ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Bùi Thị Hồng Nhung

KỸ THUẬT KHAI PHÁ MẪU DÃY VÀ MẪU THỨ TỰ BỘ PHẬN TRONG KHAI PHÁ QUY TRÌNH

Chuyên ngành: Hệ thống Thông tin
Mã số: 9480104.01

TÓM TẮT LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội – 2020

Công trình được hoàn thành tại: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Người hướng dẫn khoa học: PGS.TS. Nguyễn Trí Thành

PGS.TS. Nguyễn Cẩm Tú

Phản biện: PGS.TS. Đỗ Trung Tuấn

Phản biện: PGS.TS. Nguyễn Long Giang

Luận án sẽ được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại .....

vào hồi giờ ngày tháng năm 2020.

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam
- Trung tâm Thông tin - Thư viện, Đại học Quốc gia Hà Nội

Mở đầu

Dữ liệu đã được chứng minh là tài nguyên mới và quan trọng trong nền công nghiệp tương lai, đặc biệt là nền công nghiệp 4.0. Việc khai thác các dữ liệu đã trở thành một khâu có tác động đến lợi thế cạnh tranh của doanh nghiệp. Các hệ thống thông tin hiện đại ngày nay đã và đang tích lũy được một lượng dữ liệu khổng lồ về các quá trình thực hiện nghiệp vụ trên nhiều miền lĩnh vực khác nhau. Những dữ liệu về các sự kiện xảy ra trong quá trình thực hiện của hệ thống được thu thập và lưu trữ trong các tệp dữ liệu nhật ký sự kiện. Khai phá quy trình (process mining) là lĩnh vực cho phép sử dụng dữ liệu nhật ký sự kiện để phân tích và cải tiến các quy trình. Có hai yếu tố chính làm cho khai phá quy trình ngày càng nhận được nhiều sự quan tâm của các học giả trong lĩnh vực hàn lâm và ứng dụng. Thứ nhất, ngày càng có nhiều dữ liệu sự kiện được ghi nhận lại trong các hệ thống thông tin (như Hoạch định nguồn lực doanh nghiệp - ERP; Quản lý luồng công việc - WFM; Quản lý quan hệ khách hàng - CRM; Quản lý chuỗi cung ứng - SCM; Quản lý dữ liệu sản phẩm – PDM…) giúp cung cấp tốt hơn các thông tin chi tiết về quy trình thực tế. Thứ hai, xuất hiện ngày càng nhiều các yêu cầu đặt ra đối với các nhà quản lý về cách các quy trình của họ hoạt động trong thế giới thực nhằm hỗ trợ và cải tiến các quy trình nghiệp vụ trong môi trường kinh doanh có tính cạnh tranh cao với nhiều thay đổi nhanh chóng. Trong quản lý quy trình nghiệp vụ (BPM) các nhà quản lý đã và đang được hỗ trợ bởi các công cụ kinh doanh thông minh (BI), nhưng chúng chưa đáp ứng được kỳ vọng của các nhà quản lý trong môi trường kinh doanh hiện nay. Trọng tâm của BI là truy vấn và báo cáo các thông tin tổng hợp của doanh nghiệp dưới dạng bảng điều khiển (dashboard) sử dụng các kỹ thuật trực quan đơn giản thay vì hiểu biết sâu sắc về bản chất thực sự của quy trình khi được đưa vào thực thi trong thực tế. Một số hệ thống đã hỗ trợ khả năng khai phá dữ liệu (Data mining) hoặc hỗ trợ xử lý phân tích trực tuyến (Online Analytical Processing - OLAP) có thể xem dữ liệu đa chiều từ các góc nhìn khác nhau và có

thể tổng hợp dữ liệu để tạo báo cáo cấp cao đồng thời có thể đi sâu vào dữ liệu để tìm thông tin chi tiết. Nhưng chúng thiếu khả năng cung cấp cái nhìn về nguyên nhân gốc của sự không hiệu quả hoặc sai sót của quy trình. Khai phá quy trình được xây dựng dựa trên tiếp cận giữa học máy và khai phá dữ liệu với mô hình hóa và phân tích quy trình, cùng với sự kết hợp chặt chẽ các kỹ thuật, công cụ và phương pháp riêng nhằm thu nhận tri thức từ tập nhật ký sự kiện mô tả các bước thực hiện thực tế của các quy trình nghiệp vụ trong các hệ thống thông tin hiện thời để phân tích quy trình, phát hiện những vấn đề sai lệch từ đó đề xuất điều chỉnh, thiết kế lại quy trình một cách chính xác hơn mang lại hiệu quả công tác cao hơn. Khai phá quy trình có thể được nhúng vào các công cụ BI để cung cấp cái nhìn sâu sắc về ngữ nghĩa hoạt động thực sự của các quy trình kinh doanh, góp phần thổi sự sống vào các mô hình quy trình tĩnh với lượng dữ liệu sự kiện khổng lồ. Do đó, các xu hướng quản lý liên quan đến cải tiến quy trình hay tạo ra các quy trình thông minh có thể được giải quyết bởi khai phá quy trình. Với những lợi ích mà nó mang lại, khai phá quy trình đang trở thành một trong những hướng nghiên cứu thu hút được sự quan tâm của các nhà nghiên cứu trong lĩnh vực quản lý quy trình nghiệp vụ và khoa học máy tính. Hiện nay, khai phá quy trình đã được áp dụng trong các hệ thống BPM thương mại khác nhau.

