

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



Nguyễn Văn Quyền

**THUẬT TOÁN PHÂN CỤM TRONG KHAI PHÁ KHÍA
CẠNH TỔ CHỨC TRONG PHÁT HIỆN QUÁ TRÌNH**

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: *Công nghệ thông tin*

Hà Nội -2014

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Văn Quyền

**THUẬT TOÁN PHÂN CỤM TRONG KHAI PHÁ KHÍA
CẠNH TỔ CHỨC TRONG PHÁT HIỆN QUÁ TRÌNH**

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: *Công nghệ thông tin*

Cán bộ hướng dẫn: PGS.TS Hà Quang Thụy

Cán bộ đồng hướng dẫn: ThS. Lê Hoàng Quỳnh

Hà Nội -2014

**VIETNAM NATIONAL UNIVERSITY
UNIVERSITY OF ENGINEERING AND TECHNOLOGY**

Nguyen Van Quyen

**CLUSTERING ALGORITHMS ON ORGANIZATIONAL
PROCESS MINING**

Major: Information Technology

Supervisor: Assoc.Prof. Ha Quang Thuy

Co-Supervisor: MSc. Le Hoang Quynh

Hanoi - 2014

Lời cảm ơn

Trước tiên, em xin bày tỏ lòng biết ơn chân thành và sâu sắc tới Thầy giáo PGS.TS Hà Quang Thụy đã tận tình hướng dẫn, chỉ bảo và giúp đỡ em trong suốt quá trình làm khóa luận.

Em xin gửi lời cảm ơn sâu sắc đến các thầy cô trong Khoa Công nghệ thông tin đã truyền đạt những kiến thức quý báu cho em trong suốt quá trình học tập tại Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội.

Em cũng xin gửi lời cảm ơn tới các thầy cô, các anh chị, các bạn trong phòng thí nghiệm KTLAB đã giúp đỡ em rất nhiều trong việc hỗ trợ các kiến thức chuyên môn để hoàn thành tốt khóa luận.

Cuối cùng, em xin cảm ơn các anh chị và bạn bè, đặc biệt là các thành viên lớp K55CLC và K55CD đã ủng hộ và giúp đỡ tôi trong bốn năm học tại giảng đường cũng như trong thời gian thực hiện đề tài khóa luận.

Tôi xin chân thành cảm ơn !

Hà nội, ngày 14 tháng 5 năm 2014

Sinh viên

Nguyễn Văn Quyền

THUẬT TOÁN PHÂN CỤM TRONG KHAI PHÁ KHÍA CẠNH TỔ CHỨC TRONG PHÁT HIỆN QUÁ TRÌNH

Nguyễn Văn Quyền

Khóa QH-2010-I/CQ, Công nghệ thông tin

Tóm tắt khóa luận tốt nghiệp:

Khai phá quá trình là một chuyên ngành nghiên cứu mới tập trung vào phân tích quá trình dựa trên nhật ký sự kiện được ghi lại trong các hệ thống thông tin. Hiện nay, nghiên cứu trong lĩnh vực này ngày càng được quan tâm và nghiên cứu [2,3,4,5,12,13,14]. Bài toán khai phá quá trình thì tập trung vào khía cạnh luồng điều khiển mà bỏ qua các thông tin quan trọng như nguồn thực hiện hành động, thời gian, các trường hợp trong nhật ký sự kiện. Trong khi đó các thông tin này cũng quan trọng và có nhiều ý nghĩa cần được khai phá. Khía cạnh tổ chức là một trong những khía cạnh được nhiều nhà khoa học trên thế giới quan tâm, nghiên cứu nổi bật là nhóm của WMP Van der Aalst [4] và các nghiên cứu khác trong [2,5].

Dựa trên tìm hiểu một số nghiên cứu của Van der Aalst [3,4] và Claudia Sofia da Costa Alves [2] về khai phá khía cạnh tổ chức trong khai phá quá trình, khóa luận trình bày các thuật toán phân cụm được sử dụng trong việc phát hiện cấu trúc tổ chức trong khai phá quá trình như AHC và K-means. Ngoài ra, hai thuật toán này không có khả năng phát hiện được sự chồng chéo trong tổ chức, tức là một cá nhân thuộc về nhiều hơn một nhóm vì vậy trong khóa luận sẽ trình bày thêm thuật toán CONGA (cải tiến từ thuật toán Girvan Newman) và cải tiến của CONGA là thuật toán CONGO có thể phát hiện sự chồng chéo trong tổ chức.

Thực nghiệm giải quyết mô hình bài toán với thuật toán phân cụm phân cấp AHC cho việc phát hiện cấu trúc tổ chức không có sự chồng chéo và sử dụng công cụ phần mềm của thuật toán CONGA và các cải tiến được đưa ra bởi Steve Gregory [14] để phát hiện cấu trúc tổ chức có sự chồng chéo với dữ liệu trích xuất từ nhật ký sự kiện.

CLUSTERING ALGORITHMS ON ORGANIZATIONAL PROCESS MINING

Nguyen Van Quyen

QH-2010-I/CQ course, information technology faculty

Abstract:

Process mining emerged as a new research field focus on process analysis based on the available event log which is recorded on information systems. Today, the research in this field has received attention of many scientists around the world [2, 3, 4, 5, 12, 13, 14]. Whereas the main focus of process discovery is on the control-flow perspective, event logs may contain a wealth of information relating to other perspectives such as the organizational perspective, the case perspective and the time perspective. Furthermore, the information on these perspectives is important and meaningful. Organizational perspective is also received attention of many scientists especially the native group of WMP Van der Alast [4] and others in [2,5].

Based on the research of Van der Alast [3,4] and Claudia Sofia da Costa Alves [2] on organizational process mining, this thesis presents the clustering algorithms which are used on finding organization structure as AHC and K-means. In addition, AHC and K-means have not been to detect the overlapping organization, i.e an individual belongs to more than one group, thus this thesis presents the CONGA algorithm (based on Girvan and Newman algorithm) and the improvement of CONGA is CONGO algorithm can detect the overlapping organization.

Experimenting problem solving model with AHC algorithm for finding organizational structure without overlapping and use the CONGA software given by Steve Gregory [14] to detect the structure of overlapping organizations with data extracted from event log.

Keywords: process mining, clustering algorithm, organizational process mining, event log.

Lời cam đoan

Tôi xin cam đoan mô hình giải quyết bài toán khai phá tổ chức trong phát hiện quá trình bằng các thuật toán phân cụm và thực nghiệm trình bày trong khóa luận là do tôi thực hiện dưới sự hướng dẫn của PGS.TS Hà Quang Thụy.

Tất cả các tài liệu tham khảo từ các nghiên cứu liên quan đến khóa luận đều được nêu nguồn gốc một cách rõ ràng từ danh mục tài liệu tham khảo. Trong khóa luận, không có việc sao chép tài liệu, công trình nghiên cứu của người khác mà không ghi rõ tài liệu tham khảo.

Hà Nội, ngày 14 tháng 5 năm 2014

Sinh viên

Nguyễn Văn Quyền

Mục lục

Mở đầu	1
Chương 1. TỔNG QUAN VỀ KHAI PHÁ KHÍA CẠNH TỔ CHỨC TRONG KHAI PHÁ QUÁ TRÌNH	3
1.1 Khái quát về khai phá quá trình	3
1.2 Một số khái niệm	5
1.2.1 Nhật ký sự kiện	5
1.2.2 Lưới Petri	8
1.3 Khai phá khía cạnh bổ sung trong phát hiện quá trình	9
1.4 Khai phá khía cạnh tổ chức trong khai phá quá trình	9
1.4.1 Khai phá mô hình tổ chức	9
1.4.2 Phân tích mạng xã hội	11
1.4.3 Tìm cấu trúc cộng đồng trong mạng xã hội	12
1.5 Tóm tắt chương 1	13
Chương 2. CÁC THUẬT TOÁN PHÂN CỤM TRONG KHAI PHÁ KHÍA CẠNH TỔ CHỨC	14
2.1 Độ đo mô đun hóa để đánh giá chất lượng phân chia đồ thị	14
2.2 Thuật toán phân cụm phân cấp AHC (Agglomerative Hierarchical Clustering) ..	18
2.3 Thuật toán phân cụm K-means	21
2.4 Thuật toán CONGA	23
2.5 Thuật toán CONGO	25
2.5.1 Độ trung gian cục bộ	26
2.5.2 Các bước của thuật toán	27
2.6 Tóm tắt chương 2	31
Chương 3. MÔ HÌNH PHÁT HIỆN TỔ CHỨC TRONG PHÁT HIỆN QUÁ TRÌNH SỬ DỤNG CÁC THUẬT TOÁN PHÂN CỤM	32
3.1 Mô hình phát hiện tổ chức trong phát hiện quá trình	32
3.2 Phân tích các thành phần trong mô hình	33
3.3 Tóm tắt chương 3	38

Chương 4. THỰC NGHIỆM.....	40
4.1 Môi trường và các công cụ thực nghiệm.....	40
4.2 Dữ liệu thực nghiệm.....	43
4.3 Thực nghiệm	44
4.4 Kết quả thực nghiệm	45
4.4.1 Kết quả thực nghiệm tại bước 1	45
4.4.2 Kết quả thực nghiệm tại bước 2.....	46
4.4.3 Kết quả thực nghiệm tại bước 3.....	47
Kết luận và định hướng nghiên cứu tiếp theo	54
Tài liệu tham khảo	55

Danh sách hình vẽ

Hình 1.1 Ngữ cảnh của khai phá quá trình [3]	4
Hình 1.2 Ba bài toán trong khai phá quá trình [3].....	5
Hình 1.3 Cấu trúc của nhật ký sự kiện [3].....	7
Hình 1.4 Ví dụ về lưới Petri [3].....	8
Hình 1.5 Mạng lưới với cấu trúc cộng đồng [2].....	13
Hình 2.1 Cách tính khoảng cách giữa hai cụm.....	20
Hình 2.2 Ví dụ về thuật toán K-means [7]	22
Hình 2.3 (a) đồ thị ban đầu (b) Cách chia tốt nhất của đỉnh a có độ trung gian lớn nhất (c) và (d) Các cách chia khác của đỉnh a. [6].....	25
Hình 2.4 Mô tả thuật toán CONGO trong xóa cạnh trong vùng h [1]	28
Hình 3.1 Mô hình đề xuất giải quyết bài toán phát hiện cấu trúc tổ chức trong phát hiện quá trình.....	32
Hình 3.2 Một phần nhật ký sự kiện định dạng XES [3].....	33
Hình 3.3 Sơ đồ lớp UML cho mô hình của chuẩn XES [15]	34
Hình 3.4 Hình thể hiện số công việc giống nhau giữa hai nhân viên trích xuất từ hệ quản trị cơ sở dữ liệu	37
Hình 3.5 Kết quả phân cụm theo độ đo làm các công việc giống nhau với số cụm bằng 5 của nhật ký sự kiện [16]	38
Hình 4.1 Định dạng file đầu vào của bộ phần mềm CONGA [17]	41
Hình 4.2 Mạng xây dựng theo độ đo làm việc cùng nhau từ nhật ký sự kiện [18]	46
Hình 4.3 Mạng sinh ra từ nhật ký sự kiện theo độ đo làm việc cùng nhau (similar tasks)	47
Hình 4.4 Giá trị mô đun hóa sau mỗi lần phân cụm theo AHC	48
Hình 4.5 Mô hình tổ chức phát hiện được từ nhật ký sự kiện [18]	49
Hình 4.6 Kết quả khi phân cụm bằng thuật toán CONGA.....	50
Hình 4.7 Kết quả khi phân cụm bằng thuật toán CONGO.....	50
Hình 4.8 Giá trị mô đun hóa khi phân cụm bằng thuật toán CONGA	51
Hình 4.9 Biểu diễn giá trị mô đun hóa sau khi chạy bằng thuật toán CONGA, CONGO với h=2 và h=3.....	52
Hình 4.10 Cấu trúc tổ chức được phát hiện từ thuật toán phân cụm CONGA	53

Danh sách bảng biểu

Bảng 1.1 Cấu trúc của một nhật ký sự kiện [3]	6
Bảng 2.1 Bảng ma trận kề giữa các cá nhân theo độ đo làm việc cùng nhau [2].....	15
Bảng 2.2 Bảng thể hiện bậc của nhân viên [2]	16
Bảng 2.3 Độ trung gian của các cạnh trong vùng h bằng 2.....	29
Bảng 2.4 Độ trung gian của các cạnh trong vùng h=2 sau khi xóa cạnh	30
Bảng 2.5 Độ trung gian của các cạnh trong vùng h sau khi kết thúc thuật toán trong hình 2.4 c	31
Bảng 3.1 Bảng các công việc mà các nhân viên thực hiện và số lần thực hiện công việc [2].....	36
Bảng 3.2 Số lượng công việc chung giữa hai người thực hiện công việc [2]	36
Bảng 4.1 Cấu hình phần cứng	40
Bảng 4.2 Các phần mềm và công cụ sử dụng.....	41
Bảng 4.3 Các mở rộng (option) dùng trong thực nghiệm của bộ phần mềm CONGA [17]	43

Mở đầu

Ngày nay, rất nhiều tổ chức đã xây dựng cho mình những mô hình kinh doanh nhằm hỗ trợ hoạt động quản lý và cải tiến quá trình hoạt động kinh doanh. Trong quá trình hoạt động, các thông tin về các hoạt động kinh doanh được lưu trữ lại trong hệ thống thông tin của tổ chức đó. Hơn nữa, những dữ liệu sự kiện ghi lại trong hệ thống ngày càng tăng lên và phản ánh tốt hơn thực tiễn quá trình kinh doanh của tổ chức đó. Từ đó đặt ra những đòi hỏi mới, yêu cầu mới trong hình thành và phát triển nghiên cứu về khai phá quá trình (process mining). Khai phá quá trình là hướng nghiên cứu tích hợp khai phá dữ liệu với quản lý quá trình kinh doanh nhằm phát hiện những tri thức mới từ những dữ liệu thu thập được trong hệ thống thông tin.

Trong khai phá quá trình, ngoài việc tập trung vào khai phá khía cạnh luồng công việc, các khía cạnh bổ sung như khía cạnh tổ chức, khía cạnh thời gian và khía cạnh trường hợp cũng có ý nghĩa quan trọng, được các nhà khoa học trên thế giới quan tâm nghiên cứu [2,3,4,5].

Trong khóa luận tập trung vào một trong những khía cạnh bổ sung trong phát hiện quá trình đó là khía cạnh tổ chức. Trong nhật ký sự kiện ghi lại được chứa các sự kiện, mỗi một sự kiện tương ứng với một hành động và trong bài toán khai phá khía cạnh tổ chức sẽ tập trung vào phân tích tác nhân thực hiện hành động đó, từ đó tìm ra được cấu trúc tổ chức. Từ cấu trúc tổ chức khai phá được từ nhật ký sự kiện sẽ giúp các tổ chức biết được chính xác cơ cấu tổ chức thực tiễn của mình và mối quan hệ giữa các nhân viên, từ đó giúp quá trình quản lý tốt hơn. Dựa trên các nghiên cứu của Van der Alast [3,4] và Claudia Sofia da Costa Alves [2] trong khai phá khía cạnh tổ chức, khóa luận trình bày các thuật toán phân cụm được dùng để phân tích và tìm ra cấu trúc tổ chức như K-means hay thuật toán phân cụm phân cấp AHC. Ngoài ra, khóa luận hướng tới việc phân tích sự chồng chéo trong cấu trúc tổ chức, tức là một cá nhân có thể thuộc về nhiều hơn một cụm, nhóm do vậy khóa luận trình bày thuật toán phân cụm có khả năng phát hiện sự chồng chéo trong tổ chức như CONGA và CONGO.

Nội dung của khóa luận được chia thành các chương sau:

Chương 1: Giới thiệu khái quát về bài toán khai phá quá trình và khai phá khía cạnh bổ sung trong khai phá quá trình. Hơn nữa, trong chương này giới thiệu các nội dung của bài toán khai phá khía cạnh tổ chức trong phát hiện quá trình và các độ đo trong phân tích mạng lưới tổ chức.

Chương 2: Chương này trình bày về các thuật toán phân cụm được sử dụng để phát hiện cấu trúc tổ chức như thuật toán phân cụm phân cấp AHC, thuật toán K-means và các thuật toán có khả năng phát hiện sự chồng chéo trong tổ chức như CONGA và CONGO. Ngoài ra trong chương này khóa luận trình bày độ đo mô đun hóa để đánh giá chất lượng phân cụm từ các thuật toán trên.

Chương 3: Chương này trình bày về mô hình giải quyết bài toán phát hiện cấu trúc tổ chức sử dụng các thuật toán phân cụm và các bước thực hiện trong mô hình.

Chương 4: Trình bày giải pháp thực nghiệm trong phát hiện cấu trúc tổ chức sử dụng các thuật toán phân cụm phân cấp AHC, CONGA và thuật toán CONGO và đưa ra đánh giá và so sánh.

Phần kết luận: Tóm lược kết quả đã đạt được và đưa ra định hướng phát triển trong tương lai.

