

ĐẠI HỌC QUỐC GIA TP HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

HUỲNH NGỌC TÍN

**PHÁT TRIỂN MỘT SỐ PHƯƠNG PHÁP KHUYẾN NGHỊ
HỖ TRỢ TÌM KIẾM THÔNG TIN HỌC THUẬT
DỰA TRÊN TIẾP CẬN PHÂN TÍCH MẠNG XÃ HỘI**

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

TP HỒ CHÍ MINH – Năm 2016

ĐẠI HỌC QUỐC GIA TP HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN



HUỲNH NGỌC TÍN

**PHÁT TRIỂN MỘT SỐ PHƯƠNG PHÁP KHUYẾN NGHỊ
HỖ TRỢ TÌM KIẾM THÔNG TIN HỌC THUẬT
DỰA TRÊN TIẾP CẬN PHÂN TÍCH MẠNG XÃ HỘI**

Chuyên ngành: Khoa học Máy tính

Mã số: 62.48.01.01

Phản biện 1: PGS.TS. Đỗ Phúc

Phản biện 2: PGS.TS. Lê Hoài Bắc

Phản biện 3: PGS.TS. Quản Thành Thơ

Phản biện độc lập 1: PGS.TS. Nguyễn Đình Thúc

Phản biện độc lập 2: PGS.TS. Đỗ Năng Toàn

**NGƯỜI HƯỚNG DẪN KHOA HỌC
GS.TSKH. HOÀNG VĂN KIẾM**

TP HỒ CHÍ MINH – Năm 2016

Xin dành tặng quyền luận án này cho cha, mẹ và vợ của tôi.

LỜI CẢM ƠN

Đầu tiên, xin được gửi lời tri ân sâu sắc nhất đến GS.TSKH Hoàng Văn Kiếm, người thầy đã tận tình hướng dẫn, định hướng, và động viên em suốt thời gian học tập, nghiên cứu và thực hiện luận án này.

Xin chân thành cảm ơn GS. Atsuhiro Takasu, PGS.TS Lê Hoài Bắc, PGS.TS Đỗ Phúc, PGS.TS Lê Đình Duy, TS. Nguyễn Hoàng Tú Anh, TS. Nguyễn Anh Tuấn vì những ý kiến đóng góp quý báu cho luận án.

Xin cảm ơn Ban giám hiệu, phòng SDH-KHCN, Khoa Khoa học Máy tính, UIT-MMLab, UIT-Cloud Team và đồng nghiệp ở Trường Đại học Công nghệ Thông tin đã hỗ trợ tôi trong quá trình thực hiện và bảo vệ luận án.

Cuối cùng, tôi muốn bày tỏ lòng biết ơn sâu sắc đến Cha, Mẹ, Vợ luôn là điểm tựa vững chắc, đã chăm sóc và tiếp thêm nghị lực giúp tôi có thể hoàn thành tốt luận án này.

Tp. HCM, ngày 20 tháng 11 năm 2014

Tác giả luận án

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong luận án là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác, ngoại trừ các tư liệu được trích dẫn ghi trong mục tài liệu tham khảo.

Tác giả luận án

Mục lục

Lời cảm ơn	ii
Lời cam đoan	iii
Mục lục	1
Danh mục các ký hiệu, thuật ngữ và chữ viết tắt	7
Danh sách bảng	10
Danh sách hình vẽ	11
MỞ ĐẦU	13
Dẫn nhập	13
Mục tiêu và nội dung thực hiện của luận án	19
Đối tượng và phạm vi nghiên cứu	19
Các đóng góp chính của luận án	20
Bố cục của luận án	23
Chương 1. TỔNG QUAN: HỆ KHUYẾN NGHỊ, NHỮNG PHƯƠNG PHÁP TIẾP CẬN PHỐ BIỀN VÀ XU HƯỚNG	25
1.1 Giới thiệu	25
1.2 Khái niệm Hệ khuyến nghị	25
1.3 Phát biểu Bài toán Khuyến nghị	26
1.4 Các cách tiếp cận phổ biến	28
1.4.1 Tiếp cận nội dung (CB)	28
1.4.1.1 Kiến trúc hệ thống	30

1.4.1.2	Xây dựng và cập nhật hồ sơ người dùng	32
1.4.1.3	Phân loại tiếp cận nội dung	36
1.4.1.4	Ưu điểm và hạn chế của tiếp cận nội dung	40
1.4.2	Tiếp cận lọc cộng tác (CF)	40
1.4.2.1	Tiếp cận CF dựa trên bộ nhớ	42
1.4.2.2	Tiếp cận CF dựa trên mô hình	44
1.4.2.3	Ưu điểm và hạn chế của tiếp cận CF	50
1.4.3	Tiếp cận lai (Hybrid Approach)	51
1.4.3.1	Lai có trọng số (Weighted Hybrid)	51
1.4.3.2	Lai chuyển đổi (Switching Hybrid)	52
1.4.3.3	Lai trộn (Mixed Hybrid)	53
1.4.3.4	Lai kết hợp đặc trưng (Feature Combination Hybrid) .	54
1.4.3.5	Lai theo đợt (Cascade Hybrid)	54
1.4.3.6	Lai tăng cường đặc trưng (Feature Augmentation Hybrid)	55
1.4.3.7	Lai meta (Meta-Level Hybrid)	56
1.4.4	Tiếp cận phân tích mạng xã hội	57
1.4.4.1	Một số khái niệm cơ bản	57
1.4.4.2	Khuyến nghị xã hội (Social Recommendation)	59
1.5	Các phương pháp đánh giá hệ khuyến nghị	64
1.5.1	Phương pháp thiết lập thực nghiệm	64
1.5.2	Dộ đo đánh giá	65
1.5.2.1	Tiên đoán đánh giá	66
1.5.2.2	Tối ưu tính hữu ích của hệ thống khuyến nghị	67
1.5.2.3	Khuyến nghị các đối tượng tốt	67
1.6	Khó khăn, thách thức và xu hướng	68
1.6.1	Khó khăn, thách thức	68
1.6.2	Xu hướng mới cho hệ khuyến nghị	69
1.7	Kết chương	71

Chương 2. XÁC ĐỊNH VÀ MÔ HÌNH HÓA MẠNG XÃ HỘI HỌC THUẬT

72

2.1	Giới thiệu	72
2.2	Xây dựng và làm giàu kho dữ liệu học thuật	73
2.2.1	Tích hợp từ nhiều nguồn	74
2.2.2	Các thành phần chính của hệ thống	75
2.2.3	Rút trích thông tin bài báo từ các tập tin PDF	76
2.2.3.1	Dùng luật dựa trên GATE Framework	76
2.2.3.2	Rút trích metadata cho mục Header và mục Reference	76
2.2.4	Rút trích thông tin bài báo từ các trang web	77
2.2.5	Kết quả kho dữ liệu tích hợp	78
2.3	Xác định và mô hình hóa các mạng xã hội học thuật (ASN)	79
2.3.1	Thành phần chính của mô hình ASN	79
2.3.2	Mạng đồng tác giả CoNet giữa các nghiên cứu viên	80
2.3.2.1	Cấu trúc một nghiên cứu viên	81
2.3.2.2	Cấu trúc cung liên kết	82
2.3.3	Mạng trích dẫn giữa các nghiên cứu viên <i>CiNet_Author</i>	82
2.3.4	Mạng trích dẫn giữa các bài báo <i>CiNet_Paper</i>	82
2.3.5	Mạng cộng tác giữa các trường, viện AffNet	83
2.3.6	Các phương pháp tính toán trong mô hình ASN (Thành phần M trong mô hình ASN)	83
2.3.6.1	Các phương pháp tương tự đindh truyền thống	84
2.3.6.2	Đề xuất các phương pháp tương tự đindh mới	84
2.3.6.3	Đề xuất phương pháp lượng hóa quan hệ lòng tin	89
2.3.6.4	Đề xuất tập đặc trưng của nghiên cứu viên tiềm năng cho khuyến nghị cộng tác	90
2.4	Kết chương	93
Chương 3.	KHAI THÁC MẠNG XÃ HỘI HỌC THUẬT ĐỂ PHÁT TRIỂN CÁC PHƯƠNG PHÁP KHUYẾN NGHỊ CỘNG TÁC	94
3.1	Giới thiệu	94
3.2	Bài toán khuyến nghị cộng tác	95

3.3	Trường hợp các nghiên cứu viên có đồng tác giả (un-isolated)	97
3.3.1	Tương tự đindh dựa trên cấu trúc cục bộ	97
3.3.2	Tương tự đindh dựa trên cấu trúc toàn cục	98
3.3.3	Nhận định	98
3.3.4	Các phương pháp đề xuất	99
3.3.4.1	Tương tự đindh dựa trên đường dẫn có trọng số cực đại (MPRS)	100
3.3.4.2	Tương tự đindh dựa trên đường dẫn cực đại có xét xu hướng (MPRS+)	101
3.3.4.3	Tương tự đindh dùng phương pháp RSS+ (cải tiến từ RSS)	103
3.3.5	Thực nghiệm và đánh giá	104
3.3.5.1	Thiết lập dữ liệu thực nghiệm cho DBLP và CSPubGuru	105
3.3.5.2	Kết quả thực nghiệm	106
3.3.5.3	Kết luận	108
3.4	Trường hợp các nghiên cứu viên chưa có đồng tác giả (Isolated Researcher)	109
3.4.1	Tiếp cận của luận án	109
3.4.1.1	Tương tự nội dung nghiên cứu (Content Similarity). .	109
3.4.1.2	Quan hệ giữa các cơ quan	110
3.4.1.3	Uy tín của nghiên cứu viên	110
3.4.1.4	Độ năng động của nghiên cứu viên	110
3.4.1.5	Học máy để tiên toán liên kết đồng tác giả, phục vụ khuyến nghị	111
3.4.2	Phương pháp Đánh giá	111
3.4.2.1	Độ chính xác tiên đoán liên kết	111
3.4.2.2	Đề xuất phương pháp đánh giá chất lượng cộng tác .	113
3.4.3	Thực nghiệm, đánh giá	114
3.4.3.1	Tập dữ liệu thực nghiệm	115
3.4.3.2	Kết quả thực nghiệm	116
3.5	Kết chương	119

Chương 4. KHAI THÁC MẠNG XÃ HỘI HỌC THUẬT ĐỂ PHÁT TRIỂN CÁC PHƯƠNG PHÁP KHUYẾN NGHỊ BÀI BÁO KHOA HỌC	121
4.1 Giới thiệu	121
4.2 Bài toán Khuyến nghị bài báo khoa học	123
4.3 Khó khăn, thách thức	124
4.4 Nghiên cứu liên quan	125
4.5 Các phương pháp phổ biến cho khuyến nghị bài báo liên quan	128
4.5.1 Tiếp cận nội dung	128
4.5.1.1 CB-Baseline	128
4.5.1.2 Mô hình hóa sở thích của các nghiên cứu viên dựa trên nội dung các bài báo công bố, tham khảo, và trích dẫn (CB+R+C)	129
4.5.1.3 Phương pháp mô hình hóa xu hướng nghiên cứu của nghiên cứu viên (CB-Recent)	130
4.5.2 Tiếp cận lọc cộng tác - CF	132
4.5.3 Kết hợp tuyến tính CB và CF	134
4.6 Các phương pháp đề xuất	134
4.6.1 Kết hợp Xu hướng nghiên cứu và quan hệ lòng tin	134
4.6.1.1 Lòng tin dựa trên quan hệ đồng tác giả và quan hệ trích dẫn (CB-TrendTrust1)	135
4.6.1.2 Lòng tin dựa trên quan hệ trích dẫn tiềm ẩn (CB-TrendTrust2)	137
4.7 Thực nghiệm, đánh giá	138
4.7.1 Tập dữ liệu và thiết lập thực nghiệm	138
4.7.2 Độ đo đánh giá độ chính xác khuyến nghị	139
4.7.2.1 Độ đo NDCG (Normalized Discounted Cumulative Gain)	139
4.7.2.2 Độ đo MRR (Mean Reciprocal Rank)	140
4.7.3 Kết quả thực nghiệm	140
4.7.4 Kết luận	142
4.8 Kết chương	144

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	145
Các kết quả đạt được	145
Giá trị thực tiễn của luận án	146
Hướng phát triển	147
CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ CỦA TÁC GIẢ	148
CÁC ĐỀ TÀI KHOA HỌC CHỦ TRÌ THỰC HIỆN	151
Phụ lục A. Xây dựng và làm giàu kho dữ liệu học thuật	152
Phụ lục B. Chi tiết kho dữ liệu học thuật	163
TÀI LIỆU THAM KHẢO	167

Danh mục các ký hiệu, thuật ngữ và chữ viết tắt

Academic Social Network	: Mạng xã hội học thuật
Bayesian Classifier	: Bộ phân lớp Bayes
Bayesian Network	: Mạng Bayes
Cascade Hybrid	: Lai theo đợt
Case-Based Reasoning (CBR)	: Suy luận theo trường hợp
Centrality Measures	: Các độ đo trung tâm
Collaborative Filtering	: Lọc cộng tác
Content-Based Approach	: Tiếp cận nội dung
Collaborative Filtering Approach	: Tiếp cận lọc cộng tác
Correlation	: Tương quan
Clustering	: Gom cụm
Cold-Start	: Khởi động lạnh
Context-aware	: Nhận biết ngữ cảnh
Demographic Filtering	: Lọc dựa trên thông tin cá nhân
Feature Combination	: Kết hợp đặc trưng
Feature Augmentation	: Tăng cường đặc trưng
Heuristic-Based Collaborative Filtering	: Lọc cộng tác dựa trên kinh nghiệm
Hybrid Approach	: Tiếp cận lai
Hybrid Recommender System	: Hệ khuyến nghị lai
Information Retrieval	: Truy vấn thông tin
Internet of Things (IoT)	: Mạng tương tác thực ảo toàn cầu
Isolated researcher	: Nghiên cứu viên chưa có đồng tác giả
Item	: Đối tượng khuyến nghị

Latent Factor Model	: Mô hình đặc trưng tiềm ẩn
Matrix Factorization	: Thừa số hóa ma trận
Memory-based Collaborative Filtering	: Lọc cộng tác dựa trên bộ nhớ
Meta-Level Hybrid	: Lai meta
Mixed Hybrid	: Lai trộn
Model-based Collaborative Filtering	: Lọc cộng tác dựa trên học máy
Naive Bayes	: Bayes ngây thơ
Peers	: Những người đồng sở thích
Prediction	: Tiên đoán
Predicting Rating	: Tiên đoán đánh giá
Recommendation	: Khuyến nghị
Recommender System	: Hệ khuyến nghị
Researcher	: Nghiên cứu viên
Researcher Profile	: Hồ sơ nghiên cứu viên
Rating	: Đánh giá
Rating Matrix	: Ma trận đánh giá
Rating Prediction	: Tiên đoán đánh giá
Rating Score	: Điểm đánh giá
Social Network	: Mạng xã hội
Social Network Analysis	: Phân tích Mạng xã hội
Social Recommendation	: Khuyến nghị xã hội
Switching Hybrid	: Lai chuyển đổi
Top-N	: Lấy N phần tử đầu tiên từ một danh sách có xếp hạng
Un-isolated researcher	: Nghiên cứu viên có đồng tác giả
Unknown Rating	: Đánh giá chưa biết
User-Item	: Người dùng - Đối tượng khuyến nghị
User Profile	: Hồ sơ người dùng
User's taste	: Sở thích người dùng
Utility	: Tính hữu ích
Utility Function	: Hàm hữu ích
Utility Optimization	: Tối ưu tính hữu ích
Weighted Hybrid	: Lai có trọng số

ASN	: Academic Social Networks
CB	: Content Based
CF	: Collaborative Filtering
CT	: Công trình
DBLP	: Digital Bibliography & Library Project
MAS	: Microsoft Academic Search
MPRS	: Maximum Path base Relation Strength
MPRS+	: Maximum Path Relation Strength +
RSS	: Relation Strength Similarity
RSS+	: Relation Strength Similarity +
SNA	: Social Network Analysis
SVD	: Singular Value Decomposition
SVM	: Support Vector Machine

Danh sách bảng

1.1	Ví dụ tiên đoán đánh giá	46
1.2	Tóm tắt ưu nhược điểm những tiếp cận phổ biến và xu hướng nghiên cứu	70
2.1	Thông tin bài báo sẵn có từ DBLP, CiteSeerX	73
2.2	Các mẫu truy vấn được gởi đến CiteSeerX	77
2.3	Các mẫu truy vấn được gởi đến các thư viện không hỗ trợ OAI-PMH tương ứng với từ khóa 'Information Extraction'	78
2.4	Thông tin bài báo sẵn có từ DBLP, CiteSeerX, CSPubGuru	78
3.1	Kích thước tập dữ liệu thực nghiệm	106
3.2	Kết quả tiên đoán liên kết đồng tác giả trên tập thực nghiệm DBLP . .	107
3.3	Kết quả tiên đoán đồng tác giả trên tập thực nghiệm CSPubGuru . .	108
3.4	Độ chính xác tiên đoán đồng tác giả khi thêm đặc trưng mới	118
3.5	Chất lượng tiên đoán TopN khi thêm các đặc trưng mới	119
4.1	Tóm tắt so sánh, đánh giá các phương pháp đề xuất và các phương pháp phổ biến hiện nay	143
4.2	Đề tài khoa học đã và đang thực hiện	151
A.1	Ví dụ các bài báo nhập nhằng tên tác giả	157

Danh sách hình vẽ

0.1	Sự gia tăng dữ liệu khoa học dựa trên Cơ sở dữ liệu khoa học DBLP	14
1.1	Phân loại hệ khuyến nghị dựa trên công việc khuyến nghị	26
1.2	Các cách tiếp cận phổ biến và xu hướng hiện nay cho hệ khuyến nghị	29
1.3	Kiến trúc tổng quan của hệ khuyến nghị dựa trên tiếp cận nội dung	31
1.4	Học và cập nhật hồ sơ người dùng dựa trên thông tin phản hồi	36
1.5	Dấu ? là các giá trị cần tiên đoán trong ma trận đánh giá	41
1.6	Minh họa dùng CF để tiên đoán một người thích hay không thích xem phim.	43
1.7	Minh họa trực quan mạng xã hội	58
1.8	Minh họa khuyến nghị xã hội	60
2.1	Tích hợp dữ liệu bài báo khoa học từ nhiều nguồn không đồng nhất	74
2.2	Các bước rút trích thông tin từ header của bài báo	76
2.3	Các bước rút trích thông tin từ phần reference của bài báo	77
2.4	Kích thước kho dữ liệu tích hợp tính đến 03/2013	79
2.5	Minh họa các cấu trúc xã hội từ kho dữ liệu bài báo khoa học	80
2.6	Trực quan hàm $e^{-\delta(t)}$ ($\delta(t) \in [0, +\infty]$)	87
3.1	Những phương pháp dựa trên phân tích mạng đồng tác giả có thể khuyến nghị cộng tác cho các nghiên cứu viên có đồng tác giả (nét chấm đứt trong hình vẽ), nhưng sẽ không thực hiện được đối với các nghiên cứu viên chưa có đồng tác giả (quanh dấu chấm hỏi).	96
3.2	Minh họa cách tính mức độ quan hệ	99
3.3	Minh họa cách đánh giá độ chính xác khuyến nghị cộng tác	106
3.4	Kết quả tiên đoán đồng tác giả trên tập thực nghiệm DBLP	107
3.5	Kết quả tiên đoán đồng tác giả trên tập thực nghiệm CSPubGuru	108
3.6	Phân bố của mẫu dương (xanh) và mẫu âm (đỏ) trong không gian đặc trưng 2-chiều.	117
3.7	Độ chính xác AP khi thêm các đặc trưng mới	118
3.8	Chất lượng tiên đoán TopN khi thêm các đặc trưng mới	119
4.1	Minh họa cách tính độ chính xác khuyến nghị bài báo	139
4.2	Kết quả thực nghiệm phương pháp CB+R+C với tham số ngưỡng tương tự Th_j	141
4.3	Kết quả thực nghiệm phương pháp CB-Recent với các hệ số xu hướng alpha khác nhau	141
4.4	Kết quả thực nghiệm phương pháp lọc cộng tác CF-kNN với các giá trị k khác nhau	142

4.5	Kết quả thực nghiệm phương pháp kết hợp tuyến tính CB-Recent và CF	142
4.6	Phương pháp kết hợp xu hướng sở thích và quan hệ lòng tin	143
B.1	Mô hình ERD biểu diễn cấu trúc của tập dữ liệu đã xây dựng, CSPub-Guru dataset	166

MỞ ĐẦU

Dẫn nhập

Việc tìm kiếm bài báo, chuyên gia, thông tin khoa học để thực hiện các công việc liên quan đến nghiên cứu như khảo sát, trích dẫn, cộng tác, viết bài, gởi bài, ... là nhu cầu thường xuyên, không thể thiếu đối với những người làm nghiên cứu khoa học, đặc biệt là các nghiên cứu viên. Các hệ thống tìm kiếm, thư viện số phổ biến hiện nay trong lĩnh vực học thuật như ACM DL Portal, IEEE Xplore, Google Scholar, Microsoft Academic Search, DBLP, ... đã đáp ứng hầu hết nhu cầu tìm kiếm của các nghiên cứu viên. Tuy nhiên, đối với các nghiên cứu viên trẻ thì thường chưa đủ hiểu biết và kinh nghiệm để tự tìm ra các thông tin hữu ích liên quan đến nghiên cứu của mình. Còn đối với các nghiên cứu viên có kinh nghiệm thì phải đương đầu với tình trạng quá tải thông tin, và mất nhiều thời gian hơn để tìm được những thông tin liên quan.

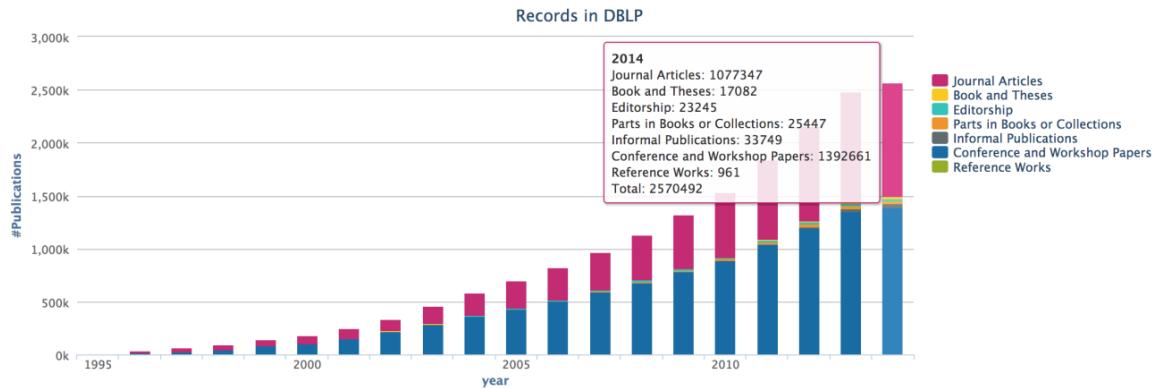
Sự bùng nổ, gia tăng một cách nhanh chóng các kho dữ liệu trên web nói chung và các kho dữ liệu học thuật nói riêng đã gây không ít khó khăn cho các nghiên cứu viên trong việc tìm kiếm thông tin liên quan. Theo thống kê từ kho dữ liệu công bố khoa học chuyên ngành khoa học máy tính DBLP¹, tháng 12/2005 DBLP có tổng cộng là 964.222 công bố khoa học; Đến tháng 12/2011 DBLP có tổng cộng 2.001.414 công bố khoa học, tăng khoảng 108% so với 2005; Đến tháng 12/2012 số công bố khoa học tổng cộng trong DBLP là 2.163.145, tăng khoảng 8% so với 2011; Và tính đến tháng 01 năm 2013 có 2.167.502 bài báo khoa học (hình 0.1). Để giúp cho những người làm nghiên cứu khoa học có thể đương đầu với tình trạng bùng nổ thông tin từ các kho dữ liệu khoa học hiện nay và có thể dễ dàng tìm thấy những thông tin hữu ích liên quan, thì hệ khuyến nghị (Recommender System) trong lĩnh vực học thuật là giải pháp đang

¹<http://dblp.uni-trier.de/> mwagner/statistics/recordsindblp(groupedbyyear).html, truy cập lần cuối ngày 5/2/2014

được quan tâm nghiên cứu trong vài năm trở lại đây. Với hệ khuyến nghị nói chung và trong lĩnh vực học thuật nói riêng thì các thông tin liên quan đến nhu cầu tìm kiếm sẽ tự động tìm đến các nghiên cứu viên, thay vì họ phải vất vả tự đi tìm thông tin như trong các hệ thống tìm kiếm thông tin truyền thống.

Records in DBLP

The diagram shows the total number of publications of the different publication types.



Hình 0.1: Sự gia tăng dữ liệu khoa học dựa trên Cơ sở dữ liệu khoa học DBLP
(Nguồn: <http://www.informatik.uni-trier.de/ley/statistics/recordsindblp.html>, truy cập lần cuối 20/03/2014)

Một số các công bố điển hình liên quan đến hệ khuyến nghị xuất hiện đầu thập niên 90 trong số đặc biệt năm 1992 của tạp chí “Communications of the ACM” về lọc thông tin có thể kể đến như công bố của Belkin N. J. và Croft B. về lọc và truy vấn thông tin [16]; công bố của Foltz P. W. và Dumais. S. T. liên quan đến việc phân tích các phương pháp lọc thông tin [43]. Theo tác giả Adomavicius và Tuzhilin, các nghiên cứu về hệ khuyến nghị đã và đang trở thành một lĩnh vực nghiên cứu rất quan trọng và thu hút nhiều quan tâm của cộng đồng [5]. Trong khoảng hai thập niên qua, có rất nhiều công việc được tiến hành trong môi trường hàn lâm, cũng như công nghiệp để phát triển những phương pháp mới cho hệ khuyến nghị. Có thể nói hệ khuyến nghị đã hình thành nên một lĩnh vực nghiên cứu mới, phong phú, có rất nhiều vấn đề khoa học, cũng như ứng dụng thực tế cần giải quyết nhằm cung cấp các dịch vụ, giúp người dùng có thể đương đầu với tình trạng ‘quá tải’ thông tin hiện nay. Các ứng dụng phổ biến có thể kể đến như khuyến nghị sách, sản phẩm của Amazon [70], hệ khuyến nghị phim cung cấp bởi MovieLens [84], hệ khuyến nghị video của YouTube [36]. Bên cạnh đó, những tổ chức, tập đoàn kinh tế lớn cũng có nhu cầu phát triển hệ khuyến nghị

thông minh, tích hợp vào hệ thống, máy chủ của họ để cung cấp thông tin tốt hơn cho người dùng. Để đề cao giá trị và thu hút sự quan tâm nghiên cứu của cộng đồng với hệ khuyến nghị, Rajaraman và Ullman đã đưa ra hai ví dụ quan trọng của hệ khuyến nghị đó là: (1) Tiên đoán sở thích của người đọc, hệ khuyến nghị cung cấp cho họ các bài báo tin tức trực tuyến; (2) Cung cấp cho khách hàng những sản phẩm từ những hệ thống bán lẻ mà có thể khách hàng cần mua, dựa trên lịch sử tìm kiếm và mua hàng của họ [97]. Adomavicius và Tuzhilin đã khảo sát và phân loại các phương pháp khuyến nghị truyền thống thành 3 nhóm chính: (1) khuyến nghị dựa trên nội dung, gọi tắt là tiếp cận nội dung (Content-Based Approach); (2) khuyến nghị dựa trên lọc cộng tác, gọi tắt là tiếp cận lọc cộng tác (Collaborative Filtering Approach) và (3) tiếp cận lai (Hybrid Approach)[5].

Tiếp cận nội dung (1) dựa trên việc so sánh nội dung của những sản phẩm quan sát với các sản phẩm mà người dùng quan tâm trong quá khứ, để tìm những sản phẩm gần với sở thích của người dùng. Ví dụ, khi cần khuyến nghị phim cho người dùng u , hệ khuyến nghị dựa trên nội dung, gọi tắt là hệ khuyến nghị nội dung, cố gắng hiểu những đặc điểm chung của những phim mà người dùng u quan tâm và có tỷ lệ bình chọn cao trong quá khứ, chẳng hạn như tên những diễn viên, đạo diễn, thể loại, chủ đề, v.v... Sau đó những phim có độ tương tự cao với sở thích của người dùng sẽ được khuyến nghị. Tiếp cận nội dung có nguồn gốc từ các nghiên cứu khai thác văn bản, truy vấn, lọc thông tin, do đó nó có một số hạn chế về việc phân tích nội dung sản phẩm như: hệ thống sẽ không thể phân biệt được chất lượng của hai bài báo là tốt hay xấu, uy tín hay không uy tín để khuyến nghị, khi hai bài báo đó được biểu diễn bằng một tập các từ khóa quan trọng như nhau. Bên cạnh đó việc rút trích đặc trưng tự động cũng khó áp dụng cho các định dạng dữ liệu khác không phải là văn bản như hình ảnh, video, âm thanh, v.v... Một hạn chế khác với tiếp cận nội dung có thể kể đến đó là: người dùng u chỉ được khuyến nghị các sản phẩm mà tương tự cao với những gì u đã bình chọn, đánh giá trong một phạm vi cụ thể. Khi vượt quá phạm vi thì hệ thống không thể thực hiện khuyến nghị được. Chẳng hạn tiếp cận nội dung sẽ thất bại khi u cần tham khảo các nhà hàng về ẩm thực Việt Nam, trong khi u chưa từng có những bình chọn và đánh giá về các nhà hàng, cũng đặc sản ẩm thực Việt Nam.

Không giống như tiếp cận nội dung, tiếp cận lọc cộng tác (2), cố gắng tiên đoán

mức độ tiềm năng của những sản phẩm sẽ khuyến nghị cho u dựa trên những sản phẩm được bình chọn bởi những người khác, có sở thích tương tự u . Ví dụ, khi cần khuyến nghị phim cho u , hệ khuyến nghị dựa trên tiếp cận lọc cộng tác, gọi tắt là hệ khuyến nghị lọc cộng tác, cố gắng xác định nhóm những người cùng sở thích với u về lĩnh vực phim (có những bình chọn tương tự cho những phim giống nhau). Sau đó hệ thống sẽ khuyến nghị cho u những phim mà những người đồng sở thích với u quan tâm nhiều nhất. Hệ thống lọc cộng tác đã và đang được ứng dụng rộng rãi trên thực tế như: khuyến nghị sách của Amazon [70], hệ khuyến nghị tin tức GroupLens [63], hệ thống Jester khuyến nghị các chuyện cười [47].

Với tiếp cận lọc cộng tác thì những sản phẩm mới chưa được bình chọn sẽ không được khuyến nghị cho người dùng, mặc dù nó có thể tương tự, tiềm năng và phù hợp với sở thích của người dùng. Một hạn chế nữa là đối với những người dùng mới, người chưa có hoặc rất ít những bình chọn về các sản phẩm liên quan. Khi đó hệ thống sẽ rất khó để có những khuyến nghị chính xác cho người dùng. Chẳng hạn trên Amazon, đối với những người dùng mới hoặc những sản phẩm chưa được bình chọn thì hệ thống không thể đưa ra các khuyến nghị chính xác cho những sản phẩm này. Do đó, đã có nhiều công trình nghiên cứu, phát triển các phương pháp lai (kết hợp hai hay nhiều phương pháp) như: Balabanovic và Shoham [11], Pazzani [95], Claypool và cộng sự [31], Nicholas [91], Li và Kim [68], và nhiều công trình khác nhằm giải quyết những hạn chế của mỗi phương pháp.

Nhìn chung, các phương pháp khuyến nghị truyền thống, phổ biến hiện nay đang gặp phải một số khó khăn, thách thức chính có thể kể đến như sau:

- Độ phức tính toán:
 - Dữ liệu lớn. Không gian người dùng và đối tượng khuyến nghị là rất lớn ảnh hưởng đến tốc độ xử lý của các thuật toán.
- Độ chính xác, chất lượng khuyến nghị: chưa cao, chưa đáp ứng thật tốt nhu cầu người dùng với một số lý do như:
 - Sở thích người dùng thay đổi theo thời gian.
 - Vấn đề ma trận đánh giá thưa, tức số đánh giá quan sát được rất ít so với số đánh giá cần tiên đoán để khuyến nghị.

-
- Vấn đề khởi động lạnh (cold start). Việc quan sát thiếu hay không quan sát được một số thông tin về sở thích, đánh giá của người dùng, cũng như các đối tượng khuyến nghị (người dùng, đối tượng khuyến nghị mới).
 - Chưa có những phương pháp thật sự tốt để đánh giá kết quả khuyến nghị.

Có thể thấy, tất cả những tiếp cận truyền thống (tiếp cận nội dung, tiếp cận lọc cộng tác, kể cả tiếp cận lai) chưa quan tâm đến các mối quan hệ xã hội của người dùng. Trên thực tế, khi cần mua một sản phẩm hay thực hiện một công việc gì đó thì chúng ta thường hỏi ý kiến bạn bè, người thân xem nên quyết định như thế nào. Chẳng hạn, chúng ta thường hỏi ý kiến người thân, bạn bè khi chọn mua một chiếc xe, máy tính, hoặc một điện thoại mới; Sinh viên thường xin ý kiến giáo sư, đồng nghiệp khi chọn một chủ đề nghiên cứu, chọn hội thảo để gửi bài, v.v... Thực chất, đó là quá trình yêu cầu bằng lời các khuyến nghị dựa trên những mối quan hệ xã hội, gọi tắt là khuyến nghị xã hội (social recommendation). Các dạng khuyến nghị xã hội như thế này diễn ra hàng ngày trong cuộc sống. Điều đó cho ta thấy những mối quan hệ xã hội đã chi phối, ảnh hưởng đến sở thích, hành vi, cũng như những quyết định của con người. Như ông bà ta thường nói “Gần mực thì đen, gần đèn thì sáng”. Thật không may, những cách tiếp cận truyền thống có "một lỗ hỏng" là chưa xem xét các mối quan hệ xã hội, cũng như ảnh hưởng của nó để thực hiện khuyến nghị cho người dùng. Trong vài năm trở lại đây, cùng với sự phát triển của web, các mạng xã hội (Social Network) đã ra đời và phát triển một cách nhanh chóng, thu hút nhiều quan tâm nghiên cứu của cộng đồng khoa học máy tính nhằm phát triển các phương pháp khuyến nghị thông minh hơn bằng cách kết hợp việc phân tích các mối quan hệ xã hội của người dùng vào quá trình khuyến nghị.

Phân tích mạng xã hội (Social Network Analysis) là phân tích định lượng những mối quan hệ giữa các cá nhân và tập thể trong mạng. Từ đó có thể đánh giá mức độ ảnh hưởng, cũng như chịu ảnh hưởng của cá nhân hay tập thể đó với cộng đồng xung quanh. Phân tích mạng xã hội được xem như một kỹ thuật chính yếu trong xã hội học hiện đại. Phân tích mạng xã hội đã và đang được dùng cho các nghiên cứu tiên tiến trong khoa học hành vi và khoa học xã hội. Trong một thập niên qua, nó đã và đang dần trở thành chủ đề phổ biến được đầu tư nghiên cứu trong lĩnh vực khoa học

máy tính. Các mối quan hệ đóng một vai trò rất quan trọng trong lan truyền, chia sẻ thông tin, tri thức. Thật khó có thể hiểu được các mối quan hệ cộng đồng xung quanh của một người có ảnh hưởng như thế nào đến hành vi, đặc điểm của người đó. Tác giả Kirchhoff và cộng sự đã nghiên cứu trình bày các độ đo trung tâm (Centrality Measures), dùng để đo mức độ quan trọng của các cá nhân trong mạng [62]. Phân tích mạng xã hội đã và đang được ứng dụng trong nhiều bài toán khác nhau như: tác giả Newman đã xây dựng mạng cộng tác khoa học và tính khoảng cách cộng tác giữa các nhà khoa học dựa trên đường đi ngắn nhất [89]; Trong một nghiên cứu khác, tác giả Newman ứng dụng phân tích mạng xã hội để rút trích các cấu trúc cộng đồng trong những mạng phức tạp [88]; Tác giả Balthrop và cộng sự ứng dụng phân tích mạng xã hội để khảo sát sự lây lan của virus máy tính [13]; Các tác giả Xu và Chen ứng dụng phân tích mạng xã hội để xác định những nhóm tội phạm, khủng bố [127]; Tác giả Kirchhoff và cộng sự nghiên cứu ứng dụng phân tích mạng xã hội để cải tiến các hệ thống truy vấn thông tin [62]; Tác giả Ma và cộng sự nghiên cứu đề xuất các phương pháp cải tiến hệ khuyến nghị dựa trên phân tích các mối quan hệ xã hội [79]; Tác giả Luong và cộng sự đã dựa trên tiếp cận khai thác mạng xã hội để phát triển các phương pháp khuyến nghị nơi gởi bài [76]. Tác giả Huynh và cộng sự đã phát triển phương pháp phân tích mạng trích dẫn cho khuyến nghị bài báo liên quan [55].

Một số nghiên cứu liên quan này cho chúng ta thấy các nghiên cứu về hệ khuyến nghị đã và đang được quan tâm thực hiện hơn một thập niên qua, trong nhiều lĩnh vực khác nhau. Tuy nhiên, các nghiên cứu khai thác thông tin quan hệ xã hội để cải tiến độ chính xác tiên đoán, thực hiện khuyến nghị thật sự thu hút nhiều nghiên cứu hơn từ khi có sự ra đời và phát triển của các mạng xã hội. Tức việc nghiên cứu, phát triển các phương pháp khuyến nghị dựa trên tiếp cận phân tích mạng xã hội đang ở những bước đi đầu tiên. Tiếp cận phân tích mạng xã hội giúp người dùng đưa ra những quyết định dựa trên tư vấn, đề xuất của những người có quan hệ. Đây là vấn đề rất tự nhiên trong cuộc sống. Trong lĩnh vực học thuật cũng vậy, các sinh viên, nghiên cứu viên thường dựa trên ý kiến đề xuất của giáo sư, đồng nghiệp, những người có kinh nghiệm để đưa ra những quyết định liên quan đến công việc nghiên cứu khoa học như: chọn hội thảo gởi bài, chọn người hợp tác, chọn bài báo để đọc, v.v... Đây chính là mục tiêu nghiên cứu của luận án. Phạm vi ứng dụng được chọn là lĩnh vực học thuật nhằm

hướng đến phục vụ cộng đồng nghiên cứu khoa học.

Mục tiêu và nội dung thực hiện của luận án

Với mục tiêu tập trung nghiên cứu phát triển các phương pháp khuyến nghị nhằm hỗ trợ nghiên cứu viên trong việc tìm kiếm thông tin học thuật dựa trên tiếp cận phân tích mạng xã hội, luận án đề ra các nội dung cụ thể như sau:

1. Xây dựng, làm giàu kho dữ liệu biên mục bài báo khoa học, chuyên ngành Khoa học Máy tính.
2. Mô hình và phân tích các mạng xã hội khoa học từ kho dữ liệu bài báo. Tập trung vào các mạng:
 - a. Mạng đồng tác giả
 - b. Mạng trích dẫn
 - c. Mạng cộng tác của các cơ quan
3. Nghiên cứu phát triển các phương pháp khuyến nghị dựa trên tiếp cận phân tích mạng xã hội, cụ thể là mạng xã hội học thuật nhằm cải tiến độ chính xác khuyến nghị. Tập trung vào giải quyết các bài toán:
 - a. Khuyến nghị cộng tác
 - b. Khuyến nghị bài báo khoa học

Đối tượng và phạm vi nghiên cứu

- Đối tượng: các bài báo khoa học dạng văn bản và thông tin biên mục của chúng.
- Phạm vi:

Lĩnh vực bài báo: Chuyên ngành Khoa học Máy tính.

Tiếp cận: dựa trên các đồ thị mạng xã hội học thuật kích thước lớn.

Các đóng góp chính của luận án

1. Đề xuất mô hình hóa các mạng xã hội học thuật nhận diện được từ kho dữ liệu học thuật, mô hình ASN [CT.6].
2. Bài toán khuyến nghị cộng tác cho nghiên cứu viên
 - Đối với nghiên cứu viên có quan hệ đồng tác giả: đề xuất, cải tiến các phương pháp phân tích xu hướng cộng tác trong mạng xã hội học thuật ASN để khuyến nghị các cộng tác viên tiềm năng. Các phương pháp đề xuất bao gồm: MPRS, MPRS+, RSS+ [CT.4, CT.1].
 - Đối với nghiên cứu viên chưa có quan hệ đồng tác giả: đề xuất tập đặc trưng để khuyến nghị những mối quan hệ cộng tác tốt, chất lượng [CT.3].
 - Đề xuất phương pháp đánh giá chất lượng cộng tác được khuyến nghị [CT.3].
3. Bài toán khuyến nghị bài báo khoa học: phát triển phương pháp khuyến nghị bài báo khoa học cho nghiên cứu viên dựa trên việc khai thác mạng trích dẫn, quan hệ lòng tin trong mô hình ASN [CT.2], [CT.8].
4. Xây dựng kho dữ liệu học thuật hơn 6 triệu bài báo và hệ thống tìm kiếm thông tin khoa học CSPubGuru (www.cspubguru.com) [CT.5, CT.7, CT.9, CT.14].

Sau quá trình nghiên cứu, thực hiện luận án, tác giả đã công bố được các công trình sau:

Tạp chí chuyên ngành

[CT.1] Tin Huynh, Kiem Hoang. New Methods for Calculating Trend- Based Vertex Similarity for Collaboration Recommendation. Journal of Computer Science and Cybernetics, vol.29, No.4, pages 338-350, (2013) (ISSN 1813-9663).

[CT.2] Huỳnh Ngọc Tín, Hoàng Kiếm. Khai thác xu hướng sở thích và quan hệ lòng tin để phát triển phương pháp khuyến nghị bài báo khoa học. Tạp chí Công nghệ thông tin và Truyền thông, Tập V-1, Số 13 (33), (2015) (ISSN 1859-3526).

Hội thảo chuyên ngành

-
- [CT.3] Tin Huynh, Atsuhiro Takasu, Tomonari Masada, Kiem Hoang. Collaborator Recommendation for Isolated Researchers. The Seventh International Symposium on Mining and Web (MAW2014) as a part of The 28th IEEE International Conference on Advanced Information Networking and Applications (AINA-2014), May 13-16, 2014, Victoria, Canada (2014). (Proceedings indexed by DBLP, EI, Scopus, and Thomson ISI. ERA Conference Ranking of AINA: B)
- [CT.4] Tin Huynh, Kiem Hoang, Dao Lam. Trend Based Vertex Similarity for Academic Collaboration Recommendation. 5th International Conference on Computational Collective Intelligence Technologies and Applications (ICCCI 2013), September 2013, Craiova, Romani, pages 11-20, (2013). (Proceedings Indexed by DBLP, EI, Scopus, ACM Digital Library, and Thomson ISI.ERA Conference Ranking: C)
- [CT.5] Tin Huynh, Kiem Hoang, Tien Do, Duc Huynh. Vietnamese Author Name Disambiguation for Integrating Publications from Heterogeneous Sources. The 5th Asian conference on Intelligent Information and Database Systems (ACIIDS 2013), Kuala Lumpur, Malaysia, pages 226-235, (2013). (Proceedings indexed by DBLP, EI, Scopus, and Thomson ISI)
- [CT.6] Tin Huynh, Kiem Hoang. Modeling Collaborative Knowledge of Publishing Activities for Research Recommendation. In Proceedings of the 4th International Conference on Computational Collective Intelligent Technologies and Applications (ICCCI 2012), November 2012, Ho Chi Minh City, VietNam, pages 28-30, (2012). (The proceedings indexed by DBLP, EI, Scopus, ACM Digital Library, and Thomson ISI.ERA Conference Ranking: C. Citation Count: 4 (không tính tự trích dẫn))
- [CT.7] Tin Huynh, Hiep Luong, and Kiem Hoang. Integrating bibliographical data of computer science publications from online digital libraries. In Proceedings of the 4th Asian conference on Intelligent Information and Database Systems (ACIIDS'12), Springer-Verlag, Berlin, Heidelberg, pages 226-235, (2012). (The proceedings indexed by DBLP, EI, and Thomson ISI. Citation Count: 1 (không tính tự trích dẫn))

-
- [CT.8] Tin Huynh, Hiep Luong, Kiem Hoang, Susan Gauch, Loc Do, Huong Tran. Scientific Publication Recommendations Based on Collaborative Citation Networks. In: Proceedings of the 3rd International Workshop on Adaptive Collaboration (AC 2012) as part of The 2012 International Conference on Collaboration Technologies and Systems (CTS 2012). Denver, Colorado, USA, pages 316-321, (2012). (The proceedings indexed by DBLP, EI, and Thomson ISI. ERA Conference Ranking: C. Citation count: 4 (không tính tự trích dẫn))
- [CT.9] Tin Huynh, Kiem Hoang. GATE framework based metadata extraction from scientific papers. In: Proceedings of the The International Conference on Education and Management Technology (ICEMT 2010), Cairo, Egypt, page 188 – 191, (2010). (The proceedings indexed by Google Scholar, IEEE Xplore Digital library, Citation count: 4, (không tính tự trích dẫn))
- [CT.10] Hung Nghiep Tran, Tin Huynh, Tien Do. Author Name Disambiguation by Using Deep Neural Network. In Proceedings of the 6th Asian conference on Intelligent Information and Database Systems, Bangkok, Thailand, April 2014 (ACIIDS'14). Springer-Verlag, Berlin, Heidelberg, pages 123-132, (2014). (The proceedings indexed by DBLP, EI, and Thomson ISI. Citation Count: 1 (không tính tự trích dẫn))
- [CT.11] Hung Nghiep Tran, Tin Huynh, Kiem Hoang. A Potential Approach to Overcome Data Limitation in Scientific Publication Recommendation. In Proceedings of the seventh international conference on knowledge and systems engineering (KSE-2015), TpHCM, Vietnam, Oct 8-10, 2015.
- [CT.12] Hiep Luong, Tin Huynh, Susan Gauch, Loc Do, and Kiem Hoang. Publication venue recommendation using author network's publication history. In Proceedings of the 4th Asian conference on Intelligent Information and Database Systems, Kaohsiung, Taiwan, March 2012 (ACIIDS'12). Springer-Verlag, Berlin, Heidelberg, pages 426-435, (2012). (The proceedings indexed by DBLP, EI, and Thomson ISI. Citation Count: 3 (không tính tự trích dẫn))
- [CT.13] Hiep Luong, Tin Huynh, Susan Gauch, Kiem Hoang. Exploiting Social Networks

for Publication Venue Recommendations. In Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval, Barcelona, Spain, pages 239 - 245, October (2012).

- [CT.14] Tien Do, Dao Lam, Tin Huynh. A Framework for integrating bibliographical data of computer science publications. 2014 International Conference on Computing, Management and Telecommunications (ComManTel 2014), 27-29 April 2014, Da nang, Viet Nam, pages 245-250, (2014).

Bố cục của luận án

Luận án được bố cục gồm các chương mục như sau:

- **Mở đầu:** Giới thiệu tóm tắt về động cơ, mục tiêu, nội dung nghiên cứu, các đóng góp chính và bối cảnh chung của luận án.
- **Chương 1:** Giới thiệu hệ khuyến nghị, phân tích ưu điểm hạn chế của các phương pháp phổ biến và các nghiên cứu liên quan.
- **Chương 2:** Tiếp cận luận án khai thác các mạng xã hội học thuật cho khuyến nghị. Chương này trình bày giải pháp cho rút trích, làm giàu kho dữ liệu bài báo khoa học. Từ kho dữ liệu bài báo, các mạng xã hội học thuật được xác định và mô hình hóa. Một phần của chương này được trình bày trong các công trình [CT5], [CT6], [CT7], [CT9].
- **Chương 3:** Trình bày, phát biểu bài toán khuyến nghị cộng tác, các phương pháp phổ biến. Để phát triển các phương pháp mới cho khuyến nghị cộng tác, luận án đã phân chia các nghiên cứu viên thành các nhóm khác nhau: (1) Nghiên cứu viên có đồng tác giả; (2) Nghiên cứu viên chưa có đồng tác giả, giải quyết trường hợp khởi động lạnh trong khuyến nghị. Các phương pháp đề xuất dựa trên phân tích các mối quan hệ xã hội rõ ràng, tiềm ẩn trong lĩnh vực học thuật có sử dụng yếu tố thời gian, xu hướng. Đồng thời, luận án cũng đề xuất tập đặc trưng cho những nghiên cứu viên cộng tác tiềm năng để học mô hình khuyến nghị. Một phần của chương này đã được trình bày trong các công trình [CT1], [CT3], [CT4].

-
- **Chương 4:** Trình bày bài toán khuyến nghị bài báo khoa học, các nghiên cứu liên quan và những phương pháp đề xuất của luận án. Tiếp cận chính của luận án là dựa trên việc khai thác các mối quan hệ từ các mạng xã hội học thuật (mô hình ASN). Điểm khác biệt trong luận án, là việc tập trung khai thác các mối quan hệ tiềm ẩn, quan hệ lòng tin kết hợp với yếu tố xu hướng sở thích. Các kết quả nghiên cứu liên quan đến bài toán ứng dụng trong chương này đã được trình bày, công bố trong công trình [CT2], [CT8].

- **Kết luận và Hướng phát triển**
- **Danh mục công trình của tác giả**
- **Mục Tài liệu Tham khảo**
- **Phụ lục A:** Trình bày chi tiết các luật JAPE để rút trích và tích hợp dữ liệu bài báo khoa học từ nhiều nguồn không đồng nhất.
- **Phụ lục B:** Thông tin chi tiết về cấu trúc kho dữ liệu học thuật đã xây dựng để phục vụ nghiên cứu, thực nghiệm.

Chương 1

TỔNG QUAN: HỆ KHUYẾN NGHỊ, NHỮNG PHƯƠNG PHÁP TIẾP CẬN PHỐ BIỀN VÀ XU HƯỚNG

1.1 Giới thiệu

Dựa trên kết quả khảo sát, chương này sẽ phát biểu lại một cách hình thức bài toán khuyến nghị trong trường hợp tổng quát, tập trung trình bày và phân tích ưu điểm, hạn chế của những phương pháp tiếp cận truyền thống cũng như xu hướng mới cho hệ khuyến nghị.

1.2 Khái niệm Hệ khuyến nghị

Hệ khuyến nghị, tiếng anh là Recommender Systems hoặc Recommendation System, là những hệ thống được thiết kế để hướng người dùng đến những đối tượng quan tâm, yêu thích, khi lượng thông tin quá lớn vượt quá khả năng xử lý của người dùng [99, 25].

Theo Ricci và cộng sự [100], hệ khuyến nghị là những công cụ phần mềm, kỹ thuật cung cấp những đề xuất các đối tượng có thể hữu ích với người dùng. Những đề xuất liên quan đến quyết định của người dùng như: sản phẩm nào nên mua, bài hát nào nên nghe, hay tin tức nào nên đọc. Tác giả Gunawardana và Shani thì cho rằng rất khó có thể đưa ra một định nghĩa cho hệ khuyến nghị, bởi vì những hệ thống với nhiều mục tiêu và hành vi khác nhau được gom nhóm lại và đặt tên là hệ khuyến nghị [49]. Tác

giả đã phân loại hệ khuyến nghị thành nhiều nhóm khác nhau dựa trên công việc mà hệ thống thực hiện (hình 1.1).



Hình 1.1: Phân loại hệ khuyến nghị dựa trên công việc khuyến nghị

Chúng ta có thể hiểu hệ khuyến nghị là những hệ thống, công cụ, kỹ thuật, được thiết kế để hướng người dùng đến những đối tượng quan tâm, yêu thích, khi lượng thông tin quá lớn vượt quá khả năng xử lý của người dùng. Khi tích hợp vào các hệ thống thương mại điện tử cũng như các hệ thống tìm kiếm, hệ khuyến nghị sẽ giúp người dùng dễ dàng hơn trong quá trình tìm kiếm thông tin liên quan, giúp thông tin liên quan tự động tìm đến người dùng thay vì người dùng phải vất vả tự đi tìm kiếm các thông tin liên quan. Hệ khuyến nghị cũng có thể xem là một trong những giải pháp hỗ trợ tìm kiếm thông minh bằng cách cố gắng hiểu sở thích của người dùng.

Tóm lại, luận án quan niệm hệ khuyến nghị là những hệ thống, công cụ, kỹ thuật thông minh tìm cách hiểu sở thích của người dùng và giúp thông tin liên quan tự động tìm đến người dùng.

1.3 Phát biểu Bài toán Khuyến nghị

Hiện nay, nhiều công trình nghiên cứu phổ biến đã trình bày các khái niệm cơ bản, định nghĩa và phát biểu cho bài toán khuyến nghị. Các nghiên cứu điển hình có thể kể đến như: Jannach và cộng sự [57], Adomavicius và Tuzhilin [5], Stefanidis và cộng

sự [109], Bobadilla và cộng sự [22]. Dựa trên các nghiên cứu liên quan, phần này sẽ hệ thống lại một số khái niệm, định nghĩa và phát biểu hình thức cho bài toán khuyến nghị.

Định nghĩa 1.1: *Không gian người dùng* [57]

Không gian người dùng là tập tất cả những người dùng mà hệ thống quan sát được, để thực hiện các phân tích, khuyến nghị. Ký hiệu là U , $U = \{u_1, u_2, u_3, \dots, u_n\}$.

