BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Cù Việt Dũng

NÂNG CAO ĐỘ CHÍNH XÁC CỦA TRA CỨU ẢNH THEO NỘI DUNG DỰA TRÊN TIẾP CẬN HỌC ĐA TẠP TỪ THÔNG TIN PHẢN HỒI CỦA NGƯỜI DÙNG

LUẬN ÁN TIẾN SĨ NGÀNH MÁY TÍNH

Hà Nội – 2023

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Cù Việt Dũng

NÂNG CAO ĐỘ CHÍNH XÁC CỦA TRA CỨU ẢNH THEO NỘI DUNG DỰA TRÊN TIẾP CẬN HỌC ĐA TẠP TỪ THÔNG TIN PHẢN HỒI CỦA NGƯỜI DÙNG

LUẬN ÁN TIẾN SĨ NGÀNH MÁY TÍNH
Mã số: 9 48 01 01

Xác nhận của Học viện Khoa học và Công nghệ	**Người hướng dẫn 1** *(Ký, ghi rõ họ tên)*	**Người hướng dẫn 2** *(Ký, ghi rõ họ tên)*

Hà Nội – 2023

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài nghiên cứu trong luận án này là công trình nghiên cứu của tôi dựa trên những tài liệu, số liệu do chính tôi tự tìm hiểu và nghiên cứu. Chính vì vậy, các kết quả nghiên cứu đảm bảo trung thực và khách quan nhất. Đồng thời, kết quả này chưa từng xuất hiện trong bất cứ một nghiên cứu nào. Các số liệu, kết quả nêu trong luận án là trung thực, nếu sai tôi hoàn toàn chịu trách nhiệm trước pháp luật

Tác giả

NCS. Cù Việt Dũng

LỜI CẢM ƠN

Luận án tiến sĩ được hoàn thiện bởi sự cố gắng của chính bản thân cùng với sự giúp đỡ tận tình của hai Thầy hướng dẫn khoa học, một số chuyên gia, đồng nghiệp, bạn bè và người thân trong gia đình.

Trước tiên, tôi xin được bày tỏ lòng biết ơn chân thành đến hai Thầy hướng dẫn khoa học PGS.TS. Nguyễn Hữu Quỳnh và PGS.TS. Ngô Quốc Tạo. Nghiên cứu sinh đã nhận được những định hướng khoa học, những bài học quý báu, sự hướng dẫn tận tình và kinh nghiệm nghiên cứu khoa học quý giá trong nghiên cứu.

Tôi xin chân thành cảm ơn phòng Ban lãnh đạo, phòng Đào tạo, các phòng chức năng của Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam đã tạo điều kiện thuận lợi trong suốt quá trình nghiên cứu và thực hiện luận án.

Tôi xin chân thành cảm ơn tới Ban giám hiệu, Ban lãnh đạo Khoa, các Thầy cô trong Bộ môn Công nghệ phần mềm và toàn thể các giảng viên Khoa Công nghệ thông tin hai trường Đại học Thủy lợi, Đại học Điện Lực đã quan tâm, giúp đỡ tôi hoàn thành nhiệm vụ.

Cuối cùng, tôi xin bày tỏ lòng biết ơn vô hạn tới mọi thành viên trong gia đình, sự khuyến khích động viên của gia đình là động lực để tôi hoàn thành luận án này

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ KÝ VIẾT TẮT

Ký hiệu	Diễn giải tiếng Anh	Diễn giải tiếng Việt
AP	Average precision	Độ chính xác trung bình
ARE	Augmented relation embedding	Nhúng quan hệ gia tăng
BSFG	Balanced sample feedback based on the graph	Mẫu phản hồi cân bằng dựa vào đồ thị
CBIR	Content-based image retrieval	Tra cứu ảnh dựa vào nội dung
CMAC	Combining multiple aspect classifier	Bộ phân lớp kết hợp đa khía cạnh
DAG-DNE	Double adjacency graph-based discriminant neighborhood embedding	Nhúng lân cận phân biệt dựa trên đồ thị lân cận kép
DGLPGE	Discriminative globality and locality preserving graph embedding	Nhúng đồ thị bảo toàn toàn cục và cục bộ phân biệt
DMINTIR	Discriminative multi-view interactive image re-ranking	Phân hạng lại ảnh tương tác đa khung nhìn phân biệt
DNE	Discriminant neighborhood embedding	Nhúng lân cận phân biệt
DSSA	Discriminative semantic subspace analysis	Phân tích không gian con ngữ nghĩa phân biệt
HMR	Heterogeneous manifold ranking	Phân hạng đa tạp không đồng nhất
HSV	Hue, saturation, value	Tông màu, độ bão hoà màu, giá trị màu.
LDA	Linear discriminant analysis	Phân tích phân biệt tuyến tính
LDP	Local discriminant embedding	Nhúng phân biệt cục bộ
LLE	Locally linear embedding	Nhúng tuyến tính cục bộ
LPP	Locality preserving projection	Chiếu bảo toàn cục bộ
LRCDP	Linear regression classification steered discriminative projection	Chiếu phân biệt định hướng phân lớp hồi quy tuyến tính

LFGBSE	Learning flexible graph-based semi-supervised embedding	Nhúng đa tạp dựa vào đồ thị linh hoạt với nhúng phân biệt bán giám sát
MFA	Marginal Fisher analysis	Phân tích lề Fisher
MMP	Maximum margin projection	Chiếu lễ cực đại
NPE	Neighborhood preserving embedding	Nhúng bảo toàn lân cận
O-SVM	Original support vector machine	Máy véc tơ hỗ trợ gốc
PCA	Principal components analysis	Phân tích thành phần chính
RBF	Radial basis function	Hàm cơ sở xuyên tâm
RF	Relevance feedback	Phản hồi liên quan
SCDP	Semantic class discriminant projection	Chiếu phân biệt lớp ngữ nghĩa
SCDPIR	Semantic class discriminant projection for image retrieval	Chiếu phân biệt lớp ngữ nghĩa cho tra cứu ảnh
SDA	Semisupervised Discriminant Analysis	Phân tích phân biệt bán giám sát
SoLPP	Supervised optimal locality preserving projection	Chiếu bảo toàn cục bộ tối ưu có giám sát
SSDL	Stable semi-supervised discriminant learning	Học phân biệt bán giám sát ổn định
SVM	Support vector machine	Máy véc tơ hỗ trợ

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

LỜI MỞ ĐẦU

1. Lý do chọn đề tài

Với sự phát triển mạnh mẽ của khoa học công nghệ, thiết bị thu nhận hình ảnh cùng mạng xã hội như facebook, twitter, instagram làm cho số lượng ảnh được lưu trữ trong các cơ sở dữ liệu và trên Internet ngày càng tăng lên. Chính vì thế, để tìm một tập ảnh phù hợp với nhu cầu của con người trong tập dữ liệu khổng lồ đó, chúng ta cần những phương pháp tra cứu ảnh hiệu quả [1]. Có hai cách tiếp cận trong bài toán tra cứu ảnh gồm tra cứu ảnh dựa vào văn bản (TBIR- Text based image retrieval) và tra cứu ảnh dựa vào nội dung (CBIR - Content based image retrieval). Trong TBIR, siêu dữ liệu (metadata) chẳng hạn như từ khóa, chú thích được sử dụng để mô tả ảnh. Mặc dù, cách tiếp cận dựa trên văn bản có thể mang lại sự linh hoạt trong việc tạo ra các truy vấn, nhưng việc tra cứu ảnh chỉ dựa trên văn bản là không hiệu quả vì các lý do sau: (1) khó tạo ra các mô tả thủ công cho một tập ảnh lớn và gia tăng từng giây, (2) sự không nhất quán giữa các mô tả của người dùng khác nhau, và (3) khó chuyển đổi từ hệ thống này sang hệ thống khác. Do đó, tra cứu ảnh dựa vào nội dung được đề xuất để khắc phục những hạn chế kể trên của cách tiếp cận tra cứu ảnh dựa vào văn bản.

Tra cứu ảnh dựa vào nội dung đã thu hút sự quan tâm của cộng đồng nghiên cứu và phát triển ứng dụng trong những thập kỷ qua. Thuật ngữ "nội dung" gắn với thị giác trực quan của con người như màu sắc, hình dạng, kết cấu hoặc các thông tin khác được lấy từ chính bức ảnh đó, không phải siêu dữ liệu như từ khóa, chú thích hay mô tả được liên kết với ảnh. Nội dung của các ảnh trong tập dữ liệu ảnh lớn sẽ được trích rút một cách tự động từ chính những ảnh đó và được lưu trữ trong cơ sở dữ liệu đặc trưng. Trong tra cứu ảnh dựa vào nội dung, một hoặc nhiều ảnh mẫu hoặc ảnh phác thảo được cung cấp làm truy vấn, trong khi đó truy vấn TBIR trực tiếp sử dụng các từ khóa, các chú thích. Khi đó đặc trưng của ảnh truy vấn sẽ được trích rút tự động theo cùng một cách thức như với các ảnh trong cơ sở dữ liệu ảnh [2]. Đặc trưng của ảnh truy vấn được đối sánh lần lượt với từng đặc trưng trong tập cơ sở dữ liệu đặc trưng sử dụng một độ đo tương tự nào đó. Tập ảnh kết quả trả về và hiển thị cho người dùng gồm các ảnh có độ tương tự cao nhất (hay có khoảng cách nhỏ nhất) so với ảnh truy vấn. Độ chính xác của hệ thống CBIR phụ thuộc chủ yếu vào hai yếu

tố: (1) biểu diễn nội dung ảnh, và (2) độ đo khoảng cách giữa đặc trưng của ảnh truy vấn đến từng ảnh trong cơ sở dữ liệu ảnh. Mặc dù đã có nhiều kỹ thuật được đề xuất nhưng đây vẫn là một thách thức lớn trong nghiên cứu tra cứu ảnh dựa vào nội dung do khoảng trống ngữ nghĩa giữa đặc trưng mức thấp (màu sắc, hình dạng, kết cấu) được trích rút từ ảnh và nhận thức của người về ảnh.

Để thu hẹp khoảng trống ngữ nghĩa này, tiếp cận phản hồi liên quan (RF - Relevant feedback) của người dùng khai thác tương tác giữa người dùng và hệ thống tra cứu ảnh để thu được thông tin về các ảnh liên quan (mẫu dương) và không liên quan (mẫu âm) so với ảnh truy vấn. Tuy nhiên, số mẫu phản hồi của người dùng thường rất nhỏ so với số chiều của đặc trưng biểu diễn ảnh. Điều này dẫn đến phải giải quyết bài toán giảm chiều đặc trưng biểu diễn ảnh, làm cho véc tơ đặc trưng mới (véc tơ đặc trưng trong không gian chiếu) có số chiều thấp hơn nhiều so với véc tơ đặc trưng gốc. Phương pháp chiếu ước lượng cả thuộc tính hình học và phân biệt của tập đặc trưng cơ sở dữ liệu trong CBIR được áp dụng. Phép chiếu ngẫu nhiên của dữ liệu dễ áp dụng nhưng có thể bỏ mất một số thông tin quan trọng của tập dữ liệu ảnh. Để giải quyết hạn chế này, phương pháp giảm chiều theo tiếp cận học máy bao gồm giảm chiều tuyến tính (không giám sát và có giám sát) đã được sử dụng, bao gồm phân tích thành phần chính (PCA - Principal component analysis), Phân tích phân biệt tuyến tính (LDA - Linear Discriminant Analysis). Các phương pháp này xác định tiêu chí đánh giá cụ thể trước khi thực hiện phép chiếu để giữ lại thông tin quan trọng theo tiêu chí đã xét. Nhờ vậy có thể đã cải thiện đáng kể độ chính xác của tra cứu. Tuy nhiên cách tiếp cận trên bỏ qua cấu trúc phi tuyến tính của dữ liệu, tức là chỉ coi tập mẫu dữ liệu nằm trên một không gian con nào đó mà không xét đến thực tế tập mẫu dữ liệu có thể nằm trên nhiều không gian con khác nhau (gọi là dữ liệu đa tạp). Các phương pháp học đa tạp được đề xuất nhằm khám phá cấu trúc phi tuyến tính của dữ liệu bằng cách xem các mẫu dữ liệu nằm trên nhiều không gian con khác nhau. Trong luận án này, thuật ngữ "Học đa tạp" được hiểu là phương pháp học máy được áp dụng trên dữ liệu đa tạp để khám phá cấu trúc phi tuyến tính của dữ liệu này. Các phương pháp học đa tạp không giám sát xử lý dữ liệu không có nhãn như: Chiếu bảo toàn cục bộ (LPP - Locality preserving projection) [3, 4], Nhúng tuyến tính cục bộ (LLE - Locally linear embedding) [5], Nhúng bảo toàn lân cận (NPE- Neighborhood

Preserving Embedding) [6], WeightedIso [7], và Supervised Isomap (S-Isomap) [8]. Các phương pháp học đa tạp có giám sát tiêu biểu gồm: Phân tích phân biệt tuyến tính [9], Nhúng phân biệt cục bộ (LDP - Local Discriminant Embedding) [10], Chiếu bảo toàn cục bộ tối ưu có giám sát (SoLPP - Supervised Optimal Locality Preserving Projection) [11], Phân tích lề Fisher (MFA - Marginal Fisher Analysis) [9], Nhúng lân cận phân biệt (DNE - Discriminant neighborhood embedding) [12], Chiếu phân biệt định hướng phân lớp hồi quy tuyến tính (LRCDP - Linear Regression Classification Steered Discriminative Projection) [13], và Nhúng đồ thị bảo toàn toàn cục và cục bộ phân biệt (DGLPGE -Discriminative Globality And Locality Preserving Graph Embedding) [14]. Một số phương pháp học đa tạp bán giám sát tiêu biểu được đề xuất bao gồm: Nhúng quan hệ gia tăng (ARE - Augmented Relation Embedding) [15], Chiếu cực đại lề cho tra cứu ảnh (MMP - Maximum Margin Projection) [16], Phân tích phân biệt bán giám sát (SDA - Semisupervised Discriminant Analysis) [17], Nhúng đa tạp dựa vào đồ thị linh hoạt với nhúng phân biệt bán giám sát (LFGBSE - Learning flexible graph-based semi-supervised embedding) [18], Học phân biệt bán giám sát ổn định (SSDL - Stable Semi-Supervised Discriminant Learning) [19]. Các phương pháp học đa tạp kể trên tuy khám phá được cấu trúc phi tuyến của dữ liệu, nhưng một số phương pháp học đa tạp không giám sát cho độ chính xác tra cứu thấp vì chúng không tận dụng được nhãn của dữ liệu, trong khi một số phương pháp học đa tạp có giám sát chưa khai thác tốt tính lân cận của các mẫu cùng lớp và các mẫu ở các lớp khác nhau. Trong thực tế, các mẫu phản hồi dương thường có số lượng hạn chế so với số lượng mẫu phản hồi âm [20].

CBIR sử dụng phản hồi liên quan có một số vấn đề sau: (1) chỉ khám phá các cấu trúc Euclide toàn cục, chỉ xem xét cấu trúc cục bộ của các mẫu trong cùng một lân cận; (2) số lượng mẫu thu được từ phản hồi của người dùng thường nhỏ và mất cân bằng giữa hai lớp dương và lớp âm; (3) Chưa quan tâm đến các khía cạnh khác nhau của dữ liệu ảnh. Do đó, độ chính xác của các phương pháp tra cứu ảnh sử dụng học máy để giảm chiều kể trên thường kém hiệu quả.

Do vậy, việc đề xuất phương pháp tra cứu ảnh hiệu quả, giải quyết được các hạn chế trên là một nhu cầu cần thiết. Luận án chọn đề tài “Nâng cao độ chính xác

của tra cứu ảnh theo nội dung dựa trên tiếp cận học đa tạp từ thông tin phản hồi của người dùng".

2. Mục tiêu của luận án

Mục tiêu chung của luận án: Nâng cao độ chính xác của tra cứu ảnh dựa trên học đa tạp để giảm chiều từ thông tin phản hồi của người dùng.

Mục tiêu cụ thể của luận án:

Đề xuất được một số giải pháp nâng cao độ chính xác tra cứu ảnh bao gồm:

-Đề xuất phương pháp tìm ma trận chiếu tối ưu theo tiếp cận học đa tạp.

-Đề xuất phương pháp tự động bổ sung mẫu dương vào tập huấn luyện, giải quyết vấn đề mất cân bằng của tập huấn luyện. Phương pháp này bổ sung các mẫu dương vào tập huấn luyện đồng thời tận dụng các khía cạnh khác nhau của đối tượng để tạo ra một bộ phân lớp mạnh.

3. Đối tượng nghiên cứu của luận án

Luận án tập trung vào nghiên cứu và tìm hiểu một số đối tượng liên quan đến tra cứu ảnh như:

- Tổng quan về Tra cứu ảnh dựa vào nội dung.
- Phản hồi liên quan, kiến trúc tổng quan của hệ thống phản hồi liên quan, các kỹ thuật và những thách thức trong phản hồi liên quan.
- Học máy, học có giám sát, học không giám sát.
- Một số phương pháp học đa tạp để giảm chiều
- Môi trường thực nghiệm, tập dữ liệu ảnh thực nghiệm và phương pháp đánh giá độ chính xác.

4. Phạm vi nghiên cứu của luận án

Trong luận án này, phạm vi nghiên cứu bao gồm:

- Nghiên cứu phương pháp theo tiếp cận học đa tạp để tìm một ma trận chiếu tối ưu mà khai thác được cấu trúc phi tuyến của dữ liệu.
- Nghiên cứu phương pháp để cân bằng tập mẫu phản hồi thông qua việc bổ sung mẫu dương sử dụng đồ thị.
- Nghiên cứu phương pháp để khai thác một số khía cạnh hữu ích của đối tượng

- Dùng tập dữ liệu ảnh màu về phong cảnh được cộng đồng nghiên cứu về tra cứu ảnh sử dụng rộng rãi để sử dụng trong thực nghiệm.

5. Các đóng góp của luận án

Nhằm mục tiêu nâng cao độ chính xác của tra cứu ảnh sử dụng học máy để giảm chiều từ thông tin phản hồi của người dùng, luận án có các đóng góp sau:

(1) Đề xuất phương pháp tìm ma trận chiếu tối ưu theo tiếp cận học đa tạp [CT5]. Phương pháp này xem xét cấu trúc cục bộ của các mẫu dương và âm thuộc hai lân cận khác nhau để học một phép chiếu mà dữ liệu có thể phân biệt trên không gian chiếu, dẫn đến cải tiến độ chính xác cho tra cứu ảnh.

(2) Đề xuất phương pháp tự động bổ sung các mẫu dương vào tập huấn luyện để giải quyết vấn đề mất cân bằng tập huấn luyện [CT4]. Phương pháp này có thể: (a) bổ sung một số mẫu dương vào tập huấn luyện; (b) tận dụng các khía cạnh khác nhau của đối tượng để tạo ra một bộ phân lớp mạnh

6. Bố cục của luận án

Luận án được tổ chức thành ba chương:

Chương 1 giới thiệu tổng quan về tra cứu ảnh dựa vào nội dung, phản hồi liên quan và phân tích ưu nhược điểm một số phương pháp phản hồi liên quan nhằm giảm khoảng cách ngữ nghĩa. Chương này cũng trình bày học đa tạp cho tra cứu ảnh, một số lý thuyết liên quan về đồ thị, máy véc tơ hỗ trợ, tập dữ liệu ảnh thực nghiệm và cách thức đánh giá độ chính xác của hệ thống tra cứu ảnh.

Chương 2 mô tả phương pháp tìm ma trận chiếu tối ưu theo tiếp cận học đa tạp trong tra cứu ảnh, gọi là chiếu phân biệt lớp ngữ nghĩa cho tra cứu ảnh (SCDPIR - Semantic class discriminant projection for image retrieval), tận dụng các thông tin hình học cục bộ của các mẫu có nhãn và không có nhãn để giảm chiều. Sau khi có được ma trận chiếu, các ảnh trong không gian gốc có số chiều lớn sẽ được chiếu sang một không gian chiếu mới có số chiều nhỏ hơn nhiều. Trong không gian chiếu mới đó, các điểm dữ liệu vẫn có thể phân biệt tốt các mẫu liên quan so với các mẫu không liên quan. Bên cạnh đó, Chương 2 cũng đưa ra thực nghiệm trên tập dữ liệu được cộng đồng CBIR sử dụng rộng rãi: Corel 10,800 ảnh và minh họa kết quả chiếu trên tập SIMPLIcity.

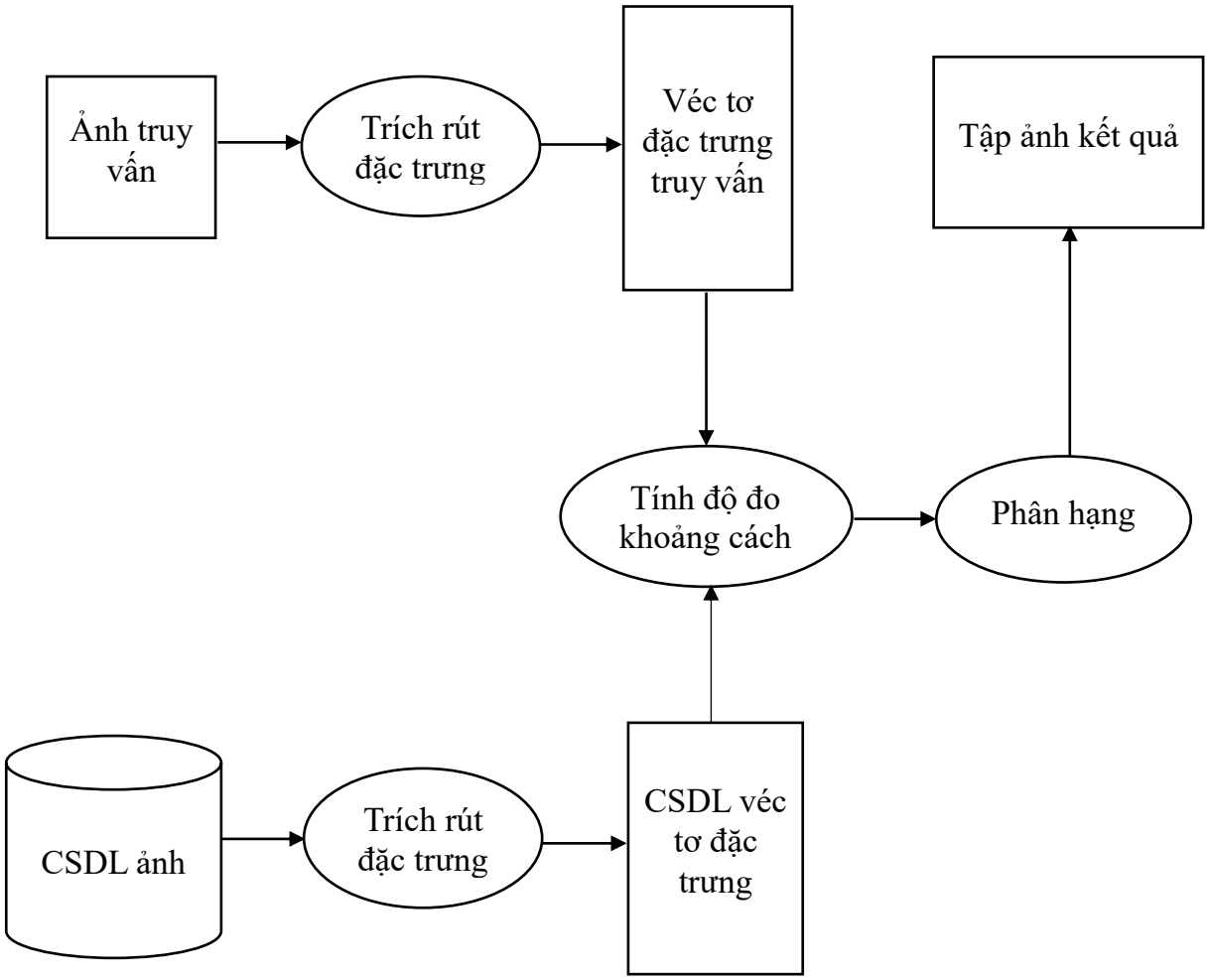
Chương 3 trình bày phương pháp cân bằng tập mẫu phản hồi và kết hợp tra cứu ảnh đa khía cạnh phương pháp thực hiện được các nội dung sau: (a) bổ sung mẫu dương (xác định nhãn cho các mẫu chưa có nhãn); (b) tận dụng thông tin của các mẫu phản hồi thuộc về hai lân cận khác nhau để xây dựng ma trận chiếu tối ưu mà trên không gian chiếu, dữ liệu có thể phân biệt hơn; (c) tận dụng các khía cạnh khác nhau của đối tượng để tạo ra một bộ phân lớp mạnh. Các kết quả thực nghiệm trên tập dữ liệu ảnh ảnh Corel 10800 ảnh chỉ ra rằng phương pháp đề xuất đã cải tiến đáng kể độ chính xác tra cứu của hệ thống. Cuối cùng, luận án đưa ra một số kết luận và định hướng nghiên cứu trong tương lai.

CHƯƠNG 1. TỔNG QUAN VỀ TRA CỨU ẢNH DỰA VÀO NỘI DUNG

1.1. Giới thiệu về tra cứu ảnh

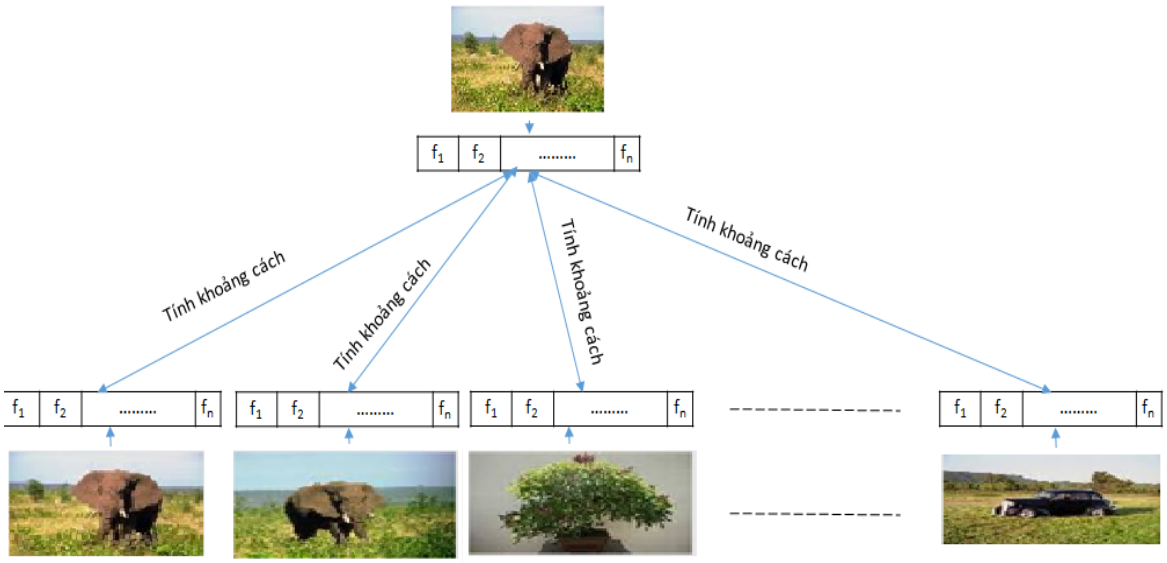
Với sự phát triển của các thiết bị thu nhận và lưu trữ ảnh, một lượng lớn dữ liệu ảnh được tạo ra. Việc tìm một ảnh đáp ứng yêu cầu của người dùng trong một bộ sưu tập lớn và đa dạng này là một nhiệm vụ khó khăn. Sự khó khăn này ngày càng gia tăng và trở thành một bài toán có nhiều thách thức. Yêu cầu khai thác dữ liệu ảnh trên một cách hiệu quả thúc đẩy sự quan tâm của cộng đồng nghiên cứu. Có hai cách tiếp cận chính cho bài toán tra cứu ảnh là tra cứu ảnh dựa vào văn bản và tra cứu ảnh dựa vào nội dung. Cách tiếp cận tra cứu ảnh dựa vào văn bản đáp ứng nhu cầu của người dùng thông qua kỹ thuật đối sánh từ khóa. Những nỗ lực đầu tiên để tổ chức ảnh dựa trên mô tả văn bản được bắt đầu từ đầu những năm 1970 [21]. Hình ảnh được lưu bằng các từ khóa theo sự kiện, địa điểm hoặc theo tên người. Các từ khóa này chủ yếu là do người dùng chú thích từ các ảnh một cách thủ công. Chú thích một tập ảnh lớn theo cách thủ công sẽ tốn nhiều công sức và chi phí thời gian lớn. Bên cạnh đó, việc chú thích này phụ thuộc vào nhận thức chủ quan của mỗi người dùng, dẫn đến cùng một ảnh, hai người khác nhau có thể có hai chú thích khác nhau. Vì thế, cách tiếp cận chú thích ảnh là không khả thi trên tập dữ liệu ảnh lớn. Cách tiếp cận tra cứu ảnh dựa vào nội dung (CBIR - content-based image retrieval) [22] được đề xuất vào đầu những năm 1980 để khắc phục vấn đề này. Cách tiếp cận này trích rút tự động nội dung ảnh, mà bao gồm đặc trưng màu, kết cấu, hình dạng, ….

Trong khi tra cứu ảnh dựa vào văn bản sử dụng một tập các từ khóa để mô tả nội dung bức ảnh, CBIR mô tả nội dung bức ảnh thông qua véc tơ đặc trưng mà thu được từ quá trình trích rút thông tin trên những điểm ảnh thô của ảnh. CBIR đã được nhiều tác giả nghiên cứu rộng rãi, nhiều phương pháp và hệ thống đã được phát triển. Nhiệm vụ của hệ thống CBIR là sử dụng một độ đo khoảng cách (hoặc độ đo tương tự) để đối sánh véc tơ đặc trưng của ảnh truy vấn với véc tơ đặc trưng của mỗi ảnh cơ sở dữ liệu (CSDL) và phân hạng chúng theo thứ tự giảm dần của độ tương tự. Hệ thống tra cứu ảnh chỉ dựa vào một độ đo khoảng cách để đối sánh ảnh truy vấn với ảnh cơ sở dữ liệu, luận án gọi là tra cứu ảnh truyền thống (hàm ý từ "truyền thống" ở đây là không có yếu tố học máy). Hình 1.1 là mô tả quá trình tra cứu ảnh dựa vào nội dung theo cách truyền thống.



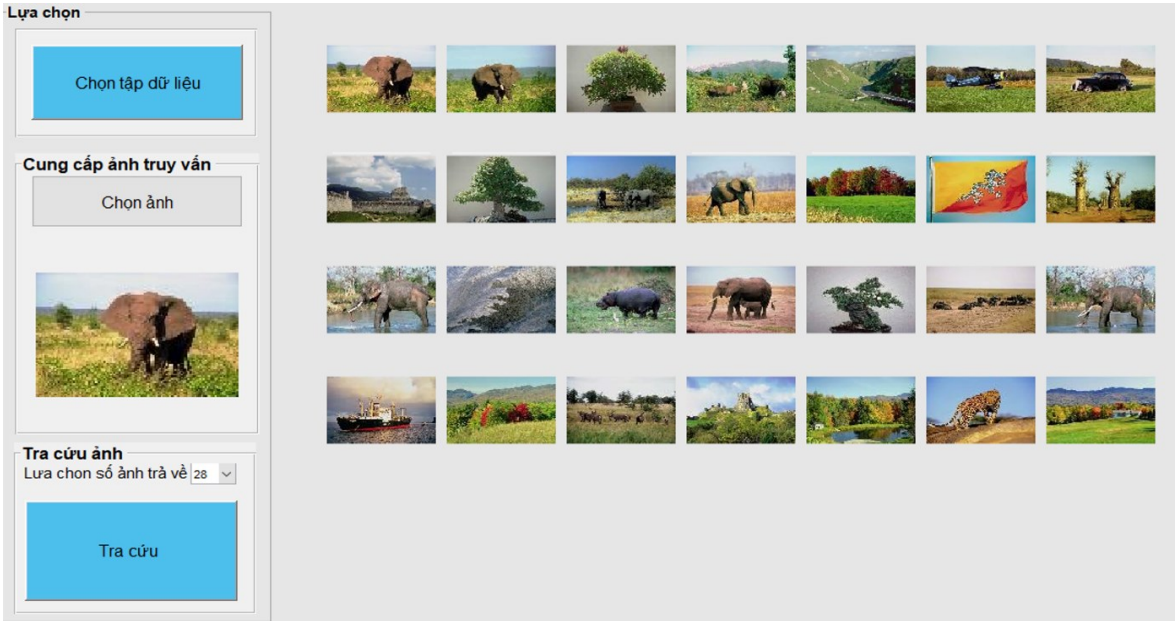
Hình 1.1. Sơ đồ tra cứu ảnh dựa vào nội dung truyền thống.

Trong Hình 1.1, tập cơ sở dữ liệu ảnh được đưa vào thủ tục trích rút đặc trưng để thu được CSDL véc tơ đặc trưng và lưu trữ chúng một cách thích hợp (pha này được thực hiện ngoại tuyến-offline). Trong pha tra cứu trực tuyến (online), người dùng cung cấp một ảnh làm ảnh truy vấn, thủ tục trích rút đặc trưng (giống như với pha offline) được thực hiện để thu được véc tơ đặc trưng truy vấn của ảnh truy vấn. Cũng lưu ý ở đây, độ dài véc tơ đặc trưng của ảnh truy vấn có cùng độ dài với véc tơ đặc trưng của ảnh CSDL. Tiếp theo, hệ thống so sánh lần lượt véc tơ đặc trưng của ảnh truy vấn với mỗi véc tơ đặc trưng của ảnh cơ sở dữ liệu theo một độ đo khoảng cách đã được xác định trước nào đó (như mô tả trong Hình 1.2) để thu được độ đo tương tự, sau đó, thủ tục phân hạng được thực hiện, tức là các ảnh cơ sở dữ liệu được sắp xếp theo thứ tự tăng dần của khoảng cách (vừa tính được) so với ảnh truy vấn. Tập ảnh kết quả thu được bao gồm K ảnh có khoảng cách nhỏ nhất với ảnh truy vấn sẽ được hiển thị cho người dùng.



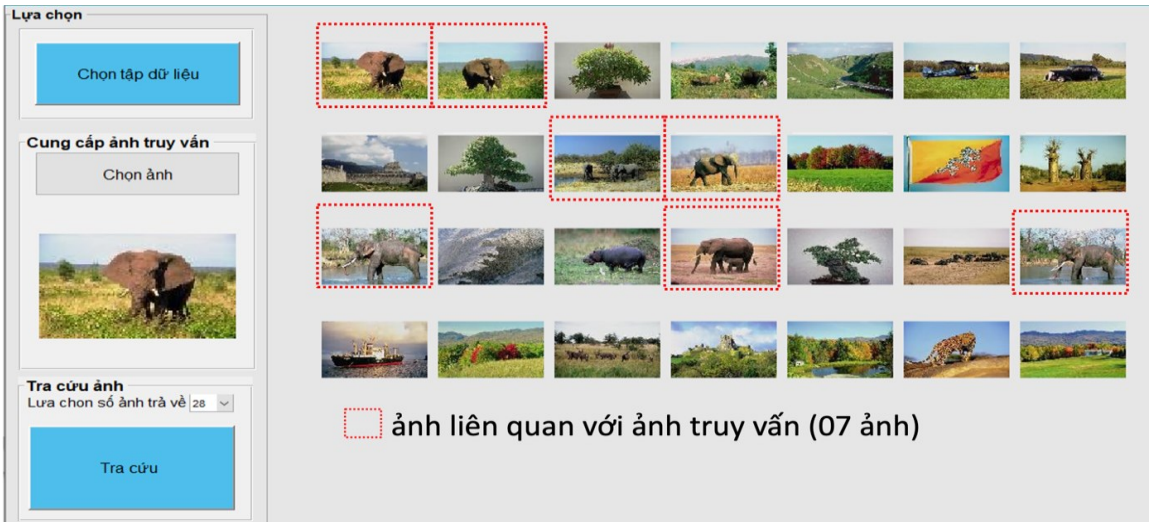
Hình 1.2. Minh họa việc đối sánh giữa ảnh truy vấn và mỗi ảnh CSDL.

Hình 1.3 là giao diện trực quan cho việc tra cứu ảnh truyền thống. Trong hình này, người dùng cung cấp ảnh con voi làm ảnh truy vấn, sau đó chúng ta thu được một tập kết quả bao gồm 28 ảnh kết quả.



Hình 1.3. Giao diện tra cứu ảnh truyền thống với ảnh truy vấn là ảnh con voi.

Trong tập kết quả thu được trên Hình 1.3, với một ảnh truy vấn là ảnh con voi, chúng ta thấy có 07 ảnh cùng chủ đề với ảnh truy vấn (hay liên quan với ảnh truy vấn). Các ảnh liên quan này được thể hiện bởi đường viền nét đứt bao quanh như Hình 1.4. Một số ảnh còn lại trong tập ảnh kết quả không có đường viền nét đứt bao là những ảnh có không liên quan với ảnh truy vấn.



Hình 1.4. Tập ảnh kết quả tra cứu bao gồm các ảnh liên quan và không liên quan.

Ban đầu, một số đặc trưng được trích rút mà thường được sử dụng trong các hệ thống CBIR bao gồm màu sắc, kết cấu, hình dạng (vùng và đường viền).... Các đặc trưng này thường được chia thành hai nhóm: thứ nhất, nhóm đặc trưng toàn cục mô tả toàn bộ hình ảnh, nhóm còn lại là đặc trưng cục bộ, mà chia ảnh thành các vùng nhỏ hơn.

Một số phương pháp tra cứu ảnh kết hợp các đặc trưng toàn cục khác nhau. Shrivastava và cộng sự [23] đã đề xuất một hệ thống CBIR gồm ba giai đoạn: Đầu tiên, trên đặc trưng màu, hệ thống thu được N ảnh từ tập dữ liệu M ảnh, tiếp theo, dựa vào đặc trưng kết cấu sử dụng bộ lọc Gabor, hệ thống thu được P ảnh liên quan trong tập N ảnh, Cuối cùng, tính toán bộ mô tả Fourier và lấy làm đặc trưng hình dạng để thu được K ảnh từ tập gồm P ảnh. Phương pháp được đánh giá thực nghiệm trên hai bộ dữ liệu Corel và Cifar với độ chính xác trung bình lần lượt là 0.77 và 0.86. Younus và cộng sự [24] đã xây dựng một hệ thống CBIR mới phụ thuộc vào đặc trưng màu và kết cấu gồm mô men màu, lược đồ màu, mô men wavelet và ma trận đồng xuất hiện. Hệ thống kết hợp thuật toán phân cụm K-mean với tối ưu bầy đàn (particle swarm optimization) trên tập dữ liệu Wang gồm 1000 ảnh với 10 chủ đề. Độ chính xác của hệ thống này có cải thiện tuy nhiên vẫn kém hiệu quả do nó không xem xét đặc trưng hình dạng và sai số trong quá trình phân cụm. Sajjad và cộng sự [25] đã đề xuất một hệ thống CBIR kết hợp các đặc trưng màu sắc và kết cấu để tạo thành một véc tơ đặc trưng 360 chiều. Để trích rút đặc trưng màu, lượng tử hóa thông qua lược đồ màu sau khi ảnh được chuyển sang không gian màu HSV. Mẫu nhị phân cục bộ

xoay được sử dụng để trích rút đặc trưng kết cấu bất biến với phép xoay. Các thực nghiệm được thực hiện trên tập Corel 1K và Corel 10K với độ chính xác tương ứng 0.67 và 0.7 trên độ thu hồi (recall) là 0.5. Nazir và cộng sự [26] đã trình bày một phương pháp tra cứu ảnh dựa trên nội dung sử dụng đặc trưng màu sắc và kết cấu. Trong phương pháp này, lược đồ màu được trích rút trong không gian HSV, và biến đổi wavalet rời rạc (discrete wavelet transform) và EHD được sử dụng làm đặc trưng kết cấu.

Ngoài ra, có nhiều phương pháp trích rút các đặc trưng không chỉ dựa trên toàn bộ ảnh mà thông qua các vùng được tách ra từ ảnh. Sharif và cộng sự [27] đề xuất một hệ thống CBIR phụ thuộc vào việc hợp nhất các từ trực quan (visual words) mà được tạo ra từ đặc trưng SIFT (scale invariant feature transform) và BRISK (binary robust invariant scalable keypoints). Yousuf và cộng sự [28] thực hiện một hệ thống CBIR dựa trên SFIT và LIOP (local intensity order pattern). LIOP đã được sử dụng để khắc phục hạn chế của SIFT trong việc thay đổi ánh sáng và các vùng có độ tương phản thấp. Việc sử dụng đặc trưng SIFT trong CBIR cho hiệu quả kém khi số chiều đặc trưng SIFT là rất lớn. Herbert và cộng sự [29] đề xuất đặc trưng SURF (speeded-up robust features) là một bộ mô tả cục bộ mạnh khác mà vượt qua giới hạn về số chiều cao của SIFT. SURF nhanh và mạnh hơn SIFT vì nó yêu cầu ít thời gian để tính toán và đối sánh các ảnh thông qua sử dụng cơ chế đánh chỉ số dựa trên tín hiệu Laplacian. Jabeen và cộng sự [30] đề xuất một hệ thống CBIR mới dựa trên việc kết hợp hai bộ mô tả SURF, FREAK (fast retina key point) để tạo thành các từ trực quan trên cơ sở của BoVW. Sau đó, phân cụm K-mean được áp dụng trên các từ trực quan đó để tính toán một lược đồ cho các từ của mỗi ảnh. Hệ thống thực nghiệm trên ba tập dữ liệu, bao gồm Caltech 256, Corel 1K và Corel1.5K, để chứng minh hiệu quả về độ chính xác trung bình, độ thu hồi (recall) trung bình.

