HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

HÀ HOÀI NAM

XÂY DỰNG CHỨC NĂNG TRA CỨU THÔNG TIN VĂN BẢN DỰA TRÊN WEB NGỮ NGHĨA CỦA HỆ THỐNG TIC-OFFICE

Chuyên ngành: Hệ Thống thông tin

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

(Theo định hướng ứng dụng)

TP. HỒ CHÍ MINH - NĂM 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS. TS Đỗ Văn Nhơn**

Phản biện 1: **PGS.TS. Trần Vĩnh Phước**

Phản biện 2: **PGS.TS. Lê Hoàng Thái**

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông
Vào lúc: 09 giờ 30 ngày 02 tháng 07 năm 2022

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Ngày nay cùng với sự phát triển của internet thì dữ liệu của ngành công nghệ thông tin ngày càng gia tăng. Nhu cầu quản lý, chia sẻ, tìm kiếm thông tin trong ngành này cũng được đặt ra và đáp ứng một phần nhờ các công cụ tìm kiếm. Một số công cụ tìm kiếm nổi tiếng hiện nay như Google hay Yahoo đều có thể cho phép người dùng tìm kiếm dữ liệu có liên quan bằng cách nhập từ khóa và tìm những tài liệu có chứa từ khóa đó. Với các hệ thống tìm kiếm này phần lớn vẫn dựa trên từ khóa và mức độ phổ biến của tài liệu. Một danh sách các từ khóa là dạng biểu diễn sơ lược nhất của nội dung, nghĩa là mỗi tài liệu được biểu diễn bởi một tập từ hay cụm từ được rút trích từ chính nội dung của tài liệu và do đó, cách biểu diễn này mang mức độ thông tin còn thấp. Do đó hệ thống tìm kiếm này có kết quả trả về không phải lúc nào cũng thỏa mãn yêu cầu tìm kiếm của người sử dụng, như là độ chính xác không cao khi kết quả trả về quá nhiều mà tỷ lệ số tài liệu hữu ích trên tổng số tài liệu trả về thấp, hoặc có thể không tìm thấy được những tài liệu liên quan khi chúng được mô tả với những từ khóa khác đồng nghĩa hoặc gần nghĩa với từ khóa mà người dùng tìm kiếm (độ bao phủ không cao) gây ra không ít khó khăn cho người sử dụng trong việc tìm kiếm chính xác thông tin cần tìm kiếm.

Xuất phát từ nhu cầu thực tế của hệ thống quản lý văn bản của Hội nông dân tỉnh Tây Ninh cùng với sự hướng dẫn tận tình của Thầy PGS.TS Đỗ Văn Nhơn, tôi quyết định chọn đề tài: "**Xây Dựng Chức Năng Tra Cứu Thông Tin Văn Bản Dựa Trên Web Ngữ Nghĩa Của Hệ Thống Tic-Office**" làm luận văn tốt nghiệp.

Nội dung của luận văn được trình bày trong 5 chương, bao gồm:

Chương 1: Giới thiệu và khảo sát các hệ thống tìm kiếm thông tin, phân tích đánh giá thực trạng, trình bày mục tiêu, giới hạn của đề tài, ý nghĩa lý luận và thực tiễn, phương pháp nghiên cứu, hướng tiếp cận giải quyết vấn đề và nội dung thực hiện của đề tài.

Chương 2: Trình bày cơ sở lý thuyết của đề tài liên quan đến vấn đề truy hồi thông tin bao mô tả cấu trúc, các phương pháp truy hồi thông tin và đánh giá hệ thống truy hồi thông tin. Các lý thuyết nền tảng về mô hình không gian vector Ontology cùng với các phương pháp xây dựng mô hình dữ liệu.

Chương 3: Mô hình và giải pháp: Chương này đề xuất các mô hình gồm một mô hình ontology mô tả tri thức về một lĩnh vực đặc biệt trong đó sử dụng keyphrase là thành phần chính để hình thành các khái niệm của ontology; Các kỹ thuật xử lý tài

liệu văn bản; Xây dựng mô hình VSM trong tra cứu tài liệu có sử dụng ngữ nghĩa cho câu truy vấn.

Chương 4: Cài đặt thử nghiệm và đánh giá: Thiết kế mô hình dữ liệu ontology hỗ trợ xử lý câu truy vấn; Xây dựng chức năng tra cứu nâng cao cho hệ thống quản lý văn bản Tic-Office. Tiến hành thực nghiệm, so sánh và đánh giá kết quả

Chương 5: Kết luận và hướng phát triển: Tổng kết những kết quả đạt được của luận văn, những hạn chế và hướng phát triển của đề tài trong tương lai.

CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

1.1 Giới thiệu tổng quan về vấn đề nghiên cứu

1.1.1 Nhu cầu và thực trạng tìm kiếm hiện nay

Bên cạnh nhu cầu về học tập, giải trí thì nhu cầu tìm kiếm thông tin là một nhu cầu không thể thiếu khi sử dụng Internet, theo [17] thống kê những trang web được có lượng truy cập nhiều nhất tại Việt Nam tháng 12 năm 2020, trong đó trang tìm kiếm Google với hơn 1 tỷ lượt truy cập trong tháng, như vậy cho thấy nhu cầu tìm kiếm của người dùng khi có sử dụng Internet là rất lớn, vì vậy nhu cầu tìm kiếm thông tin được xem quan trọng nhất trong nhu cầu sử dụng internet của người dùng.