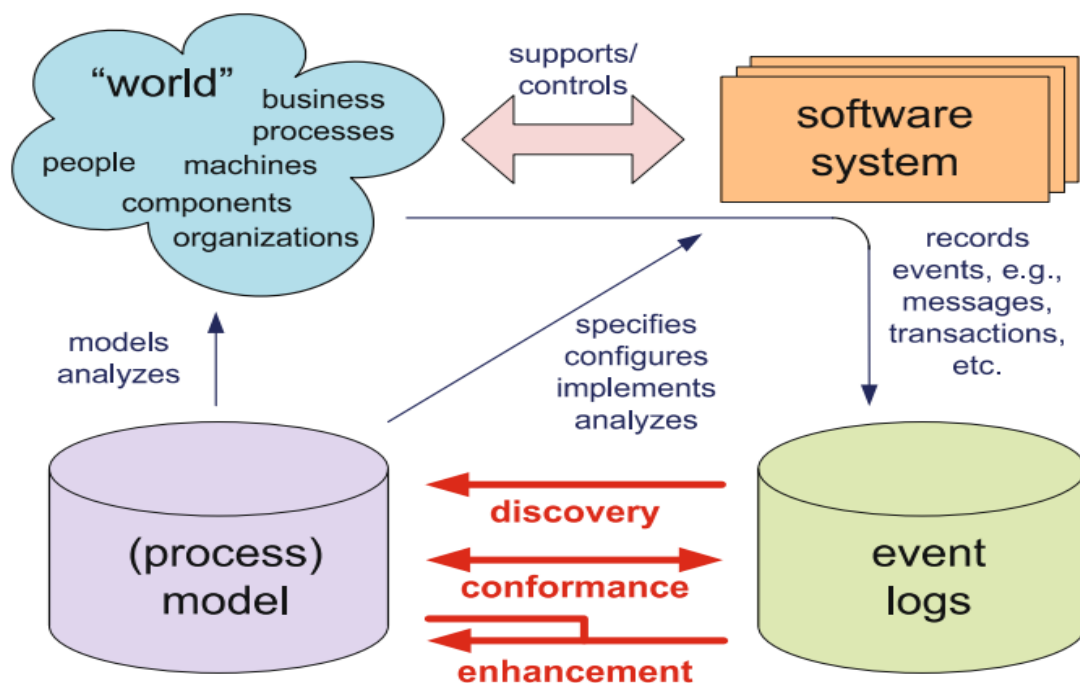
Chương 1. TÔNG QUAN VỀ KHAI PHÁ KHÍA CẠNH TỔ CHỨC TRONG KHAI PHÁ QUÁ TRÌNH

1.1 Khái quát về khai phá quá trình

Trong những năm gần đây, khai phá quá trình nổi lên như là một lĩnh vực nghiên cứu mới tập trung vào phân tích quá trình sử dụng dữ liệu sự kiện. Một trong những nguyên nhân chính khiến khai phá quá trình ngày càng được quan tâm là bởi vì ngày càng có nhiều dữ liệu sự kiện được ghi nhận lại, do đó cung cấp thông tin chi tiết về lịch sử của quá trình. Các kỹ thuật khai phá quá trình nhằm mục *đích phát hiện, giám sát và cải thiện* các quá trình thực tế bằng cách trích lọc tri thức từ các nhật ký sự kiện đã có sẵn trong các hệ thống thông tin ngày nay [4]. Các ứng dụng của khai phá quá trình đã được áp dụng vào rất nhiều miền ứng dụng khác nhau, trong đó nổi bật nhất là *quản lý quá trình kinh doanh*.

Điểm xuất phát cho mọi công việc khai phá quá trình là một nhật ký sự kiện. Nhật ký sự kiện là kết quả ghi nhận lại của hệ thống khi có một ngọvời dùng nào đó tương tác với hệ thống. Một nhật ký sự kiện bao gồm nhiều trường hợp. Mỗi trường hợp gồm nhiều các sự kiện xảy ra nối tiếp nhau. Các sự kiện trong trường hợp khác nhau có thể xảy ra xen kẽ nhau. Trường hợp và sự kiện có nhiều thuộc tính.

Theo van der Aalst [3] khai phá quá trình bao gồm 3 bài toán: phát hiện quá trình, kiểm tra sự phù hợp và tăng cường mô hình quá trình. Hình dưới đây mô tả vị trí của 3 bài toán trong khai phá quá trình.



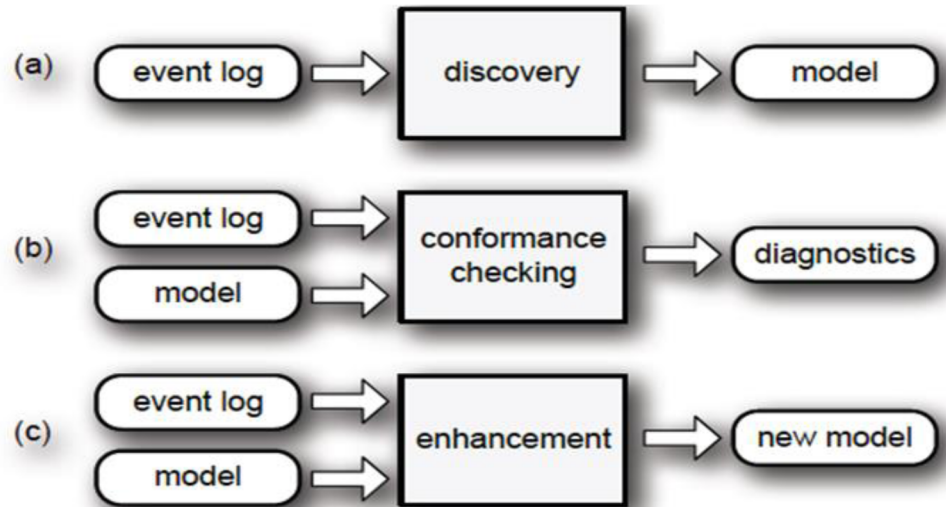
Hình 1.1 Ngữ cảnh của khai phá quá trình [3]

Bài toán đầu tiên là “**phát hiện quá trình**”: Đây là kỹ thuật phát hiện quá trình từ nhật ký sự kiện và xây dựng lên một mô hình quá trình. Đối với các doanh nghiệp, tổ chức đây là một việc rất hữu ích vì có thể phát hiện được quy trình làm việc thực tế chỉ dựa vào những thông tin được lưu lại ở nhật ký sự kiện trong các hệ thống thông tin.

Bài toán thứ hai là “**kiểm tra sự phù hợp**”: Trong bài toán này, một mô hình quá trình đã có sẽ được đem ra so sánh với nhật ký sự kiện tương ứng với nó. Kiểm tra sự phù hợp là kỹ thuật dùng để kiểm tra xem quá trình thực tế của doanh nghiệp, tổ chức được ghi lại trong nhật ký sự kiện và mô hình đã được mô hình hóa có thống nhất, phù hợp hay không từ đó phát hiện những sai lệch, chưa thống nhất để hỗ trợ cho việc cải tiến quá trình kinh doanh.

Bài toán thứ ba là “**tăng cường mô hình**”: Bài toán này hướng tới việc cải tiến hay mở rộng mô hình trước đó. Ví dụ bằng cách khai thác các khía cạnh như thời gian, bị bỏ qua trong bài toán phát hiện quá trình ta có thể biết được thời gian thực hiện giữa các hành động của quy trình, phân tích các tắc nghẽn, mức độ phục vụ từ đó mở rộng hay cải tiến mô hình cho phù hợp hơn.

Hình dưới đây mô tả theo đầu vào và đầu ra của 3 bài toán trong khai phá quá trình.



Hình 1.2 Ba bài toán trong khai phá quá trình [3]

1.2 Một số khái niệm

1.2.1 Nhật ký sự kiện

Như đã trình bày ở trên, đầu vào của bài toán khai phá quá trình là nhật ký sự kiện. Một nhật ký sự kiện là kết quả ghi nhận lại của hệ thống khi có một người nào đó tương tác với hệ thống. Bảng 1.1 đưa ra một ví dụ về một nhật ký sự kiện được trình bày trong [3] với các thông tin điển hình sử dụng cho khai phá quá trình với các trường hợp. Mỗi trường hợp (case id) chính là một phiên làm việc của người dùng được ghi lại bởi hệ thống. Trong mỗi trường hợp gồm nhiều sự kiện xảy ra nối tiếp nhau. Ngoài ra bảng 1.1 còn đưa thêm các thông tin cho mỗi sự kiện. Ví dụ mỗi sự kiện đều có thêm thuộc tính thời gian (timestamp), thông tin này rất hữu ích trong việc phân tích hiệu năng giữa các thuộc tính như thời gian chờ giữa hai hành động. Ngoài ra còn có các thuộc tính khác như nguồn (resource), chi phí của mỗi hành động (cost) ...

Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010:11.02	Register request	Pete	50	...
	35654424	31-12-2010:10.06	Examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	Check ticket	Mike	100	...
	35654426	06-01-2011:11.18	Decide	Sara	200	...
	35654427	07-01-2011:14.24	Reject request	Pete	200	...
2	35654483	30-12-2010:11.32	Register request	Mike	50	...
	35654485	30-12-2010:12.12	Check ticket	Mike	100	...
	35654487	30-12-2010:14.16	Examine casually	Pete	400	...
	35654488	05-01-2011:11.22	Decide	Sara	200	...
	35654489	08-01-2011:12.05	Pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	Register request	Pete	50	...
	35654522	30-12-2010:15.06	Examine casually	Mike	400	...
	35654524	30-12-2010:16.34	Check ticket	Ellen	100	...
	35654525	06-01-2011:09.18	Decide	Sara	200	...
	35654526	06-01-2011:12.18	Reinitiate request	Sara	200	...
	35654527	06-01-2011:13.06	Examine thoroughly	Sean	400	...
	35654530	08-01-2011:11.43	Check ticket	Pete	100	...
	35654531	09-01-2011:09.55	Decide	Sara	200	...
	35654533	15-01-2011:10.45	Pay compensation	Ellen	200	...
4	35654641	06-01-2011:15.02	Register request	Pete	50	...
	35654643	07-01-2011:12.06	Check ticket	Mike	100	...
	35654644	08-01-2011:14.43	Examine thoroughly	Sean	400	...
	35654645	09-01-2011:12.02	Decide	Sara	200	...
	35654647	12-01-2011:15.44	Reject request	Ellen	200	...
...

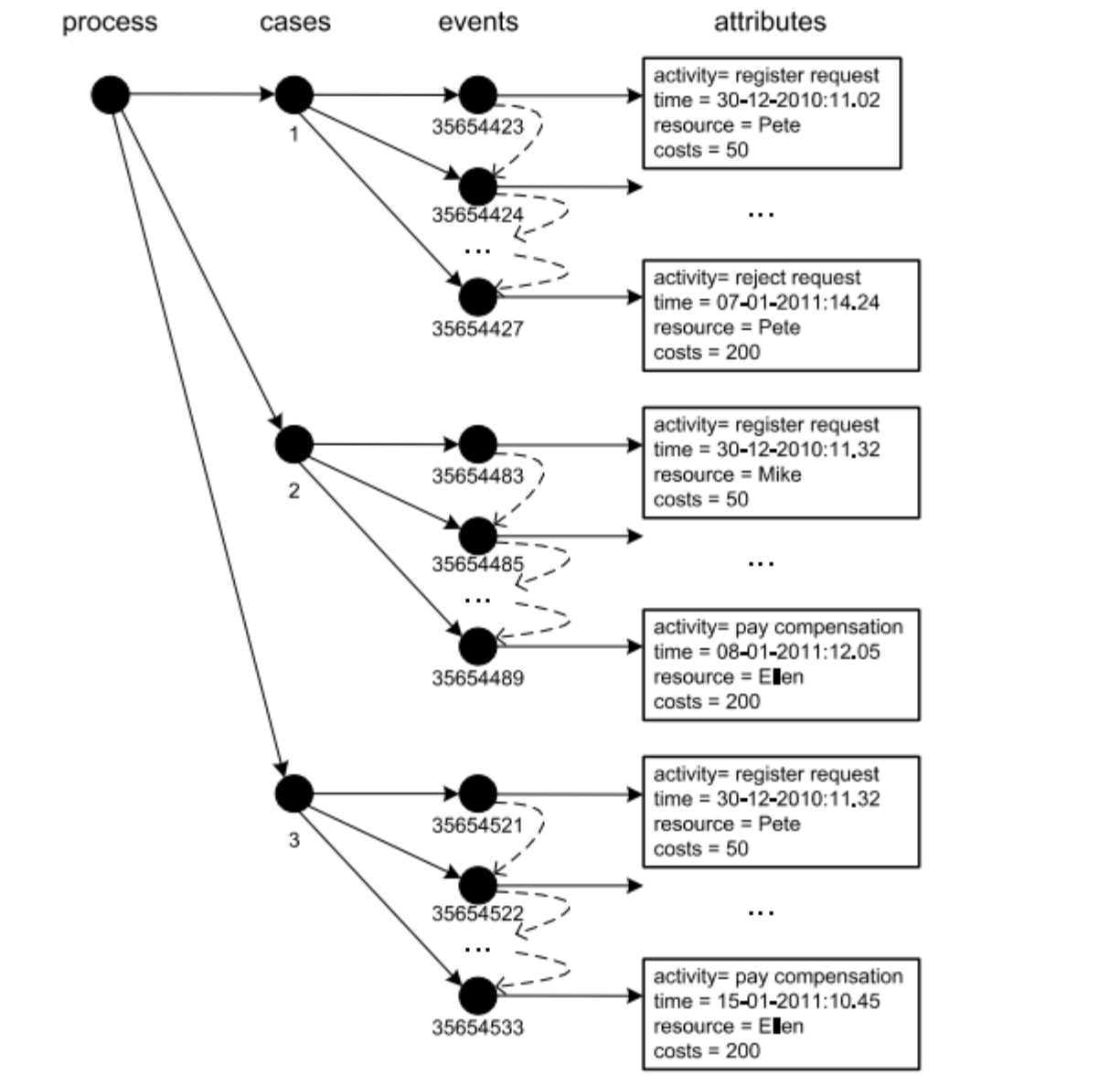
Bảng 1.1 Cấu trúc của một nhật ký sự kiện [3]

Từ bảng 1.1, Van der Aalst [3] đưa ra các giả định về nhật ký sự kiện:

- Mỗi một nhật ký sự kiện thì bao gồm nhiều trường hợp.
- Mỗi một trường hợp bao gồm nhiều sự kiện mà mỗi sự kiện liên quan đến một trường hợp.
- Sự kiện trong một trường hợp diễn ra theo thứ tự.

- Các sự kiện có nhiều thuộc tính như hành động (activity), thời gian (time), chi phí (cost), người thực hiện hành động (resource).

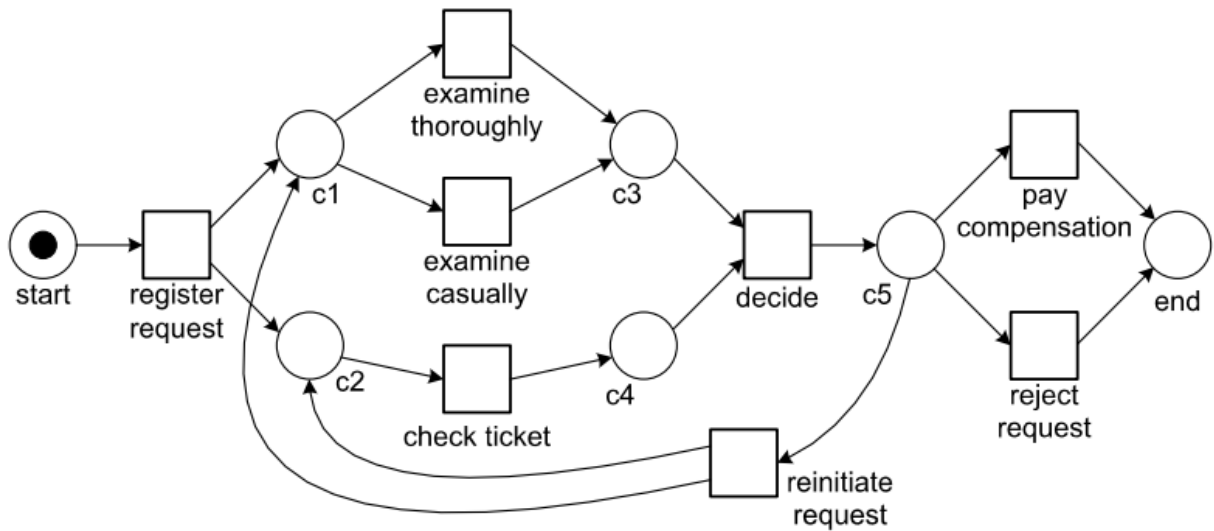
Không phải tất cả các sự kiện đều có cùng tập thuộc tính nhưng thường thì các sự kiện trong cùng một hành động thì có cùng tập các thuộc tính.



Hình 1.3 Cấu trúc của nhật ký sự kiện [3]

1.2.2 Lưới Petri

Lưới Petri [10] là một ngôn ngữ mô hình hóa cho phép biểu diễn mô hình quá trình. Lưới Petri là một đồ thị có hướng bao gồm các vị trí (place) và các thanh chuyển (transition). Trong đó thanh chuyển là một thành phần trong mô hình, liên quan đến một tác vụ hay một hành động cụ thể nào đó có thể được thực thi trong quá trình và được biểu diễn bởi hình chữ nhật. Vị trí được biểu diễn bởi vòng tròn và có thể giữ một hay nhiều thẻ (thẻ được biểu diễn bởi chấm đen). Lưới Petri có cấu trúc tĩnh nhưng các thẻ (tokens) có thể lưu thông trong mạng lưới nhờ sự điều khiển của luật cháy. Trạng thái của lưới petri được xác định bởi sự phân bố của các thẻ ở các vị trí gọi là dấu (marking) của vị trí đó. Hình 1.4 đưa ra ví dụ về một lưới petri.



Hình 1.4 Ví dụ về lưới Petri [3]

Theo van der Alast [3], ta có định nghĩa về lưới Petri. *Lưới Petri là một bộ ba $N = (P, T, F)$ trong đó P là một tập hữu hạn các vị trí (place), T là tập hữu hạn các thanh chuyển (transition) sao cho $P \cap T = \emptyset$ và $F \subseteq (P \times T) \cup (T \times P)$ là tập các cạnh có hướng hay còn được gọi là luồng quan hệ. Lưới Petri được đánh dấu là một cặp (N, M) trong đó $N = (P, T, F)$ là một lưới Petri và $M \in \mathcal{B}(P)$ là một đa tập (multi-set) trên P biểu thị dấu (marking) của lưới. Tập hợp tất cả các lưới Petri được đánh dấu được ký hiệu là \mathcal{N} .*

1.3 Khai phá khía cạnh bổ sung trong phát hiện quá trình

Bài toán khai phá quá trình tập trung chính vào khía cạnh luồng điều khiển, tức là thứ tự của các hành động. Ngoài ra nhật ký sự kiện còn bao gồm các thông tin qua trọng liên quan đến các khía cạnh khác như khía cạnh tổ chức, khía cạnh thời gian, khía cạnh quyết định.

- Khai phá khía cạnh thời gian: khía cạnh thời gian có liên quan đến việc điều phối thời gian và tần suất của sự kiện. Trong hầu hết các nhật ký sự kiện thì mỗi nhật ký sự kiện có một nhãn thời gian (timestamp). Các định dạng thời gian có nhiều loại và mức độ chi tiết, có nhật ký thì chỉ ghi thông tin về ngày, ví dụ như “20-2-2014” nhưng có những nhật ký sự kiện độ chính xác đến mili-giây. Độ chính xác của thời gian cho chúng ta phát hiện được những bế tắc trong quy trình, phân tích được mức độ phục vụ, quan sát được sử dụng nguồn và dự đoán thời gian xử lý còn lại trong các trường hợp đang được thực hiện.
- Khai phá khía cạnh quyết định: khía cạnh quyết định tập trung vào các thuộc tính của trường hợp. Mỗi trường hợp đều được xác định bằng các thuộc tính của nó và các thuộc tính của các sự kiện xảy ra trong nó.
- Khai phá khía cạnh tổ chức: khai phá khía cạnh tổ chức là đi phát hiện sự tổ chức giữa các nguồn sinh ra sự kiện. Thông tin cần khai thác ở đây là thuộc tính nguồn (resource) chính là các tác nhân thực hiện hành động.