Tuy nhiên, hiệu quả tra cứu ảnh sử dụng biểu diễn đặc trưng như trên (gọi là các đặc trưng thủ công - handcraft) là rất hạn chế bởi vì những đặc trưng thủ công này khó có thể mô tả ngữ nghĩa của ảnh. Gần đây, các hệ thống CBIR đã được chuyển sang sử dụng cách tiếp cận học sâu. Trong cách tiếp cận học sâu, một mô hình có thể xử lý dữ liệu ảnh gốc và tự khám phá ra đặc trưng tốt thông qua quá trình học. Trong [31], mô hình mạng nơ ron tích chập (CNN - Convolutional Neural Network) được

sử dụng để trích rút đặc trưng cho mỗi ảnh, nó giúp cải thiện việc tra cứu ảnh tương tự với ảnh truy vấn tốt hơn. Mô hình bao gồm các lớp tích chập (convolutional layer), các lớp gộp (pooling) và lớp kết nối đầy đủ (fully connected layer). Các lớp phía trước thường là các lớp tích chập kết hợp với các hàm kích hoạt phi tuyến và lớp pooling (được gọi chung là ConvNet), do vậy, đầu ra ở lớp gần cuối cùng trước khi chuyển qua lớp kết nối đầy đủ có thể được coi là véctơ đặc trưng hữu ích. Lớp cuối cùng là một mạng nơ ron kết nối đầy đủ và thường là một hàm softmax. Zheng và cộng sự [32] đã đề xuất một phương pháp CBIR dựa trên VGGNet [33] để trích rút đặc trưng. Phương pháp này được thực nghiệm trên ba tập, bao gồm Oxford Paris, Holidays, Caltech 101, và chỉ ra độ chính xác tốt hơn [31].

Các phương pháp tra cứu ảnh kể trên có thể nâng cao độ chính xác khi sử dụng cách tiếp cận học sâu, tuy nhiên chúng khá tốn thời gian để xử lý do số chiều của véc tơ đặc trưng thu được khá lớn và vẫn gặp phải vấn đề khoảng cách giữa đặc trưng mức thấp với cảm nhận trực quan của con người khi mô tả nội dung ảnh.

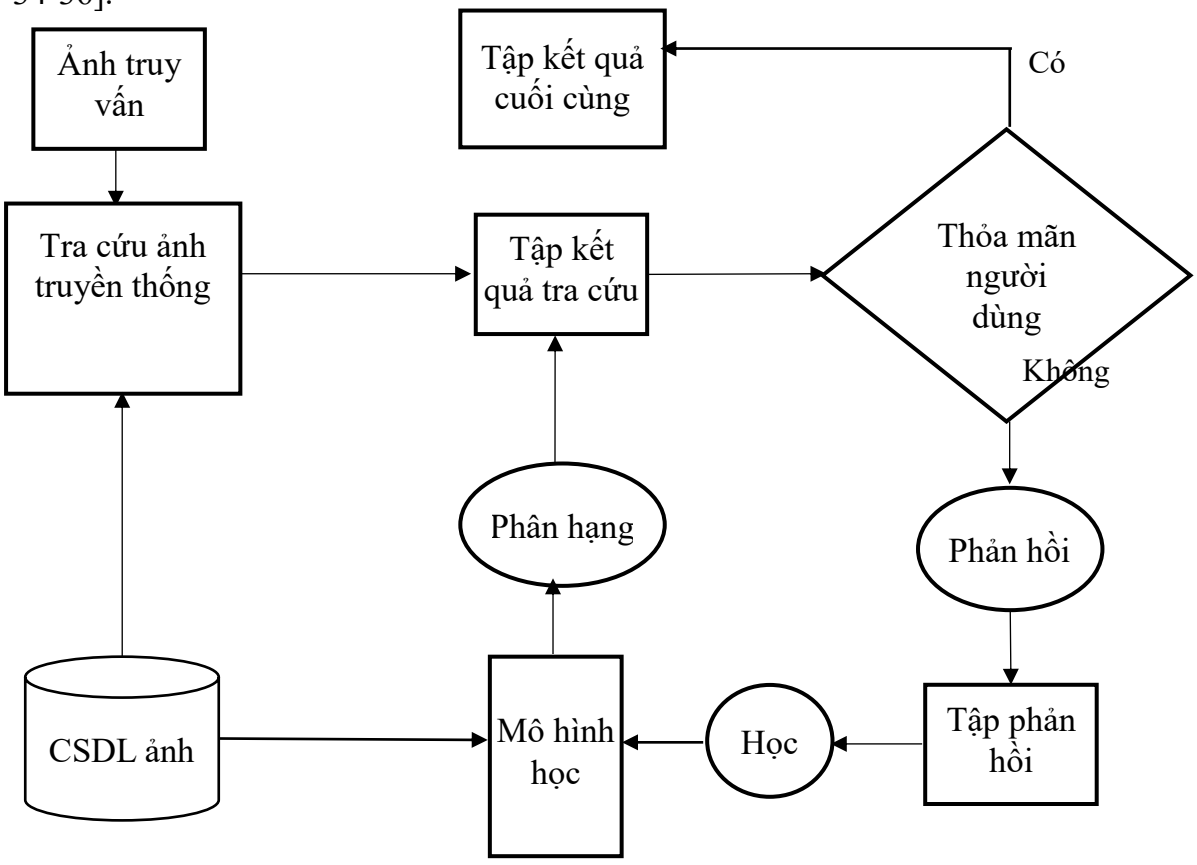
1.2. Giới thiệu về phản hồi liên quan

1.2.1. Cơ chế phản hồi liên quan

Các phương pháp CBIR theo cách tiếp cận truyền thống, dùng các đặc trưng thủ công, cho độ chính xác tra cứu không cao do khoảng trống ngữ nghĩa (Semantic gap) giữa các đặc trưng mức thấp của ảnh và nhận thức của con người về nội dung của ảnh. Trong Hình 1.5, nếu xét về khía cạnh đặc trưng mức thấp như lược đồ màu, ảnh bên trái (có ID là 644024) và ảnh bên phải (có ID là 644012) là rất khác nhau, tuy nhiên, theo nhận thức bằng mắt người, hai ảnh này là giống nhau (thực tế chúng thuộc về cùng chủ đề "khủng long" trong tập ảnh Corel).

Hình 1.5. Mình họa khoảng trống ngữ nghĩa giữa đặc trưng mức thấp và nhận thức của con người.

Để giảm khoảng trống này, cách tiếp CBIR với phản hồi liên quan (Relevant Feedback - RF) thường được sử dụng. Trong cách tiếp cận này, hệ thống bao gồm người dùng vào mỗi vòng lặp tra cứu, cụ thể, tại mỗi vòng lặp người dùng cung cấp thông tin phản hồi về sự giống và khác nhau của mỗi ảnh trong tập kết quả so với ảnh truy vấn. Người dùng cung cấp phản hồi bằng cách đánh dấu một số ảnh là "liên quan" (mẫu dương) hoặc "không liên quan" (mẫu âm) trong tập ảnh kết quả tra cứu ở phiên hiện tại. Những phản hồi này được xem là các mẫu trong tập huấn luyện để hệ thống CBIR học các đặc trưng trực quan của ảnh nhằm cải thiện độ chính xác của tập ảnh kết quả tra cứu ở phiên tiếp theo. Cơ chế phản hồi liên quan này được mô tả trên 0 ở dưới, nó được nghiên cứu rộng rãi để giảm khoảng trống ngữ nghĩa [9, 14, 34-36].



Hình 1.6. Sơ đồ tra cứu ảnh với phản hồi liên quan.

Để hiểu rõ hơn sơ đồ tra cứu ảnh với phản hồi trong 0, luận án sẽ minh họa lựa chọn phản hồi của người dùng dựa trên tập ảnh kết quả tra cứu khởi tạo trong Hình 1.4 ở Hình 1.7. Với ảnh truy vấn là ảnh con voi, tập kết quả tra cứu khởi tạo (dùng độ đo khoảng cách Euclid) bao gồm 28 ảnh, người dùng chọn 07 ảnh là liên quan (mẫu dương), 14 ảnh là không liên quan (mẫu âm), và 07 ảnh còn lại là chưa gán nhãn

như Hình 1.7. Chúng ta thấy rằng, những ảnh thuộc lớp dương (lớp liên quan với ảnh truy vấn) là thuộc cùng một chủ đề là "voi" tuy nhiên những ảnh thuộc lớp âm (lớp không liên quan với ảnh truy vấn) lại nằm rải rác ở nhiều chủ đề khác nhau còn lại.

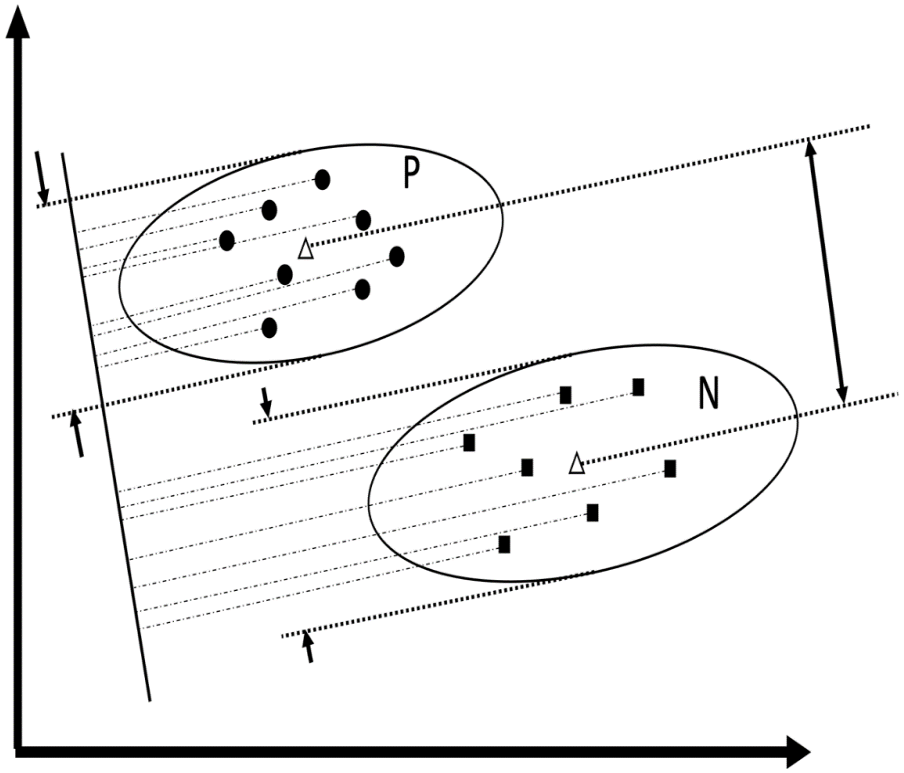
Hình 1.7. Chọn ảnh phản hồi trên tập kết quả tra cứu.

Sau khi có tập phản hồi liên quan, cách tiếp cận học máy được thực hiện để có được mô hình học máy. Áp dụng mô hình học máy cụ thể nào đó (chẳng hạn phương pháp học máy SVM) vào tra cứu, chúng ta thu được tập kết quả tra cứu như trên Hình 1.8. Tập kết quả trên Hình 1.8 bao gồm 12 ảnh liên quan đến ảnh truy vấn. Như vậy, độ chính xác tra cứu đã được cải thiện sau khi có thông tin phản hồi từ người dùng.

Hình 1.8. Kết quả tra cứu sau khi người dùng phản hồi.

1.2.2. Học đa tạp trong tra cứu ảnh dựa vào nội dung

Các bộ dữ liệu trong không gian chiều cao có thể rất khó hình dung, trong khi dữ liệu ở hai hoặc ba chiều có thể được vẽ để thể hiện cấu trúc vốn có của dữ liệu và trực quan hơn nhiều. Để hỗ trợ hình dung về cấu trúc của tập dữ liệu, kích thước phải được giảm theo một cách nào đó. Cách đơn giản nhất để thực hiện việc giảm kích thước này là thực hiện một phép chiếu dữ liệu trên không gian gốc cao chiều sang một không gian chiếu thấp chiều hơn. Một số phương pháp tra cứu ảnh với RF sử dụng phép chiếu mà coi các ảnh phản hồi nằm trong một không gian toàn cục, và LDA là một phương pháp như thế. LDA cố gắng tìm một phép chiếu đảm bảo các điểm thuộc cùng một lớp (có cùng nhãn) sẽ gần nhau và tách xa các điểm không thuộc cùng một lớp (khác nhãn). Xét ví dụ về bài toán chiếu với hai lớp được mô tả trong Hình 1.9.

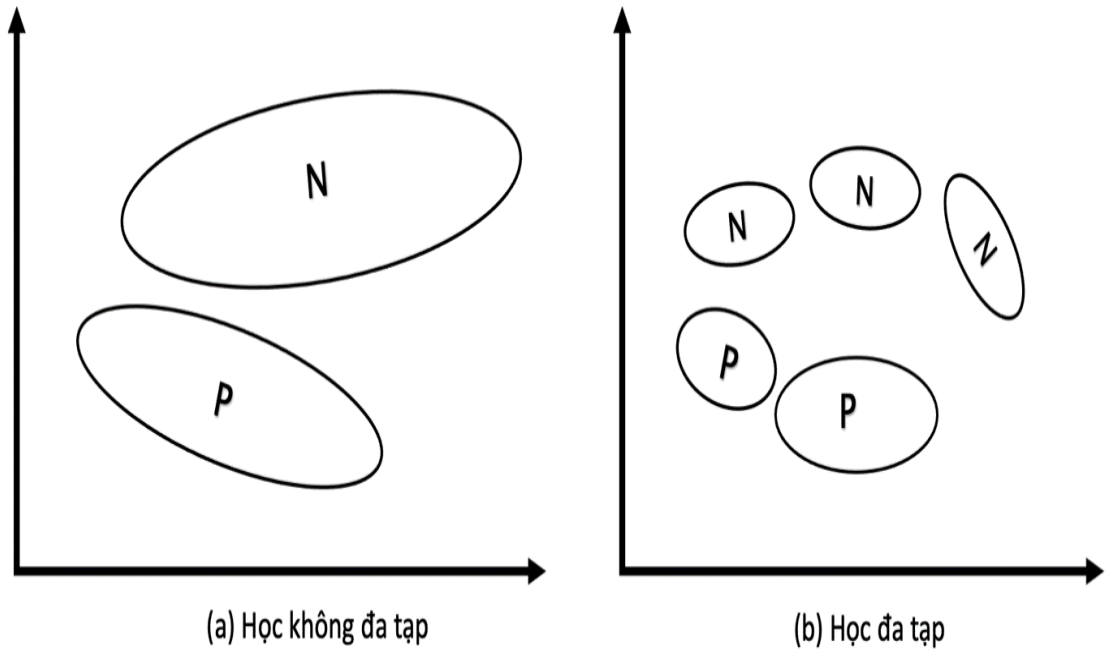


Hình 1.9. Chiếu phân tích phân biệt tuyến tính.

Hình 1.9 có hai lớp, các điểm hình tròn biểu diễn cho các điểm thuộc lớp dương (P – Positive) và các điểm hình vuông biểu thị các điểm thuộc lớp âm (N - Negative). Số chiều của dữ liệu được giảm về một chiều thông qua chiếu chúng lên một đường thẳng và mỗi điểm được đại diện bởi hình chiếu của nó lên đường thẳng

kia. LDA tìm một không gian chiếu mới sao cho hình chiếu của các điểm dữ liệu cùng lớp sẽ gần nhau trong khi hình chiếu của các điểm dữ liệu nằm ở hai lớp khác nhau sẽ xa nhau. Tức là các điểm hình tròn sẽ gần nhau trên không gian chiếu còn điểm hình vuông sẽ cách xa các điểm hình tròn bằng cách tối đa khoảng cách trọng tâm của hai lớp (điểm hình tam giác) hoặc ngược lại. Và ta thấy, LDA xem xét tất cả điểm thuộc lớp dương hoặc âm nằm trong một không gian toàn cục.

Một số phương pháp chiếu theo tiếp cận học đa tạp (được hiểu là tìm một ma trận chiếu theo tiếp cận học đa tạp) không xem xét các mẫu phản hồi dương hay âm nằm trong một không gian toàn cục. Xét Hình 1.10, trong đó Hình 1.10 (a) coi các mẫu dương nằm trong một không gian (hình elip bao quanh chữ P) trong khi các mẫu âm nằm trong một không gian khác (hình elip bao quanh chữ N). Tuy nhiên, dưới góc nhìn học đa tạp, các điểm thuộc lớp dương có thể nằm trên nhiều không gian con khác nhau (hai hình elip bao quanh hai chữ P trên Hình 1.10 (b)), trong khi các điểm thuộc lớp âm nằm trên ba không gian con khác nhau (ba hình elip bao quanh ba chữ N trên Hình 1.10 (b)).



Hình 1.10. Minh họa dữ liệu trên không gian đa tạp cho RF.

Việc học đa tạp với mục tiêu là tạo ra một không gian chiếu nơi mà các ảnh liên quan được chiếu gần nhau trong khi các ảnh không liên quan được chiếu cách xa nhau bằng cách học cấu trúc cục bộ được hình thành bởi lân cận của ảnh truy vấn và

ảnh được phản hồi. Điều này đạt được bằng cách nhúng ảnh truy vấn và tập ảnh phản hồi như tập điểm dữ liệu (các nút) trong đồ thị k lân cận gần nhất, sử dụng ma trận trọng số cho biết trọng số trên mỗi cạnh. Ánh xạ tối ưu được tìm thấy dựa trên ma trận trọng số này, sao cho các điểm lân cận trong đồ thị được ánh xạ với nhau bằng cách tối thiểu hàm chi phí. Mỗi ảnh cơ sở dữ liệu sau đó cũng được ánh xạ sang không gian chiếu mới, thu được kết quả tra cứu mới là tập ảnh lân cận gần nhất với ảnh truy vấn. Sau mỗi vòng phản hồi, cấu trúc cục bộ của không gian đa tạp lại được học lại. Thông thường không phải tất cả các ảnh trong cơ sở dữ liệu sẽ được sử dụng để xây dựng đồ thị lân cận gần nhất. Để giảm độ phức tạp tính toán, chỉ một vài chục ảnh được xếp hạng trên cùng của danh sách từ lần tra cứu trước đó được sử dụng cùng với tập ảnh đã được gán nhãn từ thông tin phản hồi của người dùng.

1.2.3. Rà soát một số nghiên cứu liên quan

Ban đầu, cách tiếp cận tra cứu ảnh với RF giả thiết rằng, tồn tại của một điểm truy vấn lý tưởng mà nếu tìm thấy được sẽ cho kết quả phù hợp với mong muốn của người dùng. Cách tiếp cận này được gọi là "dịch chuyển điểm truy vấn" (QPM - Query Point Movement). Tại mỗi vòng lặp, điểm truy vấn mới sẽ gần với điểm truy vấn lý tưởng hơn. Cũng trong cách tiếp cận này, các trọng số độ quan trọng của chiều đặc trưng trong không gian đặc trưng cũng được cập nhật theo. Cách tiếp cận học máy cũng đa dạng, chẳng hạn học máy có thể chỉ học trên các mẫu dương trong [37], dựa vào khoảng cách Mahalanobis trong [38] hoặc có thể học trên tập mẫu huấn luyện gồm cả các mẫu dương và âm như một số phương pháp trong [39-41].

Một số phương pháp tra cứu ảnh với RF sau đó thường dựa vào máy véc tơ hỗ trợ (SVM - Support Vector Machine) [20, 42-45] để phân tách các mẫu trong toàn bộ tập dữ liệu theo biên quyết định. Phương pháp máy véc tơ hỗ trợ một lớp [42] chỉ quan tâm đến các mẫu được gán nhãn là dương trong tập phản hồi và bỏ qua các mẫu còn lại. SVM hai lớp [46] đã sử dụng thông tin của cả các mẫu phản hồi dương và âm nhưng trọng số quan trọng gán cho hai nhóm này lại ngang bằng nhau. Tiếp theo, trong [44] dùng học tích cực sử dụng SVM, sau khi tìm được biên quyết định phân tách hai nhóm dương và âm, nó quan tâm đến các mẫu gần với biên cho người dùng gán nhãn đưa vào học lại mô hình. Guo và cộng sự [45] sử dụng khoảng cách Euclid, dựa trên biên quyết định tìm được, để đối sánh và xếp hạng các ảnh theo ảnh truy vấn

mà nằm trong lề. Tác giả trong [47] đề xuất một phương pháp tận dụng các mẫu được gán nhãn trong tập phản hồi, các mẫu dương coi là một nhóm còn các mẫu âm được chia thành một số nhóm nhỏ. Độ chính xác của phương pháp này đã được cải thiện đáng kể nhưng vẫn cần nghiên cứu cải tiến độ chính xác của chúng để đáp ứng đòi hỏi thực tế.

Tại Việt Nam, trong [48] tác giả Vũ Văn Hiệu đề xuất kỹ thuật chuẩn hóa $3\sigma -$ FCM với dữ liệu đặc trưng ảnh sử dụng trong CBIR giúp nâng cao chất lượng độ tương tự của các bộ đặc trưng tăng độ chính xác khi tra cứu. Kỹ thuật này đã khắc phục các hạn chế của kỹ thuật chuẩn hóa min-max và chuẩn hóa Gaussian với dữ liệu đặc trưng ảnh không đồng nhất. Bên cạnh đó, tác giả còn đề xuất một kỹ thuật trong [49] có thể nâng cao độ chính xác tra cứu ảnh dựa vào nội dung thông qua cách tiếp cận tối ưu Pareto để xây dựng tập ứng viên có kích cỡ nhỏ. [50] đề xuất một kỹ thuật chọn các mẫu không có nhãn một cách hiệu quả để gán nhãn cho quá trình học tích cực sử dụng SVM. Ngoài ra trong [51] cũng đề xuất một kỹ thuật tra cứu ảnh dựa vào nội dung có sử dụng phản hồi của người dùng với truy vấn đa đặc trưng và tích phân Choquet. Tác giả Đào Thị Thúy Quỳnh và cộng sự [52] đề xuất phương pháp tra cứu ảnh mà không đòi hỏi người dùng phải cung cấp đồng thời nhiều truy vấn đa dạng (giảm gánh nặng cho người dùng). Bên cạnh đó, phương pháp đó tận dụng sự đánh giá của người dùng để xác định độ quan trọng ngữ nghĩa của từng truy vấn và độ quan trọng của từng đặc trưng. Tuy các phương pháp kể trên có cải thiện được độ chính xác sau khi tra cứu ảnh nhưng nó vẫn gặp phải hạn chế về cỡ mẫu tập phản hồi nhỏ trong quá trình RF.

Trong tra cứu ảnh với phản hồi liên quan, các mẫu do người dùng cung cấp thường rất nhỏ so với chiều của đặc trưng, do đó chúng ta phải giải quyết bài toán gọi là "lời nguyền về số chiều - curse of dimensionality". Khi số chiều đặc trưng quá lớn so với số lượng mẫu trong tập huấn luyện, các mô hình học máy có thể rơi vào tình trạng quá khớp. Để giải quyết vấn đề này, một số tác giả đề xuất các kỹ thuật giảm chiều như phân tích thành phần chính (PCA- Principal Components Analysis) [53, 54] và phân tích phân biệt tuyến tính (LDA - Linear Discriminant Analysis) [55]. PCA tìm một phép chiếu mà trên đó phương sai là cực đại. LDA tìm một phép chiếu trên đó các mẫu có cùng nhãn được gom thành một cụm, những mẫu có nhãn khác

nhau sẽ nằm ở các cụm khác nhau. Áp dụng LDA trong bài toán giảm chiều dữ liệu cho tra cứu ảnh có độ chính xác tốt hơn PCA. Tuy nhiên khi áp dụng LDA vào tra cứu ảnh với RF, việc thu thập số các mẫu có nhãn lớn là không khả thi [56]. Cả PCA và LDA chỉ xét cấu trúc Euclide mà khám phá cấu trúc toàn cục của không gian, nên cấu trúc cục bộ được hình thành bởi ảnh truy vấn và tập mẫu phản hồi được gán nhãn bị bỏ qua, bởi vì không gian các đặc trưng trực quan mức thấp của ảnh có thể là một đa tạp [3, 15].

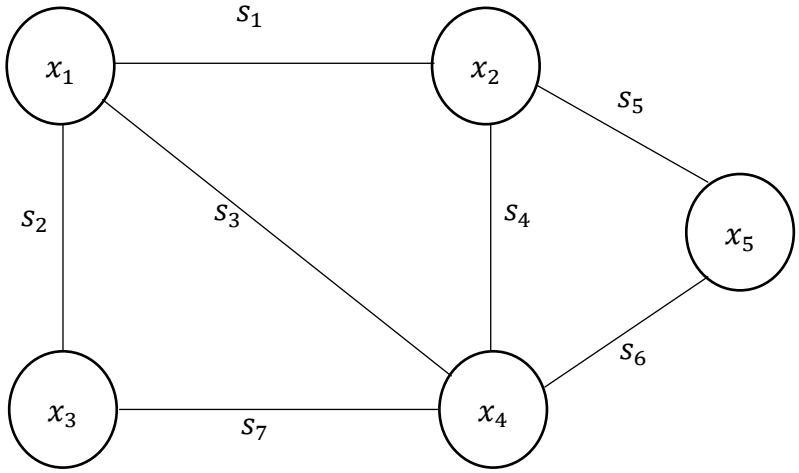
Trong những năm gần đây, có nhiều thuật toán học đa tạp đã được đề xuất để khám phá cấu trúc đa tạp. Phương pháp của He và cộng sự [3] đề xuất, có tên là phép chiếu bảo toàn tính cục bộ (Locality Preserving Projections – LPP), thực hiện phép chiếu bảo toàn cục bộ để tìm một xấp xỉ tuyến tính của đa tạp dữ liệu nội tại. Dựa trên LPP, trong [57] đề xuất một phương pháp tra cứu ảnh thực hiện so sánh ảnh trong không gian con bởi độ đo Euclide. Phương pháp tra cứu ảnh học một đa tạp ngữ nghĩa với sự nhu cầu của người dùng, có tên là nhúng quan hệ gia tăng (Augmented Relation Embedding - ARE), được đề xuất trong [15]. Tiếp theo, một phương pháp tra cứu ảnh có thể khám phá cấu trúc đa tạp cục bộ thông qua cực đại lề giữa các mẫu âm và dương ở mỗi lân cận cục bộ, có tên chiếu lề cực đại cho tra cứu ảnh (Maximum Margin Projection – MMP), đã được để xuất bởi [16]. Một số phương pháp xem xét trường hợp khi dữ liệu nằm trên một không gian con của không gian gốc bao gồm nhúng tuyến tính cục bộ (Locally Linear Embedding - LLE) [5], Isomap [58] và Laplacian Eigenmaps [59]. Phương pháp được đề xuất bởi Li và cộng sự [60] đã bảo toàn thông tin phân biệt trong việc mã hóa bởi việc kết hợp học không gian con với nguyên lý lề cực đại. Các phương pháp này ước lượng cả thuộc tính hình học và phân biệt của đa tạp con của các điểm ngẫu nhiên nằm trên đa tạp con chưa biết này. Tuy nhiên, các phương pháp này chỉ thực hiện được với các điểm dữ liệu trong tập huấn luyện, và nó không đưa ra rõ ràng phép chiếu có thể thực hiện cho các điểm dữ liệu kiểm tra mới. Bên cạnh đó, các phương pháp này chỉ xem xét tính chất hình học trong một lớp, trong khi bỏ qua mối liên hệ của các mẫu từ các lớp khác nhau. Mặt khác, các phương pháp thường không quan tâm đến các ảnh thuộc lân cận khác nhau mặc dù chúng có thể vẫn liên quan với truy vấn. Do đó, các phương pháp tra cứu ảnh này thường có hiệu quả hạn chế.

1.3. Lý thuyết liên quan đến luận án

Trong phần này, luận án trình bày một số kiến thức lý thuyết đồ thị mà phục vụ cho bài toán học bán giám sát trong phương pháp tra cứu ảnh đề xuất.

1.3.1. Giới thiệu về đồ thị

Định nghĩa 1.1 (Đồ thị vô hướng không có khuyên): Đồ thị vô hướng $\mathbf{G} = (\mathbf{X}, \mathbf{S})$ gồm hai thành phần: **X** là tập đỉnh và **S** là tập cạnh, mỗi cạnh là một cặp không có thứ tự, hay một tập gồm hai đỉnh trong **X** (tức là cặp $\{\mathbf{u}, \mathbf{v}\} \subseteq \mathbf{X^2}$ với $\mathbf{u}, \mathbf{v} \in \mathbf{X}$ và $\mathbf{u} \neq \mathbf{v}$).



Hình 1.11. Minh họa đồ thị vô hướng G_1.

Định nghĩa 1.2 (Đỉnh lân cận): Cho một đồ thị vô hướng $\mathbf{G} = (\mathbf{X}, \mathbf{S})$ với hai đỉnh $\mathbf{u}, \mathbf{v} \in \mathbf{X}$, gọi là lân cận nếu $\{\mathbf{u}, \mathbf{v}\} \in \mathbf{S}$

Nếu tất cả các đỉnh trong **G** đều lân cận với nhau, thì **G** là một đồ thị đầy đủ

Định nghĩa 1.3 (Bậc của đỉnh trong đồ thị vô hướng): Cho một đồ thị vô hướng $\mathbf{G} = (\mathbf{X}, \mathbf{S})$ với một đỉnh $\mathbf{v} \in \mathbf{X}$, bậc $\boldsymbol{d}(\mathbf{v})$ của **v** là số lượng cạnh kề với **v**

$$\boldsymbol{d}(\mathbf{v}) = |\{\mathbf{u} \in \mathbf{X} \,|\{\mathbf{u}, \mathbf{v}\} \in \mathbf{S}\}| \tag{1.1}$$

Trong đó, kí hiệu $|\mathbf{X}|$ là số lượng đỉnh của đồ thị. Có thể viết tắt $\boldsymbol{d}(\mathbf{v}_i) = \boldsymbol{d}_i, i = 1..m$.

Định nghĩa 1.4 (Ma trận bậc): Cho một đồ thị vô hướng $\mathbf{G} = (\mathbf{X}, \mathbf{S})$ với mỗi một đỉnh $\mathbf{v}_i \in \boldsymbol{X}$, $\boldsymbol{d}_i$ là bậc của đỉnh $\boldsymbol{v}_i$ thì ma trận bậc $\mathbf{D}(\mathbf{G})$ là một ma trận chéo, cụ thể ma trận bậc được biểu diễn bởi

$$\mathbf{D}(\mathbf{G}) = \boldsymbol{diag}(\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_m) \tag{1.2}$$

Có thể viết tắt $\mathbf{D}$ thay cho $\mathbf{D}(\mathbf{G})$ trừ khi có sự nhầm lẫn. Ví dụ ma trận bậc của đồ thị G_1 là:

$$\mathbf{D}(\mathbf{G}_1) = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Định nghĩa 1.5 (Ma trận liên thuộc của đồ thị vô hướng): Cho một đồ thị vô hướng $\mathbf{G} = (\mathbf{X}, \mathbf{S})$ với $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ và $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$, ma trận liên thuộc $\mathbf{B}(\mathbf{G})$ của đồ thị $\mathbf{G}$ là một ma trận có kích thước $m \times n$ trong đó mỗi phần tử b_{ij} được tính như sau:

$$b_{ij} = \begin{cases} 1 \text{ nếu đỉnh } \mathbf{x}_i \text{ là một đỉnh của cạnh } \mathbf{s}_j \\ 0 \text{ nếu ngược lại} \end{cases} \tag{1.3}$$

$$\mathbf{B}(\mathbf{G}_1) = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Định nghĩa 1.6 (Ma trận kề của đồ thị vô hướng): Cho một đồ thị vô hướng $\mathbf{G} = (\mathbf{X}, \mathbf{S})$ với $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ và $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$, ma trận kề (adjacency matrix) **A(G)** của đồ thị **G** là một ma trận đối xứng có kích thước $m \times m$ trong đó mỗi phần tử a_{ij} được cho bởi:

$$a_{ij} = \begin{cases} 1 \text{ nếu có một cạnh } \{\mathbf{x}_i, \mathbf{x}_j\} \in \mathbf{S} \\ 0 \text{ nếu ngược lại} \end{cases} \tag{1.4}$$

Nói một cách khác, trong ma trận kề mỗi phần tử tại dòng *i* cột *j* cho biết có cạnh nối từ đỉnh thứ *i* tới đỉnh thứ *j*.

Do đó, với ma trận kề **A** thì tổng trọng số $\boldsymbol{d}_i$ của đỉnh $\mathbf{v}_i$ được viết thành:

$$\boldsymbol{d}(\mathbf{v}_i) = \sum_{j=1}^{m} a_{ij} \tag{1.5}$$

Dưới đây là ma trận kề của $\mathbf{G_1}$:

$$\mathbf{A}(\mathbf{G_1}) = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Định nghĩa 1.7 (Đồ thị có trọng số không âm): Đồ thị trọng số $= (\mathbf{X}, \mathbf{S})$ với $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ và $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$, và ma trận trọng số $\mathbf{W}$ cỡ $m \times m$ là đối xứng sao cho $w_{ij} \geq 0, i, j \in \{1, 2, \dots, m\}$ và $w_{ii} = 0$ với $i = 1, 2, \dots, m$. Một cặp đỉnh $\{\mathbf{x_i}, \mathbf{x_j}\}$ là một cạnh nếu $w_{ij} > 0$.

Vì $w_{ii} = 0$ nên trong đồ thị trọng số, chúng ta có thể hiểu ma trận W như là một ma trận kề tổng quát. Trong trường hợp mà $w_{ii} \in \{0, 1\}$ thì nó trở thành ma trận kề như trong định nghĩa 1.4. Tùy từng bài toán mà ta coi trọng số w_{ij} của một cạnh $\{\mathbf{v_i}, \mathbf{v_j}\}$ là độ tương tự/ khoảng cách giữa hai đối tượng.

Với mỗi đỉnh $\mathbf{v}$ thứ i trong $\mathbf{X}$, tổng trọng số $\boldsymbol{d}(\mathbf{v})$ của $\mathbf{v}$ là tổng trọng số của các cạnh kề với $\mathbf{v}$.

$$d(\mathbf{v}) = \sum_{j=1}^{m} w_{ij} \tag{1.6}$$

1.3.2. Máy véc tơ hỗ trợ

Trong phần này, luận án trình bày ngắn gọn về các máy véc tơ hỗ trợ [43], và sử dụng nó làm cơ sở cho việc phân hạng trong pha phản hồi liên quan của hệ thống đề xuất mà được giới thiệu trong các chương sau.

Đối với tập dữ liệu huấn luyện $\mathbf{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, $\mathbf{x}_i$ là một véc tơ đặc trưng thuộc không gian đặc trưng R^m, với m là chiều của không gian đặc trưng, và $\mathrm{D_{label}} = \{y_1, y_2, \dots, y_n\}$, $y_i \in \{-1, 1\}$ tương ứng là các nhãn của chúng. Mục đích của SVM là xác định một siêu phẳng phân tách (hay biên quyết định) có thể phân chia các điểm trong tập $\mathbf{D}$ thành hai tập sao cho các điểm có cùng nhãn sẽ nằm ở cùng một phía của siêu phẳng phân tách.

$$f(\mathbf{x}) = \mathbf{w}.\mathbf{x} + b = 0, \mathbf{w} \in R^m, b \in R \tag{1.7}$$

với $\mathbf{x}$ là véc tơ đầu vào, $\mathbf{w}$ là véc tơ trọng số và b là độ lệch. Siêu phẳng phân tách này là đường thẳng nếu không gian là hai chiều, và là mặt phẳng nếu không gian là ba chiều, và tổng quát hơn với không gian R^m thì nó là một siêu phẳng phân tách m-1 chiều.

SVM xác định hai tham số **w** và *b* tương ứng là véc tơ trọng số và độ lệch; siêu phẳng phân tách tối ưu được tạo nên bằng cách cực đại lề hình học sao cho tất cả các điểm dữ liệu (x_i, y_i) đều thỏa mãn:

$$y_i(\mathbf{w}.\mathbf{x}_i + b) \geq 1, i = 1, \dots, n \quad (1.8)$$

Những điểm nằm gần biên quyết định nhất được gọi là các véc tơ hỗ trợ, và có khoảng cách bằng $1/\|\mathbf{w}\|$, tức là $y_i f(\mathbf{x}_i) = 1$, đại lượng $2/\|\mathbf{w}\|$ được gọi là lề, và biên quyết định là siêu phẳng với lề cực đại. Như vậy việc tìm siêu phẳng tối ưu tức là đi tìm siêu phẳng có giá trị $\|\mathbf{w}\|^2$ nhỏ nhất thỏa mãn:

$$\min h(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \quad (1.9)$$

thỏa mãn $y_i(\mathbf{w}.\mathbf{x}_i + b) \geq 1, i = 1, \dots, n$

Với $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$ là nhân tử Lagrange khác không. Bài toán tối ưu tổng quát trở thành:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = h(\mathbf{w}) - \sum_{i=1}^{n} \alpha_i (y_i(\mathbf{w}.\mathbf{x}_i + b) - 1) \quad (1.10)$$

$$\text{hay } \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(\mathbf{w}.\mathbf{x}_i + b) - 1) \quad (1.11)$$

Đặt đạo hàm của hàm số Lagrangian bằng không đối với các biến số **w**, và b, chúng ta có các mối quan hệ sau:

$$\frac{\partial}{\partial w}\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \iff \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad (1.12)$$

$$\frac{\partial}{\partial b}\mathcal{L}(\mathbf{w}, b, \alpha) = \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (1.13)$$

Thay thế (1.12) và (1.13) vào $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$ ta có hàm mục tiêu

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j \, y_i \, y_j \, \mathbf{x}_i \, \mathbf{x}_j \quad (1.14)$$

thỏa mãn $\sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, \dots, n$

Hàm quyết định cho phép phân lớp một mẫu z được cho bởi công thức:

$$class(z) = sign(f(z)) = sign(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i * z + b) \quad (1.15)$$

Trong công thức (1.15) dấu * ở đây là phép nhân vô hướng hai véc tơ. Khi dữ liệu là phân tách phi tuyến tính, SVM dùng hàm kernel với $\mathbf{u}, \mathbf{v}$ là hai véc tơ

$$G(\mathbf{u}, \mathbf{v}) = < \varphi(\mathbf{u}), \varphi(\mathbf{v}) > \quad (1.16)$$

Dữ liệu sẽ được ánh xạ sang một không gian mới bằng hàm φ với số chiều lớn hơn mà có thể phân tách tuyến tính. Hàm quyết định trở thành.

$$class(\mathbf{z}) = sign(f(z)) = sign(\sum_{i=1}^{s} \alpha_i y_i G(\mathbf{x}_i, \mathbf{z}) + b) \quad (1.17)$$

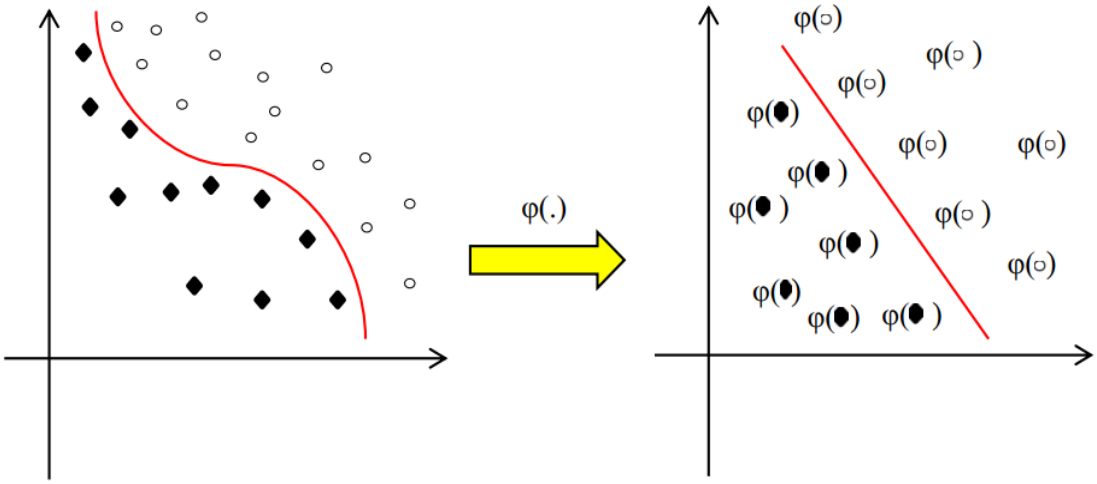
Trong đó, $f(\mathbf{z})$ là đầu ra của hàm siêu phẳng quyết định của SVM và s là số lượng các điểm véc tơ hỗ trợ. Hiệu quả của bộ phân lớp SVM phụ thuộc vào số lượng điểm véc tơ hỗ trợ.

Hàm nhân cơ sở bán kính (RBF - Radial Basis Function)

Nhân RBF hay còn gọi là nhân Gaussian thông qua hàm φ ánh xạ các véc tơ sang một không gian mới có số chiều lớn hơn mà tại đó có thể phân tách được dữ liệu bởi siêu phẳng tách. Nhân RBF được định nghĩa như sau:

$$K(x,y) =< \varphi(x), \varphi(y) >= \exp(-\|x-y\|^2) \quad (1.18)$$

Hàm nhân (1.18) tính tích vô hướng của hai véc tơ x và y. Giá trị $\|x-y\|$ sẽ nhỏ nếu hai véc tơ này tương tự, nên $-\|x-y\|^2$ cho giá trị lớn, do vậy các véc tơ nằm gần nhau thì có giá trị hàm nhân RBF lớn hơn so với các véc tơ cách xa nhau.



Hình 1.12. Minh họa hàm nhân RBF trong SVM.