Định nghĩa 1.2: *Không gian đối tượng khuyến nghị* [57]

Không gian đối tượng khuyến nghị là tập tất cả những đối tượng sẽ được khuyến nghị cho người dùng. Tùy vào ứng dụng cụ thể, các đối tượng khuyến nghị có thể là sách, báo, phim ảnh, địa điểm, nhà hàng, khách sạn, con người, v.v... Ký hiệu là P , $P = \{p_1, p_2, p_3, \dots, p_m\}$.

Định nghĩa 1.3: *Hàm hữu ích* [5]

Hàm hữu ích f là ánh xạ $f : U \times P \rightarrow \mathbb{R}$, dùng để ước lượng mức độ hữu ích của $p \in P$ với $u \in U$. Với \mathbb{R} là tập có thứ tự các số nguyên hoặc thực trong một khoảng nhất định.

Phát biểu bài toán khuyến nghị

Cho trước,

- $U = \{u_1, u_2, u_3, \dots, u_n\}$: không gian người dùng.
- $P = \{p_1, p_2, p_3, \dots, p_m\}$: không gian đối tượng khuyến nghị.

Mục đích của hệ khuyến nghị là đi tìm hàm hữu ích f , ước lượng giá trị của $f(u, p)$ (với $u \in U, p \in P$). Giá trị của $f(u, p)$ giúp tiên đoán u sẽ thích p nhiều hay ít, hay p hữu ích đối với u như thế nào. Đối với mỗi người dùng $u \in U$, hệ khuyến nghị cần chọn $TopN$ đối tượng $p \in P$ hữu ích nhất đối với người dùng u để khuyến nghị, $P_{TopN} = \langle p_{Top1}, p_{Top2}, \dots, p_{TopN} \rangle$, (với $TopN \ll m$). Việc chọn $TopN$ bao nhiêu là tùy thuộc vào nhu cầu thông tin của người dùng, cũng như mục đích cung cấp thông tin của hệ khuyến nghị. Các đối tượng $p \in P_{TopN}$, được chọn thỏa mãn các điều kiện ràng buộc sau:

- i) $\forall p_k \in P_{TopN}, f(u, p_k) \geq f(u, p_{k+1})$, với $1 \leq k \leq TopN - 1$. Tức là tập các đối tượng khuyến nghị P_{TopN} là tập có thứ tự. Đối tượng đứng trước có giá trị của

hàm hữu ích f lớn hơn hoặc bằng đối tượng đứng sau, hay đối tượng đứng trước ưu tiên khuyến nghị cho u hơn đối tượng đứng sau.

- ii) $\forall p_k \in P_{TopN}, \forall p_i \in P \setminus P_{TopN}$, thì $f(u, p_k) \geq f(u, p_i)$. Tức giá trị hữu ích của các đối tượng được khuyến nghị, được xác định thông qua hàm f , phải lớn hơn hoặc bằng những đối tượng không được khuyến nghị.

Việc xây dựng hàm hữu ích f và ước lượng giá trị hữu ích của các đối tượng khuyến nghị $p \in P$ với những người dùng $u \in U$ có thể thực hiện bằng nhiều phương pháp khác nhau như: dựa vào kinh nghiệm (heuristics), máy học, lý thuyết xấp xỉ, v.v...

Phân tiếp theo sẽ trình bày chi tiết, phân tích về những tiếp cận khuyến nghị phổ biến hiện nay, cũng như các nghiên cứu liên quan và xu hướng trên thế giới.

1.4 Các cách tiếp cận phổ biến

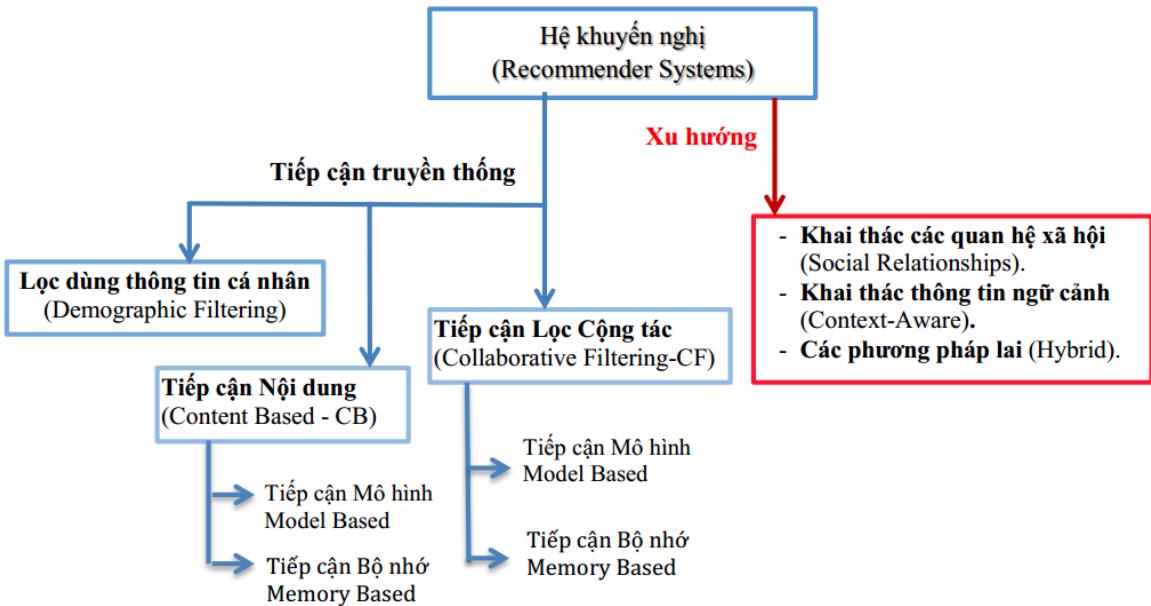
Theo Adomavicius và Tuzhilin [5], Bobadilla và cộng sự [22], các phương pháp khuyến nghị truyền thống được phân loại dựa trên cách thức mà nó thực hiện khuyến nghị. Nhìn chung, các phương pháp truyền thống có thể phân thành các nhóm như: (1) Lọc dùng thông tin cá nhân (Demographic Filtering): dùng thông tin cá nhân như tuổi, giới tính, trình độ, v.v... để xác định những nhóm người dùng nào sê thích cái gì; (2) Tiếp cận nội dung (Content-Based Filtering), gọi tắt là CB; (3) Tiếp cận lọc cộng tác (Collaborative Filtering), gọi tắt là CF; và (4) Tiếp cận lai (Hybrid Approach).

Bên cạnh đó, tác giả Bobadilla và cộng sự cũng đã khảo sát và chỉ ra xu hướng hiện nay cho hệ khuyến nghị [22]. Hình 1.2 thể hiện tóm tắt các cách tiếp cận truyền thống, phổ biến cũng như xu hướng hiện nay cho hệ khuyến nghị. Phân tiếp theo sẽ trình bày chi tiết, cũng như phân tích ưu điểm, hạn chế của một số tiếp cận chính trong phạm vi luận án.

1.4.1 Tiếp cận nội dung (CB)

Định nghĩa 1.4: *Hồ sơ người dùng*

Hồ sơ người dùng u , ký hiệu là $UserProfile(u)$, biểu diễn sở thích của u và giúp hệ khuyến nghị tiên đoán một đối tượng $p \in P$ có hữu ích hay không và mức độ hữu ích đối với u là như thế nào. $UserProfile(u)$ có thể xây dựng từ việc phân tích đặc



Hình 1.2: Các cách tiếp cận phổ biến và xu hướng hiện nay cho hệ khuyến nghị

trưng các đối tượng khuyến nghị mà u quan tâm, đánh giá trong quá khứ thông qua tương tác với hệ thống.

Tiếp cận nội dung có nguồn gốc từ cộng đồng nghiên cứu về truy vấn thông tin và lọc thông tin [16, 11]. Tiếp cận nội dung tìm cách khuyến nghị các đối tượng tương tự với những đối tượng mà người dùng quan tâm trong quá khứ. Chẳng hạn, nếu người dùng thường đọc các trang tin tức có chứa những từ như: Grand Slam, Roger Federer, Nadal, Djokovic, Australia Open, French Open, Wimbledon, US Open, v.v... thì các trang chứa tin tức liên quan đến chủ đề tennis mà người dùng chưa biết sẽ được khuyến nghị cho người dùng. Một ví dụ khác, nếu một nghiên cứu viên thường tải, đọc hay viết những bài báo khoa học về IR thì các bài báo khoa học về IR mới, cập nhật mà nghiên cứu viên đó chưa biết sẽ được ưu tiên khuyến nghị.

Để ước lượng có hay không người dùng u sẽ thích đối tượng khuyến nghị p và thích nhiều hay ít (tức việc xây dựng và ước lượng giá trị hàm hữu ích $f(u, p)$), các phương pháp dựa trên tiếp cận nội dung thông thường sẽ thực hiện các bước sau:

- Bước 1: Biểu diễn nội dung đối tượng khuyến nghị $p \in P$, ký hiệu ($Content(p)$).
- Bước 2: Mô hình hóa sở thích người dùng $u \in U$, gọi tắt là hồ sơ người dùng

(User Profile), ký hiệu ($UserProfile(u)$).

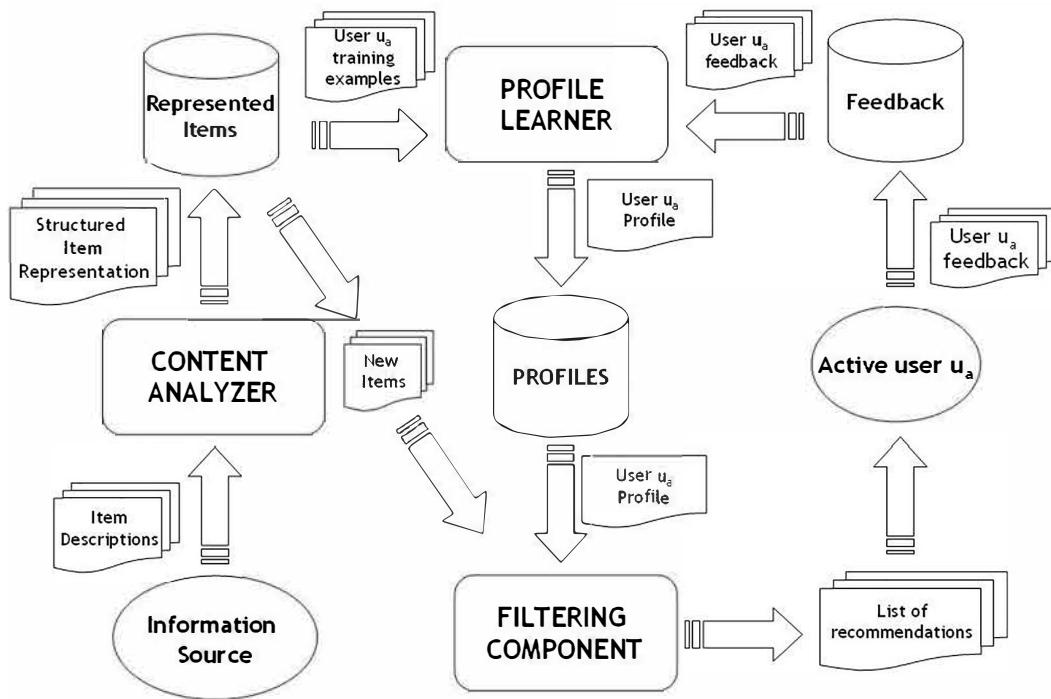
- Bước 3: Ước lượng giá trị hữu ích dựa trên độ tương tự nội dung của đối tượng khuyến nghị p với hồ sơ người dùng u . Hệ thống sẽ ưu tiên khuyến nghị những đối tượng có nội dung tương tự cao so với hồ sơ người dùng u .

$$f(u, p) = Sim(UserProfile(u), Content(p)) \quad (1.1)$$

1.4.1.1 Kiến trúc hệ thống

Pasquale Lops và cộng sự đã tiến hành khảo sát, phân tích các hệ khuyến nghị dựa trên tiếp cận nội dung [73]. Theo Pasquale Lops và cộng sự, hệ khuyến nghị dựa trên tiếp cận nội dung thường sẽ thực hiện 3 việc chính và được đảm trách bởi ba thành phần tương ứng đó là: (1) Phân tích nội dung (Content Analyzer): có nhiệm vụ phân tích, mô hình hóa nội dung của các đối tượng khuyến nghị; (2) Mô hình hóa hồ sơ người dùng (Profile Learner): xây dựng và cập nhật hồ sơ người dùng dựa trên đặc trưng các đối tượng mà người dùng quan tâm, yêu thích; (3) Lọc nội dung (Filtering Component): thực hiện khuyến nghị dựa trên việc so khớp đặc trưng nội dung của đối tượng khuyến nghị với hồ sơ người dùng. Kiến trúc tổng quan của hệ khuyến nghị dựa trên tiếp cận nội dung thể hiện qua hình vẽ 1.3.

- **Phân tích nội dung (Content Analyzer):** thành phần này có nhiệm vụ phân tích và mô hình hóa nội dung của các đối tượng khuyến nghị. Tùy vào bài toán cụ thể, các phương pháp rút trích đặc trưng sẽ được dùng để chuyển nội dung đối tượng khuyến nghị từ định dạng gốc sang không gian đặc trưng. Trường hợp dữ liệu của đối tượng khuyến nghị là không cấu trúc (ví dụ tài liệu văn bản) thì các bước tiền xử lý như loại bỏ hư từ (stop word), chuyển về gốc từ (stemming) sẽ được tiến hành trước khi rút trích đặc trưng và mô hình hóa thành những vectơ từ khóa. Mô hình biểu diễn đối tượng khuyến nghị là đầu vào cho bước học, mô hình hóa hồ sơ người dùng và bước so khớp để thực hiện khuyến nghị.
- **Mô hình hóa hồ sơ người dùng (Profile Learner):** các nghiên cứu thường dùng các phương pháp học máy giám sát để học hồ sơ người dùng dựa trên đặc trưng của các đối tượng mà người dùng thích hay không thích trong quá khứ. Qua thời gian sở thích người dùng có thể thay đổi. Dựa trên dữ liệu phản



Hình 1.3: Kiến trúc tổng quan của hệ khuyến nghị dựa trên tiếp cận nội dung
(Nguồn hình vẽ: [73])

hồi rõ ràng hay tiềm ẩn của người dùng thông qua tương tác với hệ thống, hệ thống thường sẽ định kỳ để học và cập nhật lại hồ sơ người dùng.

- **Lọc nội dung (Filtering Component):** thành phần này có nhiệm vụ so khớp hồ sơ người dùng với nội dung của các đối tượng để thực hiện khuyến nghị những đối tượng phù hợp với sở thích người dùng. Kết quả so khớp sẽ thể hiện mức độ quan tâm của người dùng $u \in U$ lên đối tượng khuyến nghị $p \in P$. Nói cách khác, giá trị hàm hữu ích $f(u, p)$ của sản phẩm p với người dùng u được ước lượng dựa trên độ tương tự nội dung của đối tượng khuyến nghị $p \in P$ với nội dung các đối tượng $p' \in P$, $\{p'\}$ là tập các đối tượng liên quan đến u hay được u quan tâm trong quá khứ.

Một trong những vấn đề ảnh hưởng đến hiệu năng của tiếp cận nội dung là kỹ thuật phân tích nội dung và phương pháp mô hình hóa hồ sơ người dùng.

1.4.1.2 Xây dựng và cập nhật hồ sơ người dùng

Hồ sơ người dùng giúp hệ thống có thể hiểu được sở thích của người dùng và tiên đoán một đối tượng có hữu ích hay không và mức độ hữu ích đối với mỗi người dùng là như thế nào. Có thể nói hồ sơ người dùng là yếu tố then chốt quyết định hiệu quả của các hệ khuyến nghị dựa trên nội dung. Vấn đề đặt ra là làm thế nào để có thể ghi nhận thông tin sở thích của người dùng và làm thế nào để mô hình hóa, cập nhật hồ sơ người dùng trong các hệ khuyến nghị dựa trên nội dung? Phần tiếp theo sẽ giúp chúng ta trả lời những câu hỏi này.

(*) Thông tin phản hồi của người dùng

Thông qua những phản hồi thì hệ thống có thể biết được những đối tượng nào được người dùng quan tâm và mức độ là nhiều hay ít. Susan Gauch và đồng nghiệp [45], cũng như Pasquale Lops và đồng nghiệp [73], đã phân thông tin phản hồi của người dùng thành hai loại: rõ ràng và tiềm ẩn khi người dùng tương tác với hệ thống. Những hình thức phản hồi rõ ràng của người dùng như: nhập trực tiếp vào hệ thống những từ khóa thể hiện sở thích, nhấn chọn thích hay không hoặc cho những điểm đánh giá trong một khoảng nào đó (thường từ 1 đến 5), đưa ra những bình luận đối với những đối tượng mà hệ thống khuyến nghị. Tuy nhiên, trên thực tế chỉ một số lượng rất ít người dùng chia sẻ những thông tin, quan điểm của họ về những đối tượng khuyến nghị khi sử dụng và tương tác với hệ thống. Vì vậy, nhiều hệ thống đã tìm cách ghi nhận thông tin phản hồi tiềm ẩn bằng việc phân tích hành vi sử dụng hệ thống của người dùng thông qua bộ nhớ đệm của trình duyệt, tập tin log, v.v... Những phản hồi tiềm ẩn có thể kể đến như: chọn xem, đánh dấu và lưu trang, thời gian xem, v.v...

Thông thường, hệ thống sẽ mô hình hóa hồ sơ người dùng dựa trên thông tin phản hồi và nội dung đối tượng. Nội dung của đối tượng khuyến nghị thường được biểu diễn bởi một tập các đặc trưng. Chẳng hạn, đối tượng là bài báo khoa học thì có thể biểu diễn bởi một số đặc trưng cơ bản như: tác giả, hội thảo, tạp chí, từ khóa thể hiện chủ đề bài báo, v.v... Tùy vào bài toán cụ thể thì các phương pháp rút trích đặc trưng sẽ được dùng để chuyển nội dung đối tượng khuyến nghị từ định dạng dữ liệu gốc sang không gian đặc trưng.

(*) Mô hình hóa hồ sơ người dùng

Hầu hết các hệ khuyến nghị nội dung áp dụng mô hình truy vấn đơn giản như so khớp từ khóa hoặc mô hình không gian vectơ. Đặc trưng của đối tượng thường là các đặc trưng dạng văn bản được rút trích từ các trang web, nội dung bài báo, thông tin mô tả sản phẩm. Trường hợp dữ liệu của đối tượng khuyến nghị là không cấu trúc, chẳng hạn tài liệu văn bản, thì các bước tiền xử lý như loại bỏ hư từ (stop word), chuyển về gốc từ (stemming) sẽ được tiến hành trước khi rút trích đặc trưng và mô hình hóa thành những vectơ từ khóa. Mô hình biểu diễn nội dung đối tượng khuyến nghị là đầu vào cho bước học hồ sơ người dùng và bước so khớp để thực hiện khuyến nghị.

Với mô hình không gian vectơ thì nội dung của đối tượng khuyến nghị $p \in P$, ký hiệu là $Content(p)$, được biểu diễn dưới dạng một vectơ đặc trưng như sau:

$$Content(p) = \vec{w}_p = (w_{1,p}, w_{2,p}, \dots, w_{k,p}) \quad (1.2)$$

Trong đó,

- k : là tổng số đặc trưng dùng để biểu diễn nội dung đối tượng. Đơn giản nhất là từ điển các từ khóa sau khi loại bỏ các stop word và thực hiện stemming.
- $w_{i,p}$ trọng số đặc trưng thứ i của đối tượng p .

Trọng số mỗi chiều trong vectơ \vec{w}_u và \vec{w}_p có thể ước lượng dựa trên tần suất xuất hiện của từ khóa bằng phương pháp TFIDF [9].

Hồ sơ người dùng thường được xây dựng dựa trên nội dung của các đối tượng mà họ thể hiện sự quan tâm, đánh giá khi tương tác, sử dụng hệ thống. Như vậy, với k đặc trưng biểu diễn nội dung các đối tượng khuyến nghị, hồ sơ người dùng u có thể biểu diễn dưới dạng một vectơ đặc trưng cũng với số chiều là k như sau:

$$UserProfile(u) = \vec{w}_u = (w_{1,u}, w_{2,u}, \dots, w_{k,u}) \quad (1.3)$$

Trong đó,

- $w_{i,u}$ trọng số đặc trưng thứ i trong hồ sơ người dùng u .

Việc ước lượng giá trị hàm hữu ích thông thường có thể dùng độ đo cosine trong

truy vấn thông tin [9].

$$f(u, p) = \text{Cosine}(\vec{w}_u, \vec{w}_p) = \frac{\vec{w}_u \bullet \vec{w}_p}{\| \vec{w}_u \| * \| \vec{w}_p \|} \quad (1.4)$$

Trong đó: dấu \bullet thể hiện tích hai vectơ, dấu $*$ thể hiện tích vô hướng và $\| . \|$ là độ dài của vectơ.

Nhiều hệ thống khuyến nghị nội dung dựa trên từ khóa đã được nghiên cứu và phát triển trong nhiều lĩnh vực ứng dụng khác nhau như: khuyến nghị phim, khuyến nghị web, khuyến nghị tin tức, v.v... Trong tài liệu [73], Pasquale Lops và cộng sự cũng đã tiến hành khảo sát và phân tích các hệ thống khuyến nghị tin tức. Đối với các hệ thống này thì hồ sơ người dùng sẽ được học dựa trên nội dung các trang mà người dùng phản hồi quan tâm hay không quan tâm. Một số nghiên cứu tìm cách mô hình hóa sở thích dài hạn của người dùng như hệ thống khuyến nghị tin NewsT [107], YourNews [7]. Bên cạnh đó, một số nghiên cứu khác như Daily Learner [21], NewsDude [20], thì xây dựng hai mô hình sở thích cho mỗi người dùng: mô hình sở thích dài hạn và mô hình sở thích ngắn hạn.

Trong lĩnh vực khuyến nghị trang web, tác giả Henry Lieberman đã đề xuất một hệ thống Letizia, hỗ trợ người dùng duyệt web [69]. Letizia có thể làm việc với các trình duyệt để lưu vết hành vi duyệt web của người dùng. Hệ thống sẽ xây dựng hồ sơ người dùng dựa trên các từ khóa rút trích từ những trang mà người dùng quan tâm. Henry Lieberman xem xét sở thích của người dùng thông qua các phản hồi tiềm ẩn, chẳng hạn hành vi lưu, đánh dấu một trang. Tương tự vậy, Dunja Mladenic đã nghiên cứu phát triển hệ thống Personal WebWatcher nhằm hỗ trợ người dùng duyệt web. Personal WebWatcher sẽ làm nổi bậc các liên kết tiềm năng trong các trang web mà người dùng duyệt qua. Tác giả đã dùng phương pháp học máy giám sát để học sở thích người dùng dựa trên nội dung các liên kết mà người quan tâm (nhấn chuột) và không quan tâm [85].

Nói chung, hầu hết những hệ thống khuyến nghị nội dung thực hiện mô hình hóa nội dung đối tượng dựa trên mô hình không gian vectơ với đặc trưng từ khóa và học mô hình người dùng dựa trên những phản hồi, tương tác rõ ràng hay tiềm ẩn của người dùng với hệ thống. Việc dùng đặc trưng từ khóa để biểu diễn nội dung đối tượng và xây dựng hồ sơ người dùng thường gặp phải một số vấn đề khó khăn liên quan đến

xử lý ngôn ngữ tự nhiên như: khác âm đồng nghĩa (synonymy), đồng âm khác nghĩa (polysemy). Để giải quyết những hạn chế liên quan đến việc mô hình hóa nội dung đối tượng dựa trên từ khóa, một số nghiên cứu khác quan tâm đến việc phát triển các phương pháp biểu diễn nội dung đối tượng và hồ sơ người dùng trên mô hình mạng ngữ nghĩa hoặc đặc trưng khái niệm thay vì đặc trưng từ khóa.

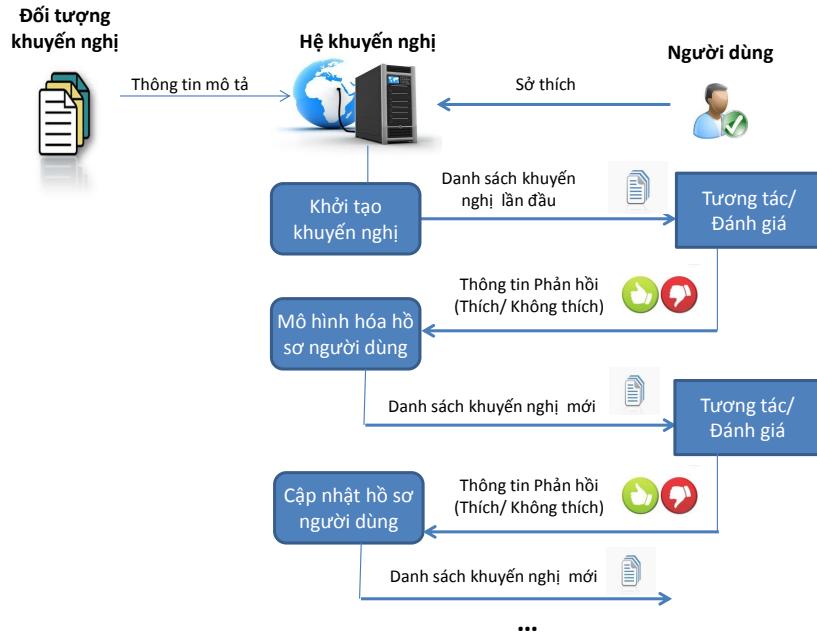
(*) Cập nhật hồ sơ người dùng

Trên thực tế, sở thích của người dùng thường sẽ thay đổi theo thời gian. Tùy vào lĩnh vực ứng dụng, mà sở thích người dùng sẽ thay đổi nhanh hay chậm. Chẳng hạn, trong lĩnh vực khuyến nghị phim, sách hay bài báo khoa học thường thì sở thích người dùng sẽ thay đổi chậm hơn so với lĩnh vực khuyến nghị tin tức. Trong khuyến nghị tin tức, đôi khi người dùng cần đọc những tin quan trọng, "nóng" mà không thuộc chủ đề họ quan tâm (tức không liên quan đến sở thích người dùng). Do đó, vấn đề thay đổi sở thích của người dùng là một trong những khó khăn, thách thức ảnh hưởng đến việc xây dựng và cập nhật hồ sơ người dùng trong các hệ khuyến nghị dựa trên nội dung.

Để đương đầu với sự thay đổi sở thích của người dùng, nhiều nghiên cứu đã đề xuất các giải pháp khác nhau cho việc xây dựng và cập nhật hồ sơ người dùng. Một số nghiên cứu liên quan đến khuyến nghị tin tức tìm cách mô hình hóa hồ sơ người dùng thành hai phần: sở thích dài hạn (thay đổi chậm) và sở thích ngắn hạn (thay đổi nhanh) như: Daily Learner [21], NewsDude [20]. Saranya.K.G và Sadhasivam đề xuất hai loại hồ sơ người dùng là: tĩnh (bao gồm thông tin do người dùng đăng ký) và động được xây dựng dựa trên thông tin tiềm ẩn mà người dùng tương tác với hệ thống. Trong lĩnh vực E-Learning, Nguyen và đồng nghiệp [90], Le và đồng nghiệp [67], đã nghiên cứu phát triển phương pháp xây dựng và cập nhật hồ sơ người dùng dựa trên luật, ứng dụng vào hệ khuyến nghị tài nguyên, dịch vụ dạy và học trong E-Learning. Các tác giả đã đề xuất mô hình α -Community để xây dựng và cập nhật hồ sơ người học dựa trên luật. Mô hình α -Community dựa trên lý thuyết tập thô và ý tưởng cơ bản là giá trị đặc trưng của hồ sơ người học sẽ được suy diễn dựa trên hồ sơ của những thành viên trong cùng nhóm học tập, cộng đồng.

Tóm lại, mỗi bài toán, lĩnh vực ứng dụng sẽ có phương pháp phù hợp để xây dựng và cập nhật hồ sơ người dùng. Với những lĩnh vực mà sở thích người dùng thay đổi

nhanh thì có thể chia hồ sơ người dùng thành 2 phần: tĩnh (dài hạn) và động (ngắn hạn). Thông thường, các hệ khuyến nghị sẽ định thời để cập nhật lại hồ sơ người dùng qua một khoảng thời gian. Hình 1.4 minh họa các bước cơ bản liên quan việc học và cập nhật hồ sơ người dùng.



Hình 1.4: Học và cập nhật hồ sơ người dùng dựa trên thông tin phản hồi

1.4.1.3 Phân loại tiếp cận nội dung

Các phương pháp truyền thống dựa trên nội dung có thể chia thành hai nhóm chính: (1) Một là các phương pháp dựa trên bộ nhớ, thực hiện tính toán độ tương tự và khoảng cách giữa $Content(p)$ và $UserProfile(u)$ dùng các độ đo Cosine, Euclidean [9]; (2) Hai là các phương pháp dựa trên mô hình, với mô hình được học từ dữ liệu dùng các kỹ thuật học máy để phân các đối tượng khuyến nghị thành những đối tượng người dùng quan tâm (1) hay không quan tâm (0) như: phân lớp SVM [60], phân lớp Bayesian [92] và các phương pháp xác suất như: Pazzani và Billsus [94], Mooney và Roy [86], Gemmis và đồng nghiệp [38]. Chẳng hạn phân lớp Bayesian có thể dùng để ước lượng xác suất một tài liệu (đối tượng khuyến nghị p) thuộc lớp C_x (quan tâm và không quan tâm), khi cho trước một tập các từ khóa mô tả tài liệu p này là $\{k_{1,p}, k_{2,p}, \dots, k_{n,p}\}$, chính là $P(C_x | k_{1,p}, k_{2,p}, \dots, k_{n,p})$ [92].

(a) Tiếp cận nội dung dựa trên bộ nhớ

Tiếp cận nội dung dựa trên bộ nhớ thường thực hiện việc ước lượng mức độ hữu ích của đối tượng khuyến nghị $p \in P$ với người dùng $u \in U$ (tức giá trị hàm hữu ích $f(u, p)$) dựa trên việc tổng hợp mức độ quan tâm của u đối với tập k đối tượng có nội dung tương tự với p , ký hiệu $P_k = \{p_k\}, P_k \subseteq P$, hoặc tổng hợp mức độ quan tâm từ tập k những người dùng có sở thích tương tự u , $U_k = \{u_k\}, U_k \subseteq U$. Tùy thuộc vào cách biểu diễn nội dung đối tượng dữ liệu và hồ sơ người dùng, chúng ta sẽ có một hàm phù hợp để tính độ tương tự và xác định tập P_k , cũng như U_k . Thông thường, các nghiên cứu dùng mô hình không gian vectơ và độ đo Cosine để biểu diễn nội dung và tính độ tương tự giữa các đối tượng.

Phương pháp dựa trên bộ nhớ được nhiều nghiên cứu sử dụng là phương pháp lân cận gần nhất, kNN. Để phát triển hệ thống khuyến nghị tin tức, Daniel Billsus và Michael J. Pazzani đã dùng phương pháp kNN xây dựng hồ sơ người dùng ngắn hạn kết hợp với hồ sơ người dùng dài hạn được xây dựng dùng phương pháp Bayesian [21]. Các tác giả đã chỉ ra rằng, việc chia hồ sơ người dùng thành hai phần giúp biểu diễn đa dạng hơn sở thích của người dùng. Trong nghiên cứu của họ, kNN được dùng để xác định chuỗi tất cả các tin tức liên quan đến một sự kiện nào đó dựa trên độ tương tự nội dung với một số tin mà người dùng quan tâm, đánh giá. Việc mô hình sở thích ngắn hạn dùng kNN sẽ dễ dàng thích nghi với những quan tâm mới của người dùng trong một khoảng thời gian ngắn, thay vì hệ thống phải cần rất nhiều dữ liệu huấn luyện để xây dựng lại toàn hồ sơ người dùng sử dụng các phương pháp học máy. Việc xác định người dùng tương tự dựa trên ma trận đánh giá thường gặp phải một số vấn đề ảnh hưởng đến độ chính xác như: ma trận đánh giá thừa, một số người dùng có chung đánh giá cho cùng đối tượng nhưng vì những lý do khác nhau. Do đó, Maria Terzi và cộng sự đã đề xuất dùng kNN để xác định nhóm những người dùng tương tự dựa trên nội dung bình luận về các đối tượng khuyến nghị (Text-based User-kNN) thay vì dựa trên sự tương quan trong ma trận đánh giá [120]. Các tác giả đã tiến hành thực nghiệm trên tập dữ liệu phim từ RottenTomatoes¹ và tập đĩa nhạc từ Amazon². Kết quả thực nghiệm cho thấy phương pháp tương tự người dùng dựa trên nội dung

¹<http://www.rottentomatoes.com/>

²<https://www.amazon.com/>

bình luận cho kết quả tốt hơn việc xác định người dùng dựa trên ma trận đánh giá. Tương tự, các tác giả Li Chen và Feng Wang [30], Claudiu Cristian Musat và cộng sự [87], cũng tiến hành xây dựng hồ sơ người dùng dựa trên nội dung văn bản rút trích từ các đối tượng mà người dùng quan tâm, đánh giá để xác định nhóm những người dùng tương tự thay vì dựa trên ma trận đánh giá. Nhìn chung, các phương pháp dựa trên bộ nhớ có những ưu điểm và hạn chế sau:

Ưu điểm:

- Chất lượng khuyến nghị thường sẽ tốt hơn do tính toán trên cả tập dữ liệu khi thực hiện khuyến nghị.
- Đơn giản, dễ hiện thực.

Hạn chế:

- Tốn bộ nhớ và tốc độ xử lý chậm do phải tính toán trên cả tập dữ liệu khi thực hiện khuyến nghị.
- Không thể tổng quát hóa tập dữ liệu.

(b) Tiếp cận nội dung dựa trên mô hình

Với các phương pháp dựa trên bộ nhớ, hệ thống thường sẽ tính giá trị hàm hữu ích dựa trên các độ đo như Cosine, Euclidean. Đối với các phương pháp dựa trên mô hình, một mô hình sẽ được huấn luyện từ dữ liệu để phân các đối tượng khuyến nghị thành những đối tượng được người dùng quan tâm (1) hay không quan tâm (0) và quan tâm nhiều hay ít dùng các phương pháp học máy giám sát: phân lớp SVM [60], phân lớp Bayesian [92] và một số các phương pháp xác suất khác. Nói cách khác, mô hình huấn luyện giúp tiên đoán giá trị hàm hữu ích $f(u, p)$ của đối tượng khuyến nghị $p \in P$ đối với người dùng $u \in U$. Chẳng hạn, phân lớp Bayesian là một phương pháp dựa trên mô hình khá phổ biến, được dùng trong khai thác dữ liệu, phân lớp Bayesian có thể dùng để ước lượng xác suất một đối tượng khuyến nghị p hữu ích với u như thế nào. Hay nói cách khác, p được u quan tâm hay không và quan tâm nhiều hay ít.

Ví dụ, xác suất một tài liệu p được một người dùng u nào đó quan tâm là bao nhiêu? Tức là, giá trị hàm hữu ích $f(u, p)$ khi đó sẽ được tính dựa trên việc ước lượng

xác suất p thuộc lớp $C_1(u)$ và $C_0(u)$ (u quan tâm và không quan tâm đến p) là bao nhiêu, khi cho trước một tập các từ khóa mô tả tài liệu p là $\{k_{1,p}, k_{2,p}, \dots, k_{n,p}\}$. Giá trị hàm hữu ích $f(u, p)$ khi đó sẽ được tính như sau:

$$f(u, p) = P(p \in C_1(u)) = P(C_1(u) | k_{1,p}, k_{2,p}, \dots, k_{n,p}) \quad (1.5)$$

Giả sử các từ khóa mô tả tài liệu là độc lập, khi đó xác suất $P(p \in C_1(u))$ sẽ là:

$$P(p \in C_1(u)) = P(C_1(u) | k_{1,p}, k_{2,p}, \dots, k_{n,p}) = P(C_1(u)) \prod_{i=1}^n P(k_{i,p} | C_1(u)) \quad (1.6)$$

Michael Pazzani và đồng nghiệp đã dùng tiếp cận dựa trên mô hình để phát triển hệ thống khuyến nghị các trang web phù hợp với sở thích người dùng, Syskill & Webert [96]. Các tác giả đã dùng một số phương pháp học máy truyền thống để xây dựng hồ sơ người dùng từ thông tin phản hồi của người dùng khi tương tác với hệ thống. Kết quả thực nghiệm của họ cho thấy việc ứng dụng phân lớp Bayesian để xây dựng hồ sơ người dùng cho kết quả tốt hơn so với một số phương pháp khác như lân cận gần nhất, mạng lan truyền ngược, cây quyết định. Trong bài báo [71], Jiahui Liu và cộng sự đã giới thiệu việc phát triển hệ thống khuyến nghị tin tức Google News. Hệ thống đã phân tích tập tin log để tìm hiểu sở thích, hành vi đọc tin của người dùng. Dựa trên kết quả phân tích, Jiahui Liu đã dùng phương pháp Bayesian để mô hình hóa hồ sơ người dùng và kết hợp với phương pháp lọc cộng tác để thực hiện khuyến nghị tin tức cho người dùng. Kết quả thực nghiệm của họ cho thấy, phương pháp đề xuất đã cải tiến chất lượng khuyến nghị tin tức, gia tăng, thu hút người dùng đến với trang Google News nhiều hơn. Nhìn chung, các phương pháp dựa trên mô hình có những ưu điểm và hạn chế sau:

Ưu điểm:

- Nhanh hơn so với các phương pháp dựa trên bộ nhớ do không phải tính trên cả tập dữ liệu mà chỉ dựa trên mô hình đã xây dựng để khuyến nghị.
- Khả năng đáp ứng tốt khi kích thước tập dữ liệu gia tăng.
- Một mô hình giúp biểu diễn tốt thế giới thực sẽ giúp tránh được vấn đề quá khớp (overfitting) so với các phương pháp dựa trên bộ nhớ.

Hạn chế:

- Phải xây dựng và cập nhật lại mô hình khi có sự thay đổi. Đây là quá trình tốn thời gian và tài nguyên.
- Chất lượng tiên đoán thấp hơn so với các phương pháp dựa trên bộ nhớ vì không được tính toán trên cả tập dữ liệu. Tuy nhiên, nó tùy thuộc vào chất lượng của mô hình được xây dựng có phản ánh tốt thế giới thực hay không.

1.4.1.4 Ưu điểm và hạn chế của tiếp cận nội dung

Bên cạnh ưu điểm là phù hợp cho các hệ thống khuyến nghị mà đối tượng khuyến nghị có thể biểu diễn dưới dạng các từ khóa như trang web, tài liệu, sách báo thì tiếp cận nội dung truyền thống có một số hạn chế có thể kể đến như sau:

- i) **Hạn chế về phân tích nội dung:** hệ thống sẽ không thể phân biệt được chất lượng của hai bài báo là tốt hay xấu, uy tín hay không uy tín để khuyến nghị, khi hai bài báo đó được biểu diễn bằng một tập các từ khóa quan trọng như nhau. Bên cạnh đó việc rút trích đặc trưng tự động cũng khó áp dụng cho các định dạng dữ liệu khác không phải là văn bản như hình ảnh, video, âm thanh, v.v...
- ii) **Bên ngoài lĩnh vực quan sát:** người dùng u chỉ được khuyến nghị các đối tượng mà tương tự cao với những gì u đã bình chọn, đánh giá trong một phạm vi cụ thể. Khi vượt quá phạm vi đánh giá của u thì hệ thống không thể thực hiện khuyến nghị được. Chẳng hạn tiếp cận nội dung sẽ thất bại khi u cần tham khảo các nhà hàng về ẩm thực Việt Nam, trong khi u chưa từng có những bình chọn và đánh giá về các nhà hàng, cũng như đặc sản ẩm thực Việt Nam.
- iii) **Người dùng mới** (khởi động lạnh): hệ thống khuyến nghị nội dung sẽ không thực hiện được cho những người dùng chưa có thông tin đánh giá trước đó. Nói cách khác, hệ thống không biết thật sự sở thích của người dùng đó là gì.

1.4.2 Tiếp cận lọc cộng tác (CF)

Một số hạn chế mà tiếp cận nội dung gặp phải phần nào giải quyết được với tiếp cận lọc cộng tác. Phần này sẽ trình bày chi tiết về tiếp cận lọc cộng tác và một số nghiên

cứu liên quan.

Định nghĩa 1.7: Ma trận đánh giá [57, 110]

Cho không gian người dùng $U = \{u_1, u_2, \dots, u_n\}$ và không gian các đối tượng khuyến nghị $P = \{p_1, p_2, \dots, p_m\}$. Ma trận A kích thước $n \times m$, chứa các giá trị đánh giá $a_{i,j}$, với $i \in 1 \dots n, j \in 1 \dots m$. Những giá trị đánh giá $a_{i,j}$ thể hiện mức độ hữu ích của đối tượng p_j với người dùng u_i (hay $f(u_i, p_j) = a_{i,j}$). Giá trị $a_{i,j}$ có thể là nguyên hay thực trong một khoảng cho trước tùy vào bài toán cụ thể. Thông thường, giá trị đánh giá $a_{i,j}$ trong một số hệ thống ứng dụng phổ biến nhận các giá trị từ 1 (không hữu ích) đến 5 (rất hữu ích). Nếu một người dùng u_i chưa thể hiện đánh giá với đối tượng p_j thì $a_{i,j} = \emptyset$ và cần được tính toán, xác định (dấu chấm hỏi (?)) trong hình 1.5).

	p_1	p_2	p_3	p_4	p_5	\dots	p_m
u_1	1	?	5	?	4	?	?
u_2	?	?	4	?	5	?	?
u_3	?	4	?	5	?	?	?
u_4	?	?	?	4	?	?	?
u_5	?	?	?	5	?	?	?
\dots	?	?	?	?	?	?	?
u_n	?	3	?	?	?	?	5

Hình 1.5: Dấu ? là các giá trị cần tiên đoán trong ma trận đánh giá

Tiếp cận CF được xem là tiếp cận thành công nhất để xây dựng các hệ thống khuyến nghị và ứng dụng rộng rãi trong lĩnh vực thương mại điện tử [110, 57]. Ý tưởng chung của tiếp cận CF là khai thác thông tin, hành vi quá khứ của người dùng dựa trên các đánh giá sẵn có từ ma trận đánh giá để tiên đoán, lượng hóa mức độ hữu ích của các đối tượng khuyến nghị mà người dùng chưa biết. Các nghiên cứu tổng quan về hệ khuyến nghị đã thực hiện khảo sát, phân loại, cũng như thực nghiệm, đánh giá các thuật toán CF. Các phương pháp CF nói chung được phân thành hai nhóm chính: (1) CF dựa trên bộ nhớ như các thuật toán tính toán tương tự, lân cận; (2) CF dựa trên mô hình như các thuật toán gom cụm, phân lớp Bayesian. Phần tiếp theo sẽ trình bày chi tiết về các nhóm thuật toán CF và các nghiên cứu liên quan.

1.4.2.1 Tiếp cận CF dựa trên bộ nhớ

Các hệ thống CF dựa trên bộ nhớ thường dùng các kỹ thuật thống kê để tìm kiếm những người dùng, hoặc các đối tượng khuyến nghị tương tự nhau dựa trên thông tin đánh giá, hành vi quá khứ của người dùng từ ma trận đánh giá. Tiếp cận CF dựa trên bộ nhớ tìm cách ước lượng giá trị hàm hữu ích $f(u, p)$ của đối tượng khuyến nghị p với người dùng u dựa trên những đánh giá của những người đồng sở thích của u đối với p (lọc dựa trên người dùng), hoặc dựa trên những đánh giá của u với các đối tượng khuyến nghị p' tương tự với p (lọc dựa trên đối tượng khuyến nghị). Về cơ bản, thì các thuật toán, kỹ thuật tính toán cho lọc cộng tác dựa trên người dùng và lọc dựa trên đối tượng khuyến nghị từ ma trận đánh giá là tương tự nhau. Có khác chăng là kích thước của không gian người dùng và không gian đối tượng khuyến nghị sẽ ảnh hưởng đến tốc độ tính toán khi xác định nhóm các đối tượng tương tự.

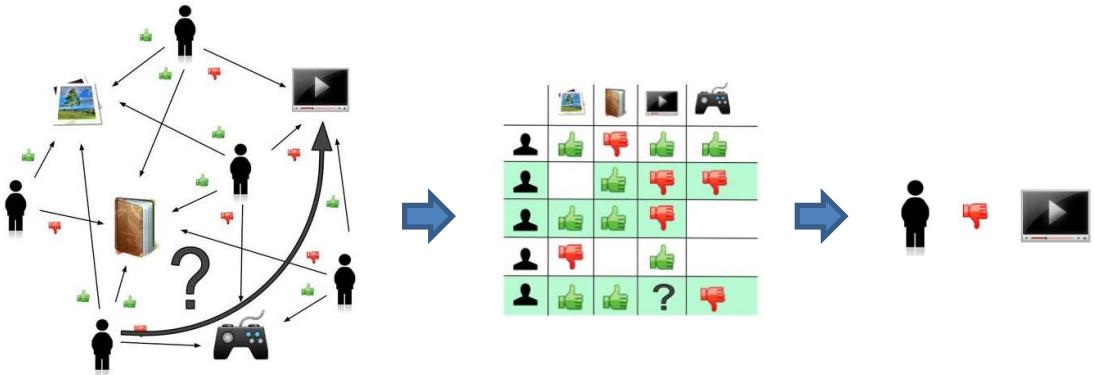
a) Lọc dựa trên người dùng

Định nghĩa 1.8: *Những người đồng sở thích*

Cho U là không gian người dùng, gọi S_u là tập những người đồng sở thích với $u \in U$, $S_u \subseteq U$. Những người đồng sở thích với u là những người có hành vi quá khứ hay các đánh giá tương tự với u trên cùng những đối tượng khuyến nghị từ ma trận đánh giá A [24, 110, 57].

(*) **Ý tưởng chính:** với các phương pháp lọc dựa trên người dùng, hệ thống thường phải thực hiện ba bước chính: một là xác định tập TopN những người có sở thích tương tự nhất với u , tức S_u ; hai là ước lượng giá trị hàm hữu ích $f(u, p)$ của đối tượng khuyến nghị p cho người dùng u bằng cách tổng hợp những đánh giá của S_u đối với p ; ba là thực hiện khuyến nghị dựa trên giá trị hàm hữu ích đã ước lượng được [57, 110]. Hình 1.6 minh họa việc thực hiện khuyến nghị dùng phương pháp lọc dựa trên người dùng.

(**) **Các bước thực hiện:** Jannach và cộng sự [57], Su và Khoshgoftaar [110] đã khảo sát, phân tích, trình bày về các phương pháp CF khác nhau. Các bước thực hiện của phương pháp CF dựa trên người dùng có thể tóm tắt dưới dạng mã giả như sau:



Hình 1.6: Minh họa dùng CF để tiên đoán một người thích hay không thích xem phim.
(Nguồn: http://en.wikipedia.org/wiki/Collaborative_filtering)

Phương pháp: CF dựa trên người dùng

Đầu vào:

- Không gian người dùng $U = \{u_1, u_2, \dots, u_n\}$, đối tượng khuyến nghị $P = \{p_1, p_2, \dots, p_m\}$
- Ma trận đánh giá $A = \{a_{i,j}\}$, với $i \in 1 \dots n, j \in 1 \dots m$

Đầu ra: $\forall u \in U$, trả về danh sách xếp hạng P_{TopN} (tức TopN những đối tượng khuyến nghị $p \in P$ dựa trên giá trị đánh giá tiên đoán được thông qua hàm $f(u, p)$)

1: $\forall u \in U, \forall u_i \in U | u_i \neq u$,

(1) Tương tự sở thích dùng độ đo cosine:

$$sim(u, u_i) = \frac{\vec{u} \bullet \vec{u}_i}{\|\vec{u}\| * \|\vec{u}_i\|} = \frac{\sum_{p \in P_{u,u_i}} a_{u,p} * a_{u_i,p}}{\sqrt{\sum_{p \in P_{u,u_i}} a_{u,p}^2} * \sqrt{\sum_{p \in P_{u,u_i}} a_{u_i,p}^2}} \quad (1.7)$$

(2) Tương tự sở thích dùng hệ số tương quan Pearson:

$$sim(u, u_i) = \frac{\sum_{p \in P_{u,u_i}} (a_{u,p} - \bar{a}_u)(a_{u_i,p} - \bar{a}_{u_i})}{\sqrt{\sum_{p \in P_{u,u_i}} (a_{u,p} - \bar{a}_u)^2} * \sqrt{\sum_{p \in P_{u,u_i}} (a_{u_i,p} - \bar{a}_{u_i})^2}} \quad (1.8)$$

Trong đó,

- $a_{u,p}$: giá trị đánh giá của u với p trong ma trận đánh giá A .
- $P_{u,u_i} = \{p \in P | a_{u,p} \neq \emptyset \text{ & } a_{u_i,p} \neq \emptyset\}$;
- $\bar{a}_u = \frac{1}{|P_u|} \sum_{p \in P_u} a_{u,p}$, với $P_u = \{p \in P | a_{u,p} \neq \emptyset\}$

2: $\forall u \in U$, xác định $S_u = \{u_i\}$: danh sách TopN những $u_i (u_i \neq u)$ tương tự nhất với u thông qua hàm $Sim(u, u_i)$.

3: $\forall u \in U, p \in P$, If ($a_{u,p} = \emptyset$ then)

4: Begin

(1) Trung bình đánh giá.

$$a_{u,p} = \frac{1}{|S_u|} \sum_{u_i \in S_u} a_{u_i,p} \quad (1.9)$$

(2) Tổng hợp đánh giá có trọng số.

$$a_{u,p} = 1 / \sum_{u_i \in S_u} |sim(u, u_i)| * \sum_{u_i \in S_u} sim(u, u_i) * a_{u_i,p} \quad (1.10)$$

(3) Tổng hợp dựa trên khoảng đánh giá.

$$a_{u,p} = \bar{a}_u + 1 / \sum_{u_i \in S_u} |sim(u, u_i)| * \sum_{u_i \in S_u} sim(u, u_i) * (a_{u_i,p} - \bar{a}_{u_i}) \quad (1.11)$$

5: End

6: $\forall u \in U$, return P_{TopN} dựa trên hàm hữu ích $f(u, p) = a_{u,p}$.

b) Lọc dựa trên đối tượng khuyến nghị

Sarwar và cộng sự đã đề xuất một phương pháp CF mới, là lọc dựa trên đối tượng khuyến nghị, thay vì lọc dựa trên người dùng như các hệ thống CF truyền thống. Tương tự lọc dựa trên người dùng, lọc dựa trên đối tượng khuyến nghị cũng bao gồm 3 bước chính: một là xác định danh sách TopN các đối tượng khuyến nghị tương tự nhất với đối tượng khuyến nghị p , I_p ; hai là ước lượng giá trị hàm hữu ích $f(u, p)$ của đối tượng khuyến nghị p cho người dùng u bằng cách tổng hợp những đánh giá của u cho $p \in I_p$; và ba là thực hiện khuyến nghị. Sarwar và cộng sự đã tiến hành thực nghiệm trên tập dữ liệu MovieLen, cho thấy các phương pháp lọc dựa trên đối tượng khuyến nghị cho kết quả tốt hơn các phương pháp lọc dựa trên người dùng [102].

1.4.2.2 Tiếp cận CF dựa trên mô hình

Theo quan điểm xác suất, thì các thuật toán CF dựa trên mô hình cần tính toán xác suất mà người dùng u có đánh giá $a_{u,p}$ cho một đối tượng khuyến nghị p , $P(a_{u,p}|u, p)$. Quá trình đó có thể xem như việc tính toán giá trị kỳ vọng cho đánh giá của người dùng u với đối tượng khuyến nghị p [24].

Khác với CF dựa trên bộ nhớ, các thuật toán CF dựa trên mô hình sẽ dùng tập các

đánh giá có sẵn trong ma trận đánh giá A để học một mô hình đánh giá cho mỗi người dùng. Sau đó, mô hình học được sẽ dùng để tiên đoán các đánh giá khác. Breese và cộng sự đã khảo sát và trình bày hai cách tiếp cận CF dựa trên mô hình gom cụm và mạng Bayes [24]. Phần tiếp theo trình bày và phân tích một số nghiên cứu liên quan.

a) Các thuật toán CF gom cụm

Một cụm bao gồm tập hợp các đối tượng dữ liệu tương tự nhau. Các đối tượng dữ liệu sẽ tương tự với các đối tượng dữ liệu khác thuộc cùng một cụm và sẽ khác biệt với các đối tượng dữ liệu trong cụm khác. Gom cụm là một kỹ thuật khá phổ biến trong lĩnh vực khai thác dữ liệu. Một số phương pháp gom cụm phổ biến có thể kể đến như: k-means [80], DBSCAN [41]. Các nghiên cứu liên quan thông thường sẽ dùng những kỹ thuật gom cụm để phân chia dữ liệu thành những cụm, sau đó áp dụng các thuật toán CF dựa trên bộ nhớ để thực hiện tiên đoán bên trong mỗi cụm [34, 103].

