nhỏ và không biểu đạt được nhiều ý nghĩa trong khi đơn vị là cả đoạn văn thì lại quá lớn. Các câu được lấy bằng cách tìm trong tệp tin DOM được xuất ở trên. Mỗi câu ở đầu ra được gán id tương ứng với id thẻ, kèm các thông tin về trang, nội dung. Các câu cũng bao gồm thông tin id của từ bắt đầu và kết thúc để dựa vào đó xác định hai câu có liên tiếp nhau hay không.

Ghi dữ liệu vào kho

Sau khi đã trích xuất nội dung tài liệu và tách thành câu, việc cần làm tiếp theo là đánh chỉ mục (index) tất cả các câu được sinh ra vào một kho lưu trữ chung cho cả nhóm tài liệu. Kho lưu trữ cần phải đảm bảo một số điều kiện dưới đây.

Một là kho lưu trữ này hỗ trợ tìm kiếm tương đối (full text search) và có thể xếp hạng kết quả (ranking) được. Tức là khi tìm kiếm một câu văn trên kho lưu trữ này, kết quả trả về phải là danh sách các câu văn (nếu có) có nội dung gần tương đồng với câu được đem đi tìm kiếm và kết quả được xếp hạng theo mức độ tương đồng theo một thang đo tương đồng nhất định.

Hai là phải lưu trữ được trên ổ cứng vì dữ liệu tài liệu có thể rất lớn, lưu trữ trên RAM sẽ tốn rất nhiều tài nguyên và có thể không đủ tài nguyên cho xử lý nhóm hàng trăm tài liệu. Thêm nữa là lưu trữ trên ổ cứng thì có thể tái sử dụng sau này được, ví dụ như khi muốn so sánh bài tập của nhóm sinh viên năm

nay với sinh viên năm trước đó thì không phải tìm và ghi lại dữ liệu sinh viên năm trước nữa.

Ba là kho lưu trữ phải cho phép dễ dàng thêm dữ liệu vào kho bất kỳ khi nào. Lý do là do một nhóm tài liệu có thể bao gồm đến hàng trăm. nghìn tài liệu và mỗi tài liệu có thể dài trăm, nghìn trang. Tài nguyên của hệ thống, bao gồm là CPU và RAM luôn bị giới hạn chỉ xử lý được đồng thời một số lượng dữ liệu nhất định. Vì vậy nên lượng dữ liệu có thể được chia nhỏ và được đánh chỉ mục (index) lần lượt vào trong câu. Ví dụ như trong một thời điểm chỉ xử lý và đánh chỉ mục (index) hai văn bản liền một lúc.

Thư viện tìm kiếm thích hợp bất kỳ đã biết trong lĩnh vực tương ứng có thể được áp dụng, chẳng hạn như Lucene là thư viện tìm kiếm được lựa chọn thỏa mãn các điều kiện nêu trên. Lucene là một thư viện tìm kiếm mã nguồn mở miễn phí được viết bằng Java và ban hành dưới giấy phép phần mềm Apache. Lucene hỗ trợ tìm kiếm tương đối và xếp hạng kết quả, lưu trữ dữ liệu trên ổ cứng và thêm bớt dữ liệu dễ dàng. Tuy nhiên, để sử dụng hiệu quả hơn, một số thành phần trong Lucene có thể được tùy biến, chẳng hạn như trong trường hợp kiểm tra nhóm thì có thể tùy biến cách lưu trữ và đánh chỉ số dữ liệu để phù hợp với cách tính độ tương đồng của hệ thống.

Độ tương đồng của hai câu văn bất kì được tính bằng cách tính toán tỉ lệ n-gram trùng nhau của hai câu văn đó. Khái niệm "n-gram" được hiểu là một chuỗi liên tục các đối tượng trong một đoạn văn bản. Các đối tượng ở đây có thể là ký tự, tiếng hoặc từ. Việc sử dụng 2-gram và 3-gram cho việc tìm kiếm cho ra kết quả phù hợp nhất. Dựa trên n-gram, độ tương đồng được tính như sau:

$$\text{Độ } tương \text{ đồng } của\ A\ so\ với\ B = \frac{Số\ n\ gram\ của\ A\ trùng\ với\ n\ gram\ của\ B}{Tổng\ số\ n\ gram\ của\ A}$$

Trong đó, n bằng 2 hoặc 3, tức là sử dụng đồng thời 2-gram và 3-gram để tính. Lưu ý, công thức tính độ tương đồng trên là bất đối xứng, tức là độ tương đồng của câu A so với câu B khác với độ tương đồng của câu B so với câu A. Thêm nữa là vì tử số luôn nhỏ hơn hoặc bằng mẫu số nên độ tương đồng có giá trị dao động trong khoảng từ 0 đến 1, trong đó giá trị bằng 1 tương đương với toàn bộ nội dung của câu A trùng khớp với một phần hay toàn bộ nội dung của câu B (câu B có thể có dài hơn câu A).