1.4 Khai phá khía cạnh tổ chức trong khai phá quá trình

1.4.1 Khai phá mô hình tổ chức

Khai phá mạng xã hội

Theo Claudia Sofia da Costa Alves [2], ý tưởng chính của kỹ thuật này là giám sát mỗi thực thể quá trình được định tuyến giữa các tác nhân (actor). Kỹ thuật này cung cấp 5 độ đo cho phép xây dựng lên mạng xã hội:

- **Chuyển giao công việc** (*Handover of work metric*): Độ đo này xác định xem ai chuyển giao công việc cho ai bằng việc trích lọc từ nhật ký sự kiện theo thứ tự thực hiện công việc trong từng trường hợp, trong đó hành động đầu tiên được hoàn thành bởi một cá thể nào đó, sau đó quy trình được tiếp tục với hành động tiếp theo và được hoàn thành, cứ như vậy một trường hợp được hoàn thành với sự chuyển giao công việc giữa các cá thể.
- **Hợp đồng con** (*Subcontracting metric*): Độ đo này tương tự với độ đo chuyển giao công việc (handover of work metric), tuy nhiên trong chuyển giao công việc thì mối quan hệ giữa hai cá thể là một chiều thì trong độ đo này mối quan hệ giữa hai cá thể là hai chiều. Ví dụ cá nhân A có hợp đồng con với cá nhân B khi giữa 2 hành động thực hiện bởi A có một hành động được thực hiện bởi B.
- **Làm việc cùng nhau** (*Working together metri*): Hai cá nhân A và B làm việc cùng nhau khi họ thực hiện các hành động trong cùng một trường hợp. Độ đo này đơn giản chỉ đếm số lượng các trường hợp mà 2 cá nhân làm việc cùng nhau.
- **Nhiệm vụ giống nhau** (*Similar task metric*): kỹ thuật này tập trung vào hành động chung, mục tiêu của kỹ thuật này là xác định xem các cá thể thực hiện các hành động giống nhau trong nhật ký sự kiện. Để thực hiện kỹ thuật này, mỗi một cá thể sẽ được thống kê số lần thực hiện các hành động cụ thể, sau đó các cá thể được so sánh với nhau để tìm ra sự tương đồng.
- **Ủy thác** (*Reassignment metric*): Kỹ thuật này phát hiện sự ủy thác hành động từ cá nhân này đến cá nhân khác. Ví dụ như nếu cá thể A thường ủy thác công việc cho cá thể B và không có việc B ủy thác công việc cho A thì có thể A là cấp trên của B.

Khai phá tổ chức

Kỹ thuật này làm việc ở mức độ cao hơn mạng xã hội. Trong khi khai phá mạng xã hội làm việc ở mức độ cá thể thì khai phá tổ chức làm việc ở mức độ nhóm, đội, bộ phận.

Claudia Sofia da Costa Alves [2] đã đưa ra một số kỹ thuật tiêu biểu trong khai phá tổ chức để xây dựng lên mạng lưới tổ chức:

- **Khai phá mặc định (Default Miner):** Kỹ thuật này chỉ đơn giản là tìm ra mối quan hệ giữa các hành động và người thực hiện.
- **Thực hiện các nhiệm vụ giống nhau (Doing Similar Tasks):** Kỹ thuật này gộp tất cả những người thực hiện hành động giống nhau vào chung một nhóm.
- **Khai phá cụm phân cấp (Hierarchical Mining Clustering):** Kỹ thuật này đưa ra một mô hình phân cấp. Nó cài đặt thuật toán phân cụm từ dưới lên dựa vào các hành động chung, tức là cụm được xác định dựa vào các hành động mà các cá nhân thực hiện.
- **Làm việc cùng nhau (Working together):** Đối lập với các độ đo trên, độ đo này dựa trên các trường hợp chung, không phải dựa trên các hành động chung. Kỹ thuật này giúp xác định các đội vì nó đặt các cá thể thực hiện chung trong một trường hợp vào một nhóm.

1.4.2 Phân tích mạng xã hội

Sau khi đã xây dựng được các biểu đồ mạng xã hội từ các độ đo trên, mạng xã hội sẽ được tiếp tục phân tích dựa vào các độ đo dưới đây.

Độ đo mức độ cá thể:

- **Bậc (Degree):** Bậc của một nút là số lượng các liên kết đến nút đó.
- **Độ trung tâm trung gian (Betweenness Centrality):** tính sự ảnh hưởng của một node đối với sự lan truyền thông tin trong mạng. Một node có betweenness centrality càng lớn thì nó có vai trò quan trọng trong mạng xã hội bởi có thể nó là cầu nối liên kết giữa 2 nhóm với nhau và nếu nó bị mất thì việc trao đổi thông tin giữa hai nhóm sẽ không thể thực hiện được.

- **Độ Trung tâm gần gũi (Closeness Centrality):** Độ gần gũi của mỗi node với các nút khác trong mạng. Nếu một nút có độ gần gũi càng thấp thì dễ giao tiếp với các nút khác nó muốn thì phải đi qua rất nhiều nút khác trong mạng.
- **Vector đặc trưng trung tâm (Eigenvector Centrality):** độ đo tương tự như bậc, tuy nhiên thay vì đếm số lượng liên kết đến node đó thì nó quan tâm đến bậc của node mà liên kết với nó. Trong một mạng, khi 2 node có cùng bậc thì đại lượng này sẽ cho biết node nào liên kết với các node quan trọng hơn ở trong mạng.
- **Hệ số phân cụm (Clustering Coefficient):** chính là số lượng liên kết trong một khu vực nút trên tổng số liên kết có thể. Ví dụ như khu vực có 4 nút mà có đến 8 liên kết thì chứng tỏ rất có thể chúng cùng trong một cụm.

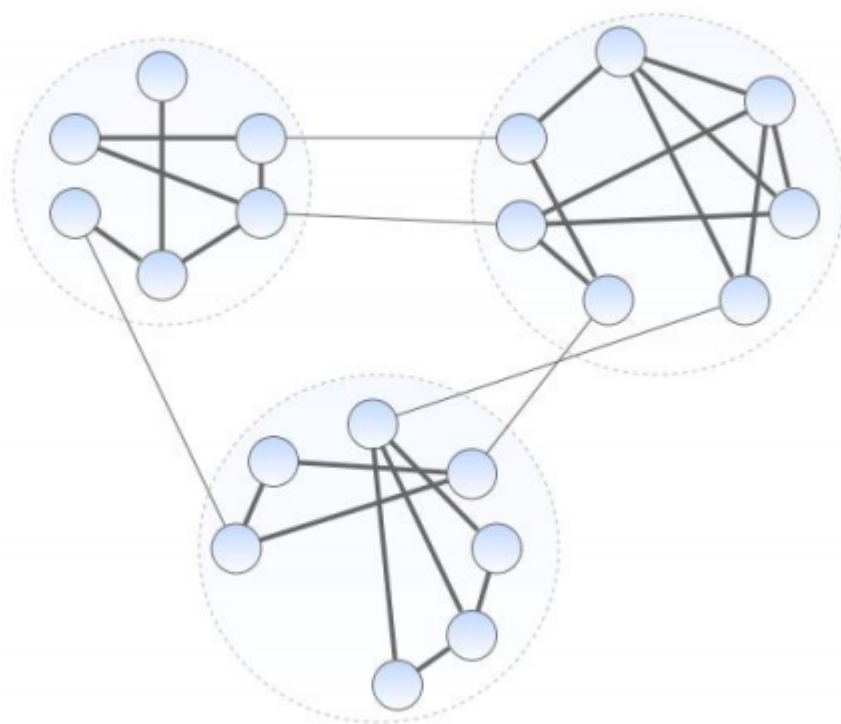
Độ đo mức độ mạng lưới:

- **Mật độ (Density):** Trong một mạng nếu có mật độ dày đặc thì chứng tỏ các node gần đều liên kết với nhau.
- **Hệ số (Coefficient):** độ đo này tính xác suất một mạng lưới có thể bị phân hoạch thành một số nhỏ hơn.
- **Độ tập trung (Centralization):** đây chính là độ tập trung của mạng. Nếu nó càng cao tức là mạng lưới bị kiểm soát bởi một hoặc một vài người. Trong trường hợp những người này bị loại khỏi mạng thì mạng lưới có thể bị bẻ gãy thành các mạng lưới nhỏ hơn.

1.4.3 Tìm cấu trúc cộng đồng trong mạng xã hội

Trong một mạng xã hội, độ mạnh của một liên kết giữa hai cá thể có thể trong vùng từ yếu đến mạnh, điều này phụ thuộc vào chất lượng, số lượng và tần số của sự thay đổi giữa các tác nhân. Liên kết mạnh được đặc tả như là việc tăng số lần trao đổi và khởi tạo liên kết giữa các cá thể. Ví dụ liên kết mạnh giữa hai người bạn thân, các thành viên trong

gia đình, nhân viên trong một tổ chức. Ngược lại, trong liên kết yếu nhiều giới hạn về thời gian và độ mật thiết được ngầm định, ví dụ như trong các tổ chức khác nhau thì thường tồn tại liên kết yếu.



Hình 1.5 Mạng lưới với cấu trúc cộng đồng [2]

Trong hình 1.5 mạng lưới được cấu thành nên từ 3 nhóm, trong mỗi nhóm chứa các cá thể. Các cá thể trong một nhóm thì có quan hệ với nhau mật thiết hơn nên có liên kết mạnh giữa các cá thể đó, được minh họa bởi liên kết đậm. Giữa các nhóm với nhau có quan hệ yếu hơn, liên kết yếu hơn nên được minh họa bởi các liên kết mờ hơn.

1.5 Tóm tắt chương 1

Trong chương này khóa luận đã trình bày nội dung tổng quan về bài toán khai phá quá trình cũng như bài toán khai phá khía cạnh tổ chức trong phát hiện quá trình, đưa ra các khái niệm cơ bản thường được dùng trong bài toán. Ngoài ra, trong chương này khóa luận cũng trình bày các độ đo để xây dựng nên mô hình mạng xã hội tổ chức, mô hình này sẽ là đầu vào cho các thuật toán phân cụm trong chương 2 để tìm ra cấu trúc tổ chức.

Chương 2. CÁC THUẬT TOÁN PHÂN CỤM TRONG KHAI PHÁ KHÍA CẠNH TỔ CHỨC

Theo van der Aalst [3] các kỹ thuật phân cụm như k-means hay AHC (agglomerative hierarchical clustering) có thể sử dụng để phát hiện mô hình tổ chức. Trong chương này, khóa luận sẽ đi sâu vào các thuật toán phân cụm để phát hiện ra mô hình tổ chức. Cũng theo Claudia Sofia da Costa Alves [2] hầu hết các phương pháp tiếp cận gần đây đều dựa trên phân cụm phân cấp, tuy nhiên mỗi cách tiếp cận đều cố gắng cải thiện để phù hợp với các độ đo phân tích mạng xã hội đã được đề cập trong chương 1. Ví dụ như thuật toán Girvan và Newman là một trong những cách tiếp cận có thể tìm kiếm hầu hết các cộng đồng tương tự nhau trong việc so sánh với các cấu trúc cộng đồng thực tế và nó mang lại kết quả với sự hài lòng tốt nhất. Trong chương này khóa luận sẽ trình bày các thuật toán được dùng để phát hiện cấu trúc tổ chức như thuật toán phân cụm AHC hay K-means và thuật toán cải tiến của Girvan và Newman trong phát hiện chồng chéo cộng đồng là CONGA, một cải tiến của CONGA là thuật toán CONGO.

2.1 Độ đo mô đun hóa để đánh giá chất lượng phân chia đồ thị

Khái niệm mô đun hóa (modularity) được đưa ra bởi Girvan và Newman [2] nhằm trở thành độ đo đánh giá chất lượng của việc chia đồ thị. Giả thiết rằng một mạng lưới có N đỉnh liên kết với nhau bởi m liên kết (cạnh). Đặt A_{ij} là một phần tử của ma trận kề đồ thị, tức là số lượng cạnh giữa đỉnh i và đỉnh j . Giả sử rằng có một phép chia các đỉnh thành c cụm ta sẽ tính được giá trị mô đun hóa của phép chia đó.

Giá trị mô đun hóa được tính như sau:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

Trong đó:

- k_i là bậc của đỉnh i và k_j là bậc của đỉnh j
- $\delta(c_i, c_j) = 1$ nếu i và j cùng một cụm và bằng 0 nếu i và j khác cụm.

Từ phép chia đồ thị này ta sẽ tính được số lượng cụm phân chia tốt nhất, tức là làm cho giá trị mô đun hóa cực đại.

Ví dụ về tính giá trị mô đun hóa của mạng lưới trong [2], dưới đây là ví dụ bảng ma trận kề theo độ đo làm việc cùng nhau (working together) giữa các nhân viên:

	Fred	Howard	John	Linda	Mona	Robert	Vincent
Fred	0	2	2	0	2	3	0
Howard	2	0	1	1	2	2	1
John	2	1	0	1	1	3	0
Linda	0	1	1	0	2	2	2
Mona	2	2	1	2	0	3	2
Robert	3	2	3	2	3	0	1
Vincent	0	1	0	2	2	1	0

Bảng 2.1 Bảng ma trận kề giữa các cá nhân theo độ đo làm việc cùng nhau [2]

Tiếp đó là bậc của các nhân viên, với mỗi nhân viên là một nút trong mạng:

Cá nhân	Bậc
Fred	9
Howard	9
John	8
Linda	8
Mona	12
Robert	14
Vincent	6

Bảng 2.2 Bảng thể hiện bậc của nhân viên [2]

Các cụm giả định hiện tại:

Cụm 0: Fred, John và Robert

Cụm 1: Howard và Mona

Cụm 2: Linda và Vincent

Khi đó giá trị mô đun hóa được tính như sau:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

$$\begin{aligned}
&= [(A_{02} - \frac{k_0 k_2}{66}) + (A_{20} - \frac{k_2 k_0}{66}) + (A_{05} - \frac{k_0 k_5}{66}) + (A_{50} - \frac{k_5 k_0}{66}) + (A_{25} - \frac{k_2 k_5}{66}) \\
&+ (A_{52} - \frac{k_5 k_2}{66}) + (A_{36} - \frac{k_3 k_6}{66}) + (A_{63} - \frac{k_6 k_3}{66}) + (A_{45} - \frac{k_4 k_5}{66}) + (A_{54} - \frac{k_5 k_4}{66})] \\
&= [(2 - \frac{9*8}{66}) * 2 + (3 - \frac{9*14}{66}) * 2 + (3 - \frac{8*14}{66}) * 2 + (2 - \frac{8*6}{66}) * 2 + (2 - \frac{9*12}{66}) * \\
&2] \\
&= 0.1827364545879942
\end{aligned}$$

Độ đo mô đun hóa do Girvan và Newman đề xuất ở trên cho kết quả tốt đối với các đồ thị không xuất hiện sự chồng chéo, tức là một nút chỉ có thể thuộc về một cụm. Do vậy, để giải quyết vấn đề một nút có thể thuộc về nhiều hơn một cụm, đánh giá chất lượng của các cụm được phân chia ra thì giá trị mô đun hóa cho mạng với sự xuất hiện chồng chéo cộng đồng cũng được tính toán và được đưa ra trong [11]. Theo tài liệu thì Chen (2010) đã đề xuất lựa chọn một nút với độ mạnh lớn nhất dựa trên hai đại lượng $B(u,c)$ (được gọi là thuộc đỉnh) và giá trị mô đun hóa cho đồ thị có trọng số Q_{ov} được định nghĩa:

$$Q_{ov} = \frac{1}{2m} \sum_c \sum_{i,j \in V} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \beta_{ic} \beta_{jc}$$

Trong đó $\beta_{ic} = k_{ic} / \sum_{c'} k_{ic'}$ là độ mạnh mà nút i thuộc về cụm c và $k_{ic} = \sum_{j \in c} w_{ij}$ là tổng trọng số của các đường liên kết từ nút i trong cụm c . $B(u,c)$ độ đo cách mà một nút u kết nối chặt chẽ với cộng đồng c so với phần còn lại của mạng. Cho hai ngưỡng B^U và B^L , khi mở rộng một cộng đồng c thì các nút lân cận thỏa mãn $B(u,c) > B^U$ sẽ được thêm vào c . Đối với các nút thỏa mãn $B^L \leq B(u,c) \leq B^U$ nếu giá trị Q_{ov} tăng khi thêm một nút như vậy thì u được thêm vào cộng đồng c . Hạn chế của cách tính này là việc lựa chọn tùy ý giá trị ngưỡng B^U và B^L và chi phí trong tính toán với độ phức tạp $O(kn^2)$ với k là số lượng cộng đồng.

Ngoài ra theo [11] Nicosia (2009) cũng đề xuất độ đo mô đun hóa dựa trên liên kết trong việc phân chia đồ thị có chồng chéo. Độ đo này được xây dựng theo hệ số của các liên kết. Cho liên kết (i, j) kết nối i tới j trong cộng đồng c là $\beta_{l(i,j),c} = F(a_{ic}, a_{jc})$ thì hệ

số liên kết mong đợi của bất kỳ liên kết có thể $l(i,j)$ từ nút i đến nút j trong cộng đồng c có thể định nghĩa là:

$$\beta_{l(i,j),c}^{out} = \frac{1}{|V|} \sum_{j \in V} F(a_{ic}, a_{jc})$$

Theo đó hệ số mong đợi của bất kỳ liên kết $l(i,j)$ trở đến nút j trong cộng đồng c được định nghĩa là:

$$\beta_{l(i,j),c}^{in} = \frac{1}{|V|} \sum_{i \in V} F(a_{ic}, a_{jc}).$$

Các hệ số được đưa ra trước đó được sử dụng như là trọng số cho xác suất của một liên kết có thể quan sát và xác suất của một liên kết bắt đầu từ i tới j trong mô hình rỗng (null model). Kết quả của độ đo mô đun hóa mới được định nghĩa:

$$Q_{ov}^{Ni} = \frac{1}{m} \sum_c \sum_{i,j \in V} \left[\beta_{l(i,j),c} A_{i,j} - \beta_{l(i,j),c}^{out} \beta_{l(i,j),c}^{in} \frac{k_i^{out} k_j^{in}}{m} \right]$$

Trong đó m là tổng số lượng các cạnh, $k_i^{out(in)}$ là số lượng liên kết đi ra ngoài (đi vào trong) của nút i . Một điểm chú ý là Q_{ov}^{Ni} phụ thuộc vào hệ số liên kết $F(a_{ic}, a_{jc})$, nó có thể là sản phẩm, trung bình, giá trị lớn nhất của a_{ic} và a_{jc} .