Những thuật toán CF dựa trên mô hình gom cụm nói chung có khả năng mở rộng tốt hơn các thuật toán CF thông thường, vì nó đã thực hiện tiên đoán bên trong các cụm có kích thước nhỏ hơn thay vì cả không gian ma trận đánh giá và cơ sở dữ liệu quan sát. Vì độ phức tạp và chi phí tính toán gom cụm, nên việc gom cụm có thể thực hiện offline. Tuy nhiên, kết quả nghiên cứu thực nghiệm cho thấy chất lượng khuyến nghị không cao khi thực hiện gom cụm [110].

b) Các thuật toán CF dựa trên xác suất Bayes

Tác giả Su và Khoshgoftaar đã trình bày thuật toán CF Bayes đơn giản nhất, đó là dùng xác suất Bayes ngây thơ để thực hiện tiên đoán [110]. Nhìn lớp trong trường hợp này thông thường là các giá trị nguyên có thể gán cho kết quả tiên đoán (nhân k=0, .., max). Giả sử những đặc trưng là độc lập với các lớp, khi đó xác suất mà đánh giá của u cho p thuộc về một lớp nào đó sẽ được tính. Lớp có xác suất cao nhất sẽ được chọn là kết quả tiên đoán. Việc tính toán xác suất, và tiên đoán được thực hiện dựa trên dữ liệu đã quan sát được.

$$ClassLabel(u, p) = \operatorname{argmax}_{k=0..max} P(class_k) * \prod_{j=1}^n P(a_j | class_k) \quad (1.12)$$

Để làm mịn giá trị xác suất cũng như tránh trường hợp xác suất điều kiện bằng

0, các tác giả Su và Khoshgoftaar đã đề xuất dùng bộ ước lượng Laplace (Laplace Estimator) khi tính xác suất có điều kiện [110]. Cụ thể như sau:

$$P(X_i = x_i | Y = y) = \frac{\#(X_i = x_i, Y = y) + 1}{\#(Y = y) + |X_i|} \quad (1.13)$$

Ví dụ minh họa: Cho ma trận đánh giá như trong bảng 1.1. Tập các giá trị đánh giá

Bảng 1.1: Ví dụ tiên đoán đánh giá

	p_1	p_2	p_3	p_4	p_5
u_1	4	1	5	5	?
u_2	?	?	4	4	5
u_3	1	?	4	2	1
u_4	?	?	4	2	3

là $\{1, 2, 3, 4, 5\}$. Yêu cầu đặt ra là cần tiên đoán đánh giá của u_1 với p_5 . Dùng thuật toán CF Bayes đơn giản nhất và phương pháp ước lượng Laplace để tính xác suất có điều kiện như trong công thức 1.13, chúng ta có thể tính như sau:

$$\begin{aligned} ClassLabel(u_1, p_5) &= \operatorname{argmax}_{c_k \in \{1,2,3,4,5\}} P(c_k | u_2 = 5, u_3 = 1, u_4 = 3) \\ &= \operatorname{argmax}_{c_k \in \{1,2,3,4,5\}} P(c_k) P(u_2 = 5 | c_k) P(u_3 = 1 | c_k) P(u_4 = 3 | c_k) \end{aligned} \quad (1.14)$$

Trong đó,

- $P(1)P(u_2 = 5|1)P(u_3 = 1|1)P(u_4 = 3|1) = (1/4)*(1/6)*(1/6)*(1/6) = 0.00116$
- $P(2)P(u_2 = 5|2)P(u_3 = 1|2)P(u_4 = 3|2) = 0$
- $P(3)P(u_2 = 5|3)P(u_3 = 1|3)P(u_4 = 3|3) = 0$
- $P(4)P(u_2 = 5|4)P(u_3 = 1|4)P(u_4 = 3|4) = (1/4)*(1/6)*(2/6)*(1/6) = 0.00231$
- $P(5)P(u_2 = 5|5)P(u_3 = 1|5)P(u_4 = 3|5) = (2/4)*(1/7)*(1/7)*(1/7) = 0.00146$

Như vậy, $ClassLabel(u_1, p_5) = \operatorname{argmax}_{c_k \in \{1,2,3,4,5\}} \{0.00116, 0, 0, 0.00231, 0.00146\} = 4$

c) Thừa số hóa ma trận (Matrix Factorization)

Bên cạnh các phương pháp CF dựa trên tính toán lân cận, gom cụm và xác suất Bayes đã đề cập ở trên thì các mô hình đặc trưng tiềm ẩn (latent factor model) là một chọn

lựa khác cho tiếp cận CF. Theo Koren và đồng nghiệp, các mô hình đặc trưng tiềm ẩn tìm cách biểu diễn thông tin của những người dùng và đối tượng khuyến nghị trong ma trận đánh giá thông qua 20 đến 100 đặc trưng tiềm ẩn rút trích từ các mẫu đánh giá. Những hiện thực thành công nhất của các mô hình đặc trưng tiềm ẩn là dựa trên kỹ thuật thừa số hóa ma trận (Matrix Factorization) [65].

Các kỹ thuật thừa số hóa ma trận về cơ bản là dựa trên các tính toán ma trận. Thừa số hóa ma trận thực hiện việc tìm kiếm hai hay nhiều ma trận mà tích của chúng chính là ma trận ban đầu. Thừa số hóa ma trận giúp hệ thống làm việc trên không gian dữ liệu có kích thước nhỏ hơn (giảm chiều). Các nghiên cứu liên quan đã chỉ ra rằng các kỹ thuật thừa số hóa ma trận hiệu quả hơn hẳn các phương pháp lọc cộng tác truyền thống dựa trên người dùng và dựa trên đối tượng khuyến nghị. Thừa số hóa ma trận đã góp phần giải quyết được tình trạng dữ liệu lớn, thưa trong ma trận đánh giá mà các phương pháp CF truyền thống phải đương đầu, giúp khám phá ra các đặc trưng tiềm ẩn thể hiện bản chất của những tương tác giữa người dùng và đối tượng khuyến nghị trong ma trận đánh giá.

Năm 2006, Netflix, một công ty cho thuê phim hàng đầu thế giới, đã thông báo tổ chức một cuộc thi nhằm cải tiến hệ thống khuyến nghị phim của họ. Netflix đã công bố một tập dữ liệu gồm 100 triệu đánh giá của khoảng 500.000 khách hàng với hơn 17.000 phim. Mỗi đánh giá sẽ nhận các giá trị từ 1 đến 5. Công ty kêu gọi và thách thức cộng đồng khoa học máy tính phát triển các phương pháp máy học, khai thác dữ liệu làm thế nào để cải tiến độ chính xác khuyến nghị của hệ thống khuyến nghị phim Netflix [17]. Năm 2008, Koren và cộng sự đã công bố phương pháp thừa số hóa ma trận cho CF và đã thắng giải thưởng Netflix trị giá 1 triệu USD [65]. Từ đó đến nay, các kỹ thuật thừa số hóa ma trận tiếp tục được phát triển và trở thành một trong những tiếp cận nổi trội cho hệ khuyến nghị nói chung và tiếp cận CF nói riêng.

Những mô hình thừa số hóa ma trận tìm cách ánh xạ không gian người dùng và đối tượng khuyến nghị (ma trận đánh giá) vào một không gian đặc trưng tiềm ẩn với số chiều là k . Ở đó, những tương tác giữa người dùng và đối tượng khuyến nghị trong ma trận đánh giá sẽ được mô hình như các tích vô hướng trong không gian đó. Mỗi đối tượng khuyến nghị $p \in P$ tương ứng với một vectơ $\vec{a}_p \in \mathbb{R}^k$. Mỗi người dùng $u \in U$ tương ứng với một vectơ $\vec{b}_u \in \mathbb{R}^k$. Với một đối tượng $p \in P$, những thành phần của

\vec{a}_p đo lường mức độ mà đối tượng p có được ở những đặc trưng đó. Đối với một người dùng $u \in U$, những thành phần của \vec{b}_u đo lường mức độ quan tâm mà u thể hiện đối với các đối tượng khuyến nghị $p \in P$ có giá trị cao ở những đặc trưng tiềm ẩn tương ứng đó. Tích vô hướng $\vec{a}_p \cdot \vec{b}_u$ thể hiện tương tác giữa người dùng u và đối tượng p (quan tâm của u trên các đặc trưng của p). Giá trị tính được trong không gian đặc trưng tiềm ẩn sẽ xấp xỉ với giá trị đánh giá của u đối với p ($r(u, p)$). Do đó, giá trị đánh giá của u với p có thể được ước lượng, tiên đoán như sau:

$$\hat{r}(u, p) = \vec{a}_p \cdot \vec{b}_u \quad (1.15)$$

Trong đó,

- $\hat{r}(u, p)$: giá trị đánh giá thể hiện quan tâm của u với p tiên đoán được trong không gian \mathbb{R}^k .
- \vec{a}_p : vectơ biểu diễn đối tượng p trong không gian đặc trưng tiềm ẩn \mathbb{R}^k .
- \vec{b}_u : vectơ biểu diễn người dùng u trong không gian đặc trưng tiềm ẩn \mathbb{R}^k .

Vấn đề đặt ra là làm thế nào để ánh xạ $p \in P, u \in U$ thành những vectơ đặc trưng tiềm ẩn $\vec{a}_p, \vec{b}_u \in \mathbb{R}^k$. Khi ánh xạ này được xác định, hệ thống có thể ước lượng giá trị đánh giá của bất kỳ $u \in U$ với $p \in P$ dùng công thức 1.15. Một mô hình như thế liên quan đến phương pháp phân tích giá trị đơn (Singular Value Decomposition), gọi tắt là SVD, một kỹ thuật dùng để xác định các đặc trưng ngữ nghĩa tiềm ẩn trong lĩnh vực truy vấn thông tin. Để áp dụng SVD cho lọc cộng tác thì ma trận đánh giá cần phải được thừa số hóa. Điều này sẽ gặp khó khăn vì ma trận đánh giá thưa (nhiều giá trị đánh giá chưa xác định). Thông thường, SVD sẽ không định nghĩa được khi thông tin trong ma trận không đầy đủ. Hơn nữa, dựa trên một số ít phần tử trong ma trận để tính toán thì dễ dẫn đến vấn đề quá khớp. Dẫn đến việc tiên đoán không chính xác.

Các nghiên cứu phổ biến thường dùng một số kỹ thuật tính toán để khởi tạo và làm đầy những chỗ thiếu thông tin trong ma trận đánh giá. Tuy nhiên, việc tính toán đó sẽ tốn nhiều chi phí khi kích thước dữ liệu gia tăng. Thêm vào, việc tính toán không chính xác có thể dẫn đến những sai lệch đáng kể trong dữ liệu. Do đó, một số nghiên cứu khác đã đề xuất việc mô hình hóa trực tiếp chỉ dựa trên những đánh giá quan sát được.

Nói chung, để học và xác định những vectơ đặc trưng tiềm ẩn \vec{a}_p, \vec{b}_u , hệ thống cần phải tối thiểu hóa lỗi bình phương dựa trên các đánh giá đã biết như sau:

$$\min \sum_{(u,p) \in K} (r(u,p) - \hat{r}(u,p))^2 + \lambda(\|\vec{a}_p\|^2 + \|\vec{b}_u\|^2) \quad (1.16)$$

Trong đó,

- K : tập các cặp (u, p) mà giá trị $r(u, p)$ được biết trước (dữ liệu huấn luyện).
- λ : hằng số kiểm soát vấn đề quá khớp khi học dựa trên dữ liệu quan sát, được xác định thông qua việc cross-validation (Thực nghiệm có thể chọn λ là 0.1, 0.2, ..., 1).

Hệ thống sẽ học một mô hình bằng cách cực thiểu hóa lỗi bình phương theo công thức 1.16 để xấp xỉ tốt nhất với những đánh giá đã quan sát được.

Phương pháp phân tích giá trị đơn SVD

SVD cho phép phân tích một ma trận R_{mn} thành tích của 3 ma trận đó là: một ma trận trực giao U , một ma trận đường chéo S và một ma trận chuyển vị của ma trận trực giao V [10].

$$R_{mn} = U_{mm}S_{mn}V_{nn}^T \quad (1.17)$$

Trong đó,

- $U^T U = I; V^T V = I$ (I : ma trận đơn vị)
- Cột của ma trận trực giao U là các vectơ riêng trực chuẩn của RR^T
- Cột của ma trận trực giao V là các vectơ riêng trực chuẩn của $R^T R$
- S là ma trận đường chéo, căn bậc 2 của những giá trị riêng từ U hay V theo thứ tự giảm dần.

Để xác định các đặc trưng tiềm ẩn (giảm chiều) khi thực nghiệm, SVD chọn số đặc trưng tiềm ẩn $k \leq \min(m, n)$. Khi đó, phương pháp SVD sẽ học để xác định 3 ma trận

U, S, V mà tích của chúng là \hat{R}_{mn} (ma trận xấp xỉ tốt nhất với ma trận ban đầu R_{mn}) bằng cách cực tiểu lỗi bình phương (công thức 1.16).

$$\hat{R}_{mn} = U_{mm} S_{mk} V_{kn}^T \quad (1.18)$$

1.4.2.3 Ưu điểm và hạn chế của tiếp cận CF

Khác với tiếp cận nội dung, các hệ thống CF không bị các hạn chế về mặt phân tích nội dung cho các đối tượng khuyến nghị có nội dung dạng văn bản. Những hệ thống CF dùng thông tin từ ma trận đánh giá quan sát được của những người dùng khác. Vì vậy, các hệ thống CF có thể áp dụng cho nhiều dạng đối tượng, nhiều kiểu nội dung khác nhau, ngay cả với các đối tượng khuyến nghị không tương tự với các đối tượng khuyến nghị quan sát trong quá khứ. Theo các tác giả Su và Khoshgoftaar, tiếp cận CF được xem là một trong những cách tiếp cận thành công nhất để thiết kế các thuật toán và xây dựng các hệ khuyến nghị [110]. Tuy nhiên, theo nghiên cứu tổng quan của các tác giả Adomavicius và Tuzhilin [5], tác giả Bobadilla và cộng sự [22], thì tiếp cận CF truyền thống cũng có một số hạn chế như sau:

- **Ma trận thưa:** Đầu vào của các hệ thống CF là ma trận đánh giá. Trên thực tế thì không gian người dùng và đối tượng khuyến nghị là rất lớn. Trong khi đó số đánh giá của người dùng với các đối tượng khuyến nghị thực chất không nhiều. Nói chung, số lượng đánh giá quan sát được là rất nhỏ so với số lượng đánh giá cần phải tiên đoán. Điều đó ảnh hưởng đến độ chính xác tiên đoán, cũng như chất lượng khuyến nghị. Chẳng hạn, trong hệ thống khuyến nghị nhà hàng, có rất nhiều nhà hàng mới chỉ được đánh giá bởi một số rất ít người dùng. Như thế, cho dù nó có chất lượng và được đánh giá rất cao bởi một ít người dùng thì nó vẫn hiếm khi được khuyến nghị cho người dùng khác. Điều đó dẫn đến hệ thống khuyến nghị "nghèo nàn".
- **Người dùng mới** (khởi động lạnh): Tương tự như tiếp cận dựa trên nội dung. Hệ thống CF cần phân tích hay học từ ma trận đánh giá để biết được sở thích của người dùng. Đối với người chưa có hoặc rất ít thông tin cá nhân cũng như thông tin đánh giá của họ, thì hệ thống không thể biết được sở thích của người dùng. Do đó không thể có những tiên đoán, khuyến nghị hữu ích.

-
- **Đối tượng khuyến nghị mới** (khởi động lạnh): Các đối tượng khuyến nghị mới được thêm vào hệ thống một cách thường xuyên. Các hệ thống CF dựa trên sở thích người dùng để thực hiện khuyến nghị, do đó nếu những đối tượng mới chưa được ai đánh giá thì nó không thể được khuyến nghị, mặc dù có thể nó rất tương tự với một số đối tượng khuyến nghị nào đã có, hoặc rất tiềm năng với người dùng.

Những hạn chế của tiếp cận nội dung cũng như tiếp cận CF truyền thống có thể giải quyết bằng cách kết hợp hai hay nhiều phương pháp khác nhau (tiếp cận lai) để phát triển các phương pháp khuyến nghị mới. Phần tiếp theo sẽ trình bày một số phương pháp lai phổ biến, cũng như các nghiên cứu, ứng dụng liên quan.

1.4.3 Tiếp cận lai (Hybrid Approach)

Những phương pháp khác nhau đều có những điểm mạnh, cũng như điểm yếu của nó (bảng 1.2). Để tận dụng những điểm mạnh và hạn chế điểm yếu của những tiếp cận khác nhau, nhiều nghiên cứu đã tập trung phát triển các hệ khuyến nghị dựa trên việc kết hợp các tiếp cận khác nhau, được gọi là tiếp cận lai (Hybrid Approach) hay hệ khuyến nghị lai (Hybrid Recommender System). Robin Burke đã khảo sát các phương pháp lai cho hệ khuyến nghị và trình bày tóm tắt 7 nhóm phương pháp tiếp cận lai khác nhau [25]. Phần tiếp theo sẽ trình bày tóm tắt 7 nhóm phương pháp lai này, các nghiên cứu liên quan cũng như điểm mạnh, yếu của mỗi nhóm phương pháp.

1.4.3.1 Lai có trọng số (Weighted Hybrid)

Mỗi phương pháp khuyến nghị phải đi tìm và xác định giá trị hàm hữu ích $f(u, p)$ của đối tượng $p \in P$ với người dùng $u \in U$. Tiếp cận lai có trọng số (Weighted Hybrid) tính toán giá trị của hàm hữu ích $f_{hybrid}(u, p)$ dựa trên kết quả của tất cả các $f(u, p)$ của các phương pháp khuyến nghị khác tồn tại trong hệ thống. Thông thường, hình thức lai có trọng số đơn giản nhất là kết hợp tuyến tính các giá trị hữu ích tính được từ các phương pháp khác nhau trong hệ thống.

Tác giả Claypool và cộng sự đã lần đầu áp dụng kết hợp tuyến tính cho hệ thống lọc tin tức trực tuyến. Họ đã khởi tạo trọng số ngang nhau cho phương pháp CB và CF để tính toán giá trị hữu ích cho tin tức khuyến nghị. Sau đó, hệ thống sẽ dần hiệu

chỉnh trọng số khi nhận được những đánh giá phản hồi từ người dùng [32]. Trong một nghiên cứu khác, tác giả Pazzani đã đề xuất phương pháp lai có trọng số bằng cách kết hợp tuyến tính kết quả tiên đoán (giá trị hữu ích) của 5 phương pháp khác nhau là lọc cộng tác dựa trên người dùng, lọc cộng tác dựa trên đối tượng, lọc nội dung, lọc dựa trên thông tin cá nhân (demographic filtering), lọc cộng tác kết hợp nội dung (Collaboration via Content). Giá trị hữu ích sau cùng là tổng giá trị hữu ích của 5 phương pháp khác nhau (các giá trị hữu ích từ 1 đến 5). Pazzani đã tiến hành thực nghiệm trên tập dữ liệu liên quan đến đánh giá của người dùng là sinh viên đối với các nhà hàng ở quận Cam, USA. Kết quả thực nghiệm của họ cho thấy việc kết hợp tuyến tính 5 phương pháp cho kết quả tốt hơn khi thiếu đi sự kết hợp của một phương pháp trong số đó [95].

- **Ưu điểm:** Tất cả khả năng, phương pháp khác nhau của hệ thống được tham gia vào quá trình khuyến nghị một cách minh bạch, tự nhiên, dễ dàng thực hiện, dễ dàng hiệu chỉnh.
- **Hạn chế:** Việc ước lượng trọng số lớn hay nhỏ cho phù hợp với những phương pháp khác nhau.

1.4.3.2 Lai chuyển đổi (Switching Hybrid)

Các hệ thống khuyến nghị thuộc nhóm lai chuyển đổi (Switching Hybrid) sử dụng một số điều kiện để chuyển đổi qua lại giữa các phương pháp khuyến nghị khác nhau. Billsus và Pazzani đã giới thiệu hệ thống DailyLearner, một nghiên cứu về hệ thống dịch vụ tin tức thích công bố trên mạng từ 05/1999 đến 06/2000, là một hệ thống sử dụng phương pháp lai chuyển đổi giữa tiếp cận nội dung và lọc cộng tác [21]. Các tác giả áp dụng phương pháp lọc nội dung trước. Sau đó, những trường hợp mà tiếp cận nội dung không thể thực hiện khuyến nghị (giá trị hữu ích thấp) thì tiếp cận lọc cộng tác sẽ được áp dụng.

Lọc cộng tác trong phương pháp lai chuyển đổi giúp hệ thống có thể khuyến nghị được các đối tượng có nội dung, ngữ nghĩa khác với các đối tượng đã được đánh giá cao. Nói cách khác, một đối tượng có thể không được khuyến nghị với tiếp cận nội dung, nhưng sau khi áp dụng lọc cộng tác thì đối tượng đó có thể được ưu tiên khuyến nghị.

-
- **Ưu điểm:** Tiếp cận này rất "nhạy" với các điểm mạnh và yếu của các phương pháp khác nhau.
 - **Hạn chế:** Tuy "nhạy" với điểm mạnh và yếu của các phương pháp khác nhau, nhưng lai chuyển đổi yêu cầu cần phải xác định điều kiện để chuyển đổi giữa các phương pháp. Điều này làm cho quá trình khuyến nghị trở nên phức tạp hơn.

1.4.3.3 Lai trộn (Mixed Hybrid)

Tiếp cận lai trộn (mixed hybrid) thực hiện các phương pháp khuyến nghị khác nhau một cách độc lập và kết hợp kết quả từ các phương pháp này thành danh sách sau cùng đề xuất cùng lúc đến người dùng. Tiếp cận lai trộn tránh được vấn đề đối tượng khuyến nghị mới (một trường hợp của khởi động lạnh). Lọc dựa trên nội dung trong tiếp cận lai trộn giúp đề xuất các đối tượng khuyến nghị mới (chưa hoặc có rất ít đánh giá) trong danh sách sau cùng dựa trên những mô tả nội dung đối tượng này trong khi phương pháp lọc cộng tác không thể thực hiện được. Bù lại, lọc cộng tác trong lai trộn giúp đề xuất các đối tượng khuyến nghị tiềm năng nhưng không tương tự về nội dung.

Các tác giả Smyth và Cotter dùng tiếp cận lai trộn để phát triển một hệ thống khuyến nghị chương trình truyền hình phù hợp với sở thích cá nhân của người dùng, hệ thống PTV [108]. Với PTV, những người dùng đăng ký vào hệ thống sẽ nhận được các khuyến nghị chương trình truyền hình mỗi ngày thông qua Internet. PTV xây dựng hồ sơ người dùng bằng cách cho người dùng tự cập nhật thông tin sở thích. Bên cạnh đó, hệ thống cũng ghi nhận đánh giá phản hồi của người dùng thông qua kết quả khuyến nghị. Kết quả khuyến nghị được tập hợp, trộn từ kết quả của hai phương pháp lọc nội dung và lọc cộng tác. Chất lượng, độ chính xác khuyến nghị của hệ thống PTV được đánh giá thông qua việc khảo sát ý kiến người dùng. Bên cạnh đó, Burke cũng đã khảo sát một số hệ thống, nghiên cứu khác có sử dụng tiếp cận lai trộn như: ProfBuilder (Wasfi, 1999), PickAFlick (Burke và đồng nghiệp 1997, 2000) [25]. Nhìn chung, tiếp cận này có những ưu điểm hạn chế như sau:

- **Ưu điểm:** Giúp đề xuất các đối tượng khuyến nghị tiềm năng mà bản thân một phương pháp riêng biệt không xác định được. Trộn lọc nội dung và lọc cộng tác giúp giải quyết được vấn đề khởi động lạnh (đối tượng khuyến nghị mới) và có thể đa dạng hóa khuyến nghị (đối tượng không tương tự về nội dung).

-
- **Hạn chế:** Tiếp cận này nhằm tận dụng nhiều đề xuất từ nhiều phương pháp khác nhau. Vì vậy, hệ thống cần cơ chế xử lý, lọc các đề xuất đúng đắn, trùng lặp từ các phương pháp khác nhau. Ở đây, nếu trộn hai phương pháp lọc nội dung và lọc cộng tác thì vẫn chưa giải quyết được vấn đề người dùng mới (một trường hợp của khởi động lạnh).

1.4.3.4 Lai kết hợp đặc trưng (Feature Combination Hybrid)

Lai kết hợp đặc trưng là tiếp cận phát triển phương pháp khuyến nghị bằng cách sử dụng kết hợp thông tin đánh giá của người dùng với nội dung của đối tượng khuyến nghị. Tác giả Basu và đồng nghiệp đã đề xuất tiếp cận học luật dựa trên việc kết hợp đặc trưng để thực hiện khuyến nghị [14]. Họ đã thử nghiệm trên tập dữ liệu hơn 45.000 phim và hơn 250 người dùng. Mỗi cặp (người dùng, phim) được mã hóa thành tập các đặc trưng bao gồm đặc trưng cộng tác (rút từ ma trận đánh giá) và các đặc trưng nội dung mô tả phim. Kết quả thực nghiệm của họ cho thấy: việc sử dụng tất cả đặc trưng về nội dung cải tiến độ đo bao phủ (Recall), nhưng không cải tiến độ chính xác (Precision); việc kết hợp đặc trưng đã cải tiến đáng kể cả độ chính xác và độ bao phủ so với không kết hợp đặc trưng.

- **Ưu điểm:** Lai kết hợp đặc trưng cho phép hệ thống xem xét dữ liệu cộng tác, nhưng không chỉ phụ thuộc duy nhất vào dữ liệu cộng tác trong ma trận đánh giá. Ngược lại, hệ thống cũng có được thông tin về sự tương tự vốn có giữa các đối tượng khuyến nghị (dựa trên đặc trưng nội dung) mà không bị ảnh hưởng bởi dữ liệu cộng tác.
- **Hạn chế:** Khó khăn trong việc xác định các đặc trưng cộng tác và đặc trưng nội dung phù hợp.

1.4.3.5 Lai theo đợt (Cascade Hybrid)

Lai theo đợt (cascade hybrid) là tiếp cận mà các phương pháp khuyến nghị khác nhau được lần lượt áp dụng theo một thứ tự ưu tiên được xác định trước tùy vào mỗi ứng dụng cụ thể. Phương pháp khuyến nghị thứ nhất sinh ra một danh sách xếp hạng các ứng viên (danh sách thô). Tiếp đó, những phương pháp khác với độ ưu tiên thấp hơn sẽ được áp dụng để lọc lại danh sách thô này. Lai theo đợt giúp phương pháp thứ hai

tránh những đối tượng có thẻ không bao giờ cần khuyến nghị vì những đối tượng này đã được lọc qua phương pháp thứ nhất. Đồng thời, các đối tượng được ưu tiên chọn với phương pháp thứ nhất sẽ được tinh lọc, chứ không bị loại bỏ thông qua phương pháp thứ hai.

Entree¹ là một hệ thống khuyến nghị nhà hàng dựa trên tri thức. Entree dùng kỹ thuật suy luận dựa trên trường hợp (case-based reasoning) để chọn và xếp hạng những nhà hàng hỗ trợ những người tham gia một hội nghị ở Chicago năm 1996. EntreeC là một cải tiến của Entree. EntreeC áp dụng tiếp cận lai theo đợt bằng cách bổ sung thêm phương pháp lọc cộng tác để thực hiện việc tinh lọc ở đợt thứ 2 so với đợt lọc đầu tiên dựa trên tri thức của Entree [25].

- **Ưu điểm:** So với tiếp cận lai có trọng số (Weighted Hybrid) và một số tiếp cận lai khác thì việc lọc lại danh sách thô làm cho tiếp cận này hiệu quả hơn bởi vì các phương pháp tiếp theo chỉ thực hiện lọc trên một không gian nhỏ hơn (danh sách thô), thay vì trên cả không gian tất cả các đối tượng khuyến nghị.
- **Hạn chế:** Khó khăn trong việc xác định độ ưu tiên giữa các phương pháp khác nhau cho mỗi ứng dụng cụ thể.

1.4.3.6 Lai tăng cường đặc trưng (Feature Augmentation Hybrid)

Với tiếp cận lai tăng cường đặc trưng, phương pháp đầu sẽ học một mô hình để sinh ra đặc trưng tăng cường cho đầu vào phương pháp tiếp theo. Tác giả Mooney và Roy đã giới thiệu một hệ thống thử nghiệm LIBRA (Learning Intelligent Book Recommending Agent) dùng cơ sở dữ liệu thông tin về sách được rút trích từ trang Amazon.com cho bài toán khuyến nghị sách [86]. LIBRA khai thác thông tin về những tác giả liên quan, tiêu đề liên quan mà Amazon đã tạo ra dựa trên phương pháp lọc cộng tác. Sau đó, những thông tin này được dùng như những đặc trưng bổ sung thêm vào những đặc trưng nội dung để học hồ sơ sở thích người dùng sử dụng thuật toán học máy Bayesian. Ở đây, bộ phân lớp xác suất Bayesian dùng để tiên đoán xác suất mỗi cuốn sách sẽ phù hợp với sở thích của người dùng là nhiều hay ít. Kết quả thực nghiệm của họ cho thấy việc tăng cường đặc trưng sinh ra bởi lọc cộng tác đã cải tiến phương pháp lọc nội dung.

¹<http://infolab.ils.nwu.edu/entree/>

Lai theo đợt và lai tăng cường đặc trưng là những tiếp cận mà hai phương pháp khác nhau sẽ được thực hiện một cách trình tự. Tức là kết quả của phương pháp thứ nhất sẽ ảnh hưởng lên phương pháp thứ hai. Tuy nhiên, về cơ bản thì hai tiếp cận lai này hoàn toàn khác nhau. Với lai tăng cường đặc trưng thì những đặc trưng được dùng trong phương pháp thứ hai bao gồm những đặc trưng sinh ra bởi phương pháp thứ nhất. Còn đối với lai theo đợt thì phương pháp thứ hai được dùng với độ ưu tiên thấp hơn phương pháp thứ nhất, nhằm lọc lại danh sách ứng viên mà phương pháp thứ nhất đã sinh ra.

- **Ưu điểm:** Việc tăng cường đặc trưng dùng các phương pháp khác giúp hệ thống có thể cải tiến độ chính xác khuyến nghị mà không thay đổi, ảnh hưởng đến phương pháp khuyến nghị chính.
- **Hạn chế:** Khó khăn trong việc xác định đặc trưng tăng cường phù hợp.

1.4.3.7 Lai meta (Meta-Level Hybrid)

Lai meta dùng mô hình được tạo ra bởi phương pháp trước làm đầu vào cho phương pháp sau. Với lai tăng cường đặc trưng (Feature Augmentation) thì phương pháp đầu sẽ học một mô hình để sinh ra đặc trưng làm đầu vào cho phương pháp tiếp theo. Trong khi lai meta thì cả mô hình của phương pháp thứ nhất sẽ làm đầu vào cho phương pháp thứ hai. Lai meta giữa lọc nội dung và lọc cộng tác phần nào giải quyết được vấn đề ma trận thừa trong tiếp cận lọc cộng tác bởi vì lai meta sẽ tìm kiếm những người dùng tương tự dựa trên các đặc trưng nội dung trước khi áp dụng phương pháp lọc cộng tác. Đối với những người dùng có quá ít đánh giá thì việc xác định nhóm những người đồng sở thích thông qua lọc cộng tác sẽ không được chính xác.

Tác giả Pazzani đã áp dụng tiếp cận lai meta để đề xuất phương pháp học hồ sơ người dùng và thực hiện khuyến nghị các trang web hay bài báo về lĩnh vực nhà hàng [95]. Đầu tiên, hồ sơ sở thích người dùng được học từ nhiều nguồn thông tin như: thông tin cá nhân, nội dung trang web mà họ quan tâm, đánh giá và biểu diễn dưới dạng các vectơ trọng số. Sau đó, phương pháp lọc cộng tác sẽ được áp dụng để tổng hợp đánh giá từ những người dùng đồng sở thích đã xác định bởi phương pháp lọc dựa trên nội dung trước đó. Tương tự vậy, Balabanovic và Shoham đã kết hợp đặc điểm nội dung

vào tiếp cận CF để phát triển Fab, một hệ khuyến nghị thông tin trên web, thuộc dự án thư viện số của trường Đại học Standford [11].

- **Ưu điểm:** Với lai meta giữa lọc nội dung và cộng tác, phương pháp lọc cộng tác sẽ dễ dàng thực hiện tính toán trên "dữ liệu dày" hơn so với dữ liệu thô trong ma trận đánh giá.
- **Hạn chế:** Khó khăn trong việc chọn phương pháp để thực hiện trước. Mỗi phương pháp được chọn vẫn gặp phải những hạn chế vốn có của nó.

Tóm lại, mỗi phương pháp đều có những ưu điểm cũng như hạn chế của nó. Tiếp cận lai giúp giảm bớt phần nào những hạn chế của các phương pháp khác nhau. Kết quả thực nghiệm của các nghiên cứu khảo sát ở trên cho thấy các phương pháp lai hầu hết cho kết quả tốt hơn một phương pháp riêng lẻ nào đó. Tuy nhiên, với các bài toán trong lĩnh vực học thuật, thì lai như thế nào vẫn là vấn đề tiếp tục được nghiên cứu, phát triển.

1.4.4 Tiếp cận phân tích mạng xã hội

1.4.4.1 Một số khái niệm cơ bản

a) Mạng xã hội

Mạng xã hội, một khái niệm đã có từ những năm đầu thế kỷ 20. Khái niệm mạng xã hội có nguồn gốc từ ngành khoa học xã hội. Mạng xã hội là một cấu trúc xã hội hình thành từ những đối tượng là các cá nhân hay tổ chức được gọi là các actor và những quan hệ, ràng buộc xã hội giữa chúng [126] (hình 1.7).

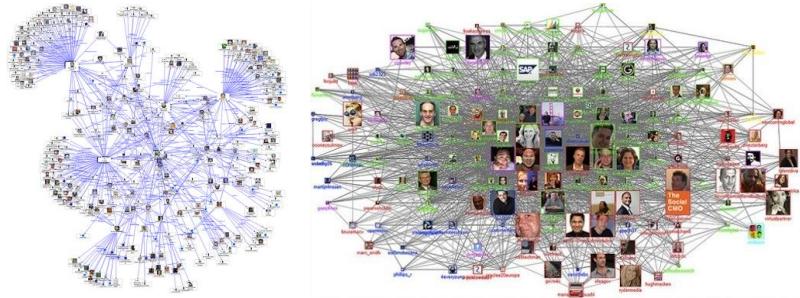
b) Biểu diễn mạng xã hội

Một mạng xã hội (Social Network), ký hiệu SN, có thể được định nghĩa là một đồ thị gồm các thành phần cơ bản như sau:

$$SN = (V, E, L(V), L(E)) \quad (1.19)$$

Trong đó,

- V: tập các actor hay còn gọi là các nút trong mạng.



Hình 1.7: Minh họa trực quan mạng xã hội
 (Nguồn: <http://mashable.com/2012/09/26/graph-databases/>
<http://www.fmsasg.com/SocialNetworkAnalysis/>
 Truy cập lần cuối 20/01/2014)

- $E \subseteq V \times V$: tập các cạnh của đồ thị hay các mối quan hệ trong mạng.
- $L(V)$: tập các nhãn nút.
- $L(E)$: tập các nhãn cạnh.

Tùy vào mỗi bài toán cụ thể, các nút, cạnh thuộc mạng có cấu trúc và ngữ nghĩa khác nhau. Tức với mỗi bài toán, chúng ta có thể dùng đơn đồ thị có hướng, vô hướng hoặc đa đồ thị để mô hình hóa các mạng xã hội tương ứng.

c) Phân tích Mạng xã hội

Nghiên cứu về các mạng xã hội còn được gọi là phân tích mạng xã hội, viết tắt là SNA. Theo tác giả Serrat [104], SNA tập trung vào phân tích cấu trúc của những quan hệ. SNA giả sử rằng các mối quan hệ là rất quan trọng và tìm cách tính toán, đo lường các mối quan hệ một cách hình thức và không hình thức để hiểu được những gì đã tạo điều kiện hay cản trở những luồng thông tin di chuyển, lan truyền trong mạng. SNA được sử dụng rộng rãi trong khoa học xã hội và khoa học hành vi, trong kinh tế, tiếp thị cũng như công nghệ và được xem như một kỹ thuật chính trong khoa học xã hội hiện đại [126].

Ngành xã hội học đã nghiên cứu và đưa ra một số nguyên tắc có thể ứng dụng trong quá trình phân tích các mạng xã hội. Trong đó homophily, proximity là những nguyên tắc cơ bản và được quan tâm.

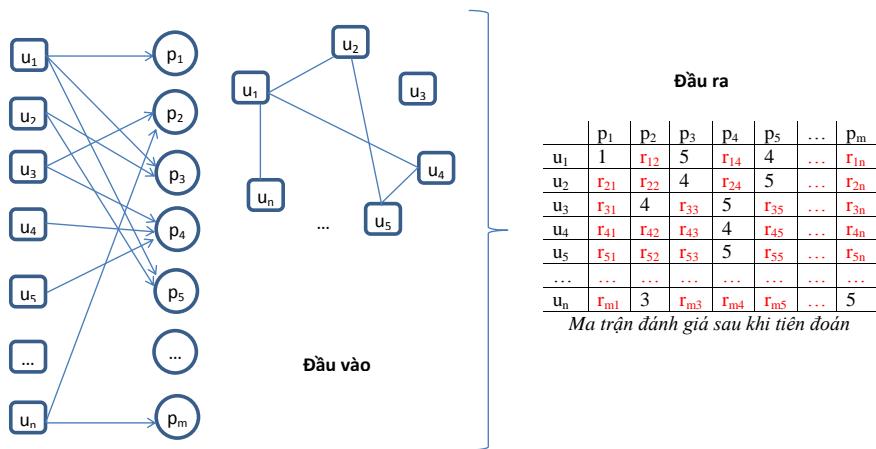
-
- **Nguyên tắc ‘homophily’.** Những nhà xã hội học đã đưa ra nguyên tắc: "Sự tương tự tạo ra những kết nối", đó chính là nguyên tắc ‘homophily’ trong khoa học xã hội và hành vi. Nguyên tắc này đã hình thành nên những kết nối và cấu trúc xã hội với nhiều kiểu quan hệ khác nhau như: bạn bè, vợ chồng, đồng nghiệp, đồng thành viên và nhiều kiểu quan hệ khác. ‘Homophily’ đã tạo ra những phân chia mạnh mẽ nhất trong môi trường cá nhân của chúng ta với tuổi tác, tôn giáo, nghề nghiệp, giới tính [83].
 - **Nguyên tắc ‘proximity’.** Trong lĩnh vực tâm lý xã hội, nguyên tắc ‘proximity’ cho biết các cá nhân có xu hướng hình thành những mối quan hệ với những người ở gần. Tức những người gặp nhau thường xuyên, sống và làm việc gần nhau thì dễ hình thành và phát triển quan hệ.

1.4.4.2 Khuyến nghị xã hội (Social Recommendation)

Hệ khuyến nghị truyền thống giả sử những người dùng trong hệ thống độc lập với nhau, chưa xem xét những tương tác cũng như quan hệ xã hội của họ. Trong khi, các mối quan hệ xã hội có vai trò rất quan trọng, ảnh hưởng đến sở thích, hành vi và quyết định của con người. Chẳng hạn, chúng ta thường xin ý kiến tư vấn, đề xuất của người thân, bạn bè, thầy cô, đồng nghiệp để thực hiện một quyết định như: mua một sản phẩm, xem một cuốn phim, tìm kiếm nhà hàng, chọn một công việc, đọc một cuốn sách, tìm một bài báo, chọn một đối tác. Thật ra, đó là quá trình yêu cầu các khuyến nghị dựa trên các mối quan hệ xã hội, gọi tắt là khuyến nghị xã hội.

Việc nghiên cứu phát triển các phương pháp khuyến nghị xã hội đã thu hút nhiều quan tâm nghiên cứu của cộng đồng từ khi các mạng xã hội ra đời và phát triển. Tiếp cận khuyến nghị xã hội bổ sung việc xem xét sở thích của người dùng dựa trên việc phân tích, khai thác các thông tin từ các mạng xã hội như: đánh giá của người dùng từ các trang mạng xã hội, mức độ ảnh hưởng của các mối quan hệ xã hội đến sở thích của người dùng. Với khuyến nghị xã hội, bên cạnh ma trận đánh giá của người dùng, thông tin đầu vào còn có ma trận thể hiện mối quan hệ giữa những người dùng (hình 1.8).

Tác giả Tang và cộng sự đã khảo sát và hệ thống lại những tiếp cận khuyến nghị truyền thống, thảo luận một số hướng nghiên cứu mới cho hệ khuyến nghị và đã tập



Hình 1.8: Minh họa khuyen nghị xã hội

trung vào hướng phân tích mạng xã hội cho khuyến nghị hay khuyến nghị xã hội (Social Recommendation) [116]. Tác giả cũng đã phát biểu một cách hình thức bài toán khuyến nghị xã hội và xác định khuyến nghị xã hội vẫn đang ở giai đoạn của những bước đi đầu tiên cần đầu tư nghiên cứu phát triển. Tác giả cũng đã chỉ ra một số hướng nghiên cứu tiềm năng nhằm cải tiến các khuyến nghị xã hội như: khai thác tính không đồng nhất của các mạng xã hội và những liên kết phụ thuộc yếu, khai thác các mối quan hệ tiềm ẩn, quan hệ lòng tin, xem xét thông tin thời gian của các đánh giá.

Tác giả Aranda và cộng sự đã trình bày một hệ thống khuyến nghị dựa trên mạng xã hội. Hệ thống được hiện thực dùng dữ liệu thu thập từ trang web trò chơi BGG.¹ Mục đích của hệ thống là giới thiệu các trò chơi mới cho các thành viên của BGG thông qua quá khứ đánh giá và các mối quan hệ của họ [8].

Tác giả Ma và cộng sự đã dựa trên quan điểm "mạng xã hội của người dùng sẽ ảnh hưởng đến hành vi của họ trên web" để đề xuất phương pháp khuyến nghị xã hội mới. Phương pháp đề xuất hợp nhất ma trận đánh giá với mạng xã hội người dùng bằng cách thừa số hóa ma trận đánh giá. Kết quả thực nghiệm của họ đã cho thấy phương pháp khuyến nghị xã hội đề xuất cho kết quả tốt hơn những thuật toán CF phổ biến, và nó có thể mở rộng cho những tập dữ liệu lớn [78].

Tác giả Esslimani và cộng sự đã đề xuất một tiếp cận lọc cộng tác mới dựa trên mạng hành vi (Behavioral Network Collaborative Filtering - BNCF) dùng những mẫu

¹<http://boardgamegeek.com/>

định liên quan đến định hướng để mô hình quan hệ giữa những người dùng và áp dụng kỹ thuật lan truyền trong mạng xã hội để khai thác thêm những liên kết tiềm ẩn thông qua mạng hành vi. Mục đích của BNCF là kết hợp cả những liên kết mới, tiềm ẩn để thực hiện tiên đoán, cải tiến chất lượng khuyến nghị. Độ chính xác tiên đoán dùng BNCF được đánh giá thông qua các tập dữ liệu thực. Kết quả thực nghiệm của họ cho thấy lợi ích của việc khai thác các liên kết mới, tiềm ẩn để tính toán tiên đoán. Các tác giả đã cho thấy BNCF cải tiến độ chính xác tiên đoán dựa trên phân tích lỗi trung bình MAE (Mean Absolute Error) và HMAE (High MAE) [40].

Tác giả Davoodi và cộng sự đã đề xuất một phương pháp lai, kết hợp những đặc điểm của các thuật toán dựa trên nội dung vào hệ thống lọc CF dựa trên mạng xã hội để khuyến nghị chuyên gia [37]. Phương pháp của họ nhằm cải tiến độ chính xác tiên đoán dựa trên việc xem xét khía cạnh xã hội, hành vi của chuyên gia. Phương pháp của họ bao gồm một số bước chính như: xác định những cộng đồng xã hội của những chuyên gia, chỉ định những thành viên đại diện trong mỗi cộng đồng dùng các độ đo trung tâm (centrality measure) (Kirchhoff et al. [62]), và cuối cùng là khuyến nghị những chuyên gia đại diện cộng đồng liên quan nhiều nhất với người dùng đang có nhu cầu tìm kiếm chuyên gia. Độ chính xác tiên đoán trong thực nghiệm của họ cho thấy, việc dùng thông tin mạng xã hội đã có cải thiện so với phương pháp không dùng thông tin mạng xã hội. Kết quả tiên đoán là 78.4% so với 75.6% với độ đo Precision cho Top5 (P@5) [37].

Các tác giả Abbasi và Altmann đã đề xuất hệ thống AcaSoNet dùng để xác định và quản lý mạng xã hội của những người nghiên cứu [1]. AcaSoNet được dùng để phân tích, đánh giá hoạt động của những người nghiên cứu. AcaSoNet đã rút trích và thu thập dữ liệu của những người nghiên cứu từ www, đồng thời cho phép người dùng cập nhật dữ liệu của họ. Mạng xã hội của những người nghiên cứu được xác định dựa trên kho dữ liệu bài báo thu thập. Hiện nay AcaSoNet chỉ cung cấp giao diện và dữ liệu phục vụ cho phân tích mạng xã hội. Sắp đến, họ dự định sẽ cung cấp thêm các dịch vụ khác như khuyến nghị bài báo, gom cụm cộng đồng nghiên cứu, v.v...

Tác giả Tang và cộng sự đã đề xuất hệ thống ArnetMiner nhằm rút trích và khai thác những mạng xã hội học thuật [118]. ArnetMiner bao gồm các module chính như: (1) Rút trích hồ sơ người nghiên cứu từ www; (2) Tích hợp dữ liệu bài báo từ các thư

viện số hiện có; (3) Mô hình hóa các mạng xã hội học thuật, chủ yếu là mạng đồng tác giả và mạng trích dẫn; (4) Cung cấp các dịch vụ tìm kiếm trực trên các mạng xã hội học thuật. Vào thời điểm công bố, ArnetMiner có 448.470 hồ sơ nghiên cứu viên, hệ thống cũng đã tích hợp hơn 1.000.000 bài báo khoa học.

Tác giả Brandao và cộng sự đã xem xét hai nguyên tắc xã hội là ‘homophily’ (mục 1.4.4.1) và ‘proximity’ (mục 1.4.4.1) để đề xuất hai độ đo mới cho khuyến nghị cộng tác trong mạng xã hội học thuật [23]. Hai độ đo đề xuất lần lượt xem xét yếu tố cơ quan công tác và thông tin vị trí địa lý của nghiên cứu viên. Đồng thời nhóm tác giả cũng dựa trên những khái niệm xã hội như: tính mới, tính đa dạng, tính bao phủ, để đề xuất các độ đo mới nhằm đánh giá kết quả khuyến nghị. Kết quả thực nghiệm của họ cho thấy các phương pháp đề xuất cho kết quả tốt hơn những phương pháp phổ biến hiện nay.

Năm 2008, Ijad Madisch và cộng sự đã xây dựng và phát triển hệ thống mạng xã hội trong lĩnh vực học thuật, đặt tên là ResearchGate¹. ResearchGate là một trang mạng xã hội học thuật, giúp các nhà khoa học, nghiên cứu viên chia sẻ bài báo khoa học, kiến thức và kinh nghiệm nghiên cứu thông qua việc cập nhật danh sách bài báo công bố, thảo luận, hỏi đáp. Bên cạnh đó ResearchGate giúp các nghiên cứu viên tìm kiếm việc làm phù hợp với kinh nghiệm và năng lực nghiên cứu, hỗ trợ tìm kiếm chuyên gia để hợp tác nghiên cứu. ResearchGate thực hiện khá hiệu quả việc chia sẻ thông tin học thuật. Thông tin bài báo trên ResearchGate được chia sẻ, khuyến nghị cho người dùng thông qua các mối quan hệ tường minh (quan hệ follow) do người dùng chủ động cập nhật. ResearchGate chứa rất nhiều thông tin học thuật liên quan các nhà khoa học, nghiên cứu viên, bài báo, các trường, viện. Tuy nhiên, theo hiểu biết của nghiên cứu sinh thì dữ liệu của ResearchGate hiện không chia sẻ, cho phép tải về để thực hiện các nghiên cứu. Bên cạnh đó, chức năng khuyến nghị của ResearchGate còn khá đơn giản, chỉ xem xét mối quan hệ tường minh, chưa xem xét các quan hệ tiềm ẩn, trung gian, cũng như các quan hệ học thuật khác như quan hệ trích dẫn, quan hệ hợp tác giữa các cơ quan và những đặc trưng khác của nghiên cứu viên.

Một vấn đề khá quan trọng khác cần xem xét khi đề cập đến các mối quan hệ xã hội đó là khái niệm lòng tin (trust). Lòng tin có thể xem là thuộc tính của quan hệ

¹<https://www.researchgate.net/about>

xã hội. Theo Touhid Bhuiyan, 2013 [18], có nhiều định nghĩa khác nhau cho khái niệm lòng tin, nhưng định nghĩa được đa số cộng đồng trích dẫn và sử dụng là định nghĩa của nhà xã hội học Dasgupta. Lòng tin là sự mong đợi của một người về những hành động của người khác mà có ảnh hưởng đến quyết định, lựa chọn của họ [35]. Theo Piotr Sztompka, 1999 [115], lòng tin gồm hai thành phần chính là tin tưởng (belief) và cam kết (commitment). Tức một người sẽ tin tưởng rằng một người khác sẽ hành động theo một cách nhất định và đặt lòng tin vào họ, nhưng sự tin tưởng không thôi thì chưa đủ để có lòng tin. Lòng tin được đặt vào một ai đó khi sự tin tưởng đạt tới mức độ làm nền tảng cho một cam kết thực hiện một hành động cụ thể. Gần đây, lòng tin đã trở thành một chủ đề nghiên cứu quan trọng trong nhiều lĩnh vực như: xã hội học, tâm lý học, và cả tin học.

Stephen Marsh là một trong những người đi tiên phong trong việc khai thác lòng tin trong tính toán khoa học [81]. Gần đây, lòng tin đã thu hút nhiều quan tâm nghiên cứu của cộng đồng trong việc phát triển các hệ thống khuyến nghị trực tuyến. Người dùng thường sẽ tin tưởng và dễ dàng chấp nhận các khuyến nghị từ bạn bè, người thân hơn là những người lạ khác, ngay cả khi hệ khuyến nghị có những đề xuất hữu ích và chất lượng. Bên cạnh đó, lòng tin được sử dụng để cải tiến các phương pháp khuyến nghị truyền thống. Việc sử dụng quan hệ lòng tin giúp các hệ khuyến nghị có thể đương đầu với những khó khăn, thách thức như: ma trận đánh giá thưa, khởi động lạnh (cold-start).

Paolo Massa và Paolo Avesani đã đề xuất thay thế bước tính toán tương tự người dùng trên ma trận đánh giá bằng độ đo lòng tin giữa những người. Họ đề xuất thuật toán lan truyền lòng tin trên mạng và tính mức độ lòng tin giữa những người dùng. Kết quả thực nghiệm trên tập dữ liệu Epinions cho thấy việc khai thác lòng tin cải tiến độ chính xác khuyến nghị [82]. Hao Ma và cộng sự đã nghiên cứu đề xuất phương pháp tối ưu dựa trên kết hợp cả các mối quan hệ lòng tin và không tin (distrust) nhằm cung cấp các khuyến nghị chính xác và thực tế cho người dùng. Nhóm tác giả cũng đã thực nghiệm trên tập dữ liệu Epinions và cho thấy hương pháp của họ tốt hơn hẳn các phương pháp hiện có trên tập dữ liệu này [77]. Lahiru S. Gallego và cộng sự đã nghiên cứu khai thác lòng tin để hướng đến phát triển hệ khuyến nghị cho các dịch phần mềm trực tuyến [44].

Trong lĩnh vực học thuật, theo hiểu biết của nghiên cứu sinh thì khái niệm lòng tin, những mối quan hệ tiềm ẩn, yếu tố xu hướng sở thích nghiên cứu, xu hướng quan hệ chưa được đề cập và khai thác để phát triển các phương pháp khuyến nghị thông tin cho nghiên cứu viên.

Nhìn chung các phương pháp khuyến nghị xã hội có những thuận lợi và khó khăn như sau:

- **Thuận lợi:** độc lập lĩnh vực, không phụ thuộc vào đối tượng khuyến nghị. Có thể đương đầu với vấn đề ma trận đánh giá thưa, người dùng mới. Kết hợp thông tin quan hệ xã hội với các thông tin khác có thể cải tiến độ chính xác tiên đoán, chất lượng khuyến nghị.
- **Khó khăn, thách thức:** Việc xác định các mối quan hệ xã hội (rõ ràng và tiềm ẩn). Làm thế nào để lượng hóa được mức độ ảnh hưởng của các mối quan hệ xã hội đến sở thích, hành vi và quyết định của người dùng. Làm thế nào để kết hợp thông tin về các mối quan hệ xã hội với những thông tin khác để phát triển các phương pháp lai.

Đây là hướng tiếp cận chính của luận án nhằm phát triển các phương pháp khuyến nghị mới để hỗ trợ nghiên cứu viên đương đầu với tình trạng quá tải thông tin trong lĩnh vực học thuật. Luận án chọn tiếp cận phân tích mạng xã hội nhằm giải quyết những khó khăn, thách thức mà các phương pháp truyền thống đang đương đầu, đó là: vấn đề người dùng mới, đối tượng khuyến nghị mới (khởi động lạnh); ma trận đánh giá thưa, thiếu hoặc không có thông tin đánh giá của người dùng; vấn đề chất lượng tiên đoán, khuyến nghị.