1.1.2 Khảo sát hệ thống tìm kiếm văn bản

Hầu hết đối với các hệ thống quản lý dữ liệu hiện nay thì các yêu cầu về quản lý, chia sẻ và tìm kiếm thông tin là chức năng cơ bản cần phải có trong hệ thống quản lý. Trong đó chức năng tra cứu thông tin quản lý thì chỉ dừng ở mức độ tìm kiếm cơ bản theo từ khóa được lưu trữ trong dữ liệu. Với phương pháp tìm kiếm theo từ khóa thì kết quả chỉ tìm được nội dung liên quan tới từ khóa chứ không tìm được các nội dung liên quan tìm ẩn trong nội dung tìm kiếm.

1.2 Mục tiêu đề tài

Để đáp ứng yêu cầu tra cứu có thể tìm kiếm đầy đủ thông tin trong tài liệu trong hệ thống thì đề tài cần thực hiện các nội dung như sau:

- Tìm hiểu về web ngữ nghĩa, xây dựng mô hình dữ liệu hỗ trợ biểu diễn câu truy vấn
- Tìm hiểu về kỹ thuật xử lý ngôn ngữ tự nhiên, kỹ thuật rút trích dữ liệu từ hình anh scan của tài liệu.
- Kỹ thuật so khớp giữa tài liệu và câu truy vấn sử dụng mô hình VSM.
- Xây dựng chức năng tra cứu nâng cao cho hệ thống Tic-Office để hỗ trợ người dùng trong tìm kiếm văn bản được đầy đủ.

1.3 Đối tượng và phạm vi nghiên cứu

Hệ thống quản lý văn bản của Hội Nông Dân, nhu cầu và hiện trạng tra cứu.

Phương pháp xây dựng mô hình ontology và sử dụng các công cụ hỗ trợ.

Phương pháp xử lý tài liệu sử dụng công cụ trích xuất nội dung OCR, rút trích thực thể có nghĩa NER.

Phương pháp so khớp tài liệu và câu truy vấn sử dụng VSM.

Tập tài liệu văn bản được lưu trữ và xử lý trên hệ thống Tic-Office.

1.4 Phương pháp nghiên cứu

Phương pháp nghiên cứu lý thuyết: Tìm hiểu kỹ thuật OCR, kỹ thuật rút trích NE, nghiên cứu các lý thuyết liên quan đến xây dựng hệ thống Web ngữ nghĩa. Thu thập, tổng hợp thông tin về văn bản của hệ thống Tic-Office.

Phương pháp khảo sát: Tìm hiểu quy trình lưu trữ, cấu trúc dữ liệu, công tác quản lý và chức năng tra cứu văn bản của hệ thống Tic-Office.

Phương pháp thực nghiệm: Xây dựng chức năng tra cứu nâng cao, so sánh với chức năng tra cứu hiện tại, đánh giá kết quả đạt được của hai chức năng tra cứu.

1.5 Ý nghĩa khoa học và thực tiễn của đề tài

Áp dụng công nghệ mới trong tìm kiếm thông tin của tài liệu của Web ngữ nghĩa. Phát triển các ứng dụng để góp phần từng bước phổ biến và làm phát triển công nghệ này.

Ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên, kỹ thuật rút trích NE, sử dụng mô hình VSM có sử dụng thực thể vào lĩnh vực tìm kiếm nội dung văn bản theo ngữ nghĩa, góp phần phục vụ tốt công tác nghiên cứu, tìm hiểu, sử dụng và khai thác tài liệu của hệ thống Tic-Office. Hỗ trợ công tác văn thư, lưu trữ và tra cứu tài liệu một cách nhanh chóng.

1.6 Nội dung thực hiện

Nghiên cứu khảo sát hiện trạng của hệ thống quản lý văn bản Tic-Office. Phân tích hiện trạng nhu cầu tìm kiếm và khả năng mở rộng nhu cầu tìm kiếm của ứng dụng.

Xây dựng mô hình ontology hỗ trợ câu truy vấn. Sử dụng các công cụ hỗ trợ xử lý nội dung tài liệu.

Xây dựng mô hình vector biểu diễn câu truy vấn và tài liệu

Xây dựng chức năng tra cứu hỗ trợ tìm kiếm theo ngữ nghĩa của hệ thống Tic-Office.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Vấn đề truy tìm thông tin

2.1.1 Cấu trúc của một hệ thống truy tìm thông tin

Hệ thống truy tìm thông tin (Information Retrieval, viết tắt IR) là một hệ thống tìm kiếm thông tin các yêu cầu của người dùng đặt ra và thực hiện tìm kiếm trong tất cả nguồn dữ liệu mà hệ thống đang lưu trữ, quản lý để trả về cho người dùng thông tin đúng với yêu cầu đưa ra.

Hệ thống IR tập trung chủ yếu vào văn bản (document) được quản lý, lưu trữ, truy xuất bằng cách nào để dễ dàng có thể truy vấn (query) nhanh chóng, kịp thời.

2.1.2 Các phương pháp truy hồi thông tin

Ý tưởng của phương pháp này là sử dụng một danh sách các thuật ngữ trong tài liệu hay câu truy vấn là một dạng biểu diễn nội dung của câu truy vấn và tài liệu đó. Khi một thuật ngữ của tài liệu được chọn thì phải mã hóa theo mô hình toán học để máy tính có thể xử lý được.