Các độ đo mô đun hóa trong việc đánh giá chất lượng phân chia đồ thị sẽ được dùng để đánh giá các đồ thị được phân chia bởi các thuật toán phân cụm được trình bày trong các tiêu mục dưới đây.

2.2 Thuật toán phân cụm phân cấp AHC (Agglomerative Hierarchical Clustering)

Theo Claudia Sofia da Costa Alves[2] trong phương pháp tiếp cận xã hội học, các thuật toán phân cấp có thể được tích tụ (từ dưới lên) hoặc chia ra (từ trên xuống). Các giải thuật tích tụ (Agglomerative algorithms) bắt đầu với mỗi một phần tử trong mạng là một cụm và kết quả là một cụm duy nhất cho tất cả các phần tử trong mạng. Trong giải thuật tích tụ mỗi một lần lặp sẽ gộp hai cụm đã có, do vậy sẽ có ít nhất là một cụm. Giải thuật

phân chia thì làm việc theo cách ngược lại, nó bắt đầu bởi một cụm bao gồm tất cả các phần tử của mạng và kết thúc là mỗi một phần tử cho mỗi cụm. Trong giải thuật này thì sau mỗi lần lặp thì một cụm sẽ được chia làm hai do vậy sau khi kết thúc sẽ được kết quả là mỗi một phần tử trong mạng sẽ thuộc một cụm. Rất nhiều mô hình tổ chức là phân cấp do vậy việc tìm kiếm cộng đồng trong mạng với giải thuật tích tụ (agglomerative) được sử dụng nhiều hơn giải thuật phân chia (divide algorithms).

Sau đây là thuật toán phân cụm phân cấp cơ bản AHC được tác giả đề cập trong [2]:

Cho mạng lưới gồm N nút.

Bước 1: Mỗi nút được coi như là một cụm, tức là N cụm gồm một phần tử.

Bước 2: Tìm cặp cụm gần nhau nhất và gộp chung chúng thành một cụm. Tính lại khoảng cách giữa cụm mới với các cụm cũ.

Bước 3: Lặp lại bước 2 cho đến khi tất cả các phần tử đã được gộp lại thành một cụm duy nhất N phần tử hoặc đã đạt số lượng cụm yêu cầu.

Thuật toán gặp phải các vấn đề sau:

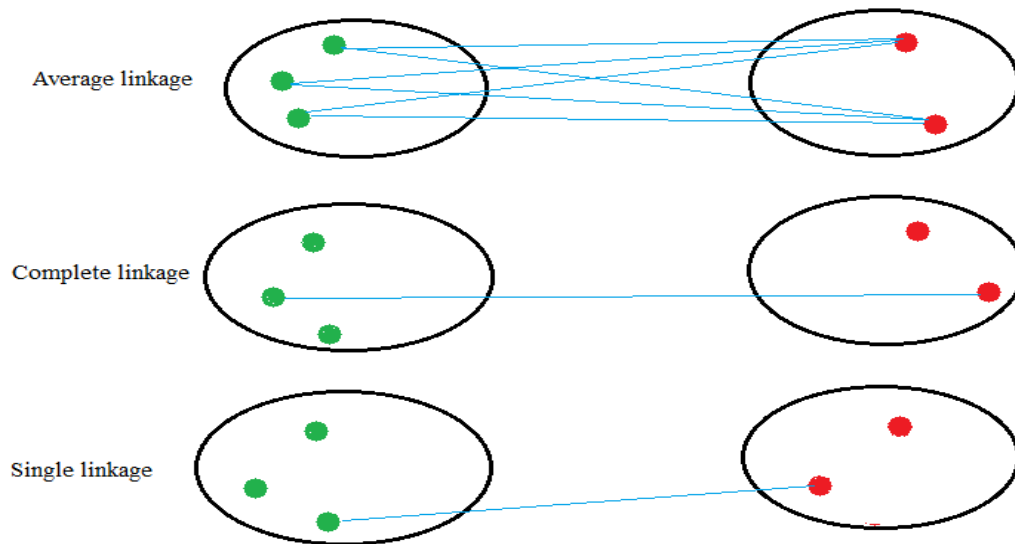
Vấn đề thứ nhất: Nếu mỗi cụm chỉ gồm một phần tử thì lúc này khoảng cách giữa các cụm đơn giản chính là khoảng cách giữa các nút. Nhưng khi thuật toán đã sang lần lặp tiếp theo khi mỗi cụm có thể có nhiều phần tử, ta cần tìm cách tính độ tương tự giữa các cụm theo độ tương tự của các phần tử trong nó. Để giải quyết vấn đề này, có 3 cách tính khoảng cách giữa 2 cụm:

Liên kết đơn (Single linkage): Khoảng cách giữa hai cụm được tính bằng khoảng cách giữa cặp phần tử gần nhất của chúng.

Liên kết hoàn chỉnh (Complete linkage): Khoảng cách giữa hai cụm được tính bằng khoảng cách giữa các cặp phần tử xa nhất của chúng.

Liên kết trung bình (Average linkage): Khoảng cách giữa 2 cụm được tính bằng trung bình tất cả các khoảng cách của các phần tử giữa hai cụm.

Hình dưới đây mô tả các cách tính khoảng cách giữa 2 cụm:



Hình 2.1 Cách tính khoảng cách giữa hai cụm

Vấn đề thứ hai: Ở bước 2 của thuật toán, nếu ta tìm được nhiều hơn một cặp cụm thỏa mãn có độ tương tự bằng nhau và lớn nhất. Ta có thể giải quyết vấn đề này bằng cách chọn ngẫu nhiên một cặp cụm trong số này. Tuy nhiên cách giải quyết này chưa tối ưu và mang lại độ chính xác không cao.

Vấn đề thứ 3: Thuật toán này không đưa ra được một con số chính xác rằng mạng lưới nên chia thành bao nhiêu cụm là hợp nhất. Vấn đề này được giải quyết bằng sử dụng giá trị mô đun hóa được đề cập trước đó, số cụm được phân chia hợp lý nhất khi giá trị modularity lớn nhất.

2.3 Thuật toán phân cụm K-means

Theo [7] thuật toán K-means là một trong những phương pháp phân cụm cổ điển nhất và cũng là quan trọng nhất trong số những thuật toán phân cụm phẳng hay phân cụm phân hoạch. Thuật toán được J. MacQueen phát triển từ năm 1967, sau đó được J. A. Hartigan và M. A. Wong đã phát triển hoàn thiện hơn. Ngày nay thì thuật toán đã được cải tiến để có chất lượng tốt hơn và phù hợp với nhiều kiểu dữ liệu như đặc biệt với kiểu dữ liệu tập mờ (fuzzy data).

Mô tả thuật toán k-means với tập dữ liệu đầu vào D:

Các bước phân cụm trong thuật toán:

Bước 1: Chọn ngẫu nhiên k trong tập dữ liệu D làm trọng tâm cho các cụm

Bước 2: Phân các phần tử dữ liệu trong tập D vào các cụm dựa vào độ tương đồng của nó với trọng tâm của các cụm, phần tử dữ liệu sẽ được phân vào cụm có độ tương đồng lớn nhất.

Bước 3: Tính lại trọng tâm của các cụm

Bước 4: Nhảy đến bước 2 cho đến khi quá trình hội tụ, tức là không có sự gán lại các phần tử dữ liệu giữa các cụm, hay trọng tâm của các cụm là không đổi.

Trong thuật toán K-means, điểm quan trọng là ở bước 2 khi các phần tử dữ liệu được di chuyển giữa các cụm để làm cực đại hóa độ tương tự giữa các phần tử dữ liệu trong một cụm (hay cực đại hóa độ tương tự trong nội tại một cụm, hay cực tiểu hóa khoảng cách giữa các phần tử dữ liệu trong nội tại một cụm). Độ đo nội tại một cụm được tính bằng công thức:

$$J = \sum_{i=1}^k \sum_{p \in C_i} sim(p, m_i)$$

Trong đó: C_i là cụm thứ i và m_i là trọng tâm của nó và $\text{sim}(p, m_i)$ là độ tương tự giữa p và m_i .

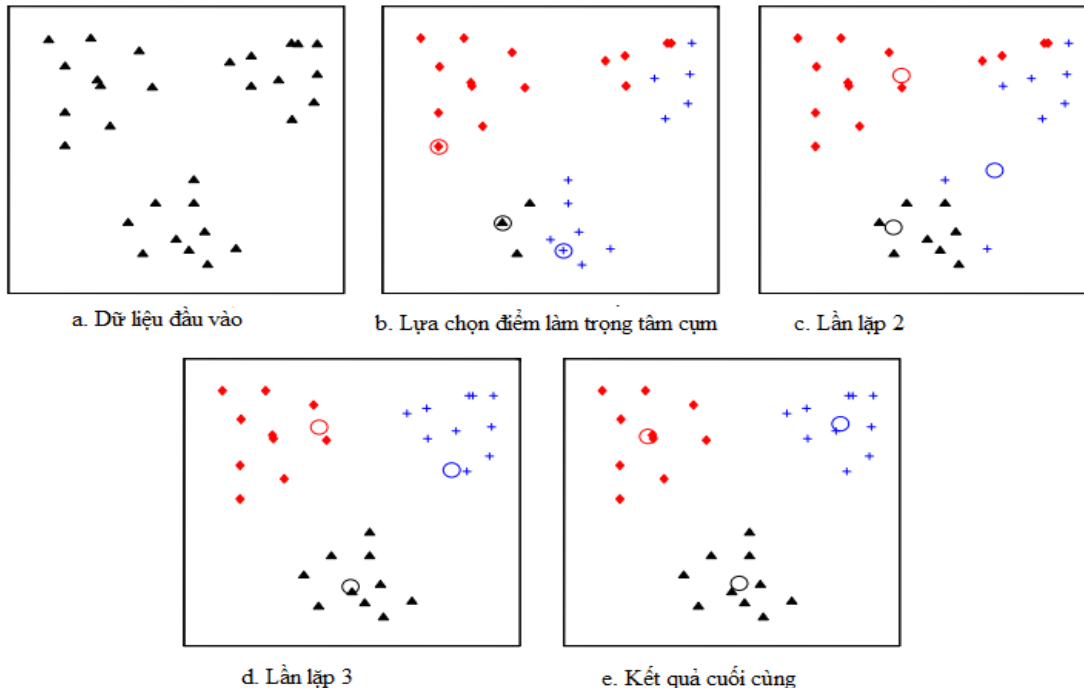
Trọng tâm của m_i của C_i được tính theo công thức:

$$m_c = \sum_{p \in C} \frac{P}{|C|}$$

Thuật toán k-means hoạt động sao cho hàm điều kiện (criterion function), thường hàm hội tụ được chọn là hàm tổng bình phương lỗi [7] được định nghĩa:

$$E = \sum_C \sum_{p \in C} |p - m_c|^2$$

Thuật toán k-means trả về số lượng cụm là tối thiểu nhưng không đảm bảo tìm được giá trị cực đại toàn cục của hàm J nhưng bằng cách chạy lại thuật toán một số lần để thu được giá trị cực đại cục bộ. Hình dưới đây mô tả ví dụ về thuật toán K-means.



Hình 2.2 Ví dụ về thuật toán K-means [7]

Kết quả cuối cùng của k-means phụ thuộc rất nhiều vào cách lựa chọn k phần tử dữ liệu ban đầu làm trọng tâm của k cụm bởi vì k được lựa chọn hoàn toàn ngẫu nhiên, nên kết quả sau mỗi lần chạy thuật toán khác nhau là khác nhau, như vậy ta có thể chạy thuật toán k-means một số lần và chọn lấy lần chạy sao cho kết quả hàm J là lớn nhất.

2.4 Thuật toán CONGA

Các thuật toán được đề cập trước đó như AHC hay K-means không giải quyết được vấn đề chồng chéo tổ chức, một người có thể thuộc về nhiều nhóm hay tổ chức. Ví dụ như cá nhân A vừa làm công việc X và kiêm nhiệm cả công việc Y nên A có thể thuộc về hai nhóm khác nhau. Để giải quyết vấn đề trên, trong khóa luận này sẽ trình bày các thuật toán phân cụm có khả năng phát hiện sự chồng chéo trong tổ chức như CONGA và cải tiến của nó là CONGO.

Thuật toán CONGA được cải tiến từ thuật toán Girvan-Newman [9] nhằm giải quyết bài toán chồng chéo cộng đồng [6], dựa trên thuật toán Girvan-Newman tác giả đề xuất thêm một ý tưởng mới đó là phép chia các đỉnh thành nhiều phần khác nhau, để mỗi phần của đỉnh được chia được xuất hiện trong các cộng đồng con. Vì được cải tiến từ thuật toán Girvan-Newman nên CONGA vẫn sử dụng khái niệm độ đo trung gian của các cạnh. Độ đo trung gian của một cạnh chính là tổng số đường đi giữa các cặp đỉnh trong toàn bộ đồ thị mà đi qua cạnh được xét tới. Một kiểu đường đi hay được sử dụng đó là “đường đi ngắn nhất giữa hai đỉnh”. Đối với đồ thị có trọng số, độ đo trung gian của một cạnh chính là độ đo trung gian của cạnh đó trong đồ thị không có trọng số chia cho trọng số của cạnh đó.

Độ đo trung gian của đỉnh v trong đồ thị là tổng số đường đi ngắn nhất giữa các cặp đỉnh của đồ thị mà đi qua v. Ta có thể dễ dàng tính được độ trung gian của đỉnh $C_B(v)$ từ các độ đo trung gian của cạnh $C_B(e)$:

$$C_B(v) = \frac{1}{2} \sum_{e \in \Gamma(v)} C_B(e) - (n - 1)$$

Trong đó $\Gamma(v)$ là tập các cạnh có v là đỉnh cuối và n là số đỉnh của đồ thị chứa v.

Các bước của thuật toán CONGA:

Bước 1: Tính độ trung gian của tất cả các cạnh trong đồ thị.

Bước 2: Tính độ trung gian của các đỉnh trong đồ thị, dựa vào độ trung gian của các cạnh dựa vào công thức trên.

Bước 3: Tìm danh sách các đỉnh mà độ trung gian của đỉnh đó lớn hơn giá trị lớn nhất của các độ trung gian cạnh.

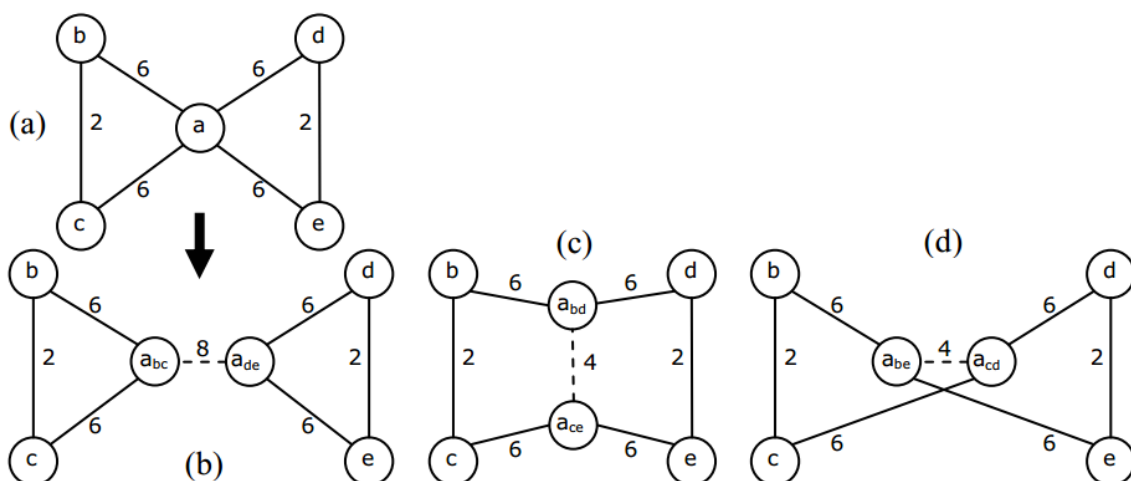
Bước 4: Nếu danh sách ở 3 không rỗng, tính các độ trung gian theo cặp của các đỉnh trong danh sách, sau đó xác định phép phân chia tối ưu nhất cho các đỉnh đó.

Bước 5: Thực hiện việc loại bỏ cạnh hoặc phân chia đỉnh để chia đồ thị thành các phần.

Bước 6: Tính lại độ trung gian của các thành phần được chia ra.

Bước 7: Lặp lại bước 2 cho đến khi không còn cạnh nào

Hình 2.3 mô tả việc phân tách một đỉnh có độ trung gian lớn nhất.



Hình 2.3 (a) đồ thị ban đầu (b) Cách chia tốt nhất của đỉnh a có độ trung gian lớn nhất (c) và (d) Các cách chia khác của đỉnh a. [6]

Thuật toán CONGA giải quyết được bài toán chồng chéo cộng đồng, nhưng nhược điểm của thuật toán là thời gian tính toán, với độ phức tạp tính toán lên đến $O(m^3)$ với m là số cạnh. CONGA chỉ phù hợp xử lý các mạng lưới với một vài nghìn cạnh và đỉnh, với số lượng các cạnh và đỉnh lớn hơn nhiều lần thì không nên sử dụng CONGA. Để giải quyết vấn đề này khóa luận sẽ đề cập đến giải thuật CONGO, được cải tiến từ CONGA.

2.5 Thuật toán CONGO

Ngày nay, dữ liệu trong các hệ thống thông tin ngày càng được mở rộng đặc biệt là về khối lượng, cũng như sự kiện sẽ ngày càng lớn hơn, chứa nhiều thuộc tính hơn. Việc dùng các thuật toán phân cụm như CONGA đối với mạng lưới có hàng triệu node mà hơn nữa giữa các node này lại có rất nhiều các liên kết để phát hiện ra mô hình tổ chức là một việc làm khó khăn và tốn nhiều chi phí.