Khoảng cách của một ảnh (một điểm hay một véc tơ) tới biên quyết định

Khoảng cách từ một điểm $\mathbf{z}$ tới biên quyết định H cần tìm với $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ là k véc tơ hỗ trợ cùng nhãn tương ứng $\{y_1, y_2, \dots, y_k\}$

$$\psi(z) = d(z,H) = |\sum_{i=1}^{k} \alpha_i y_i \mathbf{x}_i * \mathbf{z} + b| \quad (1.19)$$

với giá trị α_i và b thu được sau khi huấn luyện mô hình.

1.3.3. Độ đo khoảng cách

Hiệu quả của một hệ thống CBIR bị ảnh hưởng nhiều bởi quá trình trích rút đặc trưng cũng như phép đo xem sự giống hay khác nhau giữa các ảnh. Muốn xác

định hai ảnh có giống hay khác nhau hay không, các nhà nghiên cứu thường sử dụng một độ đo tương tự hoặc khoảng cách giữa hai ảnh đó với nhau trong hệ thống CBIR của mình [61]. Dựa vào phép đo tương tự/khoảng cách này sẽ xác định được hình ảnh nào sẽ phù hợp nhất với ảnh truy vấn và trả về một tập kết quả gồm K ảnh trên cùng từ kết quả phân hạng tập dữ liệu ảnh. Tập kết quả trả về gồm những ảnh có độ tương tự lớn nhất, tức là có khoảng cách nhỏ nhất so với ảnh truy vấn.

Một độ đo trên tập $\mathbf{X}$ là một ánh xạ $d: \mathbf{X} \times \mathbf{X} \rightarrow \boldsymbol{\mathcal{R}}$ sao cho $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{X}$ thỏa mãn tất cả các điều kiện sau:

$d(\mathbf{x}, \mathbf{y}) \geq 0, d(\mathbf{x}, \mathbf{y}) = 0$ xảy ra khi và chỉ khi $\mathbf{x} = \mathbf{y}$

$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$

$d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \geq d(\mathbf{x}, \mathbf{z})$

Trong các hệ thống CBIR, với hai ảnh $\mathbf{x} = (x_1, x_2, ..., x_n)$ và $\mathbf{y} = (y_1, y_2, ..., y_n)$ được biểu diễn trong không gian đặc trưng n chiều thì một số độ đo khoảng cách thường dùng được tính toán gồm:

Khoảng cách Manhattan còn gọi là khoảng cách L_1 được xác định như sau:

$$\text{KC}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \quad (1.20)$$

Khoảng cách Euclid còn gọi là khoảng cách L_2 (luận án sử dụng độ đo này để tính khoảng cách khi tra cứu trong pha tra cứu khởi tạo)

$$\text{KC}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1.21)$$

Khoảng cách Minkowski là tổng quát hóa của độ đo L_1 và L_2, trong đó p ($p \geq 1$) là tham số:

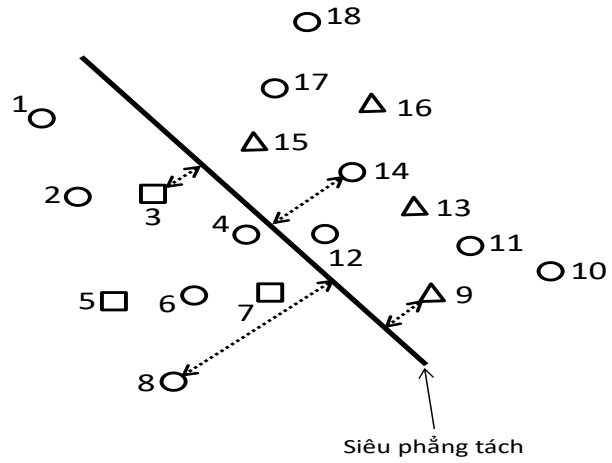
$$\text{KC}(\mathbf{x}, \mathbf{y}) = \sqrt[p]{\sum_{i=1}^{n} (x_i - y_i)^p} \quad (1.22)$$

Tùy thuộc vào giá trị của tham số p mà chúng ta có ba loại khoảng cách với $p = 1$, $p = 2$, $p = \infty$, nó tương ứng trở thành khoảng cách Manhattan, Euclid, hay Chebyshev [61]. Trong phép đo này, mỗi chiều của véc tơ đặc trưng hình ảnh là độc lập với nhau và có độ quan trọng như nhau.

Để đưa ra tập ảnh kết quả tra cứu, hệ thống tra cứu thường tính khoảng cách theo một độ đo khoảng cách nào đó của một ảnh truy vấn với toàn bộ ảnh trong tập

dữ liệu và tiến hành sắp xếp theo thứ tự tăng dần theo giá trị khoảng cách vừa tính được. Tập ảnh kết quả tra cứu bao gồm *k* ảnh trên cùng có giá trị khoảng cách nhỏ nhất.

Luận án này sử dụng biên quyết định do huấn luyện bằng máy véc tơ hỗ trợ trên tập ảnh phản hồi thu được từ tập kết quả tra cứu trước đó để phân hạng lại các ảnh trong tập dữ liệu theo khoảng cách của chúng với biên quyết định đó. Tập ảnh được phân hạng theo khoảng cách với quy luật: các ảnh thuộc lớp dương được sắp xếp theo thứ tự giảm dần của khoảng cách sẽ nằm ở phần trên, tiếp theo sẽ là các ảnh thuộc lớp âm được sắp xếp theo thứ tự tăng dần của khoảng cách. Lý do mà tập ảnh được phân hạng phụ thuộc vào siêu phẳng tách như thế là đối với các điểm thuộc lớp dương thì khoảng cách càng cách xa siêu phẳng tách sẽ có khả năng cao là mang nhãn dương, tức là độ liên quan tới ảnh truy vấn cao. Mặt khác, đối với các điểm thuộc về lớp âm thì càng gần siêu phẳng tách thì khả năng cao là giống ảnh truy vấn hơn là những điểm thuộc lớp âm có khoảng cách xa hơn so với siêu phẳng tách. Để hiểu rõ hơn về điều trên, trong Hình 1.13 sẽ minh họa việc phân hạng lại tập ảnh theo siêu phẳng tách.



Hình 1.13. Phân hạng các ảnh liên quan theo siêu phẳng tách SVM.

Giả sử tập dữ liệu ảnh **DB** = $\{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_{18}\}$ gồm 18 ảnh được biểu diễn dưới dạng 18 điểm được đánh số tương ứng với mỗi ảnh như trong Hình 1.13, trong đó có 03 ảnh $\mathbf{I}_3, \mathbf{I}_5, \mathbf{I}_7$ mang nhãn dương (điểm hình vuông), 04 ảnh $\mathbf{I}_9, \mathbf{I}_{13}, \mathbf{I}_{15}, \mathbf{I}_{16}$ mang nhãn âm (điểm hình tam giác) và 11 ảnh còn lại (điểm hình tròn) trong dữ liệu là không có nhãn (tức là không xuất hiện trong tập huấn luyện máy véc tơ hỗ trợ). Huấn

luyện mô hình SVM với tập dữ liệu gồm 7 điểm đã được gán nhãn để có siêu phẳng tối ưu (siêu phẳng tách – đường nét liền) phân tách được hai lớp dương và âm. Các điểm (1, 2, 3, 4, 5, 6, 7, 8) nằm bên phía lớp dương sẽ được sắp xếp theo thứ tự giảm dần của khoảng cách tới siêu phẳng và thu được tập các điểm theo thứ tự (8, 5, 2, 6, 1, 7, 3, 4). Ngược lại, với các điểm (9, 10, 11, 12, 13, 14, 15, 16, 17, 18) nằm bên phía lớp âm của siêu phẳng tách thì lại được sắp xếp theo thứ tự tăng dần của khoảng cách tới siêu phẳng, do đó ta có tập các điểm sau khi sắp xếp gồm (12, 9, 15, 14, 13, 17, 11, 10, 16, 18). Cuối cùng gộp hai tập đó lại thành một tập cuối cùng (8, 5, 2, 6, 1, 7, 3, 4, 12, 9, 15, 14, 13, 17, 11, 10, 16, 18).

1.4. Đánh giá độ chính xác CBIR

1.4.1. Độ chính xác và độ chính xác trung bình

Ma trận nhầm lẫn được sử dụng phổ biến để đo lường độ chính xác phân lớp cho bài toán phân lớp cũng như trong tra cứu ảnh. Do thông tin của các ảnh không thuộc tập kết quả tra cứu là không cần thiết nên các số liệu FN (False Negative, số ảnh có liên quan không tra cứu được) và TN (True Negative, số ảnh không liên quan không tra cứu được) thường sẽ được bỏ qua trong các bài toán CBIR. Người dùng sẽ chỉ quan tâm tới số ảnh có liên quan được hiển thị nhiều trong kết quả trả về, do đó TP (True Positive) là trường hợp quan trọng nhất. Bên cạnh đó, trường hợp FP (False Positive) dương tính giả là số ảnh trong kết quả trả về trên cùng mà không liên quan được hiển thị cho người dùng, thường chúng có số lượng đa số trong tập ảnh sẽ làm ảnh hưởng đến độ chính xác của hệ thống.

Để đánh giá hiệu quả của một hệ thống CBIR, độ chính xác được sử dụng trong các phương pháp tra cứu ảnh là để đảm bảo rằng trong N ảnh trên cùng được trả về là có liên quan tới ảnh truy vấn. Sự liên quan này được thể hiện bởi tập tin cậy nền (ground truth) để biết được những ảnh thuộc cùng chủ đề với nhau. Ở đây "tập tin cậy nền" được hiểu là tập ảnh cơ sở dữ liệu đã được chia thành các chủ đề (có nhãn) và dựa vào đó, hệ thống sẽ dùng cho việc đánh giá ảnh nào là liên quan đến ảnh truy vấn, ảnh nào là không liên quan đến ảnh truy vấn. Độ chính xác tại mỗi lần tra cứu là tỷ lệ giữa số lượng ảnh liên quan với ảnh truy vấn trong tập ảnh trên cùng trả về và số lượng tất cả ảnh được hiển thị trên cùng trả về [61]. Nó được tính như sau:

$$\textbf{Precision} = \frac{\text{TP}}{(\text{T P} + \text{F P})} \tag{1.23}$$

Số lượng ảnh kết quả được hiển thị trên đầu trả về cho người dùng được gọi là phạm vi (scope). Khi giá trị số lượng ảnh liên quan với ảnh truy vấn so sánh tại một phạm vi cụ thể K = TP + FP, thường được gọi là P@K (trong trường hợp K=100, chúng ta hiểu P@100 là độ chính xác trên 100 ảnh trả về). Khi độ chính xác được vẽ trên một biểu đồ với nhiều phạm vi được gọi là biểu đồ phạm vi chính xác. Khi phản hồi liên quan được sử dụng trong hệ thống tra cứu ảnh thì độ chính xác thường được tạo biểu đồ dựa trên số lần lặp, giá trị phạm vi được đặt tại một giá trị cố định.

Trong tra cứu ảnh, độ chính xác trung bình **AP** (Average Precision) [61] thường được sử dụng để đánh giá độ chính xác của toàn bộ hệ thống được đo bằng trung bình tất cả độ chính xác tại mỗi lần tra cứu. AP được tính toán như sau:

$$\textbf{AP} = \frac{\sum_{i=1}^{N} precision(i)}{N} \tag{1.24}$$

Với $precision(i)$ là độ chính xác của mỗi truy vấn và N là số lượng ảnh được đưa lần lượt làm ảnh truy vấn.

Độ chính xác của một phương pháp cụ thể trong CBIR là rất quan trọng nhưng không phải là yếu tố duy nhất để đánh giá được phương pháp đó có hoạt động tốt hơn hay kém hơn các phương pháp khác. Bên cạnh đó, tốc độ tra cứu cũng không kém phần quan trọng cho các hệ thống CBIR. Nếu một hệ thống có thể cho kết quả gần như ngay lập tức đa số các hình ảnh kết quả mà người dùng quan tâm nhưng phải mất hàng phút, hoặc có thể lên đến hàng chục phút, hàng giờ thì phương pháp đó dường như không hiệu quả. Không giống như độ chính xác, hiệu quả tính toán có thể được cải thiện theo thời gian do trong tương lai với tốc độ phát triển công nghệ làm cho bộ vi xử lý hay các phương tiện lưu trữ có thể xử lý nhanh hơn. Chính vì thế, phương pháp đề xuất tập trung để cải thiện độ chính xác, trong các thực nghiệm ở Chương 2 và Chương 3 của luận án sẽ sử dụng độ chính xác trung bình để đánh giá hiệu quả của các phương pháp. Trong luận án sử dụng đường cong độ chính xác-phạm vi (precision - scope) và tỷ lệ độ chính xác trung bình AP để đánh giá độ chính xác của các thuật toán tra cứu ảnh. Phạm vi được chỉ ra bởi K ảnh trên cùng được hiển thị cho người dùng. Độ chính xác là tỷ số của số các ảnh liên quan và K ảnh trên cùng được

hiển thị cho người dùng. Đường cong độ chính xác - phạm vi mô tả độ chính xác với nhiều phạm vi và do đó đưa ra đánh giá độ chính xác toàn bộ của các thuật toán, trong khi đó, tỷ lệ độ chính xác nhấn mạnh độ chính xác tại một giá trị cụ thể của một phạm vi.

1.4.2. Một số tập dữ liệu ảnh dùng cho tra cứu ảnh dựa vào nội dung

Tập dữ liệu ảnh COREL 10800

Bộ sưu tập hình ảnh gốc Corel Photo Gallery [62] nổi tiếng trong cộng đồng nghiên cứu bao gồm hơn 800 đĩa CD, mỗi đĩa chứa các ảnh của một khái niệm chủ đề cụ thể tiền cảnh nổi bật. Các khái niệm chủ đề này có thể rất rộng như "đại dương, mùa thu" hoặc có thể giới hạn phạm vi nhỏ hơn như "hoàng hôn, đường cao tốc". Do sự phức tạp của các chủ đề ảnh thay đổi từ loại này sang loại khác và cỡ của bộ sưu tập là rất lớn nên toàn bộ tập ảnh thường không được sử dụng trong một hệ thống tra cứu cụ thể. Điều đó dẫn đến tình trạng mỗi nhóm nghiên cứu tạo ra bộ Corel con của riêng mình trong các thực nghiệm của họ.

Để đánh giá độ chính xác của các phương pháp tra cứu, nhiều hệ thống CBIR sử dụng một tập con của Corel Photo Gallery gồm 10800 ảnh làm tập dữ liệu ảnh thử nghiệm [47]. Tập Corel 10800 bao gồm 80 chủ đề khác nhau và mỗi chủ đề thường có 100 ảnh trừ một số ít chủ đề có nhiều hơn 100 ảnh. Kích thước của mỗi ảnh trong mỗi chủ đề cố định là 120 x 80 hoặc 80 x 120. Một số ảnh minh họa thuộc các chủ đề hoa hồng, hổ, tuyết, pháo hoa và cây cảnh được trình bày như Hình 1.14.

Hình 1.14. Một số mẫu trong tập dữ liệu ảnh COREL 10800.

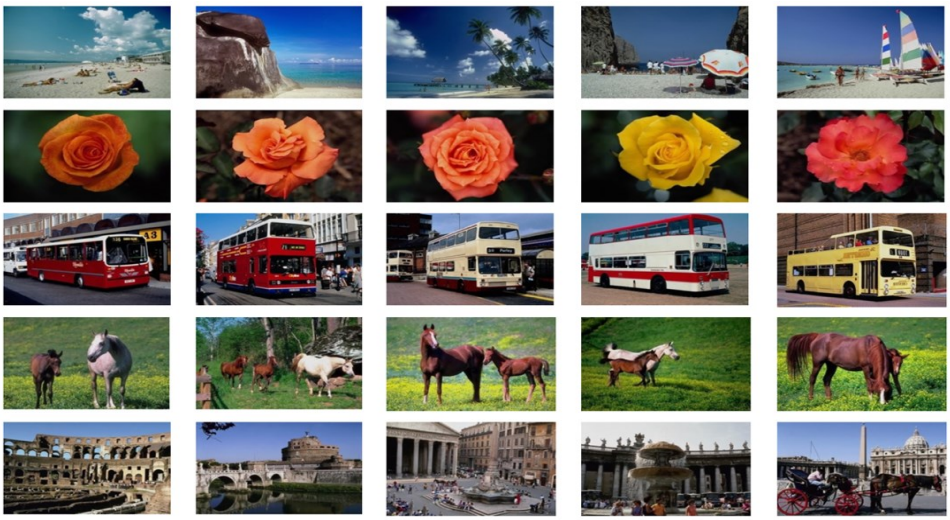
Trong thực nghiệm, luận án sử dụng tập dữ liệu này với 5 đặc trưng khác nhau của mỗi bức ảnh cho một véc tơ có độ dài 190 chiều gồm lược đồ màu [63] (32 chiều),

tương quan màu [64] (64 chiều), mô men màu [65] (6 chiều), đặc trưng Gabor [66] (48 chiều), và đặc trưng biến đổi wavalet [67] (40 chiều). Đầu tiên, đặc trưng lược đồ màu được sử dụng, mỗi ảnh được chuyển từ không gian màu RGB sang không gian màu HSV. Từng thành phần kênh màu H, S, V được lượng hóa tương ứng thành 8, 2 và 2 bins màu cho ta một véc tơ có độ dài 32 phần tử. Tiếp theo, đặc trưng tương quan màu được tạo ra bằng cách trong không gian RGB sẽ lượng hóa 4 bins cho mỗi kênh R, G, B tương ứng. Quá trình trích rút tương quan màu tạo ra một véc tơ gồm 64 phần tử. Bước tiếp theo, trong mỗi kênh màu R, G và B của không gian màu RGB, đặc trưng mô men màu sử dụng hai mô men đầu tiên là trung bình (mean) và độ lệch chuẩn (standard deviation) cho một véc tơ có 6 phần tử. Tiếp theo, véc tơ đặc trưng Gabor wavelet gồm 48 phần tử được tính toán trên ảnh đa cấp xám gồm meanSquaredEnergy và meanAmplitude cho 4 tỷ lệ (scale): “0.05, 0.1, 0.2, 0.4” và 6 hướng (orientation): “0, π/6, 2π/6, 3π/6, 4π/6, 5π/6 ”. Cuối cùng, đặc trưng biến đổi wavelet được tính toán trên mỗi ảnh với 3 mức phân tách. Giá trị trung bình và độ lệch chuẩn của các hệ số biến đổi được sử dụng tạo thành một véc tơ đặc trưng gồm 40 phần tử cho mỗi hình ảnh

Các kết quả tra cứu ban đầu với tập đặc trưng này và sử dụng khoảng cách Euclid được gọi là phương pháp “Baseline”. Phương pháp Baseline được sử dụng trong các so sánh thực nghiệm của luận án.

Tập dữ liệu ảnh SIMPLIcity

Để minh họa trực quan cho phép chiếu trong phương pháp đề xuất, thực nghiệm của luận án thực hiện trên tập con của SIMPLIcity [68].

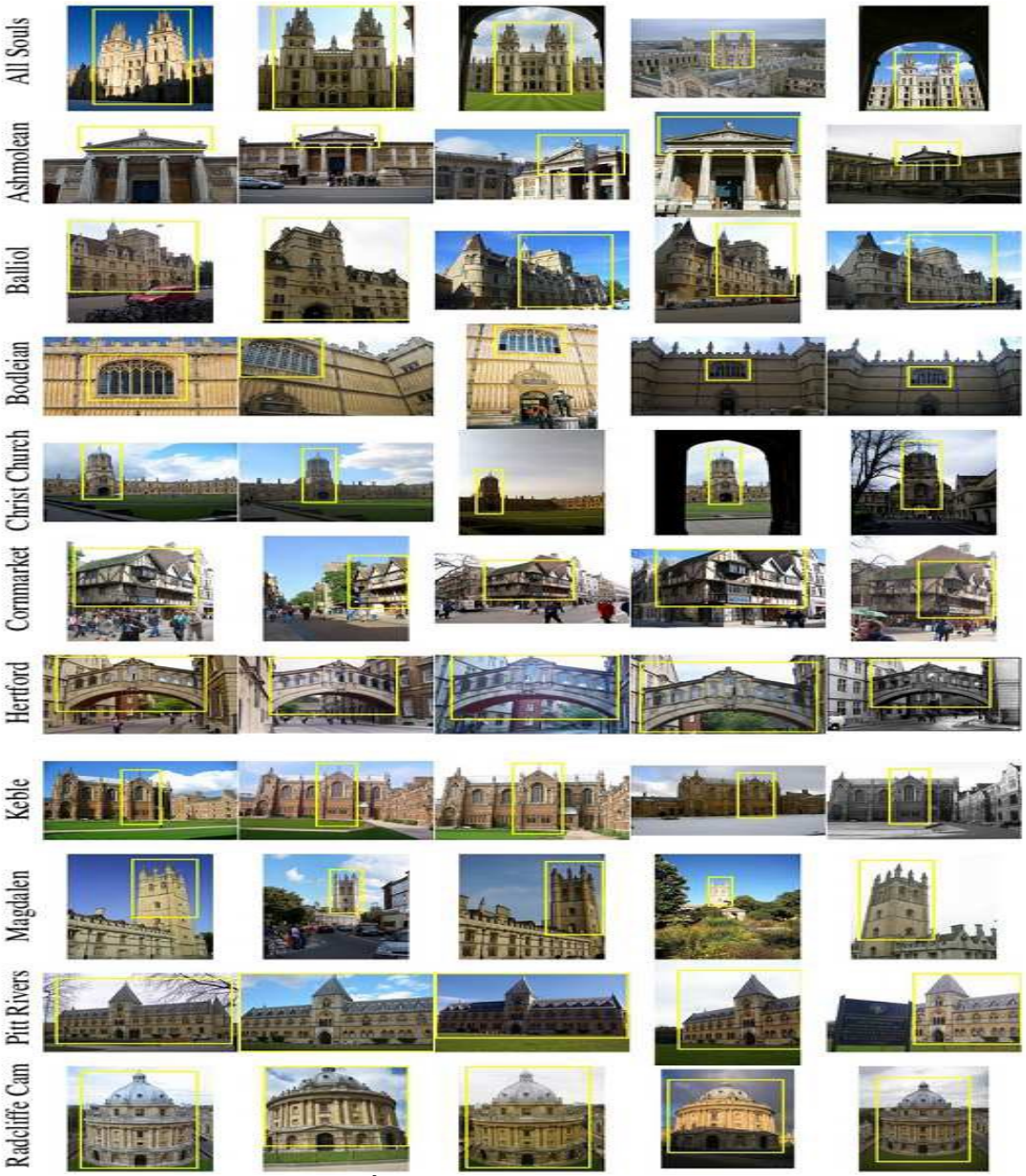


Hình 1.15. Một số ảnh mẫu trong tập dữ liệu ảnh SIMPLIcity.

Tập dữ liệu nhỏ này được cấu tạo gồm một nghìn ảnh với 10 chủ đề có kích cỡ 256 * 384 hoặc 384 * 256. Một số ảnh mẫu trong tập dữ liệu ảnh SIMPLIcity được chỉ ra trong Hình 1.15.

Tập dữ liệu ảnh Oxford Building

Tập ảnh cơ sở dữ liệu Oxford Building [69] bao gồm 5062 ảnh được thu thập từ Flickr bằng cách tìm kiếm các địa danh cụ thể của Oxford. Bộ sưu tập này được phân loại theo cách thủ công thành 11 địa danh (chủ đề) khác nhau tạo ra một tập tin cậy nền (ground truth). Tập ảnh truy vấn chứa 55 ảnh khác nhau được xây dựng bằng cách mỗi địa danh chọn 5 ảnh truy vấn khác nhau (xem chi tiết trong Hình 1.16).



Hình 1.16. Tập ảnh truy vấn chứa 55 ảnh trong tập ảnh Oxford Building

Đối với mỗi hình ảnh trong tập dữ liệu được gán nhãn là một trong bốn nhãn bao gồm: (1) Good - ảnh đẹp, rõ ràng các đối tượng/ tòa nhà, (2) OK – hơn 25% của đối tượng là nhìn thấy được, (3) Junk – ít hơn 25% của đối tượng được nhìn thấy, hoặc có một mức độ rất cao bị che lấp hoặc méo mó, (4) Bad – đối tượng không được biểu diễn. Trong [50] chỉ ra rằng tập dữ liệu Oxford bị chú thích sai và các truy vấn được lựa chọn tra cứu cho kết quả không tốt. Bên cạnh đó, bộ dữ liệu này có kích cỡ tương đối nhỏ (5062 ảnh), số lượng chủ đề cũng ít hơn so với tập Corel. Do đó, trong thực nghiệm, luận án lựa chọn tập Corel để đánh giá độ chính xác tra cứu.

Tập dữ liệu ảnh Caltech 101

Tập dữ liệu ảnh Caltech-101 [70] bao gồm các ảnh thuộc 101 chủ đề khác nhau. Mỗi ảnh được gán nhãn với một đối tượng duy nhất (xem minh họa trong Hình 1.17).

Hình 1.17. Mỗi ảnh cho một chủ đề trong số 101 chủ đề trong tập ảnh Caltech 101

Mỗi chủ đề chứa khoảng 40 đến 800 ảnh, tổng cộng khoảng 8742 ảnh. Mỗi hình ảnh có kích thước xấp xỉ 200 × 300 điểm ảnh, trong đó có bao gồm một số ảnh là ảnh đa cấp xám. Mặc dù tập dữ liệu Caltech 101 số lượng chủ đề lớn hơn tập Corel 10800 nhưng có nhiều chủ đề chứa những ảnh có nền màu đồng nhất (thực tế ảnh chụp phong cảnh có được màu nền đồng nhất là không khả thi). Bên cạnh đó số lượng ảnh trong mỗi chủ đề chênh lệch nhau khá lớn, số lượng ảnh thấp nhất là 40 ảnh, số lượng ảnh lớn nhất là 800 ảnh. Do đó, trong quá trình thực nghiệm, luận án lựa chọn thực hiện trên tập ảnh Corel 10800 thay vì tập ảnh Caltech 101

1.4.3. Kịch bản phản hồi liên quan trong thực nghiệm

Trong một hệ thống tra cứu ảnh thực tế, một ảnh truy vấn thường không có trong cơ sở dữ liệu ảnh. Để mô phỏng một môi trường như thế, luận án sử dụng bốn phần kiểm chứng chéo để đánh giá các thuật toán. Chính xác hơn, chia toàn bộ cơ sở dữ liệu ảnh theo chủ đề thành bốn tập con có cỡ ngang bằng nhau. Tại mỗi lần chạy của kiểm chứng chéo, một tập con được lựa chọn làm tập ảnh truy vấn, và ba tập còn lại được sử dụng làm cơ sở dữ liệu ảnh cho tra cứu. Đường cong độ chính xác - phạm vi và tỷ lệ độ chính xác trung bình được tính toán bởi trung bình các kết quả từ kiểm chứng chéo với bốn phần.

Khi đưa ảnh truy vấn vào hệ thống, hệ thống sẽ phân hạng theo thứ tự tăng dần của khoảng cách (luận án sử dụng khoảng cách Euclid trong pha tra cứu khởi tạo) giữa ảnh truy vấn đó với các ảnh có trong tập dữ liệu ảnh. Nếu đặc trưng của ảnh mô tả tốt nội dung ngữ nghĩa của ảnh thì khoảng cách Euclid này cũng có thể phản ánh tương đối quan hệ giữa đặc trưng và nhận thức của người về ảnh. Việc thu được kết quả theo khoảng cách Euclid đó được thực hiện bằng cách thu những ảnh nằm gần (láng giềng) với ảnh truy vấn trong không gian đặc trưng. Nhưng với các kỹ thuật đã có hiện nay theo cách tiếp cận này đều cho hiệu quả không cao do khoảng trống ngữ nghĩa. Do đó, người ta thường sử dụng thông tin nhận thức của người thông qua sự tương tác của người dùng với hệ thống để thu hẹp khoảng trống đó.

Việc thu thông tin phản hồi của người dùng trong quá trình đánh giá hệ thống tra cứu tốn nhiều công sức và thời gian. Bên cạnh đó, cảm nhận chủ quan của mỗi người dùng trong các đánh giá một ảnh, tức là hai người dùng khác nhau cảm nhận về cùng một ảnh là khác nhau, thậm chí cùng một người dùng có thể cảm nhận về

cùng một ảnh ở hai thời điểm khác nhau là khác nhau. Do đó, trong đánh giá hệ thống tra cứu ảnh, người ta thiết kế một phản hồi mô phỏng tự động để mô hình quá trình tra cứu, tức là máy tính có thể đưa ra phản hồi cho K ảnh trên cùng dựa vào thông tin của tập tin cậy nền (tập tin cậy nền cho biết một ảnh bất kỳ trong tập ảnh thuộc chủ để nào). Với mỗi truy vấn được cung cấp, hệ thống tính toán khoảng cách và sắp xếp các ảnh trong cơ sở dữ liệu theo thứ tự tăng dần của khoảng cách vừa tính được. Tập kết quả tra cứu khởi tạo gồm K ảnh trên cùng sau khi phân hạng được lựa chọn làm các ảnh phản hồi, và thông tin nhãn của chúng (liên quan hoặc không liên quan) được sử dụng cho phân hạng lại.

Người dùng tương tác với hệ thống thông qua đánh dấu trong tập kết quả tra cứu khởi tạo các ảnh có cùng chủ đề (cùng khái niệm) với ảnh truy vấn làm ảnh liên quan (mẫu phản hồi dương) và những ảnh còn lại không đánh dấu làm ảnh không liên quan (mẫu phản hồi âm) và lấy thêm $\frac{K}{2}$ ảnh tiếp theo được xếp hạng ngay sau tập kết quả tra cứu khởi tạo làm mẫu chưa được gán nhãn. Với mỗi truy vấn, cơ chế phản hồi liên quan được lựa chọn cho hai lần lặp phản hồi. Lý do cho điều này là do người dùng thường không có đủ kiên nhẫn để tham gia phản hồi nhiều vòng, cho nên kết quả thể hiện tại hai lần lặp đầu tiên là rất quan trọng.

1.5. Kết luận chương 1

Trong chương 1, luận án đã trình bày lý thuyết tổng quan về một hệ thống tra cứu ảnh dựa vào nội dung và phản hồi liên quan. Bên cạnh đó, cũng phân tích một số phương pháp phản hồi liên quan nhằm giảm khoảng trống ngữ nghĩa. Qua đó, phân tích, đánh giá ưu nhược điểm một số phương pháp CBIR hiện có để đề xuất một số phương pháp nhằm giải quyết những hạn chế đã phân tích. Trong quá trình học trong phản hồi liên quan, luận án nhận thấy rằng các hệ thống tra cứu ảnh cho hiệu quả tra cứu thấp là do một số nguyên nhân sau:

- Thứ nhất, số lượng mẫu dương (ảnh được người dùng đánh dấu là "liên quan") và số mẫu âm (ảnh được người dùng đánh dấu là "không liên quan") thường không cân bằng. Ngoài ra, số lượng mẫu chưa được gán nhãn thường bị bỏ qua trong pha phản hồi mặc dù chúng chứa nhiều thông tin có ích.

- Thứ hai, một số phương pháp tra cứu ảnh chỉ thực hiện trên một không gian toàn cục, trong khi một số phương pháp tra cứu ảnh theo tiếp cận học đa tạp chỉ xem

xét các mẫu trong cùng một lân cận. Các thống kê toàn cục như phương sai thường khó ước lượng khi số lượng mẫu hạn chế. Bên cạnh đó, số lượng mẫu dữ liệu huấn luyện rất hạn chế, thường ít hơn nhiều so với số chiều của không gian đặc trưng.

- Cuối cùng, một số phương pháp chỉ sử dụng một bộ phân lớp nên chưa thể biểu diễn tốt các khía cạnh khác nhau của một đối tượng bởi vì một đối tượng có thể bao gồm nhiều khía cạnh khác nhau.

Luận án này sẽ tập trung vào giải quyết vấn đề nâng cao độ chính xác tra cứu ảnh để giải quyết một số khó khăn ở trên.

CHƯƠNG 2. PHƯƠNG PHÁP HỌC CHIẾU PHÂN BIỆT LỚP NGỮ NGHĨA CHO TRA CỨU ẢNH VỚI PHẢN HỒI LIÊN QUAN.

Trong chương 2 này, luận án sẽ đề xuất phương pháp học chiếu phân biệt lớp ngữ nghĩa cho giảm chiều trong tra cứu ảnh [CT5] để giải quyết hạn chế: số chiều của đặc trưng thường cao hơn rất nhiều so với số mẫu trong tập phản hồi và các mẫu nằm ở hai không gian con (hai lân cận) khác nhau chưa được xét đến.

2.1. Giới thiệu

Các hình ảnh trong CBIR được thể hiện bằng véc tơ đặc trưng trực quan thường có kích thước (số chiều) rất cao, kích thước của véc tơ đặc trưng có thể từ hàng chục đến hàng nghìn. Ví dụ, đặc trưng lược đồ màu sau khi trích rút có thể cho một véc tơ đặc trưng có kích thước 256 chiều, hay một số mô hình tiền huấn luyện trong học sâu có thể trích rút véc tơ đặc trưng kích thước lên đến hơn 1000 chiều. Phương pháp CBIR truyền thống gặp phải khó khăn trong việc mô hình hóa dữ liệu ảnh trực tiếp trong một không gian đặc trưng chiều cao. Việc học từ các ảnh huấn luyện trong không gian đặc trưng chiều cao sẽ không khả thi cho việcvề mặt tính toán do số lượng ảnh cần thiết cho việc tổng quát hóa tốt phải tăng theo cấp số nhân so với số chiều của đặc trưng. Bên cạnh đó, trong quá trình phản hồi liên quan, người dùng thường mong muốn có kết quả nhanh do đó người dùng họ sẽ không đủ kiên nhẫn để chờ đợi thời gian huấn luyện của mô hình học máy với tập ảnh huấn luyện có số chiều của véc tơ đặc trưng cao quá lớn.

Khi số chiều của véc tơ đặc trưng là rất lớn (lớn hơn số lượng ảnh trong tập ảnh huấn luyện), một số phương pháp học máy có thể gặp phải vấn đề "lời nguyền củavề số chiều" (curse of dimensionality). Hãy xem xét một thuật toán phân lớp cho hai lớp dương và âm đơn giản như sau, tìm một tập các trọng số w sao cho tích vô hướng của w với một ảnh có véc tơ đặc trưng x, sẽ cho kết quả có giá trị âm tương ứng với lớp mang nhãn âm và cho kết quả có giá trị dương tương ứng với lớp mang nhãn dương. Lúc này, w sẽ có độ dài bằng với số chiều của x, do số mẫu nhỏ hơn số chiều đặc trưng nên w sẽ có nhiều tham số hơn so với các mẫu trong toàn bộ dữ liệu. Điều này có nghĩa là sẽ có thể gặp phải vấn đề quá khớp dữ liệu và do đó sẽ không tổng quát hóa tốt cho các mẫu khác không xuất hiện trong tập ảnh huấn luyện. Một cách có thể khắc phục được hạn chế này bằng cách là giảm chiều véc tơ đặc trưng

của tập ảnh để ánh xạ nó từ một không gian chiều cao sang một không gian con chiều thấp hơn và học khái niệm ngữ nghĩa mức cao được thực hiện trong không gian con chiều thấp đó. Ở đây cần lưu ý rằng học đa tạp cũng là một phương pháp giảm chiều bởi vì nó cũng thực hiện việc chiếu dữ liệu từ không gian chiều cao sang không gian chiều thấp, tuy nhiên, học đa tạp khai thác được cấu trúc phi tuyến của dữ liệu.

Theo phân loại trong học máy, các phương pháp giảm chiều có thể được chia thành ba loại chính bao gồm: phương pháp giảm chiều không giám sát, giảm chiều có giám sát và giảm chiều bán giám sát.

Các phương pháp không giám sát xử lý dữ liệu không có nhãn, bao gồm phân tích thành phần chính (PCA) [53, 71][53, 71], chiếu bảo toàn cục bộ (LPP - Locality preserving projection) [3, 4], nhúng tuyến tính cục bộ (LLE - Locally linear embedding) [5], nhúng bảo toàn lân cận (NPE- Neighborhood Preserving Embedding) [71], Laplacian Eigenmaps [59][59], và Isomap [58].[58]. PCA tìm một phép chiếu mà trên đó phương sai là cực đại. PCA chỉ xét cấu trúc Euclide mà khám phá cấu trúc toàn cục của không gian, nên cấu trúc cục bộ được hình thành bởi ảnh truy vấn và tập mẫu phản hồi được gán nhãn bị bỏ qua, bởi vì không gian các đặc trưng trực quan mức thấp của ảnh có thể là một đa tạp [3, 15]. Một số phương pháp xem xét trường hợp khi dữ liệu nằm trên một không gian con của không gian gốc bao gồm nhúng tuyến tính cục bộ (Locally Linear Embedding - LLE) [5], Isomap [58][58] và Laplacian Eigenmaps [59].[59]. Các phương pháp tra cứu ảnh kể trên theo cách tiếp cận giảm chiều không giám sát cho độ chính xác kémthấp do chỉ dựa trên những mẫu không có nhãn để tìm phép chiếu ánh xạ sang không gian mới.

Bên cạnhTrong khi đó, các phương pháp học có giám sát chiếu dữ liệu có nhãn vào một không gian con thấp chiều để thu độ chính xác phân lớp tốt hơn và độ phức tạp tính toán thấp. Các phương pháp học có giám sát tiêu biểu gồm phân tích phân biệt tuyến tính (LDA) [71], phân tích phân biệt tuyến tính [71], nhúng phân biệt cục bộ (LDP - Local Discriminant Embedding) [10], chiếu bảo toàn cục bộ tối ưu có giám sát (SoLPP - Supervised Optimal Locality Preserving Projection) [11], phân tích lề Fisher (MFA - Marginal Fisher Analysis) [71], nhúng lân cận phân biệt (DNE - Discriminant neighborhood embedding) [12], nhúng láng giềng phân biệt dựa trên đồ thị lân cận kép (DAG-DNE -Double Adjacency Graph-based DNE) [72]lân cận phân biệt dựa trên đồ thị lân cận kép (DAG-DNE -Double Adjacency Graph-based DNE)

[72], chiếu phân biệt phân lớp hồi quy tuyến tính (LRCDP - Linear Regression Classification Steered Discriminative Projection) [13], và nhúng đồ thị bảo toàn phân biệt toàn cục và cục bộ[13], và nhúng đồ thị bảo toàn toàn cục và cục bộ phân biệt (DGLPGE - Discriminative Globality And Locality Preserving Graph Embedding) [14]. LDA tìm một phép chiếu mà trên không gian chiếu mới đó các mẫu có cùng nhãn sẽ nằm gần nhau, những mẫu mang nhãn khác nhau sẽ nằm xa nhau. Tuy nhiên LDA cũng như PCA chỉ xét cấu trúc Euclide mà khám phá cấu trúc toàn cục của không gian, do đó bỏ qua cấu trúc đa tạp của không gian đặc trưng trực quan mức thấp của ảnh. Để khắc phục các hạn chế của LDA, MFA đã được đề xuất trong [71]. MFA có thể khám phá cấu trúc đa tạp bởi xây dựng đồ thị lân cận nội tại lớp và đồ thị lân cận ngoại lớp để giữ cấu trúc cục bộ của các mẫu. MFA cũng tìm ra các hướng chiếu mà cực đại phân tán nội tại lớp và cực tiểu sự phân tán ngoại lớp. Tương tự như MFA với đồ thị lân cận, DNE được đề xuất trong [12] có thể tìm các hướng chiếu tốt nhất bởi bằng việc sử dụng phân tích phổ. Trong DNE, đồ thị lân cận được xây dựng để lưu giữ cấu trúc cục bộ bởi phân biệt các lân cận thuần nhất và không thuần nhất. Tuy nhiên, DNE dường như nhận một phân tán nội lớp nhỏ trong không gian chiếu [73].[73]. Để khắc phục hạn chế của DNE, nhúng láng giềng phân biệt dựa trên đồ thị lân cận kép (DAG-DNE -Double Adjacency Graph-based DNE) [72] là để cực đại lề giữa the[72] là để cực đại lề giữa phân tán nội tại lớp và ngoại lớp, trước khi tìm các hướng chiếu. DAG-DNE có thể giữ tính nén của nội tại lớp và mở rộng sự tách biệt của ngoại lớp trong không gian con. Nhưng DAG-DNE có hạn chế là có số chiều đặc trưng bị cố định và thường là lớn. Gần đây, một số phương pháp được đề xuất bao gồm: phương pháp LRCDP [13][13] và phương pháp DGLPGE [14]. LRCDP không xử lý trực tiếp dữ liệu trên không gian đa chiều ban đầu nên không bị ảnh hưởng bởi thông tin dư thừa hoặc nhiễu. Phương pháp này cải thiện độ chính xác và giảm chi phí tính toán. Tuy nhiên, LRCDP kế thừa ý tưởng của LDA nên nó chỉ khám phá ra cấu trúc Euclide toàn cục mà không đề cập đến cấu trúc đa tạp cục bộ. DGLPGE xem xét sự phân biệt lớp, thông tin toàn cục và cục bộ của dữ liệu đa chiều theo thứ tự để cải thiện khả năng bảo toàn thông tin hình học của dữ liệu và khả năng phân biệt trong không gian con chiều thấp. Tuy nhiên, DGLPGE và các phương pháp kể trên chỉ thực hiện với các điểm dữ liệu có nhãn trong tập huấn luyện, và nó không đưa ra rõ ràng phép chiếu có thể thực hiện cho các điểm thử mới một cách rõ ràng.