Tại Việt Nam cũng không nằm ngoài xu hướng phát triển của thế giới, cải tiến quy trình nghiệp vụ nhằm rút ngắn thời gian hoàn thành dịch vụ công là một mục tiêu được đặt ra trong bốn nghị quyết của Chính phủ về cải thiện môi trường kinh doanh, nâng cao năng lực cạnh tranh quốc gia trong bốn năm vừa qua. Như vậy, việc nghiên cứu và triển khai về khai phá quy trình không chỉ phù hợp với xu thế nghiên cứu triển khai về khai phá quy trình trên thế giới mà còn phù hợp với chủ trương cải tiến quy trình nghiệp vụ của Chính phủ ta hiện nay và đây là một công việc hết sức cần thiết.

Mục tiêu nghiên cứu: Thứ nhất, luận án cung cấp một khảo sát khái quát về Khai phá quy trình. Thứ hai, luận án đề xuất

các phương pháp biểu diễn vết và phương pháp tính khoảng cách giữa các vết cập nhật những kết quả nghiên cứu hiện đại trên thế giới nhằm nâng cao hiệu quả của giải pháp phân cụm vết cải thiện chất lượng mô hình quy trình. Nghiên cứu, đề xuất thuật toán phân cụm vết khai thác được các đặc trưng riêng trong lĩnh vực khai phá quy trình là mục tiêu thứ ba của luận án. Cuối cùng, luận án xây dựng các phần mềm thử nghiệm thực thi các giải pháp biểu diễn vết, tính khoảng cách giữa các vết và thuật toán phân cụm vết được luận án đề xuất để kiểm chứng tính hiệu quả của các đề xuất đó.

Đối tượng nghiên cứu của luận án là các phương pháp biểu diễn vết, các phương pháp tính khoảng cách vết và các thuật toán phân cụm vết.

Phạm vi nghiên cứu của luận án tập trung vào giải pháp Phân cụm vết nâng cao chất lượng mô hình quy trình trong bài toán Phát hiện mô hình quy trình với ba vấn đề gồm (i) Các phương pháp biểu diễn vết; (ii) Các độ đo trong phân cụm vết; (iii) Các thuật toán phân cụm vết.

Phương pháp nghiên cứu của luận án là nghiên cứu lý thuyết kết hợp với nghiên cứu thực nghiệm để kiểm chứng đánh giá các đề xuất của luận án.

Chương 1. Phát hiện mô hình quy trình trong Khai phá quy trình và các thách thức

1.1 Khai phá quy trình-Một lĩnh vực nghiên cứu mới

Khai phá quy trình là một chuyên ngành nghiên cứu mới nổi, được phát triển mạnh mẽ trong một thập niên gần đây. Theo Van der Aalst, khai phá quy trình là một lĩnh vực nghiên cứu liên kết giữa *học máy và khai phá dữ liệu* (*machine learning and data mining*) với *mô hình hóa và phân tích quy trình* (*process modeling and analysing*), nhằm chiết xuất các tri thức có giá trị liên quan đến quy trình nghiệp vụ (*business process*) từ các nhật ký sự kiện (*event log*), bổ sung các phương pháp tiếp cận *quản lý quy trình nghiệp vụ* (*bussiness process management: BPM*).

1.2. Một số khái niệm cơ bản về nhật ký sự kiện

1.2.1 Hoạt động

Hoạt động (*activity*, còn được gọi là hành động) là một bước xử lý nghiệp vụ đã được xác định cụ thể, rõ ràng (*well-defined*), không gây nhập nhằng trong một tổ chức. Khi đề cập tới một hoạt động (chẳng hạn, tiếp nhận đơn yêu cầu: Tiếp nhận) thì mọi người có liên quan trong tổ chức đều có thể hiểu rõ và thi hành được nội dung công việc tương ứng với hoạt động.