Do sử dụng độ đo tương đồng riêng nên việc đánh chỉ số và tìm kiếm sử dụng thư viện Lucene có sẵn tốt hơn là được tùy chỉnh một số thành phần để phù hợp, đặc biệt là cách ghi dữ liệu vào cơ sở dữ liệu của Lucene. Kỹ thuật đánh chỉ mục ngược (inverted index) được sử dụng làm cách thức lưu trữ và hoạt động của Lucene. Theo đó, thay vì lưu trữ một câu bao gồm những từ nào thì sẽ lưu trữ ngược lại, tức là một từ bao gồm những câu văn nào chứa từ đó. Kỹ thuật đánh chỉ mục ngược được minh họa trên Hình 3 và sẽ được mô tả dưới đây.

Mặc định, nhiều hệ thống tìm kiếm tương đối văn bản sẽ lấy mỗi token là một từ được tách bởi dấu cách và sử dụng độ đo cosin hoặc tf-idf để tính điểm tương đồng. Do đó cần phải thay đổi việc tách token và tính điểm bên trong kho dữ liệu tìm kiếm.

Việc tách thành các "token" từ một "văn bản" theo thư viện Lucene được thông qua các bộ phân tích (hay analyzer) của thư viện Lucene. Bộ phân tích này bao gồm các bộ tách từ (tokenizer) có trách nhiệm tách một văn bản ra thành nhiều token và các bộ lọc (filter) có trách nhiệm lọc và hậu xử lý các token được xuất ra từ bộ tách từ hoặc bộ lọc khác. Theo đó, để áp dụng công thức tính độ tương đồng sử dụng n-gram nêu trên, các bộ tách từ sẽ tách một câu thành các từ bởi dấu cách và bỏ qua dấu câu, và các bộ lọc sẽ được tùy biến để thực hiện chức năng gộp các từ được tách bởi bộ tách từ nêu trên thành 2-gram và 3-gram.

Sau khi tùy biến các thành phần bên trong thư viện Lucene để tạo thành cơ sở dữ liệu phù hợp thì tất cả các câu sẽ được ghi vào trong kho. Với mỗi câu, dữ liệu được ghi vào bao gồm: nội dung câu, tiêu đề (hoặc tên tệp), id của câu, id của tài liệu. Các trường nội dung câu, tiêu đề, id của câu nhằm phục vụ cho quá trình tổng hợp lại và đánh dấu kết quả. Còn id của tài liệu để phân biệt tài liệu, loại bỏ trường hợp câu được tìm kiếm có chung tài liệu với kết quả trả về. Riêng trường nội dung câu là được xử lý qua các bộ phân tích để tách thành ký tự, còn các trường khác chỉ để lưu trữ và đính kèm trong kết quả trả về.

Quá trình ghi dữ liệu vào kho kết thúc khi tất cả tài liệu được đánh chỉ số vào trong kho. Quá trình ghi dữ liệu được thực hiện song song đa luồng, tức là việc xử lý ghi một tài liệu được thực hiện trong một luồng và có thể xử lý đồng thời nhiều tài liệu liền một lúc. Tuy nhiên, bắt buộc tất cả tài liệu được kiểm tra một lúc thì mới có thể chuyển sang giai đoạn tiếp theo.

Tìm kiếm dữ liệu trong kho và lọc kết quả

Tiếp theo, các tài liệu sẽ được kiểm tra song song với kho dữ liệu đã được ghi ở bước trước. Mỗi câu trong tài liệu được kiểm tra lần lượt và tuần tự với dữ liệu trong kho. Với mỗi câu truy vấn tìm kiếm thì kho dữ liệu sẽ trả về danh sách kết quả tìm kiếm. Mỗi kết quả bao gồm nội dung câu, id của câu, id của tài liệu, tiêu đề và kèm theo điểm tương đồng. Trong danh sách kết quả hệ thống sẽ lọc bỏ các kết quả mà có trùng tài liệu (có id tài liệu trùng nhau) và có điểm số thấp hơn ngưỡng tương đồng đã quy định trước (các trường hợp này được quy là không trùng lặp).