2.1.2.1 Mô Hình Boolean

Mô hình Boolean được tính toán bằng đại số boolean và tập hợp trong toán học nên cài đặt đơn giản, dễ sử dụng và thời gian tìm hiểu nhanh chóng. Với mô hình này, mỗi văn bản được trình bày bởi một vector nhị phân, vector chỉ có hai giá trị {0,1},

nếu từ khóa thứ k được tìm thấy trong văn bản V_l trọng số được xác định là $W_{kl} = 1$, nếu không tồn tại trong văn bản thì $W_{kl} = 0$.

Các phép toán logic "AND, OR, NOT" được sử dụng để biểu diễn nội dung câu truy vấn khi muốn tìm kiếm ngữ nghĩa chính xác.

2.1.2.2 Mô hình Boolean nâng cao (Advanced Boolean Model)

Một trong các phương pháp sử dụng của mô hình mở rộng là thay đổi giá trị của hàm boolean thay vì chỉ trả về hai giá trị 0 hoặc 1 thì kết quả trả về sẽ có giá trị từ 0 đến 1 tương ứng với độ tương đồng giữa biểu thức và văn bản.

2.1.2.3 Mô Hình Không Gian Vector (VSM)

Mô hình VSM khắc phục những hạn chế của mô hình boolean bằng cách đánh trọng số cho đối tượng đặc trưng. Trọng số đối tượng đặc trưng không giới hạn bởi hai trị 0 hoặc 1, các trọng số này được sử dụng để tính toán độ đo tương đồng của mỗi văn bản với câu truy vấn.

2.1.2.4 Mô Hình Xác Suất (Probability Model)

Mô hình xác suất là một biểu diễn toán học của một hiện tượng ngẫu nhiên. Nó được xác định bởi không gian mẫu, các sự kiện trong không gian mẫu và xác suất liên quan đến mỗi sự kiện.

2.1.3 Đánh giá một hệ thống tìm kiếm thông tin

Một hệ thống IR được đánh giá hiệu quả khi thỏa mãn hai

độ đo cơ bản là độ chính xác (Precision) và độ bao phủ (Recall).

$$\text{Độ chính xác} = \frac{|S \cap U|}{|S|} \qquad \text{Độ bao phủ} = \frac{|S \cap U|}{|U|}$$

Với S là tập các tài liệu tìm được có liên quan đến trong hệ thống.

U là tập hợp các tài liệu liên quan theo đánh giá của người dùng.

2.2 Ontology

2.2.1 Định nghĩa

2.2.1.1 Trong triết học

Ontology (Bản thể học) là sự tra vấn triết học về bản tính nền tảng của hiện hữu, thực tại, tồn tại. Các triết gia khác nhau tán thành những bản thể học khác nhau vì họ có những quan điểm khác nhau về cái đang tồn tại ở cấp độ nền tảng hay phổ biến nhất. Bản thể học của Descartes, chẳng hạn, bàn về các tinh thần, vật chất và Thượng đế, trong khi đó bản thể học của Sartre lại bàn về tồn tại và sự phủ định của nó, không tồn tại hay hư vô. Bản thể học đôi khi được mô tả là một nhánh của siêu hình học, nhưng trên thực tế nó là thuật ngữ rộng hơn siêu hình học ở chỗ có hữu thể học siêu hình học và hữu thể học phi siêu hình học.

2.2.1.2 Trong lĩnh vực Trí tuệ nhân tạo

Trong Trí tuệ nhân tạo ontology cũng đã có rất nhiều định nghĩa khác nhau từ nhiều nhà nghiên cứu trên thế giới, một

số khái niệm được xem là kinh điển và được công nhận rộng rãi như định nghĩa của Gruber (1993), Borst (1997), Studer (1998)… Nhìn chung, định nghĩa về ontology thì qua mỗi thời điểm có các khái niệm, các định nghĩa thể hiện một cách nhìn khác nhau về mô hình dữ liệu và đi cùng với khái niệm là một phương pháp luận và kỹ thuật xây dựng mô hình dữ liệu ontology.

2.2.2 Các thành phần của ontology

Ontology thông thường được thiết kế từ các thành phần như : Classs, Properties, Function, Axioms, Relation, Instance…

2.2.3 Phân loại ontology

Về cơ bản có các loại ontology sau:

- Ontology biểu diễn tri thức (Knowledge representation Ontology)
- Ontology tổng quát (Generic Ontology)
- Metadata ontology
- Ontology miền (Domain Ontology)
- Ontology tác vụ (Tast Ontology)
- Ontology lĩnh vực - tác vụ (Domain – Tast Ontology)
- Ontology ứng dụng (Application Ontology)
- Ontology chỉ mục (Index Ontology)
- Ontology hỏi và trả lời (Tell and Ask Ontology)

Ngoài ra, các ontology còn được phân loại dựa vào tính

phức tạp của mô hình biểu diễn dữ liệu như Lightweight ontology, Heavyweight ontology

2.2.4 Vai trò của Ontology

Ontology mục đích ban đầu là tạo ra các miền tri thức gồm nhiều lĩnh vực khác nhau để có được thông tin đa dạng, phục vụ cho nhu cầu xử lý thông tin của con người cũng như máy tính có thể xử lý và thao tác được. Bên cạnh đó các mô hình dữ liệu còn có thể dùng để chia sẽ thông tin giữa các hệ thống xử lý dữ liệu với nhau.