1.2.2 Sự kiện

Sự kiện *(event)* là một lần thi hành một hoạt động trong thực tế cùng với các thông tin liên quan. Ví dụ khi một khách hàng cụ thể nộp đơn yêu cầu bồi thường hàng không, một sự kiện tương ứng với hoạt động tiếp nhận đơn yêu cầu được thi hành trong một nhãn thời gian cụ thể *(timestamp)*, do một tài nguyên cụ thể thực hiện *(resource)*, với một chi phí cụ thể *(cost)*, và là một bước cụ thể trong toàn bộ quá trình xử lý đơn yêu cầu của khách hàng cụ thể đã cho.

1.2.3 Trường hợp

Trường hợp *(case)* là dãy bao gồm tất cả các sự kiện được thi hành trong một lần xử lý cụ thể đối với một nghiệp vụ. Mỗi trường hợp được định danh bằng mã trường hợp *(case id)* và các sự kiện xuất hiện được sắp xếp theo thứ tự tăng dần của nhãn thời gian.

1.2.4 Vết

Trong khai phá quy trình, khi khai phá khía cạnh liên quan đến hoạt động, các trường hợp có thể được mô tả cô đọng dưới dạng tập các vết (*trace*). Với vết là một chuỗi các hoạt động có chung mã trường hợp và được sắp xếp theo thứ tự tăng dần của nhãn thời gian

1.2.5 Biểu diễn và lưu trữ nhật ký sự kiện

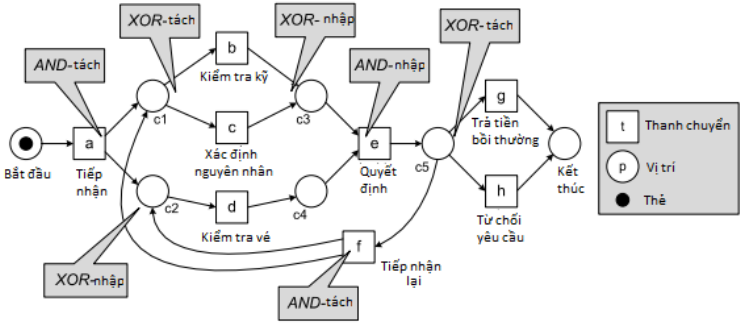
Nhật ký sự kiện của các tổ chức và hệ thống khác nhau có thể được lưu trữ dưới nhiều định dạng khác nhau như cơ sở dữ liệu, csv, excel, XES, MXML, … Đối với các ứng dụng khai phá quy trình thường sử dụng định dạng MXML bởi tính linh hoạt dễ hiểu, dễ sử dụng (Hình 1.2).

```
<Process id="L-conf">
    <ProcessInstance id="Case1">
        <AuditTrailEntry>
            <WorkflowModelElement>Tiếp nhận</WorkflowModelElement>
            <EventType>complete</EventType>
            <Timestamp>12-30-2010:11.02</Timestamp>
            <Originator>Pete</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
            <WorkflowModelElement>Kiểm tra kỹ</WorkflowModelElement>
            <EventType>complete</EventType>
            <Timestamp>12-31-2010:10.06</Timestamp>
            <Originator>Sue</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
            <WorkflowModelElement>Kiểm tra vé</WorkflowModelElement>
            <EventType>complete</EventType>
            <Timestamp>01-06-2011:15.12</Timestamp>
            <Originator>Mike</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
            <WorkflowModelElement>Quyết định</WorkflowModelElement>
            <EventType>complete</EventType>
            <Timestamp>01-07-2011:11.18</Timestamp>
            <Originator>Sara</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
            <WorkflowModelElement>Từ chối yêu cầu</WorkflowModelElement>
            <EventType>complete</EventType>
            <Timestamp>01-07-2011:14.24</Timestamp>
            <Originator>Pete</Originator>
        </AuditTrailEntry>
    </ProcessInstance>
    <ProcessInstance id="Case2">
        <AuditTrailEntry>
            <WorkflowModelElement>Tiếp nhận</WorkflowModelElement>
            <EventType>complete</EventType>
            <Timestamp>12-30-2010:11.32</Timestamp>
            <Originator>Mike</Originator>
        </AuditTrailEntry>
        .....................................
    </ProcessInstance>
    .....................................
</Process>
```

Hình 1.2 Cấu trúc file MXML biểu diễn nhật ký sự kiện Lfull

1.3Mô hình hóa quy trình nghiệp vụ và khai phá quy trình

Khai phá quy trình là bài toán chiết xuất thông tin có giá trị liên quan đến các hoạt động của một quy trình lưu vết nhật ký sự kiện và tự động đưa ra một mô hình quy trình nghiệp vụ phản ánh chính xác những thông tin chứa trong nhật ký sự kiện đó (Hình 1.4). Khác với phương pháp thủ công, khai phá quy trình không dùng tập các hoạt động và mối liên hệ của chúng về mặt lý thuyết từ các nhà phân tích mà tiến hành phát hiện và cải tiến quy trình một cách tự động dựa trên tập dữ liệu khách quan mà quy trình đã được triển khai thực hiện trong thực tế được lưu vết trong NKSK, giúp giải quyết được các hạn chế từ mô hình hóa quy trình nghiệp vụ thủ công.