1.5 Các phương pháp đánh giá hệ khuyến nghị

1.5.1 Phương pháp thiết lập thực nghiệm

Các tác giả Gunawardana và Shani đã chỉ ra rằng, thông thường sẽ có hai cách thiết lập thực nghiệm để đánh giá hệ khuyến nghị. Đó là thiết lập đánh giá online và thiết lập đánh giá off-line, gọi tắt là đánh giá online và đánh giá off-line [49].

-
- **Đánh giá Online:** Mục đích của đánh giá online là đo lường sự thay đổi hành vi người dùng khi họ tương tác với hệ khuyến nghị. Hay nói cách khác, chúng ta cần xem xét hệ khuyến nghị đã ảnh hưởng như thế nào đến việc thay đổi hành vi người dùng, khi họ tương tác với nó. Ưu điểm của đánh giá online là phản ánh đúng hiệu năng thật sự của hệ khuyến nghị khi người dùng tương tác với nó.

Với đánh giá online thì hệ khuyến nghị cần được triển khai sử dụng thật sự. Nhưng khó khăn là chúng ta cần thực nghiệm và kiểm tra một loạt nhiều thuật toán khuyến nghị xem cái nào là phù hợp trước khi đưa vào sử dụng. Đây giống như trường hợp "Quả trứng và con gà". Bên cạnh đó, việc đánh giá online cần phải xem xét vấn đề đánh giá ở nhiều góc độ khác nhau như: cố định giao diện, thay đổi thuật toán và ngược lại. Như vậy, để thực hiện đánh giá online sẽ tốn nhiều chi phí. Vì vậy hầu hết các nghiên cứu hiện nay đều sử dụng phương pháp đánh giá off-line.

- **Đánh giá off-line:** Với đánh giá off-line, các nghiên cứu cần đưa ra giả thuyết "Giả sử kết quả thực nghiệm off-line là tương quan với hành vi người dùng online". Tức chúng ta mong đợi và giả sử rằng nếu như đánh giá off-line đạt hiệu quả tốt thì khi triển khai Online cũng vậy.

Để thực hiện đánh giá off-line, cần thiết phải giả lập hay mô phỏng tốt quá trình online khi hệ thống thực hiện khuyến nghị và người dùng sử dụng các kết quả khuyến nghị. Thông thường, các nghiên cứu phải lưu lại dữ liệu lịch sử người dùng. Sau đó ẩn đi những tương tác của người dùng để xem người dùng sẽ quan tâm, hay hành động như thế nào với các đối tượng sẽ được khuyến nghị thông qua việc áp dụng những phương pháp khuyến khác nhau.

1.5.2 Độ đo đánh giá

Các độ đo đánh giá hệ khuyến nghị có nguồn gốc từ các phương pháp đánh giá trong lĩnh vực học máy và truy vấn thông tin. Cùng với việc chọn lựa cách thiết lập đánh giá, là online hay off-line, các tác giả Gunawardana và Shani đã phân loại hệ khuyến nghị thành 3 nhóm chính dựa trên công việc thực hiện khuyến nghị và chỉ ra những độ đo đánh giá phù hợp với mỗi nhóm [49].

1.5.2.1 Tiên đoán đánh giá

Tiên đoán đánh giá tức hệ thống yêu cầu tiên đoán những đánh giá của người dùng dựa trên tập hợp các đối tượng khuyến nghị cho trước. Những ứng dụng phổ biến thường thấy với tiên đoán đánh giá là các website thương mại điện tử như Netflix¹, CNET². Chẳng hạn với Netflix, khi người dùng duyệt danh sách các phim mới thì hệ thống sẽ gán một giá trị tiên đoán đánh giá cho mỗi phim này. Những phim có giá trị tiên đoán đánh giá cao xem như ưu tiên khuyến nghị cho người dùng.

Dộ đo thường dùng cho tiên đoán đánh giá là lỗi bình phương trung bình (Root of The Mean Square Error), viết tắt là RMSE (công thức 1.20) và các biến thể của nó như MAE (Mean Average Error), NMAE (Normalized Mean Average Error) [26, 105].

$$RMSE = \sqrt{\frac{\sum_{(u,p) \in K} (r_{u,p} - v_{u,p})^2}{n}} \quad (1.20)$$

Trong đó,

- $r_{u,p}$: giá trị đánh giá của người dùng u trên đối tượng p hệ thống tiên đoán được.
- $v_{u,p}$: giá trị đánh giá thật sự (đúng) của người dùng u trên đối tượng p .
- $K = \{(u, p)\}$: tập các đánh giá của người dùng u trên đối tượng p cần tiên đoán.
- $n = |K|$: kích thước tập K .

RMSE được đánh giá là phù hợp cho công việc tiên đoán. Tuy nhiên nó thật sự phù hợp khi chúng ta không cần phân biệt giữa các lỗi đánh giá. Chẳng hạn, giá trị đánh giá thật sự của người dùng u trên đối tượng p là 2, nhưng kết quả hai phương pháp khác nhau cho ra lần lượt là 1 và 3. Khi đó tính độ lệch dùng RMSE cho kết quả như nhau. Nhưng thật ra ý nghĩa khuyến nghị sẽ khác nhau. Hay nói rõ hơn, phương pháp cho ra giá trị tiên đoán là 3 sẽ ưu tiên khuyến nghị đối tượng p cho u , trong khi phương pháp cho ra giá trị đánh giá là 1 sẽ không mong muốn khuyến nghị p cho u .

¹www.netflix.com

²www.cnet.com

1.5.2.2 Tối ưu tính hữu ích của hệ thống khuyến nghị

Một nhóm các công việc khác của các hệ khuyến nghị là tối ưu tính hữu ích của hệ thống khuyến nghị. Như vậy đối với các công việc khuyến nghị dạng này là cần định nghĩa một hàm hữu ích, và xây dựng thuật toán để tối ưu hàm hữu ích này. Với mỗi bài toán khuyến nghị và mục tiêu khuyến nghị cụ thể, một hàm hữu ích sẽ được định nghĩa và tìm thuật toán để tối ưu hàm này. (Gunawardana and Shani [49]) đã phân ra hai dạng hàm hữu ích là hướng lợi nhuận và hướng người dùng.

- **Hướng lợi nhuận:** Nếu xét ở góc độ lợi nhuận của nhà cung cấp dịch vụ, chẳng hạn đối với một số trang web thương mại điện tử thì họ cần hệ thống khuyến nghị mà có thể cực đại hóa khả năng bán hàng, nhằm đạt được doanh số bán hàng cao nhất hoặc lợi nhuận nhiều nhất. Đối với các trang web tin tức quảng cáo thì hệ khuyến nghị cần khuyến nghị các mẫu tin dài, để giữ người dùng ở lại với trang web càng lâu càng tốt.
- **Hướng người dùng:** Đôi khi người dùng quan tâm đến những đối tượng mới, chưa biết hơn những đối tượng quen thuộc, đã biết. Việc khuyến nghị các đối tượng quen thuộc đôi khi là dư thừa, không cần thiết. Vì vậy, hệ thống khuyến nghị hướng người dùng cần xem xét tính mới, tính đa dạng khi thực hiện khuyến nghị các đối tượng. Gần đây đã có một số các nghiên cứu bắt đầu xem xét tính mới, đa dạng khi thực hiện khuyến nghị. Các tác giả Zhang và Hurley đã nghiên cứu về tính mới, tính đa dạng trong khuyến nghị và đề xuất một phương pháp đánh giá cho phép phân tích hiệu năng của nhiều thuật toán khác nhau dưới góc độ khuyến nghị những đối tượng không những liên quan, nhưng phải mới [129, 54]. Bên cạnh đó, Vargas và Castells cũng xác định những khái niệm nền tảng cho tính mới, đa dạng trong khuyến nghị và đề xuất một framework hỗ trợ phát triển các độ đo để đánh giá tính mới, đa dạng cho các thuật toán khuyến nghị [122].

1.5.2.3 Khuyến nghị các đối tượng tốt

Công việc chung nhất của hầu hết các hệ khuyến nghị là đề xuất một danh sách các đối tượng được cho là phù hợp với sở thích người dùng. Kiểu khuyến nghị này thường thấy

ở các trang web thương mại điện tử như: Amazon, Netflix. Khi người dùng chọn một sản phẩm nào đó, thì hệ thống sẽ khuyến nghị một danh sách các đối tượng khác được cho là liên quan đến nhu cầu, sở thích của người dùng. Theo tác giả Gunawardana và Shani [49], các công việc khuyến nghị các đối tượng tốt có thể phân thành hai loại: (1) Một là khuyến nghị một vài đối tượng tốt (recommending some good items). Đây là các bài toán như khuyến nghị sách, phim, tin tức. Những bài toán đó, có một số lượng vô cùng lớn các phim, sách, tin tức liên quan đến người dùng cần khuyến nghị. Và hệ thống khuyến nghị chỉ nên khuyến nghị một vài đối tượng liên quan. Tránh khuyến nghị những đối tượng không liên quan sở thích hơn là phải khuyến nghị tất cả các đối tượng liên quan sở thích của người dùng; (2) Hai là tất cả các đối tượng khuyến nghị phải và nên tốt (recommending all good items). Chẳng hạn, bài toán khuyến nghị bài báo trích dẫn. Hệ thống cần tìm tất cả các bài báo tốt liên quan cần trích dẫn để khuyến nghị cho người dùng, tránh khuyến nghị các bài báo không liên quan đến công việc trích dẫn.

Để đánh giá danh sách các đối tượng khuyến nghị mà hệ thống đề xuất cho người dùng có phù hợp hay không, các nghiên cứu hiện nay thường sử dụng các độ đo có nguồn gốc truy vấn thông tin như: độ chính xác Precision (P), độ bao phủ Recall (R), và những độ đo có xem xét thứ tự xếp hạng của các đối tượng khuyến nghị trong danh sách đề xuất, đó là Mean Average Precision (MAP) [128], NCDG [58].

1.6 Khó khăn, thách thức và xu hướng

1.6.1 Khó khăn, thách thức

Một số khó khăn, thách thức chính đối với các phương pháp khuyến nghị truyền thống, phổ biến hiện nay có thể kể đến như sau:

- Dữ liệu lớn. Không gian người dùng và đối tượng khuyến nghị là rất lớn.
- Độ chính xác, chất lượng khuyến nghị.
- Vấn đề ma trận đánh giá thưa, tức số đánh giá quan sát được rất ít so với số đánh giá cần tiên đoán để khuyến nghị. Điều đó ảnh hưởng đến độ chính xác tiên đoán, chất lượng khuyến nghị.

-
- Vấn đề khởi động lạnh (cold start). Quan sát thiếu hay không quan sát được một số thông tin về sở thích, đánh giá của người dùng, cũng như các đối tượng khuyến nghị. Hoặc làm thế nào để thực hiện khuyến nghị cho những người dùng mới hay đối tượng khuyến nghị mới.
 - Các phương pháp đánh giá kết quả khuyến nghị.

1.6.2 Xu hướng mới cho hệ khuyến nghị

Trong các nghiên cứu khảo sát về hệ khuyến nghị của Adomavicius và Tuzhilin [5], cũng như của Bobadilla và cộng sự [22], các tác giả đã chỉ ra những xu hướng mà cộng đồng đang quan tâm để phát triển các phương pháp khuyến nghị mới nhằm giải quyết những khó khăn, thách thức của các phương pháp truyền thống. Một số xu hướng khuyến nghị mới có thể kể đến như sau:

- Stefanidis và cộng sự cũng chỉ ra rằng, các phương pháp truyền thống chưa quan tâm xem xét sự ảnh hưởng của yếu tố thời gian, xu hướng đến kết quả khuyến nghị như thế nào [109].
- Làm thế nào để kết hợp thông tin xã hội rõ ràng, cũng như tiềm ẩn vào các phương pháp truyền thống.
- Sử dụng thông tin nhận biết ngữ cảnh (context-aware) để thực hiện khuyến nghị. Liên quan đến tiếp cận khai thác thông tin ngữ cảnh, Gediminas Adomavicius và cộng sự khảo sát phân tích các nghiên cứu khác nhau [3, 6]. Thông tin ngữ cảnh giúp các hệ khuyến nghị cung cấp những khuyến nghị phù hợp với người dùng theo thời gian, địa điểm. Cùng với sự phát triển mạnh mẽ của công nghệ di động, hướng tiếp cận khai thác thông tin ngữ cảnh được đánh giá là rất tiềm năng đối với các hệ khuyến nghị trên thiết bị di động.
- Tiếp cận lai nhằm giải quyết những hạn chế của mỗi phương pháp khác nhau [25, 5, 22].
- Thu thập, sử dụng thông tin tiềm ẩn của người dùng từ Internet để xác định sở thích của họ.

- Trong các nghiên cứu [49, 105, 106], nhóm tác giả Gunawardana và Shani đã chỉ ra rằng, với mỗi bài toán cần có phương pháp khuyến nghị phù hợp. Đồng thời với mỗi phương pháp khuyến nghị, cũng cần có phương pháp đánh giá phù hợp. Tức phương pháp đánh giá sẽ quyết định độ chính xác tiên đoán, kết quả của phương pháp khuyến nghị. Việc nghiên cứu phát triển các phương pháp đánh giá kết quả khuyến nghị cũng nằm trong xu hướng và quan tâm của cộng đồng.

Tóm lại, những ưu điểm, hạn chế của các cách tiếp cận phổ biến hiện nay và xu hướng có thể trình bày tóm tắt trong bảng 1.2 ở cuối chương.

Bảng 1.2: Tóm tắt ưu nhược điểm những tiếp cận phổ biến và xu hướng nghiên cứu

Ưu điểm & Hạn chế	Tiếp cận truyền thống và xu hướng				
	Truyền thống		Xu hướng		
	Nội dung (CB)	Lọc cộng tác (CF)	CB kết hợp CF	Phân tích mạng xã hội	Khai thác thông tin ngữ cảnh
Phù hợp đối tượng dạng văn bản	Có	Có	Có	Có	Có
Đa dạng đối tượng khuyến nghị	Không	Có	Có	Có	Có
Hạn chế về phân tích nội dung	Có	Không	Không	Không	Không
Có thể đa dạng hóa khuyến nghị. (Ngoài lĩnh vực quan sát)	Không	Có	Có	Có	Có
Người dùng mới. (Vấn đề khởi động lạnh)	Có	Có	Có	Có	Có
Đối tượng khuyến nghị mới. (Vấn đề khởi động lạnh)	Không	Có	Có	Có	Có
Vấn đề ma trận đánh giá thưa.	Không	Có	Có	Có	Có
Có khả năng giải quyết vấn đề ma trận thưa và khởi động lạnh	Không	Không	Có	Có	Có
Khó khăn chung:					
(*) Dữ liệu lớn.					
(*) Độ chính xác, chất lượng khuyến nghị chưa cao.					
(*) Dữ liệu đánh giá thưa.					
(*) Chưa có phương pháp tốt để đánh giá kết quả, chất lượng khuyến nghị.					
(*) Vấn đề khởi động lạnh.					

1.7 Kết chương

Chương này đã trình bày tổng quan về bài toán khuyến nghị trong trường hợp tổng quát và tập trung phân tích ưu điểm, hạn chế của các phương pháp khuyến nghị truyền thống. Bên cạnh những tiếp cận truyền thống cho hệ khuyến nghị, chương này cũng trình bày về xu hướng nghiên cứu, cũng như những tiếp cận mới cho hệ khuyến nghị mà cộng đồng đang quan tâm nghiên cứu, đó chính là các phương pháp tiếp cận lai, kết hợp phân tích mạng xã hội, xu hướng sử dụng thông tin nhận biết ngữ cảnh, cũng như khai thác thông tin tiềm ẩn của cá nhân từ "Internet of Things". Trên cơ sở khảo sát các nghiên cứu liên quan, mục tiêu và phạm vi luận án là tập trung pháp triển các phương pháp khuyến nghị lai dựa trên tiếp cận khai thác quan hệ xã hội, một tiếp cận có thể cải tiến độ chính xác tiên đoán, chất lượng khuyến nghị, có thể giải quyết được vấn đề dữ liệu đánh giá thừa, trường hợp người dùng mới (chưa có thông tin lịch sử về sở thích của người dùng).

Trong lĩnh vực học thuật, những kết quả thống kê (hình 0.1) cho thấy thông tin bài báo khoa học bùng nổ rất nhanh chóng, gây rất nhiều khó khăn cho những người làm nghiên cứu trong việc tìm kiếm tài liệu, chuyên gia, để thực hiện các công việc nghiên cứu. Vì vậy luận án hướng đến phát triển các phương pháp khuyến nghị nhằm hỗ trợ cộng đồng học thuật. Những bài toán đặt ra trong phạm vi của luận án thuộc nhóm phát triển các phương pháp khuyến nghị các đối tượng tốt cho người dùng. Tức là, để đánh giá kết quả khuyến nghị luận án đã sử dụng các độ đo như: độ chính xác Precision (P), độ chính xác top N (P@N), Recall (R), Mean Everage Precision (MAP).

Với dữ liệu, thông tin thu thập được, luận án chọn tiếp cận khai thác các mối quan hệ xã hội kết hợp yếu tố thời gian, không gian cho khuyến nghị. Để thực hiện được việc khai thác các mối quan hệ xã hội trong học thuật, chương tiếp theo sẽ trình bày việc rút trích, mô hình hóa các mạng xã hội trong lĩnh vực học thuật từ kho dữ liệu bài báo khoa học.

Chương 2

XÁC ĐỊNH VÀ MÔ HÌNH HÓA MẠNG XÃ HỘI HỌC THUẬT

2.1 Giới thiệu

Với mục tiêu phát triển các phương pháp khuyến nghị trong lĩnh vực học thuật dựa trên tiếp cận phân tích mạng xã hội, luận án cần xem xét: (1) Chuẩn bị kho dữ liệu học thuật đủ lớn và đủ phong phú; (2) Xác định và mô hình hóa các mối quan hệ xã hội học thuật; (3) Khai thác các mối quan hệ học thuật để phát triển các phương pháp khuyến nghị.

Về các kho dữ liệu học thuật thì các nghiên cứu phổ biến hiện nay thực hiện trên nhiều tập dữ liệu khác nhau được rút trích từ nhiều nguồn khác nhau. Chẳng hạn, Chen và cộng sự [28, 29, 27], S. D. Gollapalli và cộng sự [48], thì tiến hành thử nghiệm trên dữ liệu được trích xuất từ CiteSeerX¹; Trong khi đó, Tang và cộng sự [117], Sugiyama và cộng sự [111, 112, 113], Luong và cộng sự [75, 76], tiến hành thực nghiệm trên tập dữ liệu bài báo khoa học được trích xuất từ các hội thảo chuyên ngành và gán nhãn thủ công. Một số nghiên cứu phổ biến khác thì trích xuất từ kho dữ liệu khoa học DBLP² để xây dựng tập dữ liệu thực nghiệm. Nói chung, theo hiểu biết của chúng tôi thì hiện nay chưa có những tập dữ liệu chuẩn (benchmark) đối với các bài toán khuyến nghị trong lĩnh vực học thuật. Bên cạnh đó, cho đến nay thì những thông tin có được từ các tập dữ liệu phổ biến cho download như DBLP, CiteSeerX vẫn còn khá hạn chế, thiếu nhiều thông tin cần thiết (bảng 2.1). Vì vậy, việc xây dựng và làm giàu một kho

¹<http://csxstatic.ist.psu.edu/about/data>

²<http://dblp.uni-trier.de/xml/>

dữ liệu khoa học đủ lớn và đủ phong phú và công bố rộng rãi cho cộng đồng tham khảo để tiến hành các đánh giá thực nghiệm là cần thiết.

Bảng 2.1: Thông tin bài báo sẵn có từ DBLP, CiteSeerX

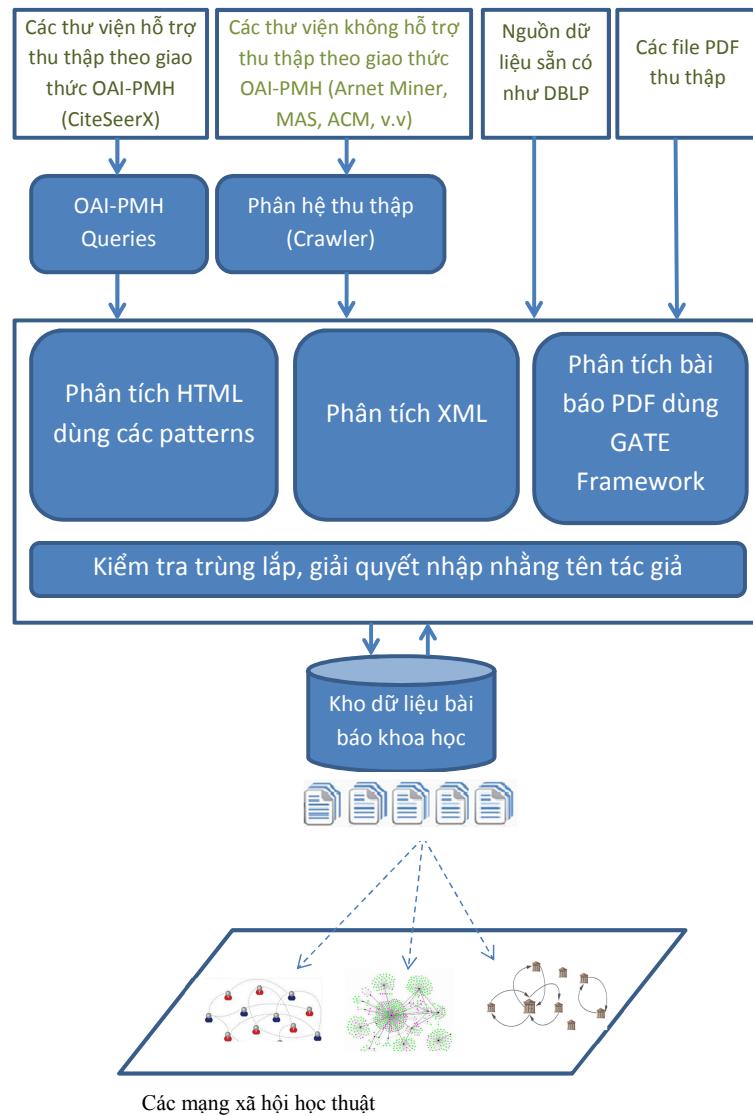
Thông tin bài báo	DBLP	CiteSeerX
Tiêu đề	Có	Có
Tác giả	Có	Có
Cơ quan công tác	Không	Không
Tóm tắt	Không	Có
Nơi công bố	Có	Có
Năm công bố	Có	Có
Từ khóa	Không	Có

Vấn đề tiếp theo là làm thế nào để xác định được các cấu trúc xã hội từ kho dữ liệu học thuật, cũng như lượng hóa được các mối quan hệ để có thể tính được mức độ ảnh hưởng của cộng đồng xung quanh đối với từng nghiên cứu viên và ngược lại. Đây vẫn là một thách thức lớn đối với các nghiên cứu hiện nay về phân tích mạng xã hội. Luận án đã tìm cách xác định và mô hình hóa các cấu trúc xã hội từ kho dữ liệu bài báo khoa học, gọi tắt là mô hình các mạng xã hội học thuật ASN (Academic Social Networks). Liên quan đến việc khai thác mô hình ASN, luận án đã phát triển các phương pháp tính toán mới (thành phần M trong mô hình ASN) làm cơ sở cho tính toán phát triển các phương pháp khuyến nghị trong lĩnh vực học thuật.

Chương này sẽ tập trung trình bày 2 phần chính: (1) Giải pháp, kết quả của việc xây dựng và làm giàu kho dữ liệu học thuật; (2) Mô hình các mạng xã hội học thuật ASN, cũng như các phương pháp lượng hóa trên các mạng xã hội học thuật ASN. Kết quả liên quan trong chương này đã được công bố trong các công trình: [CT5], [CT6], [CT7], [CT9].

2.2 Xây dựng và làm giàu kho dữ liệu học thuật

Giải pháp cho việc xây dựng và làm giàu kho dữ liệu bài báo khoa học là rút trích, tích hợp từ nhiều nguồn không đồng nhất. Sau khi có kho dữ liệu học thuật thì những mạng xã hội học thuật được xác định và mô hình hóa (hình 2.1).



Hình 2.1: Tích hợp dữ liệu bài báo khoa học từ nhiều nguồn không đồng nhất

2.2.1 Tích hợp từ nhiều nguồn

Mỗi thư viện số sở hữu một cơ sở dữ liệu và một nguồn thông tin riêng dựa vào nguồn và cách họ thu thập. Vì vậy, đôi khi một bài báo hoặc các thông tin liên quan có thể tìm thấy trong thư viện này, nhưng không thấy trong thư viện kia. Để có một nguồn thông tin đầy đủ và phong phú phục vụ cho xây dựng, phân tích các mạng xã hội học thuật, luận án đã tiến hành nghiên cứu phương pháp và xây dựng công cụ cho việc tích hợp dữ liệu khoa học từ nhiều nguồn không đồng nhất.

Những nguồn dữ liệu khoa học được xem xét thu thập như: các bài báo khoa học

dạng PDF, một số cơ sở dữ liệu khoa học và thư viện số trong lĩnh vực học thuật. Các hệ thống thư viện số này có thể phân thành hai nhóm chính: (1) Nhóm hệ thống ‘mở’ như DBLP, CiteSeerX, ArnetMiner¹ là nhóm cho phép download miễn phí các bài báo mà họ đã lập chỉ mục (2) Nhóm thu phí truy cập, bao gồm các hệ thống thư viện số như ACM, IEEE Explore, Elsevier, SpringerLink. Phần bên dưới là kiến trúc hệ thống và giải pháp mà luận án áp dụng để xây dựng công cụ tích hợp các nguồn dữ liệu học thuật.

2.2.2 Các thành phần chính của hệ thống

Hình 2.1 tổng quan kiến trúc hệ thống rút trích và tích hợp thông tin bài báo khoa học từ nhiều nguồn không đồng nhất. Hệ thống bao gồm các phân hệ chính sau:

- Rút trích thông tin từ bài báo dạng PDF. Với những tính năng xử lý văn bản, ngôn ngữ tự nhiên mạnh mẽ của GATE framework, luận án dùng GATE Framework để định nghĩa các luật cho việc rút trích thông tin từ các tập tin bài báo dạng PDF.²
- Phân hệ thu thập (Crawler): thu thập, rút trích thông tin bài báo từ các thư viện số trên internet. Phân hệ sẽ lập lịch, định thời và gởi các yêu cầu đến các trang web của các thư viện số. Sau đó sẽ tiến hành phân tích các trang HTML trả về để rút trích ra các bài báo và thông tin liên quan. Với những thư viện số có hỗ trợ giao thức truy cập OAI-PMH thì hệ thống sẽ gởi các truy vấn OAI-PMH để rút trích metadata của các bài báo khoa học từ những thư viện số này.
- Xử lý dữ liệu có sẵn từ DBLP: phân hệ sẽ phân tích file dữ liệu của DBLP dưới dạng XML được công bố trên internet và đưa vào trong Cơ sở dữ liệu của hệ thống.
- Xử lý nhập nhằng tên tác giả: hỗ trợ xử lý, giải quyết sự trùng lặp, nhập nhằng tên tác giả khi tích hợp dữ liệu từ nhiều nguồn.

¹<http://www.arnetminer.org/>

²<http://gate.ac.uk/>, truy cập lần cuối ngày 07/02/2014

2.2.3 Rút trích thông tin bài báo từ các tập tin PDF

2.2.3.1 Dùng luật dựa trên GATE Framework

GATE¹, một framework cung cấp nhiều từ điển, ontologies, thư viện, công cụ cho xử lý văn bản, ngôn ngữ tự nhiên hỗ trợ cho rút trích thông tin. Vì tính phổ biến, và những hỗ trợ mạnh mẽ của GATE cho xử lý văn bản, luận án quyết định chọn GATE để xây dựng các luật (dưới dạng các pattern) để rút trích thông tin từ bài báo khoa học dạng PDF.

2.2.3.2 Rút trích metadata cho mục Header và mục Reference

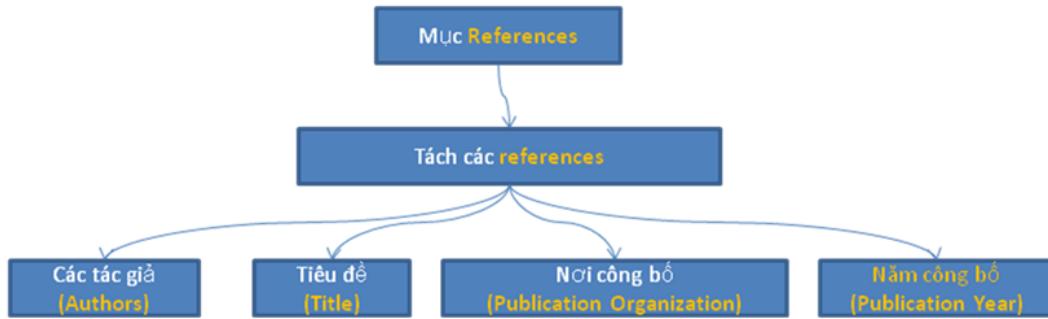


Hình 2.2: Các bước rút trích thông tin từ header của bài báo

Để rút trích thông tin từ bài báo dưới dạng PDF, luận án tiến hành phân tích phần header và reference của bài báo. Đầu tiên tập tin PDF được chuyển thành tập tin dạng văn bản. Sau đó việc phân tích mục header và mục reference được thực hiện như mô tả chi tiết trong hình 2.2 và hình 2.3 [CT9].

Phân chi tiết về các luật được định nghĩa dựa trên ngữ pháp JAPE của GATE được trình bày trong phụ lục A.

¹<https://gate.ac.uk/ie/>



Hình 2.3: Các bước rút trích thông tin từ phần reference của bài báo

2.2.4 Rút trích thông tin bài báo từ các trang web

Đối với nhóm các trang, thư viện hỗ trợ giao thức OAI-PMH như CiteSeerX¹, ArXiv² thì các hệ thống này sẽ đưa ra các thư viện và giao thức giúp ta có thể rút trích và download thông tin các bài báo khoa học từ hệ thống của họ. Ví dụ các mẫu truy vấn được gởi đến CiteSeerX được định nghĩa trong bảng 2.2.

Bảng 2.2: Các mẫu truy vấn được gởi đến CiteSeerX

Mẫu truy vấn	Mô tả
<code>http://citeseerx.ist.psu.edu/oai2?verb=Identify</code>	Truy vấn gởi đến CiteSeerX để truy vấn thông tin về kho chứa.
<code>http://citeseerx.ist.psu.edu/oai2?verb=ListMetadataFormats</code>	Truy vấn gởi đến CiteSeerX để truy vấn các định dạng có thể có từ kho chứa.
<code>http://citeseerx.ist.psu.edu/oai2?verb=ListRecords&metadataPrefix=oai_dc</code>	Truy vấn gởi đến CiteSeerX để lấy các mẫu tin.
<code>http://citeseerx.ist.psu.edu/oai2?verb=ListIdentifiers&metadataPrefix=oai_dc&from=1900-01-01&until=2013-01-01</code>	Truy vấn gởi đến CiteSeerX để truy vấn các mẫu tin metadata theo dạng Dublin Core, công bố từ 1900-01-01 đến 2013-01-01.

Đối với nhóm các thư viện số không hỗ trợ OAI-PMH như Microsoft Academic Search, tiếp cận của luận án là gởi đến các trang web này các yêu cầu dưới dạng một câu truy vấn tìm kiếm với đầu vào là một từ khóa chuyên ngành khoa học máy tính. Những từ khóa tìm kiếm lấy từ danh sách phân loại chuyên ngành Khoa học Máy tính của ACM. Bảng 2.3 mô tả các mẫu câu truy vấn được hình thành và gởi đến một số

¹<http://csxstatic.ist.psu.edu/about/data>, truy cập lần cuối ngày 07/02/2014

²<http://arxiv.org>, truy cập lần cuối ngày 07/02/2014

thư viện số phổ biến. Sau khi gửi các truy vấn đến các thư viện số, kết quả trả về là các trang dưới dạng HTML. Hệ thống sẽ tiến hành phân tích các trang HTML trả về và rút trích ra thông tin tương ứng của các bài báo khoa học [CT7].

Bảng 2.3: Các mẫu truy vấn được gửi đến các thư viện không hỗ trợ OAI-PMH tương ứng với từ khóa 'Information Extraction'

Thư viện số	Mẫu truy vấn
ACM	$http://portal.acm.org/results.cfm?query=information%20extraction&dl=ACM&coll=Portal&short=0$
IEEE Xplore	$http://ieeexplore.ieee.org/search/freeseachresult.jsp?reload=true&queryText=information%20extraction$
MAS	$http://academic.research.microsoft.com/Search?query=Information%20Extraction&SearchDomain=2$

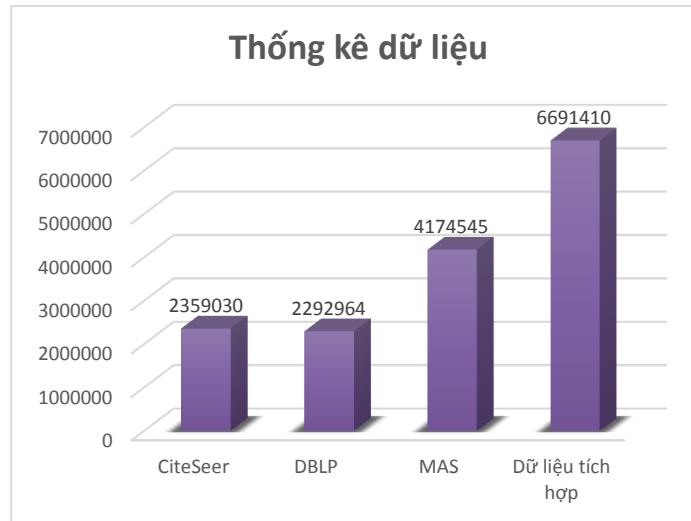
2.2.5 Kết quả kho dữ liệu tích hợp

Tính đến tháng 03/2013, luận án đã thu thập được hơn 6 triệu bài báo chuyên ngành khoa học máy tính và thông tin liên quan. Tập dữ liệu đã thu thập, tích hợp đặt tên là CSPubGuru. Kích thước và thông tin lưu trữ của CSPubGuru được trình bày trong bảng 2.4 và hình 2.4. Thông tin chi tiết về cấu trúc của kho dữ liệu CSPubGuru có thể tham khảo thêm trong phụ lục B. Tạm thời, CSPubGuru và các tập dữ liệu thực nghiệm liên quan được công bố tại: <https://sites.google.com/site/tinhuynhuit/dataset> hoặc <http://www.cspubguru.com/DownloadServlet>

Bảng 2.4: Thông tin bài báo sẵn có từ DBLP, CiteSeerX, CSPubGuru

Thông tin bài báo	DBLP	CiteSeerX	CSPubGuru
Tiêu đề	Có	Có	Có
Tác giả	Có	Có	Có
Cơ quan công tác	Không	Không	Có
Tóm tắt	Không	Có	Có
Nơi công bố	Có	Có	Có
Năm công bố	Có	Có	Có
Từ khóa	Không	Có	Có

Phần tiếp theo sẽ trình bày chi tiết về mô hình đề xuất ASN, mô hình hóa các mạng xã hội học thuật xác định từ kho dữ liệu bài báo khoa học đã tích hợp.



Hình 2.4: Kích thước kho dữ liệu tích hợp tính đến 03/2013

2.3 Xác định và mô hình hóa các mạng xã hội học thuật (ASN)

Từ kho dữ liệu học thuật thu thập được, chúng ta có thể nhận diện ra một số đối tượng nghiên cứu như: nghiên cứu viên, bài báo khoa học, các trường, các viện hay cơ quan công tác của các tác giả. Hình 2.5 minh họa các mạng xã hội có thể quan sát được từ kho dữ liệu học thuật.

Phần tiếp theo sẽ trình bày chi tiết về việc mô hình hóa các đối tượng và quan hệ giữa các đối tượng, cũng như việc khai thác mô hình ASN để phát triển các phương pháp khuyến nghị trong lĩnh vực học thuật.

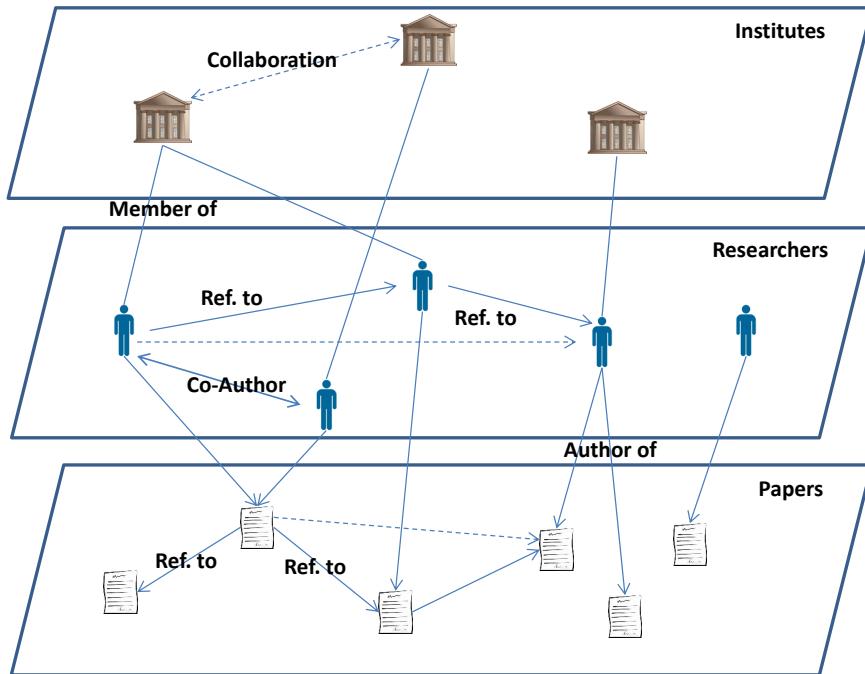
2.3.1 Thành phần chính của mô hình ASN

Mô hình đặc tả các cấu trúc mạng xã hội và lượng hóa các mối quan hệ xã hội ASN gồm các thành phần chính như sau:

$$ASN = (CoNet, CiNet_Author, CiNet_Paper, AffNet, M) \quad (2.1)$$

Trong đó:

- *CoNet*: Mạng cộng tác đồng tác giả [CT.6].
- *CiNet_Author*: Mạng trích dẫn của các tác giả [CT.6].



Hình 2.5: Minh họa các cấu trúc xã hội từ kho dữ liệu bài báo khoa học

- *CiNet_Paper*: Mạng trích dẫn của các bài báo [CT.8].
- *AffNet*: Mạng cộng tác của các trường, viện, cơ quan [CT.3].
- *M*: tập các độ đo, phương pháp tính toán mức độ quan hệ, tương quan của các đối tượng trong các mô hình ASN phục vụ cho phân tích các mạng xã hội học thuật ASN. Các phương pháp tính toán đề xuất dựa trên sự kết hợp của nhiều đặc trưng như: cấu trúc liên kết, đặc trưng nội dung của mỗi đối tượng, đặc trưng thời gian và không gian của các mối quan hệ [CT.6].

2.3.2 Mạng đồng tác giả CoNet giữa các nghiên cứu viên

Mạng đồng tác giả là một cấu trúc đồ thị gồm các thành phần như sau:

$$CoNet = (R, E_1) \quad (2.2)$$

Trong đó:

- *R*: tập hợp các nghiên cứu viên là tác giả của các bài báo khoa học.

-
- E_1 : tập các cung liên kết có hướng có trọng số chỉ mức độ quan hệ cộng tác đồng tác giả.

2.3.2.1 Cấu trúc một nghiên cứu viên

Một nghiên cứu viên $r \in R$ có cấu trúc là một bộ các thành phần, được định nghĩa như sau:

$$r = (Attr, \vec{P}_r) \quad (2.3)$$

Trong đó:

- $Attr$: tập hợp các thuộc tính (attributes) của đối tượng tác giả.
- $\vec{P}_r = (w_1, w_2, \dots, w_n)$: vector biểu diễn hồ sơ hay quan tâm nghiên cứu của một nghiên cứu viên $r \in R$.

Mô hình nghiên cứu viên: Quan tâm nghiên cứu của một nghiên cứu viên có thể tổng hợp từ các bài báo mà nghiên cứu viên công bố trong quá khứ.

$$\vec{P}_r = \sum_{i=1}^n \vec{f}_i \quad (2.4)$$

Trong đó,

- \vec{f}_p : vector biểu diễn nội dung của bài báo p .
- n : tổng số bài báo mà nghiên cứu viên r đã công bố trong quá khứ.

Mô hình nghiên cứu viên dựa trên xu hướng: Quan tâm nghiên cứu của nghiên cứu viên sẽ thay đổi theo thời gian. Chẳng hạn một nghiên cứu viên trong quá khứ quan tâm đến khai thác dữ liệu văn bản, tìm kiếm thông tin, nhưng do ảnh hưởng từ cộng đồng và nhu cầu phát triển nên gần đây lại quan tâm đến tìm kiếm thông tin thông minh, hệ khuyến nghị. Xu hướng sở thích nghiên cứu của một nghiên cứu viên thể hiện thông qua các bài báo công bố gần đây hơn là các bài quá lâu trong quá khứ. Vì vậy, luận án đề xuất mô hình quan tâm nghiên cứu của nghiên cứu viên dựa trên xu hướng như sau:

$$\vec{P}_r = \sum_{i=1}^n e^{\gamma * (t_c - t(p_i))} * \vec{f}_{p_i} \quad (2.5)$$

Trong đó,

- γ : hệ số xu hướng. ($\gamma \in [0, 1]$. Trường hợp đơn giản $\gamma = 1$)
- t_c : năm hiện tại.
- $t(p_i)$: năm công bố của bài báo p_i .
- n : Tổng số bài báo mà r công bố trong quá khứ.
- \vec{f}_p : vector biểu diễn nội dung bài báo p .

2.3.2.2 Cấu trúc cung liên kết

Một cung liên kết $e \in E_1$ là cung nối có hướng giữa 2 nghiên cứu viên r_i và r_j , ký hiệu $e(r_i, r_j)$ (với $r_i, r_j \in R$). Trọng số của $e \in E_1$ được tính dựa vào các phương pháp trong tập độ đo M .

2.3.3 Mạng trích dẫn giữa các nghiên cứu viên $CiNet_Author$

Mạng trích dẫn $CiNet_Author$ giữa các nghiên cứu viên thể hiện mức độ quan tâm của một nghiên cứu viên này với các nghiên cứu viên khác. $CiNet_Author$ là một cấu trúc đồ thị gồm các thành phần như sau:

$$CiNet_Author = (R, E_2) \quad (2.6)$$

Trong đó:

- R : tập các đỉnh của đồ thị. Mỗi đỉnh là một nghiên cứu viên.
- E_2 : tập các cặp đỉnh có hướng (x, y) thể hiện quan hệ trích dẫn. Hướng từ $x \rightarrow y$ thể hiện quan hệ nghiên cứu viên y đã được trích dẫn bởi nghiên cứu viên x , hay x có trích dẫn bài của y . Trọng số của $e \in E_2$ được tính dựa vào các phương pháp trong tập độ đo M .

2.3.4 Mạng trích dẫn giữa các bài báo $CiNet_Paper$

Mạng trích dẫn $CiNet_Paper$ giữa các bài báo là một cấu trúc đồ thị gồm các thành phần như sau:

$$CiNet_Paper = (P, E_3) \quad (2.7)$$

Trong đó:

- P : tập các đỉnh của đồ thị. Mỗi đỉnh là một bài báo.
- E_3 : tập các cặp đỉnh có hướng (x,y) thể hiện quan hệ trích dẫn. Hướng từ $x \rightarrow y$ thể hiện quan hệ bài báo y đã trích dẫn bởi bài báo x , hay bài báo x có trích dẫn bài bài báo y . Trọng số của $e \in E_3$ được tính dựa vào các phương pháp trong tập độ đo M .

2.3.5 Mạng cộng tác giữa các trường, viện AffNet

Mạng cộng tác AffNet là đồ thị có hướng và có trọng số gồm tập đỉnh có cấu trúc được mô tả chi tiết như sau:

$$AffNet = (Aff, E_4) \quad (2.8)$$

Trong đó:

- Aff : tập hợp các đỉnh của đồ thị là trường đại học, viện, cơ quan công tác.
- E_4 : tập các cặp đỉnh có hướng (x,y) thể hiện quan hệ cộng tác giữa 2 cơ quan x , y . Trước mắt chúng tôi chỉ xem xét quan hệ cộng tác viết bài (đồng tác giả). Đối với trọng số của cung quan hệ có hướng $(x,y) \in R$, có thể lượng hóa dựa vào số lần cộng tác viết bài của các cá nhân thuộc các cơ quan x và y . Cách tính trọng số của $e \in E_4$ được tính dựa vào các phương pháp tính toán trong tập M của mô hình ASN.

2.3.6 Các phương pháp tính toán trong mô hình ASN (Thành phần M trong mô hình ASN)

Thành phần M trong mô hình ASN là tập các phương pháp tính toán mức độ quan hệ, tương quan của các đối tượng trong các mô hình ASN phục vụ cho phân tích các mạng xã hội học thuật như *CoNet*, *CiNet_Author*, *CiNet_Paper*, *AffNet*. Các phương pháp tính toán đề xuất chủ yếu dựa trên việc xem xét yếu tố thời gian, cụ thể là xu hướng ảnh hưởng như thế nào đến sở thích và các mối quan hệ trong các cấu trúc mạng xã hội học thuật đã xác định.

2.3.6.1 Các phương pháp tương tự đỉnh truyền thống

Bao gồm các phương pháp tính tương tự đỉnh dựa trên thông tin lân cận cục bộ đã đề cập phần trên như là Cosine (Salton [101]), Jaccard (Jain and Dubes [56]), Adamic Adar (Adamic and Adar [2]). Các phương pháp này không quan tâm cung nối có trọng số hay không.

Độ đo Cosine. Độ đo cosine được định nghĩa như sau:

$$Sim_{Cosine}(u, v) = \frac{|n_u \cap n_v|}{\sqrt{|n_u| \cdot |n_v|}} \quad (2.9)$$

Trong đó:

- n_u : là tập những đỉnh lân cận của đỉnh u .
- $|n_u|$: số lượng những đỉnh lân cận của đỉnh u .
- $|n_u \cap n_v|$: số lượng lân cận chung của u và v .

Độ đo Jaccard. Hệ số tương tự Jaccard cho 2 đỉnh được tính như sau:

$$Sim_{Jaccard}(u, v) = \frac{|n_u \cap n_v|}{|n_u \cup n_v|} \quad (2.10)$$

Độ đo Adamic-Adar. Với Adamic-Adar thì hai đỉnh là tương tự hơn khi những lân cận chung của nó có số lân cận xung quanh ít hơn.

$$Sim_{Adamic-Adar}(u, v) = \sum_{c \in \{n_u \cup n_v\}} \frac{1}{\log |n_c|} \quad (2.11)$$

2.3.6.2 Đề xuất các phương pháp tương tự đỉnh mới

(1) Tương tự đỉnh dựa trên đường đi cực đại (MPRS) [CT.1]

Phương pháp tương tự đỉnh dựa trên “đường đi cực đại” giữa hai đỉnh u, v bất kỳ trong mạng đồng tác giả, gọi tắt là *MPRS* (Maximum Path based Relation Strength). Khi đó độ tương tự giữa u, v gọi là $Sim_{MPRS}(u, v)$ có thể được tính theo các bước sau:

Gọi $Direct_Sim_{MPRS}(u, v)$: trọng số cạnh nối giữa hai đỉnh u, v bất kỳ. Khi đó:

$$Direct_Sim_{MPRS}(u, v) = \begin{cases} \frac{f(u, v)}{\sum_{\forall c \in n_u} f(u, c)}, & \text{Nếu } u, v \text{ link trực tiếp} \\ 0, & \text{Ngược lại} \end{cases} \quad (2.12)$$

Trong đó,

- $f(u, v)$ là một hàm tính số lần đồng tác giả của u và v
- n_u : là tập các đỉnh lân cận của u .

Trong trường hợp u và v không có liên kết trực tiếp. Nếu trong mạng có một đường đi đơn p từ u đến v qua k đỉnh là $z_1, z_2, z_3, \dots, z_k$ (với z_1 là u , z_k là v), thì trọng số đường đi có thể tính như sau:

$$WeightOf_DirectPath_p(u, v) = \prod_{i=1}^k Direct_Sim_{MPRS}(z_i, z_{i+1}) \quad (2.13)$$

Trong trường hợp mạng đang xét có m đường đi đơn từ u đến v là p_1, p_2, \dots, p_m thì khi đó mức độ quan hệ của u và v , tức $Indirect_Sim_{MPRS}(u, v)$ có thể được tính như sau:

$$Indirect_Sim_{MPRS}(u, v) = \max_{i=1}^m (WeightOf_DirectPath_{p_i}(u, v)) \quad (2.14)$$

Trong những mạng kích thước lớn thì việc tính $Indirect_Sim_{MPRS}(u, v)$ có thể “quá tải”. Đồng thời, độ tương tự đỉnh của hai đỉnh bất kỳ có thể ít ý nghĩa và giá trị của $WeightOf_DirectPath_p(u, v)$ tiệm cận 0 nếu như đường đi đơn từ u đến v qua quá nhiều đỉnh trung gian. Vì thế trong quá trình thực nghiệm, chúng tôi đã sử dụng một giá trị ngưỡng là r như một thông số heuristic để kiểm soát quá trình xác định và tính trọng số các đường đi đơn từ u đến v trong những mạng kích thước lớn. Tức chúng tôi chỉ xem xét các đường đi đơn từ u đến v có “bán kính” (số đỉnh trên đường đi) nhỏ hơn hay bằng r . Như vậy $WeightOf_DirectPath_p(u, v)$ được tính như sau:

$$WeightOf_DirectPath_p(u, v) = \begin{cases} \prod_{i=1}^{k-1} Direct_Sim_{MPRS}(z_i, z_{i+1}), & \text{Nếu } k \leq r \\ 0, & \text{Ngược lại} \end{cases} \quad (2.15)$$

Tóm lại, tương tự của hai đỉnh u, v trong mạng theo phương pháp MPRS có thể tính như sau:

$$Sim_{MPRS}(u, v) = Indirect_Sim_{MPRS}(u, v) \quad (2.16)$$

(2) Tương tự đính dựa trên đường đi cực đại có xét xu hướng (MPRS+) [CT.1]

Tương tự $MPRS$, nhưng với $MPRS+$ chúng tôi xem xét yếu tố xu hướng cộng tác để lượng hóa mức độ quan hệ của hai đỉnh trong mạng đồng tác giả. Độ tương tự giữa 2 đỉnh bất kỳ theo $MPRS+$ được tính như sau: Gọi $Direct_Sim_{MPRS+}(u, v, t_0)$: trọng số cạnh nối giữa hai đỉnh u, v bất kỳ có xét đến yếu tố xu hướng. Khi đó:

$$Direct_Sim_{MPRS+}(u, v, t_0) = \begin{cases} \frac{f_{Trend}(u, v, t_0)}{\sum_{c \in n_u} f_{Trend}(u, c, t_0)}, & \text{Nếu } u, v \text{ link trực tiếp} \\ 0, & \text{Ngược lại} \end{cases} \quad (2.17)$$

Trong đó,

- n_u : là tập các đỉnh lân cận của u .
- t_0 : năm bắt đầu xem xét yếu tố xu hướng cộng tác của u và v .