Ngoài hạn chế của mỗi phương pháp, độ chính xác của các phương pháp sử dụng cách tiếp cận có giám sát kể trên có xu hướng giảm nếu chỉ sẵn có một số nhỏ mẫu được gán nhãn là sẵn có. Viễn cảnh mẫu nhỏ là phổ biến trong tra cứu ảnh với phản hồi liên quan. Trong khi thuật toán phương pháp học có giám sát cho độ chính xác tốt hơn các thuật toán học không giám sát, thu thập tập dữ liệu huấn luyện có nhãn trong học có giám sát đòi hỏi nhiều nhân công và tốn nhiều thời gian. Trong khi đó, chúng ta có thể thu dữ liệu không có nhãn rất là dễ dàng. Để khắc phục hạn chế của các phương pháp học có giám sát ở trên, cách tiếp cận học bán giám sát mà khai thác cả các mẫu phản hồi của người dùng và các mẫu chưa có nhãn được đề xuất. Một số phương pháp tiêu biểu theo cách tiếp cận giảm chiều bán giám sát đã được đề xuất bao gồm nhúng quan hệ gia tăng (ARE - Augmented Relation Embedding) [15] và, chiếu cực đại lề cho tra cứu ảnh (MMP - Maximum Margin Projection) [16], phân tích phân biệt bán giám sát (Semisupervised Discriminant Analysis - SDA) [17][17], nhúng đa tạp dựa vào đồ thị linh hoạt với nhúng phân biệt bán giám sát (LFGBSE - Learning flexible graph-based semi-supervised embedding) [18], học phân biệt bán giám sát ổn định (SSDL - Stable Semi-Supervised Discriminant Learning) [71]. Phương pháp tra cứu ảnh[71]. Phương pháp tra cứu ảnh theo tiếp cận học bán giám sát, gọi là nhúng quan hệ gia tăng (ARE) [15] học một đa tạp ngữ nghĩa và tôn trọng sở thích của người dùng. ARE xây dựng ba đồ thị quan hệ: một cáiđồ thị mô tả các quan hệ tương tự, và hai cáiđồ thị còn lại mã hóa các quan hệ liên quan/không liên quan sử dụng các phản hồi liên quan được người dùng cung cấp. Với các đồ thị được định nghĩa, học một đa tạp ngữ nghĩa có thể được biến đổi thành giải bài toán tối ưu có ràng buộc. Tuy nhiên, độ quan trọng của các ảnh có nhãn và không có nhãn là ngang bằng nhau trong quá trình tìm chiếu tối ưu. Khắc phục hạn chế của ARE, phương pháp học bán giám sát, có tên là cực đại lề cho tra cứu ảnh (MMP), được đề xuất. Nó khám phá cả các cấu trúc phân biệt và hình học. MMP xây dựng một đồ thị mô tả quan hệ lân cận đặc trưng, một đồ thị trong phạm vi lớp và một đồ thị liên lớp, và xây dựng hai hàm mục tiêu. Tuy nhiên, phương pháp này chỉ quan tâm đến nén và tách biệt các điểm thuộc cùng một lân cận mà bỏ qua việc nén và tách biệt các điểm khác lân cận, tức là không đảm bảo các điểm liên quan ngữ nghĩa mà ở các lân cận khác nhau là gần ảnh truy vấn trong không gian con chiều thấp hơn. Điều này làm giảm độ chính xác cho nhiệm vụ phân lớp trong tra cứu ảnh với phản hồi liên quan.

Trong [71], Gao và cộng sự trình bày một phương pháp học bán giám sát, có tên là học phân biệt bán giám sát ổn định (Stable Semi-Supervised Discriminant Learning - SSDL). Bằng việc xây dựng một hàm mục tiêu hợp lý, SSDL học cấu trúc nội tại mà mô tả tốt cả sự tương tự và đa dạng của dữ liệu và sau đó liên kết biểu diễn cấu trúc này thành phân tích phân biệt tuyến tính. Mặc dù các phương pháp giảm chiều theo cách tiếp cận bán giám sát đã thu được những thành công, vẫn có những vấn đề chưa được giải quyết như: nó đòi hỏi nhiều thông tin tiền nghiệm, đòi hỏi nhiều liên kết để đảm bảo độ chính xác (đòi hỏi nhiều liên kết này còn khó hơn cả gán nhãn một số điểm), khó xác định chiều tối ưu, và độ chính xác không đủ tốt khi các điểm có nhãn là không đủ. Ngoài ra, các phương pháp này chỉ xem xét tính chất hình học trong một lân cận nào đó, trong khi bỏ qua mối liên hệ của các mẫu từ hai lân cận khác nhau. Do đó, chúng giảm độ chính xác tra cứu cho tra cứu ảnh.

Để khắc phục vấn đề trên, luận án đề xuất một phương pháp học chiếu phân biệt lớp ngữ nghĩa (Semantic Class Discriminant Projection - SCDP) [CT5]. Trong SCDP, các ảnh liên quan ngữ nghĩa với ảnh truy vấn mà thuộc cùng một lân cận sẽ được nén lại ở mức độ cao nhất và cũng các ảnh liên quan ngữ nghĩa này nhưng chúng thuộc về các lân cận khác nhau sẽ có mức độ nén thấp hơn. Điều này đảm bảo rằng, các ảnh liên quan ngữ nghĩa sẽ gần với ảnh truy vấn trong không gian con chiều thấp hơn kể cả khi chúng thuộc về các lân cận khác nhau. Trong khi đó, các ảnh của các lớp khác nhau mà thuộc về hai lân cận khác nhau sẽ có mức độ tách biệt cao nhất và cũng các ảnh của các lớp khác nhau nhưng thuộc về cùng một lân cận sẽ có mức độ tách biệt thấp hơn. Do đó, SCDP có thể bảo toàn trung thực cấu trúc cục bộ của các điểm dữ liệu trong không gian đặc trưng trực quan gốc nhiều chiều và tìm một ma trận chiếu tốt cho chúng.

2.2. Nghiên cứu liên quan

Trong phần này, luận án rà soát ngắn gọn DNE, ARE, MMP, và DAG-DNE, chúng là cơ sở cho phương pháp đề xuất.

Nhúng lân cận phân biệt (DNE)

DNE [12] là một phương pháp học không gian con có giám sát, mà tạo ra một đa tạp con nén cho dữ liệu trong cùng lớp trong không gian con thấp chiều được nhúng. Đồng thời, DNE cố gắng tạo ra các khoảng trống giữa các đa tạp con cho các

lớp khác nhau là rộng nhất có thể. Đầu tiên, DNE xây dựng đồ thị lân cận sử dụng k lân cận gần nhất. Ma trận trọng số kề **W** được xác định như sau:

$$w_{ij} = \begin{cases} +1, & \text{nếu } \mathbf{x}_i \text{ và } \mathbf{x_j} \text{ là lân cận và } l(\mathrm{x}_i) = l(\mathrm{x}_j) \\ -1, & \text{nếu } \mathbf{x}_i \text{ và } \mathbf{x_j} \text{ là lân cận và } l(\mathrm{x}_i) \neq l(\mathrm{x}_j) \\ 0, & \text{ngược lại;} \end{cases} \quad (2.1)$$

Tiếp theo, DNE giải bài toán tối ưu sau:

$$\begin{gathered} min \sum_{ij} \left\| \boldsymbol{P}^T \boldsymbol{x}_i - \boldsymbol{P}^T \boldsymbol{x}_j \right\|^2 w_{ij} \\ thỏa\ mãn\ \boldsymbol{P}^T \boldsymbol{P} = \boldsymbol{I} \end{gathered} \quad (2.2)$$

ở đây **I** là ma trận đơn vị. Hàm mục tiêu (2.2) có thể được chia tiếp như sau:

$$\begin{aligned} & \sum_{ij} \left\| \mathbf{P}^T \mathbf{x}_i - \mathbf{P}^T \mathbf{x}_j \right\|^2 w_{ij} \\ & = 2 \sum_{ij} (\mathbf{x}_i^T \mathbf{P} \mathbf{P}^T \mathbf{x}_i - \mathbf{x}_i^T \mathbf{P} \mathbf{P}^T \mathbf{x}_j)\, w_{ij} \\ & = 2tr\{\mathbf{P}^T \mathbf{X}(\mathbf{D} - \mathbf{W})\mathbf{X}^T \mathbf{P}\} \\ & = 2tr\{\mathbf{P}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{P}\} \end{aligned} \quad (2.3)$$

ở đây $\mathbf{L} = \mathbf{D} - \mathbf{W}$ và $d_{ii} = \sum_j w_{ij}$. Do đó, bài toán tối ưu (2.2) có thể được viết lại như sau:

$$\begin{gathered} \min_{\mathbf{P}} tr\{\mathbf{P}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{P}\} \\ \text{thỏa mãn } \mathbf{P}^T \mathbf{P} = \mathbf{I} \end{gathered} \quad (2.4)$$

Ma trận chiếu **P** có thể được tìm thấy bởi giải bài toán giá trị riêng tổng quát sau:

$$\mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{P} = \boldsymbol{\lambda} \mathbf{P} \quad (2.5)$$

Do đó, **P** được cấu tạo gồm r véc tơ chiếu tối ưu tương ứng với r giá trị riêng nhỏ nhất.

Chiếu lỗ cực đại (MMP)

MMP [16] sử dụng cách tiếp cận học đa tạp sử dụng cả dữ liệu có nhãn và chưa có nhãn, và đánh trọng số độ quan trọng cho mẫu có nhãn và chưa có nhãn. MMP có thể thu được bởi giải một bài toán véc tơ riêng tổng quát. Đầu tiên, MMP xây dựng đồ thị lân cận gần nhất G. Ma trận trọng số của G được xác định như sau:

$$w_{ij} = \begin{cases} 1, & \text{nếu } \mathbf{x}_i \text{ và } \mathbf{x_j} \text{ là lân cận} \\ 0, & \text{ngược lại;} \end{cases} \quad (2.6)$$

Để khám phá cả cấu trúc hình học và phân biệt của dữ liệu đa tạp, MMP xây dựng hai đồ thị trong phạm vi lớp (within-class) G_w và liên lớp (between-class) G_b và xác định hai ma trận trọng số tương ứng:

$$w_{ij}^{b} = \begin{cases} +1, \text{ nếu } \mathbf{x}_i \text{ và } \mathbf{x}_j \text{ là lân cận và } l(\mathbf{x}_i) \neq l(\mathbf{x}_j) \\ 0, \text{ ngược lại;} \end{cases} \quad (2.7)$$

$$w_{ij}^{w} = \begin{cases} \gamma, \text{ nếu } \mathbf{x}_i \text{ và } \mathbf{x}_j \text{ là lân cận và } l(\mathbf{x}_i) = l(\mathbf{x}_j) \\ 1, \text{ nếu } \mathbf{x}_i \text{ và } \mathbf{x}_j \text{ là lân cận và } (\mathbf{x}_i \text{ hoặc } \mathbf{x}_j \text{ chưa nhãn}) \\ 0, \text{ ngược lại;} \end{cases} \quad (2.8)$$

Tiếp theo, MMP giải bài toán tối ưu:

$$\underset{\substack{\mathbf{P} \\ \text{thỏa mãn } \mathbf{P^T X D^w X^T P = 1}}}{\text{argmax}} \quad \mathbf{P^T X}(\alpha \mathbf{L^b} + (1-\alpha)\mathbf{W^w})\mathbf{X^T P} \quad (2.9)$$

Ma trận chiếu P mà cực đại (2.9) được cho bởi nghiệm giá trị riêng lớn nhất đối với bài toán giá trị riêng tổng quát:

$$\mathbf{X}(\alpha \mathbf{L^b} + (\mathbf{1}-\alpha)\mathbf{W^w})\mathbf{X^T P} = \lambda \mathbf{X D^w X^T P} \quad (2.10)$$

Các véc tơ cột $\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_d$ là các nghiệm của (2.10) được sắp xếp theo các giá trị riêng của chúng $\lambda_1 > \lambda_2 > \cdots > \lambda_d$.

Nhúng quan hệ gia tăng (ARE)

Phương pháp ARE [15] là phương pháp cơ sở cho MMP. Phương pháp ARE thực hiện tra cứu ảnh với phản hồi liên quan bởi sử dụng ba đồ thị bao gồm đồ thị quan hệ tương tự trên toàn bộ cơ sở dữ liệu ảnh và hai đồ thị quan hệ phản hồi, chúng liên kết các mẫu phản hồi âm và dương do người dùng cung cấp. Hàm mục tiêu của ARE được cho như sau:

$$\underset{\substack{\mathbf{P} \\ \text{thỏa mãn } \sum_{ij} \|\mathbf{P^T x}_i - \mathbf{P^T x}_j\|^2 w_{ij} = 1}}{\text{maximize}} \quad \sum_{ij} \|\mathbf{P^T x}_i - \mathbf{P^T x}_j\|^2 \left(w_{ij}^{N} - \gamma w_{ij}^{P}\right) \quad (2.11)$$

ở đây $\mathbf{P}$ là ma trận biến đổi. $\mathbf{W}^P$ mô tả các quan hệ tương tự dương với $\mathbf{R}$ gồm các mẫu dương:

$$w_{ij}^{P} = \begin{cases} 1, \text{ nếu } \mathbf{x}_i \in \mathbf{R} \wedge \mathbf{x}_j \in \mathbf{R} \\ 0, \text{ ngược lại;} \end{cases}; \quad (2.12)$$

, $\mathbf{W}^N$ mô tả các quan hệ khác nhau với $\mathbf{IR}$ gồm các mẫu âm

$$w_{ij}^{N} = \begin{cases} 1, \text{ nếu } \mathbf{x}_i \in \mathbf{R} \wedge \mathbf{x}_j \in \mathbf{IR} \text{ or } \mathbf{x}_j \in \mathbf{R} \wedge \mathbf{x}_i \in \mathbf{IR} \\ 0, \text{ ngược lại;} \end{cases} \quad (2.13)$$

, và $\mathbf{W}$ là ma trận trọng số của đồ thị lân cận gần nhất.

$$w_{ij} = \begin{cases} e^{-\frac{\rho^2(\mathbf{x}_i,\mathbf{x}_j)}{\tau}}, & \text{nếu } \mathbf{x}_i \in k-NN(\mathbf{x}_j) \\ & \text{hoặc } \mathbf{x}_j \in k-\ NN(\mathbf{x}_i) \\ 0, & \text{ngược lại;} \end{cases} \quad (2.14)$$

Nhúng lân cận phân biệt dựa trên đồ thị lân cận kép (DAG-DNE)

Khác với DNE, DAG-DNE [72] có thể cực đại lề giữa phân tán giữa các lớp (inter-class scatter) và phân tán trong lớp (intra-class scatter) bởi hai ma trận trọng số:

$$w_{ij}^{b} = \begin{cases} +1, \text{ nếu } \mathbf{x}_i \text{ và } \mathbf{x}_j \text{ là lân cận và } l(\mathbf{x}_i) \neq l(\mathbf{x}_j) \\ 0, \text{ ngược lại;} \end{cases} \quad (2.15)$$

$$\mathrm{w}_{\mathrm{ij}}^{\mathrm{w}} = \begin{cases} +1, \text{ nếu } \mathbf{x}_i \text{ và } \mathbf{x}_j \text{ là lân cận và } \mathrm{l}(\mathbf{x}_i) = \mathrm{l}(\mathbf{x}_j) \\ 0, \text{ ngược lại;} \end{cases} \quad (2.16)$$

Hàm mục tiêu của DAG-DNE là:

$$\max_{\substack{\mathrm{P} \\ \text{thỏa mãn } P^TP=I}} \mathbf{P}^{\boldsymbol{T}}\mathbf{X}(\ \mathbf{L}^{\boldsymbol{b}} - \ \mathbf{L}^{\boldsymbol{w}})\mathbf{X}^{\boldsymbol{T}}\mathbf{P} \quad (2.17)$$

Trong phương trình (2.17), nó tương đương với việc tìm các véc tơ riêng $\mathbf{P}$ đầu tiên của $\mathbf{X}(\ \mathbf{L}^{\boldsymbol{b}} - \ \mathbf{L}^{\boldsymbol{w}})\mathbf{X}^{\boldsymbol{T}}$.

2.3. Đề xuất phương pháp học chiếu phân biệt lớp ngữ nghĩa trên dữ liệu đa tạp

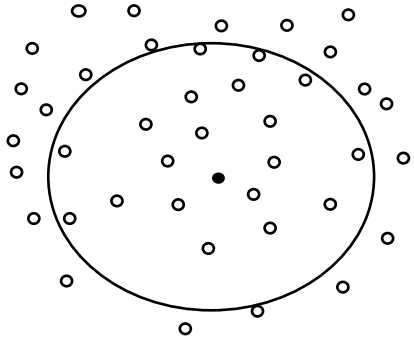
Trong phần này, luận án trình bày phương pháp SCDP đề xuất, mà khai thác cấu trúc hình học cục bộ và phân biệt trong dữ liệu để học một không gian con ngữ nghĩa.

Xây dựng hàm mục tiêu

Bài toán chung cho giảm chiều được phát biểu như sau:

Cho một tập $\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_N$ trong $\mathbb{R}^n$, tìm một ma trận biến đổi $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d)$ mà ánh xạ N điểm này thành một tập $\mathbf{y}_1, \mathbf{y}_2, \dots \mathbf{y}_N$ trong $\mathbb{R}^d$ $(d \ll n)$ sao cho $\mathbf{y}_i$ biểu diễn $\mathbf{x}_i$, ở đây $\mathbf{y}_i = \mathbf{U}^T\mathbf{x}_i$

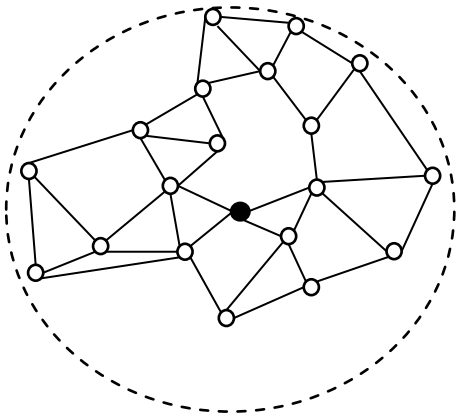
Chúng ta xem xét trường hợp khi dữ liệu nằm trên một không gian con của không gian gốc. Cho $\mathbb{Q} \subset \mathbb{R}^n$ là một không gian đặc trưng ảnh n chiều, và σ: $\mathbb{Q} \times \mathbb{Q} \to \mathbb{R}$ là một hàm khoảng cách nào đó. Cho ma trận $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_N] \in \mathbb{R}^{n \times N}$ biểu diễn N ảnh trong tập ảnh kết quả khởi tạo.



Hình 2.1. Minh họa tra cứu khởi tạo

Hình 2.1 minh họa trực quan của một không gian đặc trưng ảnh 2 chiều gồm các ảnh (điểm hình tròn rỗng). Khi người dùng đưa một ảnh truy vấn (điểm hình tròn đặc), hệ thống tra cứu ảnh truyền thống thu được tập ảnh kết quả khởi tạo thông qua khoảng cách Euclide gồm 20 ảnh (20 điểm nằm trong vòng tròn nét đứt).

Để mô hình cấu trúc hình học cục bộ của không gian con gồm ảnh truy vấn và các ảnh lân cận (láng giềng) với ảnh truy vấn đó, luận án xây dựng một đồ thị quan hệ đặc trưng hay đồ thị lân cận gần nhất G^F. Với mỗi điểm dữ liệu $\mathbf{x}_i$, tìm k lân cận gần nhất, tức là $k - NN(\mathbf{x}_i)$ và đặt một cạnh giữa $\mathbf{x}_\mathrm{i}$ và các lân cận của nó.



Hình 2.2. Đồ thị lân cận gần nhất G^F

Trên Hình 2.2 với mỗi điểm (ảnh) xác định k ảnh (trong trường hợp này k=3) có khoảng cách ngắn nhất (là lân cận hay láng giềng gần nhất) nối cạnh với nhau chúng ta thu được đồ thị quan hệ đặc trưng G^F.

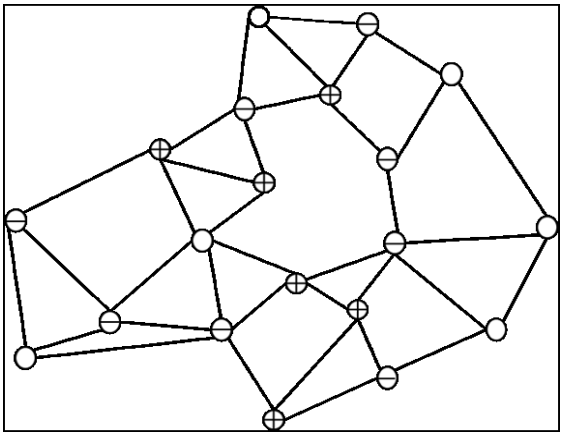
Cho ma trận mà ghi lại các trọng số trên các cạnh của G^F là $\mathbf{W}^{\boldsymbol{F}} \in \mathbb{R}^{N\times N}$**,** được xác định theo công thức (2.14) và được viết lại như sau:

$$w_{ij}^{F} = \begin{cases} e^{-\frac{\rho^2(\mathbf{x}_i,\mathbf{x}_j)}{\tau}}, & \text{nếu } \mathbf{x}_i \in k-NN(\mathbf{x}_j) \\ & \text{hoặc } \mathbf{x}_j \in k-\ NN(\mathbf{x}_i) \\ 0, & \text{ngược lại;} \end{cases} \quad (2.18)$$

ở đây $\rho^2(\mathbf{x}_i, \mathbf{x}_j)$ là độ đo khoảng cách Euclide (L_2), τ là một số vô hướng dương nào đó, và $k-NN$ là ký hiệu cho k lân cận cận gần nhất. Trọng tâm lân cận của $\mathbf{x}_i$ là trung bình của các mẫu bao gồm $\mathbf{x}_i$ và các mẫu có cạnh nối trực tiếp tới $\mathbf{x}_i$.

Đồ thị lân cận gần nhất G^F với ma trận trọng số $\mathbf{W}^{\boldsymbol{F}}$ mô tả hình học cục bộ của đa tạp dữ liệu. Nó thường được sử dụng trong các kỹ thuật theo tiếp cận học đa tạp. Tuy nhiên, đồ thị này không bao gồm thông tin nhãn trong dữ liệu.

Giả sử rằng chúng ta có N_1 điểm được gán nhãn, và N_2 điểm còn lại là chưa có nhãn, ở đây $N_1 + N_2 = N$. Trong tra cứu ảnh, các ảnh có nhãn gồm ảnh truy vấn gốc và các ảnh với phản hồi liên quan của người dùng. Với phản hồi liên quan, trong luận án sử dụng **IR** để biểu thị tập các ảnh được trả về bởi hệ thống mà không liên quan đến ảnh truy vấn, **R** gồm các ảnh được trả về bởi hệ thống mà liên quan đến ảnh truy vấn và **UL** gồm các ảnh chưa có nhãn.



Hình 2.3. Đồ thị lân cận gần nhất G^F sau phản hồi

Trong Hình 2.3 chúng ta thấy tập **IR** gồm 8 điểm điểm hình tròn chứa dấu trừ (ảnh mang nhãn âm), **R** gồm 6 điểm hình tròn chứa dấu cộng (ảnh mang nhãn dương) và **UL** gồm 6 điểm hình tròn rỗng (ảnh không có nhãn)

Để khám phá cả thông tin phân biệt và hình học của đa tạp dữ liệu, luận án xây dựng hai đồ thị mã hóa các quan hệ cặp trong phản hồi liên quan, cụ thể quan hệ tương tự liên quan G^R và không tương tự G^{IR}.

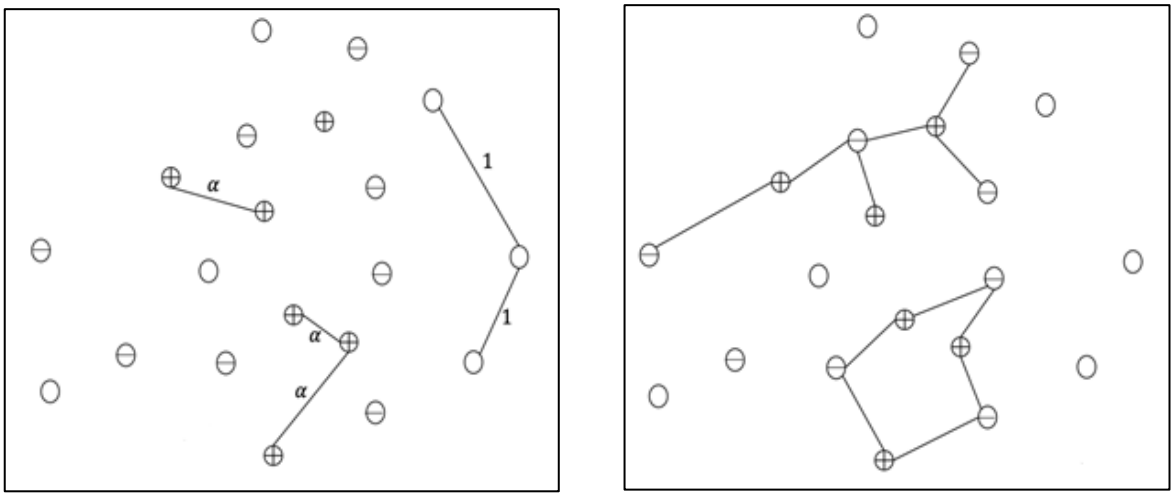
Các ma trận trọng số $\mathbf{W^R} \in \mathbb{R}^{N\times N}$ và $\mathbf{W^{IR}} \in \mathbb{R}^{N\times N}$ của G^R và G^{IR} tương ứng được định nghĩa như sau:

$$w_{ij}^R = \begin{cases} \alpha, \text{nếu } (w_{ij}^F > 0) \wedge (\mathbf{x}_i \in \mathbf{R} \wedge \mathbf{x}_j \in \mathbf{R}) \\ 1, \text{nếu } (w_{ij}^F > 0) \wedge (\mathbf{x}_i \in \mathbf{UL} \wedge \mathbf{x}_j \in \mathbf{UL}) \\ 0, \text{ngược lại;} \end{cases} \quad (2.19)$$

$$w_{ij}^{IR} = \begin{cases} 1, \text{nếu } (w_{ij}^F > 0) \wedge (\mathbf{x}_i \in \mathbf{R} \wedge \mathbf{x}_j \in \mathbf{IR}) \\ \quad \text{hoặc } (w_{ij}^F > 0) \wedge (\mathbf{x}_i \in \mathbf{IR} \wedge \mathbf{x}_j \in \mathbf{R}) \\ 0, \text{ ngược lại;} \end{cases} \quad (2.20)$$

Trong (2.19), khi hai ảnh *i* và *j* thuộc cùng một lân cận và cùng nhãn dương, chúng nên nhận một giá trị trọng số cao α.

Như trong Hình 2.4 (a) biểu thị đồ thị quan hệ tương tự liên quan G^R cho biết mối quan hệ từng cặp giữa các ảnh mang nhãn dương hoặc không có nhãn mà có quan hệ lân cận. Còn mối quan hệ từng cặp ảnh mang hai nhãn khác nhau dương và âm gồm quan hệ lân cận được biểu thị trong đồ thị G^{IR} tại Hình 2.4 (b).

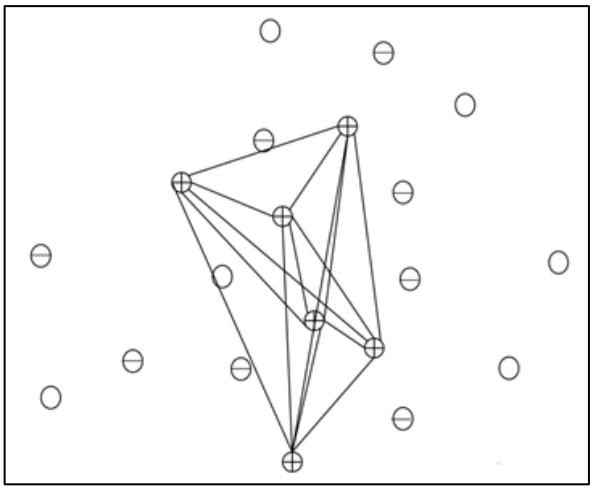


(a) G^R (b) G^{IR}

Hình 2.4. Đồ thị quan hệ G^R và G^{IR}

Chúng ta xác định ma trận $\mathbf{S\_S} \in \mathbb{R}^{N\times N}$ lưu trữ thông tin giống nhau về ngữ nghĩa liên quan với truy vấn giữa hai mẫu $\mathbf{x}_i$ và $\mathbf{x}_j$ (lưu ý rằng hai mẫu $\mathbf{x}_i$ và $\mathbf{x}_j$ không cần thiết thuộc cùng một lân cận):

$$s\_s_{ij} = \begin{cases} 1, \text{nếu } \mathbf{x}_i \in \mathrm{R} \wedge \mathbf{x}_j \in \mathrm{R} \\ 0, \text{ngược lại;} \end{cases} \quad (2.21)$$



Hình 2.5. Đồ thị quan hệ liên quan ngữ nghĩa

Cho **U** là một chiếu mà ánh xạ một mẫu x_i trong không gian gốc thành một mẫu tương ứng $\mathbf{y}_i$ trong một không gian chiều thấp hơn.

$$\mathbf{y}_i = \mathbf{U}^T \mathbf{x}_i \tag{2.22}$$

Hiển nhiên trong lân cận cục bộ của một mẫu $\mathbf{x}_i$ trên đồ thị G^F, trung bình của các mẫu thuộc cùng lân cận và cùng nhãn dương hoặc không có nhãn có thể được tính như sau:

$$\mathbf{m}_i = \sum_{j=1}^{N} \mathbf{x}_j w_{ij}^R \tag{2.23}$$

Sau khi chiếu, trung bình của các mẫu thuộc cùng lân cận và cùng nhãn dương hoặc không có nhãn có thể được tính từ (2.23) và (2.24)

$$\mathbf{m}_i^{(\mathbf{y})} = \sum_{j=1}^{N} \mathbf{y}_j w_{ij}^R \tag{2.24}$$

Một tiêu chuẩn cho chọn một ánh xạ tốt là tối ưu hai hàm mục tiêu (2.25) và (2.26) dưới các ràng buộc thích hợp.

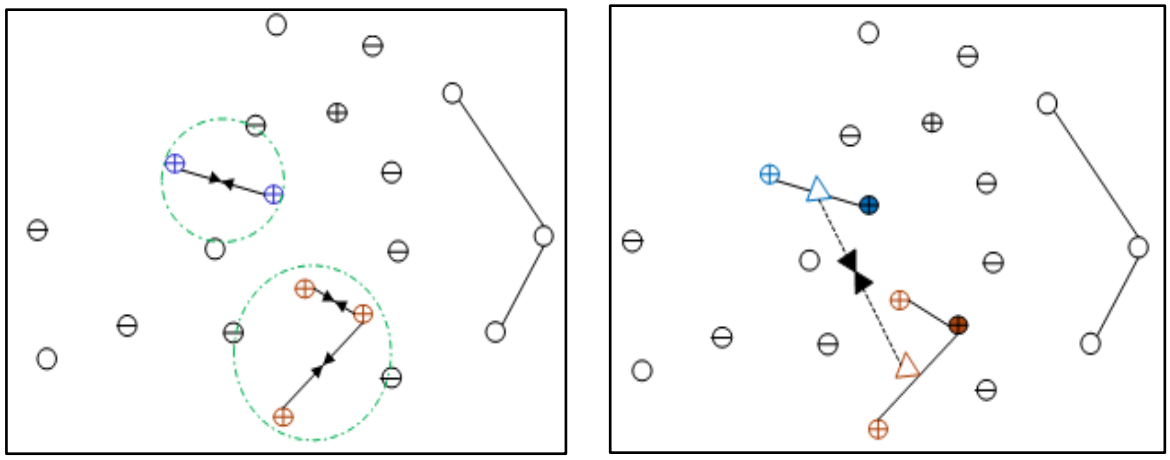
$$\min_{\mathbf{U}} \sum_{ij} \left(\left\| \mathbf{y}_i - \mathbf{y}_j \right\|^2 w_{ij}^R + \left\| \mathbf{m}_i^{(\mathbf{y})} - \mathbf{m}_j^{(\mathbf{y})} \right\|^2 s\_s_{ij} \right) \tag{2.25}$$

$$\max_{\mathbf{U}} \sum_{ij} \left(\left\| \mathbf{y}_i - \mathbf{y}_j \right\|^2 w_{ij}^{IR} + \left\| \mathbf{m}_i^{(\mathbf{y})} - \mathbf{m}_j^{(\mathbf{y})} \right\|^2 (1 - s\_s_{ij}) \right) \tag{2.26}$$

Trong (2.26) đại lượng $(1 - s\_s_{ij})$ cho chúng ta biết thông tin ảnh *i* và ảnh *j* là không cùng nhau liên quan với ảnh truy vấn (ảnh *i* và ảnh *j* là không đồng thời mang nhãn dương). Hai hàm mục tiêu (2.25) và (2.26) của SCDP có cải thiện so với một số phương pháp chiếu khác (được trình bày trong mục 2.2), SCDP tăng tính phân biệt của dữ liệu sau khi chiếu các ảnh trong không gian gốc sang không gian chiếu. Hàm mục tiêu (2.25) bổ sung đại lượng thứ hai biểu diễn các mẫu mang nhãn dương

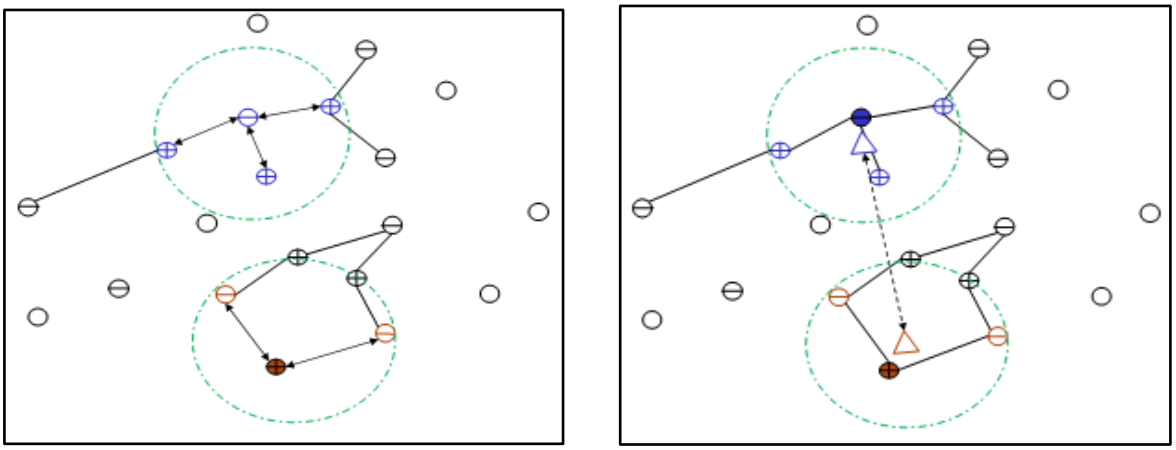
nằm ở hai lân cận khác nhau sẽ được nén gần nhau trong không gian chiếu với số chiều thấp. Mặt khác, hàm mục tiêu (2.26) bổ sung đại lượng thứ hai biểu diễn các mẫu mang nhãn khác nhau nằm ở hai lân cận khác nhau sẽ được nén xa nhau trong không gian chiếu với số chiều thấp.

Trong Hình 2.6 thể hiện ý tưởng công thức (2.25), với $\mathbf{x}_i$, $\mathbf{x}_j$ mang nhãn dương (hình tròn chứa dấu cộng) hoặc không mang nhãn (hình tròn rỗng) thuộc về cùng một lân cận (tức là có cạnh nối với $\mathbf{x}_i$ và $\mathbf{x}_j$) ở Hình 2.6 (a) thì sau khi chiếu $\mathbf{y}_i$, $\mathbf{y}_j$ đảm bảo cũng sẽ gần nhau thuộc cùng một lân cận. Còn với $\mathbf{x}_i$, $\mathbf{x}_j$ đang xét là hai điểm hình tròn chứa dấu cộng mang nhãn dương mà nằm hai lân cận khác nhau (không có cạnh giữa $\mathbf{x}_i$ và $\mathbf{x}_j$) thì cố gắng tối thiểu khoảng cách trọng tâm (điểm hình tam giác) của hai lân cận đó như trong Hình 2.6 (b).



(a) (b)

Hình 2.6. Minh họa ý tưởng công thức (2.26)



(a) (b)

Hình 2.7. Minh họa ý tưởng công thức (2.27)

Bên cạnh đó Hình 2.7 minh họa ý tưởng công thức (2.26), với $\mathbf{x}_i$, $\mathbf{x}_j$ mang nhãn khác nhau (âm hoặc dương) thuộc về cùng một lân cận Hình 2.7 (a) (có cạnh nối với $\mathbf{x}_i$ và $\mathbf{x}_j$) sẽ được ánh xạ xa nhau trong không gian chiếu. Mặt khác nếu hai điểm hình tròn đặc khác nhãn nhau nằm hai lân cận khác nhau như trong Hình 2.7 (b) thì hai lân cận đó sẽ bị đẩy xa nhau tức là tối đa khoảng cách trọng tâm hai cụm (hai điểm hình tam giác).

Phép chiếu tối ưu

Trường hợp $\mathbf{x}_i$ và $\mathbf{x}_j$ thuộc về cùng một lân cận và cùng lớp dương ($\mathbf{m}_i^{(\mathrm{y})} - \mathbf{m}_j^{(\mathrm{y})} = 0$) hoặc không có nhãn ($s\_s_{ij} = 0$) thì $(\mathbf{m}_i^{(\mathrm{y})} - \mathbf{m}_j^{(\mathrm{y})})^2 s\_s_{ij} = 0$, (2.25) suy biến thành:

$$\min_{\mathbf{U}} \sum_{ij} \left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2 w_{ij}^R \tag{2.27}$$

Tức là, hàm mục tiêu (2.25) trên đồ thị G^R chịu một mức độ phạt nặng cao nhất nếu các điểm lân cận $\mathbf{x}_i$ và $\mathbf{x}_j$ bị ánh xạ xa nhau, trong khi thực tế chúng cùng lớp dương hoặc không có nhãn mà cùng lân cận.

Trường hợp $\mathbf{x}_i$ và $\mathbf{x}_j$ có cùng lớp dương hoặc không có nhãn thuộc về hai lân cận khác nhau ($w_{ij}^R = 0$) thì $\left\|\mathbf{y}_i - \mathbf{y}_j\right\|_2^2 w_{ij}^R = 0$, (2.26) suy biến thành :

$$\min_{\mathbf{U}} \sum_{ij} \left\|\mathbf{m}_i^{(\mathrm{y})} - \mathbf{m}_j^{(\mathrm{y})}\right\|^2 s\_s_{ij} \tag{2.28}$$

Tức là, hàm mục tiêu (2.25) trên đồ thị G^F chịu một mức độ phạt nhẹ hơn nếu các điểm lân cận $\mathbf{x}_i$ và $\mathbf{x}_j$ bị ánh xạ xa nhau, trong khi thực tế chúng cùng lớp dương hoặc không có nhãn mà khác lân cận.

Số hạng thứ nhất của bài toán tối ưu (2.25) có thể viết lại như sau:

$$\begin{aligned}
&\left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2 w_{ij}^R \\
&= (\mathbf{U}^T\mathbf{x}_i - \mathbf{U}^T\mathbf{x}_j)^2 = (\mathbf{U}^T(\mathbf{x}_i - \mathbf{x}_j))^2 \\
&= trace\{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{U}\mathbf{U}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)\} \\
&= trace\{\mathbf{U}^T(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{U}\} \\
&= trace(\mathbf{U}^T\mathbf{C}_{\mathrm{x}}\mathbf{U})
\end{aligned}$$

Số hạng thứ hai của bài toán tối ưu (2.25) có thể viết lại như sau:

$$\left\|\mathbf{m}_i^{(\mathrm{y})} - \mathbf{m}_j^{(\mathrm{y})}\right\|^2 s\_s_{ij}$$

$$= (\mathbf{U}^T\mathbf{m}_i - \mathbf{U}^T\mathbf{m}_j)^2$$
$$= (\mathbf{U}^T(\mathbf{m}_i - \mathbf{m}_j))^2$$
$$= trace\{(\mathbf{m}_i - \mathbf{m}_j)^T\mathbf{U}\mathbf{U}^{\mathrm{T}}(\mathbf{m}_i - \mathbf{m}_j)\}$$
$$= trace\{\mathbf{U}^{\mathrm{T}}(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T\mathbf{U}\}$$
$$= trace(\mathbf{U}^{\mathrm{T}}\mathbf{C}_\mathbf{m}\mathbf{U})$$

Bài toán (2.25) được viết lại như sau:

$$arg\ min_{\mathbf{U}^{\mathrm{T}}\mathbf{U}=\mathbf{I}}\ trace(\mathbf{U}^T\mathbf{C}\mathbf{U}) \text{ , trong đó } \mathbf{C} = \mathbf{C}_\mathbf{x} + \mathbf{C}_\mathbf{m} \quad (2.29)$$

Công thức (2.29) cũng tương ứng với việc thu hẹp các điểm liên quan theo các lân cận khác nhau nếu chúng thuộc về các lân cận khác nhau và hàm sẽ nhận một giá nhỏ. Công thức (2.30) cũng tương ứng với việc thu hẹp giữa các điểm liên quan nếu chúng thuộc cùng một lân cận và hàm sẽ nhận một giá trị nhỏ hơn nữa.

Ngoài ra, đối với hàm mục tiêu (2.26), trường hợp x_i và x_j khác lớp và thuộc về hai lân cận khác nhau ($w_{ij}^{IR} = 0$) thì $(\mathbf{y}_i - \mathbf{y}_j)^2 w_{ij}^{IR} = 0$, (2.26) suy biến thành:

$$\max_{\mathbf{U}} \sum_{ij} \left\| \mathbf{m}_i^{(\mathrm{y})} - \mathbf{m}_j^{(\mathrm{y})} \right\|^2 (1 - s\_s_{ij}) \quad (2.30)$$

Tức là, hàm mục tiêu (2.26) trên đồ thị G^F chịu một mức độ phạt nặng nhất nếu các điểm lân cận $\mathbf{x}_i$ và $\mathbf{x}_j$ bị ánh xạ gần nhau, trong khi thực tế chúng khác lớp và khác lân cận.