2.2.5 Các ứng dụng dựa trên Ontology

Ngày nay ontology không chỉ dừng lại trong việc chia sẽ thông tin dữ liệu mà nó còn không ngừng phát triển và được áp dụng vào hầu hết các lĩnh vực khác nhau trong môi trường có liên quan đến dữ liệu điển hình như hệ thống xử lý ngôn ngữ tự nhiên, truy hồi thông tin, mua bán trên sàn thương mại điện tử, quản trị cơ sở dữ liệu, công nghệ phần mềm, mạng và an toàn bảo mật…

2.2.6 Các hướng tiếp cận xây dựng ontology

Một trong nhưng phương pháp xây dựng ontology thông dụng hiện nay là rút trích thông tin nội dung từ các nguồn dữ liệu khác nhau như từ internet. Kỹ thuật xử lý được áp dụng để rút trích thông tin nội dung bằng vào việc áp dụng phương pháp học máy, xử lý ngôn ngữ tự nhiên và phương pháp đơn giản nhất là

thống kê theo từ khóa.

2.3 Mô hình Không gian Vector (VSM)

2.3.1 Giới thiệu

Vector space model (Mô hình không gian vector) là một mô hình đại số (algebraic model) thể hiện thông tin văn bản như một vector, các phần tử của vector này thể hiện mức độ quan trọng của một từ và cả sự xuất hiện hay không xuất hiện (Bag of words) của nó trong một tài liệu [16].

2.3.2 Mô hình không gian Vector

Ý tưởng của Vector Space Model là biểu diễn văn bản và các câu truy vấn dưới dạng Vector, Rep(d) của docs và Rep(q) của query sẽ cho kết quả là các vector. Sau đó tính độ tương đồng của query với từng văn bản theo công thức Sim(Rep(q), Rep(d)) để tìm ra docs nào phù hợp nhất với query [15].

CHƯƠNG 3: MÔ HÌNH VÀ GIẢI PHÁP

3.1 Giới thiệu hệ thống Tic-Office

Hệ thống sẽ quản lý được văn bản gửi đến và văn bản chuyển đi của Hội Nông Dân từ các Sở ban ngành, huyện, thành phố trong tỉnh. So với các hệ thống quản lý văn bản khác thì hệ thống Tic-Office chỉ có một số chức năng cơ bản liên quan đến xử lý, điều hành văn bản. Hệ thống tập trung chủ yếu vào ba chức năng chính như: quản lý văn bản đến , quản lý văn bản đi và chức năng tra cứu theo từ khóa trích yếu của hệ thống.

3.2 Mô hình ontology cho ngữ nghĩa của câu truy vấn

Trong đề tài này tôi sử dụng mô hình CK_ONTO đơn giản để biểu diễn nội dung của câu truy vấn, mô hình gồm ba thành phần:

$$\mathbf{(C, K, R_{KK})}$$

Trong đó:

- K: Một tập hợp các keyphrase
- Một tập hợp C các lớp keyphrase
- Một tập hợp R_{KK} các quan hệ giữa các keyphrase

-

Bảng 3.1: bảng ví dụ mối quan hệ tương đương

Equivalent keyphrase	Selected keyphrase	
UBND	Ủy ban nhân dân	"is a acronym of"
TP	Thành phố	"is a acronym of"
Giấy ủy quyền	Giấy chuyển quyền	"is a synonym of"
Hiếm xảy ra	Không thường xuyên	"is a synonym of"
Năng động	Hoạt bát	"is a synonym of"
Tôn kính	Kính trọng	"is a synonym of"

3.3 Công cụ hỗ trợ xử lý tài liệu văn bản

3.3.1 Phương pháp nhận dạng văn bản

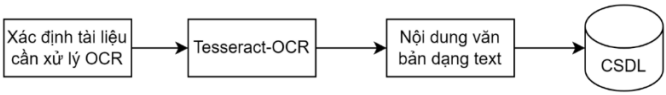
3.3.1.1 Giới thiệu

Hiện nay, nhu cầu trích xuất từ hình ảnh ngày càng tăng, bên cạnh sự gia tăng nhu cầu là sự phát triển của công nghệ nhận dạng ký tự quang học (Optical Character Recognition), còn được gọi là nhận dạng ký tự quang học viết tắt là OCR. Đây là một công nghệ chuyển đổi hình ảnh chữ viết tay hoặc đánh máy thành các ký tự được mã hóa bằng máy tính.

3.3.1.2 Phương pháp nhận dạng văn bản

Trong công trình [20] tác giả nghiên cứu một hệ thống nhận dạng văn bản được tổ chức bao gồm bốn thành phần [10]: Tiền xử lý, phân tích bố cục, nhận dạng văn bản và hậu xử lý.

3.3.1.3 Quy trình xử lý Tesseract OCR của tài liệu của hệ thống Tic-Office



Hình 3.9: mô tả quy trình xử lý tài liệu văn bản

Các bước thực hiện:

- Bước 1: Xác định tập tài liệu cần rút trích nội dung
- Bước 2: Sử dụng công cụ Tesserract-OCR để xử lý hình ảnh văn bản
- Bước 3: Lưu nội dung đã được rút trích vào CSDL

3.3.2 Phương pháp rút trích nội dung thực thể

3.3.2.1 Định nghĩa:

Thực thể là các đối tượng của thế giới thực bao gồm cả đối tượng có thể nhìn thấy hoặc không nhìn thấy được.