$f_{Trend}(u, v, t_0)$ là một hàm thể hiện xu hướng cộng tác của u với v . Trên thực tế, những mối quan hệ cộng tác gần đây của 2 nghiên cứu viên sẽ quan trọng và ảnh hưởng nhiều đến việc hình thành các quan hệ cộng tác mới hơn những mối quan hệ cộng tác quá lâu trong quá khứ. Dựa trên đặc điểm đó và đặc tính của hàm $e^{-\delta(t)}$ ($\delta(t) \in [0, +\infty]$: là khoảng thời gian từ quá khứ đến hiện tại), luận án đề xuất dùng hàm xu hướng như sau:

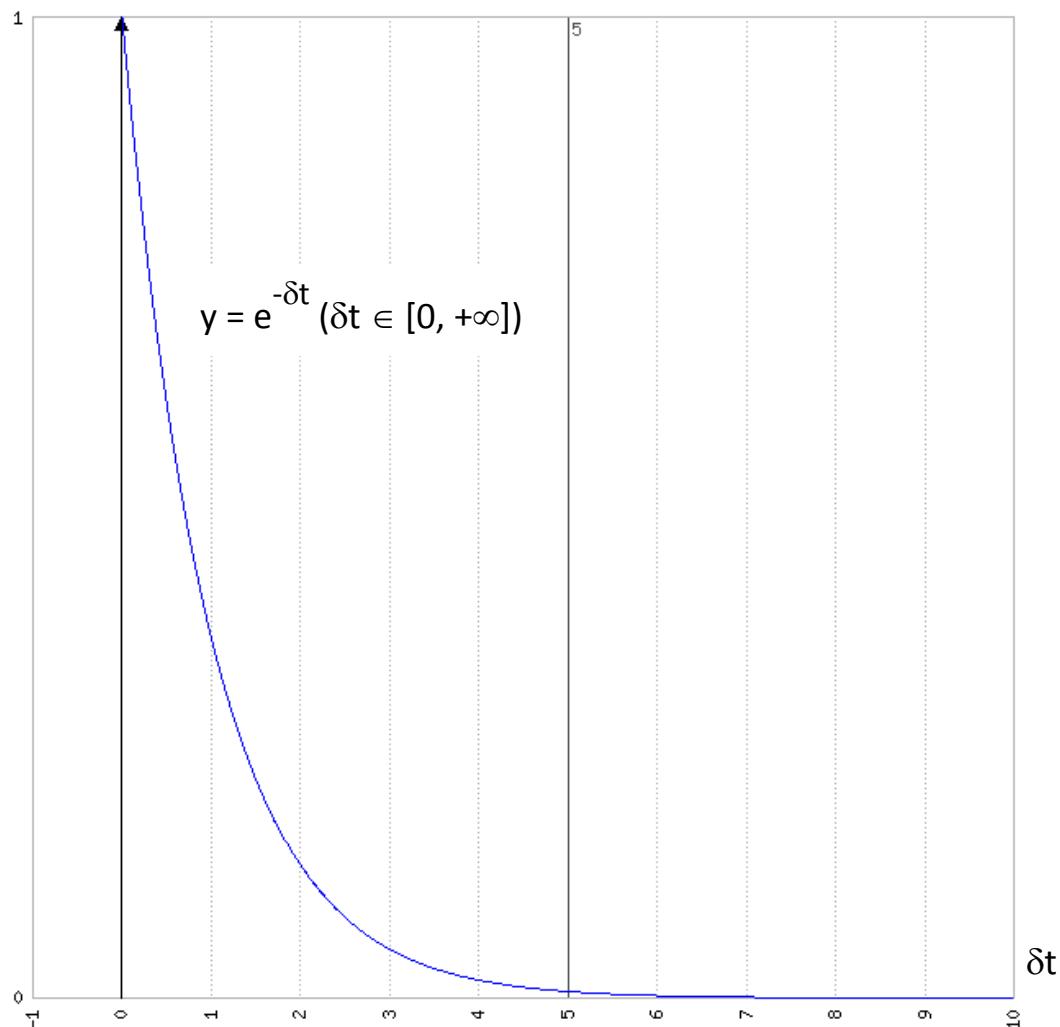
$$f_{Trend}(u, v, t_0) = \sum_{t_i=t_0}^{t_c} n(u, v, t_i) * \frac{1}{e^{\gamma*(t_c-t_i)}} \quad (2.18)$$

Trong đó,

- $n(u, v, t_i)$: là số bài báo mà u và v đồng tác giả tại thời điểm t_i .
- t_0 : năm bắt đầu xem xét yếu tố xu hướng cộng tác của u và v .
- t_c : năm hiện tại.
- γ : hệ số xu hướng. ($\gamma \in [0, 1]$. Trường hợp đơn giản $\gamma = 1$)

Lý do dùng hàm $e^{-\delta(t)}$ để thể hiện hệ số ảnh hưởng khi xét yếu tố xu hướng:

- i) $e^{-\delta(t)}$ là một hàm giảm đơn điệu và có tốc độ giảm nhanh dựa trên hệ số góc. Khi $\delta(t) \rightarrow +\infty$ thì $e^{-\delta(t)} \rightarrow 0$. Tức các mối quan hệ quá lâu trong quá khứ sẽ có hệ số ảnh hưởng không đáng kể (bằng 0) so với các mối quan hệ vừa xảy ra (bằng 1).
- ii) Khi $\delta(t) \rightarrow 0$ thì $e^{-\delta(t)} \rightarrow 1$. Tức các mối quan hệ gần đây sẽ có hệ số ảnh hưởng lớn nhất (bằng 1), lớn hơn nhiều so với các mối quan hệ quá lâu trong quá khứ (bằng 0). Hình 2.6 minh họa đặc điểm của hàm $e^{-\delta(t)}$.



Hình 2.6: Trực quan hàm $e^{-\delta(t)}$ ($\delta(t) \in [0, +\infty]$)

(3) Tương tự đindh dùng phương pháp RSS+ (cải tiến từ RSS) [CT.1]

Tác giả Chen và cộng sự đã đề xuất một phương pháp tương tự đindh dựa trên mức độ quan hệ giữa hai đindh bất kỳ trong mạng đồng tác giả, gọi là RSS (Relation Strength

Similarity) [28]. RSS là một độ đo bất đối xứng, áp dụng cho mạng có trọng số. Trong một số nghiên cứu liên quan thì Chen và cộng sự cũng đã dùng độ đo RSS này để khám phá các liên kết tiềm năng trong mạng đồng tác giả [27, 29]. Tuy nhiên, yếu tố xu hướng thì chưa được họ quan tâm trong RSS. Chúng tôi cũng đã đưa xu hướng cộng tác vào RSS và cải tiến RSS thành RSS+. Với RSS+ thì độ tương tự đỉnh có thể được tính như sau: Gọi $Direct_Sim_{RSS+}(u, v)$: trọng số cạnh nối giữa hai đỉnh u, v bất kỳ có xét đến yếu tố xu hướng. Khi đó:

$$Direct_Sim_{RSS+}(u, v, t_0) = \begin{cases} \frac{f_{Trend}(u, v, t_0)}{\sum_{c \in n_u} f_{Trend}(u, c, t_0)}, & \text{Nếu } u, v \text{ link trực tiếp} \\ 0, & \text{Ngược lại} \end{cases} \quad (2.19)$$

Trong đó,

- $f_{Trend}(u, v, t_0)$: là một hàm phụ thuộc yếu tố xu hướng cộng tác. Hàm $f_{Trend}(u, v, t_0)$ được tính tương tự như phương pháp $MPRS+$.
- n_u : là tập các đỉnh lân cận của u .

Trong trường hợp u và v không có liên kết trực tiếp. Nếu trong mạng có một đường đi đơn p từ u đến v qua k đỉnh là $z_1, z_2, z_3, \dots, z_k$ (với z_1 là u , z_k là v), thì trọng số đường đi có thể tính như sau:

$$WeightOf_DirectPath_p(u, v, t_0) = \prod_{i=1}^{k-1} Direct_Sim_{RSS+}(z_i, z_{i+1}, t_0) \quad (2.20)$$

Trong trường hợp mạng đang xét có m đường đi đơn từ u đến v là p_1, p_2, \dots, p_m thì khác với phương pháp $MPRS$ là chọn đường đi có trọng số cực đại (đường đi cộng tác có tích xác suất cộng tác qua các đỉnh trung gian là lớn nhất). Với phương pháp RSS , cũng như $RSS+$ mức độ quan hệ của u và v trong trường hợp này là tổng của các phân bố xác suất cộng tác qua các con đường cộng tác trung gian có thể, tức $Indirect_Sim_{RSS+}(u, v, t_0)$ có thể được tính như sau:

$$Indirect_Sim_{RSS+}(u, v, t_0) = \sum_{i=1}^m (WeightOf_DirectPath_{p_i}(u, v, t_0)) \quad (2.21)$$

Tương tự như $MPRS+$, với những mạng có kích thước lớn chúng tôi chỉ xem xét các đường đi đơn từ X đến Y có “bán kính” (số đỉnh trên đường đi) nhỏ hơn hay bằng

r. Như vậy $WeightOf_DirectPath_{RSS+}(u, v)$ được tính như sau:

$$WeightOf_DirectPath_p(u, v, t_0) = \begin{cases} \prod_{i=1}^{k-1} Direct_Sim_{RSS+}(z_i, z_{i+1}), & \text{Nếu } k \leq r \\ 0, & \text{Ngược lại} \end{cases} \quad (2.22)$$

Tóm lại, độ đo $RSS+$ của 2 đỉnh u, v bất kỳ trong mạng có thể tính như sau:

$$Sim_{RSS+}(u, v, t_0) = Direct_Sim_{RSS+}(u, v, t_0) + Indirect_Sim_{RSS+}(u, v, t_0) \quad (2.23)$$

2.3.6.3 Đề xuất phương pháp lượng hóa quan hệ lòng tin

(1) Lòng tin dựa trên quan hệ đồng tác giả và quan hệ trích dẫn [CT.2]

Giả sử rằng, lòng tin của một nghiên cứu viên r_i đối với nghiên cứu viên r_j , ký hiệu là $w_{trust}(r_i, r_j, t_0)$, phụ thuộc vào mức độ lòng tin của chính r_i kết hợp với lòng tin của những đồng tác giả của r_i đối với việc trích dẫn r_j bắt đầu xét từ thời điểm t_0 đến hiện tại. Chi tiết phương pháp có thể tính như sau:

$$w_{cite}(r_i, r_j, t_0) = \frac{\sum_{t_i=t_0}^{t_c} NumCitation(r_i, r_j, t_i)}{e^{\gamma*(t_c-t_i)} * TotalCitation(r_i, t_0)} \quad (2.24)$$

$$w_{trust}(r_i, r_j, t_0) = w_{cite}(r_i, r_j, t_0) + \frac{\sum_{r_u \in CoAuthor(r_i)} w_{coauthor}(r_i, r_u, t_0) * w_{cite}(r_u, r_j, t_0)}{|CoAuthor(r_i)|} \quad (2.25)$$

Trong đó,

- $NumCitation(r_i, r_j, t_i)$: số lần mà r_i đã trích dẫn r_j trong năm t_i .
- $TotalCitation(r_i, t_0)$: Tổng số trích dẫn của r_i tính từ thời điểm t_0 đến thời điểm hiện tại.
- t_c : năm hiện tại.
- t_0 : thời điểm bắt đầu xem xét yếu tố xu hướng.
- γ : hệ số xu hướng, $\gamma \in [0, 1]$. Trường hợp đơn giản $\gamma = 1$.

-
- $w_{coauthor}(r_i, r_u, t_0)$: mức độ quan hệ đồng tác giả giữa r_i với r_j tính ừ thời điểm t_0 đến hiện tại.
 - $|CoAuthor(r_i)|$: số đồng tác giả của r_i .

(2) Lòng tin dựa trên quan hệ trích dẫn tiềm ẩn [CT.2]

Trên thực tế, một nghiên cứu viên thường sẽ lần theo các bài báo trong mục tham khảo của các bài báo mà họ quan tâm để tìm kiếm các bài báo tiềm năng liên quan. Hành động đó thể hiện một quan hệ trích dẫn tiềm ẩn của một nghiên cứu viên đối với các nghiên cứu viên khác thông qua việc bắc cầu quan hệ trích dẫn. Nếu xét ở góc độ lòng tin, có thể nói, nghiên cứu viên có thể đặt lòng tin vào những nghiên cứu viên khác dựa trên việc bắc cầu quan hệ lòng tin.

Giả sử rằng, lòng tin của một nghiên cứu viên r_i đối với nghiên cứu viên r_j , ký hiệu là $w_{trust}(r_i, r_j, t_0)$. Chi tiết phương pháp lượng hóa lòng tin dựa trên quan hệ trích dẫn tiềm ẩn có thể tính như sau:

$$w_{trust}(r_i, r_j, t_0) = w_{cite}(r_i, r_j, t_0) + \frac{\sum_{r_u \in CitedAuthor(r_i)} w_{cite}(r_i, r_u, t_0) * w_{cite}(r_u, r_j, t_0)}{|CitedAuthor(r_i)|} \quad (2.26)$$

Trong đó,

- $|CitedAuthor(r_i)|$: số nghiên cứu viên mà r_i đã trích dẫn (đặt lòng tin).

(3) Lòng tin của nghiên cứu viên với bài báo [CT.2]

Mức độ lòng tin của một nghiên cứu viên r_i đối với bài báo p_j , ký hiệu $w_{trust}(r_i, p_j, t_0)$, có thể tính như sau:

$$w_{trust}(r_i, p_j, t_0) = MAX(w_{trust}(a_i, a_j, t_0)) \quad (2.27)$$

2.3.6.4 Đề xuất tập đặc trưng của nghiên cứu viên tiềm năng cho khuyến nghị cộng tác

(1) Uy tín của nghiên cứu viên [CT.3]

Chúng tôi giả sử rằng uy tín của các nghiên cứu viên là một trong những nhân tố chính giúp các nghiên cứu viên có thể hình thành nhiều hay ít các mối quan hệ cộng tác mới.

Vấn đề đặt ra là làm thế nào để lượng hóa được uy tín của các nghiên cứu viên? Chúng tôi đưa ra giả thuyết rằng: "Uy tín của các nghiên cứu viên càng cao khi họ được nhiều trích dẫn của các nghiên cứu viên uy tín khác". Đây là một giả thuyết có tính đê qui. Vì vậy, để lượng hóa uy tín của các nghiên cứu viên, luận án đã xây dựng mạng trích dẫn và đề xuất dùng thuật toán random-walk-with-restart (RWR) (Tong et al. [121]) trên mạng trích dẫn. Cụ thể uy tín của một nghiên cứu viên r_i là $I.Rate(r_i)$ được tính như sau:

$$I.Rate(r_i) = \frac{1-d}{N} + d * \left(\sum_{r_j \text{ LinkTo } r_i} \frac{I.Rate(r_j)}{|OutLink(r_j)|} + \sum_{r_j \text{ has no}} \frac{I.Rate(r_j)}{N} \right) \quad (2.28)$$

Trong đó,

- N : Tổng số các nghiên cứu viên trong mạng trích dẫn (citation network)
- $|OutLink(r)|$: số lượng các out-link của r
- d : nhân tố thâm thấu dùng trong RWR (damping factor). Thông thường trong thực nghiệm thì d được thiết lập 0.85 [121].

(2) Độ năng động của nghiên cứu viên [CT.3]

Bên cạnh các yếu tố về lĩnh vực quan tâm, quan hệ của cơ quan, uy tín của các nghiên cứu viên thì mức độ năng động của nghiên cứu viên là một yếu tố khá quan trọng giúp hình thành các mối quan hệ cộng tác mới. Thông thường một nghiên cứu viên càng năng động thì càng có nhiều cơ hội để hình thành các cộng tác mới. Câu hỏi đặt ra là làm thế nào để lượng hóa được mức độ năng động của một nghiên cứu viên. Luận án đề xuất phương pháp lượng hóa dựa trên tần xuất công bố các bài báo của các nghiên cứu viên. Luận án đưa ra giả thuyết: 'Một nghiên cứu viên năng động nếu ngày càng cho ra nhiều bài báo'. Khi đó, mức độ năng động của một nghiên cứu viên r xét từ một mốc thời gian t_0 , ký hiệu là $ActiveScore(r, t_0)$, có thể lượng hóa thông qua như sau:

$$f_{active}(r, t_0) = \sum_{i=0}^c N(r, t_i) * \frac{1}{e^{(t_c - t_i)}} \quad (2.29)$$

Trong đó,

- t_c : năm hiện tại.
- t_0 : năm trong quá khứ mà bắt đầu xem xét quá trình, cũng như xu hướng nghiên cứu của các nghiên cứu viên. Trong thực nghiệm chúng tôi chọn $t_0 = t_c - 5$. Vì theo kinh nghiệm của chúng tôi, 5 năm là một khoảng thời gian đủ lớn để đánh giá xu hướng và quá trình hoạt động nghiên cứu gần đây của một nghiên cứu viên.
- $N(r, t_i)$: số lượng bài báo của nghiên cứu viên r tại thời điểm t_i .

Chuẩn hóa: để chuẩn hóa giá trị mức độ năng động của các nghiên cứu viên r là $ActiveScore(r)$ về khoảng $[0,1]$, chúng ta có thể tính như sau:

$$ActiveScore(r, t_0) = \frac{f_{active}(r, t_0) - \min_{r_i \in R}(f_{active}(r_i, t_0))}{\max_{r_i \in R}(f_{active}(r_i, t_0)) - \min_{r_i \in R}(f_{active}(r_i, t_0))} \quad (2.30)$$

(3) Mức độ quan hệ giữa các cơ quan [CT.3]

Chúng tôi đưa ra giả thuyết: "Những quan hệ cộng tác mới, tiềm năng thường sẽ hình thành từ những cơ quan có quan hệ cộng tác tốt". Vì thế, hai cơ quan có quan hệ cộng tác tốt sẽ dễ dàng hình thành các quan hệ đồng tác giả nhiều hơn so với hai cơ quan ít hoặc không có quan hệ. Vấn đề đặt ra là làm thế nào để lượng hóa được mức độ quan hệ của các cơ quan. Luận án đề xuất cách lượng hóa quan hệ cộng tác giữa các cơ quan dựa trên số lượng các quan hệ đồng tác giả của các cá nhân thuộc những cơ quan này. Khi đó mức độ quan hệ cơ quan của hai người nghiên cứu r, r' bất kỳ là $(OrgRS(r, r'))$ được tính như sau:

Trong mạng, p là một đường đi công tác đơn từ o đến o' thông qua k cơ quan khác $o \equiv o_1, o_2, \dots, o_k \equiv o'$.

$$w(o_i, o_{i+1}) = P(o_i, o_{i+1}) = \frac{Coll_Num(o_i, o_{i+1})}{Total_Coll_Num(o_i)} \quad (2.31)$$

Trong đó,

- $Coll_Num(o_i, o_{i+1})$: Số lượng các quan hệ đồng tác giả của các cá nhân nghiên cứu viên thuộc o_i và o_{i+1} .
- $Total_Coll_Num(o_i)$: Tổng số quan hệ đồng tác giả của các cá nhân nghiên cứu

viên trong o_i với các cá nhân thuộc các cơ quan khác.

Chúng ta có thể dùng các độ đo cơ bản từ lý thuyết đồ thị và lý thuyết xác suất để tính mức độ quan hệ giữa các cơ quan o, o' ($OrgRS(o, o')$) như sau:

$$OrgRS(o, o') = \sum_{i=1}^m (Path_Weight_{p_i}(o, o')) \quad (2.32)$$

Trong đó,

- p_1, \dots, p_m : Tất cả các đường đi có hướng không chu trình giữa o, o' .

- $Path_Weight_p(o, o') = \prod_{i=1}^k w(o_i, o_{i+1})$.

Nói chung, mức độ quan hệ cơ quan của hai cá nhân nghiên cứu viên r, r' được tính dựa trên mức độ quan hệ của các cơ quan, nơi mà r, r' đang làm việc.

$$OrgRS(r, r') = OrgRS(o_r, o_{r'}) \quad (2.33)$$

2.4 Kết chương

Chương này đã trình bày giải pháp cho rút trích, xây dựng kho dữ liệu bài báo khoa học bằng cách tích hợp từ nhiều nguồn không đồng nhất. Với kho dữ liệu bài báo khoa học thu thập, các mạng xã hội học thuật sẽ được rút trích, mô hình, luận án gọi mô hình đề xuất cho việc mô hình hóa các mạng xã hội học thuật là ASN. Tiếp cận của luận án để phát triển các phương pháp khuyến nghị trong lĩnh vực học thuật là tập trung phân tích các mạng xã hội học thuật. Hiện nay, luận án tập trung vào mạng đồng tác giả CoNet, mạng quan hệ cộng tác giữa các cơ quan AffNet, mạng trích dẫn giữa những người nghiên cứu *CiNet_Author*. Một phần của chương này được trình bày trong các công trình [3](#), [1](#), [4](#).

Chương 3

KHAI THÁC MẠNG XÃ HỘI HỌC THUẬT ĐỂ PHÁT TRIỂN CÁC PHƯƠNG PHÁP KHUYẾN NGHỊ CỘNG TÁC

3.1 Giới thiệu

Có nhiều định nghĩa khác nhau cho cộng tác, nhưng nhìn chung cộng tác là hành động hay quá trình hai hay nhiều cá nhân, tổ chức làm việc cùng nhau để thực hiện một mục đích chung¹. Trong nghiên cứu khoa học, có thể quan niệm cộng tác nghiên cứu là quá trình làm việc cùng nhau của những nghiên cứu viên để đạt được một mục đích chung trong việc tìm ra các tri thức khoa học mới (Katz et al. [61]). Cộng tác nghiên cứu giúp các nghiên cứu viên có cơ hội để trao đổi kiến thức, kinh nghiệm. Những nghiên cứu viên càng có nhiều quan hệ công tác tốt thì càng có khả năng tạo ra nhiều tri thức mới trong khoa học (Katz et al. [61], Lotka [74]).

Có thể nói đối tác hay người cộng tác là một trong những yếu tố then chốt quyết định chất lượng, kết quả đạt được của quá trình cộng tác. Câu hỏi đặt ra là làm thế nào có thể tìm được những người cộng tác phù hợp với một mục đích công việc cụ thể? Với những nghiên cứu viên trẻ chưa có kinh nghiệm thì thật khó có thể biết được ai là người cộng tác phù hợp. Còn những nghiên cứu viên kinh nghiệm thì phải đương đầu với tình trạng quá tải thông tin. Sự gia tăng một cách nhanh chóng kích thước của các kho dữ liệu học thuật, đã gây không ít khó khăn trong việc xác định, tìm kiếm những

¹<http://oxforddictionaries.com/definition/english/collaboration>, truy cập lần cuối 07/02/2014

chuyên gia, người cộng tác phù hợp. Do đó, nghiên cứu phát triển các phương pháp, hệ thống khuyến nghị cộng tác sẽ là giải pháp cho vấn đề này. Bên cạnh đó, khuyến nghị cộng tác cũng giúp nghiên cứu viên tăng kết nối trong cộng đồng học thuật.

Mục đích của chương này là trình bày, phát biểu bài toán khuyến nghị cộng tác trong nghiên cứu khoa học và phát triển các phương pháp mới dựa trên tiếp cận khai thác các mối quan hệ xã hội học thuật từ mô hình ASN (đã đề cập trong chương trước) để giải quyết bài toán này cho từng nhóm nghiên cứu viên khác nhau.

3.2 Bài toán khuyến nghị cộng tác

Định nghĩa 3.1: *Nghiên cứu viên có đồng tác giả (un-isolated researcher)*

Nghiên cứu viên có đồng tác giả là các nghiên cứu viên mà tồn tại ít nhất một bài báo đã công bố trong quá khứ có đồng tác giả với một nghiên cứu viên khác.

Định nghĩa 3.2: *Nghiên cứu viên chưa có đồng tác giả (isolated researcher)*

Nghiên cứu viên chưa có đồng tác giả là các nghiên cứu viên mà trong quá khứ, tính tới thời điểm hiện tại chưa có bài báo công bố nào có đồng tác giả với một nghiên cứu viên khác.

Khuyến nghị cộng tác trong nghiên cứu khoa học là bài toán tự động liệt kê những người, nhóm cộng tác tiềm năng ứng với đầu vào là một hay nhóm những nghiên cứu viên. Khuyến nghị cộng tác đóng vai trò quan trọng và gần đây đã bắt đầu thu hút nhiều quan tâm. (Chen et al. [28]) đã phát triển hệ thống tìm kiếm chuyên gia cộng tác CollabSeer dựa trên cấu trúc mạng đồng tác giả. (Tang et al. [119]) đã nghiên cứu đề xuất các phương pháp so khớp chuyên gia dựa trên nhiều ràng buộc khác nhau và ứng dụng vào bài toán khuyến nghị chuyên gia phản biện bài báo khoa học, khuyến nghị giảng viên cho một môn học. Trong một nghiên cứu khuyến nghị cộng tác khác, (Tang et al. [117]) đề xuất các phương pháp khuyến nghị cộng tác cho các chuyên gia trong các nghiên cứu liên ngành. Bên cạnh đó các nghiên cứu liên quan đến bài toán tìm kiếm, so khớp chuyên gia cũng cung cấp các phương pháp nền tảng cho khuyến nghị cộng tác nghiên cứu. (Hofmann et al. [52], Balog and de Rijke [12], và Gollapalli et al. [48]) đã nghiên cứu, trình bày, thực nghiệm, đánh giá các phương pháp phổ biến mà dùng để biểu diễn thông tin và tính toán tương tự, so khớp chuyên gia.

Trong phạm vi luận án này, chúng tôi xem xét giải quyết bài toán khuyến nghị cộng tác với đầu vào là một nghiên cứu viên, hệ thống có nhiệm vụ sinh ra danh sách xếp hạng những người cộng tác tiềm năng. Bài toán có thể được định nghĩa một cách hình thức như sau:

- **Đầu vào:**

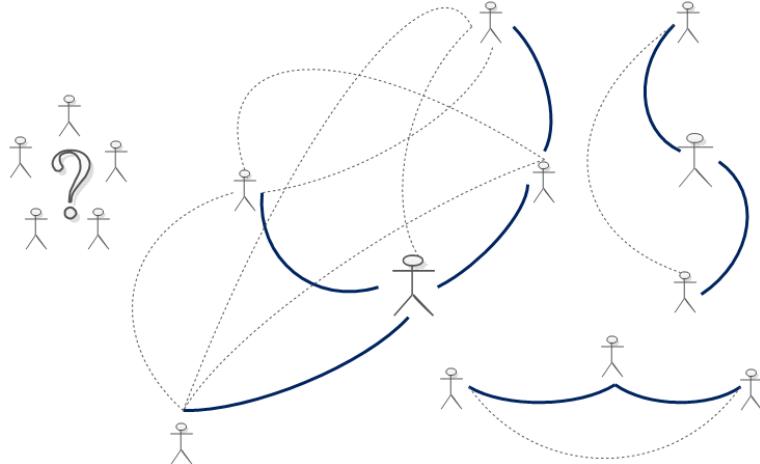
$R = \{r\}$: tập tất cả các nghiên cứu viên.

$P = \{p\}$: tập tất cả các bài báo trong kho dữ liệu.

$O = \{o\}$: danh sách các cơ quan nơi các nghiên cứu viên đang làm việc.

- **Đầu ra:**

Xác định hàm $f(r_i, r_j)$ để ước lượng tiềm năng quan hệ cộng tác của $r_i \in R$ với $r_j \in R, r_i \neq r_j$. $\forall r \in R$, dựa trên hàm f chọn $TopN$ các NCV tiềm năng nhất, $R_{TopN} \subset R$, $R_{TopN} = < r_1, r_2, \dots, r_{TopN} >$, (với $TopN << |R|, r_i \in R_{TopN}, r_i \neq r$) để khuyến nghị cho r.



Hình 3.1: Những phương pháp dựa trên phân tích mạng đồng tác giả có thể khuyến nghị cộng tác cho các nghiên cứu viên có đồng tác giả (nét chấm đứt trong hình vẽ), nhưng sẽ không thực hiện được đối với các nghiên cứu viên chưa có đồng tác giả (quanh dấu chấm hỏi).

Với các nghiên cứu viên khác nhau (trẻ, có kinh nghiệm, nghiên cứu viên mới) sẽ cần có những phương pháp khác nhau để thực hiện việc khuyến nghị. Chẳng hạn, đối với các nghiên cứu viên có quan hệ trong mạng đồng tác giả thì tiếp cận nổi trội đang

được các nghiên cứu phổ biến hiện nay quan tâm là dựa trên việc phân tích các mối quan hệ trong mạng đồng tác giả để tìm ra những người cộng tác tiềm năng (Konstas et al. [64], Davoodi et al. [37], Chen et al. [28, 27, 29], Lopes et al. [72], Brandão et al. [23]). Nhưng đối với các nghiên cứu viên mới, không có các mối quan hệ trong mạng đồng tác giả, thì theo hiểu biết của tôi, hiện nay vẫn chưa có cách giải quyết (Hình 3.1). Vì vậy luận án đã phân thành hai nhóm nghiên cứu viên chính và phát triển các phương pháp khuyến nghị cho hai nhóm nghiên cứu viên này: (1) nhóm các nghiên cứu viên đã có đồng tác giả; (2) Nhóm các nghiên cứu viên chưa có đồng tác giả. Phần còn lại của chương sẽ trình bày về các phương pháp phổ biến và phương pháp đề xuất dựa trên khai thác các mối quan hệ xã hội học thuật từ mô hình ASN để khuyến nghị cộng tác cho hai nhóm nghiên cứu viên này.

3.3 Trường hợp các nghiên cứu viên có đồng tác giả (un-isolated)

Dối với các nghiên cứu viên có quan hệ trong mạng đồng tác giả thì tiếp cận nổi trội đang được các nghiên cứu phổ biến hiện nay quan tâm là dựa trên việc phân tích các mối quan hệ đồng tác giả để tiên đoán những người cộng tác tiềm năng (Konstas et al. [64], Davoodi et al. [37], Chen et al. [28, 27, 29], Lopes et al. [72], Brandão et al. [23]). Trong phân tích mạng đồng tác giả, việc lượng hóa mức độ quan hệ giữa các đỉnh là bước khá quan trọng giúp tiên đoán, khám phá những liên kết cộng tác tiềm năng. Các phương pháp tính toán tương tự đỉnh truyền thống có thể chia thành hai nhóm: các phương pháp dựa trên cấu trúc cục bộ và các phương pháp dựa trên cấu trúc toàn cục của mạng.

3.3.1 Tương tự đỉnh dựa trên cấu trúc cục bộ

Các phương pháp dựa trên cấu trúc cục bộ dùng thông tin lân cận cục bộ để tính độ tương tự của hai đỉnh bất kỳ trong mạng. Ý tưởng chung của các độ đo cục bộ là “Hai đỉnh càng tương tự nhau nếu chúng có chung nhiều lân cận”. Tức chỉ có những đỉnh lân cận trực tiếp của hai đỉnh được xét đến khi tính toán sự tương tự, trong khi các đỉnh khác không được quan tâm xem xét. Một số phương pháp tương tự đỉnh cục bộ phổ biến có thể kể đến như hệ số Jaccard (Chen et al. [28]), độ đo Cosine (Salton [101]),

độ đo Adamic-Adar (Adamic and Adar [2]).

Các phương pháp tương tự đỉnh này cũng đã được trình bày trong thành phần M của mô hình ASN trong chương 2. Đối với bài toán khuyến nghị công tác (trường hợp cho các nghiên cứu viên đã có quan hệ đồng tác giả), thì những phương pháp này được dùng như phương pháp cơ sở (baseline) để so sánh với các phương pháp phổ biến hiện nay cũng như các phương pháp đề xuất khác trong mô hình ASN.

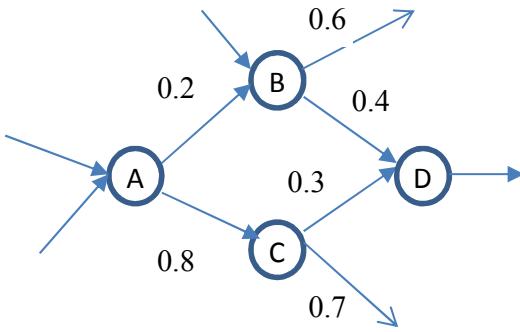
3.3.2 Tương tự đỉnh dựa trên cấu trúc toàn cục

Các phương pháp này dựa trên cấu trúc toàn cục của cả mạng thay vì chỉ xét cấu trúc cục bộ như các phương pháp kể trên. Một số phương pháp toàn cục phổ biến thường dùng như SimRank (Jeh and Widom [59]), P-Rank (Zhao et al. [130]). Các phương pháp này dựa trên ý tưởng “Hai đỉnh càng giống nhau nếu như những lân cận trực tiếp trong mạng là tương tự nhau”. Vì thế việc tính độ tương tự đỉnh dùng SimRank, P-Rank là một quá trình đẽ qui.

Theo hiểu biết của tác giả, thì một trong những phương pháp phổ biến nhất hiện nay cho bài toán tiên đoán liên kết đồng tác giả là RSS (Relation Strength Similarity) do Chen và cộng sự đề xuất [28]. RSS là một phương pháp tương tự đỉnh dựa trên cấu trúc toàn cục của mạng đồng tác giả, là một độ đo bất đối xứng, áp dụng cho mạng có trọng số. Trong một số nghiên cứu liên quan thì Chen và cộng sự cũng đã dùng độ đo RSS này để khám phá các liên kết tiềm năng trong mạng đồng tác giả [27, 29]. RSS cũng được đánh giá, so sánh với các tương tự đỉnh cục bộ và các phương pháp đề xuất khác trong mô hình ASN cho bài toán khuyến nghị công tác (trường hợp nghiên cứu viên đã có quan hệ đồng tác giả).

3.3.3 Nhận định

Chúng ta có thể áp dụng các phương pháp cục bộ lẫn toàn cục kể trên cho việc tính toán tương tự đỉnh trong mạng đồng tác giả để tìm ra các ứng viên tiềm năng cho khuyến nghị công tác. Các phương pháp tương tự đỉnh đã chứng tỏ được những thành công của nó trong các bài toán tiên đoán liên kết trong mạng xã hội nói chung, cũng như tiên đoán liên kết đồng tác giả để khuyến nghị. Tuy nhiên, hầu hết các nghiên cứu này đều chưa quan tâm đến yếu tố xu hướng cộng tác. Trong khi yếu tố xu hướng có



Hình 3.2: Minh họa cách tính mức độ quan hệ

ảnh hưởng lớn đến việc hình thành các mối quan hệ cộng tác mới. Do đó, luận án đã đề xuất dùng yếu tố xu hướng để phát triển các phương pháp tính toán mới (thành phần M của mô hình ASN) và khai thác chúng cho việc tiên đoán đồng tác giả, khuyến nghị cộng tác (trường hợp nghiên cứu viên đã có quan hệ đồng tác giả).

3.3.4 Các phương pháp đề xuất

Phần này trình bày các phương pháp đề xuất để tính toán mức độ quan hệ giữa các đỉnh trong mạng đồng tác giả. RSS+ là phương pháp cải tiến từ RSS (Relation Strength Similarity) (Chen et al. [28]). RSS tính mức độ quan hệ của hai đỉnh u, v bất kỳ trong mạng bằng tổng trọng số các đường đi có thể từ u đến v . MPRS dựa trên đường đi có trọng số cực đại [CT.4, CT.1]. Hai phương pháp RSS+ và MPRS+ dùng yếu tố xu hướng cộng tác để lượng hóa mức độ quan hệ giữa 2 đỉnh trong mạng đồng tác giả [CT.4, CT.1].

Ví dụ: Trong hình vẽ 3.2 thì A có viết tổng cộng 10 bài báo. Trong đó A viết chung với B là 2 bài, và với C là 8 bài. Khi đó trọng số các cung AB và AC trong mạng đồng tác giả có thể gán lần lượt là: 0.2, 0.8. Hình 3.2 minh họa việc tính tương tự dựa trên MPVS và RSS cho hai đỉnh A và D.

- $Sim_{MPRS}(A, D) = MAX((0.2 * 0.4), (0.8 * 0.3)) = 0.24$
- $Sim_{RSS}(A, D) = SUM((0.2 * 0.4), (0.8 * 0.3)) = 0.32$
- Đối với các phương pháp RSS+ và MPVS+ thì trọng số của một cung nối trong mạng đồng tác giả không chỉ phụ thuộc vào số bài viết chung, mà còn phụ thuộc vào thời gian viết chung bài đó là lúc nào.

3.3.4.1 Tương tự đỉnh dựa trên đường dẫn có trọng số cực đại (MPRS)

Phương pháp dựa trên “đường dẫn cực đại” giữa hai đỉnh u, v bất kỳ trong mạng đồng tác giả, gọi tắt là *MPRS* [CT.1, CT.4]. Phương pháp MPRS thuộc thành phần tính toán M của mô hình mạng xã hội học thuật ASN đã được trình bày chi tiết trong chương 2. Phần này sẽ trình bày tóm tắt việc áp dụng phương pháp MPRS cho bài toán khuyến nghị cộng tác (trường hợp nghiên cứu viên đã có quan hệ đồng tác giả).

Phương pháp: MPRS cho khuyến nghị cộng tác

Đầu vào:

- $R = \{r\}$: tập tất cả các nghiên cứu viên có đồng tác giả.
- $CoNet = (R, E_1)$: mạng đồng tác giả giữa các nghiên cứu viên trong R.

Đầu ra:

- Xác định hàm $f(r_i, r_j)$ để ước lượng mức độ tiềm năng cho quan hệ cộng tác của $r_j \in R$ với $r_i \in R, r_i \neq r_j$.
- $\forall r_i \in R$, chọn TopN các nghiên cứu viên $r_j \in R, r_j \neq r_i$ để khuyến nghị cộng tác cho r_i dựa trên giá trị hàm $f(r_i, r_j)$.

1: **Bước 1:** Tính trọng số cho cạnh nối có hướng giữa 2 đỉnh u, v bất kỳ trong mạng đồng tác giả *CoNet* theo công thức (2.12).

2: **Bước 2:** Tìm tất cả các đường đi đơn $p \in P_{u,v}$ có độ dài (tổng số cạnh nối) nhỏ hơn 4 giữa tất cả các cặp đỉnh (u, v) trong *CoNet*.

- $\forall u \in R$:

Duyệt theo chiều sâu từ đỉnh u, qua k đỉnh z_1, z_2, \dots, z_k

(z_1 là u , z_k là $v, \forall v \in R, v \neq u, k < 5$)

Thêm $p = z_1, z_2, \dots, z_k$ vào tập $P_{u,v}$.

3: **Bước 3:** Áp dụng công thức (2.13) để tính trọng số cho tất cả các đường đi đơn $p \in P_{u,v}$, ký hiệu là $WeightOf_DirectPath_p(u, v)$.

- $\forall u \in R, \forall v \in R, v \neq u$:

$\forall p \in P_{u,v}$, tính $WeightOf_DirectPath_p(u, v)$.

4: **Bước 4:** Áp dụng công thức (2.14) để tính trọng số quan hệ giữa 2 đỉnh u, v bất kỳ trong *CoNet*, ký hiệu $Indirect_Sim_{MPRS}(u, v)$.

5: **Bước 5:** Thực hiện khuyến nghị

- $\forall r_i \in R, \forall r_j \in R, r_j \neq r_i$

Chọn TopN các r_j , có $f(r_i, r_j) = Indirect_Sim(r_i, r_j)$ cao nhất để khuyến nghị cho r_i .

Phân tích độ phức tạp tính toán của phương pháp MPRS:

- Xét bước 1: Ta cần tính trọng số cho tất cả các cạnh trong *CoNet*. Với mỗi cạnh, cần tính $f(u, v)$. Vậy số phép tính của $f(u, v)$ là cố định và độ phức tạp là $\mathcal{O}(|E_1|)$.
- Xét bước 2: Ta cần tìm tất cả các đường đi giữa mọi cặp đỉnh trong *CoNet*. Ta cần duyệt theo chiều sâu từ mỗi đỉnh tối đa 3 cấp. Giả sử các NCV có số đồng tác giả đồng đều, bậc trung bình trong *CoNet* = $d = 2 * |E_1|/|R|$. Khi đó độ phức tạp để duyệt từ mỗi đỉnh là $\mathcal{O}(d^3)$. Ở đây, chúng ta cần duyệt qua mỗi đỉnh. Vậy độ phức tạp là $\mathcal{O}(|R|d^3)$.
- Xét bước 3: Ta cần tính cho tất cả các đường đi giữa tất cả các cặp đỉnh trong *CoNet*. Mỗi đường đi có chiều dài tối đa = 3. Số đường đi giữa các cặp đỉnh tối đa là d^3 . Vậy độ phức tạp là $\mathcal{O}(|R|^2d^3)$.
- Xét bước 4: Tương tự bước 3, ta cần tính trọng số cho tất cả các cặp đỉnh trong *CoNet* dựa trên tất cả các đường đi giữa chúng. Vậy độ phức tạp là $\mathcal{O}(|R|^2d^3)$.
- Xét bước 5: Ta cần xét tất cả các cặp đỉnh. Vậy độ phức tạp là $\mathcal{O}(|R|^2)$.
- **Tóm lại**, độ phức tạp của phương pháp MPRS là $\mathcal{O}(|R|^2d^3)$ (với d là bậc trung bình của một đỉnh trong mạng đồng tác giả).

3.3.4.2 Tương tự đỉnh dựa trên đường dẫn cực đại có xét xu hướng (MPRS+)

Tương tự *MPRS*, nhưng với *MPRS+* luận án xem xét yếu tố xu hướng cộng tác để lượng hóa mức độ quan hệ của hai đỉnh trong mạng đồng tác giả [CT.1, CT.4]. Chi tiết về cách tính độ tương tự giữa 2 đỉnh bất kỳ theo *MPRS+* đã được trình bày chi tiết trong mục 2.3.6.2 trong chương 2. Phần này sẽ trình bày tóm tắt việc áp dụng phương pháp *MPRS+* cho bài toán khuyến nghị cộng tác (trường hợp nghiên cứu viên đã có quan hệ đồng tác giả).

Phương pháp: MPRS+ cho khuyến nghị cộng tác

Đầu vào:

- $R = \{r\}$: tập tất cả các nghiên cứu viên có đồng tác giả (un-isolated)
- $CoNet = (R, E_1)$: mạng đồng tác giả giữa các nghiên cứu viên trong R.

Đầu ra:

- Xác định hàm $f(r_i, r_j)$ để ước lượng mức độ tiềm năng cho quan hệ cộng tác của $r_j \in R$ với $r_i \in R, r_i \neq r_j$.
- $\forall r_i \in R$, chọn TopN các nghiên cứu viên tiềm năng nhất để khuyến nghị cộng tác cho r_i dựa trên giá trị hàm $f(r_i, r_j), r_j \in R, r_j \neq r_i$.

-
- 1: **Bước 1:** Tính trọng số theo xu hướng cho cạnh nối có hướng giữa 2 đỉnh u, v bất kỳ trong mạng đồng tác giả $CoNet$, ký hiệu $Direct_Sim_{MPRS+}(u, v, t_0)$, theo công thức (2.17).
 - 2: **Bước 2:** Tìm tất cả các đường đi đơn $p \in P_{u,v}$ có độ dài (tổng số cạnh nối) nhỏ hơn 4 giữa tất cả các cặp đỉnh (u, v) trong $CoNet$.

- $\forall u \in R$:

Duyệt theo chiều sâu từ đỉnh u , qua k đỉnh z_1, z_2, \dots, z_k

(z_1 là u , z_k là $v, \forall v \in R, v \neq u, k < 5$)

Thêm $p = z_1, z_2, \dots, z_k$ vào tập $P_{u,v}$.

- 3: **Bước 3:** Tính trọng số theo xu hướng cho tất cả các đường đi đơn $p \in P_{u,v}$, $WeightOf_DirectPath_p(u, v, t_0)$

- $\forall u \in R, \forall v \in R, v \neq u$:

- $\forall p \in P_{u,v}$, tính:

$$WeightOf_DirectPath_p(u, v, t_0) = \prod_{i=1}^{k-1} Direct_Sim_{MPRS+}(z_i, z_{i+1}, t_0) \quad (3.1)$$

- 4: **Bước 4:** Tính trọng số quan hệ giữa 2 đỉnh u, v bất kỳ trong $CoNet$.

$$Indirect_Sim(u, v, t_0) = Indirect_Sim_{MPRS+} = \max_{p_i \in P_{u,v}} (Weight_Of_DirectPath_{p_i}(u, v, t_0)) \quad (3.2)$$

- 5: **Bước 5:** Thực hiện khuyến nghị

- $\forall r_i \in R, \forall r_j \in R, r_j \neq r_i$.

Chọn TopN các r_j , có $f(r_i, r_j) = Indirect_Sim(r_i, r_j)$ cao nhất để khuyến nghị cho r_i .

Phân tích độ phức tạp tính toán của phương pháp MPRS+:

- Xét bước 1: Ta cần tính trọng số cho tất cả các cạnh trong CoNet. Với mỗi cạnh, cần tính $f_{Trend}(u, v, t_0)$ (trong công thức 2.17) trong một khoảng thời gian cho trước từ t_0 đến hiện tại. Vậy số phép tính của $f_{Trend}(u, v, t_0)$ là cố định và độ phức tạp tính toán là $\mathcal{O}(|E_1|)$.
- Xét bước 2, 3, 4, 5: tương tự phương pháp MPRS.
- **Tóm lại**, độ phức tạp của phương pháp MPRS+ là $\mathcal{O}(|R|^2 d^3)$ (với d là bậc trung bình của một đỉnh trong mạng đồng tác giả).

3.3.4.3 Tương tự đính dùng phương pháp RSS+ (cải tiến từ RSS)

RSS là một độ đo bất đối xứng, áp dụng cho mạng có trọng số do Chen và cộng sự đề xuất để tính toán độ tương tự của các nghiên cứu viên trong mạng đồng tác giả dựa trên mức độ quan hệ [28]. Chen và cộng sự đã dùng độ đo này để khám phá các liên kết tiềm năng trong mạng đồng tác giả. Kết quả thực nghiệm của họ cho thấy RSS đã khá thành công trong việc khám phá các mối quan hệ tiềm ẩn trong mạng đồng tác giả [29, 27]. Tuy nhiên, yếu tố xu hướng quan hệ có ảnh hưởng như thế nào đến các quan hệ tiềm ẩn thì chưa được họ quan tâm xem xét. Luận án đã đưa xu hướng cộng tác vào RSS và cải tiến RSS thành RSS+ [CT.1, CT.4]. Chương 2 đã trình bày chi tiết về phương pháp tính toán RSS+ trong thành phần M của mô hình ASN. Phần này sẽ tóm tắt việc áp dụng phương pháp RSS+ cho bài toán khuyến nghị cộng tác.

Phương pháp: RSS+ cho khuyến nghị cộng tác

Đầu vào:

- $R = \{r\}$: tập tất cả các nghiên cứu viên có đồng tác giả (un-isolated)
- $CoNet = (R, E_1)$: mạng đồng tác giả giữa các nghiên cứu viên trong R.

Đầu ra:

- Xác định hàm $f(r_i, r_j)$ để ước lượng mức độ tiềm năng cho quan hệ cộng tác của $r_j \in R$ với $r_i \in R, r_i \neq r_j$.
- $\forall r_i \in R$, chọn TopN các nghiên cứu viên $r_j \in R, r_j \neq r_i$ để khuyến nghị cộng tác cho r_i dựa trên giá trị hàm $f(r_i, r_j)$.

-
- 1: **Bước 1:** Tính trọng số theo xu hướng cho cạnh nối có hướng giữa 2 đỉnh u, v bất kỳ trong mạng đồng tác giả $CoNet$, ký hiệu $Direct_Sim_{RSS+}(u, v, t_0)$, theo công thức (2.17).
 - 2: **Bước 2:** Tìm tất cả các đường đi đơn $p \in P_{u,v}$ có độ dài (tổng số cạnh nối) nhỏ hơn 4 giữa tất cả các cặp đỉnh (u, v) trong $CoNet$.

- $\forall u \in R$:

Duyệt theo chiều sâu từ đỉnh u , qua k đỉnh z_1, z_2, \dots, z_k

(z_1 là u , z_k là $v, \forall v \in R, v \neq u, k < 5$)

Thêm $p = z_1, z_2, \dots, z_k$ vào tập $P_{u,v}$.

- 3: **Bước 3:** Tính trọng số theo xu hướng cho tất cả các đường đi đơn $p \in P_{u,v}$, $WeightOf_DirectPath_p(u, v, t_0)$ (Thực hiện tương tự phương pháp MPRS+)
- 4: **Bước 4:** Tính mức độ quan hệ giữa 2 đỉnh u, v bất kỳ trong mạng CoNet.

$$Indirect_Sim(u, v, t_0) = Indirect_Sim_{RSS+} = \sum_{p_i \in P_{u,v}}^{max} (Weight_Of_DirectPath_{p_i}(u, v, t_0)) \quad (3.3)$$

- 5: **Bước 5:** tương tự phương pháp MPRS+.
-

Phân tích độ phức tạp tính toán của phương pháp RSS+: tương tự phương pháp MPRS+, RSS+ có độ phức tạp tính toán là $O(|R|^2d^3)$ (Với $|R|$ là số lượng nghiên cứu viên, d: bậc trung bình của một nghiên cứu viên).

3.3.5 Thực nghiệm và đánh giá

Hiện nay chưa có tập dữ liệu chuẩn để đánh giá cho bài toán khuyến nghị cộng tác. Hầu hết các nhóm nghiên cứu đều tiến hành thực nghiệm trên tập dữ liệu do họ thu thập và

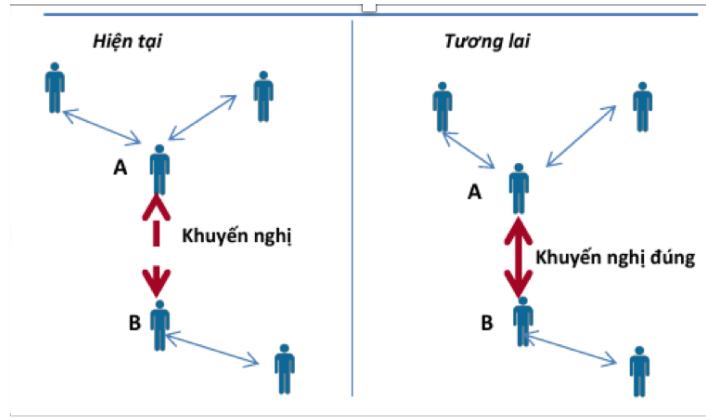
xây dựng. Tang và cộng sự đã thực nghiệm trên tập dữ liệu của hệ thống ArnetMiner cho bài toán khuyến nghị cộng tác liên ngành [117]. Chen và cộng sự triển khai các thực nghiệm của họ trên tập dữ liệu của CiteSeerX cho bài toán khuyến nghị cộng tác, cũng như khám phá các liên kết tiềm năng trên mạng đồng tác giả [28, 27, 29]. Các nhóm nghiên cứu kể trên chỉ đề cập đến số liệu thực nghiệm rút ra như thế nào, chứ họ chưa công bố tập dữ liệu. Với tính phổ biến của DBLP, trong nghiên cứu của mình chúng tôi chọn thực nghiệm trên tập DBLP và tập dữ liệu do chúng tôi rút trích, tích hợp từ nhiều nguồn, CSPubGuru (Phụ lục B). Dữ liệu sử dụng trong thực nghiệm của chúng tôi sẽ được công bố tại trang web <https://sites.google.com/site/tinhuynhuit/dataset> hoặc www.cspubguru.com/DownloadServlet.

Về phương pháp đánh giá cho hệ khuyến nghị, đây là một vấn đề vẫn đang được nghiên cứu. Dáng tin cậy nhất là khảo sát người dùng, phân tích phản hồi của người dùng thông qua hệ thống, hoặc lấy ý kiến chuyên gia. Tuy nhiên, để làm được điều đó thì chúng ta cần phải có hệ thống triển khai sử dụng trên thực tế. Một số nghiên cứu liên quan hiện nay dùng kết quả tiên đoán liên kết đồng tác giả để đánh giá hiệu năng của các phương pháp khuyến nghị cộng tác (Chen et al. [28, 29], Tang et al. [117], Chen et al. [27]). Nếu hệ thống khuyến nghị tiên đoán A cộng tác với B và mối quan hệ đồng tác giả này xảy ra trong tương lai thì xem như đây là một tiên đoán đúng, ngược lại là sai (hình 3.3). Tương tự các nghiên cứu phổ biến hiện nay, luận án cũng dùng độ chính xác tiên đoán liên kết đồng tác giả để đánh giá, so sánh hiệu năng các phương pháp đề xuất với các phương pháp khác.

3.3.5.1 Thiết lập dữ liệu thực nghiệm cho DBLP và CSPubGuru

Chúng tôi dùng dữ liệu các bài báo công bố từ năm 2001 đến năm 2008 để tiến hành thực nghiệm. Dữ liệu 5 năm đầu [2001-2005] được dùng để xây dựng mạng huấn luyện (training network). Dữ liệu các năm sau, từ [2006-2008] dùng để đánh giá độ chính xác khuyến nghị. Kích thước của các tập dữ liệu được trình bày trong bảng 3.1.

Để khách quan, chúng tôi chia mạng huấn luyện thành ba nhóm bậc khác nhau: cao, trung bình và thấp. Những nghiên cứu viên bậc cao là những nghiên cứu viên có số bậc thuộc nhóm 1/3 bậc cao nhất của tất cả các bậc, những nghiên cứu viên bậc thấp là những nghiên cứu viên có số bậc thuộc nhóm 1/3 bậc thấp nhất của tất cả các



Hình 3.3: Minh họa cách đánh giá độ chính xác khuyến nghị cộng tác

Bảng 3.1: Kích thước tập dữ liệu thực nghiệm

Tập CSPubGuru	Kích thước	Tập DBLP	Kích thước
Huấn luyện [2001-2005]		Huấn luyện [2001-2005]	
Số nghiên cứu viên	563.788	Số nghiên cứu viên	369.704
Số bài báo	982.462	Số bài báo	453.980
Dánh giá [2006-2008]		Dánh giá [2006-2008]	
Số nghiên cứu viên	608.443	Số nghiên cứu viên	301.862
Số bài báo	898.325	Số bài báo	458.357

bậc, còn lại là những nghiên cứu viên thuộc nhóm nghiên cứu viên bậc trung bình. Với mỗi loại nhóm bậc nghiên cứu viên, chúng tôi chọn ngẫu nhiên 100 nghiên cứu viên để tiến hành thực nghiệm. Với mỗi nghiên cứu viên, mức độ tương tự với tất cả những người còn lại trong mạng được tính và danh sách TopN những người tương tự nhất được trả về theo các phương pháp khác nhau. Độ chính xác Precision cho tiên đoán liên kết đồng tác giả được tính dựa vào mạng đồng tác giả giai đoạn [2006-2008].