Ngoài ra, để loại bỏ các câu văn phổ thông và không thuộc vào nội dung chính như tiêu đề, lời cảm ơn thì sẽ dựa trên việc tìm kiếm với một kho dữ liệu phổ thông. Trong đó, kho dữ liệu này chứa tất cả các lời cảm ơn cũng như các câu văn phổ thông khác đã được tổng hợp trước đó. Các câu văn tìm thấy được độ tương đồng cao với các câu trong kho dữ liệu phổ thông này sẽ được đánh giá là không tương đồng.

Sau khi đã có được danh sách câu được đánh giá là tương đồng, bước tiếp theo là tổng hợp lại các câu tương đồng thành các đoạn tương đồng. Các đoạn tương đồng là tập hợp các câu tương đồng liền kề nhau và có chung nguồn tương đồng. Việc xác định hai câu có liền kề hay không dựa trên id của hai câu. Sau khi có các đoạn tương đồng, một số đoạn tương đồng có độ dài rất ngắn và đứng độc lập sẽ bị loại bỏ.

Đánh dấu và hiển thị kết quả

Việc đánh dấu và hiển thị kết quả sử dụng kết hợp các dữ liệu được tạo ra ở bước trên bao gồm: dữ liệu các câu trùng lặp, các lỗi chính tả, ảnh được xuất từ PDF và dữ liệu vị trí, kích thước của các từ. Kết quả đầu ra là tệp HTML có khả năng hiển thị ngay trên web của người dùng. Trong đó, với mỗi trang kết quả, nền là ảnh được xuất từ PDF, các phần đánh dấu là các đối tượng có thể nhấn (click) và hiển thị thông tin nguồn tương đồng. Việc đánh dấu mỗi câu văn dựa trên vị trí, hình dạng của tất cả các từ có trong câu đó và nối với nhau. Việc hiển thị dữ liệu không giới hạn và ràng buộc phải là HTML và hiển thị trên web, với dữ liệu đã sẵn có ở trên, hoàn toàn có thể hiển thị bằng nhiều cách khác như trên máy tính, thiết bị di động.

Ví dụ minh họa

Tiếp theo, một ví dụ về xác định tính trùng lặp của hai câu văn dựa trên tính toán tỉ lệ n-gram trùng nhau của hai câu văn đó sẽ được mô tả dựa trên Hình 4.Đặt vào tình huống, khi tìm kiếm sử dụng từ khóa "Tôi là sinh viên đại học" thì kết quả tìm kiếm sẽ phải trả về danh sách kết quả như "Tôi chính là sinh viên đại học", 'Tôi là một sinh viên đại học", "Tôi là sinh viên cao đẳng" hay "Tôi là sinh viên" (các câu này đã được đánh chỉ mục sẵn trước đó). Các kết quả trả về phải xếp hạng được, ví dụ như câu "Tôi là một sinh viên đại học" và "Tôi chính là sinh viên đại học" phải được xếp hạng cao hơn các câu "Tôi là sinh viên" hay "Tôi là sinh viên cao đẳng".

Đối với câu văn "Tôi là sinh viên" khi tách thành 2-gram, với mỗi gram tương ứng với một tiếng thì sẽ bao gồm "Tôi là", "là sinh", "sinh viên" còn khi tách thành 3-gram thì sẽ là "Tôi là sinh", "là sinh viên". So với việc tìm kiếm dựa trên các từ đơn lẻ thì n-gram cho hiệu quả hơn ở chỗ các từ trùng lặp mà liền nhau có độ tương đồng cao hơn là các từ trùng lặp mà ở vị trí rời rạc với nhau. Ví dụ như khi so sánh câu "Tôi là một sinh viên" với câu "Tôi đợi một học sinh đi ra công viên". Rõ ràng là hai câu này không thể đánh giá là tương đồng được. Câu thứ nhất có năm tiếng và có tận 4 tiếng cũng xuất hiện ở câu thứ hai nhưng các tiếng đó không ở gần nhau. Khi tách cả 2 câu thành dạng 2-gram thì không có bất cứ 2-gram nào trùng nhau cả. Thêm một ví dụ nữa là hai câu "Tôi là một sinh viên" với câu "Tôi là một cậu sinh viên đại học năm nhất". Có thể nói câu thứ nhất có tương đồng với câu thứ hai và ở đây số 2-gram trùng nhau là 3 trong tổng số 4 các 2-gram của câu thứ nhất. Do đó có thể thấy rằng sử dụng n-gram là cách đơn giản để đánh giá mức độ tương đồng.