Trường hợp $\mathbf{x}_i$ và $\mathbf{x}_j$ thuộc về cùng một lân cận và khác lớp $(1 - s_{s_{ij}} = 0)$ thì $(\mathbf{m}_i^{(\mathrm{y})} - \mathbf{m}_j^{(\mathrm{y})})^2(1 - s\_s_{ij}) = 0$, (2.26) suy biến thành:

$$\max_{\mathbf{U}} \sum_{ij} \left\| \mathbf{y}_i - \mathbf{y}_j \right\|^2 w_{ij}^{IR} \quad (2.31)$$

Tức là, hàm mục tiêu (2.26) trên đồ thị G^{IR} chịu một mức độ phạt nhẹ hơn nếu các điểm lân cận $\mathbf{x}_i$ và $\mathbf{x}_j$ bị ánh xạ gần nhau, trong khi thực tế chúng khác lớp và cùng lân cận.

Số hạng thứ nhất của hàm mục tiêu (2.26) có thể viết lại như sau:

$$\left\| \mathbf{y}_i - \mathbf{y}_j \right\|^2 w_{ij}^{IR}$$
$$= (\mathbf{U}^T\mathbf{x}_i - \mathbf{U}^T\mathbf{x}_j)^2$$
$$= (\mathbf{U}^T(\mathbf{x}_i - \mathbf{x}_j))^2$$
$$= trace\{(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{U}\mathbf{U}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)\}$$

$$= trace\{\mathbf{U}^T(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{U}\}$$
$$= trace(\mathbf{U}^T\boldsymbol{B}_{\mathbf{x}}\mathbf{U})$$

Số hạng thứ hai của hàm mục tiêu (2.26) có thể viết lại như sau:

$$\left\|\mathbf{m}_i^{(\mathbf{y})} - \mathbf{m}_j^{(\mathbf{y})}\right\|^2 (1 - s\_s_{ij})$$
$$= (\mathbf{U}^T\mathbf{m}_i - \mathbf{U}^T\mathbf{m}_j)^2$$
$$= (\mathbf{U}^T(\mathbf{m}_i - \mathbf{m}_j))^2$$
$$= trace\{(\mathbf{m}_i - \mathbf{m}_j)^T\mathbf{U}\mathbf{U}^T(\mathbf{m}_i - \mathbf{m}_j)\}$$
$$= trace\{\mathbf{U}^T(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T\mathbf{U})\}$$
$$= trace(\mathbf{U}^T\mathbf{B}_{\mathbf{m}}\mathbf{U})$$

Bài toán tối ưu (2.26) có thể viết lại [73] như sau:

$$arg\ max_{\mathbf{U}^T\mathbf{U}=\mathbf{I}}\ trace(\mathbf{U}^T\mathbf{B}\mathbf{U}), \text{ trong đó } \mathbf{B} = \mathbf{B}_{\mathbf{x}} + \boldsymbol{B}_{\mathbf{m}} \qquad (2.32)$$

Công thức (2.32) tương ứng với việc tách biệt giữa các điểm khác ngữ nghĩa nếu chúng thuộc cùng một lân cận và hàm sẽ nhận một giá trị lớn. Thêm nữa, (2.33) cũng tương ứng với việc tách biệt các điểm khác ngữ nghĩa theo các lân cận khác nhau nếu chúng thuộc về các lân cận khác nhau và hàm sẽ nhận một giá lớn hơn nữa.

Từ hàm mục tiêu (2.29) và (2.32), vấn đề tìm phép chiếu $\mathbf{y} = \mathbf{U}^T\mathbf{x}$ sẽ được đưa về bài toán tối ưu sau:

$$\mathbf{U} = arg\ max_{\mathbf{U}} \frac{trace(\mathbf{U}^T\mathbf{B}\mathbf{U})}{trace(\mathbf{U}^T\mathbf{C}\mathbf{U})} \qquad (2.33)$$

Để đơn giản ta xét trường hợp $(\mathbf{U}^T\mathbf{C}\mathbf{U}) = \mathbf{I}$, bài toán tối ưu trở thành:

$$max_{\mathbf{U}} = trace(\mathbf{U}^T\mathbf{B}\mathbf{U}) \qquad (2.34)$$

thỏa mãn $(\mathbf{U}^T\mathbf{C}\mathbf{U}) = \mathbf{I}$, $\mathbf{B} \in \mathbb{R}^{n\times n}$, $\mathbf{C} \in \mathbb{R}^{n\times n}$

Áp dụng Lagrangian lên (2.34),

$\mathcal{L} = trace(\mathbf{U}^T\mathbf{B}\mathbf{U}) - trace(\wedge^T(\mathbf{U}^T\mathbf{C}\mathbf{U} - \mathbf{I}))$, $với \wedge \in \mathbb{R}^{n\times n}$ là nhân tử Lagranger nhiều biến.

Giải đạo hàm của $\mathcal{L}$ bằng 0 chúng ta được:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{U}} = 2\mathbf{B}\mathbf{U} - 2\mathbf{C}\mathbf{U}\wedge = 0$$
$$\Rightarrow \mathbf{B}\mathbf{U} = \mathbf{C}\mathbf{U}\wedge$$
$$\Rightarrow \mathbf{C}^{-1}\mathbf{B}\mathbf{U} = \mathbf{U}\wedge \qquad (2.35)$$

Vậy $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d)$ là k véc tơ lớn nhất tương ứng với các trị riêng $\wedge = diag\ (\lambda_1, \lambda_{2,} \dots, \lambda_d)$ của ma trận $(\mathbf{C}^{-1}.\mathbf{B})$ với điều kiện $\mathbf{C}$ khả nghịch.

Trong trường hợp ma trận $\mathbf{C}$ không khả nghịch, chúng ta sẽ áp dụng phân tích nhân tử Cholesky $\mathbf{C} = \mathbf{L}\mathbf{L}^{+}$ lên (2.34) như sau:

$$[\mathbf{L}^{\dagger}\mathbf{B}(\mathbf{L}^{\dagger})^{-1}][\mathbf{L}^{\dagger}\mathbf{U}] = \wedge [\mathbf{L}^{\dagger}\mathbf{U}]$$

Khi đó, nghiệm cần tìm $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d)$ là d véc tơ riêng tương ứng với các trị riêng lớn nhất $\wedge = diag\ (\lambda_1, \lambda_{2,} \dots, \lambda_d)$ của ma trận $\mathbf{L}^{\dagger}\mathbf{B}(\mathbf{L}^{\dagger})^{-1}$.

Do đó, để nhúng một ảnh truy vấn $\mathbf{q}$ vào trong không gian đặc trưng, chúng ta sẽ có véc tơ đặc trưng ảnh truy vấn $\mathbf{q}^{(\mathrm{x})} \in \mathbb{Q}$, chúng ta ánh xạ nó vào đa tạp bởi $\mathbf{q}^{(\mathrm{y})} = \mathbf{U}^{T}\mathbf{q}^{(\mathrm{x})}$. Tìm các điểm lân cận của $\mathbf{q}^{(\mathrm{y})}$ sử dụng khoảng cách Euclid, và các ảnh tương ứng với lân cận gần nhất của nó sẽ được phân hạng ở đỉnh trong danh sách trả về.

Thuật toán 2.1 [CT5] dưới đây là thuật toán SCDP (Semantic Class Discriminant Projection), nó thực hiện chiếu dữ liệu từ không gian chiều cao sang không gian con chiều thấp.

Thuật toán 2.1. Thuật toán chiếu phân biệt lớp ngữ nghĩa (SCDP).

Input: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_N\} \in \mathbb{R}^n$ gồm N ảnh với **R, IR, UL** $\subset \mathbf{X}$

R : tập ảnh có nhãn dương,

IR: tập ảnh có nhãn âm,

UL: tập ảnh không có nhãn,

d**:** số chiều không gian chiếu

k, α: các tham số.

Output: Ma trận chiếu $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d)$

Bước 1: $w_{ij}^{F} \leftarrow \begin{cases} e^{-\frac{\sigma^2(\mathbf{x}_i, \mathbf{x}_j)}{\tau}}, & \text{nếu } \mathbf{x}_i \in k - NN(\mathbf{x}_j) \\ & \text{hoặc } \mathbf{x}_j \in k - NN(\mathbf{x}_i) \\ 0, & \text{ngược lại;} \end{cases}$

Bước 2: $w_{ij}^{R} \leftarrow \begin{cases} \alpha, \text{nếu } (w_{ij}^{F} > 0) \wedge (\mathbf{x}_i \in \mathbf{R} \wedge \mathbf{x}_j \in \mathbf{R}) \\ 1, \text{nếu } (w_{ij}^{F} > 0) \wedge (\mathbf{x}_i \in \mathbf{UL} \wedge \mathbf{x}_j \in \mathbf{UL}) \\ 0, \text{ngược lại;} \end{cases}$

$$w_{ij}^{IR} \leftarrow \begin{cases} 1, \text{nếu } (w_{ij}^{F} > 0) \wedge (\mathbf{x}_i \in \mathbf{R} \wedge \mathbf{x}_j \in \mathbf{IR}) \\ \text{hoặc } (w_{ij}^{F} > 0) \wedge (\mathbf{x}_i \in \mathbf{IR} \wedge \mathbf{x}_j \in \mathbf{R}) \\ 0, \text{ ngược lại;} \end{cases}$$

$$s\_s_{ij} \leftarrow \begin{cases} 1, \text{ if } \mathbf{x}_i \in \mathbf{R} \wedge \mathbf{x}_j \in \mathbf{R} \\ 0, \text{ ngược lại;} \end{cases}$$

Bước 3:

$$\mathbf{B} \leftarrow (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T + (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \text{ với } \mathbf{x}_i, \mathbf{x}_j \in w_{ij}^{IR} \text{ và } \mathbf{m}_i = \sum_j \mathbf{x}_j w_{ij}^{IR}$$

$$\mathbf{C} \leftarrow (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T + (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \text{ với } \mathbf{x}_i, \mathbf{x}_j \in w_{ij}^{R} \text{ và } \mathbf{m}_i = \sum_j \mathbf{x}_j w_{ij}^{R}$$

Bước 4: $\mathbf{U} = arg\ max_{\boldsymbol{U}} \frac{trace(\mathbf{U}^T \mathbf{B} \mathbf{U})}{trace(\mathbf{U}^T \mathbf{C} \mathbf{U})}$ với $(\mathbf{U}^T \mathbf{C} \mathbf{U}) = \mathbf{I}$

$\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d)$ với mỗi cột là véc tơ riêng tương ứng với các trị riêng $\lambda_1 > \lambda_2 > \cdots > \lambda_d$.

Độ phức tạp của thuật toán SCDP:

Độ phức tạp của thuật toán SCDP là $O((n+d)n^2$ trong đó *n* là số đặc trưng, *d* là số chiều trong không gian chiếu.

Chứng minh:

Rõ ràng, độ phức tạp của thuật toán SCDP là thời gian thực hiện của các Bước 1, 2, 3 và 4.

Độ phức tạp của Bước 1 là thời gian để xây dựng đồ thị G^F, là $\boldsymbol{O(\frac{1}{2}N^2 n + N^2 logN)}$. Bởi vì $\boldsymbol{\frac{1}{2}N^2 n}$ là thời gian để tính toán khoảng cách các cặp và $\boldsymbol{N^2 logN}$ là thời gian cho *N* lần sắp xếp. Cũng cần lưu ý rằng vì $k \ll N$ (k là số lân cận), vì vậy chúng ta bỏ qua tham số k này trong tính chi phí thời gian.

Độ phức tạp trong Bước 2 là thời gian để tính trọng số của cặp, là $(\frac{1}{2}N^2n)$. Bởi vì chúng ta chỉ cần tính trọng số của cặp, thời gian để xây dựng đồ thị quan hệ phản hồi G^R, là $\frac{1}{2}N^2n$.

Độ phức tạp của Bước 3 phụ thuộc vào thời gian tính toán của (2.30) và (2.33). Độ phức tạp của (2.30) là $O((n+d)n^2)$. Kích thước của ma trận chiếu U là $d \times n$ và của C là $n \times n$. Do đó, ma trận $\mathbf{U}^T\mathbf{CU}$ có kích thước $d \times d$. Độ phức tạp để tính ma trận $\mathbf{U}^T\mathbf{CU}$ là $O(nd^2)$. Do đó, độ phức tạp để tính $trace(\mathbf{U}^T\mathbf{CU})$ là $O(d^3)$. Độ phức tạp để tính toán phân tích giá trị kỳ dị (SVD - Singular Value Decomposition) của ma trận $\mathbf{U}^T\mathbf{CU}$ là $O(d^3)$ và chiếu các điểm vào không gian d chiều dựa trên việc tính toán d giá trị riêng nhỏ nhất của ma trận $\mathbf{U}^T\mathbf{CU}$ với chi phí thời gian là $O(dn^2)$. Do đó, độ phức tạp giải vấn đề (2.30) là $O((n+d)n^2)$. Chi phí thời gian của (2.33) là $O((n+d)n^2)$. Kích thước của ma trận B là $n \times n$ và kích thước của ma trận $\mathbf{U}^T\mathbf{BU}$ là $d \times d$. Độ phức tạp để tính toán ma trận $\mathbf{U}^T\mathbf{BU}$ là $O(nd^2)$. Do đó, độ phức tạp để tính $trace(\mathbf{U}^T\mathbf{BU})$ là $O(d^3)$. Độ phức tạp để tính SVD của ma trận $\mathbf{U}^T\mathbf{BU}$ là $O(d^3)$, và chiếu các điểm vào không gian d chiều dựa trên việc tìm d các giá trị riêng nhỏ nhất của ma trận $\mathbf{U}^T\mathbf{BU}$ có chi phí thời gian là $O(dn^2)$. Vì vậy, giải vấn đề (2.33) cần độ phức tạp là $O((n+d)n^2)$. Do đó, độ phức tạp tính toán của Bước 3 là $O((n+d)n^2)$.

Độ phức tạp tính toán của Bước 4 là thời gian cần thiết để giải bài toán tối ưu hóa (2.37), là $O((n+d)n^2)$. Độ phức tạp cần thiết để tính toán $trace(\mathbf{U}^T\mathbf{BU})$ là $O(d^3)$. Thời gian cần thiết để tính toán ma trận $\mathbf{U}^T\mathbf{CU}$ là $O(nd^2)$. Độ phức tạp để tính SVD của ma trận $\mathbf{U}^T\mathbf{CU}$ là (d^3). Phép chiếu các điểm vào không gian d chiều và tìm d giá trị riêng nhỏ nhất của ma trận $\mathbf{U}^T\mathbf{CU}$ có chi phí thời gian là $O(dn^2)$. Do đó, (2.37) có độ phức tạp là $O((n+d)n^2)$.

Trong ngữ cảnh $n \gg N$, độ phức tạp của thuật toán SCDP là $O((n+d)n^2)$, ngược lại, đó là $O(N^2(n+logN))$. Bên cạnh đó, trong các ứng dụng thực tế, đồ thị G^F thường được xây dựng ngoại tuyến nên chúng ta có thể bỏ qua thời gian xây dựng đồ thị G^F. Vì vậy, áp dụng quy tắc tổng, độ phức tạp của thuật toán SCDP là $O((n+d)n^2)$. Kết luận của mệnh đề đã được chứng minh.

2.4. Tra cứu ảnh với học chiếu phân biệt lớp ngữ nghĩa

Luận án đề xuất phương pháp tra cứu ảnh với học một phép chiếu phân biệt lớp ngữ nghĩa cho giảm chiều (SCDPIR) thông qua phản hồi liên quan với chiếu phân biệt lớp ngữ nghĩa (SCDP) (SCDP đã được trình bày trong 2.3). SCDP được áp dụng trong SCDPIR (Bước 2.4 trong Thuật toán 2.2) để chiếu tập dữ liệu ảnh sang không gian mới thu được tập ảnh kết quả tra cứu gồm nhiều ảnh liên quan với ảnh truy vấn hơn trong không gian gốc (tức là trong không gian chiếu với số chiều thấp, các ảnh mang nhãn dương sẽ gần nhau hơn và các ảnh mang nhãn âm sẽ xa nhau hơn so với trong không gian gốc với số chiều lớn). Thuật toán 2.2 [CT5] ở dưới thực hiện tra cứu ảnh với học chiếu phân biệt lớp ngữ nghĩa cho giảm chiều (SCDPIR).

Thuật toán 2.2. Tra cứu ảnh với học chiếu phân biệt lớp ngữ nghĩa (SCDPIR)**.**

Input: **DB**: Tập dữ liệu ảnh,
q: Ảnh truy vấn khởi tạo,
N: Số lượng ảnh trả về tại mỗi lần lặp
d: số chiều không gian chiếu

Output: **S***:* Tập ảnh kết quả

Bước 1: $\mathbf{X} \leftarrow$Retrieval-Init(**q**, **DB**, *N*>;

Bước 2: **Repeat**

Bước 2.1: $\mathbf{IR} \leftarrow \boldsymbol{Feedback}(\mathbf{X}, -1)$;

Bước 2.2 $\mathbf{R} \leftarrow \boldsymbol{Feedback}(\mathbf{X}, 1)$*;*

Bước 2.3 $\mathbf{UL} \leftarrow \mathbf{X} - (\mathbf{IR} \cup \mathbf{R})$

Bước 2.4 $\mathbf{U} \leftarrow \boldsymbol{SCDP}(\mathbf{X}, \mathbf{R}, \mathbf{IR}, \mathbf{UL}, d, k, \alpha)$;

Bước 2.5 $\mathbf{DB}^{(\mathbf{y})} \leftarrow \boldsymbol{Mapping}(\mathbf{DB}, \mathbf{U})$; $\mathbf{q}^{(\mathbf{y})} \leftarrow \boldsymbol{Mapping}(\mathbf{q}, \mathbf{U})$

Bước 2.6 **S**←Retrieval< $\mathbf{q}^{(y)}$, $\mathbf{DB}^{(y)}$, *N*>;

until (Người dùng dừng phản hồi);

Bước 3. **Return S**;

SCDPIR được thực hiện như sau: Để giảm thời gian tra cứu ảnh, với cơ sở dữ liệu ảnh **DB** đã có, việc xây dựng đồ thị G^F được thực hiện ngoại tuyến (offline) từ trước. Khi bắt đầu quá trình tra cứu, người dùng đưa ảnh truy vấn **q** vào. Hệ thống tiến hành tra cứu khởi tạo trên không gian đặc trưng nhiều chiều (Bước 1). Trong bước này, các ảnh trong cơ sở dữ liệu được phân hạng theo các khoảng cách Euclide của chúng đối với ảnh truy vấn **q**, và *N* ảnh trên cùng (mỗi ảnh được biểu diễn bởi một véc tơ cột của ma trận **X**) được trình bày cho người dùng. Tiếp theo, quá trình lặp phản hồi được thực hiện. Trên tập ảnh được trả về bởi Bước 1, người dùng phản hồi các ảnh liên quan/không liên quan thông qua hàm ***Feedback***(,) và gán cho **IR** hoặc **R** tùy thuộc vào tham số thứ hai là -1 hay 1 (Bước 2.1 và 2.2). Các ảnh chưa được gán nhãn còn lại của tập ảnh trả về sẽ được gán cho **UL** (Bước 2.3). Trên cơ sở tập **R, IR** và **UL**, thuật toán xây dựng các đồ thị G^F, G^R và G^{IR}. Bằng việc áp dụng thuật toán chiếu **SCDP**(, , , ,), chúng ta chiếu các ảnh vào một không gian con chiều thấp để thu được ma trận chiếu **U** (Bước 2.4). Lưu ý rằng trong không gian con chiều thấp này, các ảnh mà thuộc cùng một lớp dương và cùng một lân cận trong không gian chiều cao sẽ có xu hướng gần nhau nhất, các ảnh thuộc cùng lớp dương và khác lân cận sẽ có xu hướng gần nhau thứ hai. Ngược lại, các ảnh mà khác lớp nhau và khác lân cận trong không gian cũ sẽ có xu hướng xa nhau nhất, các ảnh khác lớp nhau mà cùng lân cận trong không gian cũ sẽ có xu hướng xa nhau thứ nhì. Trên cơ sở ma trận chiếu **U** vừa nhận được, hàm ***Mapping***(,) sẽ ánh xạ toàn bộ tập ảnh cơ sở dữ liệu **DB** (hay ảnh truy vấn **q**) sang không gian con có chiều thấp hơn để được $\mathbf{DB}^{(y)}(\mathbf{q}^{(y)})$ (Bước 2.5). Trên không gian con chiều thấp, thuật toán tiến hành tìm *N* ảnh lân cận gần nhất và phân hạng các ảnh theo thứ tự khoảng cách tăng dần. Quá trình này được tiếp tục cho đến khi người dùng dừng phản hồi. Thuật toán trả về tập kết quả gồm các ảnh của tập **S.**

Độ phức tạp của thuật toán SCDPIR:

Độ phức tạp của thuật toán SCDPIR là $O(l + (\mathrm{n} + \mathrm{d})\mathrm{n}^2)$ trong đó l là số ảnh trong tập ảnh, *n* là số chiều của không gian đặc trưng gốc và *d* là số chiều của không gian chiếu.

Chứng minh:

Rõ ràng, độ phức tạp của thuật toán SCDPIR là độ phức tạp tính toán của Bước 1 và Bước 2 trong thuật toán.

Rõ ràng, Bước 1 thực hiện đối sánh giữa hình ảnh truy vấn và mỗi hình ảnh trong cơ sở dữ liệu ảnh, do đó độ phức tạp là $O(l)$ với $l = |DB|$.

Trong Bước 2, số lần lặp là số lượng phản hồi của người dùng, thường là nhỏ và có thể được coi là một hằng số. Do đó độ phức tạp của Bước 2 là thời gian thực hiện các lệnh trong vòng lặp **repeat**… **until**. Phần thân của Bước 2 bao gồm các bước 2.1, 2.2, 2.3, 2.4, 2.5 và 2.6. Xem xét Bước 2.1, bước này thực hiện phản hồi của người dùng về *N* đối tượng hàng đầu của tập **DB**, do đó độ phức tạp là $O(N)$. Tương tự, Bước 2.2 cũng cần thời gian thực hiện là $O(N)$. Độ phức tạp của Bước 2.3 là $O(1)$.

Bước 2.4 gọi hàm **SCDP**(), do đó độ phức tạp là $O((n + d)n^2)$ (vì số lượng phản hồi của người dùng thường rất nhỏ so với số chiều của không gian đặc trưng). Bước 2.5 thực hiện ánh xạ từng ảnh trong tập ảnh **DB**, dựa trên ma trận chiếu **U** nên thời gian là $O(l)$. Tương tự như Bước 1, trên không gian chiếu, Bước 2.6 thực hiện phép so sánh giữa ảnh truy vấn và từng ảnh trong cơ sở dữ liệu ảnh nên thời gian là $O(l)$. Độ phức tạp phần thân của **repeat… until** là $O(l + (n + d)n^2)$. Kể từ khi thân vòng lặp **repeat… until** được thực hiện *e* lần, áp dụng quy tắc nhân, chúng ta có thời gian thực hiện là $O(e(l + (n + d)n^2))$. Vì *e* nhỏ nên chúng ta coi nó như một hằng số và do đó độ phức tạp bước 2 là $O(l + (n + d)n^2)$. Do đó, áp dụng quy tắc cộng, chúng ta nhận được độ phức tạp của thuật toán SCDPIR là $O(l + (n + d)n^2)$. Kết luận của mệnh đề đã được chứng minh.

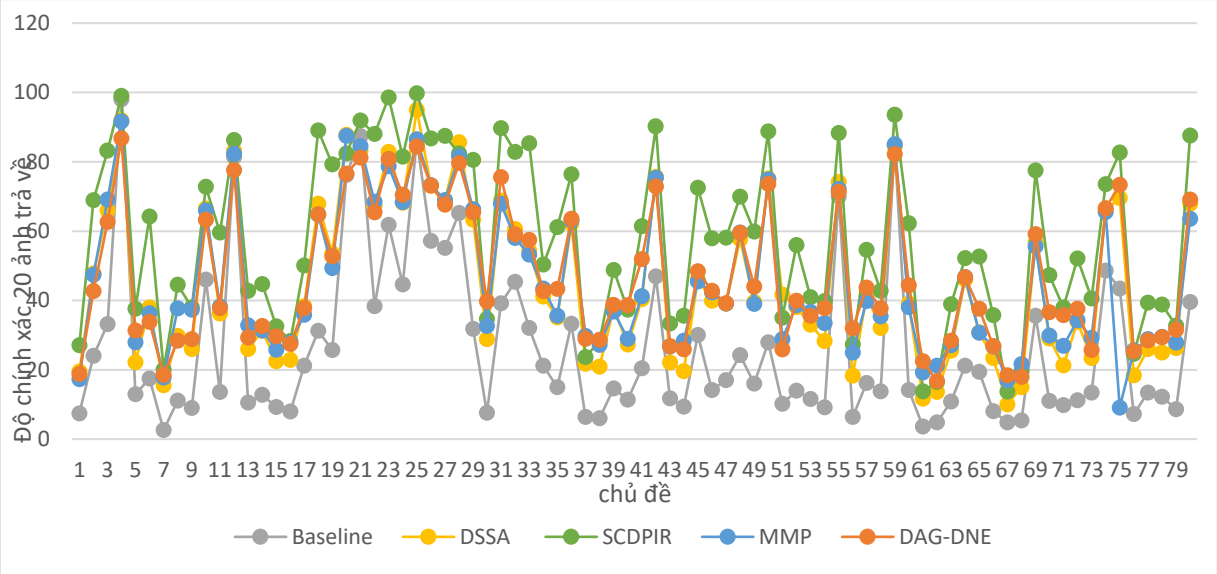
2.5. Đánh giá hiệu năng tra cứu ảnh với học chiếu phân biệt lớp ngữ nghĩa

2.5.1. Độ chính xác tra cứu ảnh

Để minh chứng độ chính xác của thuật toán tra cứu ảnh đề xuất (SCDPIR), luận án so sánh SCDPIR với ba thuật toán tra cứu ảnh sử dụng ba phương pháp chiếu khác nhau, tức là, MMP, DSSA (discriminative semantic subspace analysis) [74] và DAG-DNE. Độ chính xác tra cứu ảnh của SCDPIR được cải thiện so với ba thuật toán kể trên là do SCDPIR sử dụng thuật toán học chiếu phân biệt lớp ngữ nghĩa SCDP giúp tăng tính phân biệt của các ảnh trong tập dữ liệu ảnh trong không gian chiếu. Lý do so sánh thuật toán SCDP với MMP, DAG-DNE và DSSA là vì chúng đưa cấu trúc đa tạp vào bản miêu tả và cố gắng tìm một không gian con mà tại đó (tại không gian chiếu) các khoảng cách Euclide có thể phản ánh tốt hơn ngữ nghĩa của các ảnh. SCDPIR sử dụng thuật toán SCDP dựa trên phản hồi của người dùng xây dựng tập mẫu phản hồi gồm một số ảnh mang nhãn dương hoặc âm và một số mẫu chưa có nhãn. Mỗi lần lặp phản hồi với tập gồm N ảnh trên cùng trả về, luận án sử dụng 2/3 số ảnh trên cùng trong tập N để gán nhãn dương hoặc âm, và 1/3 số ảnh còn lại trong tập N là không gán nhãn. Trong thế giới thực, người dùng thường không cung cấp nhiều vòng lặp phản hồi. Độ chính xác tra cứu sau hai vòng lặp phản hồi đầu tiên (đặc biệt là vòng lặp đầu tiên) là quan trọng nhất. Trong thực nghiệm này, đặt giá trị cho tham số k=12 và $\alpha = 50$.

Kết quả của tập dữ liệu ảnh Corel

Tại lúc bắt đầu tra cứu, các khoảng cách Euclid trong không gian 190 chiều gốc được sử dụng để phân hạng các ảnh trong cơ sở dữ liệu (baseline). Sau khi người dùng cung cấp các phản hồi liên quan, các phương pháp MMP, DSSA, DAG-DNE, và SCDPIR được áp dụng để phân hạng lại các ảnh trong cơ sở dữ liệu. Độ chính xác trong hình này được thực hiện với các kích thước tối ưu là 6 cho SCDPIR, 2 cho MMP, 8 cho DSSA và 12 cho DAG-DNE (xem chi tiết về số lượng chiều tối ưu trong mục 2.5.2) sau vòng lặp phản hồi đầu tiên của 80 chủ đề. Từ kết quả trong Hình 2.8, ta thấy rằng phương pháp truyền thống cho độ chính xác rất thấp so với các phương pháp còn lại. Lý do cho điều này là vì phương pháp truyền thống sử dụng trực tiếp độ đo khoảng cách Euclid trong không gian đặc trưng chiều cao và gặp phải sự chênh lệch giữa các đặc trưng mức thấp và các khái niệm ngữ nghĩa mức cao. Cũng trong hình này, thấy rằng độ chính xác của phương pháp đề xuất là cao nhất so với các phương pháp còn lại.



Hình 2.8. Độ chính xác 5 phương pháp ở 20 ảnh trả về.

Trong Bảng 2.1, các kết quả chi tiết được chỉ ra cho thấy được độ chính xác tra cứu của các phương pháp này thay đổi với các chủ đề khác nhau.

Bảng 2.1. Độ chính xác trung bình tại 20 ảnh trả về của các thuật toán sau vòng lặp phản hồi đầu tiên (%).

STT	Chủ đề	Baseline	DSSA	MMP	DAG-DNE	SCDPIR
1	'art_1'	7.4	19.5	17.3	18.8	**27.1**
2	'art_antiques'	24	47.8	47.4	42.7	**68.9**
3	'art_cybr'	33.2	66	69.1	62.6	**83.2**
4	'art_dino'	98	92	91.6	86.7	**99**
5	'art_mural'	13	22.2	27.9	31.3	**37.6**
6	'bld_castle'	17.5	38	36.3	33.8	**64.2**
7	'bld_lighthse'	2.6	15.5	17.8	18.5	**20.1**
8	'bld_modern'	11.1	29.8	37.7	28.4	**44.5**
9	'bld_sculpt'	9	26	37.3	28.8	**38**
10	'eat_drinks'	46	66.5	66	63.3	**72.8**
11	'eat_feasts'	13.6	36.2	38.3	37.8	**59.6**
12	'fitness'	81.4	83.1	82.3	77.6	**86.3**
13	'obj_234000'	10.6	26	32.8	29.3	**42.8**
14	'obj_aviation'	12.8	31.2	31.3	32.6	**44.7**
15	'obj_balloon'	9.3	22.5	25.8	29.7	**32.5**
16	'obj_bob'	8	22.8	27.8	27.5	**28.3**
17	'obj_bonsai'	21.2	38.4	35.9	37.8	**50.1**

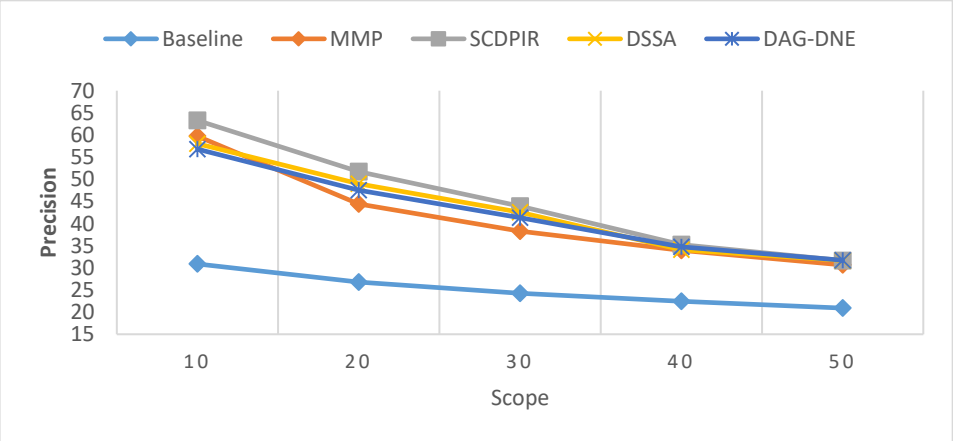
STT	Chủ đề	Baseline	DSSA	MMP	DAG-DNE	SCDPIR
18	'obj_bus'	31.2	67.9	64.8	64.9	**89.1**
19	'obj_car'	25.7	53.6	49.3	52.8	**79.3**
20	'obj_cards'	76.2	**87.8**	87.5	76.6	82.4
21	'obj_decoys'	87.4	83.2	84.5	81.2	**91.9**
22	'obj_dish'	38.4	67.3	68.5	65.4	**88**
23	'obj_doll'	61.8	82.9	78.7	80.8	**98.6**
24	'obj_door'	44.6	68.2	68.4	70.5	**81.5**
25	'obj_eastregg'	86.4	95	86.5	84.4	**99.8**
26	'obj_flags'	57.2	73.1	73.3	73.2	**86.7**
27	'obj_mask'	55.2	68.9	68.9	67.6	**87.5**
28	'obj_mineral'	65.2	**85.6**	81.5	79.6	82.4
29	'obj_moleculr'	31.8	63.3	66.3	65.4	**80.6**
30	'obj_orbits'	7.6	28.8	32.6	**39.7**	34.7
31	'obj_ship'	39.2	68.8	67.9	75.6	**89.7**
32	'obj_steameng'	45.4	60.6	58	59	**82.9**
33	'obj_train'	32.1	54.3	53.2	57.5	**85.4**
34	'pet_cat'	21.2	41.1	43.4	42.7	**50.4**
35	'pet_dog'	15	35.1	35.6	43.3	**61.2**
36	'pl_flower'	33.3	62.2	63.1	63.6	**76.4**
37	'pl_foliage'	6.4	21.7	**29.8**	28.9	23.7
38	'pl_mashroom'	6	20.9	27.2	28.7	**28.3**
39	'sc_'	14.6	37.8	36.8	38.7	**48.8**
40	'sc_autumn'	11.4	27.3	28.9	**38.6**	37.3
41	'sc_cloud'	20.4	40.4	41.2	51.9	**61.4**
42	'sc_firewrk'	46.9	75.7	75.4	72.9	**90.2**
43	'sc_forests'	11.8	22.1	26.7	26.8	**33.4**
44	'sc_iceburg'	9.4	19.6	28.4	25.9	**35.6**
45	'sc_indoor'	30	46.1	45.5	48.4	**72.5**
46	'sc_mountain'	14.2	39.9	42.3	42.8	**57.9**
47	'sc_night'	17	39.1	39.1	39.2	**58.1**
48	'sc_rockform'	24.2	57.7	59.6	59.5	**69.9**
49	'sc_rural'	16	39.6	39	44	**59.9**
50	'sc_sunset'	27.9	75.2	74.9	73.6	**88.8**
51	'sc_waterfal'	10.2	**41.7**	28.8	25.9	35

STT	Chủ đề	Baseline	DSSA	MMP	DAG-DNE	SCDPIR
52	'sc_waves'	14	38.1	38.6	39.9	**56**
53	'sp_ski'	11.6	33	36.7	35.6	**40.9**
54	'texture_1'	9.2	28.3	33.5	37.8	**39.8**
55	'texture_2'	70	74.2	72.2	71.3	**88.3**
56	'texture_3'	6.4	18.3	25	**32**	27.3
57	'texture_4'	16.2	41.4	39.8	43.7	**54.6**
58	'texture_5'	13.8	32.1	35.4	37.7	**42.8**
59	texture_6'	84.8	84.5	85.1	82.2	**93.5**
60	'wl_buttrfly'	14.2	39.1	38.1	44.4	**62.2**
61	'wl_cat'	3.6	11.7	19.2	**22.5**	13.8
62	'wl_cougr'	4.8	13.6	**21.2**	16.7	16.5
63	'wl_deer'	10.8	25.5	27.1	28.4	**38.9**
64	'wl_eagle'	21.2	46	46.8	46.7	**52.2**
65	'wl_elephant'	19.4	30.8	30.7	37.6	**52.7**
66	'wl_fish'	8.1	23.4	26.7	26.8	**35.8**
67	'wl_fox'	4.8	10	16.9	**18.4**	13.7
68	'wl_goat'	5.4	15	**21.6**	17.9	19.6
69	'wl_horse'	35.6	56.6	55.6	59.1	**77.5**
70	'wl_lepoad'	11	29	29.9	36.6	**47.3**
71	'wl_lion'	9.8	21.3	26.9	35.8	**38**
72	'wl_lizard'	11.2	33.8	34.3	37.6	**52.1**
73	'wl_nests'	13.4	23.4	29.2	25.7	**40.6**
74	'wl_owls'	48.6	66.1	65.5	66.6	**73.5**
75	'wl_porp'	43.4	69.6	9.1	73.4	**82.7**
76	'wl_primates'	7.2	18.4	**25.6**	25.3	24.6
77	'wl_roho'	13.4	26	28.8	28.5	**39.4**
78	'wl_tiger'	12.2	25	29.5	29.4	**38.8**
79	'wl_wolf'	8.6	26.3	27.8	31.5	**32.7**
80	'woman'	39.6	67.9	63.6	69.1	**87.6**

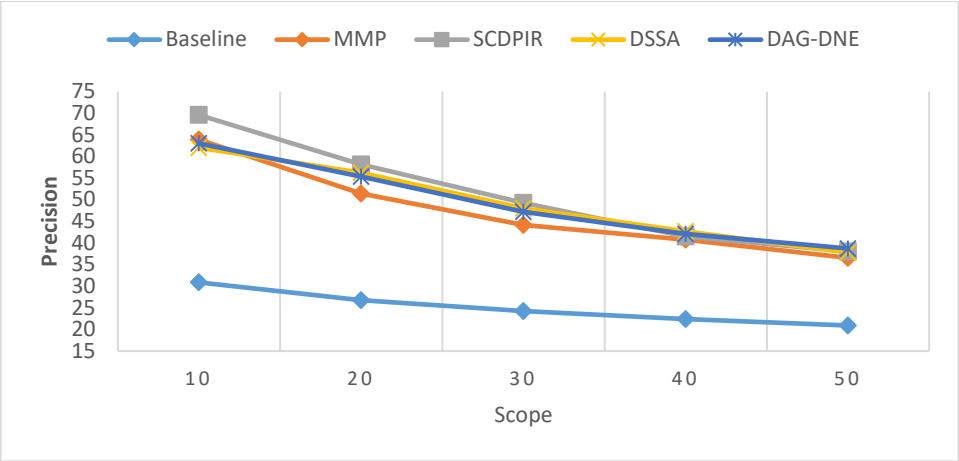
Trong Bảng 2.1, có thể dễ nhận thấy rằng một số chủ đề dễ, tất cả các thuật toán thực hiện tốt, và một số chủ đề khó, tất cả các thuật toán thực hiện cho kết quả nghèo nàn. Trong số 80 chủ đề, phương pháp SCDPIR thực hiện tốt nhất trên 68 chủ đề (giá trị được in đậm). Độ chính xác của phương pháp đề xuất là cao hơn vì nó đảm bảo rằng, trong không gian chiếu, các hình ảnh của cùng một chủ đề, thuộc cùng một

vùng lân cận sẽ gần nhau nhất (ở cấp độ đầu tiên) và các hình ảnh thuộc các vùng lân cận khác nhau sẽ gần nhau hơn (ở cấp độ thứ hai). Ngoài ra, phương pháp được đề xuất cũng đảm bảo rằng, trong không gian chiếu, các hình ảnh không thuộc cùng một chủ đề ,nhưng trong cùng một vùng lân cận sẽ cách xa nhau nhất (ở cấp độ 1) và các hình ảnh thuộc cùng một vùng lân cận sẽ xa hơn xa nhau (ở cấp độ thứ hai). Với 12 chủ đề còn lại, DSSA thực hiện tốt nhất trên 03 chủ đề, MMP thực hiện tốt nhất trên 04 chủ đề, và DAG-DNE thực hiện tốt nhất trên 05 chủ đề.

Các đường cong trung bình độ chính xác - phạm vi (average precision-scope curves) được thể hiện trong Hình 2.9, với các phạm vi số ảnh hàng đầu trả về lần lượt là 10, 20, 30, 40 và 50, của các thuật toán khác nhau cho hai lần lặp phản hồi đầu tiên. Hình 2.9 (a) và Hình 2.9 (b) lần lượt là độ chính xác cho lần lặp phản hồi thứ nhất và thứ hai. Tại phạm vi 10, 20, và 30, thuật toán SCDPIR thực hiện tốt hơn các thuật toán còn lại. Độ chính xác của SCDPIR là tương tự với các phương pháp so sánh ở phạm 40 và 50. Các độ chính xác của DSSA và DAG-DNE là rất gần nhau. Tại phạm vi 10, MMP thực hiện tốt hơn DSSA và DAG-DNE. Nhưng cả DSSA và DAG-DNE đều thực hiện tốt hơn MMP ở các phạm vi sau đó. Tất cả bốn phương pháp SCDPIR, DSSA, DAG-DNE, và MMP là tốt hơn đáng kể so với Baseline, nó chỉ ra rằng các phản hồi liên quan được người dùng cung cấp là rất hữu ích trong cải tiến độ chính xác tra cứu. Do đó, chúng ta có thể kết luận được rằng phương pháp đề xuất SCDPIR đã cải thiện độ chính xác trong việc học một không gian con lớp ngữ nghĩa với phản hồi liên quan.



a) lần lặp phản hồi thứ nhất.



(b) lần lặp phản hồi thứ hai.

Hình 2.9. Các đường cong precision-scope trung bình của các thuật toán khác nhau cho hai lần lặp đầu tiên.

Bảng 2.2 cho thấy thời gian thực hiện trung bình của các phương pháp khác nhau. Thời gian thực hiện là thời gian từ lúc hệ thống nhận được ảnh truy vấn tra cứu thu được tập ảnh tra cứu khởi tạo phản hồi và lấy thông tin đánh giá phản hồi của người dùng cho đến lúc hệ thống trả về kết quả mới. Thời gian thực hiện truy vấn trung bình cho bốn phương pháp này rất nhanh (nhỏ hơn 0,05s cho lần lặp phản hồi đầu tiên). Thời gian thực hiện phương pháp SCDPIR chậm hơn MMP và nhanh hơn DAG-DNE và tương đương với DSSA. Phương pháp SCDPIR chậm hơn một chút so với phương pháp MMP vì nó phải tính toán nhiều thông tin hơn cho các đối tượng trong cùng một chủ đề nhưng thuộc các hàng xóm khác nhau và ngược lại. Cấu hình máy tính cho thí nghiệm là máy Intel Core i5 Catalina 3,1 GHz Dual-Core với bộ nhớ LPDDR3 8 GB 2133 MHz.