3.3.5.2 Kết quả thực nghiệm

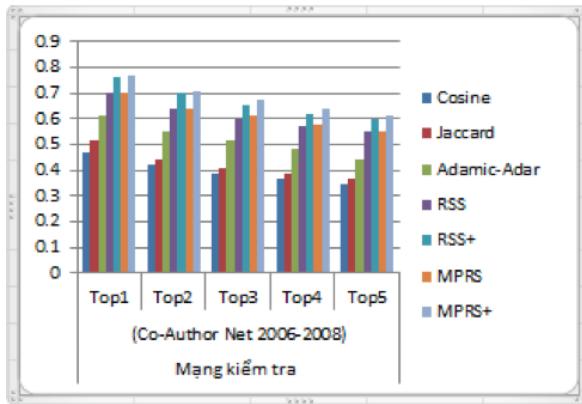
Thực nghiệm được tiến hành chạy trên hệ thống máy chủ UIT-Cloud¹, trên máy ảo có cấu hình như sau:

- Số CPU: 32 processors, Intel (R) Xeon(R) CPU E5-2690 0 @ 2.9GHz
- Bộ nhớ: 80.0 GB Hệ điều hành: Windows 7, 64 bits.

¹<http://mmlab.uit.edu.vn/home/uit-cloud>, truy cập lần cuối 01/09/2014

Thực nghiệm trên tập DBLP

Với việc thiết lập dữ liệu thực nghiệm được mô tả như trong mục 3.3.5.1 và hàm $f_{(Trend)}$ theo công thức 2.18. Trong thực nghiệm này, năm hiện tại t_c là năm 2005, t_0 là năm 2001 là thời điểm bắt đầu quan sát yếu tố xu hướng giữa các mối quan hệ. Khoảng thời gian [2006-2008] dùng để rút trích dữ liệu làm tập đánh giá (Ground Truth). Chúng tôi đánh giá độ chính xác tiên đoán liên kết đồng tác giả (Precision) với Top1, Top2, Top3, Top4, Top5 những đỉnh tương tự nhất được trả về. Với Top5 những đỉnh tương tự nhất được trả về thì MPRS+, RSS+ có độ chính xác lần lượt là 0.61, 0.60 cho tiên đoán đồng tác giả trong tương lai gần [2006-2008]. Trong khi các phương pháp tương tự đỉnh phổ biến hiện nay cao nhất là RSS chỉ đạt 0.55 với tiên đoán tương lai gần (bảng 2, hình 3). Như vậy, kết quả thực nghiệm trên tập DBLP cho thấy các phương pháp đề xuất RSS+, MPRS, MPRS+ cho kết quả tốt hơn so với các phương pháp tương tự đỉnh phổ biến hiện nay (bảng 3.2 và hình 3.4).



Hình 3.4: Kết quả tiên đoán đồng tác giả trên tập thực nghiệm DBLP

Bảng 3.2: Kết quả tiên đoán liên kết đồng tác giả trên tập thực nghiệm DBLP

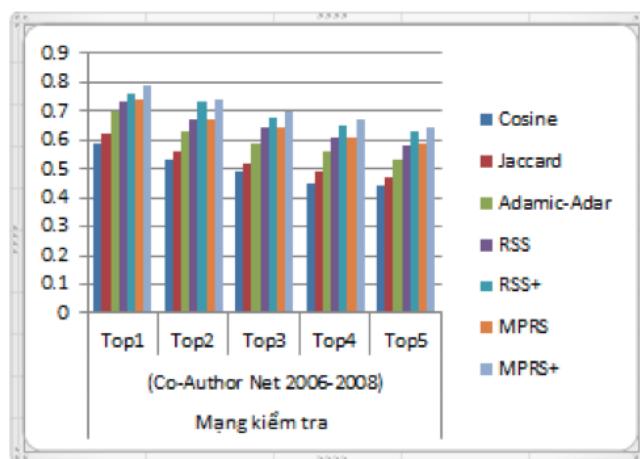
Phương pháp	Giai đoạn đánh giá [2006-2008]				
	P@1	P@2	P@3	P@4	P@5
Cosine	0.47	0.42	0.39	0.37	0.35
Jaccard	0.52	0.44	0.41	0.39	0.37
AdamicAdar	0.61	0.55	0.52	0.48	0.44
RSS	0.70	0.64	0.60	0.57	0.55
RSS+	0.76	0.70	0.65	0.62	0.60
MPRS	0.70	0.64	0.61	0.58	0.55
MPRS+	0.77	0.71	0.67	0.64	0.61

Thực nghiệm trên tập CSPubGuru

Tương tự với việc tiến hành thực nghiệm trên tập DBLP như mô tả ở trên mục 3.3.5.2. Phần này trình bày kết quả thực nghiệm trên một tập dữ liệu khác được nghiên cứu sinh xây dựng bằng cách rút trích, tích hợp từ một số ngôn phổ biến (Phụ lục B). Với Top5 những đỉnh tương tự nhất được trả về thì MPRS+, RSS+ có độ chính xác (Precision) lần lượt là 0.64, 0.63 cho tiên đoán đồng tác giả trong tương lai gần [2006-2008]. Trong khi các phương pháp tương tự đỉnh phổ biến hiện nay cao nhất là RSS chỉ đạt 0.58 với tiên đoán tương lai gần (tham khảo bảng 3.3 và hình 3.5).

Bảng 3.3: Kết quả tiên đoán đồng tác giả trên tập thực nghiệm CSPubGuru

Phương pháp	Giai đoạn đánh giá [2006-2008]				
	P@1	P@2	P@3	P@4	P@5
Cosine	0.59	0.53	0.49	0.45	0.44
Jaccard	0.62	0.56	0.52	0.49	0.47
AdamicAdar	0.70	0.63	0.59	0.56	0.53
RSS	0.73	0.67	0.64	0.61	0.58
RSS+	0.76	0.73	0.68	0.65	0.63
MPRS	0.74	0.67	0.64	0.61	0.59
MPRS+	0.79	0.74	0.70	0.67	0.64



Hình 3.5: Kết quả tiên đoán đồng tác giả trên tập thực nghiệm CSPubGuru

3.3.5.3 Kết luận

Như vậy, kết quả thực nghiệm trên cả hai tập dữ liệu thực nghiệm là DBLP và CSPubGuru (mục 3.3.5.2 và mục 3.3.5.2) đều cho thấy các phương pháp đề xuất RSS+,

MPRS, MPRS+ cho kết quả tốt hơn so với các phương pháp tương tự đính phổ biến hiện nay. Đặc biệt yếu tố xu hướng cộng tác trong hai phương pháp MPRS+ và RSS+ đã giúp cải tiến đáng kể kết quả dựa trên đánh giá tiên đoán liên kết đồng tác giả.

3.4 Trường hợp các nghiên cứu viên chưa có đồng tác giả (Isolated Researcher)

Nghiên cứu viên Isoloated researcher được định nghĩa là những nghiên cứu viên không có bất cứ quan hệ, liên kết nào trong mạng đồng tác giả (định nghĩa 3.2).

3.4.1 Tiếp cận của luận án

Không có các thông tin đồng tác giả, các phương pháp khuyến nghị cộng tác dựa trên việc phân tích mạng đồng tác giả phổ biến hiện nay chắc chắn không thể làm việc được. Để giải quyết vấn đề này, luận án đã đề xuất khai thác các thông tin hỗ trợ khác có trong mạng ASN như: tương tự nội dung dựa trên một số bài báo quan sát được (nếu có), quan hệ của các cơ quan, uy tín và mức độ năng động của các nghiên cứu viên. Các thông tin hỗ trợ này từ mô hình mạng xã hội học thuật ASN được dùng như tập đặc trưng để học và tiên đoán các liên kết đồng tác giả tiềm năng cho các nghiên cứu viên chưa có đồng tác giả [CT.3].

3.4.1.1 Tương tự nội dung nghiên cứu (Content Similarity).

Mặc dù chưa có quan hệ đồng tác giả, nhưng khi quan sát từ kho dữ liệu thực nghiệm thì những nghiên cứu viên này (luận án gọi là isolated researcher) có công bố một vài bài báo. Do đó, những từ khóa thể hiện quan tâm nghiên cứu của họ có thể tồn tại. Vì vậy, độ tương tự về mặt nội dung nghiên cứu của nghiên cứu viên này với những người khác có thể xem như đặc trưng cơ bản (baseline) để so sánh khi đề xuất thêm các đặc trưng khác trong mô hình mạng ASN.

Để tính mức độ tương tự về nội dung nghiên cứu, chúng tôi dùng phương pháp truyền thống trong mã hóa và so khớp văn bản là TF-IDF (Baeza-Yates and Ribeiro-Neto [9]). Các vector đặc trưng w_r và $w_{r'}$ ứng với những nghiên cứu viên r và r' được xây dựng bằng cách tính giá trị TF-IDF cho các thuật ngữ dùng trong các bài báo của

r và r' . Khi đó Độ tương tự nội dung nghiên cứu của r và r' được tính như sau:

$$ContentSim(r, r') = \frac{(w_r \cdot w_{r'})}{\|w_r\| \cdot \|w_{r'}\|} \quad (3.4)$$

3.4.1.2 Quan hệ giữa các cơ quan

Thông thường, những quan hệ cộng tác mới, tiềm năng sẽ hình thành từ những cơ quan có quan hệ cộng tác tốt. Vì thế, hai cơ quan có quan hệ cộng tác tốt sẽ dễ dàng hình thành các quan hệ đồng tác giả mới nhiều hơn so với hai cơ quan ít hoặc không có quan hệ. Dựa trên quan điểm đó, phương pháp tính mức độ quan hệ giữa các cơ quan trong mô hình ASN được sử dụng như một trong những đặc trưng giúp tiên đoán các mối quan hệ cộng tác tiềm năng cho các nghiên cứu viên chưa có quan hệ đồng tác giả. Chi tiết phương pháp, công thức tính mức độ quan hệ giữa các cơ quan được trình bày trong chương 2, mục 2.3.6.4, công thức 2.31, 2.32, 2.33.

3.4.1.3 Uy tín của nghiên cứu viên

Có thể nói rằng, uy tín của các nghiên cứu viên là một trong những nhân tố chính giúp các nghiên cứu viên hình thành nhiều hay ít các mối quan hệ cộng tác mới. Luận án đưa ra giả thuyết: ‘Uy tín của các nghiên cứu viên càng cao khi họ được nhiều trích dẫn của các nghiên cứu viên uy tín khác’. Đây là một giả thuyết có tính đệ qui. Vì vậy, để lượng hóa uy tín của các nghiên cứu viên, luận án đã xây dựng mạng trích dẫn dùng xuất dùng thuật toán random-walk-with-restart (RWR) trên mạng trích dẫn (Tong et al. [121]). Chi tiết về phương pháp, công thức tính uy tín của nghiên cứu viên được trình bày chi tiết trong thành phần M của mô hình ASN chương 2, mục 2.3.6.4, công thức 2.28.

3.4.1.4 Độ năng động của nghiên cứu viên

Thông thường, một nghiên cứu viên càng năng động thì càng có nhiều cơ hội để hình thành các mối quan hệ cộng tác mới. Do đó, mức độ năng động của nghiên cứu viên được bổ sung vào tập đặc trưng để học một mô hình tiên đoán các quan hệ cộng tác tiềm năng cho các nghiên cứu viên chưa có quan hệ đồng tác giả. Phương pháp lượng hóa mức độ năng động cho một nghiên cứu viên r được trình bày chi tiết trong thành phần M của mô hình ASN chương 2, mục 2.3.6.4, công thức 2.29.

3.4.1.5 Học máy để tiên đoán liên kết đồng tác giả, phục vụ khuyến nghị

Chúng ta đưa ra giả thuyết rằng, những cộng tác xảy ra trong tương lai (growth truth) là những cộng tác tốt cần được khuyến nghị. Vì vậy chúng ta có thể dùng kết quả tiên đoán liên kết đồng tác giả để đánh giá kết quả khuyến nghị. Để tiên đoán liên kết đồng tác giả giữa nghiên cứu viên độc lập với các nghiên cứu viên khác, chúng tôi áp dụng phương pháp học giám sát. Chúng tôi dùng SVM, do SVM hiện được đánh giá là phương pháp học giám sát phù hợp nhất cho phân lớp dữ liệu thành hai lớp. Việc học và tiên đoán liên kết đồng tác giả dựa trên tập dữ liệu được gán nhãn cộng tác (mẫu dương) và không cộng tác (mẫu âm).

3.4.2 Phương pháp Đánh giá

Về phương pháp đánh giá cho hệ khuyến nghị, đây là một vấn đề vẫn đang được nghiên cứu. Dáng tin cậy nhất là khảo sát người dùng, phân tích phản hồi của người dùng thông qua hệ thống, hoặc lấy ý kiến chuyên gia. Để làm được điều đó thì chúng ta cần phải có hệ thống triển khai sử dụng trên thực tế để lấy ý kiến phản hồi. Đây là vấn đề khó khăn đối với các nghiên cứu hiện nay. Để giải quyết vấn đề này thì hầu hết các nghiên cứu liên quan đều dùng độ chính xác của kết quả tiên đoán liên kết để đánh giá (Chen et al. [27, 28], Tang et al. [117]).

Bên dưới sẽ trình bày chi tiết hơn về các phương pháp đánh giá đang được dùng phổ biến hiện nay, cũng như phương pháp mới là chất lượng khuyến nghị cộng tác do tác giả luận án nghiên cứu đề xuất.

3.4.2.1 Độ chính xác tiên đoán liên kết

Tương tự với (Chen et al. [28], Tang et al. [117]), để lượng hóa độ chính xác tiên đoán liên kết cho các nghiên cứu viên chưa có đồng tác giả cần được khuyến nghị với các nghiên cứu viên khác, chúng tôi dùng các độ đo phổ biến trong truy vấn thông tin như độ chính xác (Precision), độ bao phủ (Recall), độ đo F, độ chính xác trung bình AP (Average Precision) (Baeza-Yates and Ribeiro-Neto [9]). Nếu hệ thống tiên đoán một cặp (một nghiên cứu viên chưa có đồng tác giả và một nghiên cứu viên khác) sẽ là một cộng tác đồng tác giả và mối quan hệ đồng tác giả này xảy ra trong tương lai thì xem như đây là một tiên đoán đúng, ngược lại là sai (hình 3.3). Chi tiết về các độ đo này

như sau:

- *Precision*: Độ chính xác hay xác suất một tiên đoán cộng tác là đúng (thật sự có xảy ra).
- *Recall*: Độ bao phủ hay xác suất một cộng tác đúng xảy ra trong kết quả tiên đoán.
- *Average Precision*: Trong tuy vấn thông tin, không chỉ bao nhiêu tài liệu liên quan mà thử tự kết quả tìm kiếm trả về cũng đóng vai trò quan trọng giúp định hướng tìm kiếm. Bằng cách tính Precision và Recall ở mỗi vị trí trong danh sách xếp hạng trả về, chúng ta có thể vẽ ra một đường cong Precision-Recall. Vẽ Precision $p(r)$ như một hàm của recall r là $p(r)$. Khi đó, Average Precision tính giá trị trung bình của $p(r)$ qua các khoảng giá trị của r biến thiên từ 0 đến 1.
 - *tp (true positive)*: Số liên kết thật sự đúng được tiên đoán.
 - *fp (false positive)*: Số liên kết không đúng nhưng vẫn được tiên đoán.
 - *fn (false negative)*: Số liên kết thật sự đúng nhưng được không tiên đoán.
 - *tn (true negative)*: Số liên kết không đúng và không được tiên đoán.

Khi đó,

$$Precision = p = \frac{tp}{(tp + fp)} \quad (3.5)$$

$$Recall = r = \frac{tp}{(tp + fn)} \quad (3.6)$$

Dộ đo *F-Measure* là sự kết hợp, cân bằng giá trị của *Precision* và *Recall*.

$$F\text{-Measure} = 2 * \frac{p * r}{p + r} \quad (3.7)$$

Average Precision tính giá trị trung bình của Precision qua các khoảng giá trị của Recall biến thiên từ 0 đến 1.

$$AP = \int_0^1 p(r)dr \quad (3.8)$$

Các phương pháp này phù hợp cho việc đánh giá hiệu quả của tiên đoán liên kết, và đã được nhiều nhóm nghiên cứu sử dụng (Chen et al. [27], Chen et al. [28], Tang et al. [117]). Tuy nhiên khuyến nghị cộng tác không giống như tiên đoán cộng tác. Chúng ta nên xem xét thêm các yếu tố chất lượng, hiệu quả về sau của các mối quan hệ cộng tác được khuyến nghị. Không chỉ đơn thuần là xem xét khuyến nghị đó thực tế có xảy ra hay không, mà còn phải xem nó xảy ra như thế nào. Phần sau sẽ trình bày các phương pháp lượng hóa chất lượng cộng tác mà luận án đề xuất.

3.4.2.2 Đề xuất phương pháp đánh giá chất lượng cộng tác

Khuyến nghị cộng tác khác với tiên đoán cộng tác. Với tiên đoán cộng tác thì chúng ta chỉ cần xem xét bao nhiêu liên kết được tiên đoán đúng. Còn với khuyến nghị cộng tác, chúng ta không chỉ xem xét bao nhiêu liên kết được tiên đoán đúng mà còn xem xét chất lượng của các liên kết tiên đoán đúng dùng để khuyến nghị đó như thế nào. Nó có phải là các cộng tác thật sự tiềm năng để khuyến nghị hay không? Làm thế nào để lượng hóa được chất lượng của các liên kết tiềm năng này. Một số yếu tố có thể xem xét để lượng hóa chất lượng cộng tác như: (1) Số lượng bài báo mới có thể tạo ra; (2) Khoảng thời gian mỗi quan hệ cộng tác được khuyến nghị có thể duy trì; (3) Khả năng mở rộng thêm các mối cộng tác khác từ người được khuyến nghị. Để giới hạn phạm vi thực nghiệm, luận án sử dụng yếu tố thứ nhất và đề xuất lượng hóa chất lượng cộng tác dựa trên số bài báo mới có thể tạo ra từ các cộng tác được khuyến nghị.

(1) Chất lượng cộng tác dựa vào số bài báo mới được tạo ra

Luận án đưa ra giả thuyết: "Một quan hệ cộng tác tốt hơn những quan hệ cộng tác khác nếu nó tạo ra nhiều bài báo hơn". Điều này có nghĩa là, hệ thống nên khuyến nghị những người mà có thể làm việc hiệu quả với các nghiên cứu viên chưa có đồng tác giả và tạo ra càng nhiều công trình, bài báo khoa học càng tốt. Khi đó, chất lượng của TopN những người cộng tác tiềm năng được khuyến nghị, ký hiệu R_{TopN} , có thể lượng hóa như sau:

$$TopNQuality_BasedOn_NumOfPublications(r, R_{TopN}) = \sum_{i=Top1}^{TopN} 1/ranked_order(r_i) * NumOfPublications(r, r_i) \quad (3.9)$$

Trong đó,

- $ranked_order(r_i)$: vị trí xếp hạng của r_i trong danh sách xếp hạng $R_{TopN} = < r_{Top1}, r_{Top2}, \dots, r_{TopN} >$ ($TopN$ người tiềm năng nhất) .
- $NumOfPublications(r, r_i)$: số bài báo đồng tác giả của r và r_i được khuyến nghị.

(2) Chất lượng cộng tác dựa vào khả năng mở rộng nhiều quan hệ mới

Một yếu tố khác có thể được xem xét để đánh giá chất lượng của các mối quan hệ cộng tác, đó là khả năng mở rộng thêm các quan hệ khác từ quan hệ công tác được khuyến nghị. Trong trường hợp này, luận án đưa ra giả thuyết: "Một quan hệ cộng tác tốt hơn những quan hệ cộng tác khác nếu quan hệ đó giúp tạo ra nhiều mối quan hệ cộng tác mới". Khi đó, chất lượng cộng tác của danh sách TopN những người cộng tác tiềm năng được khuyến nghị, ký hiệu R_{TopN} , có thể lượng hóa dựa trên số lượng cộng tác mới tạo ra thông qua quan hệ công tác được khuyến nghị. Một cách hình thức có thể mô hình như sau:

$$TopNQuality_BasedOn_NumOfNewCollaborators(r, R_{TopN}) = \sum_{i=Top1}^{TopN} 1/ranked_order(r_i) * NumOfNewCollaborators(r, r_i) \quad (3.10)$$

Trong đó,

- $NumOfNewCollaborators(r, r_i)$: số lượng người cộng tác mới mà r có được thông qua việc cộng tác với r_i .

3.4.3 Thực nghiệm, đánh giá

Thực nghiệm được tiến hành chạy trên hệ thống máy chủ UIT-Cloud, trên máy ảo có cấu hình như sau:

- Số CPU: 32 processors, Intel (R) Xeon(R) CPU E5-2690 0 @ 2.9GHz
- Bộ nhớ: 80.0 GB Hệ điều hành: Windows 7, 64 bits.

3.4.3.1 Tập dữ liệu thực nghiệm

Theo hiểu biết của chúng tôi, thì hiện nay chưa có tập dữ liệu chuẩn để đánh giá các phương pháp khuyến nghị cộng tác trong nghiên cứu khoa học, đặc biệt đối với trường hợp các nghiên cứu viên chưa có đồng tác giả. Hầu hết các nghiên cứu phổ biến hiện nay tiến hành trên các tập dữ liệu do họ tự xây dựng. (Chen et al. [27, 28]) thực hiện trên tập dữ liệu của hệ thống CiteSeer (Giles et al. [46]). (Tang et al. [117]) triển khai thực nghiệm liên quan đến khuyến nghị cộng tác trên tập dữ liệu thu thập của hệ thống ArnetMiner (Tang et al. [118]). Các tập dữ liệu này hiện nay chưa được công bố công khai. Vì vậy trong nghiên cứu của mình, chúng tôi đã tiến hành thu thập và rút trích các thông tin các bài báo khoa học máy tính từ trang web Microsoft Academic Search¹ để xây dựng tập dữ liệu thực nghiệm của mình. Tập dữ liệu thực nghiệm của chúng tôi, hiện công bố tại địa chỉ <https://sites.google.com/site/tinhuyhnuit/>

Tập dữ liệu thu thập được có 807.005 researchers và 1.266.790 bài báo công bố trong khoảng thời gian 2001 đến 2011. Trong đó, các bài báo và thông tin tác giả trong khoảng [2001-2005] dùng làm dữ liệu huấn luyện và [2006-2011] làm dữ liệu kiểm tra. Những tác giả với thông tin bài báo của họ trong khoảng [2001-2005] dùng để xây dựng mạng huấn luyện đồng tác giả G_0 và những tác giả công bố trong khoảng [2006-2011] dùng để xây dựng mạng đồng tác giả cho kiểm tra và đánh giá G_1 .

Ở đây, chúng tôi lọc các nghiên cứu viên chưa có đồng tác giả là các nghiên cứu viên không có liên kết nào trong G_0 nhưng xuất hiện trong G_1 . Tổng cộng, chúng tôi rút ra được 23.651 nghiên cứu viên chưa có đồng tác giả rừ G_0 , trong đó 1.491 nghiên cứu viên chưa có đồng tác giả có những liên kết mới trong G_1 . Vì lý do tài nguyên tính toán trên hệ thống server đang dùng để chạy các thực nghiệm bị hạn chế, nên chúng tôi chọn ngẫu nhiên 300 nghiên cứu viên chưa có đồng tác giả từ danh sách 1.491 để thực nghiệm.

Từ trong G_1 , chúng tôi rút trích tất cả các liên kết đồng tác giả của 300 nghiên cứu viên chưa có đồng tác giả đã chọn để xây dựng tập mẫu dương (+). Tổng cộng chúng tôi có 1.263 liên kết được gán nhãn như tập mẫu dương. Để đánh giá độ chính xác tiên đoán liên kết dùng phân lớp nhị phân SVM, tập mẫu âm được xây dựng bằng cách chọn ngẫu nhiên các nghiên cứu viên không có liên kết với các nghiên cứu viên chưa có

¹<http://academic.research.microsoft.com/>, truy cập lần cuối 07/02/2014

đồng tác giả trong G_1 . Để cân bằng trong tập dữ liệu số lượng mẫu âm và mẫu dương được chọn bằng nhau. Tổng cộng, tập dữ liệu có 1.263 mẫu dương và 1.263 mẫu âm.

Tiếp đến, khoảng 50% (631) mẫu dương và 50% (631) mẫu âm được dùng để huấn luyện bộ phân lớp SVM. Phần còn lại 632 mẫu dương, 632 mẫu âm dùng để đánh giá kết quả tiên đoán. Các vector đặc trưng cho tất cả các mẫu được xây dựng. Phân bố giá trị của các mẫu trong không gian đặc trưng được thể hiện trực quan trong hình 3.6.

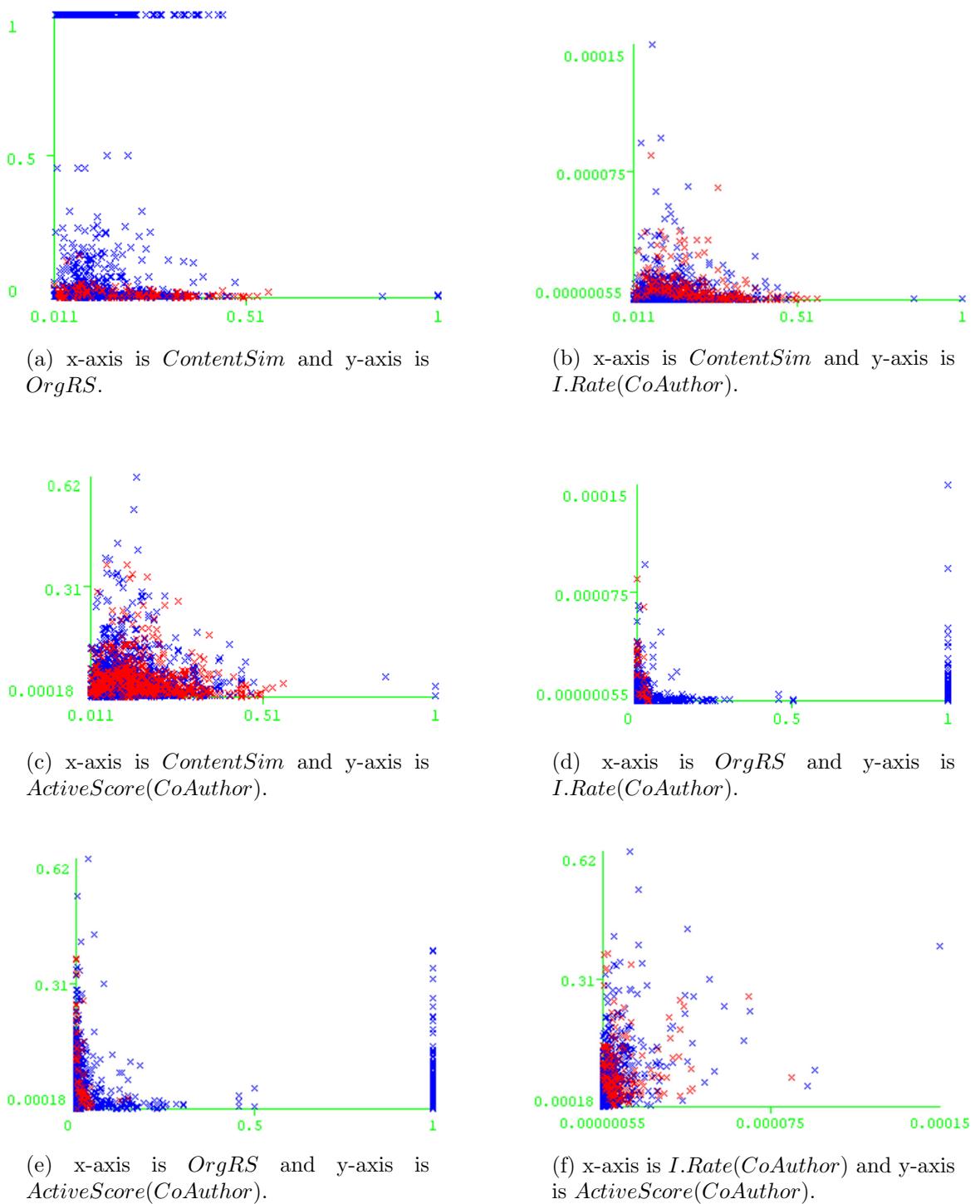
3.4.3.2 Kết quả thực nghiệm

Mục tiêu của tiên đoán liên kết đồng tác giả trong luận án là để đánh giá kết quả khuyến nghị. Do đó, chúng tôi, hướng đến phân tích độ chính xác cũng như chất lượng phân lớp đối với các mẫu dương (các cặp tác giả được tiên đoán có cộng tác) hơn là phân tích các mẫu âm (các cặp tác giả tiên đoán là không có cộng tác). Bảng 3.4 và 3.5 trình bày kết quả thực nghiệm liên quan đến số lượng (độ chính xác tiên đoán) và chất lượng (chất lượng cộng tác dựa trên các tiên đoán cộng tác) khuyến nghị cộng tác khi lần lượt được thay đổi, bổ sung thêm các đặc trưng mới.

Số lượng: Kết quả thực nghiệm thể hiện một quan sát khác thú vị. Đó là đặc trưng về nội dung (tương tự nội dung nghiên cứu ContentSim) không ảnh hưởng đáng kể đến việc tiên đoán, nhận diện các liên kết đồng tác giả. Độ chính xác tiên đoán AP khi chỉ áp dụng duy nhất đặc trưng nội dung là 0.5328 (bảng 3.4).

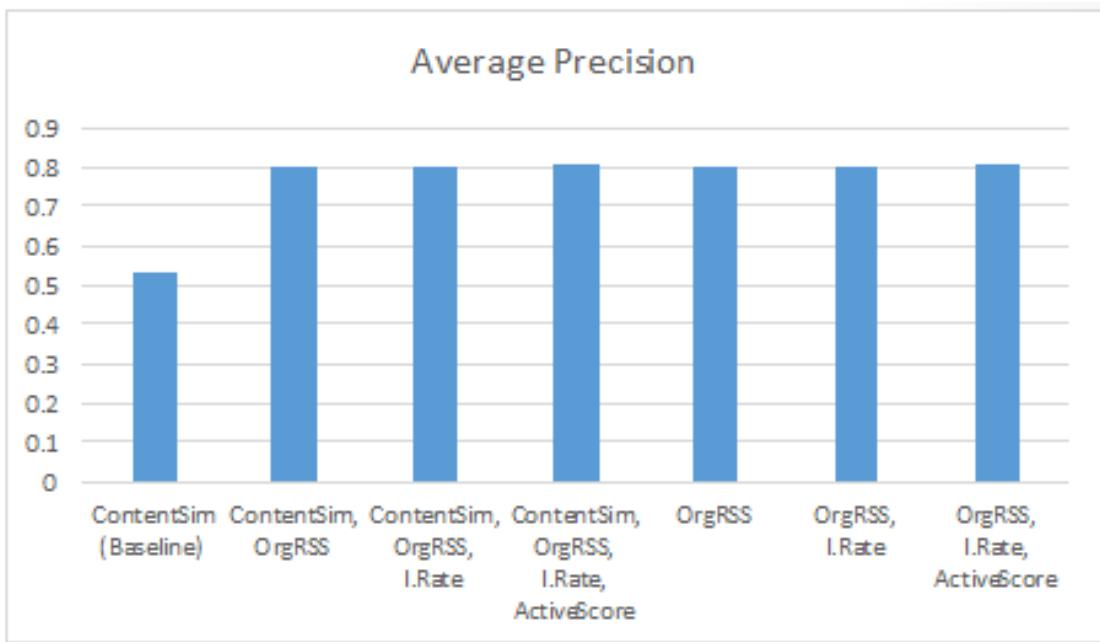
Khi bổ sung thêm đặc trưng "quan hệ giữa các cơ quan" (OrgRS) thì đã có một sự thay đổi đáng kể. Độ chính xác AP tăng từ 0.5328 lên 0.8039. Bên cạnh đó, quan sát thực nghiệm cũng cho thấy việc thêm đặc trưng "uy tín của nghiên cứu viên" (I.Rate) và "độ năng động của nghiên cứu viên" (ActiveScore) không có ảnh hưởng đáng kể đến độ chính xác của tiên đoán liên kết đồng tác giả (Bảng 3.4 và Hình 3.7).

Chất lượng: Việc phân tích khía cạnh chất lượng là nhằm phục vụ tốt hơn cho khuyến nghị cộng tác. Luận án đã tiến hành phân tích chất lượng của TopN các tiên đoán liên kết đồng tác giả. Kết quả phân tích cho thấy, mặc dù đặc trưng "độ năng động của nghiên cứu viên" (ActiveScore) không có ảnh hưởng đáng kể đến độ chính



Hình 3.6: Phân bố của mẫu dương (xanh) và mẫu âm (đỏ) trong không gian đặc trưng 2-chiều.

xác của tiên đoán liên kết đồng tác giả (Bảng 3.4 và Hình 3.7), nhưng nó có ảnh hưởng rất đáng kể chất lượng công tác đối với các liên kết đồng tác giả tiên đoán. Việc bỏ



Hình 3.7: Độ chính xác AP khi thêm các đặc trưng mới

sung đặc trưng OrgRS tăng "chất lượng cộng tác" từ 19.64 lên 74.22 với Top10, và từ 95.30 lên 398.82 với Top50 các liên kết đồng tác giả được tiên đoán sẽ cộng tác (Bảng 3.5).

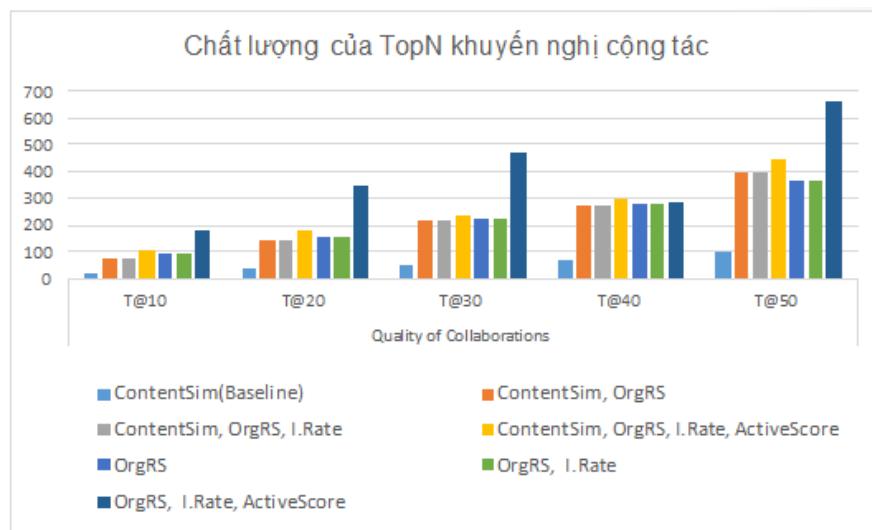
Bên cạnh đó, việc bỏ không xem xét đặc trưng nội dung ContentSim sẽ làm tăng "giá trị chất lượng cộng tác" từ 102.04 lên 178.75 với Top10 và từ 446.19 lên 662.89 với Top50 các liên kết đồng tác giả được tiên đoán sẽ cộng tác. Quan sát thực nghiệm trên tập dữ liệu của chúng tôi lại cho thấy "uy tín của nghiên cứu viên" (I.Rate) không ảnh hưởng đáng kể đối với độ chính xác (AP) cũng như chất lượng ("giá trị chất lượng

Bảng 3.4: Độ chính xác tiên đoán đồng tác giả khi thêm đặc trưng mới

Đặc trưng	Độ chính xác (Precision)	Độ bao phủ (Recall)	Average Precision
<i>ContentSim (Baseline)</i>	0.5113	0.7896	0.5328
<i>ContentSim, OrgRS</i>	0.9079	0.4367	0.8039
<i>ContentSim, OrgRS, I.Rate</i>	0.9079	0.4367	0.8039
<i>ContentSim, OrgRS, I.Rate, ActiveScore</i>	0.8792	0.4953	0.8122
<i>OrgRS</i>	0.9133	0.4335	0.8048
<i>OrgRS, I.Rate</i>	0.9133	0.4335	0.8042
<i>OrgRS, I.Rate, ActiveScore</i>	0.8864	0.4446	0.8113

Bảng 3.5: Chất lượng tiên đoán TopN khi thêm các đặc trưng mới

Đặc trưng	Chất lượng cộng tác dựa trên số bài báo mới được tạo ra				
	T@10	T@20	T@30	T@40	T@50
<i>ContentSim (Baseline)</i>	19.64	36.89	50.50	65.68	95.30
<i>ContentSim, OrgRS</i>	74.22	140.30	213.95	275.75	398.82
<i>ContentSim, OrgRS, I.Rate,</i>	74.23	140.31	213.95	275.75	398.80
<i>ContentSim, OrgRS, I.Rate, ActiveScore</i>	102.04	175.56	233.97	292.16	446.19
<i>OrgRS</i>	91.57	154.52	221.18	278.60	370.69
<i>OrgRS, I.Rate</i>	91.56	154.52	221.18	278.59	370.67
<i>OrgRS, I.Rate, ActiveScore</i>	178.75	349.76	469.04	585.74	662.89



Hình 3.8: Chất lượng tiên đoán TopN khi thêm các đặc trưng mới

cộng tác") của các liên kết đồng tác giả được tiên đoán (chi tiết trong bảng 3.4 3.5). Biểu diễn trực quan cho đánh giá chất lượng cộng tác dựa trên số lượng bài báo mới được tạo ra thể hiện trong hình 3.8.

3.5 Kết chương

Chương này đã giới thiệu, phát biểu bài toán khuyến nghị cộng tác cho các nghiên cứu viên. Luận án đã phân loại các nghiên cứu viên thành 2 nhóm chính: (1) nghiên cứu viên có đồng tác giả; (2) nghiên cứu viên chưa có đồng tác giả và đã phát triển các phương pháp khuyến nghị mới cho hai nhóm nghiên cứu viên này dựa trên việc khai

thác các cấu trúc, mối quan hệ xã hội từ mạng ASN.

Dối với các nghiên cứu viên có đồng tác giả, luận án đã cải tiến phương pháp phổ biến hiện nay, bằng cách đề xuất việc mô hình quan hệ xã hội dựa trên xu hướng. Thực nghiệm được tiến hành trên các tập dữ liệu khoa học như: DBLP, tập dữ liệu tự xây dựng bằng cách tích hợp từ nhiều nguồn (Phụ lục A, B). Kết quả cho thấy các phương pháp dựa trên yếu tố xu hướng đề xuất cho kết quả tốt hơn hẳn các phương pháp khác trong cả hai tập thực nghiệm. Chẳng hạn, với tập DBLP thì độ chính xác tiên đoán đồng tác giả trong tương lai gần với các phương pháp dựa trên xu hướng cho Top5 là 0.60 cho phương pháp RSS+, 0.61 cho phương pháp MPRS+. Trong khi các phương pháp hiện tại chỉ đạt 0.55 (cao nhất) cho Top5 với phương pháp RSS (bảng 3.2).

Dối với các nghiên cứu viên chưa có đồng tác giả (vẫn đề khởi động lạnh trong khuyến nghị), đóng góp của luận án là đã phát biểu bài toán, và đề xuất tập đặc trưng mới để học một mô hình khuyến nghị dựa trên khai thác các mối quan hệ giữa các cơ quan và những đặc trưng ảnh hưởng đến chất lượng cộng tác như: uy tín của nghiên cứu viên, mức độ năng động của nghiên cứu viên bên cạnh đặc trưng truyền thống là độ tương tự về quan tâm nghiên cứu.

Các phương pháp đề xuất cho khuyến nghị cộng tác trong chương này đã cải tiến đáng kể những phương pháp phổ biến hiện nay. Kết quả của chương này đã được trình bày trong các công trình 1, 4.

Chương 4

KHAI THÁC MẠNG XÃ HỘI HỌC THUẬT ĐỂ PHÁT TRIỂN CÁC PHƯƠNG PHÁP KHUYẾN NGHỊ BÀI BÁO KHOA HỌC

4.1 Giới thiệu

Tìm kiếm bài báo khoa học liên quan đến quan tâm nghiên cứu để đọc, tham khảo, trích dẫn là việc làm thường xuyên của những người làm nghiên cứu khoa học, cụ thể là các nghiên cứu viên. Hiện nay, các hệ thống tìm kiếm, thư viện số phổ biến trong lĩnh vực học thuật như ACM DL Portal, IEEE Xplore, Google Scholar, Microsoft Academic Search, DBLP, ... đã đáp ứng hầu hết nhu cầu tìm kiếm tài liệu khoa học của các nghiên cứu viên. Tuy nhiên, đối với các nghiên cứu viên trẻ thì thường chưa đủ hiểu biết và kinh nghiệm để tự tìm ra các thông tin, bài báo hữu ích cho nghiên cứu của mình. Còn đối với các nghiên cứu viên có kinh nghiệm thì phải đương đầu với tình trạng quá tải thông tin, và mất nhiều thời gian hơn để tìm được những tài liệu liên quan. Bên cạnh đó, có thể có nhiều thông tin bài báo liên quan đến quan tâm nghiên cứu mà họ đã bỏ qua, hoặc không tìm thấy. Như vậy, vấn đề đặt ra là “Làm thế nào để hầu hết các bài báo tiềm năng, liên quan đến quan tâm nghiên cứu của các nghiên cứu viên sẽ chủ động tìm đến họ, thay vì họ phải vất vả tự đi tìm kiếm thông tin liên quan?”. Đây chính là động cơ thúc đẩy việc nghiên cứu phát triển các phương pháp, hệ khuyến nghị bài báo khoa học của cộng đồng trong những năm gần đây.

Liên quan đến khuyến nghị bài báo khoa học, có một số dạng bài toán phổ biến có

thể kể đến như sau:

- Nhóm thứ 1, có thể kể đến là các nghiên cứu liên quan đến việc khuyến nghị bài báo liên quan đến bài báo. Tức khi người dùng duyệt tìm kiếm các bài báo, thì hệ thống sẽ khuyến nghị các bài báo tương tự với bài báo mà người dùng đang duyệt. Vấn đề chính của bài toán này là biểu diễn và so khớp nội dung của bài báo đang duyệt với các bài báo trong kho dữ liệu. Một số nghiên cứu đi vào phân tích, biểu diễn và so khớp nội dung (Ohta et al. [93], El-Arini and Guestrin [39]). Một số nghiên cứu khác thì so khớp tương tự dựa trên số lượng các bài báo tham khảo và trích dẫn chung (Lawrence et al. [66], Wanjanuk [125], Huynh et al. [55]).
- Nhóm thứ 2, là các nghiên cứu liên quan đến việc khuyến nghị bài báo trích dẫn bên trong bài báo. Tức là, các nghiên cứu tìm cách đề xuất các bài báo tiềm năng có thể trích dẫn cho một câu hay một đoạn trong bài báo đang viết. Tiếp cận chủ yếu dựa trên phân tích và so khớp nội dung của câu, đoạn trong bài viết với nội dung các bài báo quan sát được trong kho dữ liệu (He et al. [51, 50], Huang et al. [53]).
- Nhóm thứ 3, là các nghiên cứu liên quan đến việc khuyến nghị các bài báo nghiên cứu tiềm năng liên quan đến quan tâm nghiên cứu của một nghiên cứu viên (Wang and Blei [124], Sugiyama and Kan [111, 112, 113], Sun et al. [114]).

Nhằm giúp các nghiên cứu viên có thể nắm bắt, cập nhật tình hình nghiên cứu liên quan, cũng như tổng hợp thông tin để thực hiện nghiên cứu, viết bài, trong phạm vi luận án này chúng tôi tập trung phát triển các phương pháp khuyến nghị bài báo khoa học tiềm năng phù hợp với quan tâm nghiên cứu của nghiên cứu viên (bài toán thuộc nhóm thứ 3 kể trên). Mục đích của chương này là trình bày chi tiết bài toán khuyến nghị bài báo khoa học, cũng như việc áp dụng tiếp cận khai thác các mối quan hệ xã hội học thuật, được gọi là mạng xã hội học thuật để phát triển các phương pháp mới cho bài toán con này. Kết quả của chương này đã được công bố trong cách công trình: [CT2], [CT8].

4.2 Bài toán Khuyến nghị bài báo khoa học

Trong phạm vi luận án, khuyến nghị bài báo khoa học cho nghiên cứu viên là bài toán với đầu vào là một hay nhiều nghiên cứu viên và tập các bài báo khoa học quan sát được. Hệ thống sẽ trả về danh sách xếp hạng các bài báo khoa học tiềm năng, ứng với quan tâm nghiên cứu của mỗi nghiên cứu viên. Dựa trên phần phát biểu hình thức bài toán khuyến nghị tổng quát trong chương 1, phần này sẽ trình bày một số định nghĩa và phát biểu một cách hình thức cụ thể cho bài toán Khuyến nghị bài báo khoa học.

Định nghĩa 4.1: *Không gian nghiên cứu viên*

Không gian nghiên cứu viên là tập tất cả những nghiên cứu viên mà hệ thống cần thực hiện các phân tích, khuyến nghị. Ký hiệu là R , $R = \{r_1, r_2, r_3, \dots, r_n\}$.

Định nghĩa 4.2: *Không gian bài báo khoa học (đối tượng khuyến nghị)*

Không gian bài báo khoa học là tập tất cả những bài báo khoa học mà hệ thống quan sát được. Tập các bài báo này sẽ được dùng để phân tích, tính toán mức độ phù hợp đến quan tâm nghiên cứu của mỗi nghiên cứu viên để thực hiện khuyến nghị. Ký hiệu là P , $P = \{p_1, p_2, p_3, \dots, p_m\}$.

Định nghĩa 4.3: *Hàm ước lượng mức độ tiềm năng và hữu ích của bài báo*

Mức độ tiềm năng của một bài báo $p \in P$ ứng với một nghiên cứu viên $r \in R$, được xác định bởi một ánh xạ $f : R \times P \rightarrow \mathbb{R}$, ký hiệu $f(r, p)$ hoặc $rating(r, p)$. Với \mathbb{R} là tập có thứ tự các số nguyên hoặc thực trong một khoảng nhất định. Trong lĩnh vực học thuật, mức độ hữu ích này có thể thể hiện thông qua việc trích dẫn, tham chiếu.

Khi đó, nếu $f(r, p) \neq null$ thì $f(r, p)$ chính là đánh giá (thật sự) của r đối với p , còn nếu $f(r, p) = null$ tức cần phải ước lượng mức độ hữu ích của p đối với r .

Định nghĩa 4.4: *Phát biểu hình thức bài toán khuyến nghị bài báo*

Cho trước,

- $R = \{r\}$: tập tất cả những nghiên cứu viên.
- $P = \{p\}$: tập tất cả các bài báo đã quan sát.

-
- $R_p \subseteq R$: tập những nghiên cứu viên r đã thể hiện đánh giá, quan tâm với các bài báo khoa học p
 - $P_r \subseteq P$: tập những bài báo khoa học được các nghiên cứu viên r đánh giá, thể hiện sự quan tâm.
 - $Existed_Rating = v(r', p')$. Các đánh giá quan sát được, thể hiện mức độ liên quan của bài báo $p' \in P_r$ với nghiên cứu viên $r' \in R_p$.

Mục đích của hệ khuyến nghị bài báo khoa học là xây dựng hàm hữu ích $f(r, p)$ và ước lượng giá trị của hàm f để tiên đoán xem r sẽ quan tâm đến p nhiều hay ít, hay p tiềm năng và hữu ích đối với r như thế nào. Đối với mỗi nghiên cứu viên r_i , hệ khuyến nghị cần chọn $TopN$ bài báo khoa học, $P_{TopN} = \langle p_1, p_2, \dots, p_{TopN} \rangle$, tiềm năng và hữu ích nhất đối với nghiên cứu viên r_i để khuyến nghị. Các bài báo $P_{TopN} = \langle p_1, p_2, \dots, p_{TopN} \rangle$ được chọn thỏa mãn các điều kiện sau:

- $\forall p_k \in P_{TopN}, v(r_i, p_k) \notin Existed_Rating$. Tức phải khuyến nghị những bài báo p_k mà nghiên cứu viên r_i chưa biết.
- $\forall p_k \in P_{TopN}, f(r_i, p_k) \geq f(r_i, p_{k+1})$, với $1 \leq k \leq n - 1$. Tức là tập các bài báo khuyến nghị P_{TopN} là tập có thứ tự. Bài báo đứng trước có giá trị của hàm hữu ích f lớn hơn hoặc bằng và ưu tiên khuyến nghị cho r_i hơn bài báo đứng sau. Trường hợp dấu bằng xảy ra thì mức độ ưu tiên khuyến nghị là như nhau.
- $\forall p_k \in P_{TopN}, \forall p_{no_rec} \in P \setminus P_{TopN}$, thì $f(r_i, p_k) \geq f(r_i, p_{no_rec})$. Tức giá trị hữu ích của các bài báo được khuyến nghị, được xác định thông qua hàm f , phải lớn hơn hoặc bằng những bài báo không được khuyến nghị.

4.3 Khó khăn, thách thức

Tương tự các hệ khuyến nghị khác, hệ khuyến nghị bài báo khoa học, cũng có những thách thức khó khăn như sau:

- Dữ liệu lớn. Không gian nghiên cứu viên R và bài báo P là rất lớn.
- Ma trận đánh giá thừa. Ma trận thể hiện sự đánh giá, quan tâm của các nghiên cứu viên đến các bài báo là rất thừa.

-
- Vấn đề khởi động lạnh. Quan sát thiếu hay không thể quan sát được các thông tin về nghiên cứu viên, cũng như bài báo khoa học.
 - Độ chính xác khuyến nghị.
 - Các phương pháp đánh giá kết quả bài báo khuyến nghị.

4.4 Nghiên cứu liên quan

Như đã trình bày phần trên, liên quan đến khuyến nghị bài báo khoa học, có một số bài toán con khác nhau mà các nghiên cứu hiện nay đang quan tâm: (1) Bài toán khuyến nghị bài báo trích dẫn cho các nghiên cứu viên khi viết bài. Một số nghiên cứu điển hình có thể kể đến như: (He et al. [51, 50], Huang et al. [53]). Các nghiên cứu này nhằm phát triển mô hình cho phép ánh xạ giữa các câu trong bài báo với tài liệu trích dẫn, nhằm hỗ trợ trích dẫn trong lúc viết bài báo khoa học. (2) Hai là nhóm các nghiên cứu khác nhằm phát triển các thuật toán khuyến nghị các bài báo tương tự khi người dùng duyệt qua một bài báo trong thư viện số (Ohta et al. [93], El-Arini and Guestrin [39], Lawrence et al. [66], Wanjantuk [125], Huynh et al. [55]). Và cuối cùng là các nghiên cứu liên quan đến bài toán nhằm tìm kiếm, lọc ra danh sách những bài báo phù hợp với quan tâm nghiên cứu của nghiên cứu viên để khuyến nghị. Với bài toán này thì các nghiên cứu của Wang and Blei [124], Sugiyama and Kan [111, 112, 113], Sun et al. [114] là các nghiên cứu tương tự nhất với vấn đề nghiên cứu và trình bày trong chương này.

Đối với hầu hết những bài toán có đối tượng khuyến nghị dạng văn bản thì tiếp cận dựa trên nội dung, gọi tắt là tiếp cận nội dung được xem là tiếp cận phù hợp nhất (Adomavicius et al. [4]). Với tiếp cận nội dung, hệ thống sẽ mô hình hóa sở thích nghiên cứu của các nghiên cứu viên dựa trên việc tích hợp nội dung các bài báo mà họ đã công bố trong quá khứ. Sau đó, sở thích của các nghiên cứu viên sẽ được so khớp với nội dung của các bài báo quan sát được. Một danh sách xếp hạng các bài báo liên quan trả về sẽ được đề xuất cho các nghiên cứu viên.

Chủ đề của một bài báo có thể xác định thông qua nội dung của nó và các bài báo tham khảo, trích dẫn. Hay nói cách khác, nội dung các bài báo tham khảo, trích dẫn đóng góp vào việc hình thành, xác định chủ đề của một bài báo. Vì vậy, (Sugiyama

and Kan [111, 112]) đã đề các xuất các phương pháp tiếp cận nội dung mới cho khuyến nghị bài báo khoa học phù hợp với quan tâm nghiên cứu của các nghiên cứu viên. Dóng góp chính của họ là khai thác quan tâm tiềm ẩn trong hồ sơ sở thích của các nghiên cứu viên từ bài báo trong quá khứ kết hợp với các bài báo tham khảo và bài báo trích dẫn của các nghiên cứu viên từ mạng trích dẫn. Họ đã thu thập 597 bài báo từ hội nghị ACL và lấy ý kiến 28 nghiên cứu viên. 28 nghiên cứu viên này sẽ xem danh sách 597 bài báo và cho biết bài báo nào liên quan hay không liên quan đến quan tâm nghiên cứu của họ. Tác giả đã dùng tập dữ liệu gán nhãn này để xây dựng tập đánh giá (Ground Truth). Bản chất của mạng trích dẫn này là một mạng rất thưa. Do đó, Sugiyama et al., 2013 đã tìm cách giảm bớt dữ liệu thưa bằng lọc cộng tác để khám phá bài báo trích dẫn tiềm năng và dùng các bài trích dẫn tiềm năng để tinh chỉnh việc dùng bài báo trích dẫn để mô hình hóa bài báo ứng viên. Kết quả thực nghiệm của họ cho thấy việc khai thác bài báo trích dẫn tiềm năng đã cải tiến độ chính xác khuyến nghị (Sugiyama and Kan [113]).