Hình 1.4 Mô hình quy trình NKSK Lfull sử dụng lưới Petri.

1.4 Ba bài toán chính trong khai phá quy trình

1.4.1 Bài toán Phát hiện mô hình quy trình

Phát hiện mô hình quy trình là bài toán đầu tiên trong khai phá quy trình, đại diện cho phương pháp mô hình hóa quy trình một cách tự động. Bài toán nhận đầu vào là Tập nhật ký sự kiện của một quy trình nghiệp vụ và cho đầu ra là một mô hình quy trình có khả năng đại diện cho các hoạt động thấy được trong nhật ký sự kiện đó.

1.4.2 Bài toán Kiểm tra sự phù hợp

Kiểm tra sự phù hợp là bài toán thứ hai của khai phá quy trình, bài toán nhận đầu vào là Tập nhật ký sự kiện và Mô hình quy trình. Kết quả đầu ra sẽ chẩn đoán và định lượng sự không phù hợp giữa hoạt động được mô hình hóa trong mô hình quy trình và hoạt động được quan sát trong NKSK.

1.4.3 Bài toán cải tiến mô hình

Cải tiến mô hình là bài toán thứ ba của khai phá quy trình, được thực hiện sau khi có kết quả từ bài toán kiểm tra sự phù hợp là mô hình quy trình không phản ánh đúng thực tế. Cải tiến mô hình nhận tập Nhật ký sự kiện và Mô hình quy trình làm dữ liệu đầu vào và cho đầu ra là một mô hình quy trình mới được sửa hay mở rộng từ mô hình trước đó.

Luận án tập trung nghiên cứu chuyên sâu về bài toán Phát hiện mô hình quy trình với các phương pháp cải tiến chất lượng của mô hình quy trình được sinh ra. Đây là bài toán có vai trò quan trọng, là đầu vào và cũng là yếu tố quyết định tới

chất lượng cũng như hiệu quả của hai bài toán Kiểm tra sự phù hợp và Cải tiến mô hình. Nếu ngay từ đầu mô hình quy trình được sinh ra không có độ chính xác cao thì việc đánh giá sự phù hợp cũng như cải tiến mô hình quy trình đều không có giá trị thực tiễn.

1.5. Thách thức và giải pháp phân cụm vết nâng cao chất lượng bài toán Phát hiện mô hình quy trình

1.5.1 Thách thức dữ liệu từ NKSK và nhóm giải pháp

Trong Phát hiện mô hình quy trình nói riêng và Khai phá quy trình nói chung, nhật ký sự kiện đóng một vai trò quan trọng, đây không chỉ là dữ liệu đầu vào mà còn là đối tượng nghiên cứu chính mở ra nhiều hướng phát triển, nhiều bài toán ứng dụng khác nhau của khai phá quy trình. Hai thách thức về kích thước nhật ký sự kiện quá lớn và các sự kiện trong nhật ký sự kiện quá cụ thể với mức trừu tượng thấp có những ảnh hưởng to lớn tới chất lượng mô hình quy trình được sinh ra. Cụ thể hai thách thức này làm nảy sinh hai vấn đề nổi bật. Thứ nhất, nhật ký sự kiện quá lớn tạo ra các khó khăn đối với các công cụ khai phá quy trình hiện có, chẳng hạn, công cụ ProM 5.3 không làm việc được với một số bộ dữ liệu sẵn có. Thứ hai, các sự kiện trong nhật ký sự kiện quá cụ thể với mức trừu tượng rất thấp dẫn tới mô hình quy trình kết quả có độ chính xác thấp và rất phức tạp để diễn giải.

Tiền xử lý dữ liệu sự kiện là nhóm các hướng giải pháp được nhiều nhà nghiên cứu quan tâm gồm kỹ thuật chia để trị phát hiện mẫu nhằm phân tách các bản ghi sự kiện dựa trên việc phân vùng các hoạt động, nâng mức trừu tượng của dữ liệu sự kiện,… giúp cải tiến kết quả của bài toán trong phát hiện quy trình. Cụ thể nhóm giải pháp gồm ba bài toán Trừu tượng hóa hoạt động; Trôi khái niệm; Phân cụm vết.