Bảng 2.2. Trung bình thời gian thực thi khi tra cứu một truy vấn

Phương pháp	**Trung bình thời gian thực thi (s)**	
	Phản hồi lần 1	**Phản hồi lần 2**
SCDPIR	0.045	0.063
DAG-DNE	0.049	0.067
DSSA	0.046	0.059

MMP	0.034	0.052

Bảng 2.3 cho thấy thời gian thực hiện của mỗi bước trong thuật toán SCDPIR cho ba hình ảnh thử nghiệm. Ba hình ảnh thử nghiệm này, tương ứng với ba hình ảnh truy vấn, được chọn ngẫu nhiên bởi chương trình trong tập dữ liệu ảnh COREL được mô tả trong phần 1.4.2. Những hình ảnh thử nghiệm này có ID 80, 331 và 1572 và chúng thuộc về các chủ đề art_1, art_cybr và fitness tương ứng.

Bảng 2.3. Thời gian thực hiện từng bước trong thuật toán SCDPIR.

Bước	**Thời gian thực hiện theo số vòng lặp (s)**					
	ID 80		**ID 331**		**ID 1572**	
	Lần 1	**Lần 2**	**Lần 1**	**Lần 2**	**Lần 1**	**Lần 2**
1	0.0012	0.0012	0.0018	0.0018	0.0013	0.0014
2	0.04872	0.06144	0.04648	0.06269	0.04683	0.06059
2.1	0.0015	0.0015	0.0011	0.0011	0.0015	0.0015
2.2	0.0016	0.0016	0.0014	0.0014	0.0012	0.0012
2.3	0.0017	0.0017	0.0015	0.0015	0.0018	0.0018
2.4	0.0394	0.0523	0.0389	0.0551	0.0594	0.05181
2.5	0.0035	0.0033	0.0025	0.0024	0.0019	0.0032
2.6	0.00102	0.00104	0.00108	0.00119	0.00103	0.00108
Tổng thời gian	0.04992	0.06264	0.04828	0.06449	0.04813	0.06199

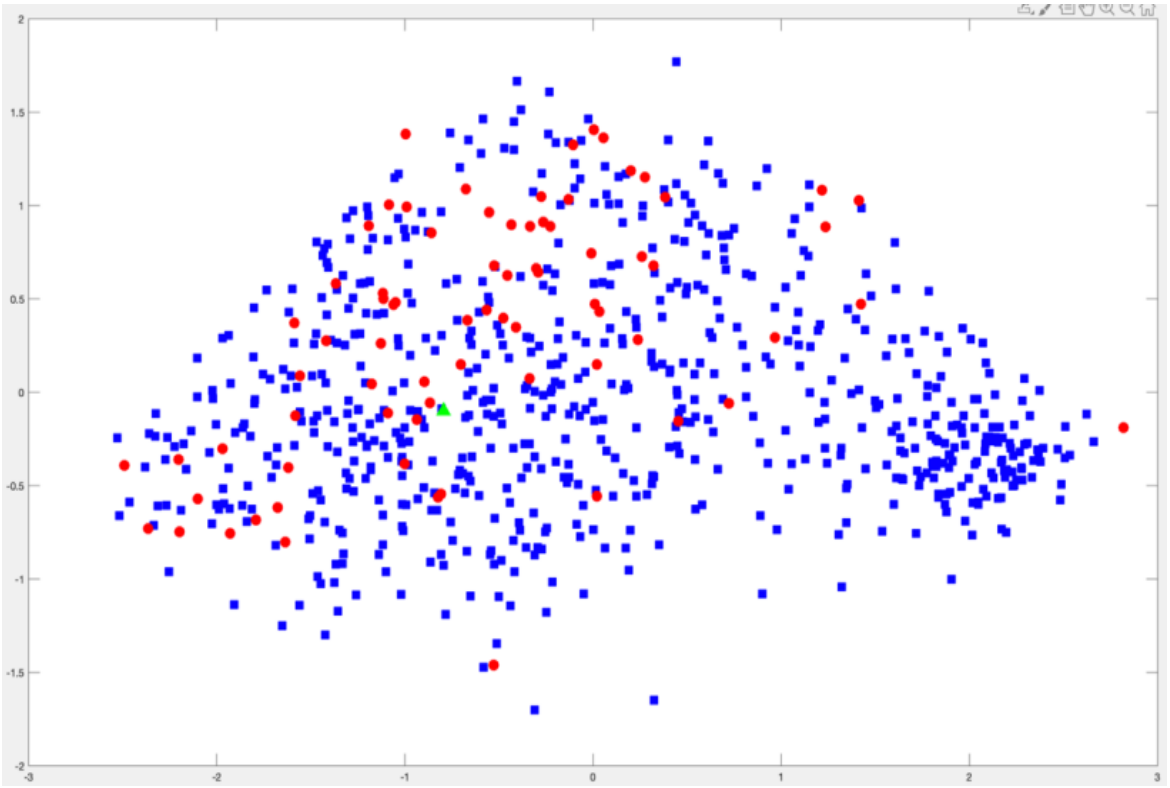
Kết quả của tập dữ liệu ảnh SIMPLIcity

Với tập Corel 10800 cho ta thấy độ chính xác của phương pháp đề xuất đã cải thiện đáng kể, nhưng để trực quan hóa phép chiếu phương pháp đề xuất tập Corel không tối ưu vì số lượng ảnh quá nhiều. Do đó trong phần này, các thực nghiệm được thực hiện trên tập dữ liệu ảnh SIMPLIcity có 1000 ảnh để trình bày việc trực quan hóa kết quả của bốn phương pháp. Sau đó, độ chính xác của phương pháp đề xuất so với các phương pháp khác được mô tả.

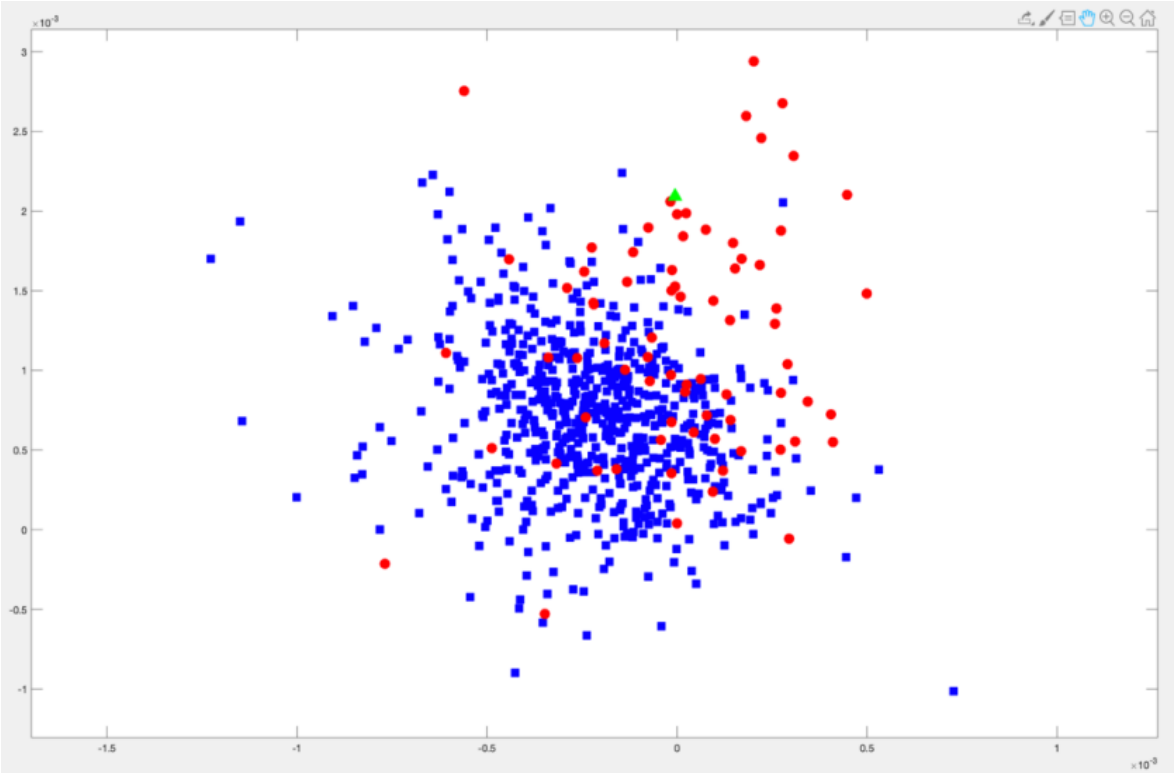
Thực nghiệm lấy ngẫu nhiên một ảnh có ID là 243 từ lớp ‘Buildings’ của tập dữ liệu SIMPLIcity. Do số chiều của ảnh là cao và vì mục đích trực quan hóa nên thực nghiệm sẽ minh họa trên hai chiều của véc tơ đặc trưng. Hình 2.10 minh họa các phân bố mẫu trên mặt phẳng 2D cho một truy vấn với ID là 243 trong lớp ‘Buildings’.

Vị trị của ảnh truy vấn và các ảnh cơ sở dữ liệu, sau các giai đoạn tra cứu được chỉ ra trong Hình 2.10 (a). Các kết quả tra cứu khởi tạo được chỉ ra trong Hình 2.10 (b), trong hình này ảnh truy vấn được biểu thị bởi hình tam giác màu xanh lá. Các đường tròn màu đỏ biểu diễn các mẫu trong cơ sở dữ liệu là liên quan với ảnh truy vấn trong khi dấu hình vuông màu xanh biển biểu thị các ảnh cơ sở dữ liệu không liên quan với ảnh truy vấn. Trong giai đoạn khởi tạo, ảnh truy vấn được trộn với các mẫu liên quan và không liên quan. Sau đó, phản hồi của người dùng được tận dụng để sinh ra một biến đổi mới cho vòng tra cứu tiếp theo. Các kết quả tra cứu sau vòng lặp phản hồi đầu tiên được chỉ ra trong Hình 2.10 (c), (d), (e), và (f) cho các phương pháp MMP, DSSA, DAG-DNE và SCDPIR tương ứng.

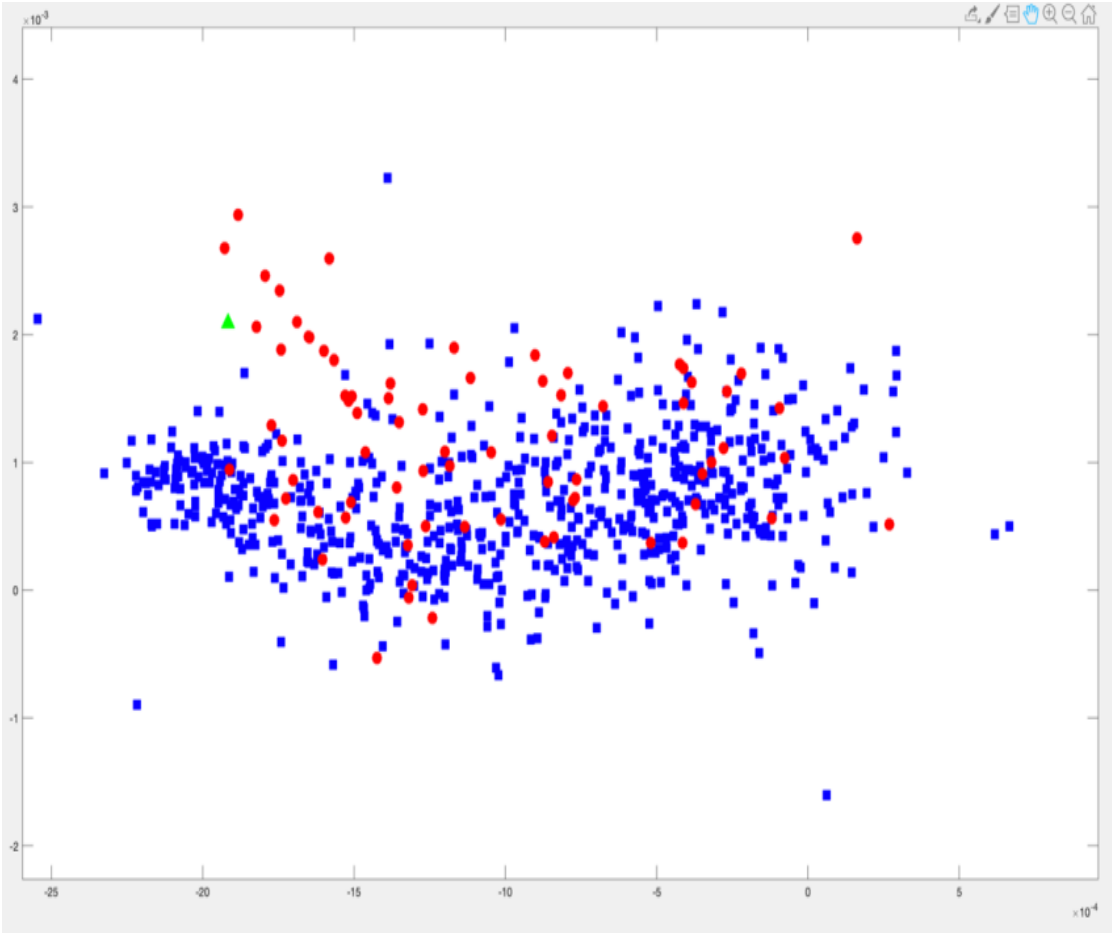
(a)



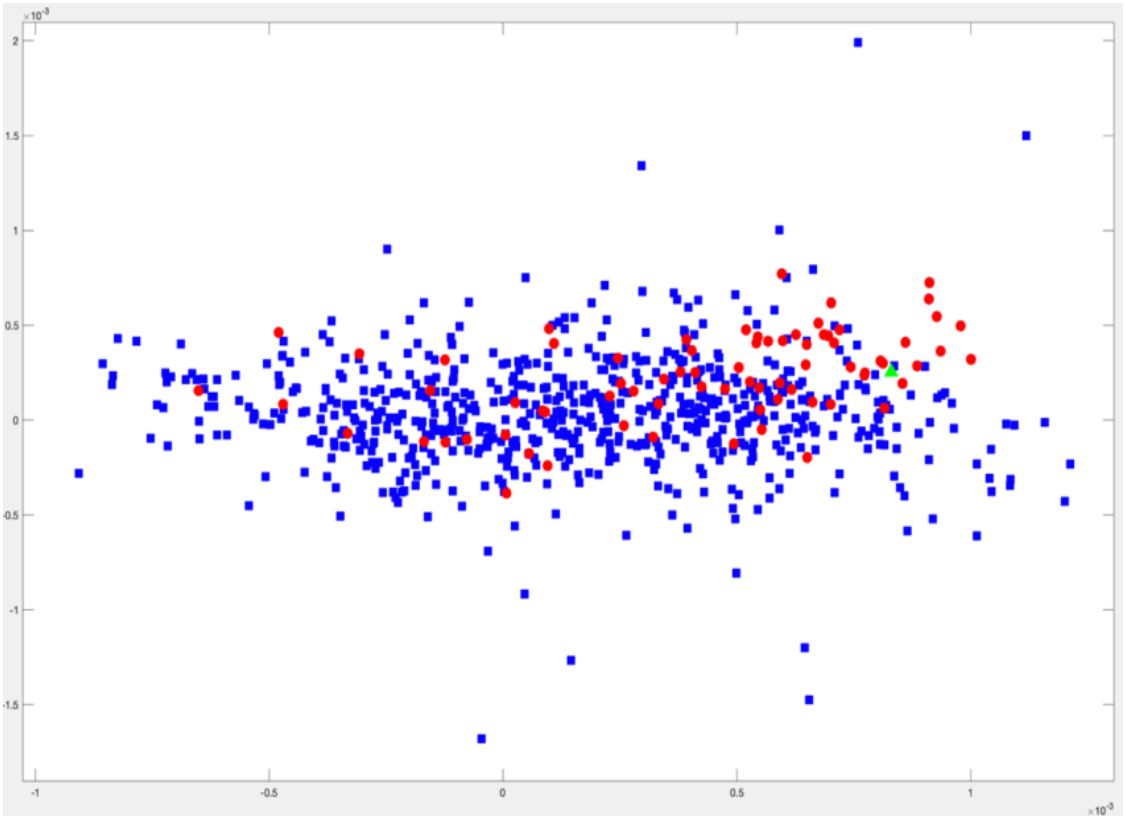
(b)



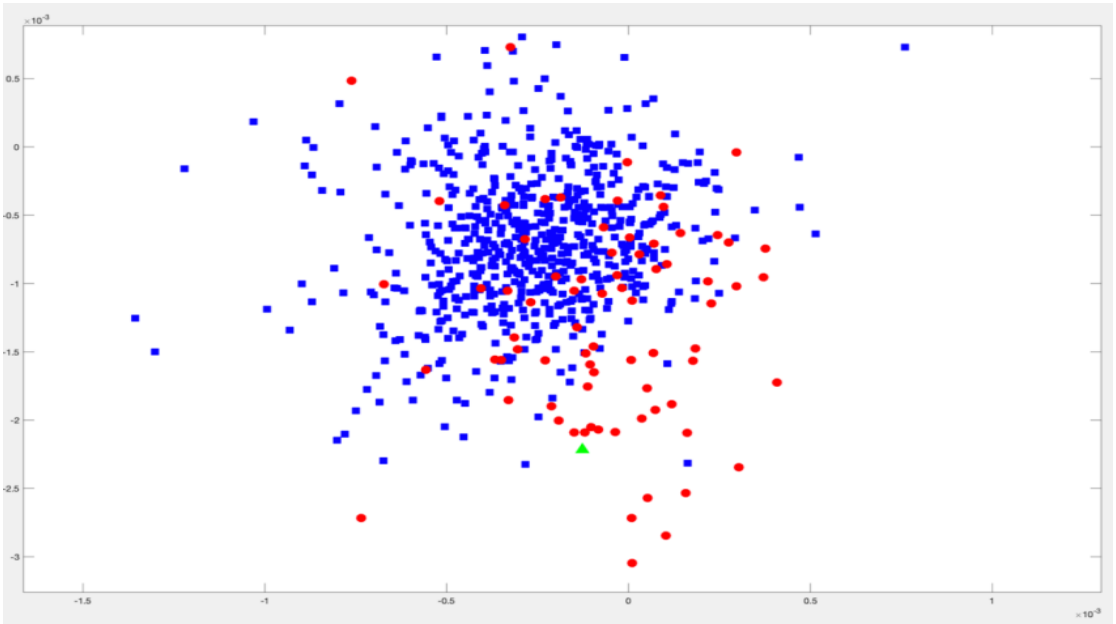
(c)



(d)



(e)



(f)

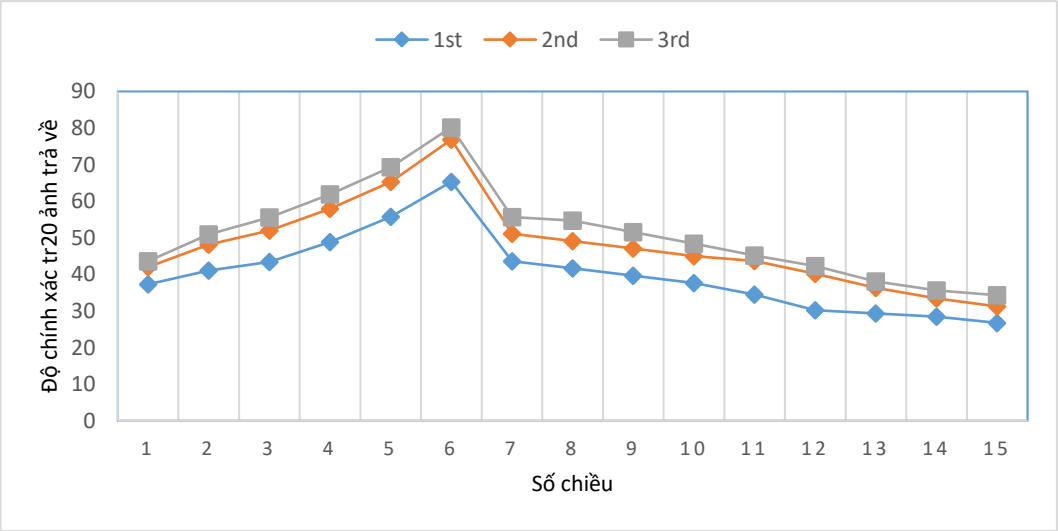
Hình 2.10. Phân phối mẫu cho ảnh truy vấn id 243 (a), chủ đề "Building" với các phương pháp baseline (b), MMP (c), DSSA (d), DAG-DNE (e), và SCDPIR (f).

Xem xét các mẫu trên mặt phẳng chiếu sử dụng các phương pháp MMP, DSSA, DAG-DNE, và SCDPIR sau vòng lặp thứ nhất như được chỉ ra trong các Hình 2.10 (c), (d), (e), và (f); Các điểm liên quan (hình tròn) có xu hướng gần mẫu truy

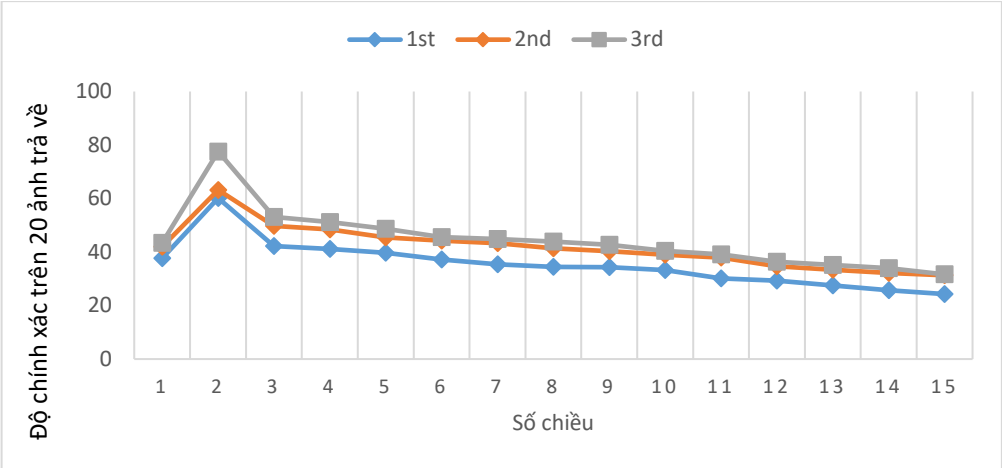
vấn (hình tam giác) cho các phương pháp. Tuy nhiên, khi sử sụng phương pháp SCDPIR thì xung quanh mẫu truy vấn có nhiều mẫu liên quan hơn. Mặt khác, các ảnh không liên quan (dấu vuông xanh biển) được đẩy ra xa hơn từ truy vấn bởi phép biến đổi. Từ các kết quả này, chúng ta thấy phương pháp SCDPIR thực hiện tốt hơn pháp MMP, DSSA, DAG-DNE.

2.5.2. Chiều của không gian chiếu phân biệt lớp ngữ nghĩa

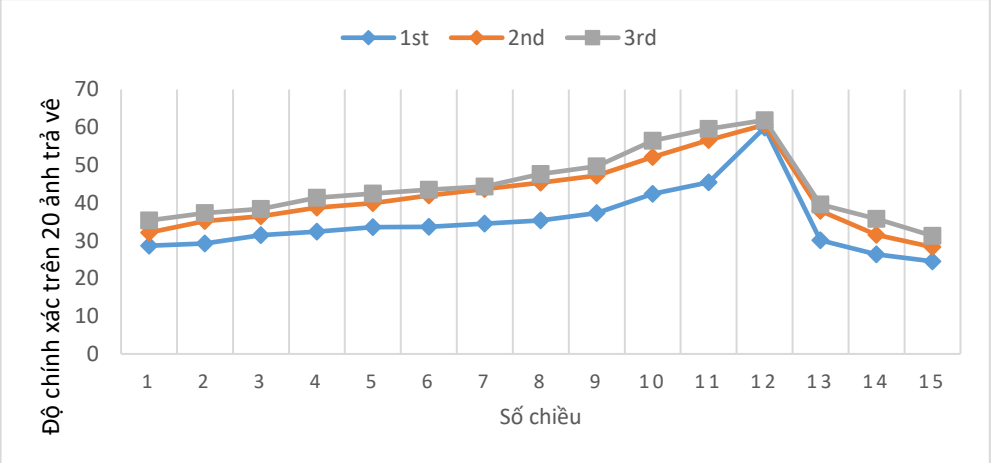
Bốn phương pháp SCDP, DSSA, DAG-DNE và MMP đều theo tiếp cận học đa tạp. Vậy chiều của không gian con của bốn thuật toán này như thế nào? Hình 2.11 chỉ ra độ chính xác tra cứu của bốn phương pháp này theo số chiều trên tập dữ liệu ảnh COREL. Mỗi đường cong trong Hình 2.11 tương ứng với mỗi vòng lặp thứ nhất, thứ hai và thứ ba sau khi phản hồi. Chúng ta thấy rằng độ chính xác của MMP luôn nhận được độ chính xác tốt nhất tại hai chiều (Hình 2.11 (b)), độ chính xác của SCDP luôn có độ chính xác tốt nhất tại sáu chiều (Hình 2.11 (a)), DSSA đạt độ chính xác tốt nhất tại số chiều rất lớn là 8 chiều (Hình 2.11 (d)), và DAG-DNE đạt độ chính xác tốt nhất tại số chiều rất lớn là 12 chiều (Hình 2.11 (c)). Như vậy, số chiều chiếu tối ưu của SCDPIR cao hơn của MMP nhưng thấp hơn của DAG-DNE và DSSA. Nhưng, hiệu suất của SCDPIR cao hơn nhiều so với MMP khi nó ở số chiều tương đối thấp và điều này có thể chấp nhận được trong các ứng dụng thực tế. Ngoài ra, với thuật toán DAG-DNE, độ chính xác đạt được tốt nhất với số chiều tương đối lớn cao và nó sẽ bị vấn đề quá khớp khi áp dụng trong các ứng dụng tại thế giới thực.



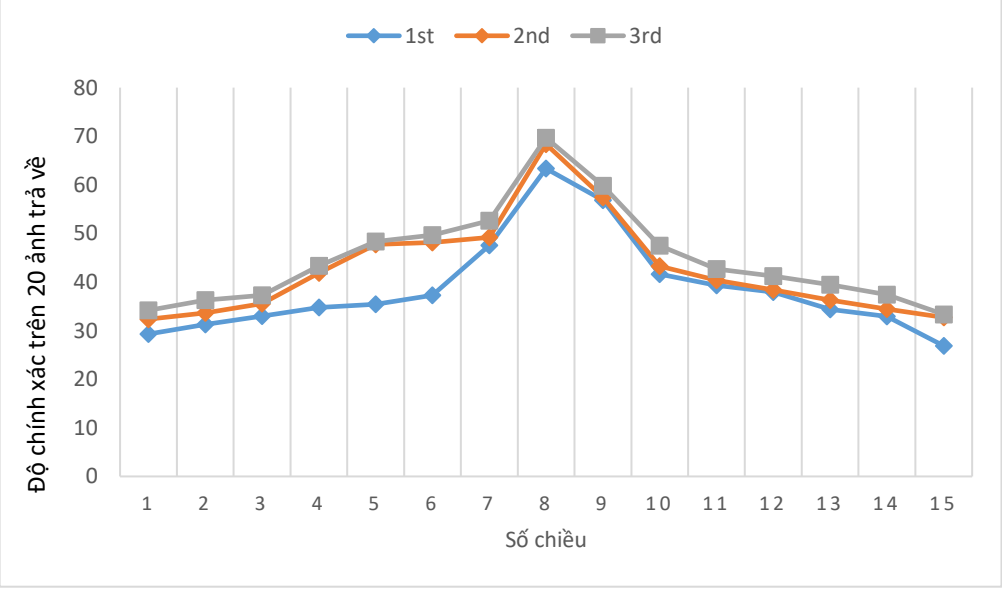
(a) Chiều của không gian con (SCDP)



(b) Chiều của không gian con (MMP)



(c) Chiều của không gian con (DAG-DNE)



(d) Chiều của không gian con (DSSA)

Hình 2.11. Độ chính xác của bốn phương pháp theo số chiều.

2.6. Kết luận chương 2

Trong chương này, luận án trình bày phương pháp SCDP có thể khám phá được

cấu trúc phi tuyến của dữ liệu trên không gian gốc để tìm được ma trận chiếu. Sau khi có được mà trận chiếu, các ảnh trong không gian gốc chiều lớn sẽ được ánh xạ chiếu sang một không gian mới có số chiều thấp hơn rất nhiều. Do đó nó giải quyết vấn đề quá khớp của mô hình phân lớp cho pha phản hồi trong tra cứu ảnh. Bên cạnh đó, trong chương 2 đã đánh giá thực nghiệm trên hai tập dữ liệu được cộng đồng CBIR sử dụng rộng rãi là Corel 10800 và SIMPLIcity. Kết quả trên các đồ thị đã chỉ ra rằng độ chính xác của phương pháp đề xuất đã được cải thiện và đáng tin cậy.

CHƯƠNG 3. CÂN BẰNG TẬP MẪU PHẢN HỒI VÀ KẾT HỢP TRA CỨU ẢNH ĐA KHÍA CẠNH

Trong chương 2, luận án đề xuất một phương pháp tra cứu ảnh [CT5] để giải quyết vấn đề "lời nguyền về số chiều". Tuy nhiên, nó vẫn còn gặp phải một số hạn chế sau: (1) chưa giải quyết được vấn đề mất cân bằng số mẫu giữa hai lớp dương và âm trong quá trình phản hồi, do số các mẫu nhãn âm thuộc nhiều chủ đề khác nhau trong kho số các mẫu nhãn dương thuộc về một chủ đề; (2) Chỉ sử dụng một bộ phân lớp để tạo ra kết quả tra cứu nên cho độ chính xác chưa cao bởi vì một bộ phân lớp không thể biểu diễn hết các khía cạnh hữu ích khác nhau của một đối tượng (một đối tượng có thể bao gồm nhiều khía cạnh hữu ích khác nhau). Để giải quyết hai hạn chế ở trên, chương này đề xuất phương pháp tra cứu ảnh học bán giám sát dựa vào đồ thị để giải quyết vấn đề mất cân bằng mẫu và khai thác được các khía cạnh hữu ích khác nhau của một đối tượng.

3.1. Giới thiệu

Như đã trình bày ở Chương 1 và Chương 2, bài toán tra cứu ảnh dựa trên học máy với thông tin phản hồi liên quan có nhiều điểm khác so với bài toán phân lớp, hồi quy dựa trên học máy. Sự khác nhau thể hiện rõ rệt nhất là về số lượng ảnh có nhãn trong tập ảnh huấn luyện, trong RF, số lượng ảnh có nhãn thu được bởi người dùng phản hồi là hạn chế [70]. Do đó, những phương pháp học máy không đòi hỏi số lượng mẫu huấn luyện lớn sẽ phù hợp với RF. RF dựa vào máy véc tơ hỗ trợ là một trong những cách tiếp cận học với số mẫu nhỏ cho hiệu quả tốt vì khả năng tổng quát tốt của nó [20, 42-45]. Như đã trình bày trong mục 1.2.3, một số phương pháp áp dụng SVM cho quá trình RF đã cải thiện độ chính xác tra cứu với nhiều hướng khác nhau nhưng nhìn chung vẫn gặp phải một số hạn chế. Hầu hết những phương pháp tra cứu ảnh sử dụng SVM thường không quan tâm đến những ảnh chưa được gán nhãn (nhãn dương hoặc âm) dù chúng rất hữu ích cho quá trình học phản hồi hay giảm chiều để nâng cao độ chính xác tra cứu. Việc kết hợp thông tin hữu ích của các ảnh chưa gán nhãn vào RF là một vấn đề thách thức lớn vì để thu thập xây dựng tập phản hồi có nhãn mất rất nhiều công sức về thời gian và có thể dẫn đến nhầm lẫn trong các hệ thống thực tế. Mặt khác, chúng còn bỏ qua sự mất cân bằng số mẫu dương và âm trong tập phản hồi trong khi với RF thì điều này thường xuyên xảy ra

do sự khác nhau rất rõ rệt trong hai nhóm phản hồi mang nhãn dương và âm. Tức là trong tập ảnh phản hồi mang nhãn dương thì chỉ gồm những ảnh có cùng một khái niệm (cùng một chủ đề) nhưng với tập phản hồi mang nhãn âm thì lại không phải thế, chúng bao gồm những ảnh mang nhiều khái niệm khác nhau (nằm rải rác ở rất nhiều chủ đề khác nhau).

Để minh chứng về độ chính xác tra cứu kém hiệu quả khi áp dụng SVM vào quá trình RF cho tra cứu ảnh do sự mất cân bằng mẫu trong quá trình RF, luận án thực hiện một thực nghiệm đánh giá trên tập Corel. Trong thực nghiệm này, luận án sử dụng năm đặc trưng (lược đồ màu, tương quan màu, mô men màu, đặc trưng Gabor, đặc trưng biến đổi wavelet) được trích rút tự động từ tập ảnh cơ sở dữ liệu Corel[1] 10800 như đã trình bày trong mục 1.4.2 tạo thành tập véc tơ đặc trưng, mỗi véc tơ đặc trưng có độ dài 190 chiều. Thực nghiệm tiến hành lấy ngẫu nhiên 30 ảnh trong tập dữ liệu ảnh Corel 10800 để lấy từng ảnh đó đưa vào hệ thống tra cứu làm ảnh truy vấn. Với mỗi ảnh truy vấn, hệ thống so sánh với toàn bộ tập véc tơ đặc trưng trong tập dữ liệu ảnh thông qua độ đo khoảng cách Euclid cho chúng ta tập ảnh kết quả tra cứu khởi tạo gồm 100 ảnh trên cùng có khoảng cách nhỏ nhất. Tiếp theo, hệ thống tự động gán nhãn dương hoặc âm cho 100 ảnh đó dựa trên tập tin cậy nền (cho biết chủ đề của mỗi ảnh) thu được tập ảnh phản hồi. Sau đó, áp dụng SVM hai lớp (đã trình bày tại mục 1.3.2) trên tập ảnh phản hồi đã được gán nhãn đó tìm được siêu phẳng phân tách hai lớp dương và âm. Cuối cùng, hệ thống thực hiện phân hạng lại tập ảnh trong cơ sở dữ liệu theo khoảng cách từng ảnh tới siêu phẳng phân tách (xem chi tiết tại mục 1.3.2) để thu được tập kết quả mới. Các số liệu kết quả thực nghiệm được thể hiện như trong Bảng 3.1

Bảng 3.1. Độ chênh lệch giữa hai nhóm dương âm của mỗi truy vấn.

STT	Mã ảnh truy vấn	Số mẫu dương	Số mẫu âm	Tỉ lệ chênh lệch (%)
1	8114	17	83	79.5
2	2755	27	73	63

[1] https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval (Download lúc 6:32 AM ngày 25/12/2016)

STT	Mã ảnh truy vấn	Số mẫu dương	Số mẫu âm	Tỉ lệ chênh lệch (%)
3	**5464**	**5**	**95**	**94.7**
4	*7548*	*46*	*54*	*14.8*
5	**9619**	**4**	**96**	**95.8**
6	*10356*	*42*	*58*	*27.6*
7	5907	17	83	79.5
8	**1497**	**8**	**92**	**91.3**
9	1612	78	22	71.8
10	2779	14	86	83.7
11	**9072**	**1**	**99**	**99**
12	2744	23	77	70.1
13	8785	15	85	82.4
14	2627	25	75	66.7
15	**10024**	**3**	**97**	**96.9**
16	*3775*	*47*	*53*	*11.3*
17	**2121**	**4**	**96**	**95.8**
18	2708	16	84	81
19	**6643**	**3**	**97**	**96.9**
20	*5103*	*46*	*54*	*14.8*
21	3791	13	87	85.1

STT	Mã ảnh truy vấn	Số mẫu dương	Số mẫu âm	Tỉ lệ chênh lệch (%)
22	**8956**	**5**	**95**	**94.7**
23	6308	14	86	83.7
24	*5925*	*59*	*41*	*30.5*
25	**9884**	**9**	**91**	**90.1**
26	3080	28	72	61.1
27	**8159**	**3**	**97**	**96.9**
28	8120	23	77	70.1
29	4099	31	69	55.1
30	**6117**	**8**	**92**	**91.3**

Nhìn vào số liệu trong Bảng 3.1 ta thấy một số ảnh truy vấn được in đậm có tỉ lệ chênh lệch của hai lớp cao (> 90%) với mã (số thứ tự ảnh trong tập dữ liệu) là 5464, 9619, 1497, 9072,10024, 2121, 6643, 8956, 9884, 8159, 6117. Và một số ảnh truy vấn có mã 7548, 10356, 3775, 5103, 5925 là có tỉ lệ chênh lệch của hai lớp thấp (< 31%). Kết quả độ chính xác P@100 của 30 ảnh truy vấn đó sau khi áp dụng SVM hai lớp trên tập ảnh phản hồi có nhãn được chỉ ra trong Bảng 3.2 ở dưới:

Bảng 3.2. Độ chính xác tra cứu của 30 truy vấn sau phản hồi SVM.

STT	Mã ảnh truy vấn	Tỉ lệ chênh lệch (%)	Độ chính xác sau phản hồi SVM (%)
1	8114	79.5	28
2	2755	63	39
3	**5464**	**94.7**	**6**
4	*7548*	*14.8*	*68*

STT	Mã ảnh truy vấn	Tỉ lệ chênh lệch (%)	Độ chính xác sau phản hồi SVM (%)
5	**9619**	**95.8**	**5**
6	*10356*	*27.6*	*77*
7	5907	79.5	27
8	**1497**	**91.3**	**9**
9	1612	71.8	92
10	2779	83.7	19
11	**9072**	**99**	**1**
12	2744	70.1	42
13	8785	82.4	19
14	2627	66.7	48
15	**10024**	**96.9**	**5**
16	*3775*	*11.3*	*50*
17	**2121**	**95.8**	**4**
18	2708	81	22
19	**6643**	**96.9**	**8**
20	*5103*	*14.8*	*85*
21	3791	85.1	14
22	**8956**	**94.7**	**7**

STT	Mã ảnh truy vấn	Tỉ lệ chênh lệch (%)	Độ chính xác sau phản hồi SVM (%)
23	6308	83.7	28
24	*5925*	*30.5*	*86*
25	**9884**	**90.1**	**11**
26	3080	61.1	47
27	**8159**	**96.9**	**4**
28	8120	70.1	36
29	4099	55.1	43
30	**6117**	**91.3**	**17**

Nhìn vào Bảng 3.2, chúng ta thấy rằng những ảnh truy vấn mà tỉ lệ chênh lệch giữa hai lớp cao sẽ cho một độ chính xác tra cứu thấp (trung bình 7%), còn những ảnh truy vấn với tỉ lệ chênh lệch giữa hai lớp thấp có độ chính tra cứu cao (trung bình 73.2%). Minh chứng thực nghiệm này một lần nữa cho ta thấy rằng "Sự cân bằng giữa hai lớp trong quá trình phản hồi là rất quan trọng giúp tăng độ chính xác tra cứu trong CBIR".

Như các vấn đề đã trình bày thì ta thấy độ chính xác về độ chính xác của các hệ thống tra cứu ảnh sử dụng phản hồi liên quan dựa vào SVM chưa được như kỳ vọng là do:

- Thứ nhất, các phương pháp thường bỏ qua các mẫu chưa có nhãn và vấn đề cân bằng mẫu. Trong tra cứu ảnh sử dụng phản hồi liên quan, người dùng chỉ gán nhãn một số ít ảnh và không đảm bảo gán nhãn mỗi mẫu phản hồi chính xác cho tất cả các lần dẫn đến thu thập tập phản hồi có nhãn lớn là rất khó. Mặt khác, trong RF số lượng mẫu ở hai nhóm dương âm thường mất cân bằng do số các mẫu có nhãn âm thường nhiều hơn nhiều số mẫu có nhãn dương. Chính vì thế phải bổ sung thêm mẫu dương cho tập phản hồi sử dụng thông tin mẫu chưa có nhãn

- Thứ hai, cấu trúc lân cận cục bộ của các ảnh có thể được khám phá qua các mẫu chưa có nhãn. Các thống kê toàn cục như phương sai thường khó ước lượng khi số mẫu không đủ. Do chỉ dựa vào số các mẫu do người dùng phản hồi và các cấu trúc Euclidean toàn cục làm cho bộ phân lớp của SVM không ổn định, trong khi cấu trúc đa tạp cục bộ của các đặc trưng trực quan mức thấp bị bỏ qua.