Trong một nghiên cứu khác, Jianshan Sun et al., 2013 đã đề xuất các phương pháp mới cho khuyến nghị bài báo khoa học liên quan đến quan tâm nghiên cứu của nghiên cứu viên bằng cách kết hợp thông tin nội dung của các bài báo quan tâm và các mối quan hệ xã hội của nghiên cứu viên (Sun et al. [114]). Họ đã rút trích danh sách các bài báo liên quan và các mối quan hệ xã hội của những nghiên cứu viên từ trang mạng trực tuyến CiteULike để xây dựng tập dữ liệu thực nghiệm bao gồm tập đánh giá (ground truth), tập huấn luyện (training set), cũng như tập kiểm tra (testing set). Kết quả thực nghiệm cho thấy phương pháp kết hợp thông tin nội dung và quan hệ xã hội rút trích từ các mạng trực tuyến CiteULike đã cải tiến chất lượng khuyến nghị so với phương pháp tiếp cận nội dung.

Về các tập dữ liệu và phương pháp đánh giá thì Joeran Beel et al., 2013 đã thực hiện một khảo sát 176 bài báo phổ biến nhất hiện nay và chỉ ra rằng: cho đến bây giờ vẫn chưa có sự đồng thuận, thống nhất về các tập dữ liệu cũng như phương pháp đánh giá khi thực hiện so sánh các phương pháp khuyến nghị bài báo khoa học khác nhau (Beel et al. [15]). Điều đó dẫn đến một tình trạng chung, đó là chưa thể biết được những điểm mạnh và yếu thật sự của những phương pháp đề xuất hiện có.

Mặc dù là cách tiếp cận phù hợp và đã chứng tỏ được ưu điểm vượt trội trong nhiều

nghiên cứu, nhưng các phương pháp dựa trên tiếp cận nội dung phổ biến hiện nay vẫn còn những hạn chế nhất định. Một số hạn chế có thể nhận thấy và kể đến như sau:

- **Xu hướng sở thích:** thời gian, xu hướng chưa được quan tâm xem xét khi mô hình hóa sở thích nghiên cứu của nghiên cứu viên. Ví dụ, một nghiên cứu viên trong quá khứ quan tâm đến vấn đề khai thác dữ liệu, nhưng do ảnh hưởng của đồng nghiệp, xu hướng nghiên cứu của cộng đồng, nên gần đây lại bắt đầu quan tâm đến truy vấn thông tin thông minh, hệ khuyến nghị.
- **Tính mới của bài báo:** thật sự không phù hợp nếu chọn một bài báo có nội dung liên quan, nhưng quá cũ để ưu tiên khuyến nghị.
- **Chất lượng của bài báo:** không nên chỉ xem xét các bài báo có nội dung tương tự sở thích của nghiên cứu viên để ưu tiên khuyến nghị. Bên cạnh mức độ tương tự với sở thích của nghiên cứu viên, cần xem xét những bài báo có chất lượng tốt, của những chuyên gia có uy tín để ưu tiên khuyến nghị.

Câu hỏi đặt ra là làm thế nào để mô hình hóa xu hướng sở thích của một nghiên cứu viên, tính mới, cũng như chất lượng bài báo. Trên thực tế, những chuyên gia uy tín thường là những người sẽ sản sinh ra nhiều công trình tốt, có chất lượng được công đồng đón nhận, trích dẫn và đặt lòng tin. Như vậy câu hỏi tiếp theo là, làm thế nào để lượng hóa được uy tín của chuyên gia, lòng tin của một nghiên cứu viên đối với các nghiên cứu viên khác? Và lòng tin ảnh hưởng như thế nào đến việc quyết định chọn bài báo để đọc, trích dẫn? Do đó, luận án đã tiếp cận bằng cách mô hình quan tâm nghiên cứu đối tượng nghiên cứu viên và các mối quan hệ của nghiên cứu viên trong mạng xã hội học thuật ASN dựa trên yếu tố xu hướng. Khai thác cấu trúc mạng xã hội học thuật (dùng cấu trúc đồng tác giả *CoNet* và cấu trúc trích dẫn *CiNet_Author*) để lượng hóa uy tín của nghiên cứu viên, lòng tin của một nghiên cứu viên đặt lên những nghiên cứu viên khác. Từ đó phát triển các phương pháp khuyến nghị bài báo khoa học tiềm năng cho nghiên cứu viên.

Phần bên dưới, sẽ trình bày các tiếp cận, phương pháp truyền thống, cũng như những phương pháp đề xuất cho khuyến nghị bài báo khoa học.

4.5 Các phương pháp phổ biến cho khuyến nghị bài báo liên quan

4.5.1 Tiếp cận nội dung

Tiếp cận nội dung (Content-based methods) hiện đang được đánh giá là tiếp cận phù hợp nhất cho các đối tượng khuyến nghị dạng văn bản (Adomavicius and Tuzhilin [5]).

Với tiếp cận nội dung, vector biểu diễn hồ sơ nghiên cứu của các nghiên cứu viên và vector biểu diễn nội dung bài báo sẽ được xây dựng và so khớp. Trong phạm vi bài toán này, chúng tôi dùng phương pháp mô hình hóa sở thích của nghiên cứu viên dựa trên nội dung các bài báo đã công bố trong quá khứ như phương pháp cơ sở để đánh giá so sánh với các phương pháp đề xuất.

4.5.1.1 CB-Baseline

Dây là phương pháp mô hình hóa quan tâm nghiên cứu của nghiên cứu viên dựa trên việc tổng hợp nội dung các bài báo công bố trong quá khứ, gọi tắt là CB-Baseline, có thể tóm tắt qua các bước như sau:

Phương pháp: CB-Baseline

Đầu vào:

- Tập các nghiên cứu viên $R = \{r\}$
- Không gian các bài báo quan sát được $P = \{p\}$

Đầu ra: $\forall r \in R$, trả về P_{TopN} dựa trên giá trị hữu ích tiên đoán để khuyến nghị cho r .

1: **Bước 1:** Tiền xử lý các bài báo $p \in P$

- Rút trích phần tiêu đề và tóm tắt.
- Loại bỏ stopwords, và stemming.

2: **Bước 2:** Vector hóa nội dung các bài báo dùng TFIDF.

- $\forall p \in P$: xây dựng vector biểu diễn nội dung bài báo p là \vec{f}_p dùng phương pháp gán trọng số TFIDF.

3: **Bước 3:** Vector hóa sở thích các nghiên cứu viên.

- $\forall r \in R$: xây dựng vector profile \vec{P}_r cho mỗi nghiên cứu viên r dựa vào các bài báo mà r đã công bố.

$$\vec{P}_r = \sum_{i=1}^n \vec{f}_p \quad (4.1)$$

Trong đó, n : Tổng số bài báo mà r đã công bố.

4: **Bước 4:** So khớp nội dung bài báo với sở thích của nghiên cứu viên.

- Lặp $\forall r \in R$,

$\forall p \in P$, xếp hạng và chọn $TopN$ những bài báo có độ tương tự cao nhất với r dựa trên giá trị của $SimCB(r,p)$, mà r chưa biết đến trước đây để thực hiện khuyến nghị cho r .

$$SimCB(r,p) = Cosine(\vec{w}_p, \vec{w}_r) \quad (4.2)$$

- Cuối lặp
-

4.5.1.2 Mô hình hóa sở thích của các nghiên cứu viên dựa trên nội dung các bài báo công bố, tham khảo, và trích dẫn (CB+R+C)

Phương pháp này được đề xuất bởi Sugiyama and Kan [111]. Họ quan niệm, quan tâm nghiên cứu của nghiên cứu viên không chỉ thể hiện thông qua nội dung của các bài báo mà họ công bố, mà còn được thể hiện thông qua nội dung của các bài báo mà họ tham khảo (ký hiệu R), được trích dẫn (ký hiệu C). Do đó, Sugiyama et al. đã tổng hợp vector đặc trưng của tất cả các bài báo công bố kết hợp với vector đặc của bài tham khảo, trích dẫn để mô hình hóa quan tâm nghiên cứu của các nghiên cứu viên.

Thuật toán CB+R+C của họ có thể tóm tắt như sau:

Phương pháp: CB+R+C

Đầu vào:

- Tập các nghiên cứu viên $R = \{r\}$
- Không gian các bài báo quan sát được $P = \{p\}$

Đầu ra: $\forall r \in R$, trả về P_{TopN} dựa trên giá trị hữu ích tiên đoán để khuyến nghị cho r .

1: **Bước 1:** Tương tự phương pháp 1.

2: **Bước 2:** Mô hình hóa nội dung bài báo bằng cách tổng hợp các vector đặc trưng của bài tham khảo, bài được trích dẫn.

$$\overrightarrow{F_p} = \overrightarrow{f_p} + \sum_{i=1}^m Sim(p, Ref_i(p)) * \overrightarrow{f_{Ref_i(p)}} + \sum_{i=1}^n Sim(p, Cite_i(p)) * \overrightarrow{f_{Cite_i(p)}} \quad (4.3)$$

Trong đó,

- m : Tổng số bài mà p đã tham khảo,
- n : Tổng số bài đã trích dẫn bài p ,
- $Ref_i(p)$: bài báo tham khảo thứ i của p ,
- $Cite_i(p)$: bài báo thứ i đã trích dẫn bài p .

3: **Bước 3:** Vector hóa quan tâm nghiên cứu của các nghiên cứu viên.

$\forall r \in R$: xây dựng vector profile $\overrightarrow{P_r}$

$$\overrightarrow{P_r} = \sum_{i=1}^n \overrightarrow{F_{p_i}} \quad (4.4)$$

Trong đó, n : Tổng số bài báo mà r đã công bố.

4: **Bước 4:** Tương tự phương pháp 1.

Để lọc bỏ những bài báo không liên quan khi xem xét các bài báo tham khảo và trích dẫn, Sugiyama et al. đã đề xuất sử dụng một tham số ngưỡng tương tự ($Th_j \in [0, 1]$) để quyết định chọn ra những bài tham khảo, trích dẫn dùng để kết hợp với các bài báo khác khi xây dựng mô hình sở thích của nghiên cứu viên (Sugiyama and Kan [111]). Tức $Sim(p, Ref_i(p)) > Th_j$, $Sim(p, Cite_i(p)) > Th_j$, thì khi đó vector đặc trưng của $Ref_i(p)$ và $Cite_i(p)$ sẽ được kết hợp với vector đặc trưng của p .

4.5.1.3 Phương pháp mô hình hóa xu hướng nghiên cứu của nghiên cứu viên (CB-Recent)

Các phương pháp mô hình hóa sở thích nghiên cứu của các nghiên cứu viên thông thường chỉ tập trung vào việc mã hóa nội dung các bài báo mà họ công bố, tham khảo hoặc được trích dẫn. Trên thực tế, sở thích của người dùng sẽ dần thay đổi theo thời gian. Vì vậy, quan tâm nghiên cứu của nghiên cứu viên sẽ bị chi phối bởi những bài báo công bố trong thời gian gần đây hơn những bài mà nghiên cứu viên công bố quá lâu trong quá khứ. Sugiyama và đồng nghiệp cũng đã khai thác yếu tố thời gian và đề xuất phương pháp mô hình quan tâm nghiên cứu gần đây của nghiên cứu viên cho bài

toán khuyến nghị bài báo khoa học, gọi tắt là phương pháp CB-Recent [111]. Phương pháp CB-Recent có thể tóm tắt như sau:

Phương pháp: CB-Recent

Đầu vào:

- Tập các nghiên cứu viên $R = \{r\}$
- Không gian các bài báo quan sát được $P = \{p\}$

Đầu ra: $\forall r \in R$, trả về P_{TopN} dựa trên giá trị hữu ích tiên đoán để khuyến nghị cho r .

- 1: **Bước 1:** Tương tự phương pháp CB-Baseline.
- 2: **Bước 2:** Tương tự bước 2 của phương pháp CB+R+C.
- 3: **Bước 3:** Vector hóa sở thích các nghiên cứu viên dựa trên xu hướng. $\forall r \in R$: xây dựng vector profile \vec{P}_r cho mỗi nghiên cứu viên r dựa trên yếu tố xu hướng.

$$\vec{P}_r = \sum_{i=1}^n e^{\gamma * (t_c - t(p_i))} * \vec{f}_{p_i} \quad (4.5)$$

Trong đó,

- γ : hệ số xu hướng. ($\gamma \in [0, 1]$). Trường hợp đơn giản $\gamma = 1$)
 - t_c : năm hiện tại thực hiện khuyến nghị.
 - $t(p_i)$: năm công bố của bài báo p_i .
 - n : Tổng số bài báo mà r công bố trong quá khứ.
 - \vec{f}_{p_i} : vector biểu diễn nội dung bài báo p_i .
- 4: **Bước 4:** Thực hiện khuyến nghị. $\forall r \in R, \forall p \in P$,
- $f(r, p) = Sim_{CB}(r, p) = Cosine(\vec{F}_{p_i}, \vec{P}_r)$
 - Chọn TopN các $p \in P$ có $f(r, p)$ lớn nhất, tức P_{TopN} , để khuyến nghị cho r .
-

Phân tích độ phức tạp tính toán của phương pháp CB-Recent:

- Các bước 1 và 2 là tiền xử lý.
- Xét bước 3: Cần tính vector cho mọi nghiên cứu viên. Với mỗi nghiên cứu viên, ta cần tính tổng vector của n bài báo. Vậy độ phức tạp là $\mathcal{O}(|R|n)$ (n : Tổng số bài báo mà r công bố trong quá khứ).
- Xét bước 4: Cần xét tất cả các NCV và tất cả các bài báo. Vậy độ phức tạp là $\mathcal{O}(|R||P|)$. Do $|P| >> n$, nên tóm lại, độ phức tạp của thuật toán là $\mathcal{O}(|R||P|)$

(Với $|R|$: số lượng nghiên cứu viên, $|P|$: số lượng bài báo).

4.5.2 Tiếp cận lọc cộng tác - CF

Khác với tiếp cận nội dung, tiếp cận lọc cộng tác (tiếp cận CF) không bị hạn chế về mặt phân tích nội dung văn bản. Những phương pháp CF dùng thông tin từ ma trận đánh giá quan sát được từ người dùng và đối tượng khuyến nghị. Tiếp cận CF có thể áp dụng cho nhiều dạng đối tượng, nhiều kiểu nội dung khác nhau, ngay cả với những đối tượng khuyến nghị không tương tự với những đối tượng quan sát trong quá khứ. Các phương pháp CF được đánh giá là các phương pháp thành công nhất trong việc xây dựng các hệ thống khuyến nghị nói chung (Su and Khoshgoftaar [110]).

Với bài toán khuyến nghị bài báo khoa học liên quan cho các nghiên cứu viên, giả sử các bài báo được các nghiên cứu viên tham khảo, trích dẫn là các bài có liên quan đến quan tâm nghiên cứu của họ. Khi đó, chúng ta có thể xây dựng ma trận đánh giá A dựa trên quan hệ trích dẫn, nhằm thể hiện sự quan tâm của các nghiên cứu viên đối với các bài báo trong kho dữ liệu. A có dòng là các nghiên cứu viên và cột là các bài báo. Giá trị ở dòng i , cột j trong ma trận A thể hiện sự quan tâm của nghiên cứu viên r_i với bài báo p_j .

$$f(r_i, p_j) = \frac{\text{CitationCount}(r_i, p_j)}{\text{TotalCount}(r_i)} \quad (4.6)$$

Trong đó,

- $\text{CitationCount}(r_i, p_j)$: số lần mà nghiên cứu viên r_i đã trích dẫn bài báo p_j trong quá khứ.
- $\text{TotalCitation}(r_i)$: tổng số trích dẫn của r_i .

Dựa trên quan điểm này, chúng ta có thể xây dựng phương pháp lọc cộng tác cho bài toán khuyến nghị bài báo khoa học liên quan.

Lọc cộng tác với CF-kNN

Phương pháp tiên đoán mức độ liên quan của các bài báo khoa học với các nghiên cứu viên dựa trên tiếp cận CF, có thể tóm tắt như sau:

Phương pháp: CF-kNN

Đầu vào:

- Tập các nghiên cứu viên $R = \{r\}$
- Không gian các bài báo quan sát được $P = \{p\}$

Đầu ra: $\forall r \in R$, trả về P_{TopN} dựa trên giá trị hữu ích tiên đoán để khuyến nghị cho r .

1: **Bước 1:** Xây dựng ma trận A có giá trị tại dòng i , cột j thể hiện mức độ liên quan của các $p_j \in P$ với $r_i \in R$, $f(r_i, p_j)$.

2: **Bước 2:** Xác định k người đồng sở thích.

Đầu lặp: $\forall r_i \in R$

- Dùng thuật toán kNN để xác định k người có sở thích tương tự r_i , ký hiệu $kNN(r_i)$. Độ tương tự sở thích của $r \in R$ với r_i có thể tính theo hệ số tương quan Pearson dựa trên ma trận A như sau:

$$Sim_{Pearson}(r_i, r) = \frac{\sum_{p_j \in P_{r,r_i}} (v(r_i, p_j) - \bar{v}(r_i)) * (v(r, p_j) - \bar{v}(r))}{\sqrt{\sum_{p_j \in P_{r,r_i}} (v(r_i, p_j) - \bar{v}(r_i))^2} * \sqrt{\sum_{p_j \in P_{r,r_i}} (v(r, p_j) - \bar{v}(r))^2}} \quad (4.7)$$

Trong đó,

- P_{r,r_i} : Tập các bài báo mà r, r_i đồng trích dẫn trong quá khứ.
- $\bar{v}(r_i)$: giá trị trung bình trích dẫn của nghiên cứu viên r_i trên các bài báo p_j .

Cuối lặp

3: **Bước 3:** Tổng hợp giá trị từ k người đồng sở thích, để tiên đoán những giá trị $v(i, j)$ chưa xác định trong A và chọn TopN để khuyến nghị.

Đầu lặp: $\forall r_i \in R, p_j \in P : v(r_i, p_j) = 0$

$$v(r_i, p_j) = k * \sum_{r' \in kNN(r_i)} Sim_{Pearson}(r', r_i) * v(r', p_j) \quad (4.8)$$

Trong đó,

- k: hệ số chuẩn hóa, $k = \frac{1}{\sum_{r' \in kNN(r_i)} |Sim_{Pearson}(r', r_i)|}$
- Chọn ra TopN những $v(r_i, p_j)$ chưa xác định để khuyến nghị cho r_i . (Không khuyến nghị lại các bài báo p_j mà r_i đã biết)

Cuối lặp

Độ phức tạp tính toán của CF-kNN: $\mathcal{O}(mn)$ (với m: số lượng nghiên cứu viên, n: số lượng bài báo) [98].

Mặc dù được đánh giá là tiếp cận thành công trong việc phát triển các phương pháp, hệ thống khuyến nghị, nhưng các phương pháp CF cũng có những hạn chế của nó. (Adomavicius and Tuzhilin [5], Bobadilla et al. [22]), đã chỉ ra những hạn chế của các phương pháp CF như sau:

- **Ma trận đánh giá thừa:** ảnh hưởng nhiều đến việc phân tích ma trận để tiên đoán những giá trị đánh giá chưa xác định trong ma trận.
- **Đối tượng khuyến nghị mới:** không thể thực hiện khuyến nghị cho người dùng những đối tượng khuyến nghị mới. Tức đối tượng khuyến nghị chưa được ai quan tâm đánh giá, mặc dù có thể đối tượng mới đó rất gần với sở thích của người dùng.
- **Người dùng mới:** không thể khuyến nghị cho những người dùng mới chưa có thông tin quan sát trong ma trận đánh giá.

Việc áp dụng tiếp cận CF cho bài toán khuyến nghị bài báo khoa học liên quan đã gặp phải những hạn chế đã đề cập, đặc biệt ma trận đánh giá thể hiện sự quan tâm của các nghiên cứu viên với các đối tượng khuyến nghị bài báo khoa học là một ma trận rất thừa. Như vậy, mặc dù rất tiềm năng nhưng tiếp cận CF không phải là tiếp cận phù hợp cho bài toán khuyến nghị bài báo khoa học liên quan cho các nghiên cứu viên.

4.5.3 Kết hợp tuyến tính CB và CF

Hình thức kết hợp đơn giản nhất là kết hợp tuyến tính kết quả của CB-Recent và CF-kNN.

$$Sim_{Hybrid}(r_i, p_j) = \alpha * Sim_{CB}(r_i, p_j) + (1 - \alpha) * v(r_i, p_j) \quad (4.9)$$

$$\forall r_i \in R, p_j \in P$$

4.6 Các phương pháp đề xuất

4.6.1 Kết hợp Xu hướng nghiên cứu và quan hệ lòng tin

Việc chọn một bài báo để tham khảo, bên cạnh yếu tố nội dung bài báo có liên quan, các nghiên cứu viên còn quan tâm đến uy tín của những tác giả của bài báo đó. Hay

nói cách khác, nghiên cứu viên đang đặt lòng tin vào một số nghiên cứu viên, chuyên gia uy tín khác trong lĩnh vực. Đây là những khiếm khuyết của các phương pháp phổ biến hiện nay. Ở đây, chúng tôi đề xuất kết hợp khai thác nội dung bài báo với các quan hệ lòng tin của nghiên cứu viên để phát triển các phương pháp mới cho khuyến nghị bài báo khoa học tiềm năng cho nghiên cứu viên.

4.6.1.1 Lòng tin dựa trên quan hệ đồng tác giả và quan hệ trích dẫn (CB-TrendTrust1)

Giả sử rằng, lòng tin của một nghiên cứu viên đối với một bài báo phụ thuộc vào mức độ lòng tin của chính nghiên cứu viên đó kết hợp với lòng tin của những đồng tác giả của họ đối với việc trích dẫn các tác giả của bài báo đang xem xét. Chi tiết phương pháp có thể tóm tắt qua các bước như sau:

Phương pháp: CB-TrendTrust1

Đầu vào:

- Tập các nghiên cứu viên $R = \{r\}$
- Không gian các bài báo quan sát được $P = \{p\}$

Đầu ra: $\forall r \in R$, trả về P_{TopN} dựa trên giá trị hữu ích tiên đoán để khuyến nghị cho r .

1: **Bước 1:** Xây dựng mạng $CiNet_Author(R, E_2)$ gồm 2 thành phần chính là R, E_2 .

- R : Tập các đỉnh, mỗi đỉnh là một nghiên cứu viên.
- E_2 : Tập các cạnh (cặp đỉnh) có hướng thể hiện quan hệ trích dẫn, hướng từ $x \rightarrow y$ thể hiện quan hệ x đã trích dẫn y , hay x đặt lòng tin lên y , khi trích dẫn y . Trọng số của cạnh có thể lượng hóa như sau:

$$w_{cite}(r_i, r_j, t_0) = \frac{\sum_{t_i=t_0}^{t_c} NumCitation(r_i, r_j, t_i)}{e^{\gamma*(t_c-t_i)} * TotalCitation(r_i, t_0)} \quad (4.10)$$

Trong đó,

- $NumCitation(r_i, r_j, t_i)$: số lần mà r_i đã trích dẫn r_j trong năm t_i .
 - $TotalCitation(r_i, t_0)$: Tổng số trích dẫn của r_i tính từ thời điểm t_0 đến thời điểm hiện tại.
 - t_c : năm hiện tại.
 - t_0 : thời điểm bắt đầu xem xét yếu tố xu hướng.
 - γ : hệ số xu hướng. Trường hợp đơn giản $\gamma = 1$.
-

2: **Bước 2:** Xây dựng mạng đồng tác giả $CoNet(R, E_1)$.

- R : Tập các đỉnh, mỗi đỉnh là một nghiên cứu viên.
- E_1 : tập các cặp đỉnh có hướng thể hiện quan hệ đồng tác giả, hướng từ $x \rightarrow y$ thể hiện quan hệ x đồng tác giả với y.

3: **Bước 3:** Kết hợp quan hệ trích dẫn của nghiên cứu viên r_i với quan hệ trích dẫn của các đồng tác giả của r_i để lượng hóa quan hệ lòng tin giữa 2 nghiên cứu viên là r_i và r_j tính từ thời điểm t_0 , $w_{trust}(r_i, r_j, t_0)$. $w_{trust}(r_i, r_j, t_0) = w_{cite}(r_i, r_j, t_0) +$

$$+ \frac{\sum_{r_u \in CoAuthor(r_i)} w_{coauthor}(r_i, r_u, t_0) * w_{cite}(r_u, r_j, t_0)}{|CoAuthor(r_i)|} \quad (4.11)$$

4: **Bước 4:** Lượng hóa mức độ tin tưởng của một nghiên cứu viên r_i với bài báo p_j .

$$w_{trust}(r_i, p_j, t_0) = MAX(w_{trust}(r_i, r_j, t_0)) \quad (4.12)$$

(Với $r_j \in R_{p_j}$: tập các tác giả của bài báo p_j)

5: **Bước 5:** Kết hợp trọng số lòng tin với độ tương tự quan tâm nghiên cứu gần đây của nghiên cứu viên. $\forall r_i \in R, p_j \in P : RatingValue(r_i, p_j) = 0$,

$$RatingValue(r_i, p_j, t_0) = \alpha * w_{trust}(r_i, p_j, t_0) + (1 - \alpha) * Sim_{CB}(r_i, p_j) \quad (4.13)$$

6: **Bước 6:** Khuyến nghị. $\forall r_i \in R$,

- Chọn TopN bài báo có $RatingValue(r_i, p_j)$ cao nhất khuyến nghị cho r_i .
-

Phân tích độ phức tạp tính toán của phương pháp CB-TrendTrust1:

- Các bước 1 và 2 xem như bước tiền xử lý.
- Xét bước 3: Cần lượng hóa lòng tin cho mọi cặp nghiên cứu viên. Ngoài ra, với mỗi nghiên cứu viên $r \in R$, cần xét $|CoAuthor(r)|$ đồng tác giả. Gọi số đồng tác giả trung bình của một nghiên cứu viên là k , thì độ phức tạp ở bước này là $\mathcal{O}(|R|^2k)$.
- Xét bước 4: Cần lượng hóa lòng tin giữa mỗi nghiên cứu viên với mọi bài báo. Với mỗi bài báo $p \in P$, cần xét $|Authors(p)|$ tác giả. Gọi số tác giả trung bình của một bài báo p là l . Khi đó, độ phức tạp tính toán ở bước này là $\mathcal{O}(|R||P|l)$.
- Xét bước 5: Cần xem xét mọi bài báo để quyết định có khuyến nghị cho nghiên cứu viên hay không. Vậy độ phức tạp tính toán ở bước này là $\mathcal{O}(|R||P|)$.

- Tóm lại, do $|P| >> |R|$, nên tóm lại, độ phức tạp tính toán của phương pháp này là $\mathcal{O}(|R||P|l)$ (Với $|R|$: số lượng nghiên cứu viên, $|P|$: số lượng bài báo, l : số tác giả trung bình của một bài báo).

4.6.1.2 Lòng tin dựa trên quan hệ trích dẫn tiềm ẩn (CB-TrendTrust2)

Trên thực tế, một nghiên cứu viên thường sẽ lần theo các bài báo trong mục tham khảo của các bài báo mà họ quan tâm để tìm kiếm các bài báo tiềm năng liên quan. Hành động đó thể hiện một quan hệ trích dẫn tiềm ẩn của các nghiên cứu viên đối với các bài báo liên quan dựa trên việc bắc cầu quan hệ trích dẫn. Nếu xét ở góc độ lòng tin, có thể nói, nghiên cứu viên có thể đặt lòng tin vào những nghiên cứu viên khác dựa trên việc bắc cầu quan hệ lòng tin. Chi tiết của phương pháp khai thác quan hệ lòng tin dựa trên quan hệ trích dẫn tiềm ẩn có thể tóm tắt như sau:

Phương pháp: CB-TrendTrust2

Đầu vào:

- Tập các nghiên cứu viên $R = \{r\}$
- Không gian các bài báo quan sát được $P = \{p\}$

Đầu ra: $\forall r \in R$, trả về P_{TopN} dựa trên giá trị hữu ích tiên đoán để khuyến nghị cho r .

1: **Bước 1:** Tương tự phương pháp CB-TrendTrust1

2: **Bước 2:** Tổng hợp quan hệ trích dẫn của nghiên cứu viên r_i với quan hệ trích dẫn của các tác giả mà r_i đã trích dẫn để lượng hóa quan hệ lòng tin giữa 2 nghiên cứu viên là r_i và r_j tính từ thời điểm t_0 , $w_{trust}(r_i, r_j, t_0)$.

$$w_{trust}(r_i, r_j, t_0) = w_{cite}(r_i, r_j, t_0) +$$

$$+ \frac{\sum_{r_u \in CitedAuthor(r_i)} w_{cite}(r_i, r_u, t_0) * w_{cite}(r_u, r_j, t_0)}{|CitedAuthor(r_i)|} \quad (4.14)$$

3: **Bước 3:** Áp dụng tương tự các bước 4, 5, 6 của phương pháp CBTrendTrust2.

Độ phức tạp tính toán của CB-TrendTrust2: tương tự với phương pháp CBTrendTrust1, phương pháp này cũng có độ phức tạp tính toán là $\mathcal{O}(|R||P|l)$ (Với $|R|$: số lượng nghiên cứu viên, $|P|$: số lượng bài báo, l : số tác giả trung bình của một bài báo).

4.7 Thực nghiệm, đánh giá

Phần này trình bày kết quả đánh giá, so sánh các phương pháp khác nhau cho khuyến nghị bài báo khoa học liên quan cho nghiên cứu viên trên tập dữ liệu lớn thu thập từ trang web Microsoft Academic Search.

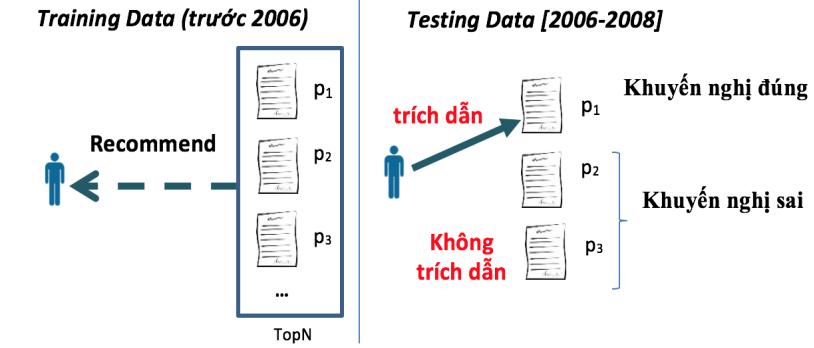
Thực nghiệm được tiến hành chạy trên hệ thống máy chủ UIT-Cloud, trên máy ảo có cấu hình như sau:

- Số CPU: 32 processors, Intel (R) Xeon(R) CPU E5-2690 0 @ 2.9GHz
- Bộ nhớ: 80.0 GB Hệ điều hành: Windows 7, 64 bits.

4.7.1 Tập dữ liệu và thiết lập thực nghiệm

Joeran Beel et al., 2013, đã chỉ ra rằng: đến bây giờ vẫn chưa có sự thống nhất về các tập dữ liệu cũng như phương pháp đánh giá khi thực hiện so sánh các phương pháp khác nhau cho khuyến nghị bài báo khoa học (Beel et al. [15]). Trong nghiên cứu này, chúng tôi đã thu thập thông tin các bài báo khoa học từ trang Microsoft Academic Search để xây dựng tập dữ liệu thực nghiệm. Để cùng góp phần với cộng đồng trong việc đa dạng, và dần chuẩn hóa các tập dữ liệu thực nghiệm cho bài toán này, chúng tôi đã phổ biến tập dữ liệu tại sites.google.com/site/tinhuynhuit/dataset.

Trong thực nghiệm, chọn ngẫu nhiên 1000 nghiên cứu viên có bài báo công bố trước 2006 và sau 2006 như dữ liệu đầu vào. Các bài báo của họ công bố trước năm 2006 (xem như dữ liệu quá khứ) được chọn làm dữ liệu huấn luyện. Các bài báo được 1000 nghiên cứu viên trích dẫn từ 2006 đến 2008 xem như dữ liệu trong tương lai làm Ground-Truth để kiểm chứng chất lượng các phương pháp khuyến nghị. Tức là, nếu phương pháp khuyến nghị một bài báo tiềm năng cho nghiên cứu viên, mà trong tương lai nghiên cứu viên có trích dẫn bài báo này thì xem như đó là một khuyến nghị đúng, ngược lại là sai (hình 4.1). Ground-Truth bao gồm 52.254 bài được 1000 nghiên cứu viên này trích dẫn trong năm từ 2006 đến 2008. Cách chia trực thời gian thành dữ liệu quá khứ và dữ liệu tương lai, sau đó dùng dữ liệu tương lai làm Ground-Truth để đánh giá chất lượng phương pháp khuyến nghị được áp dụng phổ biến trong những nghiên cứu hiện nay như: Tang et al. [117], Sugiyama and Kan [111, 113], Sun et al. [114].



Hình 4.1: Minh họa cách tính độ chính xác khuyến nghị bài báo

4.7.2 Độ đo đánh giá độ chính xác khuyến nghị

Thông thường, TopN những đối tượng tiềm năng trả về từ hệ thống sẽ được dùng để đánh giá độ chính xác của phương pháp khuyến nghị. Hầu hết các độ đo đánh giá được dùng phổ biến trong các nghiên cứu hiện nay đều có nguồn gốc từ lĩnh vực truy vấn thông tin (IR). Tương tự các nghiên cứu của [111, 112, 113], ở đây chúng tôi tập trung phân tích kết quả thực nghiệm với độ đo NDCG (Järvelin and Kekäläinen [58]) và MRR (Voorhees [123]).

4.7.2.1 Độ đo NDCG (Normalized Discounted Cumulative Gain)

DCG là một độ đo liên quan đến chất lượng xếp hạng. DCG đo lường tính hữu ích của đối tượng dựa trên vị trí của nó trong danh sách xếp hạng trả về. Tính hữu ích sẽ được tích lũy từ đầu cho đến cuối danh sách xếp hạng trả về. Và giá trị trung bình của DCG (tức NDCG) qua tất cả các người dùng sẽ được dùng để thể hiện độ chính xác khuyến nghị.

Ở đây chúng ta chỉ quan tâm TopN những kết quả trả về là có liên quan hay không liên quan. Vì vậy, NDCG@TopN được dùng để đánh giá. Với TopN là số lượng các bài báo trong danh sách xếp hạng được khuyến nghị cho các nghiên cứu viên.

$$DCG(i) = \begin{cases} G(1), \text{ nếu } i = 1 \\ DCG(i - 1) + \frac{G(i)}{\log(i)}, \text{ với } i \neq 1 \end{cases} \quad (4.15)$$

Trong đó, i là vị trí xếp hạng thứ i . Ở đây $G(i) = 1$, nếu kết quả khuyến nghị là

liên quan (đúng), ngược lại $G(i) = 0$.

4.7.2.2 Độ đo MRR (Mean Reciprocal Rank)

Reciprocal Rank (RR) là một độ đo xem xét vị trí xếp hạng của đối tượng liên quan đầu tiên được trả về. MRR là trung bình của RR thông qua nhiều truy vấn khác nhau. Hay trong bài toán của chúng ta MRR là trung bình kết quả khuyến nghị xét qua nhiều nghiên cứu viên.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^Q \frac{1}{Rank_i} \quad (4.16)$$

Trong đó,

- $|Q|$: Tổng số nghiên cứu viên được thực hiện khuyến nghị
- $Rank_i$: vị trí xuất hiện đầu tiên của bài báo khuyến nghị liên quan trong danh sách xếp hạng trả về.

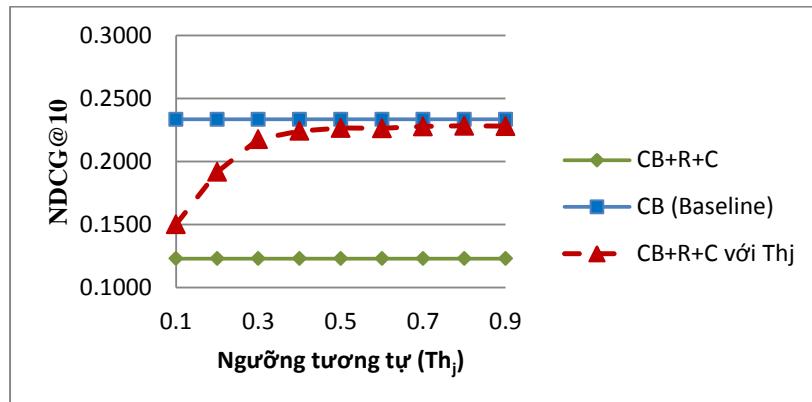
4.7.3 Kết quả thực nghiệm

Phần này trình bày kết quả thực nghiệm so sánh, phân tích các phương pháp phổ biến, cũng như các phương pháp đề xuất bao gồm: CB, CB+R+C, CF-kNN, CB-Recent, CB-Recent+CF, CB-TrendTrust1, CB-TrendTrust2.

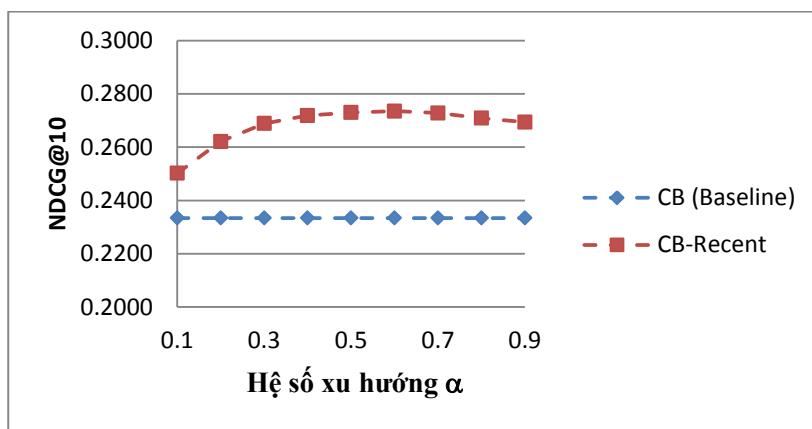
Với phương pháp CB+R+C, để quyết định chọn bài báo tham khảo (R), trích dẫn (C) kết hợp với bài báo công bố dựa trên ngưỡng tương tự (Th_j), chúng tôi cũng đã tiến hành thay đổi Th_j , Th_j nhận các giá trị rời rạc 0.1, 0.2, ..., 0.9. Kết quả tốt nhất đạt được tại $Th_j = 0.8$, với NDCG@10 = 0.2282, vẫn thấp hơn so với phương pháp cơ sở CB có NDCG@10=0.2334 (Hình 4.2).

Phương pháp mô hình hóa quan tâm nghiên cứu của nghiên cứu viên dựa trên thời gian của nhóm Sugiyama và đồng nghiệp (CB-Recent) đã cho thấy ưu điểm vượt trội, cải tiến đáng kể độ chính xác khuyến nghị trong thực nghiệm so sánh. Trong thực nghiệm với CB-Recent, chúng tôi thay đổi hệ số xu hướng α , nhận các giá trị 0.1, 0.2, ..., 0.9. Kết quả tốt nhất đạt được tại $\alpha = 0.6$, với NDCG@10 = 0.2735, cao hơn hẳn so với phương pháp cơ sở CB-Baseline với NDCG@10 = 0.2334 (Hình 4.3).

Đối với phương pháp lọc cộng tác CF-kNN, ở bước gom cụm các nghiên cứu viên đồng sở thích với kNN, để chọn được giá trị k tốt nhất, chúng tôi đã tiến hành thay đổi



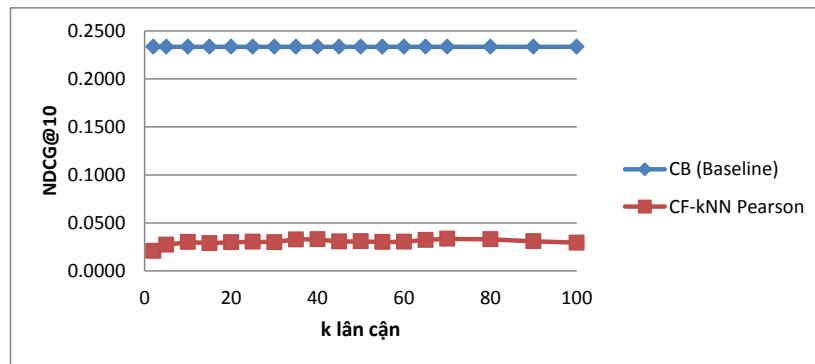
Hình 4.2: Kết quả thực nghiệm phương pháp CB+R+C với tham số ngưỡng tương tự Th_j



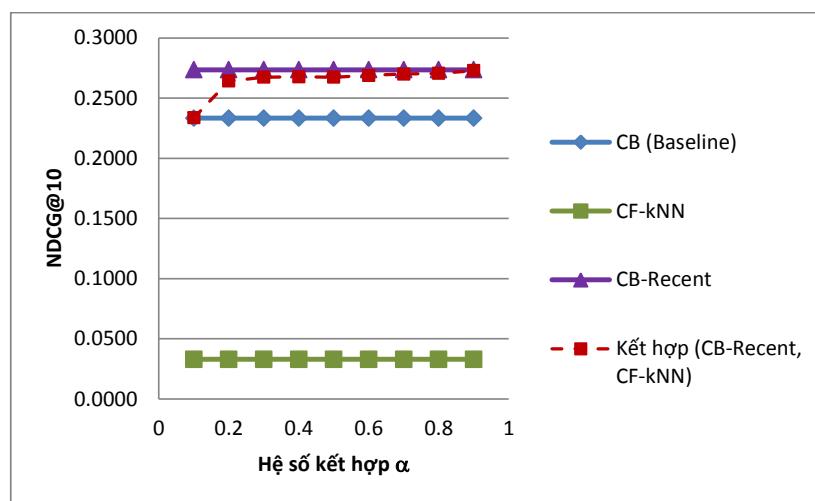
Hình 4.3: Kết quả thực nghiệm phương pháp CB-Recent với các hệ số xu hướng alpha khác nhau

k với các giá trị khác nhau từ 3 đến 100. Với mỗi giá trị k, chúng tôi xem xét độ chính xác khuyến nghị với độ đo NDCG@5, NDCG@10, MRR. Hình 4.4 là kết quả áp dụng phương pháp CF với các hệ số k khác nhau so với phương pháp nội dung CB-Baseline. Kết quả thực nghiệm cho thấy tiếp cận lọc cộng tác CF không phải là tiếp cận phù hợp cho bài toán này. Ma trận trích dẫn (dòng tập các nghiên cứu viên và cột là tập bài báo) quá thưa đã ảnh hưởng lớn đến độ chính xác của phương pháp lọc cộng tác. Việc kết hợp tuyến tính CB-Recent (tốt nhất trong các phương pháp CB) và CF-kNN cho kết quả trong hình 4.5.

Sau khi lượng hóa quan hệ lòng tin thông qua việc tổng hợp lòng tin của những quan hệ đồng tác giả và những quan hệ trích dẫn, chúng tôi kết hợp tuyến tính với



Hình 4.4: Kết quả thực nghiệm phương pháp lọc cộng tác CF-kNN với các giá trị k khác nhau



Hình 4.5: Kết quả thực nghiệm phương pháp kết hợp tuyến tính CB-Recent và CF

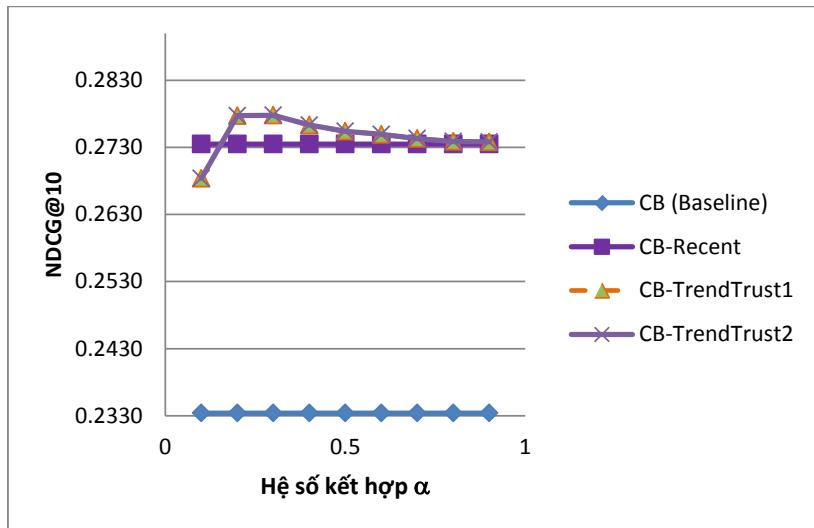
xu hướng sở thích (CB-Recent). Trong thực nghiệm chúng tôi thay đổi hệ số alpha để tìm giá trị tốt nhất cho kết hợp. Hình 4.6 trực quan kết quả kết hợp. Phương pháp CB-TrendTrust2 cho kết quả trội hơn so với CB-TrendTrust1 và CB-Recent.

Tổng hợp so sánh đánh giá các phương pháp đề xuất và phương pháp phổ biến hiện nay trình bày trong bảng 4.1.

4.7.4 Kết luận

Thông qua kết quả thực nghiệm trên tập dữ liệu tương đối lớn, chúng ta có thể đưa ra các nhận định đối với các phương pháp khuyến nghị bài báo liên quan như sau:

- Tiếp cận lọc cộng tác CF cho thấy không phải là tiếp cận phù hợp cho bài toán



Hình 4.6: Phương pháp kết hợp xu hướng sở thích và quan hệ lòng tin

Bảng 4.1: Tóm tắt so sánh, đánh giá các phương pháp đề xuất và các phương pháp phổ biến hiện nay

Phương pháp Khuyến nghị	NDCG@5	NDCG@10	MRR
CB-Baseline	0.2945	0.2334	0.5128
CB+R+C	0.1464	0.1230	0.3061
CB+R+C, $Th_j = 0.8$	0.2877	0.2282	0.4985
CB-Recent, $\alpha = 0.6$	0.3577	0.2735	0.6142
CF-kNN	0.0357	0.0330	0.0934
CB-Recent+CF, $\alpha = 0.9$	0.3570	0.2728	0.6140
CB-TrendTrust1	0.3610	0.2778	0.6164
CB-TrendTrust2	0.3610	0.2778	0.6164

này, trong khi tiếp cận nội dung là tiếp cận phù hợp nhất mà các nghiên cứu hiện nay đang thực hiện.

- Yếu tố xu hướng (sở thích gần đây của nghiên cứu viên) đã cải tiến đáng kể độ chính xác khuyến nghị.
- Tiếp cận kết hợp nội dung với các mối quan hệ xã hội tiềm ẩn, cụ thể ở đây quan hệ lòng tin đã góp phần cải tiến độ chính xác khuyến nghị bài báo liên quan cho nghiên cứu viên, nhưng không đáng kể.

4.8 Kết chương

Mục đích của chương này là phát biểu bài toán khuyễn nghị bài báo khoa học. Trong phạm vi thực hiện, luận án đã nghiên cứu, đánh giá các phương pháp khuyễn nghị phổ biến, đồng thời đề xuất phương pháp mới dựa trên tiếp cận khai thác thông tin xu hướng, các mối quan hệ xã hội học thuật được khai thác từ mạng xã hội học thuật ASN. Các phương pháp đề xuất dựa trên việc kết hợp xu hướng sở thích nghiên cứu và quan hệ lòng tin của nghiên cứu viên. Thực nghiệm tiến hành trên tập dữ liệu khoa học rút trích từ hệ thống Microsoft Academic Search. Độ chính xác khuyễn nghị được đánh giá dựa trên dữ liệu gán nhãn là các bài báo mà nghiên cứu viên tham khảo, trích dẫn trong tương lai, nhưng chưa trích dẫn trong quá khứ. Kết quả thực nghiệm cho thấy phương pháp đề xuất đã cải tiến độ chính xác khuyễn nghị so với phương pháp cơ sở và các phương pháp phổ biến nhất hiện nay ([4.1](#)).

Mặc dù việc kết hợp quan hệ lòng tin với xu hướng sở thích của nghiên cứu viên đã cải tiến độ chính xác khuyễn nghị. Tuy nhiên, bên cạnh nội dung, lòng tin vào chuyên gia, đôi khi nghiên cứu viên còn quan tâm đến tính mới, chất lượng bài báo. Một vấn đề nữa là phương pháp kết hợp các yếu tố này như thế nào cho hiệu quả cũng cần phải nghiên cứu. Vấn đề này hiện đang được nghiên cứu sinh tiếp tục nghiên cứu giải quyết.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Các kết quả đạt được

Nhằm hỗ trợ các nghiên cứu viên dễ dàng hơn trong việc tìm kiếm, khai thác các thông tin học thuật, luận án đã tập trung nghiên cứu và phát triển các phương pháp khuyến nghị cho hai bài toán con chính: (1) Khuyến nghị cộng tác; (2) Khuyến nghị bài báo khoa học. Dựa trên quan điểm quan điểm, nghiên cứu viên thường hỏi, tham khảo ý kiến của đồng nghiệp, thầy cô khi tìm một bài báo để đọc, người cộng tác. Hay nói cách khác đồng nghiệp, thầy, cô là những người sẽ tư vấn giới thiệu các tài liệu hữu ích, những người cộng tác tiềm năng mà phù hợp với quan tâm nghiên cứu của nghiên cứu viên. Thực chất đó chính là việc yêu cầu khuyến nghị từ những người có quan hệ. Dựa trên quan điểm đó, luận án đã tiếp cận khai thác các mối quan hệ xã hội học thuật để phát triển các phương pháp cho khuyến nghị cộng tác và khuyến nghị bài báo khoa học.

Sau quá trình nghiên cứu, luận án đã thu thập được một số kết quả có ý nghĩa khoa học như sau:

1. Khảo sát, phân tích, đánh giá các phương pháp khuyến nghị nói chung và khuyến nghị trong lĩnh vực học thuật nói riêng.
2. Đề xuất mô hình hóa các mạng xã hội học thuật nhận diện được từ kho dữ liệu học thuật, mô hình ASN [CT.6].
3. Bài toán khuyến nghị cộng tác cho nghiên cứu viên

-
- Đối với nghiên cứu viên có quan hệ đồng tác giả: đề xuất, cải tiến các phương pháp phân tích xu hướng cộng tác trong mạng xã hội học thuật ASN để khuyến nghị các cộng tác viên tiềm năng. Các phương pháp đề xuất bao gồm: MPRS, MPRS+, RSS+ [CT.4, CT.1].
 - Đối với nghiên cứu viên chưa có quan hệ đồng tác giả: đề xuất tập đặc trưng để khuyến nghị những mối quan hệ cộng tác tốt, chất lượng [CT.3].
 - Đề xuất phương pháp đánh giá chất lượng cộng tác được khuyến nghị [CT.3].
4. Bài toán khuyến nghị bài báo khoa học: phát triển phương pháp khuyến nghị bài báo khoa học cho nghiên cứu viên dựa trên việc khai thác mạng trích dẫn, quan hệ lòng tin trong mô hình ASN [CT.8].
 5. Xây dựng kho dữ liệu học thuật hơn 6 triệu bài báo và hệ thống tìm kiếm thông tin khoa học CSPubGuru (www.cspubguru.com) [CT.5, CT.7, CT.9, CT.14].

Giá trị thực tiễn của luận án:

- Ứng dụng các phương pháp khai thác mạng xã hội học thuật vào các bài toán khuyến nghị trong lĩnh vực học thuật, hỗ trợ cộng đồng làm nghiên cứu khoa học. Một số bài toán ứng dụng đã được thử nghiệm như: khuyến nghị cộng tác, khuyến nghị bài báo liên quan.
- Kết quả nghiên cứu của luận án về hệ khuyến nghị có thể áp dụng cho nhiều lĩnh vực ứng dụng khác nhau.

Việc nghiên cứu phát triển các phương pháp, hệ thống khuyến nghị, giải pháp thông minh giúp người dùng dễ dàng hơn trong việc tìm kiếm thông tin liên quan là một vấn đề lớn, còn khá sơ khai và nhiều thách thức đang thu hút nhiều nghiên cứu của cộng đồng khoa học trên khắp thế giới. Những kết quả đạt được bước đầu giúp nghiên cứu sinh có được nền tảng tri thức để bước vào lĩnh vực nghiên cứu nhiều thách thức. Phần tiếp theo là một số vấn đề có thể tiếp tục thực hiện cho hướng phát triển của luận án.

Hướng phát triển

Tiếp cận phân tích mạng xã hội đã cho thấy những ưu điểm, tiềm năng trong việc nghiên cứu, phân tích hành vi con người nhằm phát triển các phương pháp, tìm kiếm, khuyến nghị thông minh. Thực tế nghiên cứu cho thấy, tiếp cận phân tích mạng xã hội có thể giải quyết được một số hạn chế tồn đọng chính của những tiếp cận truyền thống như: độ chính xác tiên đoán, ma trận đánh giá thưa. Tuy nhiên, các mối quan hệ xã hội, thông tin từ các trang mạng xã hội, cũng như việc nghiên cứu phát triển các phương pháp khuyến nghị còn rất nhiều vấn đề cần đầu tư nghiên cứu. Một số vấn đề chính trong hướng phát triển của luận án có thể kể đến như sau:

- Nghiên cứu tính đa dạng của những mối quan hệ và khả năng ảnh hưởng của chúng.
- Nghiên cứu các phương pháp mô hình hóa hành vi con người, cũng như xem xét mức độ ảnh hưởng của nó trong việc ra quyết định.
- Tiếp tục nghiên cứu phát triển các phương pháp lai nhằm giải quyết một số hạn chế mà tiếp cận hiện nay đang gặp phải như: ma trận thưa, khởi động lạnh, v.v...
- Độ chính xác của các phương pháp khuyến nghị vẫn còn là một thách thức lớn. Cần cải tiến độ chính xác của các phương pháp khuyến nghị, phát triển các phương pháp khuyến nghị mới kết hợp thông tin từ nhiều nguồn: cá nhân, lịch sử, quan hệ xã hội, ngữ cảnh, thời gian, vị trí và các thông tin khác từ “Internet of Things”, chứ không đơn thuần là lịch sử và quan hệ xã hội.
- Mở rộng bài toán khuyến nghị cộng tác cho nhóm, viện, quốc gia.
- Phát triển các phương pháp cho khuyến nghị ý tưởng dựa trên khai thác dữ liệu khoa học.
- Mở rộng các phương pháp khuyến nghị trong luận án ra nhiều lĩnh vực khác.