1.5.2 Tổng quan về giải pháp Phân cụm vết nâng cao chất lượng mô hình quy trình

Một cách tiếp cận để khắc phục điều này là thay vì sinh một mô hình quy trình lớn từ toàn bộ nhật ký sự kiện để giải thích mọi thứ, người ta tiến hành phân cụm nhật ký sự kiện thành một tập các cụm sự kiện con sao cho dữ liệu trong mỗi cụm

sự kiện con tương đồng với nhau và sinh các mô hình quy trình con từ tập các cụm sự kiện này. Các mô hình quy trình con được sinh ra từ tập các bản ghi nhật ký sự kiện con đồng nhất có thể dẫn đến các mô hình quy trình đơn giản, dễ hiểu, dễ phân tích có độ đo phù hợp cao và độ phức tạp về cấu trúc thấp. Phương pháp phân cụm vết được coi là phương pháp đơn giản, linh hoạt và hiệu quả giúp làm giảm độ phức tạp cho bài toán phát hiện quy trình.

1.5.3 Các vấn đề nghiên cứu trong giải pháp phân cụm vết

Tổng hợp từ các nghiên cứu về giải pháp Phân cụm vết, các nhà khoa học đã đưa ra ba vấn đề cơ bản xuất hiện trong bài toán phân cụm vết nhật ký sự kiện gồm:

(i) Vấn đề đầu tiên là phương pháp biểu diễn các vết bao gồm hai nội dung lựa chọn đặc trưng và phương pháp biểu diễn dữ liệu. Mỗi trường hợp bao gồm một dãy các sự kiện, mỗi sự kiện bao gồm một tập các thuộc tính. Lựa chọn đặc trưng liên quan đến việc xem xét sử dụng các thuộc tính nào để tạo các vết sao cho phù hợp và có thể đại diện tốt nhất cho nhật ký sự kiện đang xét. Với các đặc trưng được lựa chọn, cần tìm ra phương thức biểu diễn dữ liệu phù hợp. Tồn tại hai tiếp cận cho vấn đề thứ nhất là tiếp cận véc-tơ và tiếp cận ngữ cảnh. Các nghiên cứu của luận án về vấn đề này được trình bày tại chương 2 và chương 5.

(ii) Vấn đề thứ hai là độ đo tương tự giữa các phần tử dữ liệu trong các thuật toán phân cụm. Nội dung nghiên cứu vấn đề này của luận án được trình bày tại chương 3.

(iii) Vấn đề thứ ba là thuật toán phân cụm được áp dụng. Nội dung này được luận án trình bày tại chương 4.

Chương 2. Đồ thị khoảng cách trong biểu diễn vết nâng cao chất lượng mô hình quy trình

2.1 Các phương pháp biểu diễn vết truyền thống

2.1.1 Túi các hoạt động – Bag of activities

Túi các hoạt động - Bag of activities (BOA) là một trong những phương pháp biểu diễn vết cơ bản nhất. Trong phương pháp này, mỗi một vết trong tập nhật ký sự kiện được chuyển đổi

thành một véc-tơ số (véc-tơ nhị phân hoặc véc-tơ tần số) dựa theo véc-tơ đặc trưng của tập nhật ký sự kiện đó.

2.1.2 k-gram

Mô hình biểu diễn dữ liệu k-gram được sử dụng rộng rãi trong các lĩnh vực xử lý ngôn ngữ tự nhiên, khai phá dữ liệu. Ý tưởng của nó là chia một chuỗi ban đầu thành các chuỗi con liên tiếp độ dài k bằng cách sử dụng một cửa sổ trượt độ dài k trượt từ trái sang phải qua từng phần tử xuất hiện trong chuỗi. k-gram với $k = 1, 2, 3, 4$ được gọi lần lượt là unigram, bigram, trigram và tetra-gram.

2.1.3 Lặp cực đại - Maximal Repeats

Lặp cực đại – Maximal Repeats (MR) với mục đích tìm kiếm tất cả các chuỗi hoạt động chung lớn nhất xuất hiện ít nhất hai lần trong toàn bộ nhật ký sự kiện. Mục đích này được xuất phát từ ý tưởng nhận xét rằng sự phân bố của các chuỗi hoạt động chung lớn nhất giữa các vết trong nhật ký sự kiện biểu thị sự giống nhau hoặc biểu thị một mối liên kết nào đó giữa các vết.