Từ thực trạng trên, Steve Gregory [1] đã đưa ra cải tiến cho thuật toán CONGA dựa trên dạng cục bộ (local form) của betweenness, theo tác giả thì nó mang lại kết quả tốt và nhanh hơn, đặc biệt hiệu quả trong việc phát hiện các cộng đồng nhỏ trong các mạng lớn với độ phức tạp chỉ $O(n \log n)$ cho các mạng thưa. Steve cũng giới thiệu về độ đo trung gian cục bộ “local betweenness” được sử dụng trong CONGO.

2.5.1 Độ trung gian cục bộ

Tính toán độ trung gian rất tốn kém bởi việc đếm tất cả đường đi ngắn nhất trong mạng lưới. Một cách để tránh điều này thì tác giả chỉ đếm duy nhất các đường đi ngắn nhất (short shortest paths). Trong trường hợp này, độ đo trung gian của cạnh e được định nghĩa lại là số đường đi ngắn nhất đi qua e mà độ dài của nó nhỏ hơn hoặc bằng h , với h là một biến của thuật toán CONGO.

Độ trung gian cặp (pair betweenness) của đỉnh v cho $\{u, w\}$ là số đường đi ngắn nhất đi qua $\{u, v\}$ và $\{v, w\}$ mà độ dài của đường đi đó không lớn hơn h . Độ trung gian của đỉnh được suy ra từ độ trung gian cặp tương tự như trong CONGA.

Trong thuật toán CONGA tại bước 1 thực hiện tìm kiếm theo bề rộng từ tất cả các đỉnh để tìm ra độ trung gian của các cạnh. Khi sử dụng độ trung gian cục bộ (local betweenness) thì độ sâu của tìm kiếm này được giới hạn là h , nó sẽ nhanh hơn việc phải đi qua tất cả các cạnh trong mạng. Độ trung gian cục bộ có ảnh hưởng đặc biệt trong bước 3 của thuật toán CONGA, khi đó độ trung gian sẽ không phải tính toán cho toàn bộ mạng lưới mà chỉ tại phần cục bộ trong một đồ thị con quanh các cạnh được xóa đi hoặc là đỉnh được phân chia trong bước 2. Steve Gregory định nghĩa một đồ thị con gọi là vùng h (h-region).

Vùng h của cạnh $\{u, v\}$ là vùng bị ảnh hưởng bởi việc loại bỏ $\{u, v\}$, nó là đồ thị con nhỏ nhất chứa tất cả các đường đi ngắn nhất không dài hơn h mà cũng đi qua $\{u, v\}$. Đây là một đồ thị con được tạo thành từ các đỉnh:

$$V_{u,v,h} = \{w: d(u, w) < h \vee d(v, w) < h\}$$

Trong đó $d(u, w)$ là độ dài đường đi ngắn nhất giữa u và v . Vùng h của đỉnh v , tức là vùng bị ảnh hưởng của việc chia v là đồ thị con nhỏ nhất chứa tất cả các đường đi ngắn nhất không dài hơn h mà đi qua v hoặc bắt đầu/kết thúc tại v . Đồ thị con này được tạo thành từ các đỉnh:

$$V_{v,h} = \{w: d(v, w) \leq h\}$$

2.5.2 Các bước của thuật toán

Thuật toán CONGO tương tự như thuật toán CONGA nhưng sử dụng độ đo trung gian cục bộ. Thuật toán CONGO gồm các bước sau:

Bước 1: Tính toán độ trung gian của tất cả các cạnh và độ trung gian của các đỉnh.

Bước 2: Tìm cạnh có độ trung gian lớn nhất và đỉnh có độ trung gian lớn nhất, nếu lớn hơn

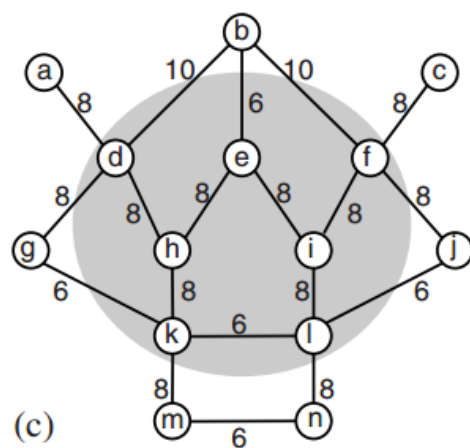
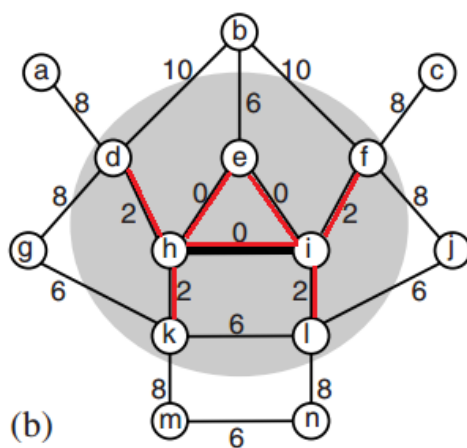
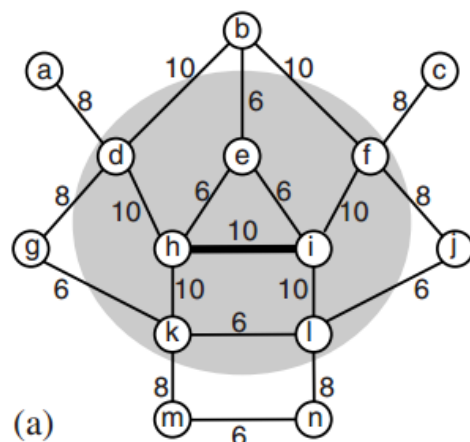
Bước 3: Tính toán lại độ trung gian của cạnh và độ đo trung gian đỉnh:

- a. Trừ đi độ trung gian của vùng h , tập trung vào cạnh bị xóa hay đỉnh được phân chia.
- b. Xóa cạnh hoặc phân chia đỉnh.
- c. Thêm độ đo trung gian cho vùng giống nhau.

Bước 4. Quay lại bước 2 cho đến khi nào mà không còn lại cạnh nào.

Trong bước 3 của thuật toán, để tính toán lại độ trung gian của cạnh và độ trung gian của đỉnh trong vùng h (h-region), đầu tiên tìm tất cả đường đi ngắn nhất không dài hơn h mà nằm trong vùng sau đó trừ đi độ trung gian đã được tính toán trước đó của các cạnh mà nó đi qua và độ trung gian cặp (pair betweenness) mà đi qua đỉnh phân chia. Điều này làm cho độ trung gian của cạnh được chọn hoặc đỉnh được chọn để phân chia bằng 0. Sau khi xóa cạnh hoặc phân chia đỉnh, tiếp tục tìm tất cả đường đi ngắn nhất không dài hơn h trong vùng và thêm độ trung gian của các cạnh mà nó đi qua và độ trung gian cặp của các đỉnh mà nó đi qua.

- (a) Cạnh $\{h, i\}$ được chọn để xóa. Vùng với $h=2$ của $\{h,i\}$ được đánh dấu trong vòng tròn đậm
- (b) Các đường đi ngắn nhất được tìm thấy và bị trừ bởi độ trung gian trong hình a
- (c) Cạnh $\{h,i\}$ được xóa đi, các đường đi ngắn nhất trong vùng được tìm thấy và thêm vào độ trung gian



Hình 2.4 Mô tả thuật toán CONGO trong xóa cạnh trong vùng h [1]

Tại mạng lưới (a) sau khi $\{h,i\}$ được chọn làm cạnh bị xóa, với $h = 2$ thì ta có vùng được chọn sẽ bao gồm các đỉnh $\{h,i,d,e,g,k,l\}$. Tại bước thứ 3a của thuật toán, ta tính được độ trung gian của các cạnh trong vùng được chọn qua bảng dưới đây.

Tên cạnh	Độ trung gian trong vùng $h=2$	Độ trung gian trước đó (mạng a)	Độ trung gian sau khi trừ (nếu có)
$\{d,h\}$	8	10	2
$\{h,e\}$	6	6	0
$\{h,k\}$	8	10	2
$\{e,i\}$	6	6	0
$\{h,i\}$	10	10	0
$\{i,f\}$	8	10	2
$\{i,l\}$	8	10	2
$\{k,l\}$	6	6	6

Bảng 2.3 Độ trung gian của các cạnh trong vùng h bằng 2

Tiếp theo xóa cạnh $\{h,i\}$ ta có độ trung gian của các cạnh trong vùng $h=2$ như sau:

Tên cạnh	Độ trung gian
$\{d,h\}$	6
$\{h,e\}$	8
$\{h,k\}$	6
$\{e,i\}$	8
$\{i,f\}$	6
$\{i,l\}$	6
$\{k,l\}$	6

Bảng 2.4 Độ trung gian của các cạnh trong vùng $h=2$ sau khi xóa cạnh

Từ kết quả của bảng trên, thêm vào độ đo trung gian của cạnh tương ứng nếu cạnh đó cùng cạnh được chọn $\{h,i\}$ tạo nên vùng $h=2$ (cạnh được tô đỏ), kết quả này như trong hình c theo bảng sau:

Tên cạnh	Độ trung gian
{d,h}	8
{h,e}	8
{h,k}	8
{e,i}	8
{i,f}	8
{i,l}	8
{k,l}	6

Bảng 2.5 Độ trung gian của các cạnh trong vùng h sau khi kết thúc thuật toán trong hình 2.4 c

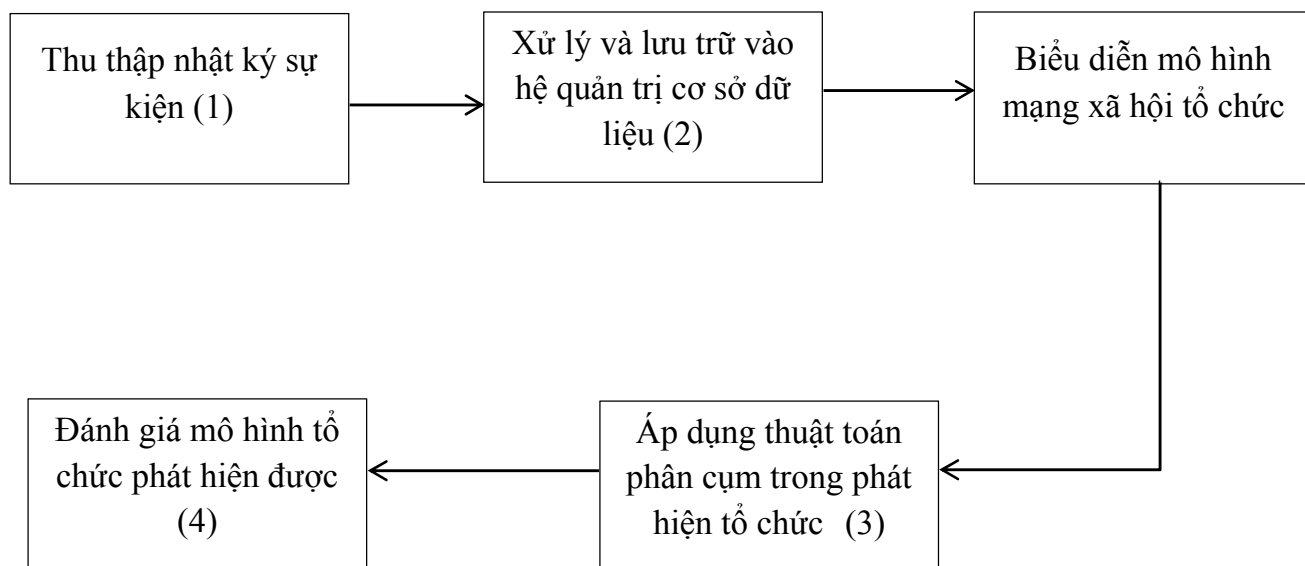
2.6 Tóm tắt chương 2

Trong chương 2 khóa luận đã trình bày các độ đo mô đun hóa trong việc đánh giá chất lượng phân chia đồ thị không xuất hiện chồng chéo, tức là một nút chỉ thuộc về duy nhất một cộng đồng hay một cụm, và hơn nữa là cho việc phân chia đồ thị có sự xuất hiện chồng chéo. Trong chương này, khóa luận cũng tập trung phân tích các thuật toán phân cụm được sử dụng để xác định cấu trúc cộng đồng, nhóm, tổ chức trong mạng, đi sâu vào phân tích các thuật toán phân cụm như AHC hay k-means để phát hiện cấu trúc cộng đồng, ngoài ra còn đề cập các thuật toán phân cụm có xử lý được việc chồng chéo cộng đồng như CONGA và cải tiến của nó là CONGO.

Chương 3. MÔ HÌNH PHÁT HIỆN TỔ CHỨC TRONG PHÁT HIỆN QUÁ TRÌNH SỬ DỤNG CÁC THUẬT TOÁN PHÂN CỤM

3.1 Mô hình phát hiện tổ chức trong phát hiện quá trình

Sau đây là mô hình chúng tôi đề xuất nhằm xây dựng cấu trúc tổ chức trong phát hiện mô hình mạng xã hội trong tổ chức và cài đặt thuật toán CONGO đã được trình bày trong chương 3 để phát hiện cấu trúc tổ chức trong mạng xã hội đó.



Hình 3.1 Mô hình đề xuất giải quyết bài toán phát hiện cấu trúc tổ chức trong phát hiện quá trình

Mô hình giải quyết bài toán đưa ra ở trên gồm 4 pha:

- Pha 1: Thu thập nhật ký sự kiện thực tế của các tổ chức đã được công bố trên các trang web.
- Pha 2: Xử lý dữ liệu từ nhật ký sự kiện, trích chọn các thuộc tính cần thiết cho quá trình khai phá tổ chức mà cần thiết cho việc phát hiện tổ chức như hành động và nguồn thực hiện hành động đó.

- Pha 3: Áp dụng các thuật toán phân cụm vào trong việc tìm ra cấu trúc tổ chức.
- Pha 4: Sử dụng độ đo mô đun hóa đã đề cập trong Chương 2 để tìm ra cấu trúc tổ chức tốt nhất, tức là kết quả của thuật toán cho số cụm là tốt nhất dựa trên độ đo trên.

3.2 Phân tích các thành phần trong mô hình

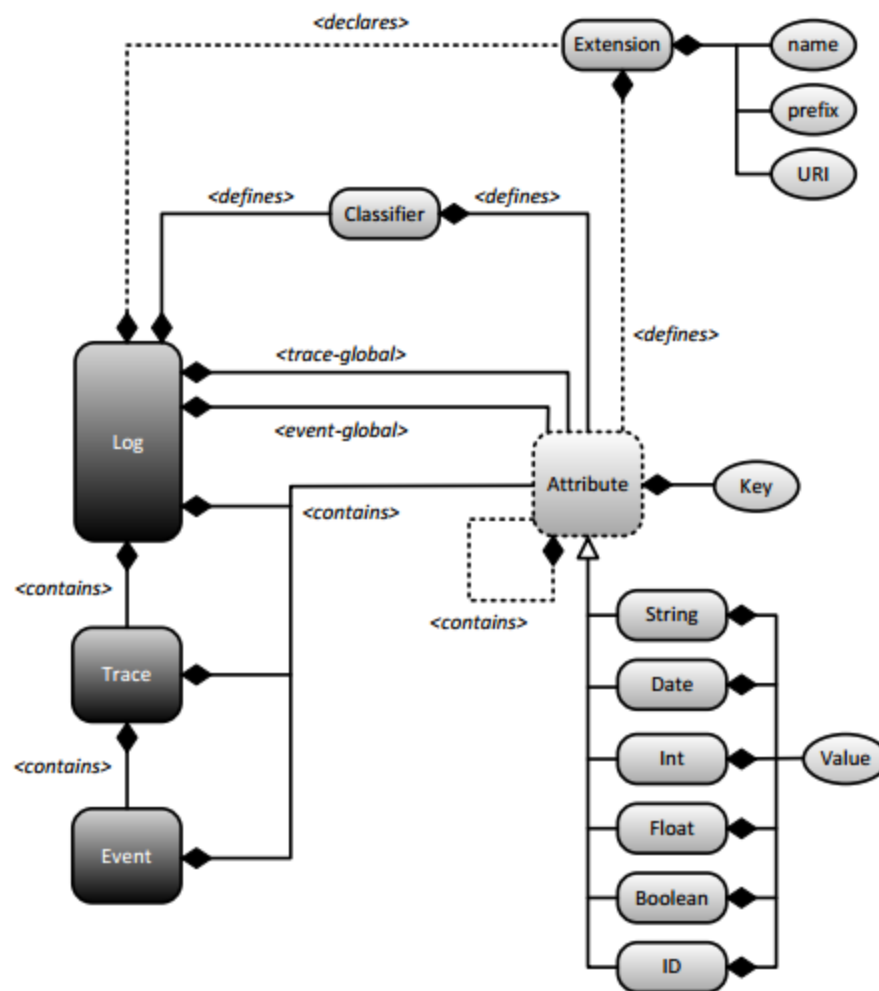
Pha 1: Thu thập nhật ký sự kiện.