CÁC CÔNG TRÌNH ĐÃ CÔNG

BỘ CỦA TÁC GIẢ

Tạp chí chuyên ngành

- [CT.1] Tin Huynh, Kiem Hoang. New Methods for Calculating Trend- Based Vertex Similarity for Collaboration Recommendation. Journal of Computer Science and Cybernetics, vol.29, No.4, pages 338-350, (2013).
- [CT.2] Huỳnh Ngọc Tín, Hoàng Kiếm. Khai thác xu hướng sở thích và quan hệ lòng tin để phát triển phương pháp khuyến nghị bài báo khoa học. Tạp chí Công nghệ thông tin và Truyền thông, Tập V-1, Số 13 (33), (2015).

Hội thảo chuyên ngành

- [CT.3] Tin Huynh, Atsuhiro Takasu, Tomonari Masada, Kiem Hoang. Collaborator Recommendation for Isolated Researchers. The Seventh International Symposium on Mining and Web (MAW2014) as a part of The 28th IEEE International Conference on Advanced Information Networking and Applications (AINA-2014), May 13-16, 2014, Victoria, Canada (2014). (Proceedings indexed by DBLP, EI, Scopus, and Thomson ISI. ERA Conference Ranking of AINA: B)
- [CT.4] Tin Huynh, Kiem Hoang, Dao Lam. Trend Based Vertex Similarity for Academic Collaboration Recommendation. 5th International Conference on Computational Collective Intelligence Technologies and Applications (ICCCI 2013), September 2013, Craiova, Romani, pages 11-20, (2013). (Proceedings Indexed by DBLP, EI, Scopus, ACM Digital Library, and Thomson ISI.ERA Conference Ranking: C)

-
- [CT.5] Tin Huynh, Kiem Hoang, Tien Do, Duc Huynh. Vietnamese Author Name Disambiguation for Integrating Publications from Heterogeneous Sources. The 5th Asian conference on Intelligent Information and Database Systems (ACIIDS 2013), Kuala Lumpur, Malaysia, pages 226-235, (2013). (Proceedings indexed by DBLP, EI, Scopus, and Thomson ISI)
- [CT.6] Tin Huynh, Kiem Hoang. Modeling Collaborative Knowledge of Publishing Activities for Research Recommendation. In Proceedings of the 4th International Conference on Computational Collective Intelligent Technologies and Applications (ICCCI 2012), November 2012, Ho Chi Minh City, VietNam, pages 28-30, (2012). (The proceedings indexed by DBLP, EI, Scopus, ACM Digital Library, and Thomson ISI. ERA Conference Ranking: C. Citation Count: 4 (không tính tự trích dẫn))
- [CT.7] Tin Huynh, Hiep Luong, and Kiem Hoang. Integrating bibliographical data of computer science publications from online digital libraries. In Proceedings of the 4th Asian conference on Intelligent Information and Database Systems (ACIIDS'12), Springer-Verlag, Berlin, Heidelberg, pages 226-235, (2012). (The proceedings indexed by DBLP, EI, and Thomson ISI. Citation count: 1 (không tính tự trích dẫn))
- [CT.8] Tin Huynh, Hiep Luong, Kiem Hoang, Susan Gauch, Loc Do, Huong Tran. Scientific Publication Recommendations Based on Collaborative Citation Networks. In: Proceedings of the 3rd International Workshop on Adaptive Collaboration (AC 2012) as part of The 2012 International Conference on Collaboration Technologies and Systems (CTS 2012). Denver, Colorado, USA, pages 316-321, (2012). (The proceedings indexed by DBLP, EI, and Thomson ISI. ERA Conference Ranking: C. Citation count: 4 (không tính tự trích dẫn))
- [CT.9] Tin Huynh, Kiem Hoang. GATE framework based metadata extraction from scientific papers. In: Proceedings of the The International Conference on Education and Management Technology (ICEMT 2010), Cairo, Egypt, page 188 – 191, (2010). (The proceedings indexed by Google Scholar, IEEE Xplore Digital library,

Citation count: 4, (không tính tự trích dẫn))

- [CT.10] Hung Nghiep Tran, Tin Huynh, Tien Do. Author Name Disambiguation by Using Deep Neural Network. In Proceedings of the 6th Asian conference on Intelligent Information and Database Systems, Bangkok, Thailand, April 2014 (ACIIDS'14). Springer-Verlag, Berlin, Heidelberg, pages 123-132, (2014). (The proceedings indexed by DBLP, EI, and Thomson ISI. Citation Count: 1 (không tính tự trích dẫn))
- [CT.11] Hung Nghiep Tran, Tin Huynh, Kiem Hoang. A Potential Approach to Overcome Data Limitation in Scientific Publication Recommendation. In Proceedings of the seventh international conference on knowledge and systems engineering (KSE-2015), TpHCM, Vietnam, Oct 8-10, 2015.
- [CT.12] Hiep Luong, Tin Huynh, Susan Gauch, Loc Do, and Kiem Hoang. Publication venue recommendation using author network's publication history. In Proceedings of the 4th Asian conference on Intelligent Information and Database Systems, Kaohsiung, Taiwan, March 2012 (ACIIDS'12). Springer-Verlag, Berlin, Heidelberg, pages 426-435, (2012). (The proceedings indexed by DBLP, EI, and Thomson ISI. Citation Count: 3 (không tính tự trích dẫn))
- [CT.13] Hiep Luong, Tin Huynh, Susan Gauch, Kiem Hoang. Exploiting Social Networks for Publication Venue Recommendations. In Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval, Barcelona, Spain, pages 239 - 245, October (2012).
- [CT.14] Tien Do, Dao Lam, Tin Huynh. A Framework for integrating bibliographical data of computer science publications. 2014 International Conference on Computing, Management and Telecommunications (ComManTel 2014), 27-29 April 2014, Da nang, Viet Nam, pages 245-250, (2014).

CÁC ĐỀ TÀI KHOA HỌC

CHỦ TRÌ THỰC HIỆN

Bảng 4.2: Đề tài khoa học đã và đang thực hiện

Tên đề tài	Năm thực hiện	Cấp quản lý	Vai trò tham gia	Kết quả nghiệm thu
Truy vấn thông minh dựa trên rút trích tự động và mô hình cấu trúc metadata văn bản	2011	DHQG không trọng điểm	Chủ trì	Tốt
Tích hợp dữ liệu biên mục các bài báo khoa học máy tính từ nhiều nguồn không đồng nhất	2013	Cơ sở	Chủ trì	Khá
Tiếp cận phân tích mạng xã hội để phát triển hệ khuyến nghị thông tin trong lĩnh vực học thuật	2014	DHQG không trọng điểm	Chủ trì	Đã báo cáo giữa kỳ
Tư vấn việc làm dựa trên phân cụm người dùng	2015	Cơ sở	Chủ trì	Đang thực hiện

Phụ lục A. Luật rút trích thông tin bài báo khoa học

A.1 Các luật JAPE để rút trích thông tin từ bài báo khoa học PDF

Luật JAPE để xác định tên tác giả

```
Rule: Author
(
  (
    {Person} |
    (
      {Token.string != ",",
       Token.string != "and",
       Token.kind != "number"})
    )+
  ): author
)
-->
: author.Author = {rule= "Author"}
```

Luật JAPE để xác định các ‘Email’

```

Rule : LineEmailAnnotation
( ( {Token . string == "{"}
( {Token}
( {SpaceToken . kind == "space" }) ?
) +
( {SpaceToken . kind == "control" }) ?
)?
( {Token}
( {SpaceToken . kind == "space" }) ?
) +
( {Token . string == "@"} |
{Address . kind == "email"} |
{Token . string == "}"})
)
( {SpaceToken . kind == "space" }) ?
(
{Token}
( {SpaceToken . kind == "space" }) ?
) +
): lineEmailAnnotation
—>
: lineEmailAnnotation . LineEmailAnnotation =
{rule = "LineEmailAnnotation"}

```

Luật JAPE để xác định cơ quan công tác ‘Affiliation’

```

Rule : LineAffiliationAnnotation
((      {Token . string=="Dept"} | 
{Token . string=="dept"} | 
{Token . string=="University"} | 
{Token . string=="university"} | 
{Token . string=="Faculty"} | 
{Token . string=="FACULTY"} | 
{Lookup . majorType=="location"} | 
{Lookup . majorType=="org_key"} | 
{Lookup . majorType=="org_base"} | 
{Lookup . majorType=="cdg"} | 
{Lookup . majorType=="facility_key",
!Token . string=="Hall"} | 
((      {Token . kind=="number", Token . length>=3} 
{SpaceToken . kind=="space"} 
) | 
(      {Token . kind=="number"} 
({SpaceToken . kind=="space"})? 
({Token . kind== "punctuation",
Token . subkind == "dashpunct"}) 
({SpaceToken . kind=="space"})? 
{Token . kind=="number"} 
) 
)
)
({SpaceToken . kind=="space"})? 
(      {Token} 
({SpaceToken . kind=="space"})? 
)*
): lineAffiliationAnnotation
-->

```

```
:lineAffiliationAnnotation . LineAffiliationAnnotation =
{rule = "LineAffiliationAnnotation"}
```

Luật JAPE để xác định vùng ‘Abstract’

```
Rule: AbstractKeyword
(
    ({SpaceToken . kind=="control"})+
    (
        {Token . string=="Abstract ." } |
        {Token . string=="ABSTRACT ." } |
        {Token . string=="Abstract" } |
        {Token . string=="ABSTRACT" }
    )
    ({Token . string==".")?
):abstract_Keyword
--->
:abstract_Keyword . AbstractKeyword =
{rule = "AbstractKeyword"}
```

Luật JAPE để xác định, tách vùng ‘Reference’

```
Rule: ReferencesKeyword
(
    ({SpaceToken . kind=="control"})+
    (
        {Token . kind=="number" }
        ({Token . string==".")?
        ({SpaceToken . kind=="space" })+
    )?
    (
        {Token . string=="References" } |
        {Token . string=="REFERENCES" } |
    )
)
```

```

{Token.string=="reference"} |
{Token.string=="REFERENCE"} |
)
): referencesKeyword
--->
: referencesKeyword . ReferencesKeyword =
{rule= "ReferencesKeyword" }

```

A.2 Giải quyết nhập nhằng tên tác giả cho kho bài báo tích hợp

Bài toán nhập nhằng tên tác giả là vấn đề đã và đang thu hút nhiều quan tâm nghiêm cứu trong lĩnh vực thư viện số cũng như các hệ thống tìm kiếm tài liệu, chuyên gia. Nhập nhằng tên tác giả xảy ra khi: một tác giả sử dụng nhiều bút danh khác nhau (synonyms), hoặc nhiều tác giả khác nhau nhưng có cùng bút danh (polysems) trong các bài báo khoa học [42]. Thông thường những tác giả người Châu Á rất dễ bị nhầm lẫn tên, lý do chính là thứ tự trong cách viết họ tên. Bình thường thì họ viết họ trước, tên sau. Đôi khi tên đứng trước, họ đứng sau và tên lót (middle name) có lúc được dùng, có lúc không. Bảng A.1 trình bày một ví dụ minh họa về trường hợp nhập nhằng tên tác giả trong kho dữ liệu bài báo khoa học. Tên tác giả '**Tuan Nguyen**' trong bài báo số 1 và '**Anh-Tuan Nguyen**' trong bài báo số 3 là cùng một người (synonyms); Trong khi tác giả '**Tuan Nguyen**' trong bài số 1 và '**Tuan Nguyen**' trong bài số 2 là hai người khác nhau (polysems).

Phần bên dưới sẽ trình cách tiếp cận và đóng góp của luận án đối với đề giải quyết nhầm lẫn tên tác giả khi tích hợp dữ liệu khoa học từ nhiều nguồn không đồng nhất.

Bảng A.1: Ví dụ các bài báo nhập nhằng tên tác giả

STT	Bài báo
1	Multiagent Place-Based Virtual Communities for Pervasive Computing. Conference: PERCOM '08 Proceedings of the 2008 Sixth Annual IEEE International Conference on Pervasive Computing and Communications. Authors: Tuan Nguyen , Seng Loke; Torabi, T.; Hongen Lu. Dept. of Comput. Sci. & Comput. Eng., La Trobe Univ., Bundoora, VIC.
2	Stationary points of a kurtosis maximization algorithm for blind signal separation and antenna beamforming. Journal: Journal IEEE Transactions on Signal Processing Authors: Zhi Ding, Tuan Nguyen Dept. of Electr. & Comput. Eng., Iowa Univ., Iowa City, IA.
3	Semantic-PlaceBrowser: Understanding Place for Place-Scale Context-Aware Computing Conference: The Eighth International Conference on Pervasive Computing (Pervasive 2010), Helsinki, Finland, 2010. Authors: Authors: Anh-Tuan Nguyen , Seng Wai Loke, Torab Torabi, Hongen Lu. Department of Computer Science and Computer Engineering, La Trobe University, Victoria, 3086, Australia

A.2.1 Tiếp cận đề xuất

Làm thế nào để giải quyết nhập nhằng tên tác giả trong 2 bài báo khác nhau. [42] đã phân các phương pháp thành hai nhóm chính: (1) gom nhóm các bài báo của cùng một tác giả bằng các phương pháp tính toán tương tự (học tính synonyms hay polysems của hai bài báo); (2) Học và xây dựng mô hình hồ sơ cá nhân (profile) cho mỗi tác giả, và gán những bài báo tương ứng với tác giả nào đó dựa trên mô hình đã học. Cả hai cách tiếp cận này đều có gắng xác định tập đặc trưng dựa trên sự tương tự của các thuộc tính metadata rút trích từ các bài báo khoa học. Tập đặc trưng sẽ khác nhau tùy thuộc vào đặc tính của tập dữ liệu thực nghiệm. Với các dữ liệu thu thập từ nhiều nguồn không đồng nhất, luận án đã tiếp cận dùng phương pháp học giám sát và đề xuất tập đặc trưng để giải quyết vấn đề nhập nhằng tên giữa hai bài báo khác nhau.

A.2.1.1 Các phương pháp so khớp chuỗi phổ biến

Bilenko et al. [19], Cohen et al. [33], đã khảo sát các độ đo, phương pháp tính toán sự tương tự của hai chuỗi ký tự nhằm xác định sự trùng lặp. Các độ đo này cơ bản được chia thành 3 nhóm chính: (1) Khoảng cách biến đổi (Edit distance); (2) Tương tự dựa

trên từ (token); và (3) Độ đo kết hợp giữa từ và khoảng cách biến đổi.

Khoảng cách biến đổi (Edit Distance)

Khoảng cách biến đổi giữa chuỗi X và chuỗi Y là chi phí những thao tác thay đổi mà có thể chuyển chuỗi X thành chuỗi Y. Một số độ đo khoảng cách biến đổi phổ biến như Levenshtein, Monger-Elkan, Jaro, Jaro-Winkler [33]. Levenshtein được biết đến như độ đo khoảng cách biến đổi phổ biến nhất. Với độ đo Levenshtein, chi phí biến đổi chuỗi ký tự được tính bởi 3 loại thao tác: (1) Thao tác chèn một ký tự mới; (2) Thao tác xóa một ký tự; (3) Thao tác thay thế một ký tự bằng một ký tự khác.

$$Sim_{levenshtein}(X, Y) = 1 - \frac{d(X, Y)}{[max(length(X), length(Y))]} \quad (17)$$

Trong đó,

- $d(X, Y)$: Số thao tác biến đổi tối thiểu để chuyển chuỗi X thành chuỗi Y.
- $length(x)$: Chiều dài chuỗi X.

Tương tự dựa trên từ (token)

Trong một số trường hợp thì thứ tự của từ không thật sự quan trọng. Khi đó hệ số Jaccard và TF-IDF là những độ đo dựa trên từ hiệu quả và được dùng phổ biến [9], [33].

$$Sim_{jaccard}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (18)$$

- $|X \cap Y|$: Số lượng từ giống nhau giữa chuỗi X và chuỗi Y.
- $|X \cup Y|$: Số lượng từ phân biệt nhau đôi một trong cả chuỗi X và chuỗi Y.

Đo đố kết hợp

Một độ đo phổ biến cho sự kết hợp là độ đo Mogne-Elkan [33]. Với độ đo Mogne-Elkan, hai chuỗi X, Y sẽ được tách thành các chuỗi con dưới dạng các từ là $X = x_1 \dots x_K$ và $Y = y_1 \dots y_L$. Khoảng cách biến đổi sẽ được tính cho mỗi từ trong chuỗi con. Cụ thể Độ

do Mogne-Elkan được tính như sau:

$$Sim_{monge-elkan}(X, Y) = \frac{1}{K} \sum_{i=1}^K \max_{j=1}^L Sim'(x_i, y_j) \quad (19)$$

Trong đó,

- Sim' : là khoảng cách biến đổi (Edit Distance) như Levenshtein.

A.2.1.2 Đè xuất tập đặc trưng

Với bài toán nhập nhằng tên tác giả trong kho dữ liệu tích hợp, các chuỗi metadata của bài báo khoa học cần xem xét đó là: tên tác giả, tên đồng tác giả, tên cơ quan công tác, độ tương tự của các từ khóa trong bài báo. Với các chuỗi metadata này thì những độ đo tương tự ở mức từ, không quan tâm đến thứ tự của từ trong chuỗi là phù hợp, chẳng hạn như Jaccard. Các độ đo khoảng cách biến đổi (edit distance) thông thường chỉ phù hợp cho việc kiểm tra lỗi đánh máy nhầm. Do đó, luận án đã dùng Jaccard để tính toán độ tương tự của các metadata trong hai bài báo khoa học. Bên dưới là tập đặc trưng và cách tính để giải quyết nhập nhằng tên tác giả bằng phương pháp học giám sát.

(1) Tên tác giả

Chúng ta đưa ra giả thuyết: ‘Nếu hai chuỗi tên tác giả càng tương tự thì khả năng hai chuỗi này liên quan đến một người càng cao’. Độ tương tự hai chuỗi tên tác giả được tính như sau:

$$Author_Name_Sim(A, B) = \frac{|Author_Name_A \cap Author_Name_B|}{|Author_Name_A \cup Author_Name_B|} \quad (20)$$

Trong đó,

- $Author_Name_A$: Tên tác giả A trong bài báo đang xét.
- $|A \cap B|$: Số lượng từ giống nhau giữa tên tác giả A và B.
- $|A \cup B|$: Số lượng từ đôi một khác nhau trong tên của A và B.

Ví dụ:

$Author_Name_A = "Tuan Nguyen"$ và $Author_Name_B = "Nguyen Anh Tuan"$
 $|Author_Name_A \cap Author_Name_B| = 2$
 $|Author_Name_A \cup Author_Name_B| = 3$, và
 $Author_Name_Sim(A, B) \approx 0,67$.

(2) Tên đồng tác giả

Trong trường hợp này, giả thuyết sẽ là: ‘Nếu hai tác giả nhập nhằng, có càng nhiều đồng tác giả cùng tên trong hai bài báo khác nhau thì khả năng hai người này là một càng cao’. Giá trị đặc trưng dựa trên độ tương tự đồng tác giả được tính như sau:

$$CoAuthors_Names_Sim(A, B) = MAX(Author_Name_Sim(A_i, B_j)) \quad (21)$$

Trong đó,

- $A_i \in CoAuthors(A)$
và $CoAuthors(A)$: tập các đồng tác giả của A trong bài báo P_1 .
- $B_j \in CoAuthors(B)$
và $CoAuthors(B)$: tập các đồng tác giả của B trong bài báo P_2 .

(3) Cơ quan Công tác của tác giả

Cơ quan công tác của hai tác giả nhập nhằng trong hai bài báo khác nhau là một trong những đặc trưng quan trọng giúp nhận diện có hay không hai tác giả nhập nhằng này là một người. Độ tương tự cơ quan công tác của hai tác giả nhập nhằng cũng được tính dùng hệ số Jaccard như sau:

$$Aff_Sim(A, B) = \frac{|Aff_A \cap Aff_B|}{|Aff_A \cup Aff_B|} \quad (22)$$

Trong đó,

- Aff_A : Cơ quan của A trong bài báo đang xét.

-
- $|Aff_A \cap Aff_B|$: Số lượng từ giống nhau giữa tên cơ quan của A và cơ quan của B.
 - $|Aff_A \cup Aff_B|$: Số lượng từ phân biệt đôi một trong tên cơ quan của A và tên cơ quan của B.

(4) Cơ quan Công tác của Đồng tác giả

Giả thuyết được đưa ra trong trường hợp này là: 'Nếu hai tác giả nhập nhằng càng có nhiều đồng tác giả làm cùng cơ quan thì khả năng hai người này là một càng cao'. Giá trị đặc trưng này được tính như sau:

$$CoAuthor_Affs_Sim(A, B) = MAX(Aff_Sim(Aff_i_P1, Aff_j_P2)) \quad (23)$$

Trong đó,

- $Aff_i_P1 \in CoAuthors_Affs(A)$
i=1..n (n: số lượng đồng tác giả của A trong bài báo P_1)
và CoAuthors_Affs(A): là các cơ quan của các đồng tác giả của A trong bài báo P_1 .
- $Aff_i_P2 \in CoAuthors_Affs(B)$
j=1..m (m: số lượng đồng tác giả của B trong bài báo P_2)
và CoAuthors_Affs(B): là các cơ quan của các đồng tác giả của B trong bài báo P_2 .

(5) Từ khóa bài báo

Hai bài báo liên quan đến hai tác giả nhập nhằng càng chứa nhiều từ khóa tương tự nhau, thì khả năng hai tác giả này là một người càng cao. Đặc trưng tương tự dựa trên từ khóa được tính như sau:

$$Paper_Keywords_Sim(A, B) = \frac{|Paper_Keywords_A \cap Paper_Keywords_B|}{|Paper_Keywords_A \cup Paper_Keywords_B|} \quad (24)$$

Trong đó,

-
- $Paper_Keywords_A$: tập các từ khóa trong bài báo A.
 - $|Paper_Keywords_A \cap Paper_Keywords_B|$: số từ (token) giống nhau của các từ khóa trong bài báo A và bài báo B.
 - $|Paper_Keywords_A \cup Paper_Keywords_B|$: tổng số từ (token) đỏi một khác nhau trong các từ khóa của bài báo A và bài báo B.

Với tập đặc trưng đề xuất, các vector tương ứng với hai bài báo có tác giả nhập nhằng được xây dựng. Luận án áp dụng các phương pháp học giám để học và nhận diện tác giả nhập nhằng dựa trên tập dữ liệu được gán nhãn cho huấn luyện và kiểm tra.

Phụ lục B. Chi tiết kho dữ liệu học thuật

B.1 Nguồn dữ liệu thu thập

Dề có được kho dữ liệu khoa học đủ lớn và đủ phong phú, luận án đã tiến hành thu thập, tích hợp dữ liệu, thông tin bài báo khoa học từ nhiều nguồn khác nhau, những nguồn đáng tin cậy như DBLP, CiteSeerX, Microsoft Academic Search. Cơ sở dữ liệu khoa học xây dựng được, tạm gọi là CSPubguru, hiện đang được nghiên cứu sinh sử dụng để phát triển hệ thống tìm kiếm thông tin khoa học CSPubGuru (www.cspubguru.com). Tính đến tháng 12/2013, luận án đã tích hợp dữ liệu từ các nguồn sau:

DBLP - Digital Bibliography & Library Project

DBLP (<http://www.informatik.uni-trier.de/~ley/db/>) là một cơ sở dữ liệu khoa học mở cung cấp thông tin chỉ mục các bài báo trong lĩnh vực khoa học máy tính, dữ liệu này được xây dựng và thu thập bởi giáo sư Michael Ley từ trường đại học Universität Trier, Đức. Dữ liệu DBLP được cung cấp miễn phí dưới dạng file XML. Luận án sử dụng dữ liệu từ file XML mà DBLP cung cấp tháng 8/2013 có 2.356.294 bài báo và thông tin metadata đi kèm, được tải về từ liên kết <http://dblp.uni-trier.de/xml/>

Microsoft Academic Search (MAS)

Microsoft Academic Search, gọi tắt MAS (<http://academic.research.microsoft.com/>) là dự án thư viện số được phát triển bởi Microsoft Academic Research Asia và được tài

trợ bởi Microsoft. Dự án được thực hiện từ năm 2009, MAS đã thu thập và lập chỉ mục dữ liệu khoa học cho nhiều ngành, lĩnh vực khác nhau. Luận án đã xây dựng công cụ để trích, thu thập (crawling) từ trang web của MAS. Dữ liệu được thu thập từ tháng 10/2011 đến tháng 03/2012. Tổng cộng thu thập được 4.174.546 bài báo và thông tin metadata đi kèm.

CiteSeerX

CiteSeerX (<http://citeseerx.ist.psu.edu/>), là một hệ thống thư viện số được phát triển bởi giáo sư C. Lee Giles và cộng sự, thuộc trường Đại học bang Pennsylvania, Hoa Kỳ. CiteSeerX thu thập, rút trích metadata của các bài báo khoa học chuyên ngành khoa học máy tính dưới dạng các file PDF từ trang web của các nghiên cứu viên. CiteSeerX cho phép chúng ta thu thập dữ liệu từ cơ sở dữ liệu của họ thông qua giao thức OAIH. Luận án đã xây dựng công cụ và thu thập từ CiteSeerX. Tính đến tháng 05/2013 đã thu thập được 2.375.228 bài báo khoa học chuyên ngành khoa học máy tính và thông tin metadata đi kèm.

B.2 Cấu trúc của tập dữ liệu CSPubGuru

Hiện luận án tổ chức lưu trữ kho dữ liệu khoa học, CSPubGuru, tích hợp từ nhiều nguồn trong một cơ sở dữ liệu quan hệ có cấu trúc như mô tả trong hình B.1:

Một số bảng dữ liệu chính của CSPubGuru dataset có thể kể đến như:

- + **Paper**: Lưu thông tin metadata các bài báo khoa học.
- + **Author**: thông tin các tác giả của bài báo.
- + **Conference**: thông tin hội nghị.
- + **Journal**: thông tin tạp chí.
- + **Keyword**: từ khóa trong bài báo.
- + **Domain**: thông tin lĩnh vực của bài báo.
- + **Org**: cơ quan công tác của các tác giả.
- + **Author_paper**: quan hệ nghiên cứu viên là tác giả của bài báo.
- + **Paper_paper**: thông tin bài báo trích dẫn bài báo.

B.3 Kích thước của tập dữ liệu CSpubGuru

Sau khi tích hợp từ nhiều nguồn. Tính đến tháng 04/2014, tập dữ liệu khoa học CSpubGuru có kích thước như sau:

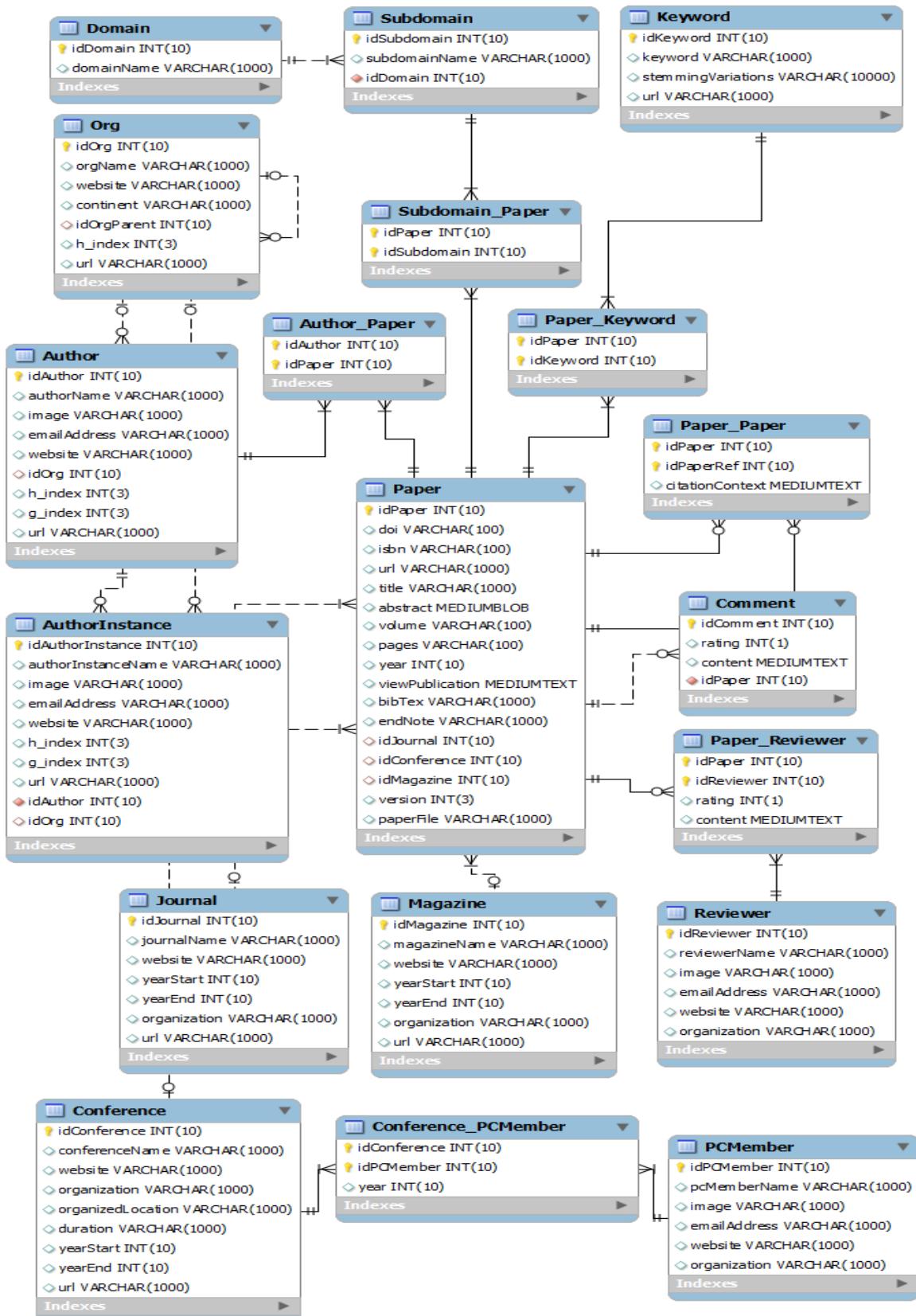
- **Paper:** 6.691.410
- **Author:** 1.931.898
- **Conference:** 9743
- **Journal:** 2719
- **Org:** 9557
- **Author_paper:** 12627886
- **Paper_paper:** 22749073

B.4 Download và trích dẫn

Tập dữ liệu khoa học CSpubGuru hiện công bố để phục vụ nghiên cứu khoa học và có thể download từ địa chỉ: www.cspubguru.com/DownloadServlet.

Các đề tài, bài báo nghiên cứu khác sử dụng tập dữ liệu CSpubGuru, cần trích dẫn bài báo bên dưới:

- Tin Huynh, Hiep Luong, and Kiem Hoang. Integrating bibliographical data of computer science publications from online digital libraries. In Proceedings of the 4th Asian conference on Intelligent Information and Database Systems (ACI-IDS'12), Springer-Verlag, Berlin, pages 226-235, (2012).



Hình B.1: Mô hình ERD biểu diễn cấu trúc của tập dữ liệu đã xây dựng, CSPubGuru dataset

TÀI LIỆU THAM KHẢO

- [1] Abbasi, A. and Altmann, J. (2010). A social network system for analyzing publication activities of researchers. TEMEP Discussion Papers 201058, Seoul National University; Technology Management, Economics, and Policy Program (TEMEP). [61](#)
- [2] Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211 – 230. [84, 98](#)
- [3] Adomavicius, G., Mobasher, B., Ricci, F., and Tuzhilin, A. (2011). Context-aware recommender systems. *AI Magazine*, 32(3):67–80. [69](#)
- [4] Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Syst.*, 23(1):103–145. [125](#)
- [5] Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17:734–749. [14, 15, 26, 27, 28, 50, 69, 128, 134](#)
- [6] Adomavicius, G. and Tuzhilin, A. (2011). Context-aware recommender systems. In *Recommender Systems Handbook*, pages 217–253. Springer US. [69](#)
- [7] Ahn, J.-w., Brusilovsky, P., Grady, J., He, D., and Syn, S. Y. (2007). Open user profiles for adaptive news systems: Help or harm? In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 11–20, New York, NY, USA. ACM. [34](#)
- [8] Aranda, J., Givoni, I., Handcock, J., and Tarlow, D. (2007). An online social network-based recommendation system. *Toronto, Ontario, Canada*. [60](#)

-
- [9] Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 33, 34, 36, 109, 111, 158
- [10] Baker, K. (2005). Singular value decomposition tutorial. 49
- [11] Balabanovic, M. and Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Commun. ACM*, 40(3):66–72. 16, 29, 57
- [12] Balog, K. and de Rijke, M. (2007). Finding similar experts. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’07, pages 821–822, New York, NY, USA. ACM. 95
- [13] Balthrop, J., Forrest, S., Newman, M. E. J., and Williamson, M. M. (2004). Technological networks and the spread of computer viruses. *CoRR*, cs.NI/0407048. 18
- [14] Basu, C., Hirsh, H., and Cohen, W. (1998). Recommendation as classification: Using social and content-based information in recommendation. In *In Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 714–720. AAAI Press. 54
- [15] Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., and Nürnberg, A. (2013). Research paper recommender system evaluation: A quantitative literature survey. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, RepSys ’13, pages 15–22, New York, NY, USA. ACM. 126, 138
- [16] Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Commun. ACM*, 35(12):29–38. 14, 29
- [17] Bennett, J. and Lanning, S. (2007). The netflix prize. In *Proceedings of the KDD Cup Workshop 2007*, pages 3–6, New York. ACM. 47
- [18] Bhuiyan, T. (2013). *Trust for Intelligent Recommendation*. Springer Publishing Company, Incorporated. 63

-
- [19] Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., and Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23. [157](#)
- [20] Billsus, D. and Pazzani, M. J. (1999). A hybrid user model for news story classification. In *Proceedings of the Seventh International Conference on User Modeling*, UM ’99, pages 99–108, Secaucus, NJ, USA. Springer-Verlag New York, Inc. [34](#), [35](#)
- [21] Billsus, D. and Pazzani, M. J. (2000). User modeling for adaptive news access. *User Modeling and User-Adapted Interaction*, 10(2-3):147–180. [34](#), [35](#), [37](#), [52](#)
- [22] Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 46(0):109 – 132. [27](#), [28](#), [50](#), [69](#), [134](#)
- [23] Brandão, M. A., Moro, M. M., Lopes, G. R., and Oliveira, J. P. (2013). Using link semantics to recommend collaborations in academic social networks. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW ’13 Companion, pages 833–840, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. [62](#), [97](#)
- [24] Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI’98, pages 43–52, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. [42](#), [44](#), [45](#)
- [25] Burke, R. D. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-adapted Interaction*, 12:331–370. [25](#), [51](#), [53](#), [55](#), [69](#)
- [26] Celma, O. (2010). Evaluation metrics. In *Music Recommendation and Discovery*, pages 109–128. Springer Berlin Heidelberg. [66](#)
- [27] Chen, H.-H., Gou, L., Zhang, X., and Giles, C. L. (2011a). Capturing missing edges in social networks using vertex similarity. In *Proceedings of the sixth international conference on Knowledge capture*, K-CAP ’11, pages 195–196, New York, NY, USA. ACM. [72](#), [88](#), [97](#), [98](#), [103](#), [105](#), [111](#), [113](#), [115](#)

-
- [28] Chen, H.-H., Gou, L., Zhang, X., and Giles, C. L. (2011b). Collabseer: a search engine for collaboration discovery. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 231–240, New York, NY, USA. ACM. [72](#), [88](#), [95](#), [97](#), [98](#), [99](#), [103](#), [105](#), [111](#), [113](#), [115](#)
- [29] Chen, H.-H., Gou, L., Zhang, X. L., and Giles, C. L. (2012). Discovering missing links in networks using vertex similarity measures. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 138–143, New York, NY, USA. ACM. [72](#), [88](#), [97](#), [98](#), [103](#), [105](#)
- [30] Chen, L. and Wang, F. (2013). Preference-based clustering reviews for augmenting e-commerce recommendation. *Knowl.-Based Syst.*, 50:44–59. [38](#)
- [31] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., and Sartin, M. (1999a). Combining content-based and collaborative filters in an online newspaper. In *Proceedings of the ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*, Berkeley, California. ACM. [16](#)
- [32] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., and Sartin, M. (1999b). Combining content-based and collaborative filters in an online newspaper. [52](#)
- [33] Cohen, W. W., Ravikumar, P. D., and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *IIWeb*, pages 73–78. [157](#), [158](#)
- [34] Connor, M. and Herlocker, J. (1999). Clustering items for collaborative filtering. In *Proceedings of the ACM SIGIR Workshop on Recommender Systems*, SIGIR '99, New York, NY, USA. ACM. [45](#)
- [35] Dasgupta, P. (1988). Trust as a commodity. In Gambetta, D., editor, *Trust: Making and Breaking Cooperative Relations*, pages 49–72. Blackwell. [63](#)
- [36] Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., and Sampath, D. (2010). The youtube video recommendation system. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 293–296, New York, NY, USA. ACM. [14](#)

-
- [37] Davoodi, E., Afsharchi, M., and Kianmehr, K. (2012). A social network-based approach to expert recommendation system. In *Proceedings of the 7th international conference on Hybrid Artificial Intelligent Systems - Volume Part I*, HAIS'12, pages 91–102, Berlin, Heidelberg. Springer-Verlag. [61](#), [97](#)
- [38] de Gemmis, M., Lops, P., Semeraro, G., and Basile, P. (2008). Integrating tags in a semantic content-based recommender. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 163–170, New York, NY, USA. ACM. [36](#)
- [39] El-Arini, K. and Guestrin, C. (2011). Beyond keyword search: Discovering relevant scientific literature. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 439–447, New York, NY, USA. ACM. [122](#), [125](#)
- [40] Esslimani, I., Brun, A., and Boyer, A. (2009). From social networks to behavioral networks in recommender systems. In *Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in*, pages 143–148. [61](#)
- [41] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Simoudis, E., Han, J., and Fayyad, U. M., editors, *KDD*, pages 226–231. AAAI Press. [45](#)
- [42] Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *SIGMOD Rec.*, 41(2):15–26. [156](#), [157](#)
- [43] Foltz, P. W. and Dumais, S. T. (1992). Personalized information delivery: An analysis of information filtering methods. *Commun. ACM*, 35(12):51–60. [14](#)
- [44] Gallego, L. S., Gamage, D. U., Hill, J. H., and Raje, R. R. (2014). Towards trust-based recommender systems for online software services. In *Proceedings of the 9th Annual Cyber and Information Security Research Conference*, CISR '14, pages 61–64, New York, NY, USA. ACM. [63](#)

-
- [45] Gauch, S., Speretta, M., Chandramouli, A., and Micarelli, A. (2007). The adaptive web. chapter User Profiles for Personalized Information Access, pages 54–89. Springer-Verlag, Berlin, Heidelberg. [32](#)
- [46] Giles, C. L., Bollacker, K. D., and Lawrence, S. (1998). Citeseer: an automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, DL ’98, pages 89–98, New York, NY, USA. ACM. [115](#)
- [47] Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retr.*, 4(2):133–151. [16](#)
- [48] Gollapalli, S. D., Mitra, P., and Giles, C. L. (2012). Similar researcher search in academic environments. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, JCDL ’12, pages 167–170, New York, NY, USA. ACM. [72](#), [95](#)
- [49] Gunawardana, A. and Shani, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10:2935–2962. [25](#), [64](#), [65](#), [67](#), [68](#), [70](#)
- [50] He, Q., Kifer, D., Pei, J., Mitra, P., and Giles, C. L. (2011). Citation recommendation without author supervision. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM ’11, pages 755–764, New York, NY, USA. ACM. [122](#), [125](#)
- [51] He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, L. (2010). Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW ’10, pages 421–430, New York, NY, USA. ACM. [122](#), [125](#)
- [52] Hofmann, K., Balog, K., Bogers, T., and de Rijke, M. (2010). Contextual factors for finding similar experts. *J. Am. Soc. Inf. Sci. Technol.*, 61(5):994–1014. [95](#)
- [53] Huang, W., Kataria, S., Caragea, C., Mitra, P., Giles, C. L., and Rokach, L. (2012). Recommending citations: Translating papers into references. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM ’12, pages 1910–1914, New York, NY, USA. ACM. [122](#), [125](#)

-
- [54] Hurley, N. and Zhang, M. (2011). Novelty and diversity in top-n recommendation – analysis and evaluation. *ACM Trans. Internet Technol.*, 10(4):14:1–14:30. [67](#)
- [55] Huynh, T., Luong, H., Hoang, K., Gauch, S., Do, L., and Tran, H. (2012). Scientific publication recommendations based on collaborative citation networks. In *Proceedings of the 3rd International Workshop on Adaptive Collaboration (AC 2012) as part of The 2012 International Conference on Collaboration Technologies and Systems (CTS 2012)*, pages 316 – 321, Denver, Colorado, USA. [18](#), [122](#), [125](#)
- [56] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. [84](#)
- [57] Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2010). *Recommender Systems: An Introduction*. Cambridge University Press, New York, NY, USA, 1st edition. [26](#), [27](#), [41](#), [42](#)
- [58] Järvelin, K. and Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’00, pages 41–48, New York, NY, USA. ACM. [68](#), [139](#)
- [59] Jeh, G. and Widom, J. (2002). Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ’02, pages 538–543, New York, NY, USA. ACM. [98](#)
- [60] Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML ’98, pages 137–142, London, UK, UK. Springer-Verlag. [36](#), [38](#)
- [61] Katz, J. S., Katz, J. S., Martin, B. R., and Martin, B. R. (1997). What is research collaboration. *Research Policy*, 26:1–18. [94](#)
- [62] Kirchhoff, L., Stanoevska-Slabeva, K., Nicolai, T., Fleck, M., and Stanoevska, K. (2008). Using social network analysis to enhance information retrieval systems. In

-
- Applications of Social Network Analysis (ASNA) (Zurich)*, volume 7, pages 1–21. 18, 61
- [63] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J. (1997). GroupLens: Applying collaborative filtering to usenet news. *Commun. ACM*, 40(3):77–87. 16
- [64] Konstas, I., Stathopoulos, V., and Jose, J. M. (2009). On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’09, pages 195–202, New York, NY, USA. ACM. 97
- [65] Koren, Y. (2008). Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’08, pages 426–434, New York, NY, USA. ACM. 47
- [66] Lawrence, S., Giles, C. L., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *Computer*, 32:67–71. 122, 125
- [67] Le, D.-L., Nguyen, A.-T., Nguyen, D.-T., and Hunger, A. Building learner profile in adaptive e-learning systems. 35
- [68] Li, Q. and Kim, B. M. (2003). An approach for combining content-based and collaborative filters. In *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages - Volume 11*, AsianIR ’03, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics. 16
- [69] Lieberman, H. (1995). Letizia: An agent that assists web browsing. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI’95, pages 924–929, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 34
- [70] Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80. 14, 16

-
- [71] Liu, J., Dolan, P., and Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. In Rich, C., 0001, Q. Y., Cavazza, M., and Zhou, M. X., editors, *IUI*, pages 31–40. ACM. [39](#)
- [72] Lopes, G. R., Moro, M. M., Wives, L. K., and De Oliveira, J. P. M. (2010). Collaboration recommendation on academic social networks. In *Proceedings of the 2010 international conference on Advances in conceptual modeling: applications and challenges*, ER’10, pages 190–199, Berlin, Heidelberg. Springer-Verlag. [97](#)
- [73] Lops, P., de Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 73–105. Springer US. [30](#), [31](#), [32](#), [34](#)
- [74] Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12):317 – 324. [94](#)
- [75] Luong, H. P., Huynh, T., Gauch, S., Do, L., and Hoang, K. (2012a). Publication venue recommendation using author network’s publication history. In *ACIIDS (3)*, pages 426–435. [72](#)
- [76] Luong, H. P., Huynh, T., Gauch, S., and Hoang, K. (2012b). Exploiting social networks for publication venue recommendations. In *KDIR*, pages 239–245. [18](#), [72](#)
- [77] Ma, H., Lyu, M. R., and King, I. (2009). Learning to recommend with trust and distrust relationships. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys ’09, pages 189–196, New York, NY, USA. ACM. [63](#)
- [78] Ma, H., Yang, H., Lyu, M. R., and King, I. (2008). Sorec: Social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM ’08, pages 931–940, New York, NY, USA. ACM. [60](#)
- [79] Ma, H., Zhou, D., Liu, C., Lyu, M. R., and King, I. (2011). Recommender systems with social regularization. In *Proceedings of the Fourth ACM International Confer-*

-
- ence on Web Search and Data Mining, WSDM '11, pages 287–296, New York, NY, USA. ACM. 18
- [80] Macqueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability*, volume 1, pages 281–297. University of California Press. 45
- [81] Marsh, S. P. (1994). *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling. 63
- [82] Massa, P. and Avesani, P. (2007). Trust-aware recommender systems. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 17–24, New York, NY, USA. ACM. 63
- [83] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444. 59
- [84] Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., and Riedl, J. (2003). MovieLens unplugged: Experiences with an occasionally connected recommender system. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, pages 263–266, New York, NY, USA. ACM. 14
- [85] Mladenic, D. (1999). Machine learning used by personal webWatcher. In *Proceedings of ACAI-99 Workshop on Machine Learning and Intelligent Agents*, Chania, Crete, Greece. 34
- [86] Mooney, R. J. and Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 195–204, New York, NY, USA. ACM. 36, 55
- [87] Musat, C. C., Liang, Y., and Faltings, B. (2013). Recommendation using textual opinions. In Rossi, F., editor, *IJCAI. IJCAI/AAAI*. 38
- [88] Newman, M. (2004). Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):321–330. 18

-
- [89] Newman, M. E. J. (2001). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132+. [18](#)
- [90] Nguyen, A.-T., Thi, B.-T. D., and Le, D.-L. A tool for instructional recommendation in e-learning. [35](#)
- [91] Nicholas, I. S. C. and Nicholas, C. K. (1999). Combining content and collaboration in text filtering. In *In Proceedings of the IJCAI'99 Workshop on Machine Learning for Information Filtering*, pages 86–91. [16](#)
- [92] Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Mach. Learn.*, 39(2-3):103–134. [36](#), [38](#)
- [93] Ohta, M., Hachiki, T., and Takasu, A. (2011). Related paper recommendation to support online-browsing of research papers. In *Applications of Digital Information and Web Technologies (ICADIWT), 2011 Fourth International Conference*, pages 130–136. [122](#), [125](#)
- [94] Pazzani, M. and Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Mach. Learn.*, 27(3):313–331. [36](#)
- [95] Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artif. Intell. Rev.*, 13(5-6):393–408. [16](#), [52](#), [56](#)
- [96] Pazzani, M. J., Muramatsu, J., and Billsus, D. (1996). Syskill & webert: Identifying interesting web sites. In Clancey, W. J. and Weld, D. S., editors, *AAAI/IAAI, Vol. 1*, pages 54–61. AAAI Press / The MIT Press. [39](#)
- [97] Rajaraman, A. and Ullman, J. D. (2012). *Mining of massive datasets*. Cambridge University Press. [15](#)
- [98] Rashid, A. M., Lam, S. K., LaPitz, A., Karypis, G., and Riedl, J. (2007). Towards a scalable knn cf algorithm: Exploring effective applications of clustering. In *Proceedings of the 8th Knowledge Discovery on the Web International Conference on Advances in Web Mining and Web Usage Analysis*, WebKDD’06, pages 147–166, Berlin, Heidelberg. Springer-Verlag. [133](#)

-
- [99] Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of The ACM*, 40:56–58. [25](#)
- [100] Ricci, F., Rokach, L., and Shapira, B. (2011). Introduction to recommender systems handbook. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 1–35. Springer US. [25](#)
- [101] Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. [84](#), [97](#)
- [102] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW ’01, pages 285–295, New York, NY, USA. ACM. [44](#)
- [103] Sarwar, B. M., Karypis, G., Konstan, J., and Riedl, J. (2002). Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *5th International Conference on Computer Information Technology (ICCIT)*. [45](#)
- [104] Serrat, O. (2009). Social network analysis. *Knowledge Solutions*. [58](#)
- [105] Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender Systems Handbook*, pages 257–297. Springer US. [66](#), [70](#)
- [106] Shani, G. and Gunawardana, A. (2013). Tutorial on application-oriented evaluation of recommendation systems. *AI Commun.*, 26(2):225–236. [70](#)
- [107] Sheth, B. and Maes, P. (1993). Evolving agents for personalized information filtering. In *Artificial Intelligence for Applications, 1993. Proceedings., Ninth Conference on*, pages 345–352. [34](#)
- [108] Smyth, B. and Cotter, P. (2000). A personalized television listings service. *Commun. ACM*, 43(8):107–111. [53](#)
- [109] Stefanidis, K., Ntoutsi, I., Nørvåg, K., and Kriegel, H.-P. (2012). A framework for time-aware recommendations. In Liddle, S., Schewe, K.-D., Tjoa, A., and Zhou,

-
- X., editors, *Database and Expert Systems Applications*, volume 7447 of *Lecture Notes in Computer Science*, pages 329–344. Springer Berlin Heidelberg. 27, 69
- [110] Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2. 41, 42, 45, 46, 50, 132
- [111] Sugiyama, K. and Kan, M.-Y. (2010). Scholarly paper recommendation via user’s recent research interests. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL ’10, pages 29–38, New York, NY, USA. ACM. 72, 122, 125, 126, 129, 130, 131, 138, 139
- [112] Sugiyama, K. and Kan, M.-Y. (2011). Serendipitous recommendation for scholarly papers considering relations among researchers. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL ’11, pages 307–310, New York, NY, USA. ACM. 72, 122, 125, 126, 139
- [113] Sugiyama, K. and Kan, M.-Y. (2013). Exploiting potential citation papers in scholarly paper recommendation. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL ’13, pages 153–162, New York, NY, USA. ACM. 72, 122, 125, 126, 138, 139
- [114] Sun, J., Ma, J., Liu, Z., and Miao, Y. (2013). Leveraging content and connections for scientific article recommendation in social computing contexts. *The Computer Journal*, page bxt086. 122, 125, 126, 138
- [115] Sztompka, P. (1999). *Trust: A Sociological Theory*. Cambridge University Press, Cambridge. 63
- [116] Tang, J., Hu, X., and Liu, H. (2013). Social recommendation: a review. *Social Network Analysis and Mining*, 3(4):1113–1133. 60
- [117] Tang, J., Wu, S., Sun, J., and Su, H. (2012a). Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ’12, pages 1285–1293, New York, NY, USA. ACM. 72, 95, 105, 111, 113, 115, 138

-
- [118] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 990–998, New York, NY, USA. ACM. [61](#), [115](#)
- [119] Tang, W., Tang, J., Lei, T., Tan, C., Gao, B., and Li, T. (2012b). On optimization of expertise matching with various constraints. *Neurocomput.*, 76(1):71–83. [95](#)
- [120] Terzi, M., Rowe, M., Ferrario, M.-A., and Whittle, J. (2014). Text-based user-knn: Measuring user similarity based on text reviews. In Dimitrova, V., Kuflík, T., Chin, D., Ricci, F., Dolog, P., and Houben, G.-J., editors, *User Modeling, Adaptation, and Personalization*, volume 8538 of *Lecture Notes in Computer Science*, pages 195–206. Springer International Publishing. [37](#)
- [121] Tong, H., Faloutsos, C., and Pan, J.-Y. (2006). Fast random walk with restart and its applications. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 613–622, Washington, DC, USA. IEEE Computer Society. [91](#), [110](#)
- [122] Vargas, S. and Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 109–116, New York, NY, USA. ACM. [67](#)
- [123] Voorhees, E. M. (1999). The TREC-8 question answering track report. In *TREC*. [139](#)
- [124] Wang, C. and Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 448–456, New York, NY, USA. ACM. [122](#), [125](#)
- [125] Wanjantuk, P.; Keane, J. (2004). Finding related documents via communities in the citation graph. In *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium*, pages 445–450. [122](#), [125](#)

-
- [126] Wasserman, S., Faust, K., and Iacobucci, D. (1994). *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press. [57](#), [58](#)
- [127] Xu, J. J. and Chen, H. (2005). Crimenet explorer: a framework for criminal network knowledge discovery. *ACM Trans. Inf. Syst.*, 23(2):201–226. [18](#)
- [128] Yue, Y., Finley, T., Radlinski, F., and Joachims, T. (2007). A support vector method for optimizing average precision. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 271–278, New York, NY, USA. ACM. [68](#)
- [129] Zhang, M. and Hurley, N. (2009). Evaluating the diversity of top-n recommendations. In *ICTAI*, pages 457–460. [67](#)
- [130] Zhao, P., Han, J., and Sun, Y. (2009). P-rank: a comprehensive structural similarity measure over information networks. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 553–562, New York, NY, USA. ACM. [98](#)