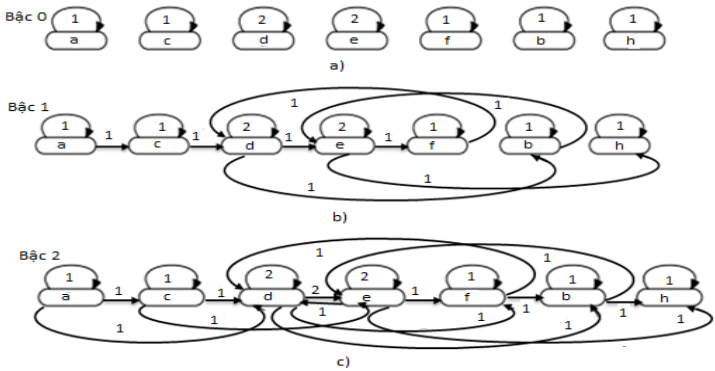
2.2 Biểu diễn vết sử dụng đồ thị khoảng cách

2.2.1 Đồ thị khoảng cách

Đồ thị khoảng cách - Distance Graph (DG) là một mô hình biểu diễn văn bản thông qua cấu trúc đồ thị có thể mô tả thông tin về thứ tự và khoảng cách giữa các từ trong văn bản. Đồ thị khoảng cách bậc k mô tả thông tin về các cặp từ cách nhau tối đa k vị trí trong văn bản. Đồ thị khoảng cách bậc k $(k \geq 0)$ của một văn bản D trong một tập văn bản C được định nghĩa là đồ thị $G(C, D, k) = (N(C), A(D, k))$, trong đó $N(C)$ là tập các nút của đồ thị và $A(D, k)$ là tập các cung.

2.2.2 Ứng dụng đồ thị khoảng cách trong biểu diễn vết

Để ứng dụng đồ thị khoảng cách trong biểu diễn vết, luận án ánh xạ tập A các *hoạt động* trong nhật ký sự kiện như là tập các *từ* riêng biệt trong tập văn bản C và một vết T trong nhật ký sự kiện được ánh xạ như là một văn bản D. Xét vết $T = \langle acdefdbeh \rangle$ chúng ta có các biểu diễn của T theo đồ thị khoảng cách bậc 0,1,2 như sau (Hình 2.2):

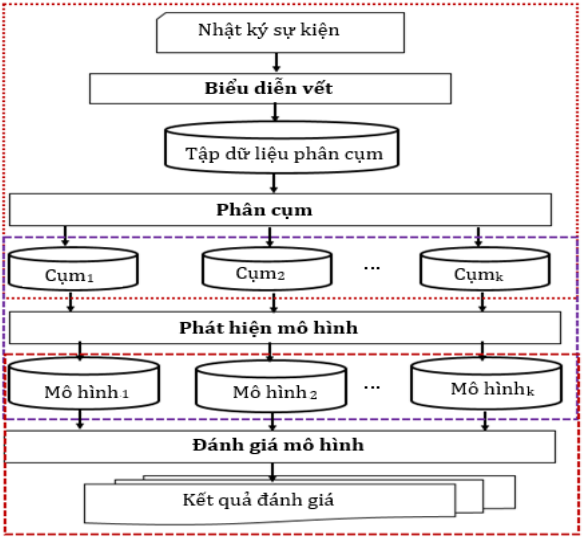


Hình 2.2. Đồ thị khoảng cách của vết $T = \langle acdefdbeh \rangle$

2.3 Mô hình ứng dụng Đồ thị khoảng cách trong biểu diễn vết

2.3.1 Mô hình ba pha Phát hiện mô hình quy trình

Luận án đề xuất một mô hình ba pha ứng dụng cho bài toán Phát hiện mô hình quy trình sử dụng giải pháp phân cụm vết dựa trên Đồ thị khoảng cách gồm: *Biểu diễn vết và Phân cụm; Phát hiện mô hình; Đánh giá mô hình* (Hình 2.3).



Hình 2.3. Khung mô hình ứng dụng đồ thị khoảng cách phát hiện mô hình quy trình

2.3.2 Thực nghiệm

Dữ liệu thực nghiệm: Bộ dữ liệu thực nghiệm luận án sử dụng ba tập nhật ký sự kiện: Lfull, prAm6 và prHm6. Trong đó Lfull có nhiều vết trùng nhau (ví dụ vết $\langle acdeh \rangle$ xuất hiện 455 lần)và có sự lặp lại các hoạt động trong cùng một vết (ví dụ hoạt động d và e xuất hiện 2 lần trong vết $\langle acdefdbeh \rangle$); prAm6 có các vết trùng nhau và không có các hoạt động lặp trong một vết; prHm6 không có các vết trùng nhau và không có các hoạt động lặp trong một vết.

Bảng 2.5. Các phương pháp biểu diễn vết và chất lượng mô hình quy trình sử dụng thuật toán K-means

Phương pháp biểu diễn vết	**Lfull**		**prAm6**		**prHm6**	
	Fitness	Precision	Fitness	Precision	Fitness	Precision
BOA	0.991	0.754	**0.968**	**0.809**	**0.902**	**0.660**
2GR	0.951	0.958	0.968	0.809	0.902	0.660
3GR	0.955	0.962	0.968	0.809	0.902	0.660
MR	0.948	0.929	0.968	0.809	0.902	0.660
DG1	0.952	0.967	0.968	0.809	0.902	0.660
DG2	**0.992**	**1**	0.968	0.809	0.902	0.660
DG3	**0.992**	**1**	0.968	0.809	0.902	0.660

So với các phương pháp biểu diễn vết trước đó, cách biểu diễn dựa trên đồ thị khoảng cách có hiệu suất tốt hơn trong trường hợp nhật ký sự kiện có chứa các hoạt động lặp lại (Lfull).