Ví dụ về kỹ thuật đánh chỉ mục ngược được minh họa trên Hình 3, thay vì lưu trữ câu "Tôi là sinh viên" có id = 123 bao gồm các từ "Tôi", "là", "sinh", "viên" thì sẽ lưu trữ theo cách từ "Tôi" có trong các câu văn có id là 123, 234,.... Khi tìm kiếm từ khóa "Tôi là sinh viên" thì đầu tiên sẽ tách câu văn thành từ "Tôi", "là", "sinh", "viên". Sau đó, theo cách lưu trữ đánh chỉ mục ngược trên thì sẽ lấy được tất cả id của các câu có từ "Tôi", tất cả id của các câu có từ "là",... Cuối cùng, bằng một độ đo tương đồng thì sẽ lấy id có điểm cao nhất (ví dụ như id xuất hiện nhiều nhất) và trả về kết quả.

Hiệu quả có thể đạt được bởi sáng chế

Rõ ràng là với các đặc điểm được mô tả trên đây, quy trình kiểm tra trùng lặp trong nhóm văn bản theo sáng chế có thể mang lại các hiệu quả dưới đây.

(i) Về hiệu năng, quy trình theo sáng chế có khả năng song song hóa và chạy đa luồng độc lập với nhau chỉ với ràng buộc duy nhất là ở sau bước thứ ba ghi dữ liệu phải hoàn thành tất cả tài liệu thì mới chuyển sang bước kế tiếp. Do

đó, một hệ thống sử dụng quy trình có thể dưới dạng cụm (cluster, bao gồm nhiều máy tính khác nhau) để tăng tốc độ xử lý. Một ưu điểm nữa ở hiệu năng khi so sánh với phương pháp kiểm tra theo ghép cặp tài liệu thì một tài liệu chỉ cần được xử lý một lần duy nhất ở mỗi bước và không phải xử lý thêm về sau.

(ii) Độ đo tương đồng là bất đối xứng dựa trên mức độ tương đồng về nội dung của câu kiểm tra với các câu khác trong kho dữ liệu. Do đó đánh giá tốt hơn về mức độ tương đồng khi hai câu có độ dài chênh lệch nhiều. Sử dụng n-gram có hiệu quả tốt khi kiểm tra văn bản tiếng Việt.

(iii) Việc đánh dấu và hiển thị kết quả trực quan và nhất quán do sử dụng tài liệu PDF là mục tiêu xử lý chính, đánh dấu dựa trên vị trí, hình dạng của từng từ trong câu.

(iv) Không giới hạn định dạng tài liệu được hỗ trợ, chỉ cần tài liệu đó có khả năng chuyển sang PDF (là định dạng chuẩn phổ biến) thì hoàn toàn có thể kiểm tra và hiển thị kết quả được.

Yêu cầu bảo hộ

1. Quy trình kiểm tra trùng lặp trong nhóm văn bản để chỉ ra các nội dung trùng lặp hoặc tương đồng của văn bản cần kiểm tra khi so sánh với các văn bản khác trong nhóm văn bản, sử dụng các kỹ thuật trích xuất nội dung tài liệu, kho dữ liệu tìm kiếm và công thức tính điểm, quy trình này bao gồm các bước:

chuẩn hóa các văn bản trong nhóm văn bản trước khi trích xuất nội dung tài liệu của từng văn bản trong nhóm các văn bản này, trong đó các văn bản này được chuẩn hóa theo định dạng tệp tin pdf;

trích xuất nội dung tài liệu của từng văn bản trong nhóm các văn bản và ghi vào kho dữ liệu có thể tìm kiếm được, trong đó nội dung tài liệu được trích xuất là các câu văn có ý nghĩa và kho dữ liệu có thể tìm kiếm được sử dụng kỹ thuật đánh chỉ mục ngược;

ghi tất cả các câu văn đã được trích xuất vào kho dữ liệu có thể tìm kiếm được, bổ sung thông tin cần thiết và gán định danh để phân biệt, và kết thúc quá trình này khi tất cả các văn bản đều đã được ghi vào kho dữ liệu;

tính giá trị độ tương đồng, trong đó việc so sánh tính trùng lặp dựa trên giá trị độ tương đồng được tính toán theo công thức:

độ tương đồng_AB = số n-gram_AB/số n-gram_A

trong đó:

n bằng 2 hoặc 3,

độ tương đồng_AB là giá trị độ tương đồng của câu A so với câu B,

số n-gram_AB là số n-gram của câu A trùng với số n-gram của câu B,

số n-gram_A là tổng số n-gram của câu A,

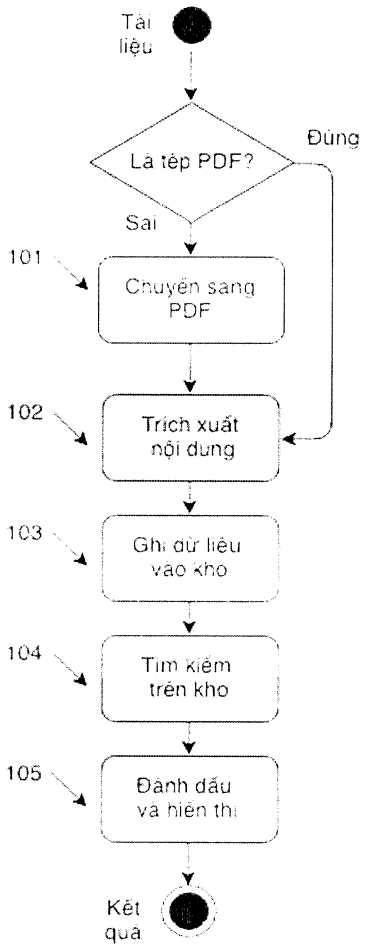
n-gram là một chuỗi liên tục các đối tượng trong một câu văn; và

tìm kiếm tất cả câu văn của từng văn bản với kho dữ liệu, tính lại giá trị điểm tương đồng, kiểm tra và so sánh tính trùng lặp của từng câu văn của văn bản cần kiểm tra với kho dữ liệu có thể tìm kiếm được và

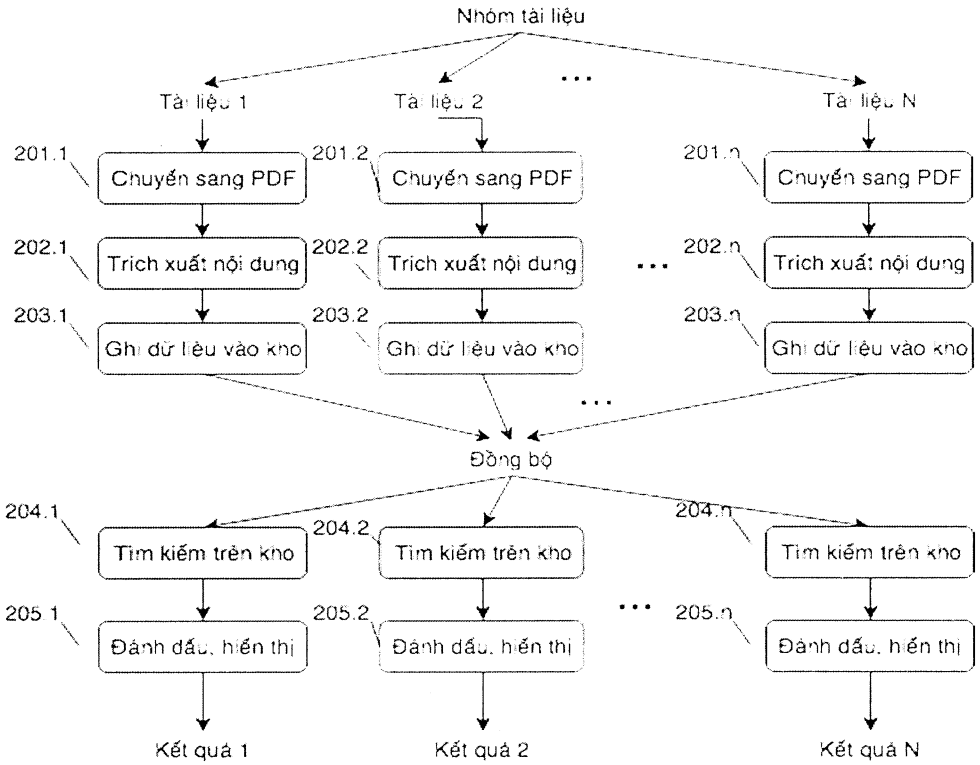
xác định các câu văn trong các văn bản là câu tương đồng khi kết quả giá trị độ tương đồng tính được lớn hơn giá trị ngưỡng tương đồng xác định trước và