Thực thể trong văn bản thì được thể hiện trong các dạng: Tên riêng, Danh từ hoặc cụm danh từ, Đại từ.

Nhận dạng thực thể có tên (Named Entity Recognition – NER) nhằm rút trích các từ, cụm từ trong văn bản là tên của một đối tượng nào đó, điển hình như tên người, tên tổ chức, tên địa danh, thời gian v.v.

3.3.2.2 Quy trình xử lý rút trích thực thể

Các bước thực hiện:

- Bước 1: Sử dụng công cụ OCR xử lý văn bản lưu vào CSDL
- Bước 2: Sử dụng công cụ Underthesea để phân tách nội dung thành các thực thể
- Bước 3 Lưu nội dung đã phân tách thành các tập tin nội dung với tên tập tin theo cấu trúc

3.3.3 Mô hình Conditional Random Fields (CRFs)

Conditional random fields là một probabilistic framework (theo xác suất) cho việc gán nhãn và phân đoạn dữ liệu tuần tự. Thay vì sử dụng xác suất độc lập trên chuỗi nhãn và chuỗi quan sát, CRFs sử dụng xác suất có điều kiện $P(Y \mid X)$ trên toàn bộ chuỗi nhãn được đưa bởi chuỗi mỗi chuỗi quan sát X. CRF là một mô hình đồ thị vô hướng định nghĩa một phân bố tuyến tính đơn trên các chuỗi nhãn được đưa ra bởi các chuỗi quan sát được. CRFs thuận lợi hơn các mô hình Markov và MEMM và làm tốt hơn cả của MEMM và HMM trên số lượng chuỗi gán nhãn lớn [13].

3.4 Xây dựng mô hình VSM trong tra cứu tài liệu có sử dụng ngữ nghĩa cho câu truy vấn

3.4.1 Số hóa văn bản theo mô hình không gian vector

Giả sử tập tài liệu D = $\{d_1, d_2, \ldots, d_n\}$ có n văn bản và tập

C = $\{c_1, c_2, \ldots, c_m\}$ có *m* từ chỉ mục biểu diễn cho tập văn bản. Vậy không gian vector biểu diễn tập chỉ mục C có m tập chỉ mục và tập văn bản D có n tập văn bản là một vector *m x n* chiều

Hàm tính trọng số của từ chỉ mục

$w_{ij} = t_{ij} \times T_i \times n_j$

Trong đó:

- t_{ij}: tổng số lần xuất hiện của từ chỉ mục trong một văn bản
- T_i : tổng số lần xuất hiện của từ trong toàn bộ văn bản
- n_j: là hệ số điều chỉnh chiều dài của văn bản trong tập văn bản.

3.4.2 Ma trận biểu diễn tập văn bản

Để biểu diễn tập văn bản D có n văn bản và có m từ chỉ mục được vector hóa thành mô hình vector A, Vector A được gọi là vector của chỉ mục văn bản. Trong đó số tập văn bản n được biểu diễn thành n cột, còn số chỉ mục m được biểu diễn thành m dòng, do đó số chỉ trong toàn bộ văn bản lúc nào cũng lớn hơn nhiều so với tập văn bản đang xét.

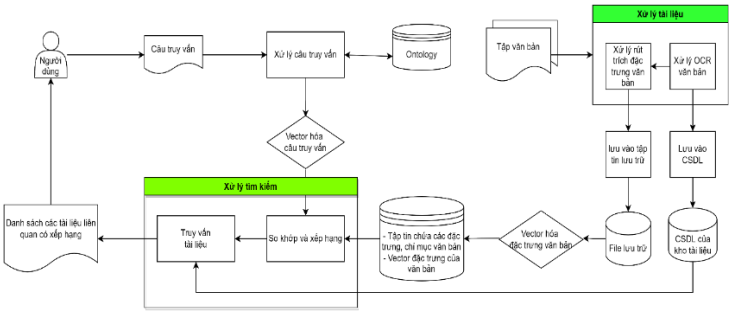
Công thức so khớp câu truy vấn và tài liệu văn bản

$$\cos\theta_j = \frac{d_j^T q}{\|d_j\|_2 \|q\|_2} = \frac{\sum_{i=1}^{m} d_{ij} q_i}{\sqrt{\sum_{i=1}^{m} d_{ij}^2} \sqrt{\sum_{i=1}^{m} q_i^2}}$$

Trong đó: d_{ij} là từ chỉ mục tại vị trí *ij* trong vector A

q_i : từ chỉ mục vị trí i của vector truy vấn

3.4.3 Kiến trúc mô hình tìm kiếm tài liệu VSM



Hình 3.11: Quy trình xử lý câu truy vấn của hệ thống VSM

Mô tả các bước thực hiện

- Bước 1: người dùng nhập vào nội dung câu truy vấn
- Bước 2: Xử lý câu truy vấn dựa vào mô hình dữ liệu ontology
- Bước 3: Xử lý rút trích đặc trưng, chỉ mục của của tập văn bản
- Bước 4: Tạo tập tin đặc trưng và chỉ mục của văn bản
- Bước 4: Tạo tập tin ma trận đặc trưng của văn bản
- Bước 5: Lưu các tập tin đặc trưng, chỉ mục, ma trận đặc trưng vào kho chờ yêu cầu xử lý
- Bước 6: Xử lý ma trận nội dung truy vấn và ma trận đặc trưng văn bản
- Bước 7: Trả về kết quả các tài liệu có xếp hạng cho người dùng