Đầu vào của bài toán khai phá dữ liệu là nhật ký sự kiện, như đã trình bày trong chương 1. Nhật ký sự kiện là những ghi chép lại trong quá trình hoạt động của tổ chức trong hệ thống. Ban đầu, dữ liệu được ghi lại có thể dưới nhiều dạng khác nhau như file dữ liệu đơn giản, các bảng tính trong Excel, một nhật ký giao dịch hay cơ sở dữ liệu dạng bảng. Do đó đặt ra yêu cầu là cần phải tiền xử lý dữ liệu và chuẩn hóa dữ liệu nhật ký sự kiện về một chuẩn nhất định. Một trong những dạng lưu trữ của nhật ký sự kiện là XES (eXtensible Event Stream) và MXML (Mining eXtensible Markup Language).

```
<?xml version="1.0" encoding="UTF-8" ?>
  <extension name="Concept" prefix="concept" uri="http://.../concept.xesext"/>
  <extension name="Time" prefix="time" uri="http://.../time.xesext"/>
  <extension name="Organizational" prefix="org" uri="http://.../org.xesext"/>
  <global scope="trace">
    <string key="concept:name" value="name"/>
  </global>
  <global scope="event">
    <date key="time:timestamp" value="2010-12-17T20:01:02.229+02:00"/>
    <string key="concept:name" value="name"/>
    <string key="org:resource" value="resource"/>
  </global>
  <classifier name="Activity" keys="concept:name"/>
  <classifier name="Resource" keys="org:resource"/>
  <classifier name="Both" keys="concept:name org:resource"/>
  <trace>
    <string key="concept:name" value="1"/>
    <event>
      <string key="concept:name" value="register request"/>
      <string key="org:resource" value="Pete"/>
      <date key="time:timestamp" value="2010-12-30T11:02:00.000+01:00"/>
      <string key="Event_ID" value="35654423"/>
      <string key="Costs" value="50"/>
    </event>
    <event>
      <string key="concept:name" value="examine thoroughly"/>
      <string key="org:resource" value="Sue"/>
    </event>
  </trace>
</xml>
```

Hình 3.2 Một phần nhật ký sự kiện định dạng XES [3]

Hình 3.2 đưa ra ví dụ về một phần của một nhật ký sự kiện được lưu trữ theo định dạng XES. Trong nhật ký sự kiện theo định dạng XES lưu một số tùy ý các trường hợp (trace), mỗi trường hợp tương ứng với một phiên làm việc. Trong mỗi trường hợp có một số tùy ý các sự kiện (event), mỗi một sự kiện đại diện cho một hành động được thực hiện. Ví dụ như trong hình 3.2 các hành động là “register request” hay “examine thoroughly”. Trong mỗi event lưu các thuộc tính như tên hành động, người thực hiện hành động, thời gian thực hiện hành động, chi phí. Tùy vào từng nhật ký sự kiện mà có thể có các thuộc tính khác nhau được lưu trữ. Hình 3.3 mô tả cấu trúc của một file định dạng theo chuẩn XES.



Hình 3.3 Sơ đồ lớp UML cho mô hình của chuẩn XES [15]

Trong phạm vi của khóa luận, khóa luận sử dụng các nhật ký sự kiện đã được các nhà nghiên cứu và các tổ chức chia sẻ. Một trong những nguồn dữ liệu sự kiện được sử dụng trong bài toán khai phá quá trình đã được sưu tầm, liệt kê trong trang web <http://www.processmining.org> của nhóm khai phá quá trình đến từ Đại học Công nghệ Eindhoven. Trang web trên đưa ra các nguồn dữ liệu mà những người nghiên cứu về bài toán khai phá quá trình có thể khai thác:

- Các nhật ký sự kiện trong các sách về khai phá quá trình [3].
- Các nhật ký sự kiện trong các khóa học về khai phá quá trình được đưa ra bởi TU/e trong năm 2007-2008.
- Các nhật ký sự kiện mẫu được sử dụng trong các hướng dẫn của bộ công cụ ProM (công cụ khai phá dữ liệu phổ biến nhất hiện nay).
- Bộ sưu tập nhật ký sự kiện trong trang web <http://data.3tu.nl/repository/> sưu tập một số lượng lớn nhật ký sự kiện chia làm hai loại: các nhật ký sự kiện thực tế và các nhật ký sự kiện tự tổng hợp.

Tùy từng bài toán trong khai phá quá trình mà ta sẽ lựa chọn bộ nhật ký sự kiện phù hợp cho mục đích nghiên cứu.

Pha 2: Xử lý nhật ký sự kiện và lưu trữ vào hệ quản trị cơ sở dữ liệu

Để thuận tiện cho các pha sau, trong pha này sẽ tiến hành trích lọc các thông tin cần thiết về các trường hợp và sự kiện. Đối với bài toán khai phá khía cạnh tổ chức, tùy thuộc vào độ đo được lựa chọn đã trình bày trong Chương 1 trong khai phá tổ chức mà lựa chọn các thông tin cần lưu trữ. Ví dụ khi sử dụng độ đo làm việc cùng nhau (working together) tức là những cá nhân nào làm việc cùng trong một trường hợp (trace) thì xếp họ vào một nhóm, do vậy thông tin cần lưu trữ ở đây là tên hành động, người thực hiện hành động và trường hợp (trace) của hành động đó. Một ví dụ khác khi sử dụng độ đo làm công việc giống nhau (similar tasks), từ nhật ký sự kiện cần xác định các cá nhân mà làm các công việc giống nhau và số công việc giống nhau giữa họ, sau đó lưu trữ lại trong hệ cơ sở dữ liệu để làm đầu vào cho các pha sau. Từ các thông tin được lưu trữ theo các độ đo trong hệ quản trị cơ sở dữ liệu, biểu diễn trực quan mạng được xây dựng từ các độ đo trên.

Dưới đây là ví dụ phân tích từ nhật ký sự kiện và đưa ra mối quan liên hệ giữa các nguồn thực hiện công việc theo độ đo làm các công việc giống nhau (similar tasks) theo

[2]. Đầu tiên, xây dựng bảng các công việc và người thực hiện các công việc đó như bảng 3.1.

	Archive claim	Check all	Check policy only	Evaluate claim	Issue payment	Register Claim	Send approval letter	Send rejection letter
Fred	0	2	0	6	0	0	2	0
Howard	0	0	2	0	4	0	0	0
John	2	2	0	0	0	2	0	2
Linda	2	0	0	6	0	0	2	2
Mona	4	2	2	0	0	4	0	0
Robert	4	2	0	0	0	6	2	0
Vincent	0	0	0	0	2	0	0	2

Bảng 3.1 Bảng các công việc mà các nhân viên thực hiện và số lần thực hiện công việc [2]

Từ bảng các công việc và số lần thực hiện công việc trong bảng 3.1 có thể đưa ra sự liên hệ giữa các thành viên, đưa ra số công việc chung giữa hai nhân viên theo bảng 3.2 dưới đây.

	Fred	Howard	John	Linda	Mona	Robert	Vincent
Fred	0	0	1	2	1	2	0
Howard	0	0	0	0	1	0	1
John	0	0	0	2	3	3	1
Linda	0	0	0	0	1	2	1
Mona	0	0	0	0	0	3	0
Robert	0	0	0	0	0	0	0
Vincent	0	0	0	0	0	0	0

Bảng 3.2 Số lượng công việc chung giữa hai người thực hiện công việc [2]

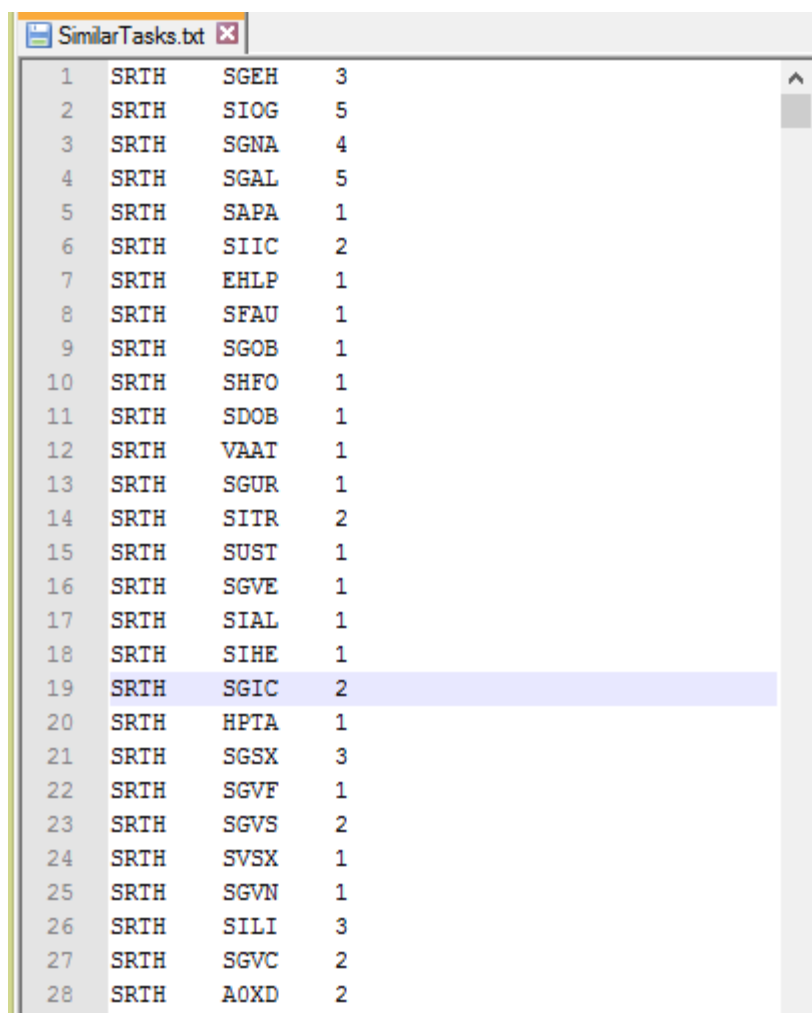
Mối quan hệ giữa các nguồn thực hiện hành động được lưu trữ trong hệ quản trị cơ sở dữ liệu cũng tương tự như đối bảng 3.2.

Pha 3: Áp dụng các thuật toán phân cụm vào phát hiện cấu trúc tổ chức.

Từ kết dữ liệu được xử lý trong pha thứ 2, tùy thuộc vào độ đo được chọn để xây dựng mạng tổ chức như làm việc cùng nhau (doing together) hay làm các công việc giống nhau (similar tasks) của các nhân viên trong tổ chức mà ta có được mối liên hệ giữa các thành viên trong tổ chức. Dữ liệu trên sẽ được sử dụng để làm đầu vào cho các thuật

toán phân cụm đã đề cập trong Chương 2. Tùy vào mục đích của bài toán phát hiện cấu trúc tổ chức, nếu chỉ muốn phát hiện cấu trúc tổ chức không có sự chồng chéo thì áp dụng các thuật toán phân cụm như K-means hay thuật toán phân cụm phân cấp AHC. Mở rộng hơn, để tìm cấu trúc tổ chức và xem xét đến sự chồng chéo trong đó, tức là một nhân viên có thể thuộc về nhiều hơn một nhóm, một đội có thể sử dụng thuật toán có khả năng phát hiện chồng chéo đã được khóa luận đề cập trong Chương 2 như CONGA, CONGO.

Ví dụ đầu vào của thuật toán CONGA là file trích xuất từ bảng làm các công việc giống nhau trong pha 2 cho bộ dữ liệu bệnh viện [16].



ID	Task Name	Count
1	SRTH SGEH	3
2	SRTH SIOG	5
3	SRTH SGNA	4
4	SRTH SGAL	5
5	SRTH SAPA	1
6	SRTH SIIC	2
7	SRTH EHLP	1
8	SRTH SFAU	1
9	SRTH SGOB	1
10	SRTH SHFO	1
11	SRTH SDOB	1
12	SRTH VAAT	1
13	SRTH SGUR	1
14	SRTH SITR	2
15	SRTH SUST	1
16	SRTH SGVE	1
17	SRTH SIAL	1
18	SRTH SIHE	1
19	SRTH SGIC	2
20	SRTH HPTA	1
21	SRTH SGSX	3
22	SRTH SGVF	1
23	SRTH SGVS	2
24	SRTH SVSX	1
25	SRTH SGVN	1
26	SRTH SILI	3
27	SRTH SGVC	2
28	SRTH AOXD	2

Hình 3.4 Hình thể hiện số công việc giống nhau giữa hai nhân viên trích xuất từ hệ quản trị cơ sở dữ liệu

Sau khi tiến hành phân cụm bằng thuật toán CONGA ta có kết quả là kết quả trong mỗi lần phân cụm. Hình 3.5 đưa ra ví dụ về số cụm được phân ra là 5.

1	ANES	CHFO	D0DC	F3NO	F5NO	F5NS	F6DV	F6NO	F7NO	F7ZU	G5NO	H1NO	H3ZU	H4ZU	H5ZU	ICU1	ICU2	ICU3	ICU4	ICU6
2	BLOB	CHE1	CHE2	CITC	CITH	CRLA	CRLE	CRPO	HAEM	HMPA	LAKB	LBAC	LVIR	SKIL	SLAN	SLVE	URIN			
3	A0XD	DAGC	EHLF	F5NF	H3ZU	H5ZU	HPTA	INFZ	NIZI	OKU1	SAPA	SAPY	SDOB	SFAU	SGAL	SGEC	SGEH	SGIC	SGKO	SGNA
4	DRAD	NGIV	RATH	SRA1	SRA2	SRA3	SRA4	SRA5	SRA6	SRAE	SRAS									
5	KLFY	SLV1																		
6																				

Hình 3.5 Kết quả phân cụm theo độ đo làm các công việc giống nhau với số cụm bằng 5 của nhật ký sự kiện [16]

Trong hình 3.5, các phần tử thuộc cùng trên một dòng là thuộc về một cụm được phân ra.

Pha 4: Đánh giá mô hình tổ chức phát hiện được từ thuật toán phân cụm.

Từ kết quả của pha 3, tức là cấu trúc tổ chức đã được phát hiện khi chạy xong các thuật toán phân cụm. Vấn đề đặt ra là phải tìm ra kết quả phân cụm có chất lượng tốt nhất, để giải quyết vấn đề này khóa luận đề cập đến việc sử dụng độ đo mô đun hóa đã trình bày trong Chương 2. Tùy thuộc vào thuật toán phân cụm được sử dụng mà áp dụng độ đo mô đun hóa phù hợp như độ đo mô đun hóa do Girvan Newman đề xuất cho tổ chức không chồng chéo hay độ đo mô đun hóa Nicosia cho đánh giá chất lượng phân cụm có xuất hiện sự chồng chéo.

Đối với các thuật toán phân cụm, sau mỗi lần phân cụm sẽ cho kết quả là một số lượng cụm cụ thể và các phần tử trong mỗi cụm đó. Mỗi lần như vậy ta có thể tính được giá trị mô đun hóa để đánh giá chất lượng phân cụm. Mục tiêu của việc đánh giá chất lượng phân cụm là tìm ra giá trị mô đun hóa lớn nhất, tức là tìm được các cụm, nhóm có chất lượng tốt nhất.

3.3 Tóm tắt chương 3

Trong chương này, khóa luận trình bày các bước trong mô hình đề xuất việc xây dựng đồ thị mô phỏng mô hình mạng xã hội trong phát hiện quá trình và cài đặt thuật toán phân cụm trên mô hình này để tìm ra mô hình tổ chức trên đồ thị vừa xây dựng được.

Trong đó khóa luận tập trung vào khai thác khía cạnh chồng chéo tổ chức, tức là sử dụng thuật toán có khả năng phát hiện sự chồng chéo tổ chức trong khi phân thành các cụm, nhóm.

Trong chương tiếp theo, khóa luận trình bày về quá trình thực nghiệm với mô hình giải quyết bài toán và thực nghiệm phát hiện cấu trúc tổ chức trong phát hiện quá trình từ nhật ký sự kiện.

Chương 4. THỰC NGHIỆM

Để triển khai mô hình đã được đề xuất trong chương 3, chương 4 khóa luận sẽ trình bày quá trình thực nghiệm và quá trình xây dựng mô hình mạng xã hội dựa trên các độ đo trong khai phá khía cạnh tổ chức đã trình bày trong chương 1, sau đó áp dụng thuật toán phân cụm vào trong mô hình thu được để tìm ra cấu trúc tổ chức.

4.1 Môi trường và các công cụ thực nghiệm

Cấu hình phần cứng:

Thành phần	Chỉ số
CPU	Intel(R) Core i3-2310M CPU 2.10 GHz
RAM	4GB
OS	Window 8 64-bit
Bộ nhớ ngoài	500GB

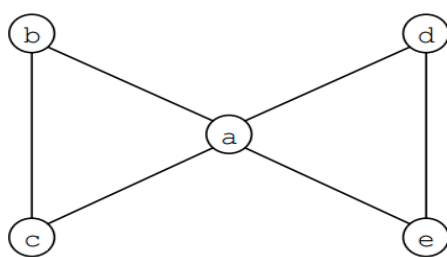
Bảng 4.1 Cấu hình phần cứng

Các phần mềm sử dụng:

STT	Tên phần mềm	Nguồn gốc
1	CONGA Software	http://www.cs.bris.ac.uk/~steve/networks/software/conga.html
2	Eclipse IDE	https://www.eclipse.org/
3	MySQL server	http://dev.mysql.com/downloads/mysql/
4	NodeXL plugin	http:// nodexl.codeplex.com/

Bảng 4.2 Các phần mềm và công cụ sử dụng

Để thực nghiệm phát hiện cấu trúc tổ chức trong khóa luận sử dụng bộ phần mềm CONGA. Đầu vào của phần mềm là một file text có định dạng:



Định dạng CONGA	Danh sách các cạnh
<pre># Network 2 a -- b -- c -- d -- e b -- a -- c c -- a -- b d -- a -- e e -- a -- d</pre>	<pre># Network 2 a b a c a d a e b c d e</pre>

Hình 4.1 Định dạng file đầu vào của bộ phần mềm CONGA [17]

Sau khi tải gói phần mềm CONGA từ địa chỉ [14] được gói “conga.jar”. Để chạy phần gói phần mềm này ta bật cửa sổ command line:

java -cp conga.rar CONGA file.txt [option]

Trong đó file.txt là dữ liệu đầu vào theo một trong hai định dạng đã được trình bày ở trên. Kết quả hiển thị trả về của bộ phần mềm gồm 4 phần:

- Giải thích các lựa chọn
- Tìm kiếm các cụm (Finding clusters)
- Kết quả (Results)
- Thống kê (Statistics)

Khi mạng được phân cụm, một file với tên bắt đầu “clustering-” sẽ bao gồm thông tin phân cụm.

Các mở rộng trong thuật toán được dùng trong khóa luận (option) theo hướng dẫn [14]:

-e	Nếu file đầu vào theo danh sách các cạnh, nếu không thì mặc định file theo định dạng CONGA.
-r	Khi chạy lại thuật toán với cùng file đầu vào để làm mới kết quả.
-n nC	Cho phép quan sát các cụm với nC cụm được phân chia, file trả về sẽ bắt đầu với “cluster-”.
-mo c	Độ đo mô đun hóa theo Nicosia cho ít nhất nC và nhiều nhất là c cụm.

-h h	Chạy thuật toán CONGO, sử dụng h là giá trị của vùng h.
-GN	Chạy thuật toán Girvan và Newman, đối với thuật toán này thì các cụm không chồng chéo.
-w eW	Chấp nhận đồ thị có trọng số trong mạng theo file đầu vào dạng danh sách các cạnh và sử dụng chúng trong tính toán độ trung gian.