Chương 3. Trọng số vết – Độ đo khoảng cách vết mới

3.1 Các phương pháp tính khoảng cách truyền thống

Trong bài toán tính khoảng cách giữa các vết, các nghiên cứu chủ yếu tập trung vào một số phương pháp cơ bản đã được sử dụng trong các lĩnh vực như Xử lý ngôn ngữ tự nhiên, Khai phá dữ liệu,… gồm: Khoảng cách Euclid, Hamming, Jaccard, Levenshtein, Hệ số tương quan Correlation, Độ đo Cosine…

Đặc điểm của các phương pháp này là tính toán khoảng cách giữa hai vết chỉ dựa trên mối quan hệ nội tại của chúng mà không tính tới mối quan hệ với các vết khác trong NKSK.

3.2 Đo khoảng cách vết sử dụng độ đo Google chuẩn hóa

3.2.1 Độ đo Google chuẩn hóa

Độ đo Google chuẩn hóa (*Normalized Google Distance - NGD*) là khoảng cách ngữ nghĩa tương đối nhằm tính toán mối quan hệ tương đồng giữa hai *thuật ngữ* trong ngôn ngữ tự nhiên dựa trên ngữ cảnh sử dụng của chúng trên mạng internet thông qua công cụ tìm kiếm Google hoặc một công cụ tìm kiếm bất kỳ cho phép trả về tổng số trang các thuật ngữ xảy ra độc lập và xảy ra đồng thời cùng nhau.

3.2.2 Ứng dụng độ đo Google chuẩn hóa tính khoảng cách giữa các vết

Luận án đề xuất các định nghĩa về Độ đo trọng số chuẩn hóa tương ứng với trường hợp tính trọng số ảnh hưởng một hoạt động, của cặp hai hoạt động và cặp ba hoạt động đối với toàn bộ nhật ký sự kiện và định nghĩa về Độ đo trọng số vết chuẩn hóa tính trọng số ảnh hưởng trung bình của một vết đối với toàn bộ nhật ký sự kiện.

Định nghĩa 3.2 (NW(x)). Độ đo trọng số chuẩn hóa của hoạt động x đối với toàn bộ NKSK:

$$NW(x) = \frac{\log f(x)}{\log N - \log f(x)} \quad (3.6)$$

Định nghĩa 3.3 (NW(x,y)). Độ đo trọng số chuẩn hóa của cặp hai hoạt động xy đối với toàn bộ NKSK:

$$NW(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y))}{\log N - \min\{\log f(x), \log f(y)\}} \quad (3.7)$$

Định nghĩa 3.4 (NW(x,y,z)). Độ đo trọng số chuẩn hóa của cặp ba hoạt động xyz đối với toàn bộ NKSK:

$$NW(x,y,z) = \frac{\max\{\log f(x), \log f(y), \log f(z)\} - \log f(x,y,z))}{\log N - \min\{\log f(x), \log f(y), \log f(z)\}} \quad (3.8)$$

Trong đó $f(x)$ là số các vết trong NKSK chứa hoạt động x; $f(x,y)$ là số các vết trong NKSK chứa đồng thời cả hai hoạt động x và y; $f(x,y,z)$ là số các vết trong NKSK chứa đồng thời cả ba hoạt động x, y và z; N là tổng số các vết trong NKSK.

Quy ước 1. Độ đo trọng số chuẩn hóa $NW(.) = 0$ khi mẫu của các công thức tương ứng (9), (10) hoặc (11) có giá trị = 0.

Định nghĩa 3.5 (NTW(t)). Độ đo trọng số vết chuẩn hóa (Normalized Trace Weight) của vết t đối với toàn bộ NKSK:

$$NTW(t) = \frac{\sum_{pt_i} NW(pt_i)}{|pt_i \in t|} \quad (3.9)$$

3.3. Ứng dụng Độ đo trọng số vết chuẩn hóa trong bài toán Phân cụm vết

Để đánh giá sự ảnh hưởng của Độ đo trọng số vết chuẩn hóa đối với kết quả phân cụm vết và chất lượng các mô hình quy trình được sinh ra, luận án tiến hành thực hiện các thực nghiệm chuyên sâu như sau.