CHƯƠNG 4: CÀI ĐẶT, THỬ NGHIỆM VÀ ĐÁNH GIÁ

4.1 Cài đặt

4.1.1 Xây dựng mô hình dữ liệu ontology

Xây dựng mô hình Ontology CK_ONTO đơn giản gồm ba thành phần {C, K, R_{KK}}

Các bước thực hiện thiết kế mô hình ontology

- Bước 1: Xây dựng các Class (Lớp đối tượng)
- Bước 2: Xây dựng các thuộc tính cho các lớp.
- Bước 3: Xây dựng mối quan hệ giữa các thực thể của các lớp

4.1.2 Module trích xuất nội dung của tài liệu sử dụng Tesseract OCR

Sử dụng công cụ Tesseract OCR để trích xuất nội dung tài liệu thành nội dung văn bản sau đó được lưu vào CSDL.

4.1.3 Module rút trích đặc trưng của tài liệu

Sử dụng công cụ Underthesea xử lý nội dung đã trích xuất từ tài liệu xử lý thành các từ, cụm từ có nghĩa được lưu trữ dưới dạng tập tin được tên tập tin theo cấu trúc.

4.1.4 Module API kết nối đến hệ thống Tic-Office

Module thực hiện giao tiếp giữa hệ thống Tic-Office và module so khớp tài liệu và câu truy vấn, sau đó trả về nội dung thõa mãn điều kiện câu truy vấn và có xếp hạng tìm kiếm.

4.1.5 Cài đặt phân hệ tìm kiếm văn bản

Phân hệ tìm kiếm văn bản gồm 2 module chính:

- Module xây dựng vector đặc trưng của tài liệu
- Module xử lý tìm kiếm thông tin

Cài đặt phân hệ tìm kiếm văn bản VSM

Phân hệ tìm kiếm tài liệu văn bản được cài đặt dựa trên quy trình xử lý tìm kiếm được nêu ở Hình 3.11

- Dữ liệu đầu vào
- Chạy module tạo ma trận đặc trưng văn bản
- Chạy module xử lý tìm kiếm văn bản trả về kết quả cho người dùng
- Giao diện chức năng tra cứu nâng cao được giới thiệu ở Hình 4.6

4.2 Kết quả thử nghiệm

Đề tài tiến hành thử nghiệm các các tài liệu trong khoảng thời gian 3 từ tháng 01/2021 đến tháng 03/2021 bao gồm 672 tập tin tài liệu văn bản. Ứng với tập tài liệu thực hiện khảo sát trên 20 câu truy vấn có chọn lọc và tính toán các độ đo Recall, Precision tương ứng, với ngưỡng tương ứng là 0.03. Hệ thống

hiện tại tìm được hầu hết các tài liệu có liên quan đến nội dung cần tìm và được sắp xếp theo thứ tự độ liên quan giảm dần chính xác hơn so với thứ tự độ liên quan trong hệ thống cũ.

Bảng 4.1: Thống kê kết quả tìm kiếm trên chức năng tra cứu mới

STT	Query	S	T	U	P (S/T)	R (S/U)
1	công nghệ thông tin	43	55	59	78%	73%
2	nông sản	42	61	61	69%	69%
3	quy hoạch giao thông	60	78	82	77%	73%
4	dự toán ngân sách	73	80	81	91%	90%
5	hội nhập	44	50	55	88%	80%
6	chi phí	55	69	73	80%	75%
7	không khí ô nhiễm	35	41	51	85%	69%
8	giới thiệu du lịch chuyên nghiệp	47	59	60	80%	78%
9	giải quyết khiếu nại	50	55	59	91%	85%
10	chuyên canh hè thu	12	13	13	92%	92%
11	hàm lượng khoa học công nghệ cao	38	44	46	86%	83%
12	đài phát thanh truyền hình	28	32	36	88%	78%
13	trung tâm thương mại	50	59	61	85%	82%
14	mật độ chăn nuôi	40	45	53	89%	75%
15	tăng trưởng kinh tế	41	48	61	85%	67%
16	sở tài chính	78	88	91	89%	86%
17	ngân sách	53	58	62	91%	85%
18	phát triển nông thôn	61	66	66	92%	92%

19	an toàn giao thông	69	77	77	90%	90%
20	kinh nghiệm sản xuất	60	68	72	88%	83%
21	khai thác khoáng sản	33	38	41	87%	80%
22	tình hình kinh tế xã hội	42	47	49	89%	86%

Kết quả thực nghiệm với độ đo Precision trung bình của hệ thống hiện tại là 85.9% và độ đo Recall trung bình là 80.5% trên tập thử nghiệm so với hệ thống cũ.