- Cuối cùng, chỉ khai thác được một số ít các khía cạnh của đối tượng trong khi một đối tượng có thể bao gồm nhiều khía cạnh hữu ích khác nhau, sử dụng một bộ phân lớp không thể biểu diễn hết các khía cạnh hữu ích khác nhau của một đối tượng

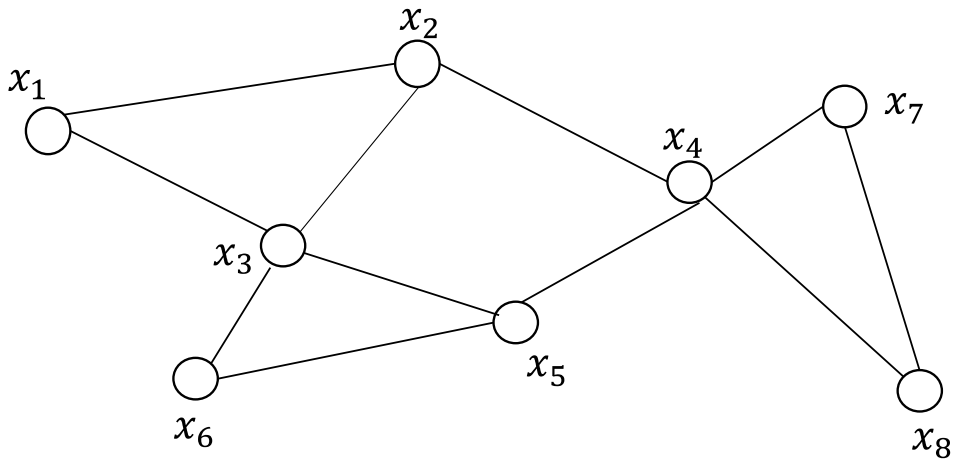
Trong luận án này, đề xuất một phương pháp tra cứu ảnh kết hợp chiếu phân biệt lớp ngữ nghĩa đa khía cạnh (Combine semantic class discriminant multiple aspect projection for image retrieval – **CIR**) sử dụng SVM cho quá trình phản hồi với một số quan sát: 1) một mẫu nằm trong một không gian con mà có mật độ tập trung những điểm mang nhãn dương nhiều thì khả năng cao điểm đó cũng sẽ mang nhãn dương và áp dụng phân lớp theo phân hoạch đồ thị với tiêu chuẩn cân bằng Ncut sẽ làm cho khẳng định càng vững chắc; 2) một tập huấn luyện có thể có nhiều khía cạnh hữu ích khác nhau, chúng ta cần có nhiều bộ phân lớp để thể hiện được các khía cạnh hữu ích khác nhau đó; và 3) các cấu trúc Euclidean toàn cục chưa phản ánh đầy đủ sự đa tạp của dữ liệu, cần xét cấu trúc đa tạp cục bộ của các đặc trưng trực quan mức thấp của ảnh. Với việc kết hợp ba quan sát này với mô hình SVM, phương pháp CIR đề xuất sẽ cải tiến được độ chính xác nâng cao độ chính xác cho tra cứu ảnh sử dụng SVM cho quá trình phản hồi. Yếu tố bán giám sát của phương pháp CIR thể hiện thông qua việc sử dụng thông tin của một số ảnh đã được gán nhãn dương hoặc âm và cả một số ảnh chưa có thông tin nhãn vào quá trình học để cải thiện độ chính xác tra cứu.

Phương pháp tra cứu ảnh kết hợp chiếu phân biệt lớp ngữ nghĩa đa khía cạnh [CT4] đề xuất thực hiện (a) bổ sung một số mẫu dương nhằm xây dựng tập mẫu cân bằng dựa vào đồ thị (BSFG - balanced sample feedback based on the graph) thông qua xác định nhãn của một số ảnh chưa gán nhãn; (b) tận dụng thông tin hình học trong việc giảm chiều hiệu quả (SCDP) (đã trình bày trong chương 2); (c) tận dụng các khía cạnh của đối tượng để xây dựng bộ phân lớp mạnh (CMAC)

3.2. Kỹ thuật cân bằng tập mẫu phản hồi sử dụng học bán giám sát đồ thị

Như đã trình bày, một trong những khó khăn gặp phải trong RF chính là xử lý vấn đề mất cân bằng giữa hai nhóm được gán nhãn dương và âm của tập phản hồi. Một tập phản hồi được cho là mất cân bằng nếu một lớp (trong RF chính là lớp mẫu nhãn âm) vượt trội về số lượng so với lớp còn lại (mẫu nhãn dương). Do đó, trong phần này trình bày một phương pháp cân bằng tập ảnh phản hồi dựa vào đồ thị cho tra cứu ảnh thông qua việc tăng số lượng mẫu dương cho tập huấn luyện.

Cho $\boldsymbol{G} = (\boldsymbol{X}, \boldsymbol{S})$ là một đồ thị vô hướng (theo định nghĩa 1.1) với tập đỉnh $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots \boldsymbol{x_N}\} \in \boldsymbol{R^n}$ thể hiện lân cận (láng giềng) gần nhất của $\boldsymbol{N}$ đỉnh (ảnh) thu được từ kết quả tra cứu với một truy vấn.

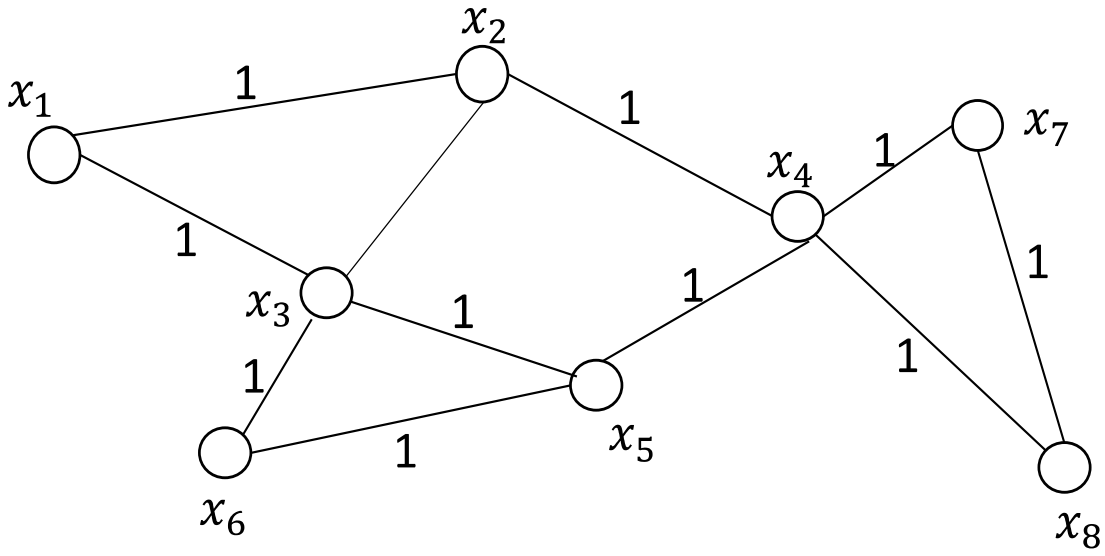


Hình 3.1. Đồ thị lân cận gần nhất G.

Với đồ thị **G** được đánh trọng số (định nghĩa 1.7) thì sẽ có một trọng số không âm $\boldsymbol{s_{ij}} \geq \boldsymbol{0}$ trên mỗi cạnh nối hai đỉnh x_i và x_j nối với nhau (tức là $\boldsymbol{x_i}$ và $\boldsymbol{x_j}$ là lân cận với nhau theo định nghĩa 1.2). Ma trận kề $\boldsymbol{S} = (\boldsymbol{s_{ij}})_{\boldsymbol{i,j=1,\dots N}}$ có trọng số của đồ thị vô hướng G theo định nghĩa 1.6 là ma trận sau:

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\cdots$	$\boldsymbol{x_N}$
$\boldsymbol{x_1}$	s_{11}	s_{12}	s_{13}	$\cdots$	s_{1N}
$\boldsymbol{x_2}$	s_{21}	s_{22}	s_{23}	$\cdots$	s_{2N}
$\boldsymbol{x_3}$	s_{31}	s_{32}	s_{33}	$\cdots$	s_{3N}
					$\vdots$
$\boldsymbol{x_N}$	s_{N1}	s_{N2}	s_{N3}		s_{NN}

Gọi $\boldsymbol{k - NN(x_i)}$ gồm k láng giềng (lân cận) gần nhất của điểm $\boldsymbol{x_i}$. Ta gán $\boldsymbol{s_{ij} = 1}$, nếu $\boldsymbol{x_i}$ hoặc $\boldsymbol{x_j}$ là lân cận của nhau (tức là $\boldsymbol{x_i} \in \boldsymbol{k - NN(x_j)}$ hoặc $\boldsymbol{x_j} \in \boldsymbol{k - NN(x_i)}$), trong trường hợp khác thì gán $\boldsymbol{s_{ij} = 0}$. Do chỉ xét với đồ thị vô hướng nên $\boldsymbol{s_{ij} = s_{ji}}$.

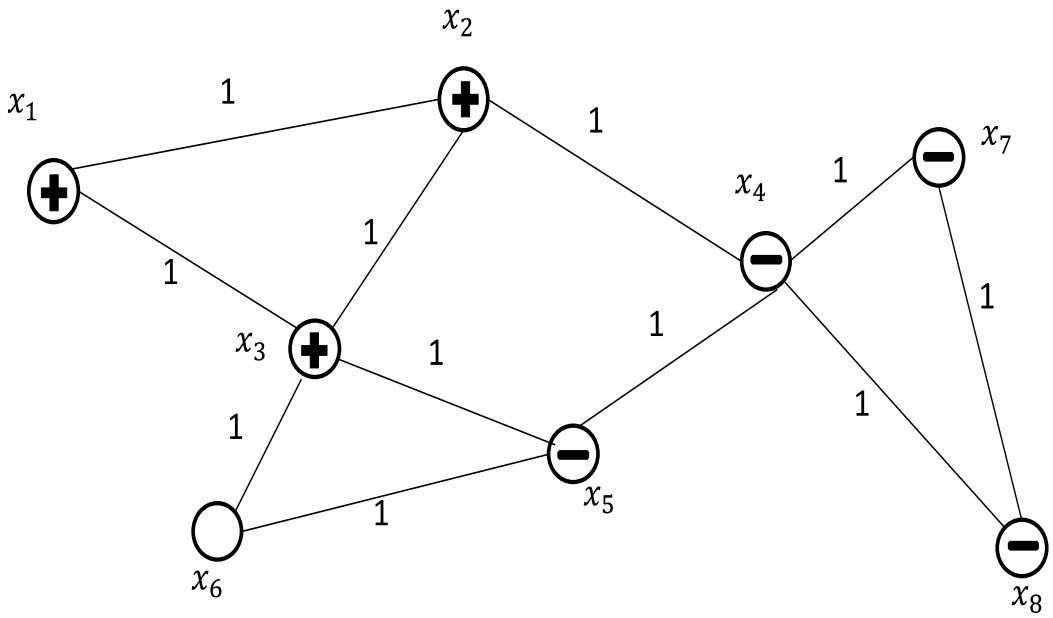


Hình 3.2. Đồ thị G với trọng số trên k-NN.

Ma trận **S** là ma trận kề tương ứng với đồ thị **G** trên Hình 3.3:

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$	$\boldsymbol{x_6}$	$\boldsymbol{x_7}$	$\boldsymbol{x_8}$
$\boldsymbol{x_1}$	0	1	1	0	0	0	0	0
$\boldsymbol{x_2}$	1	0	1	1	0	0	0	0
$\boldsymbol{x_3}$	1	1	0	0	1	1	0	0
$\boldsymbol{x_4}$	0	1	0	0	1	0	1	1
$\boldsymbol{x_5}$	0	0	1	1	0	1	0	0
$\boldsymbol{x_6}$	0	0	1	0	1	0	0	0
$\boldsymbol{x_7}$	0	0	0	1	0	0	0	1
$\boldsymbol{x_8}$	0	0	0	1	0	0	1	0

Trên tập kết quả tra cứu **X** gồm **N** ảnh trả về, người dùng gán nhãn $\boldsymbol{m}$ ($\boldsymbol{m < N}$) điểm cho ta tập $\boldsymbol{L = \{x_1, x_2, \dots x_m\} \in R^n}$, còn lại $\boldsymbol{N - m}$ ảnh là chưa có nhãn gọi là tập $\boldsymbol{UL = \{x_{m+1}, x_{m+2}, \dots, x_N\} \in R^n}$. Để xác định nơi mà những ảnh thuộc lớp dương có mật độ cao, ta xây dựng đồ thị $\boldsymbol{G^{label}}$ trên cơ sở của đồ thị $\boldsymbol{G}$ cùng với nhãn đã được lựa chọn từ người dùng.



Hình 3.3. Đồ thị $\boldsymbol{G^{label}}$. Các nút được gán nhãn (+) hoặc (-) hoặc chưa nhãn.

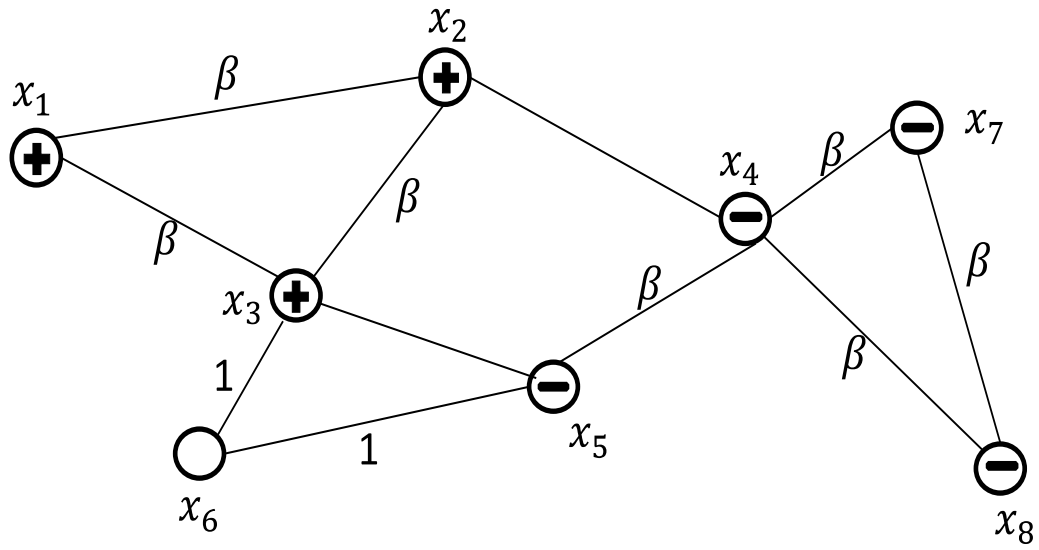
Dựa vào đồ thị $\boldsymbol{G^{label}}$ xây dựng ma trận có trọng số S^{label}. Với $\boldsymbol{label(x_i)}$ cho biết điểm x_i mang nhãn dương (là liên quan với ảnh truy vấn) hoặc là nhãn âm (là không liên quan với ảnh truy vấn). Với mỗi điểm x_i, xác định tập $\boldsymbol{k - NN^{label}(x_i)}$ theo:

$$k - NN^{label}(x_i) = \{x | \ label(x) = label(x_i) \ hoặc \ x \in UL\} \quad (3.1)$$

Trong (3.1) thể hiện với mỗi điểm x_i thì xung quanh nó có những điểm nào có liên quan với bản thân nó (nhãn giống với x_i) hoặc là chưa được gán nhãn. Chúng ta xác định ma trận trọng số $\boldsymbol{S^{label}}$ (với $i = j$ thì ${s_{ij}}^{label} = 0$) của đồ thị $\boldsymbol{G^{label}}$ như sau:

$${s_{ij}}^{label} = \begin{cases} \beta \ , \ nếu \ label(x_i) = label(x_j) \\ 1, \ nếu \ x_i \in UX, \ x_i \in k - NN^{label}(x_j) \\ hoặc \ x_j \in UX, \ x_j \in k - NN^{label}(x_i) \\ 0, \ ngược \ lại \end{cases} \quad (3.2)$$

Trong (3.2), giá trị β có ý nghĩa hai ảnh là cùng nhãn (tức cùng ngữ nghĩa).



Hình 3.4. Đồ thị G^{label} sau khi cập nhật trọng số.

Ma trận trọng số $\boldsymbol{S^{label}}$ của đồ thị sau khi cập nhật trọng số:

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$	$\boldsymbol{x_6}$	$\boldsymbol{x_7}$	$\boldsymbol{x_8}$
$\boldsymbol{x_1}$	0	β	β	0	0	0	0	0
$\boldsymbol{x_2}$	β	0	β	0	0	0	0	0
$\boldsymbol{x_3}$	β	β	0	0	0	1	0	0
$\boldsymbol{x_4}$	0	0	0	0	β	0	β	β
$\boldsymbol{x_5}$	0	0	0	β	0	1	0	0
$\boldsymbol{x_6}$	0	0	1	0	1	0	0	0
$\boldsymbol{x_7}$	0	0	0	β	0	0	0	β
$\boldsymbol{x_8}$	0	0	0	β	0	0	β	0

Trên đồ thị $\boldsymbol{G^{label}}$, trong định nghĩa 1.7 bậc mỗi đỉnh $\boldsymbol{x_i} \in \boldsymbol{X}$ được xác định bởi:

$$d_i^{label} = \sum_{j=1}^{N} s_{ij}^{label} \tag{3.3}$$

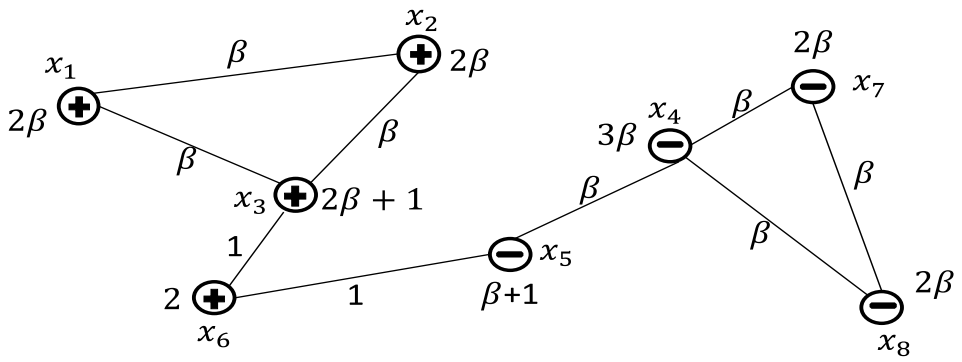
Ma trận bậc $\boldsymbol{D^{label}}$ theo định nghĩa 1.4 là:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	2β	β	β	0	0	0	0	0
x_2	β	2β	β	0	0	0	0	0
x_3	β	β	$2\beta+1$	0	0	1	0	0
x_4	0	0	0	3β	β	0	β	β
x_5	0	0	0	β	$\beta+1$	1	0	0
x_6	0	0	1	0	1	2	0	0
x_7	0	0	0	β	0	0	2β	β
x_8	0	0	0	β	0	0	β	2β

Với mỗi điểm $\boldsymbol{x_i}$ chưa có nhãn thuộc tập $\boldsymbol{UL}$, lấy nhãn của điểm $\boldsymbol{x}^*$ có bậc $\boldsymbol{d_i^{label}}$ lớn nhất trong số những điểm thuộc lân cận $\boldsymbol{k-NN^{label}(x_i)}$ làm nhãn tạm thời cho $\boldsymbol{x_i}$. Cụ thể hơn nhãn $\boldsymbol{x_i}$ tạm thời mang theo nhãn của của $\boldsymbol{x}^*$, và $\boldsymbol{x}^*$ được tìm bởi công thức:

$$x^* = \underset{x_j \in kNN^{label}(x_i)}{\operatorname{argmax}} (d_j{}^{label}) \tag{3.4}$$

Theo định nghĩa 1.4, trong ma trận bậc $\boldsymbol{D^{label}}$ cho biết mật độ xung quanh của mỗi điểm dữ liệu hiện tại. Nếu xung quanh $\boldsymbol{x_i}$ chưa có nhãn mà có nhiều điểm có nhãn thì bậc $\boldsymbol{d_i^{label}}$ sẽ lớn, nên nhãn tạm thời của nó sẽ theo số đông của những điểm mang nhãn dương hoặc âm. Hình 3.5 minh họa việc xác định nhãn tạm thời của điểm x_6 trong đồ thị G là (+) (bởi vì $d_3{}^{label}$ lớn hơn $d_5{}^{label}$):



Hình 3.5. Minh họa xác định nhãn tạm thời

Ý tưởng để xác định nhãn cuối cùng của một điểm $\boldsymbol{x_i}$ như sau. Đầu tiên, phân hoạch đồ thị thành hai lớp: lớp âm và lớp dương. Sau đó, kiểm tra xem điểm $\boldsymbol{x_i}$ thuộc lớp nào, nếu $\boldsymbol{x_i}$ thuộc lớp dương thì nhãn cuối cùng của điểm $\boldsymbol{x_i}$ là dương, bỏ qua $\boldsymbol{x_i}$ trong trường hợp ngược lại.

Tiêu chuẩn phân hoạch đồ thị được sử dụng ở đây là Ncut [75]. Lý do sử dụng Ncut là bởi vì nó sử dụng hai nguyên lý cơ bản cho phân hoạch. Nguyên lý thứ nhất là cực tiểu số các kết nối liên lớp và nguyên lý thứ hai là cực đại số các kết nối trong lớp. Do đó, nó sinh ra các phân hoạch cân bằng hơn. Bên cạnh đó, chúng ta cũng đã có phương pháp tìm nghiệm xấp xỉ với thời gian chấp nhận được. Dưới đây là mô tả ngắn gọn tiêu chuẩn Ncut và cách tìm giá trị Ncut cực tiểu.

Trên đồ thị $\boldsymbol{G^{label}}$, để xem xét liên lớp, chúng ta xác định một cut là một tập các cạnh với chỉ một đỉnh trong một lớp (dương hoặc âm).

$$cut(P,N) = \sum_{i\in P, j\in N} {s_{ij}}^{label} \tag{3.5}$$

, $\boldsymbol{P} \in \boldsymbol{G^{label}}$ là biểu thị lớp dương và $\boldsymbol{N} \in \boldsymbol{G^{label}}$ là biểu thị lớp âm.

Để xem xét trong phạm vi nhóm, chúng ta xác định một $\boldsymbol{vol(P)}$ là tổng trọng số các cạnh với ít nhất một điểm cuối trong P.

$$vol(P) = \sum_{i\in P | j\in P} {s_{ij}}^{label} \tag{3.6}$$

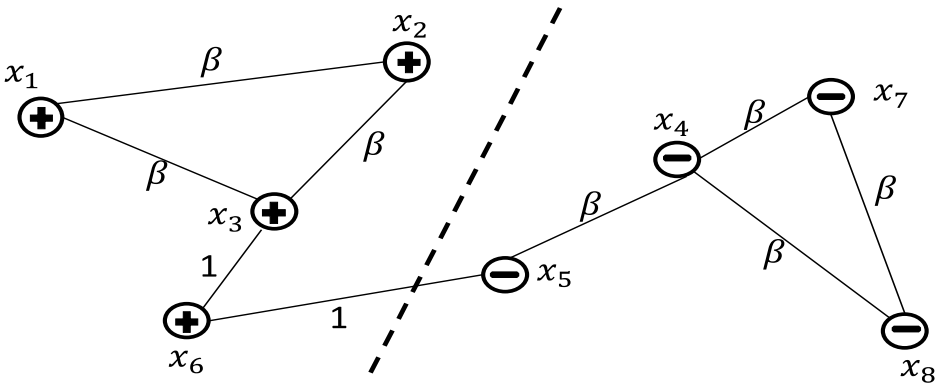
Tiêu chuẩn Ncut được xác định như sau:

$$NCut(P,N) = \frac{cut(P,N)}{vol(P)} + \frac{cut(P,N)}{vol(N)} \tag{3.7}$$

Bài toán tìm giá trị Ncut cực tiểu cho phân hoạch đồ thị là một bài toán NP-complete. May thay, phương pháp được đề xuất bởi Shi và Malik [75] tìm được nghiệm xấp xỉ bằng việc giải bài toán tìm giá trị riêng tổng quát có thể giúp chúng ta tính được giá trị Ncut cực tiểu.

$$(D^{label} - S^{label})\vec{y} = \lambda D^{label}\vec{y} \tag{3.8}$$

Việc xác định nhãn cuối cùng được minh họa trên Hình 3.6. Trên Hình 3.6, x_6 sẽ thuộc về lớp dương vì tiêu chí Ncut tạo ra một phân vùng cân bằng, vì vậy x chính thức được gắn nhãn là (+)



Hình 3.6. Đồ thị G^{label} được phân chia theo tiêu chí Ncut.

Thuật toán cân bằng tập mẫu phản hồi dựa vào đồ thị được mô tả dưới Thuật toán 3.1 [CT4].

Thuật toán 3.1. Thuật toán cân bằng tập mẫu phản hồi dựa vào đồ thị (BSFG)

Input: $X = \{x_1, x_2, \dots x_N\} \in R^n$: gồm N ảnh

$L = \{x_1, x_2, \dots x_m\} \in R^n$: gồm m ảnh đã được gán nhãn

Ouput: TS: Tập mẫu phản hồi cân bằng

Bước 1: $SP \leftarrow \{x |\ x \in L \ \wedge label(x) = 1\}$

$SN \leftarrow \{x |\ x \in L \ \wedge label(x) = -1\}$

Bước 2: G ← Graph(X);

Bước 3: Repeat

Bước 3.1: $G^{label} \leftarrow Graph\_Label(G)$;

$S^{label} \leftarrow Matrix\_Label(G^{label})$;

Bước 3.2: $D^{label} \leftarrow Degree\_Label(S^{label})$;

Bước 3.3: foreach x_i in UL:

$$x^* \leftarrow \underset{x_j \in k-NN^{label}(x_i)}{\operatorname{argmax}} (d_j{}^{label});$$

$if\ (label(x^*) == 1)\ \{$

$label(x_i) = 1;$

$SP \leftarrow SP \cup \{x_i\};$

$\}$

until $|SP|$ xấp xỉ $|SN|$

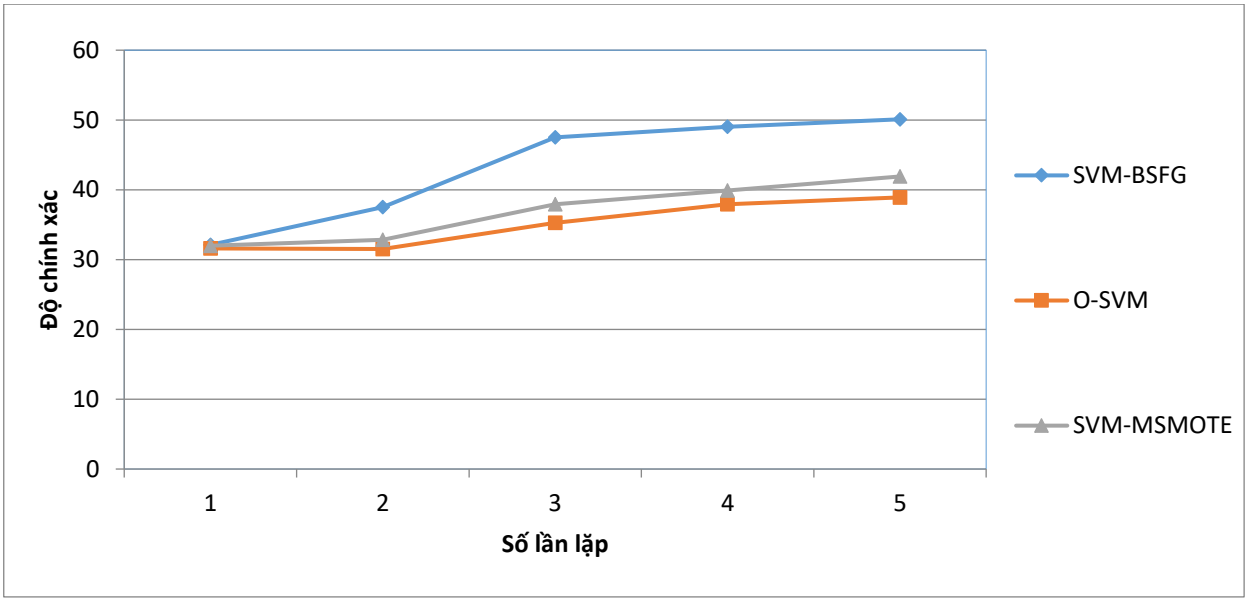
Bước 4: $TS \leftarrow SP \cup SN$

Đầu vào của Thuật toán BSFG là một tập gồm N ảnh kết quả tra cứu trả về của truy vấn ngay trước với $\boldsymbol{m}$ điểm có nhãn thuộc tập $\boldsymbol{L}$ và $\boldsymbol{N-m}$ điểm chưa có nhãn. Thuật toán cho đầu ra là một tập mẫu phản hồi kết quả $\boldsymbol{TS}$ cân bằng về số lượng mẫu ở hai nhóm dương và âm sau khi bổ sung mẫu chưa có nhãn vào tập nhãn dương. Bước đầu thuật toán tách tập $\boldsymbol{L}$ gồm m ảnh đã được gán nhãn thành hai tập con $\boldsymbol{SP}$ chứa các ảnh mang nhãn dương và tập $\boldsymbol{SN}$ gồm các ảnh mang nhãn âm. Tiếp theo, hàm **Graph**() thực hiện xây dựng đồ thị vô hướng G từ *N* ảnh trên cùng trong kết quả tra cứu trả về của lần truy vấn trước đó bao gồm cả tập $\boldsymbol{L}$ và tập $\boldsymbol{UL}$ gồm $\boldsymbol{N-m}$ ảnh chưa được gán nhãn. Sau đó tại Bước 3.1, xây dựng đồ thị $\boldsymbol{G^{label}}$ và ma trận $\boldsymbol{S^{label}}$ có trọng số theo (3.9) tương ứng thông qua hàm ***Graph_Label()*** và ***Maxtrix_Label()***; từ đó xác định ma trận bậc $\boldsymbol{D^{label}}$ (Bước 3.2). Tại bước 3.3, duyệt mỗi điểm chưa có nhãn ta xác định được điểm $\boldsymbol{x^*}$ đã có nhãn mà bậc lớn nhất trong số những điểm thuộc lân cận. Lúc này $\boldsymbol{x^*}$ nếu mang nhãn dương thì x_i cũng sẽ được gán nhãn dương và bổ sung $\boldsymbol{x_i}$ vào tập ảnh mang nhãn dương SP. Bước 3 lặp đi lặp lại cho đến khi số lượng mẫu ở hai tập $\boldsymbol{SP}$ và $\boldsymbol{SN}$ là xấp xỉ cân bằng nhau.

Độ chính xác tra cứu ảnh sử dụng (kỹ thuật cân bằng mẫu phản hồi) BSFG

Hình 3.7 thể hiện các kết quả về độ chính xác tra cứu khi áp dụng phương pháp phản hồi trên tập ảnh phản hồi của người dùng trước và sau khi bổ sung mẫu dương để cân bằng tập ảnh phản hồi gồm hai lớp dương và âm như SVM gốc (tên gọi là O-SVM, tập mẫu chưa bổ sung mẫu dương để cân bằng mẫu) [43] được mô tả tại mục 1.3.2, phương pháp cân bằng mẫu thông qua MSMOTE (Modified Synthetic Minority Oversampling Technique) [76], mà ảnh thuộc lớp dương được bổ sung thông qua từng cặp hai láng giềng gần nhất mang nhãn dương trong tập ảnh phản hồi (tên gọi là SVM-MSMOTE) và phương pháp cân bằng mẫu dựa vào đồ thị luận án đề xuất (tên gọi là SVM-BSFG).

Nhìn vào kết quả thực nghiệm này ta thấy rằng độ chính xác của phương pháp SVM gốc kém hơn phương pháp SVM-MSMOTE do sự mất cân bằng giữa hai lớp của SVM gốc. Tuy nhiên, với MSMOTE bổ sung các mẫu dương tính giả do đó không gian dữ liệu có thể bị sai ngữ cảnh nên cho hiệu quả không cao bằng so với SVM-BSFG.



Hình 3.7. Độ chính xác của ba phương pháp O-SVM, SVM-MSMOTE,và SVM-BSFG.

3.3. Kỹ thuật kết hợp các bộ phân lớp theo khía cạnh

Như được đề cập ở trên, vấn đề cân bằng mẫu đã giải quyết nhưng nó chưa khám phá được thuộc tính thống kê cho phân lớp dữ liệu. Với nhận định rằng, không có một bộ phân lớp nào có thể biểu diễn được tất cả các khía cạnh hữu ích của dữ liệu đầu vào. Như vậy, muốn biểu diễn các khía cạnh hữu ích khác nhau của dữ liệu đầu vào cần phải kết hợp bộ phân lớp khác nhau tạo nên một bộ phân lớp tốt. Với mỗi khía cạnh khác nhau của một mẫu đang xét, xây dựng một bộ phân lớp được huấn luyện độc lập sau đó có thể được kết hợp thành một bộ phân lớp mạnh. Do đó, quá trình này có thể biểu diễn được các khía cạnh hữu ích khác nhau của đối tượng và dẫn đến nâng cao độ chính xác của hệ thống tra cứu. Khía cạnh được xác định tùy theo từng mục tiêu cụ thể. Khía cạnh có thể được xác định tùy theo ứng dụng cụ thể. Trong luận án, một khía cạnh được xác định là một trong năm đặc trưng gồm đặc trưng mô men màu, đặc trưng lược đồ màu, đặc trưng tương quan màu, đặc trưng Gabor và đặc trưng wavelet.

Thuật toán 3.2 [CT4] dưới đây là thuật toán kết hợp các bộ phân lớp theo khía cạnh (Combine Multiple Aspect Classifiers - CMAC).

Thuật toán 3.2 Thuật toán kết hợp bộ phân lớp theo khía cạnh (CMAC)

Input: reduced_Aspect$_i$, i = 1,…, k **:** Các tập mẫu theo khía cạnh đã giảm chiều:

Output: β: Bộ phân lớp được kết hợp:

Bước 1: For i=1, ..., k

$C^i \leftarrow$ Aspect Classifiers (reduced_Aspect$_i$);

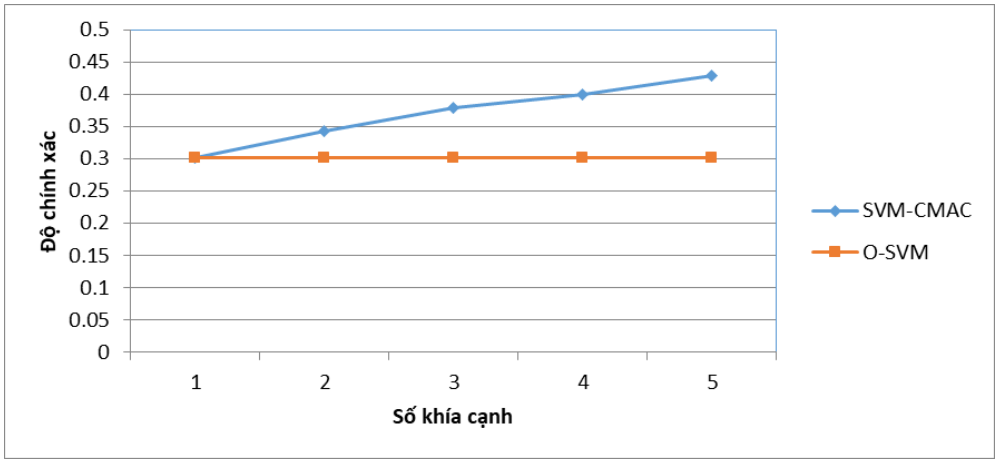
Bước 2:

$$\beta(x) = \underset{y\in\{-1,\ 1\}}{\operatorname{argmax}} \sum_b \delta_{sgn\left(C^i(x)\right),y}$$

Thuật toán CMAC nhận đầu vào là tập mẫu phản hồi đã được cân bằng và giảm chiều theo từng khía cạnh và cho đầu ra là một bộ phân lớp mạnh. Bước đầu tiên, thuật toán duyệt lần lượt từng tập khía cạnh đã giảm chiều để xây dựng bộ phân lớp C^i(reduced_Aspect$_i$) cho từng khía cạnh. Sau khi có được từng bộ phân lớp trên mỗi khía cạnh, thuật toán tiến hành kết hợp nhiều bộ phân lớp theo khía cạnh sử dụng kỹ thuật bầu cử đa số [77], chúng ta kỳ vọng thu được các kết quả tốt dựa trên niềm tin rằng số đông các chuyên gia có khả năng đưa ra quyết định đúng đắn hơn. Vậy nếu quyết định của m bộ phân lớp được kết hợp, và nhiều hơn một nửa chúng quyết định rằng quan sát x thuộc về lớp A thì toàn thể quyết định rằng $x \in A$.

Độ chính xác của CMAC

Luận án thực nghiệm đánh giá độ chính xác tra cứu của phương pháp tra cứu ảnh kết hợp 5 bộ phân lớp SVM theo CMAC trình bày tại 3.3 (SVM-CMAC) cho 5 khía cạnh tương ứng với 5 đặc trưng được trích rút gồm mô men màu, lược đồ màu, tương quan màu, đặc trưng Gabor và đặc trưng wavelet) so với SVM gốc (O-SVM).



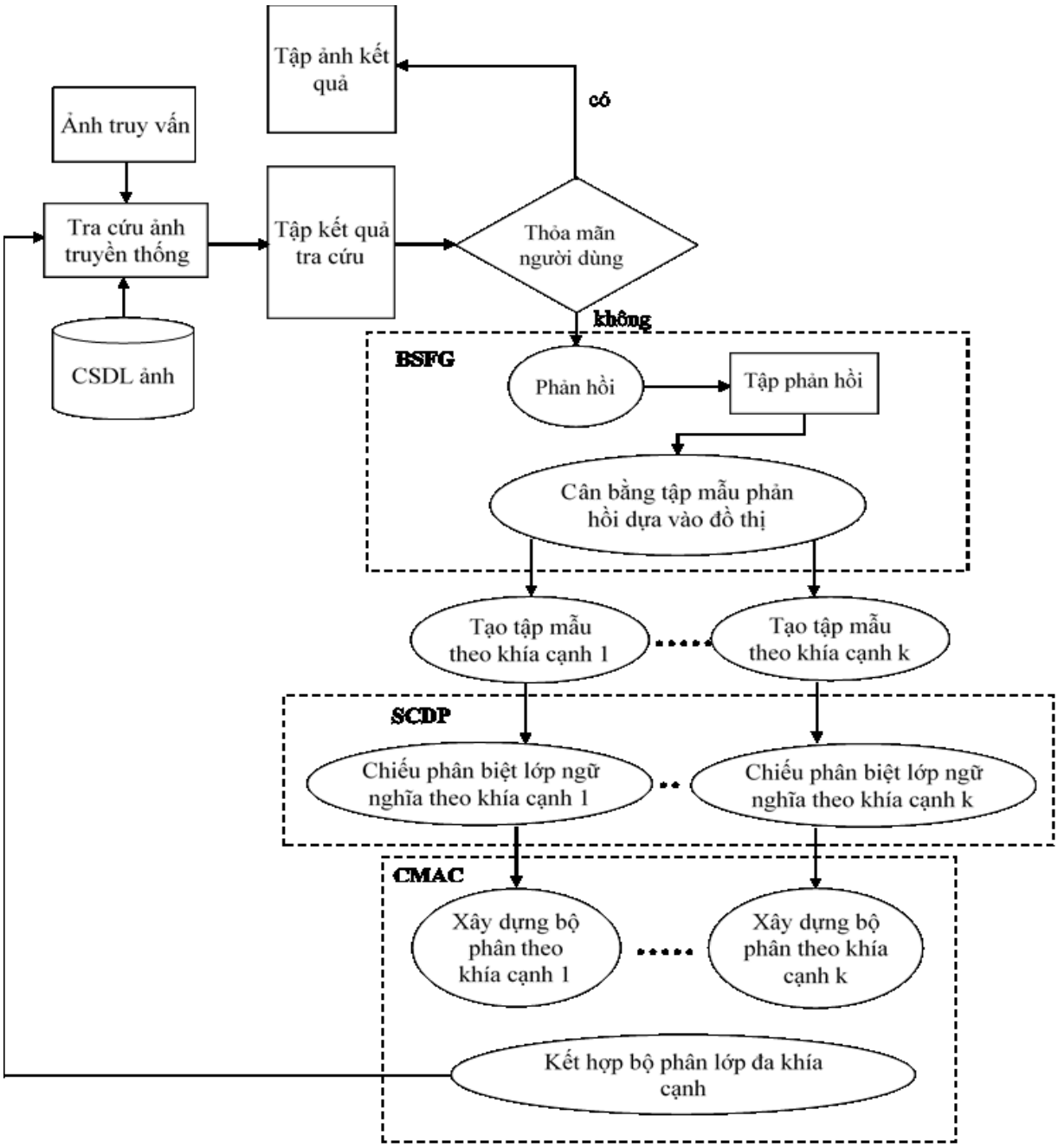
Hình 3.8. Độ chính xác của O-SVM và SVM-CMAC

Các kết quả trong Hình 3.8 chỉ ra độ chính xác của SVM-CMAC là cao hơn độ chính xác của O-SVM. Từ kết quả thực nghiệm cho thấy hệ thống có thể tra cứu

ảnh được nhiều ảnh liên quan với ảnh truy vấn hơn ở nhiều không gian con khác nhau (mỗi không gian con là một không gian đặc trưng theo từng khía cạnh) so với khi tra cứu trong một không gian chung của toàn bộ khía cạnh. Lý do của điều này là các bộ phân lớp chung đã khai thác được các khía cạnh khác nhau của đối tượng.

3.4. Phương pháp tra cứu ảnh kết hợp chiếu phân biệt lớp ngữ nghĩa đa khía cạnh.

Trong luận án đề xuất một phương pháp tra cứu ảnh được mô tả như Hình 3.9.



Hình 3.9. Sơ đồ phương pháp tra cứu ảnh kết hợp chiếu phân biệt lớp ngữ nghĩa đa khía cạnh

Trong sơ đồ Hình 3.9 giai đoạn đầu tiên, một ảnh truy vấn được cho bởi người dùng được trích rút tự động thu được véc tơ đặc trưng mức thấp. Sau đó, các ảnh được sắp xếp tăng dần theo một độ đo khoảng cách nào đó lấy về một tập ảnh hàng đầu gồm N ảnh để hiển thị cho người dùng. Nếu các kết quả trả về bởi hệ thống thỏa mãn với nhu cầu của người dùng, quá trình tra cứu kết thúc. Tuy nhiên, thực tế trong lần đầu như thế kết quả tra cứu khởi tạo này thường không đáp ứng tốt nhu cầu của người dùng, do đó quá trình phản hồi liên quan là tất yếu.