Bảng 4.3 Các mở rộng (option) dùng trong thực nghiệm của bộ phần mềm CONGA [17]

4.2 Dữ liệu thực nghiệm

Trong thực nghiệm, khóa luận sử dụng 2 bộ nhật ký sự kiện.

Bộ nhật ký sự kiện thứ nhất:

Đây là bộ dữ liệu tải từ [18] là nhật ký sự kiện mẫu cho công cụ phá vỡ phổ biến hiện nay là ProM. Phân tích nhật ký sự kiện này theo độ đo làm việc cùng nhau (working together) để xây dựng lên mô hình mạng xã hội tổ chức, sau đó sử dụng thuật toán phân cụm AHC do tôi cài đặt để tìm ra cấu trúc tổ chức.

Bộ nhật ký sự kiện này bao gồm:

- 1104 trường hợp.
- 11855 sự kiện.

Nguồn: <http://www.promtools.org/prom6/downloads/example-logs.zip>

Bộ nhật ký sự kiện thứ hai:

Theo [16] đây là bộ nhật ký sự kiện thực tế của bệnh viện thuộc học viện Dutch, dự định ban đầu sử dụng cho BPIC 2011 đầu tiên (Business Process Intelligence Contest), tác giả của bộ dữ liệu này là Van Dongen thuộc Đại học công nghệ Eindhoven.

Bộ nhật ký sự kiện này bao gồm:

- 1143 trường hợp (trace)
- 150291 sự kiện (event)
- Số sự kiện trung bình trong một trường hợp: 131 sự kiện
- Số sự kiện ít nhất trong một trường hợp: 1 sự kiện
- Số sự kiện nhiều nhất trong một sự kiện: 1814 sự kiện

Nguồn: <http://data.3tu.nl/repository/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54>

4.3 Thực nghiệm

Các bước tiến hành thực nghiệm:

Bước 1: Xử lý dữ liệu và lưu các thông tin cần thiết vào hệ quản trị CSDL MySQL.

Bước 2: Xây dựng mô hình mạng xã hội tổ chức dựa trên độ đo làm việc cùng nhau (working together) cho bộ dữ liệu thứ nhất và làm công việc giống nhau (similar tasks) cho bộ dữ liệu thứ hai đã trình bày trong chương I.

Bước 3:

- (a) Dùng thuật toán AHC để phát hiện cấu trúc tổ chức trong bộ dữ liệu thứ nhất dựa vào đầu vào là kết quả của bước 2.
- (b) Dùng thuật toán CONGA để phát hiện cấu trúc tổ chức trong bộ dữ liệu thứ hai dựa vào đầu vào là kết quả của bước 2.

4.4 Kết quả thực nghiệm

4.4.1 Kết quả thực nghiệm tại bước 1

Dữ liệu từ nhật ký sự kiện được xử lý và lưu vào các bảng trong hệ quản trị CSDL để tiện cho quá trình xử lý.

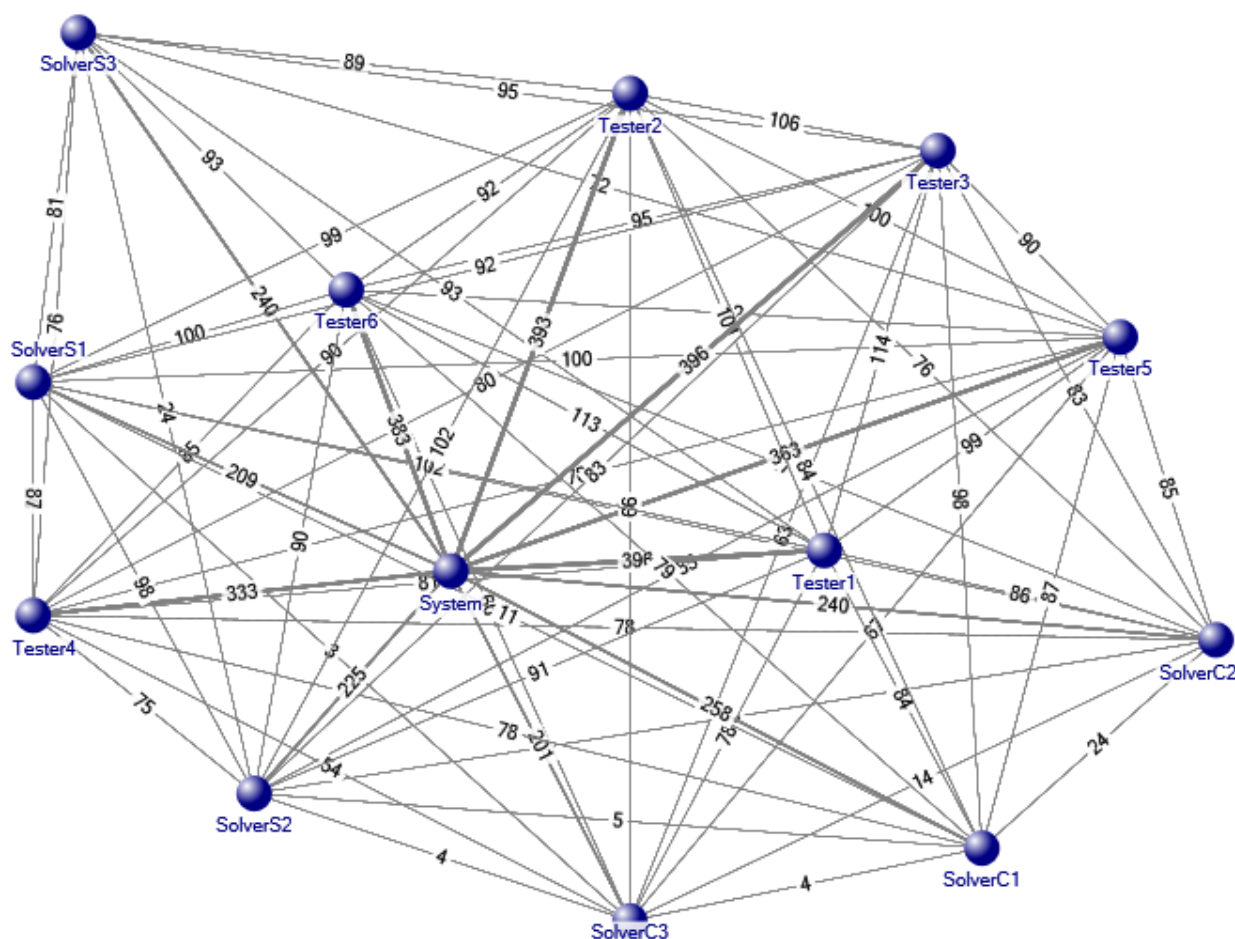
Bảng nhật ký sự kiện (events): Lưu lại tất cả các sự kiện trong nhật ký sự kiện ban đầu

Bảng hành động chung giữa các cá thể (similartasks): lưu lại những người thực hiện các hành động giống nhau và số công việc chung giữa họ.

Bảng làm việc cùng nhau giữa các cá thể (workingtogethers): lưu lại những người cùng thực hiện hành động trong cùng một trường hợp và số lần thực hiện.

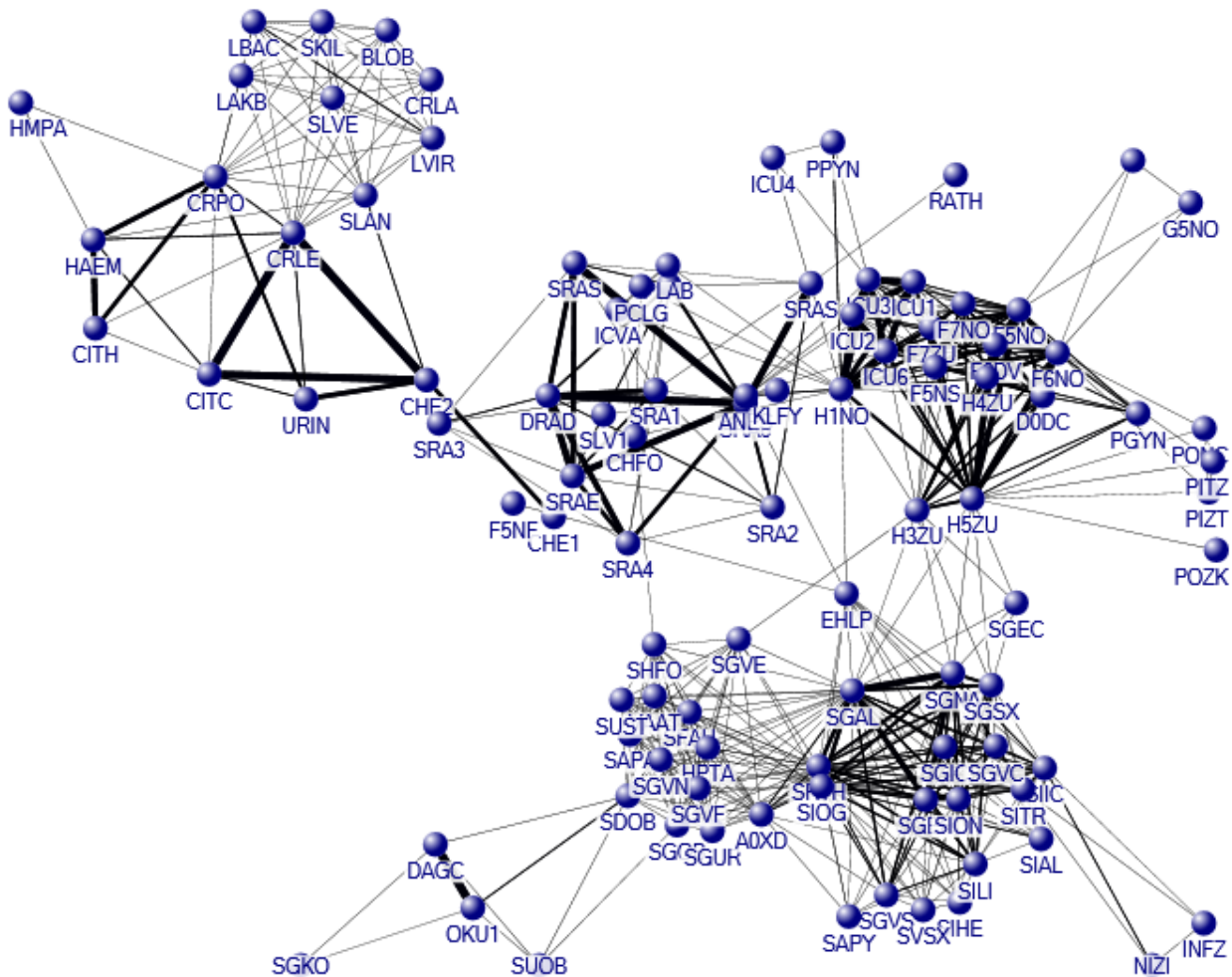
4.4.2 Kết quả thực nghiệm tại bước 2

Đối với bộ dữ liệu thứ nhất [18]: dựa trên độ đo làm việc cùng nhau (working together) trong bước 1, mô hình mạng xã hội theo độ đo này áp dụng cho bộ nhật ký sự kiện thứ nhất, mô hình mạng xã hội theo độ đo này như hình bên dưới:



Hình 4.2 Mạng xây dựng theo độ đo làm việc cùng nhau từ nhật ký sự kiện [18]

Đối với bộ dữ liệu thứ hai [16]: dựa trên độ đo làm công việc giống nhau (similar tasks), mô hình mạng xã hội theo độ đo này như trong hình bên dưới:



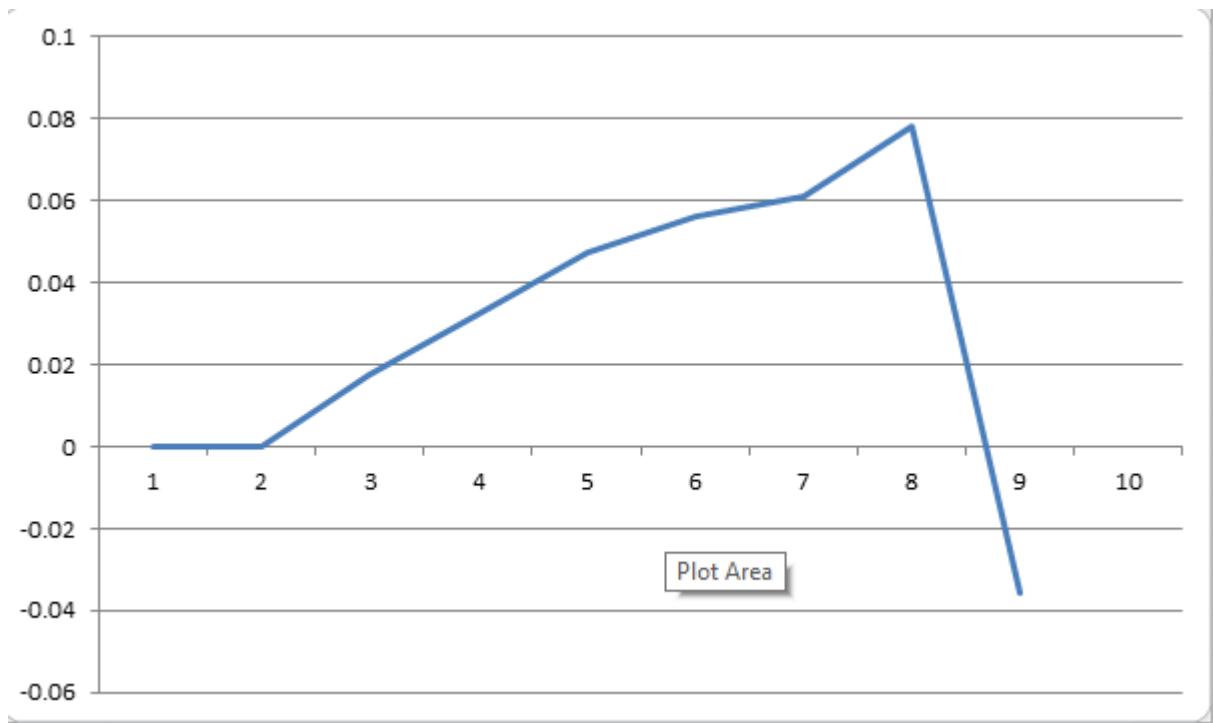
Hình 4.3 Mạng sinh ra từ nhật ký sự kiện theo độ đo làm việc cùng nhau (similar tasks)

Trong mạng trên: giữa hai cá nhân có số công việc thực hiện giống nhau nhiều, tức là có mối liên hệ lớn thì đường nối giữa hai cá nhân đó sẽ đậm hơn. Ngược lại, giữa hai cá nhân có số công việc thực hiện giống nhau ít hơn thì đường nối sẽ mờ hơn.

4.4.3 Kết quả thực nghiệm tại bước 3

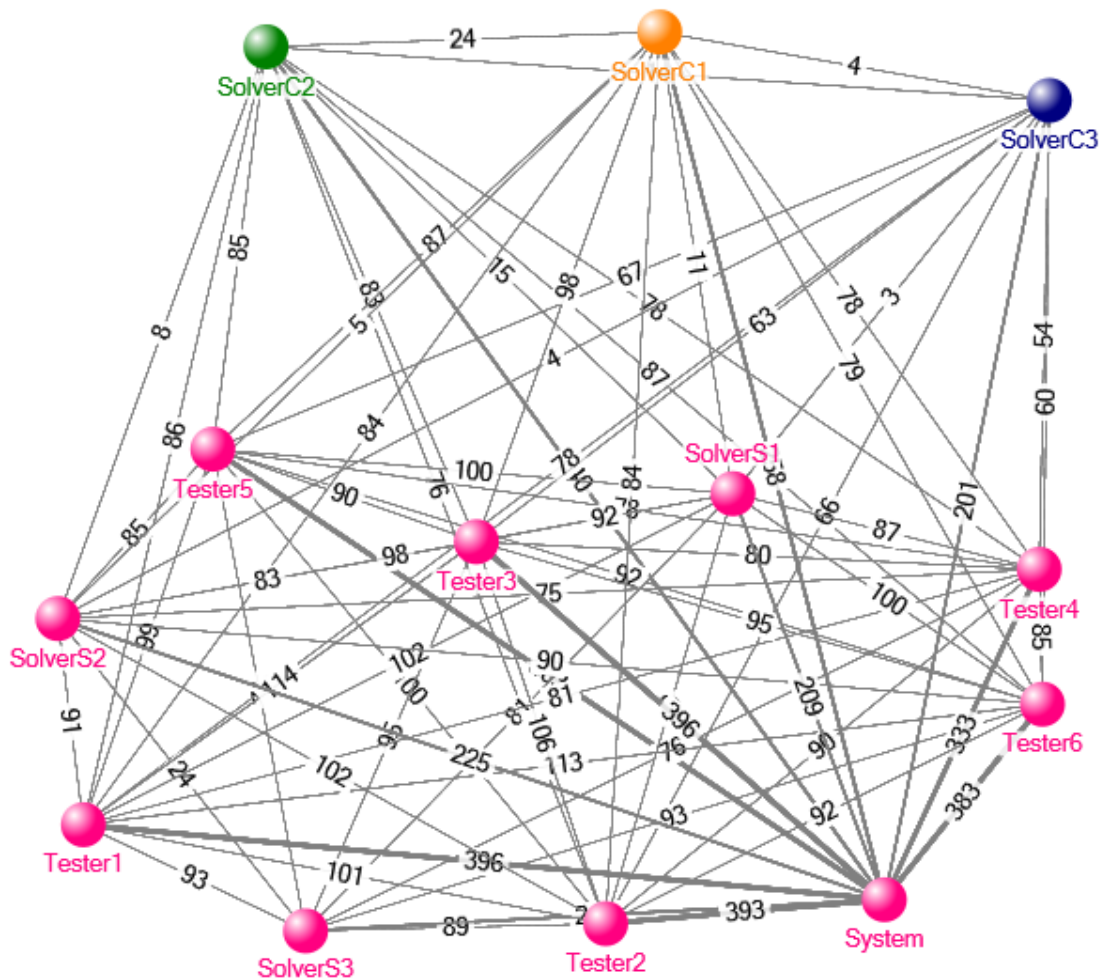
Thực nghiệm với bộ dữ liệu thứ nhất

Đối với bộ dữ liệu thứ nhất [18], sau khi cài đặt thuật toán phân cụm phân cấp AHC, kết quả giá trị mô đun hóa (modularity).