4.3 Đánh giá

Với mô hình tìm kiếm chuyên biệt có tính tập trung cao vào nội dung đặc trưng của tài liệu, kết quả trả về có độ chính xác khá cao trên tập dữ liệu thử nghiệm, được đánh giá là thoả mãn tốt nhu cầu khai thác thông tin của người sử dụng. Từ kết quả thực tế cho thấy, việc sử dụng mô hình ontology cùng với các kỹ thuật xử lý liên quan đã giúp cho hệ thống tìm kiếm với độ chính xác và độ phủ trung bình cao hơn so với hệ thống cũ trên cùng một bộ dữ liệu thử nghiệm.

Ngoài kết quả thực nghiệm đã được trình bày ở trên. Việc triển khai và thử nghiệm mô hình giải pháp mới đã đem lại thành công rất đáng khích lệ. Hệ thống cho thấy tính khả thi và thực nghiệm của giải pháp kết hợp mô hình ontology,VSM và các kỹ thuật xử lý khác.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết quả đạt được của đề tài

Với những mục tiêu đề ra ban đầu, luận văn đã thực hiện hoàn chỉnh nội dung nghiên cứu. Nhóm nghiên cứu đã đề xuất được một mô hình ontology biểu diễn nội dung câu truy vấn cùng với các kỹ thuật xử lý ngữ nghĩa để cải thiện giải pháp thiết kế hệ hỗ trợ tìm kiếm theo ngữ nghĩa cho hệ thống Tic-Office. Mô hình ontology đã giúp hệ thống có khả năng mở rộng việc xử lý câu truy vấn và xác định độ tương đồng ngữ nghĩa giữa các keyphrase một cách tự động. Từ đó, cải thiện độ chính xác và độ bao phủ của tập kết quả trả về trong quá trình tìm kiếm tài liệu theo ngữ nghĩa.

Luận văn đã phân tích và đánh giá các phương pháp tiếp cận trong việc tổ chức lưu trữ và xử lý ngữ nghĩa của tài liệu đã biết, đặc biệt là giải pháp "Kết hợp sử dụng mô hình VSM với mô hình ontology biểu diễn nội dung truy vấn". Từ đó, về mặt lý thuyết, luận văn đã đóng góp trong việc phát triển mô hình biểu diễn tri thức của lĩnh vực và các kỹ thuật xử lý liên quan tới ngữ nghĩa.

Mở rộng kỹ thuật xử lý câu truy vấn, tận dụng các thông tin mô tả cấu trúc của một khái niệm được định nghĩa trong

ontology, tiến hành phân tích và nhận dạng mẫu câu truy vấn để đưa câu truy vấn về dạng cụ thể hơn. Nếu không có kết quả nào được trả về thì ta tiến hành xử lý bằng cách rút trích tự động các đặc trưng diễn đạt nội dung chính muốn tìm kiếm, biểu diễn câu truy vấn thành mô hình vector để thuận tiện trong quá trình so khớp nội dung.

Bên cạnh việc nêu lên các ưu thế và lợi ích của việc nghiên cứu, phát triển mô hình cùng với các đặc trưng dựa trên ngữ nghĩa, chúng tôi đã cài đặt và xây dựng được một ứng dụng thử nghiệm từ những cải tiến này. Đây là một hệ thống quản lý văn bản Tic-Office của Hội nông dân tỉnh Tây Ninh, với yêu cầu sử dụng bao gồm các tác vụ chính là tổ chức lưu trữ, quản lý và tìm kiếm, đặc biệt là chức năng tìm kiếm theo ngữ nghĩa liên quan đến nội dung của tài liệu. Từ đó là cơ sở để đánh giá tính hiệu quả của việc cải tiến so với chức năng tra cứu đã có của hệ thống. Kết quả đạt được là hệ thống đã cho ra kết quả tìm kiếm có độ chính xác trung bình là 85.9% và độ phủ trung bình là 80.5% so với hệ thống cũ trên tập dữ liệu là 672 tài liệu được phân bố trong thời gian thử nghiệm là 3 tháng.

Luận văn đã đạt được mục tiêu là xây dựng chức năng tra cứu theo ngữ nghĩa của hệ thông văn bản Tic-Office để hỗ trợ người dùng tìm kiếm các văn bản liên quan với nội dung truy vấn và cũng đã xây dựng được mô hình dữ liệu ontology từ

đồng nghĩa, từ viết tắt hỗ trợ cho câu truy vấn của người dùng. Giúp tìm được các nội dung liên quan về ngữ nghĩa trong tài liệu

Luận văn đã áp dụng các kỹ thuật rút trích dữ liệu đặc trưng từ hình ảnh góp phần tăng thêm nội dung của văn bản được quản lý, ngoài ra đề tài còn áp dụng mô hình so khớp văn bản sử dụng Vector space model kết hợp mô hình dữ liệu hỗ trợ cho câu truy vấn làm tăng kết quả chính xác của câu truy vấn.

5.2. Những hạn chế của đề tài

Các kỹ thuật đề xuất trong luận văn như OCR, rút trích đặc trưng vẫn còn nhiều hạn chế như văn bản quét không chính xác bị lệch, mờ, thiếu chữ dẫn đến tình trạng rút trích còn nhiều khó khăn trong việc xử lý. Công cụ rút trích đặc trưng văn bản vẫn chưa phân loại được một số đối tượng danh từ riêng, từ viết tắt vào nhóm đối tượng dẫn đến bị thiếu sót trong xử lý dữ liệu.