Với giai đoạn tiếp theo, người dùng gán nhãn cho m ảnh là mang nhãn dương (liên quan với truy vấn) hoặc nhãn âm (không liên quan với truy vấn) trên tập kết quả trước đó, những ảnh còn lại là chưa được gán nhãn. Với mỗi ảnh chưa được gán nhãn đó thực hiện xác định nhãn của chúng nếu dương bổ sung vào tập có nhãn dựa trên phương pháp **BSFG** để cân bằng mẫu. Lúc này tập phản hồi đã cân bằng, phương pháp tạo ra k tập mẫu theo khía cạnh (trong hệ thống, chọn ba khía cạnh là màu, hình dạng và kết cấu). Với mỗi tập mẫu theo khía cạnh, thực hiện phép chiếu phân biệt lớp ngữ nghĩa **SCDP** để giảm số chiều của đặc trưng để được tập mẫu theo mỗi khía cạnh với số chiều giảm. Tiếp theo tiến hành huấn luyện phản hồi dựa trên thuật toán học máy **SVM** trên tập mẫu theo mỗi khía cạnh với số chiều giảm này để được bộ phân lớp theo khía cạnh. Phương pháp **CMAC** (Combine Multiple Aspect Classifiers) kết hợp nhiều bộ phân lớp theo khía cạnh vừa tìm được thu được bộ phân lớp mạnh khám phá thuộc tính thống kê cho phân lớp. Sau đó, tất cả các ảnh được sắp xếp lại dựa trên độ đo khoảng cách với siêu phẳng phân tách của bộ phân lớp kết hợp để hiển thị kết quả mới cho người dùng. Quá trình này lặp lại cho đến khi tập kết quả thỏa mãn nhu cầu tra cứu của người dùng.

Thuật toán 3.3 [CT4, CT5] dưới đây trình bày thuật toán đề xuất tra cứu ảnh học bán giám sát dựa vào đồ thị.

Thuật toán 3.3. Thuật toán tra cứu ảnh kết hợp chiếu phân biệt lớp ngữ nghĩa đa khía cạnh (CIR**)**

Input:

S : Tập ảnh cơ sở dữ liệu

Q*:* Ảnh truy vấn

N: Số ảnh trả về tại mỗi lần lặp

k: Số lượng khía cạnh

Output:

R: Tập ảnh kết quả tra cứu

Bước 1: X ←**Retrieval**$_{Init}$(**Q, S**, *N*);

Bước 2: Repeat

Bước 2.1: LX ←***Feedback***(**X**) //Phản hồi liên quan

Bước 2.2: TS ← ***BSFG***(**S, X, LX**); //Cân bằng tập mẫu

Bước 2.3: For *i*=1, ..., *k*//Tách tập mẫu thành k tập mẫu theo khía cạnh

Aspect_i ←**Take_Aspect**(**TS**);

Bước 2.4: For *i*=1, ..., *k* //Giảm chiều tập mẫu theo khía cạnh

$\mathbf{U} \leftarrow \mathbf{SCDP}(\text{Aspect}_i)$;

$\text{reduced\_Aspect}_i \leftarrow \mathbf{reduced}(\text{Aspect}_i, \mathbf{U})$

Bước 2.5: $\beta \leftarrow$ **CMAC(**reduced_Aspect_1,...reduced_Aspect_k, β);//Kết hợp các bộ phân lớp con

Bước 2.6: R ← **Retrieval**(β, **S**, *N*); //Tra cứu theo bộ phân lớp đã kết hợp

until (Người dùng thỏa mãn);

Bước 3: Return **R**;

Thuật toán tra cứu ảnh đề xuất trên Thuật toán 3.3 thực hiện như sau:

Trong không gian đặc trưng gốc nhiều chiều thì mỗi một ảnh trong tập dữ liệu **S** được biểu diễn là một điểm. Khi một người dùng đưa một ảnh **Q** làm truy vấn khởi tạo, thuật toán sẽ biểu diễn ảnh **Q** thành một điểm trong không gian đặc trưng nhiều chiều như đã thực hiện với các ảnh trong tập **S**. Thực hiện tra cứu $\text{Retrieval}_{\text{init}}$(**Q, S,** *N*) với truy vấn khởi tạo **Q** (Bước 1), với *N* là số lượng ảnh trả về cho mỗi lần tra cứu thu về một tập ảnh kết quả gán cho **X**. Người dùng gán nhãn một số ảnh trong tập **X** trên một giao diện đồ họa thông qua hàm **Feedback***($Result_{init}(Q)$)* để được tập **LX** gồm m ảnh mang nhãn (dương và âm) (Bước 2.1). Hàm **BSFG**(**S, X, LX**) (Bước 2.2) xây dựng một tập phản hồi cân bằng **TS** về số lượng mẫu nhãn dương và âm bằng cách bổ sung thêm các mẫu dương từ tập ảnh chưa gán nhãn. Hàm **Take_Aspect**(TS) (Bước 2.3) sẽ thực hiện việc tách khía cạnh của tập **TS** để được tập khía cạnh Aspect_i. Với mỗi tập khía cạnh thứ *i*, thủ tục $\mathbf{SCDP}(\text{Aspect}_i)$ sẽ học cho một phép chiếu **U** để thực hiện giảm chiều trên tập Aspect_i thu được tập reduced_Aspect_i (Bước 2.4). Với mỗi tập khía cạnh reduced_Aspect_i (*i* = 1, ..., *k),* **CMAC** (reduced_Aspect_1, ...,

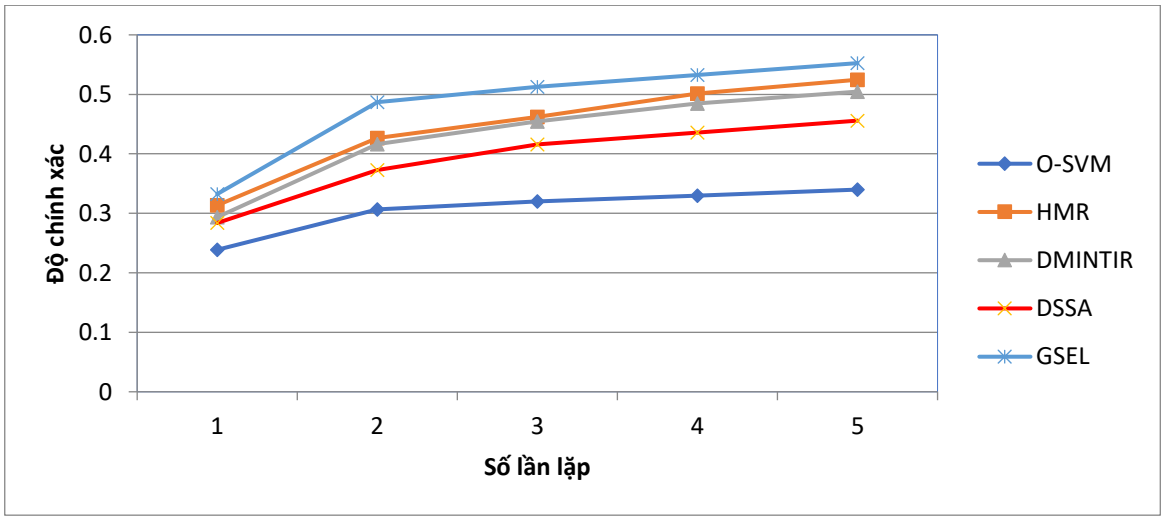
reduced_Aspect$_k$, β) (Bước 2.5) tạo ra một bộ phân lớp con $\mathbf{C}^i$(reduced_Aspect$_i$) và thực hiện kết hợp các bộ phân lớp đó để được một bộ phân lớp mạnh β. Dựa vào bộ phân lớp mạnh β vừa có, hàm **Retrieval**(β, **S**, N) (Bước 2.6) thực hiện tính toán khoảng cách lại và phân hạng các ảnh trong tập **S** cho ra tập **R** gồm N ảnh ở trên đỉnh. Quá trình trong bước 2 được thực hiện lại nhiều lần nếu người dùng chưa thấy phù hợp với nhu cầu tra cứu của mình.

3.5. Đánh giá độ chính xác của phương pháp tra cứu ảnh kết hợp

Độ chính xác của phương pháp đề xuất **CIR** được đánh giá thông qua độ chính xác tra cứu (công thức chỉ rõ trong mục 1.4.1) dựa trên tập ảnh COREL 10800 đã trình bày tại mục 1.4.2 và trên một tập ảnh thực tế tự sưu tầm gồm các ảnh danh lam, địa điểm ở thủ đô Hà Nội, Việt Nam.

Độ chính xác của CIR trên tập dữ liệu COREL 10800

Để đánh giá hiệu quả về độ chính xác của phương pháp đề xuất CIR, thực nghiệm so sánh nó với bốn phương pháp khác trên tập ảnh COREL 10800 bao gồm thuật toán phản hồi liên quan dựa vào SVM gốc O-SVM [43], phân tích không gian con ngữ nghĩa phân biệt (DSSA) [74], phân hạng lại ảnh tương tác đa khung nhìn phân biệt (discriminative multi-view interactive image re-ranking - DMINTIR) [60] và phân hạng đa tạp không đồng nhất (heterogeneous manifold ranking - HMR) [78]. Tất cả các thuật toán được đánh giá trên 5 lần lặp. Các đường cong độ chính xác được báo cáo trong Hình 3.10. Từ hình này, chúng ta thấy rằng CIR thực hiện tốt hơn bốn phương pháp khác, O-SVM, DSSA, DMINTIR và HMR.



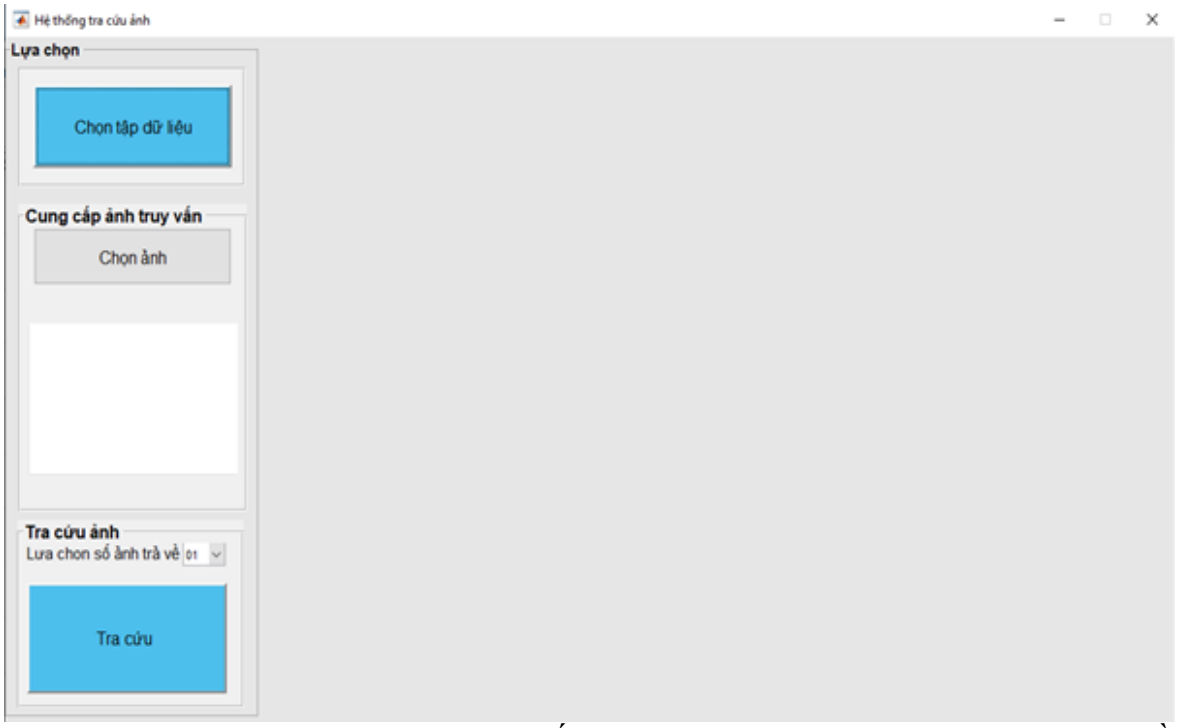
Hình 3.10. Độ chính xác của năm phương pháp.

Trong Hình 3.10 thấy rằng độ chính xác của DSSA cao hơn O-SVM, bởi vì nó có thể học một không gian con ngữ nghĩa từ các cặp ràng buộc tương tự và không tương tự mà không sử dụng thông tin nhãn lớp. Tuy nhiên độ chính xác của nó lại kém hơn DMINTIR, bởi vì nó không khai thác được các góc nhìn khác nhau của đối tượng. Độ chính xác của DMINTIR thấp hơn một chút độ chính xác của HMR bởi vì HMR khai thác được tính chất cục bộ của đa tạp dữ liệu. CIR tận dụng được các ưu điểm bao gồm học bán giám sát cho cân bằng mẫu, học đa tạp cho giảm chiều, khai thác các khía cạnh hữu ích khác nhau của đối tượng. Do đó, nó đưa ra kết quả cao nhất

Độ chính xác của CIR trên tập dữ liệu ảnh tự sưu tầm

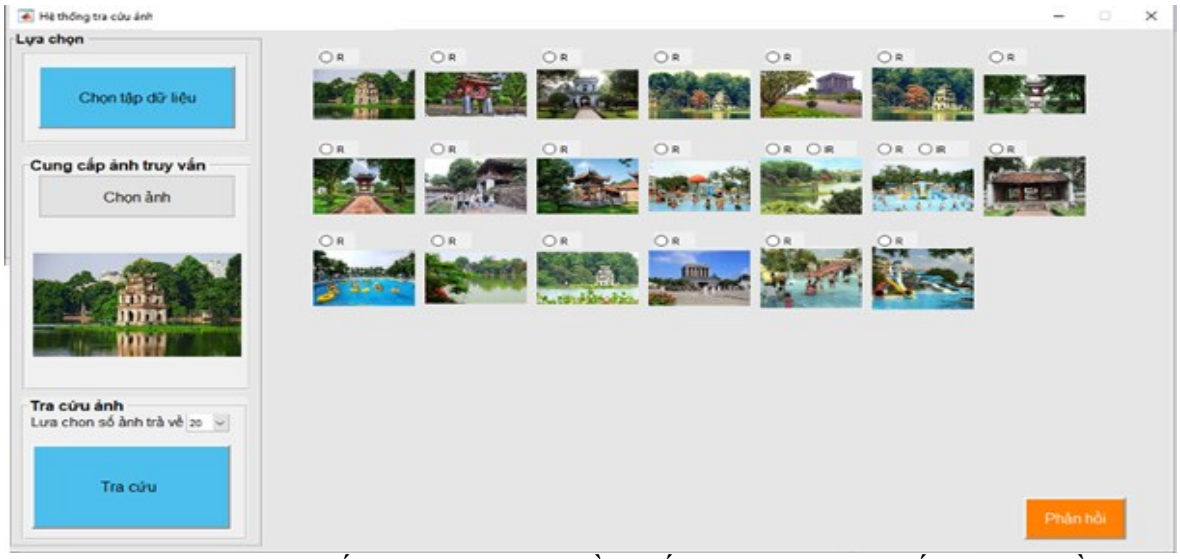
Luận án thu thập trên internet một tập dữ liệu ảnh gồm 100 ảnh phong cảnh chụp một số danh lam, địa điểm trong thủ đô Hà Nội. Tập dữ liệu ảnh này được chia đều cho 5 chủ đề bao gồm: Lăng Chủ tịch Hồ Chí Minh, Văn Miếu Quốc Tử Giám, Hồ Hoàn Kiếm, cầu Nhật Tân. Trong thực nghiệm, luận án trích rút véc tơ đặc trưng của mỗi bức ảnh gồm 5 đặc trưng (được mô tả tại mục 1.4.2 phần tập dữ liệu ảnh COREL) cho một véc tơ đặc trưng có độ dài 190 chiều.

Luận án cung cấp một giao diện có đồ họa cho người dùng thực tế có thể thực hiện tra cứu thông qua ảnh truy vấn mới đưa vào được minh họa bởi Hình 3.11.

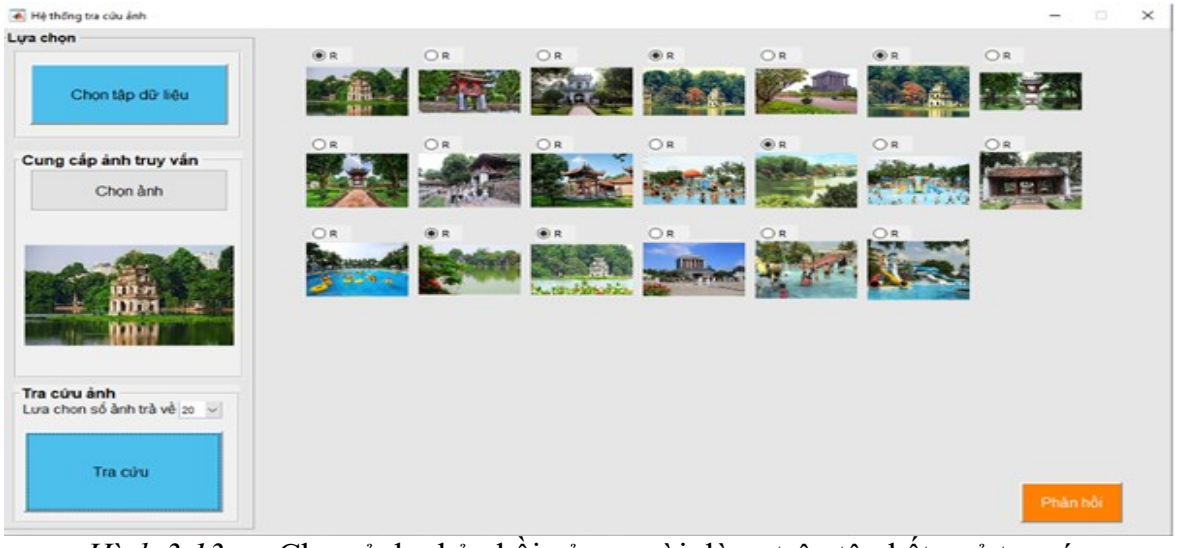


Hình 3.11. Giao diện trực quan hệ thống tra cứu ảnh học bán giám sát dựa vào đồ thị

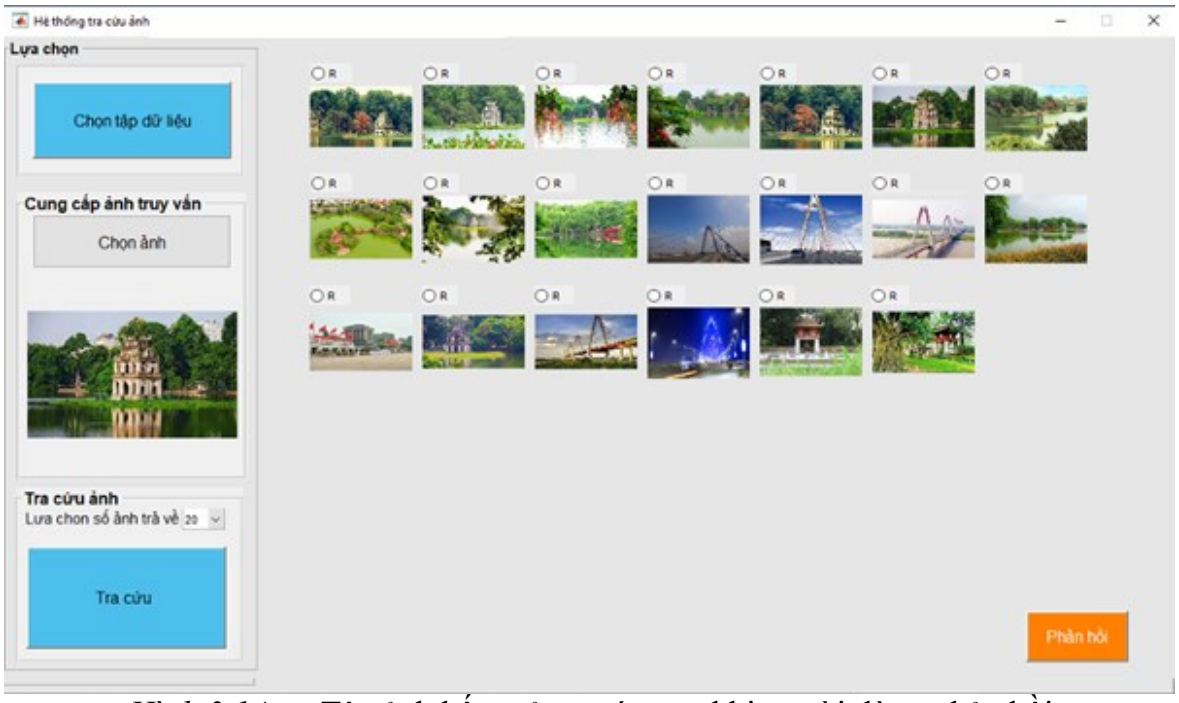
Với ảnh truy vấn đưa vào là một ảnh Hồ Hoàn Kiếm có tên "02.jpeg" trong tập dữ liệu thu thập được, tập ảnh kết quả tra cứu truyền thống (dùng độ đo khoảng cách Euclide) bao gồm 20 ảnh như trong Hình 3.12. Trong Hình 3.13, người dùng lựa chọn 06 ảnh là liên quan (chọn ô R) mang nhãn dương, còn lại 14 ảnh là không liên quan mang nhãn âm và hệ thống lấy thêm 10 ảnh ngay sau 20 ảnh trên cùng kết quả khởi tạo làm tập ảnh không có nhãn. Sau khi người dùng phản hồi thông tin, hệ thống áp dụng phương pháp tra cứu ảnh học bán giám sát dựa vào đồ thị thông qua tập ảnh huấn luyện gồm 30 ảnh trên. Kết quả tra cứu sau khi phản hồi được chỉ ra trong Hình 3.14 bao gồm 12 ảnh là liên quan (cùng chủ đề Hồ Hoàn Kiếm) với ảnh truy vấn. Chúng ta thấy rằng độ chính xác khi tra cứu truyền thống là 0.3 đã được nâng cao lên 0.6 sau khi tra cứu ảnh thông qua học bán giám sát dựa vào đồ thị.



Hình 3.12. Tập ảnh kết quả tra cứu truyền thống với ảnh truy vấn là ảnh Hồ Hoàn Kiếm



Hình 3.13. Chọn ảnh phản hồi của người dùng trên tập kết quả tra cứu



Hình 3.14. Tập ảnh kết quả tra cứu sau khi người dùng phản hồi

Trong thực nghiệm, luận án lựa chọn ngẫu nhiên 5 ảnh thuộc 5 chủ đề khác nhau để làm ảnh truy vấn. Để đánh giá độ chính xác tra cứu của phương pháp CIR có chạy tốt trên một tập dữ liệu thực tế tự sưu tầm hay không, luận án tiến hành thực nghiệm tra cứu với mỗi ảnh truy vấn đó bằng cách sử dụng phương pháp CIR. Đánh giá độ chính xác dựa trên tập ảnh kết quả tra cứu gồm 20 ảnh trên cùng có nội dung liên quan với ảnh truy vấn nhất. Các số liệu kết quả thu được như trong Bảng 3.3. Chúng ta thấy rằng kết quả chỉ ra trong Bảng 3.3 thể hiện việc độ chính xác tra cứu khi áp dụng học bán giám sát dựa vào đồ thị có cải thiện trong tập dữ liệu ảnh tự sưu tầm. Độ chính xác trung bình của 5 truy vấn ngẫu nhiên trên tăng từ 0.41 khi tra cứu ảnh truyền thống lên 0.7 sau khi áp dụng tra cứu ảnh thông qua CIR**.**

Bảng 3.3. Độ chính xác 5 ảnh truy vấn ngẫu nhiên trong tập ảnh sưu tầm

STT	**Tên ảnh**	**Chủ đề**	**Chọn phản hồi**		**Độ chính xác**	
			Số ảnh liên quan	**Số ảnh không liên quan**	**Baseline**	**CIR**
1	02.jpg	Hồ Hoàn Kiếm	06	14	0.3	0.6
2	18.jpg	Cầu Nhật Tân	04	16	0.2	0.55
3	10.jpg	Công viên nước Hồ Tây	07	13	0.35	0.65
4	16.jpg	Lăng Chủ Tịch	13	07	0.65	0.9

5	04.jpg	Văn Miếu Quốc Tử Giám	11	09	0.55	0.8
Độ chính xác trung bình					**0.41**	**0.7**

3.6. Kết luận chương 3

Tra cứu ảnh với phản hồi liên quan dựa vào SVM đã được sử dụng rộng rãi để giảm khoảng cách ngữ nghĩa và cải thiện độ chính xác của hệ thống tra cứu ảnh dựa vào nội dung. Tuy nhiên, với hướng tiếp cận này có ba hạn chế. Thứ nhất, dựa chính vào phản hồi của người dùng để xây dựng tập huấn luyện và nó thường bị vấn đề mất cân bằng dẫn đến bộ phân lớp SVM không ổn định. Thứ hai, bỏ qua cấu trúc phi tuyến của dữ liệu. Cuối cùng, không khai thác được các khía cạnh hữu ích khác nhau của đối tượng. Trong chương này, luận án đề xuất phương pháp CIR để nâng cao độ chính xác của hệ thống tra cứu sử dụng RF. Phương pháp có các ưu điểm sau: (1) tận dụng được thông tin của các mẫu chưa có nhãn; (2) khai thác được cấu trúc phi tuyến của dữ liệu đa tạp và (3) tận dụng được các khía cạnh hữu ích khác nhau của đối tượng. Các kết quả thực nghiệm trên tập dữ liệu ảnh ảnh Corel đã chỉ ra rằng phương pháp đề xuất đã cải tiến đáng kể độ chính xác tra cứu.

KẾT LUẬN

Độ chính xác của một hệ thống tra cứu ảnh dựa vào nội dung đã và đang được cộng đồng nghiên cứu quan tâm cải tiến. Nhiều phương pháp đã được đề xuất trong thời gian qua. Tuy nhiên, sự chênh lệch giữa đặc trưng mức thấp của ảnh và cảm nhận trực quan từ người dùng về nội dung ảnh làm cho độ chính xác của hệ thống tra cứu ảnh vẫn còn khoảng cách với nhu cầu của người dùng. Các đóng góp chính trong luận án này cũng theo định hướng sử dụng cơ chế phản hồi liên quan để thu hẹp sự chênh lệch khoảng cách này.

Luận án đã có các đóng góp sau:

(1) Đề xuất phương pháp tìm ma trận chiếu tối ưu theo tiếp cận học đa tạp [CT5]. Phương pháp này xem xét cấu trúc cục bộ của các mẫu dương và âm thuộc hai lân cận khác nhau để học một phép chiếu mà dữ liệu có thể phân biệt trên không gian chiếu, dẫn đến cải tiến độ chính xác cho tra cứu ảnh.

(2) Đề xuất phương pháp tự động bổ sung các mẫu dương vào tập huấn luyện để giải quyết vấn đề mất cân bằng tập huấn luyện [CT4]. Phương pháp này có thể: (a) bổ sung một số mẫu dương vào tập huấn luyện; (b) tận dụng các khía cạnh khác nhau của đối tượng để tạo ra một bộ phân lớp mạnh

Tra cứu ảnh dựa vào nội dung vẫn còn nhiều vấn đề cần tiếp tục nghiên cứu. Trong giới hạn của một luận án chưa thể giải quyết được hết mọi vấn đề, luận án chỉ giải quyết một phần trong các vấn đề tìm phép chiếu tối ưu mà khai thác được cấu trúc phi tuyến của dữ liệu, cân bằng tập mẫu phản hồi, khai thác một số khía cạnh hữu ích của đối tượng. Một số vấn đề cần được nghiên cứu tiếp trong tương lai:

- Nghiên cứu mạng nơ ron tích chập để nâng cao độ chính xác tra cứu trên tập ảnh lớn hơn.
- Nghiên cứu áp dụng cơ chế băm sâu để nâng cao tốc độ tra cứu.
- Từng bước tiến đến việc đưa hệ thống vào áp dụng một số lĩnh vực trong cuộc sống.

DANH MỤC CÔNG TRÌNH CỦA TÁC GIẢ

Trong nước:

[CT1] **Cù Việt Dũng**, Nguyễn Hữu Quỳnh, An Hồng Sơn, Đào Thị Thúy Quỳnh, Cải tiến tra cứu ảnh thông qua kết hợp các bộ phân lớp không gian con ngẫu nhiên, *Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin*, 2018**,** 72- 78

[CT2] **Cù Việt Dũng**, Nguyễn Hữu Quỳnh, Ngô Quốc Tạo, Trần Thị Minh Thu, Một phương pháp tra cứu ảnh học biểu diễn và học đa tạp cho giảm chiều với thông tin từ người dùng, *Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin*, 2019**,** 307-314

[CT3] **Cù Việt Dũng**, An Hồng Sơn, Nguyễn Hữu Quỳnh, Ngô Quốc Tạo, Đào Thị Thúy Quỳnh, Phương pháp học bán giám sát dựa vào đồ thị xây dựng tập mẫu cân bằng cho tra cứu ảnh, *Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin*, 2021**,** 143-149

Quốc tế:

[CT4] Nguyen Huu Quynh, **Cu Viet Dung**, Dao Thi Thuy Quynh, Ngo Quoc Tao, Phuong Van Canh, Graph-based semisupervised and manifold learning for image retrieval with SVM-based relevant feedback, *Journal of Intelligent & Fuzzy Systems(SCIE,IF=1.637)*, 2019, *37*, 711–722

[CT5] Nguyen Huu Quynh, **Cu Viet Dung**, Dao Thi Thuy Quynh, (2021), Semantic class discriminant projection for image retrieval with relevance feedback. *Multimedia Tools and Applications (SCIE, IF = 2.313, Q1),* 2021**,** *80*, 15351–15376

DANH MỤC TÀI LIỆU THAM KHẢO

1. H. Wang, et al., Texture image retrieval based on fusion of local and global features. *Multimedia Tools and Applications*, 2022. 81(10): p. 14081-14104.
2. N. B. Mohite and A. B. Gonde, Deep features based medical image retrieval. *Multimedia Tools and Applications*, 2022. 81(8): p. 11379-11392.
3. X. F. He and a. P. Niyogi, Locality preserving projections. *Proc. Advances in Neural Information Processing Systems*, 2003: p. pages 153–160.
4. Y. Xu, et al., Lpp solution schemes for use with face recognition. *Pattern Recognition*, 2010. vol 43: p. pages 4165–4176.
5. S. T. Roweis, Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science, vol. 290, no. 5500*, 2000: p. pp. 2323–2326.
6. X. He, et al., Neighborhood preserving embedding. *in Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2005: p. pages 1208–1213.
7. M. Vlachos, et al., Non-linear dimensionality reduction techniques for classification and visualization. *in: Proceedings of ACM Int. Conf. Knowl. Discovery Data Mining*, 2002.
8. X. Geng, D. C. Zhan, and a. Z. H. Zhou, Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Trans. Syst., Man, Cybern. B, Cybern. 35* 2005: p. pages 1098–1107.
9. S. Yan, et al., Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007. vol. 29, no. 1: p. pages 40–51.
10. H. T. Zhao, et al., Local structure based supervised feature extraction. *Pattern Recognition*, 2006. vol 39: p. pages 1546–1550.
11. W. K. Wong and a. H. T. Zhao, Supervised optimal locality preserving projection. *Pattern Recognition*, 2012. vol 45: p. pages 186–197.
12. W. Zhang, et al., Discriminant neighborhood embedding for classification. *Pattern Recognition*, 2006. vol 39: p. pages 2240–2243.
13. Z Liu, et al., Linear regression classification steered discriminative projection for dimension reduction. *Multimedia Tools and Applications*, 2020: p. pages 11993-12005.
14. J. Gou, et al., Discriminative globality and locality preserving graph embedding for dimensionality reduction. *Expert Systems with Applications*, 2020. vol 144: p. page 113079.
15. Y. Y. Lin, T. L. Liu, and a. H. T. Chen, Semantic Manifold Learning for Image Retrieval. *Proc. 13th Ann. ACM Int'l Conf. Multimedia (Multimedia '05)*, 2005.
16. X. He, D. Cai, and a. J. Han, Learning a maximum margin subspace for image retrieval. *IEEE Trans, Knowl. Data Eng*, 2008. vol. 20, no. 2: p. pp. 189–201.
17. D. Cai, X. He, and J. Han, Semi-supervised discriminant analysis. *Computer Vision, ICCV 2007*, 2007.

18. F. Dornaika and a. Y. E. Traboulsi, Learning flexible graph-based semi-supervised embedding. *IEEE transactions on cybernetics*, 2015. 46(1): p. pages 206-218.
19. Q Gao, et al., A novel semi-supervised learning for face recognition. *Neurocomputing*, 2015. 152: p. 69-76.
20. C Hoi, et al., Biased support vector machine for relevance feedback in image retrieval. *in Proc. IJCNN*, 2004: p. pp. 3189–3194.
21. H Tamura, S Mori, and T. Yamawaki, Texture Features Corresponding to Visual Perception. *IEEE Trans. Systems, Man, and Cybernetics*, 1978. vol. 8, no. 6: p. pages 460-473.
22. F Long, H Zhang, and D D Feng, *Fundamentals of content-based image retrieval*, in Multimedia information retrieval and management. 2003, Springer. p. 1-26.
23. N Shrivastava and V Tyagi, An efficient technique for retrieval of color images in large databases. *Computers & Electrical Engineering*, 2015. 46: p. 314-327.
24. Z S Younus, et al., Content-based image retrieval using PSO and k-means clustering algorithm. *Arabian Journal of Geosciences*, 2015. 8(8): p. 6211-6224.
25. M Sajjad, et al., Integrating salient colors with rotational invariant texture features for image representation in retrieval systems. *Multimedia Tools and Applications*, 2018. 77(4): p. 4769-4789.
26. A Nazir, et al., Content based image retrieval system by using HSV color histogram, discrete wavelet transform and edge histogram descriptor. *2018 international conference on computing, mathematics and engineering technologies (iCoMET)*, 2018: p. 1-6.
27. U Sharif, et al., Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artificial Intelligence Review*, 2019. 52(2): p. 901-925.
28. M Yousuf, et al., A novel technique based on visual words fusion analysis of sparse features for effective content-based image retrieval. *Mathematical Problems in Engineering*, 2018. 2018.
29. H Bay, et al., Speeded-up robust features (SURF). *Computer vision and image understanding*, 2008. 110(3): p. 346-359.
30. S Jabeen, et al., An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model. *PloS one*, 2018. 13(4): p. e0194526.
31. J Wan, et al. Deep learning for content-based image retrieval: A comprehensive study. in *Proceedings of the 22nd ACM international conference on Multimedia*. 2014.
32. Q Zheng, et al., Differential Learning: A Powerful Tool for Interactive Content-Based Image Retrieval. *Engineering Letters*, 2019. 27(1).
33. K Simonyan and A Zisserman, Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

34. T Kurita and T Kato. Learning of personal visual impression for image database systems. in *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93)*. 1993. IEEE.
35. T. Huang, et al., Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 1998: p. pages 25– 36.
36. L. Shao, F. Zhu, and a. X. Li, Transfer learning for visual categorization: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2015. vol. 26, no. 5: p. pages 1019–1034.
37. S. Sclaroff, L. Taycher, and M. La Cascia. Imagerover: A content-based image browser for the world wide web. in *1997 Proceedings IEEE workshop on content-based access of image and video libraries*. 1997. IEEE.
38. Y. Ishikawa, R. Subramanya, and C. Faloutsos, MindReader: Querying Databases Through Multiple Examples. *In VLDB '98: Proceedings of the 24rd International Conference on Very Large Data Bases*, 1998: p. pages 218–227.
39. Y. Rui, T. Huang, and S. Mehrotra, Content-Based Image Retrieval with Relevance Feedback in MARS. *In ICIP '97: Proceedings of the IEEE International Conference On Image Processing*, 1997: p. pages 815–818.
40. C. Nastar, M. Mitschke, and C. Meilhac. Efficient query refinement for image retrieval. in *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231)*. 1998. IEEE.
41. K. Porkaew and K. Chakrabarti. Query refinement for multimedia similarity retrieval in MARS. in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. 1999.
42. Y. Chen, X. S. Zhou, and a. T. S. Huang, One-class SVM for learning in image retrieval. *in Proceedings of IEEE International Conference on Image Processing*, 2001: p. pages. 34 –37.
43. L. Zhang, F. Lin, and B. Zhang, Support Vector Machine Learning For Image Retrieval. *in Image Processing. Proceedings. International Conference*, 2001: p. pages 721 - 724
44. S. Tong and E. Chang, Support Vector Machine Active Learning for Image Retrieval. *Proc. ACM Int'l Conf. Multimedia*, 2001: p. pages 107-118.
45. G. Guo, et al., Learning similarity measure for natural image retrieval with relevance feedback. *IEEE Trans. Neural Netw.*, 2002. vol. 13, no. 4: p. pages 811–820.
46. P. Hong, Q. Tian, and a. T. S. Huang, Incorporate support vector machines to content-based image retrieval with relevant feedback. *in Proc. IEEE ICIP, Vancouver, BC, Canada*, 2000: p. pages 750–753.
47. D. Tao, et al., Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, 2006. vol. 28, no. 7: p. pages 1088 –1099.
48. Vu Van Hieu, et al., Một phương pháp mới chuẩn hoá dữ liệu và hiệu chỉnh trọng số cho tổ hợp đặc trưng trong tra cứu ảnh theo nội dung. *Các công trình nghiên cứu, phát triển và ứng dụng Công nghệ Thông tin và Truyền thông*, 2016: p. 63-63.
49. Vu Van Hieu, Content based image retrieval using multiple features and Pareto approach. *Journal of Computer Science and Cybernetics*, 2016. 32(2): p. 169-187.
50. Ngo Truong Giang, et al. Batch mode active learning for interactive image retrieval. in *2014 IEEE International Symposium on Multimedia*. 2014. IEEE.
51. Ngo Truong Giang, et al., Learning interaction measure with relevance feedback in image retrieval. *Journal of Computer Science and Cybernetics*, 2016. 32(2): p. 113-131.
52. Dao Thi Thuy Quynh, et al., An efficient semantic–related image retrieval method. *Expert Systems with Applications*, 2017. 72: p. 30-41.
53. R.O. Duda, P.E. Hart, and a. D. G. Stork, Pattern Classification. *Wiley-Interscience*, 2000: p. pages 831–836.
54. I. T. Jolliffe, Principal Component Analysis. *2nd ed. New-York: Springer-Verlag*, 2002.
55. D. Tao, et al., Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. 31(2): p. 260-274.
56. X. S. Zhou and a. T. S. Huang, Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 2003. vol. 8, no. 6: p. pages 536–544.
57. X He. Incremental semi-supervised subspace learning for image retrieval. in *Proceedings of the 12th annual ACM international conference on Multimedia*. 2004.
58. J Tenenbaum, V D Silva, and J. Langford, A global geometric framework for nonlinear dimensionality reduction. *science*, 2000. 290(5500): p. 2319-2323.
59. M Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. in *Nips*. 2001.
60. J Li, et al., Discriminative multi-view interactive image re-ranking. *IEEE Transactions on Image Processing*, 2017. 26(7): p. 3113-3127.
61. I M Hameed, S H Abdulhussain, and B M Mahmmod, Content-based image retrieval: A review of recent trends. *Cogent Engineering*, 2021. 8(1): p. 1927469.
62. J. Wang, et al., Semantics-sensitive retrieval for digital picture libraries. *D-Lib Magazine*, 1999. 5(11).
63. M.J. Swain and D.H. Ballard. Indexing via color histograms. in *Active perception and robot vision*. 1992. Springer.

64. J. Huang, et al. Image indexing using color correlograms. in *Proceedings of IEEE computer society conference on Computer Vision and Pattern Recognition*. 1997. IEEE.
65. H. Yu, et al. Color texture moments for content-based image retrieval. in *Proceedings. International Conference on Image Processing*. 2002. IEEE.
66. T.S. Lee, Image representation using 2D Gabor wavelets. *IEEE Transactions on pattern analysis and machine intelligence*, 1996. 18(10): p. 959-971.
67. S. Manjunath and W. Ma, Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence*, 1996. 18(8): p. 837-842.
68. J. Wang, J. Li, and G. Wiederhold, SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on pattern analysis and machine intelligence*, 2001. 23(9): p. 947-963.
69. J. Philbin, et al. Object retrieval with large vocabularies and fast spatial matching. in *2007 IEEE conference on computer vision and pattern recognition*. 2007. IEEE.
70. L Fei-Fei, R Fergus, and P Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. in *2004 conference on computer vision and pattern recognition workshop*. 2004. IEEE.
71. P. Hew, Geometric and Zernike Moments (1996), 'Diary', Department of Mathematics, The University of Western Australia. *http://citeseer.ist.psu.edu/hew96 geometric.html*, 1996.
72. C. Ding and L. Zhang, Double adjacency graphs-based discriminant neighborhood embedding. *Pattern Recognition*, 2015: p. 1734–1742.
73. M. Masaeli, J. G. Jennifer, and M. F. Glenn. From transformation-based dimensionality reduction to feature selection. in *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010.
74. L Zhang, H Shum, and L. Shao, Discriminative semantic subspace analysis for relevance feedback. *IEEE Transactions on image processing*, 2016. 25(3): p. 1275-1287.
75. J. Shi and a. J. Malik, Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal*, 2000. vol. 22, no. 8: p. pages 888–905.
76. L Nanni, C Fantozzi, and N Lazzarini, Coupling different methods for overcoming the class imbalance problem. *Neurocomputing*, 2015. 158: p. 48-61.
77. D. Bahler and L. Navarro, Methods for Combining Heteroge-neous Sets of Classifiers. *Proc. 17th Nat'l Conf. Am. Assoc. for Artificial Intelligence.*, 2000.
78. J Wu, et al., Heterogeneous manifold ranking for image retrieval. *IEEE Access*, 2017. 5: p. 16871-16884.