loại bỏ các câu tương đồng có nội dung là các câu văn phổ thông, trong đó câu văn phổ thông được tổng hợp thành kho dữ liệu phổ thông đã được tổng hợp trước đó, các câu tương đồng có tính trùng lặp cao với các câu văn trong kho dữ liệu phổ thông này được loại bỏ khỏi kết quả;

đánh dấu và hiển thị các nội dung tương đồng gồm (các) đoạn tương đồng và/hoặc (các) câu tương đồng, trong đó văn bản có các nội dung tương đồng này được chuyển từ định dạng tệp tin pdf thành định dạng ảnh, tô màu lên hình ảnh tại vị trí có các nội dung tương đồng dựa trên các thông tin vị trí, hình dạng của các từ trong văn bản, bổ sung thêm thông tin và hiển thị.


Tài liệu
Là tệp PDF?
Đúng
Sai
101
Chuyển sang PDF
102
Trích xuất nội dung
103
Ghi dữ liệu vào kho
104
Tìm kiếm trên kho
105
Đánh dấu và hiển thị
Kết quả

Hình 1

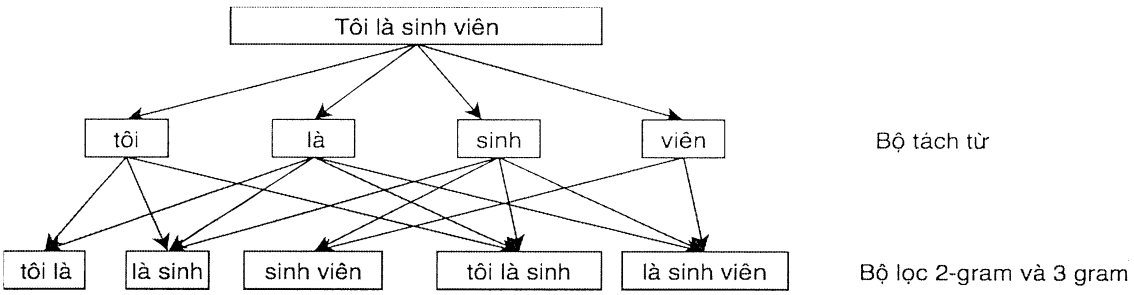

Nhóm tài liệu
Tài liệu 1
Tài liệu 2
Tài liệu N
201.1
Chuyển sang PDF
201.2
Chuyển sang PDF
201.n
Chuyển sang PDF
202.1
Trích xuất nội dung
202.2
Trích xuất nội dung
202.n
Trích xuất nội dung
203.1
Ghi dữ liệu vào kho
203.2
Ghi dữ liệu vào kho
203.n
Ghi dữ liệu vào kho
Đồng bộ
204.1
Tìm kiếm trên kho
204.2
Tìm kiếm trên kho
204.n
Tìm kiếm trên kho
205.1
Đánh dấu, hiển thị
205.2
Đánh dấu, hiển thị
205.n
Đánh dấu, hiển thị
Kết quả 1
Kết quả 2
Kết quả N

Hình 2

Từ	Câu
tôi	[1, 2, 4]
là	[1, 2, 3]
sinh	[1, 2, 3]
viên	[1, 2, 3]
bạn	[2, 5]
đại	[3, 4, 5]
học	[3, 4, 5]
ấy	[5]

ID	Câu
1	Tôi là sinh viên
2	Bạn là sinh viên
3	Tôi là sinh viên đại học
4	Tôi học đại học
5	Bạn ấy học đại học

Hình 3



Hình 4