Quá trình xử lý câu truy vấn đã được mở rộng, tuy nhiên vẫn còn khá đơn giản, chưa tận dụng đầy đủ thông tin trong cấu trúc của lớp. Nếu thông tin của lớp được khai thác đầy đủ thì hệ thống sẽ có khả năng xử lý các cấu truy vấn phức tạp hơn như ở dạng câu hỏi hay câu diễn đạt mệnh đề. Mặc dù còn đơn giản nhưng đây là cơ sở cho việc mở rộng xử lý câu truy vấn so với việc chỉ tìm theo từ khóa trong hệ thống cũ. Đề tài mới chỉ dừng ở mức xây dựng mô hình dữ liệu hỗ trợ trong câu truy vấn, chưa hỗ trợ biểu diễn cho nội dung văn bản vì vậy thông tin so khớp

giữa văn bản và câu truy vấn chưa đầy đủ so với nội dung thực tế của văn bản.

5.3. Hướng phát triển

Tiếp tục phát triển, hoàn thiện các mô hình biểu diễn tri thức, biểu diễn ngữ nghĩa cho tài liệu văn bản, mô hình xử lý ngôn ngữ tự nhiên để rút trích nội dung từ văn bản được chính xác. Nghiên cứu các thuật toán nhằm hỗ trợ tìm kiếm nhanh chóng hơn, dễ dàng hơn, cho kết quả chính xác hơn với nhu cầu tìm kiếm của người dùng. Nghiên cứu các công cụ hỗ trợ tự động, tự động hóa càng cao càng tốt trong từng khâu xử lý chẳng hạn như các mô hình và giải pháp rút trích các đặc trưng từ tài liệu kết hợp với mô hình dữ liệu của tài liệu, các kỹ thuật trong xác suất thống kê, máy học.... Nghiên cứu các giải pháp mới trong lĩnh vực tìm kiếm ngữ nghĩa để tìm ra khả năng tìm kiếm trong nhiều lĩnh vực tri thức khác. Đưa đến một giải giải pháp xây dựng một hệ thống tra cứu xử lý tổng hợp toàn bộ quy trình quản lý văn bản.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng anh

[1] Bilal Ahmad Abu-Salih, "*Applying Vector Space Model (VSM) Techniques in information Retrieval for Arabic Language"*

[2] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, "*An Introduction to Information Retrieval*", Cambridge University Press Cambridge, England, 2009.

[3] Carola Eschenbach, Michael Gruninger (FOIS 2008), "*Formal Ontology in Information Systems"*

[4] Faisal Shafait, Ray Smith (2010), *"Table detection in heterogeneous documents"*

[5] Mindy Bokser,(1992*) "Omnidocument Technologies"*

[6] Nordianah Ab Samat, Masrah Azrifah Azmi Murad, Muhamad Taufik Abdullah, Rodziah Atan (2008), "*Term Weighting Schemes Experiment Based on SVD for Malay Text Retrieval"*, Faculty of Computer Science and Information Technology University Putra Malaysia, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.10, October 2008.

[7] Rajendra Prasath and Sudeshna Sarkar, *"Cross-Language Information Retrieval with Incorrect Query Translations"*

[8] Ray Smith, Daria Antonova, Dar-Shyang Lee, (2009), *"Adapting the Tesseract Open Source OCR Engine for Multilingual OCR"*

[9] Ray Smith, (2009) ,*"Hybrid Page Layout Analysis via Tab-Stop Detection"*

[10] Simone Marinai, (2008),*"Introduction to Document Analysis and Recognition"*

Tài liệu tiếng việt

[11] Phạm Tuấn Dạt, Nguyễn Văn Thủy (2016), *Ứng dụng thư viện lập trình mã nguồn mở xây dựng chương trình nhận dạng văn bản chữ việt, anh từ ảnh số*

[12] Nguyễn Đình Ngọc (2015), "*Xây dựng phần mềm nhận dạng ký tự quang học sử dụng mã nguồn mở Tesseract ocr"*

[13] Lê Thúy Ngọc, (2008), "Xây dựng hệ thống tìm kiếm thông tin theo hướng tiếp cận ngữ nghĩa", Trường đại học Khoa Học Tự Nhiên TP.HCM.

[14] Huỳnh Thị Thanh Thương (2012), *"Nghiên cứu mô hình tổ chức và kỹ thuật tìm kiếm có ngữ nghĩa trên kho tài nguyên học tập lĩnh vực CNTT"*, Trường đại học Khoa Học Tự Nhiên TP.HCM.

Tài liệu website

[15] https://blog.duyet.net/2019/08/ir-vector-space-model.html

[16] https://butchiso.com/2013/10/tim-hieu-ve-mo-hinh-khong-gian-vector.html

[17] Chuyển đổi số tại Việt Nam và những thống kê ấn tượng đầu năm 2021 | Visual Story - Báo Lao Động (laodong.vn)

[18] HỮU THỂ HỌC / BẢN THỂ HỌC (Ontology) - Triết học (triethoc.edu.vn)

[19] DCMI: Ontology (dublincore.org)

[20] Deep learning ứng dụng trong nghiệp vụ nhận dạng văn bản - An Toàn Thông Tin (antoanthongtin.vn)

[21] Conditional Random Fields - Trí tuệ nhân tạo (trituenhantao.io)

[22] Lp space - Wikipedia

[23] Precision and recall - Wikipedia