Hình 4.4 Giá trị mô đun hóa sau mỗi lần phân cụm theo AHC

Giá trị mô đun hóa đạt giá trị cực đại tại bước lặp thứ 8 theo thuật toán phân cụm phân cấp AHC với 4 cụm. Ngoài ra khi chạy bằng thuật toán phân cụm Girvan Newman trong bộ phần mềm CONGA cũng cho kết quả giống với thuật toán AHC mà tôi cài đặt. Kết quả phân cụm như hình phía dưới, những nút có cùng màu là thuộc cùng về một nhóm, những nút khác màu thuộc về các nhóm khác nhau.



Hình 4.5 Mô hình tổ chức phát hiện được từ nhật ký sự kiện [18]

Thực nghiệm với bộ dữ liệu thứ hai

Đối với bộ dữ liệu thứ 2 [16], sau bước xây dựng mô hình mạng tổ chức dựa trên độ đo làm các công việc giống nhau (similar tasks), tiến hành phân cụm bằng thuật toán CONGA và CONGO, thuật toán dừng lại khi tất cả các cạnh trong mạng được xóa hết. Kết quả được đưa ra như hình phía sau:

```
C:\Windows\System32\cmd.exe
Remove SRA5/SRA6.SRA5.SRA2 0.25
Remove SGNA.A0XD.H3ZU.H5ZU.SGUS.SGEC.SGAL.SRTH.SAPY.SUSX.SIHE/SGAL.SGNA.H3ZU.SGS
X.SGUS.SGEC.SAPY.SUSX.SIHE 0.25
Remove CITC/CHE2.CRLE.URIN.CITC 0.22
Remove CRLE.CRPO.CITH.URIN.CHE2.CITC.HAEM/CHE2.CRLE.URIN.CITC 0.33
Remove SRA6.DRAD.SRAE.SRAS/DRAD.SRAE.SRAS 0.22
Remove ICU3.ICU4.H1NO.PPYN.ANES/H1NO.ICU3.ICU4.PPYN 0.22
Remove DAGC/OKU1 0.20
Remove CITC/CRLE.CRPO.CITH.URIN.CHE2.CITC.HAEM 0.13

===== Results =====

===== Statistics =====
Initial graph: 98 vertices, 517 edges
Initial graph: 4 components, with sizes: [68, 17, 11, 2]
Clustering: Final graph size: 180
Clustering: Vertices split: 82 total, 34 distinct
Clustering: Vertices split: {H5ZU=7, H3ZU=1, CRPO=2, CRLE=1, SIIC=1, DRAD=1, SRA
E=1, SGSX=1, SDOB=1, URIN=1, SILI=2, H4ZU=2, D0DC=2, ICU3=1, SI0G=5, H1NO=5, SGN
A=3, A0XD=1, ICU6=3, F7NO=1, ICU1=2, ICU2=2, F7ZU=3, F6DU=4, F5NO=4, F5NS=2, F6N
O=4, SRA2=1, CHE2=1, SRA6=3, SGAL=4, SION=3, SRTH=3, SGEH=4}
Clustering: Betweenness phases: 599
Clustering: Total time: 1763ms

C:\Users\cfdcom3g\Desktop\Conga>
```

Hình 4.6 Kết quả khi phân cụm bằng thuật toán CONGA

Khi chạy với thuật toán CONGO:

```
C:\Windows\System32\cmd.exe
Remove SRA5/SRA6.SRA5.SRA2 0.25 r2
Remove SGNA.A0XD.SGUS.SGEC.SGAL.SRTH.SAPY.SUSX.H3ZU.SIHE/SGAL.SGNA.SGSX.SGUS.SGE
C.SAPY.SUSX.H3ZU.SIHE 0.25 r2
Remove CITC/CHE2.CRLE.URIN.CITC 0.22 r3
Remove CRLE.CRPO.CITH.URIN.CHE2.CITC.HAEM/CHE2.CRLE.URIN.CITC 0.33 r3
Remove SRA6.DRAD.SRAE.SRAS/DRAD.SRAE.SRAS 0.22 r2
Remove ICU3.H1NO.ICU6/H1NO.ICU3.ICU2.ICU1.ICU6 0.22 r2
Remove DAGC/OKU1 0.20 r2
Remove CITC/CRLE.CRPO.CITH.URIN.CHE2.CITC.HAEM 0.13 r2

===== Results =====

===== Statistics =====
Initial graph: 98 vertices, 517 edges
Initial graph: 4 components, with sizes: [68, 17, 11, 2]
Clustering: Final graph size: 172
Clustering: Vertices split: 74 total, 35 distinct
Clustering: Vertices split: {H5ZU=6, H3ZU=1, CRPO=2, CRLE=1, SIIC=1, DRAD=1, SRA
E=1, SGSX=1, SDOB=1, URIN=1, SILI=1, H4ZU=2, D0DC=1, ICU3=1, H1NO=4, SI0G=3, SGN
A=3, A0XD=1, ICU6=3, F7NO=1, ICU1=2, PGYN=1, ICU2=1, F7ZU=4, F6DU=5, F5NO=4, F6N
O=3, F5NS=1, CHE2=1, SGUS=1, SRA6=3, SION=2, SGAL=4, SGEH=3, SRTH=3}
Clustering: Betweenness phases: 591
Clustering: Total time: 1554ms

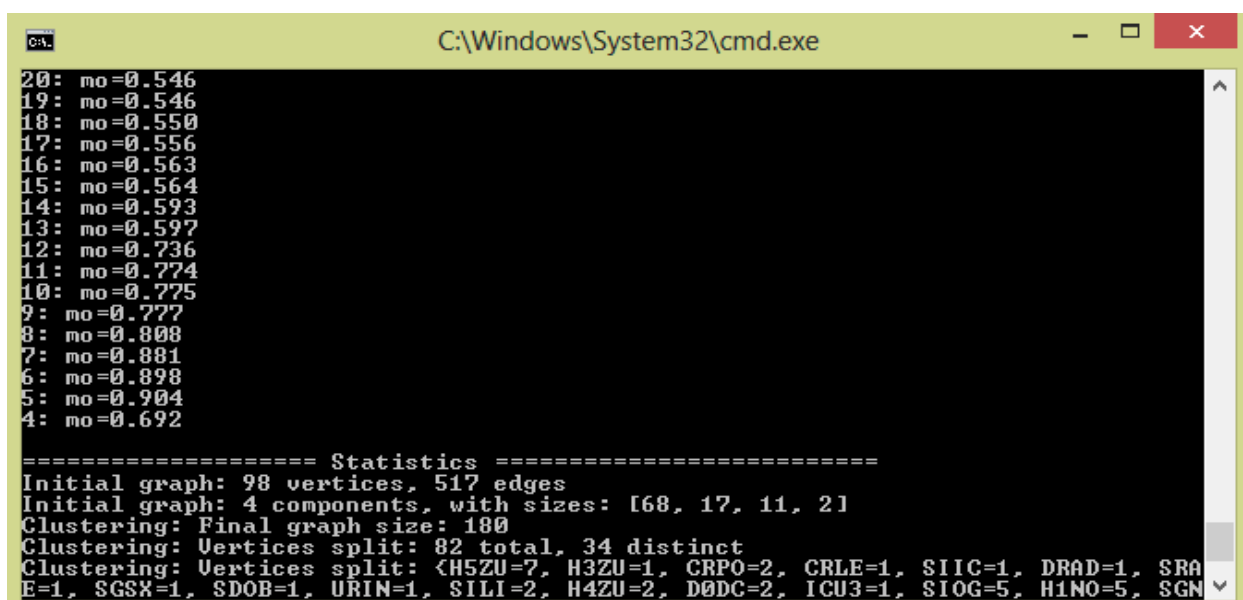
C:\Users\cfdcom3g\Desktop\Conga>
```

Hình 4.7 Kết quả khi phân cụm bằng thuật toán CONGO

Khi chạy với cả 2 thuật toán CONGA và CONGO với $h = 2$ (vùng với giá trị của h là 2) thấy rằng thuật toán CONGO (mất 1554ms) chạy nhanh hơn so với thuật toán CONGA (1763ms), kết quả này là khi thuật toán đã chạy xong, tức là tất cả các cạnh trong đồ thị đã bị xóa.

Để đánh giá chất lượng phân cụm, tức là tìm ra cụm tốt theo hướng dẫn sử dụng phần mềm CONGA [17] có thể sử dụng độ đo mô đun hóa Nicosia, mở rộng cho xử lý chồng chéo cộng đồng.

Ví dụ cho đánh giá chất lượng phân cụm khi tiến hành chạy bằng CONGA:



```

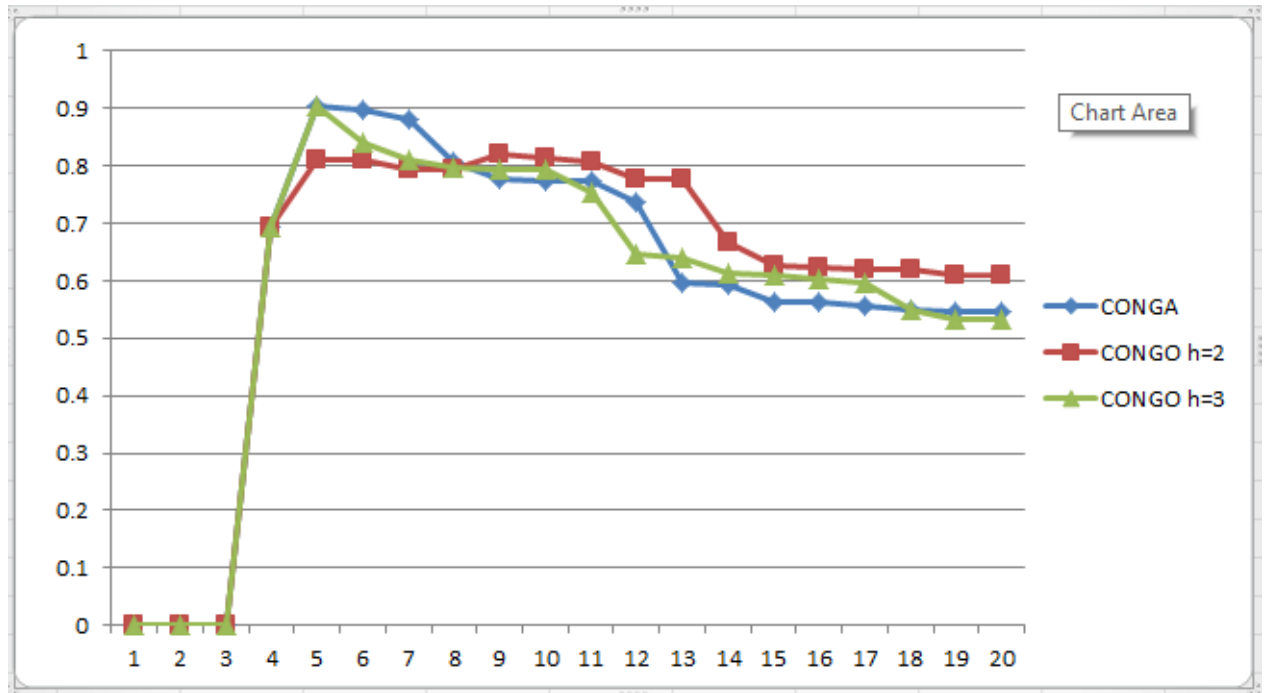
C:\Windows\System32\cmd.exe
20: mo=0.546
19: mo=0.546
18: mo=0.550
17: mo=0.556
16: mo=0.563
15: mo=0.564
14: mo=0.593
13: mo=0.597
12: mo=0.736
11: mo=0.774
10: mo=0.775
9: mo=0.777
8: mo=0.808
7: mo=0.881
6: mo=0.898
5: mo=0.904
4: mo=0.692

===== Statistics =====
Initial graph: 98 vertices, 517 edges
Initial graph: 4 components, with sizes: [68, 17, 11, 2]
Clustering: Final graph size: 180
Clustering: Vertices split: 82 total, 34 distinct
Clustering: Vertices split: <H5ZU=7, H3ZU=1, CRPO=2, CRLE=1, SIIC=1, DRAD=1, SRA
E=1, SGSX=1, SDOB=1, URIN=1, SILI=2, H4ZU=2, D0DC=2, ICU3=1, SI0G=5, H1NO=5, SGN

```

Hình 4.8 Giá trị mô đun hóa khi phân cụm bằng thuật toán CONGA

Sau khi chạy bằng thuật toán CONGA, CONGO với $h=2$ và $h=3$ và dùng độ đo mô đun hóa để đánh giá chất lượng phân chia. Kết quả so sánh được đưa ra dưới đây:



Hình 4.9 Biểu diễn giá trị mô đun hóa sau khi chạy bằng thuật toán CONGA, CONGO với $h=2$ và $h=3$

Cũng theo Steve Gregory[1] cho rằng, so với thuật toán CONGA thì thuật toán CONGO nhanh hơn đáng kể, đặc biệt là với $h=2$. Đối với các mạng thực tế quá lớn thì CONGA gặp hạn chế về mặt thời gian nhưng CONGA có độ chính xác cao hơn CONGO.

Để tìm cấu trúc tổ chức từ bộ dữ liệu thứ hai, khóa luận lựa chọn thuật toán CONGA cho kết quả cuối cùng, do số lượng các cạnh và số lượng đỉnh trong đồ thị cần phân cụm không quá lớn, thuật toán CONGO hoạt động tốt với các đồ thị với số lượng hàng triệu đỉnh và cạnh và theo [1] thì độ chính xác của CONGA tốt hơn CONGO. Sau khi phân cụm đối với bộ dữ liệu thứ 2, giá trị mô đun hóa đạt cực đại tại $n = 5$ như trong hình 4.9, kết quả 5 cụm như trong hình 4.10.

Kết luận và định hướng nghiên cứu tiếp theo

Khai phá quá trình là một chủ đề nghiên cứu thời sự, có ý nghĩa khoa học và thực tiễn, trong đó việc khai phá các khía cạnh bổ sung trong phát hiện quá trình cũng đang được rất nhiều các nhà nghiên cứu quan tâm. Khóa luận tập trung vào các thuật toán phân cụm để giải quyết bài toán phát hiện tổ chức trong phát hiện quá trình.

Khóa luận đã trình bày các nội dung cơ bản của bài toán phát hiện mô hình tổ chức trong phát hiện quá trình, nêu được phương pháp để xây dựng lên mô hình mạng xã hội tổ chức dựa vào các độ đo sau đó vận dụng các thuật toán phân cụm để tìm ra cấu trúc tổ chức.

Trong khóa luận, hướng giải quyết bài toán dựa trên các phương pháp phân tích mạng xã hội và các giải thuật phân cụm đã được giới thiệu trong các nghiên cứu trước đó của các nhà nghiên cứu mà nổi bật là Van der Aalst.

Khóa luận đã vận dụng các phương pháp khai phá cấu trúc tổ chức để xây dựng và tìm ra cấu trúc tổ chức từ nhật ký sự kiện, tiến hành phân tích và thực nghiệm với một số thuật toán cơ bản được sử dụng để tìm ra cấu trúc tổ chức như thuật toán phân cụm phân cấp, thuật toán Girvan Newman. Ngoài ra, khóa luận tập trung nghiên cứu về vấn đề chồng chéo trong cấu trúc tổ chức, tức là một cá nhân thuộc về nhiều hơn một nhóm bằng việc áp dụng các thuật toán phân cụm như CONGA và CONGO.

Trong thời gian tới, tôi sẽ tiếp tục mở rộng khóa luận bằng cách nghiên cứu thêm về các phương pháp phân tích cấu trúc tổ chức và các thuật toán phân cụm khác để tìm ra cấu trúc tổ chức từ nhật ký sự kiện. Trong quá trình thực hiện khóa luận, việc sưu tầm nhật ký sự kiện của các tổ chức mà có thể khai thác vấn đề chồng chéo tổ chức còn gặp nhiều khó khăn vì vậy kết quả thực nghiệm chưa thực sự được như mong đợi. Vì vậy, trong thời gian tới tôi sẽ cố gắng cải thiện kết quả thực nghiệm và nghiên cứu để có thể ứng dụng bài toán này vào một số doanh nghiệp của Việt Nam.

Tài liệu tham khảo

- [1] Steve Gregory (2008) : A Fast Algorithm to Find Overlapping Communities in Networks, Department of Computer Science University of Bristol, BS8 1UB, England
- [2] Claudia Sofia da Costa Alves (2010). Social Network Analysis for Business Process Discovery, The Technical University of Lisbon
- [3] Wil M.P. van der Aalst (2011): Discovery, Conformance and Enhancement of Business Processes.
- [4] Minseok Song and Wil M.P. van der Aalst (2008): Towards Comprehensive Support for Organizational Mining
- [5] Mahdi ABDELKAFI, Lotfi BOUZGUENDA, Faiez GARGOURI (2012): Discop Flow: A new Tool for Discovering Organizational Structures and Interaction Protocols in WorkFlow
- [6] Steve Gregory (2007): An Algorithm to Find Overlapping Community Structure in Networks. PKDD
- [7] Anil K. Jain (2010) : Data clustering: 50 years beyond K-means
- [8] A. LÁZÁR, D. ÁBEL and T. VÍCSEK (2010): Modularity measure of networks with overlapping communities
- [9] M. Girvan, M. E. J. Newman (2002). Community structure in social and biological networks, Proc. Natl. Acad. Sci., 99(12), 7821 (2002)
- [10] W. Reisig and G. Rozenberg (editors, 1998). Lectures on Petri Nets I: Basic Models, Lecture Notes in Computer Science, 1491, Springer-Verlag, Berlin.
- [11] JIERUI XIE, STEPHEN KELLEY, BOLESŁAW K. SZYMANSKI (2013): Overlapping Community Detection in Networks: The State-of-the-Art, ACM Computing Surveys, Vol. 45, No. 4, Article 43. and Comparative Study.
- [12] R.P. Jagadeesh Chandra Bose (2012): Process mining in the large, Eindhoven University of Technology, India

- [13] R.P. Jagadeesh Chandra Bose and Wil M.P. van der Aalst (2009): Trace Clustering based on Conserved Patterns:Towards Achieving Better Process Models, BPM 2009 International Workshops, Ulm, Germany, September 7, 2009. Revised Papers
- [14] <http://www.cs.bris.ac.uk/~steve/networks/software/conga.html>
- [15] C.W. Günther: XES Standard Definition.www.xes-standard.org, 2012
- [16] <http://data.3tu.nl/repository/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54>
- [17] Steve Gregory (2012): CONGA User's Guide, v1.67 – November 6, 2012
- [18] <http://www.promtools.org/prom6/downloads/example-logs.zip>