Kịch bản thực nghiệm: Luận án thực hiện ba kịch bản thực nghiệm phân cụm vết nhật ký sự kiện với các độ đo khoảng cách vết khác nhau:

Thực nghiệm 1: Phân cụm vết sử dụng véc-tơ nhị phân theo biểu diễn vết k-gram (k=1,2,3) và các khoảng cách Euclid, khoảng cách Jaccard, hệ số tương quan, độ đo cosine như là phương pháp cơ sở để đánh giá hiệu quả của các kịch bản thực nghiệm.

Thực nghiệm 2: Phân cụm vết sử dụng phương pháp biểu diễn vết k-gram (k=1,2,3) và Độ đo trọng số vết chuẩn hóa không xét thứ tự thực hiện của các hoạt động trong vết (NTW*).

Thực nghiệm 3: Phân cụm vết sử dụng phương pháp biểu diễn vết k-gram (k=1,2,3) và Độ đo trọng số vết chuẩn hóa có xét thứ tự thực hiện của các hoạt động trong vết (NTW**).

Kết quả thực nghiệm:

Bảng 3.3 Kết quả độ đo mô hình sử dụng thang đo truyền thống và NTW

NKSK / Thang đo	**Lfull**		**prAm6**		**prHm6**	
	Fitness	*Precision*	*Fitness*	*Precision*	*Fitness*	*Precision*
Khoảng cách Euclid						
1-gram	0.991	0.754	0.968	0.809	0.902	0.660
2-gram	0.951	**0.958**	0.968	0.809	0.902	0.660
3-gram	0.955	**0.962**	0.968	0.809	0.902	0.660
Độ đo Cosine						
1-gram	0.991	0.754	0.789	0.868	0.902	0.660
2-gram	0.951	**0.958**	0.789	0.868	0.902	0.660
3-gram	0.955	**0.962**	0.789	0.868	0.898	0.634
Khoảng cách Jaccard						
1-gram	0.953	0.796	0.722	**0.904**	0.898	0.634
2-gram	0.944	0.929	0.702	0.863	0.898	0.634

3-gram	0.910	0.736	0.722	0.904	0.898	0.634
Hệ số tương quan						
1-gram	0.953	0.796	0.722	**0.904**	0.898	0.634
2-gram	0.944	0.929	0.699	0.878	0.898	0.634
3-gram	0.930	0.707	0.702	0.863	0.898	0.634
Độ đo trọng số vết chuẩn hóa NTW*						
1-gram	**0.994**	**0.806**	**0.970**	0.606	**0.919**	**0.795**
2-gram	**0.995**	0.913	**0.972**	**0.995**	**0.921**	**0.809**
3-gram	**0.9999**	0.930	**0.973**	**0.933**	**0.911**	**0.861**
Độ đo trọng số vết chuẩn hóa NTW**						
1-gram	0.989	**0.806**	**0.970**	0.722	0.899	0.791
2-gram	0.989	0.867	**0.972**	0.756	0.903	0.583
3-gram	0.940	0.815	**0.973**	0.693	0.901	0.640

Các kết quả thực nghiệm trong Bảng 3.3 cho thấy NTW* có hiệu suất tốt nhất trong đa số các trường hợp, đặc biệt giá trị thang đo Fitness luôn cao hơn so với trường hợp dùng khoảng cách Euclid, cao hơn hoặc bằng so với NTW**.

Chương 4. Thuật toán phân cụm vết mới theo ngữ cảnh ContextTracClus

4.1 Hướng tiếp cận ngữ cảnh trong phân cụm vết

4.1.1 Khái niệm ngữ cảnh trong khai phá quy trình

Trong khai phá quy trình khái niệm ngữ cảnh cũng đã được đề cập với những đặc thù riêng: ngữ cảnh là môi trường xung quanh một quy trình nghiệp vụ, ví dụ như điều kiện thời tiết hoặc mùa nghỉ lễ; thời gian, địa điểm và tần suất thực hiện của các sự kiện cũng như mối liên kết hay các công cụ, thiết bị hỗ trợ liên quan…

4.1.2 Khái niệm ngữ cảnh vết

Xuất phát từ thực tế, tập các vết của mỗi một quy trình hay của một biến thể của một quy trình có thể được bắt đầu bởi một chuỗi các hoạt động chung tạo ra một ngữ cảnh thực hiện riêng của chúng. Trong nghiên cứu này, luận án đề xuất khái niệm ngữ cảnh vết chính là tập các chuỗi hoạt động chung đó.

4.1.3 Cây ngữ cảnh

Cây ngữ cảnh là một cây có: Một gốc được gắn nhãn "*root*" để tạo thành một cây hoàn chỉnh. *Bảng tiêu đề* giúp truy cập cây

nhanh hơn trong quá trình xây dựng và duyệt cây. Mỗi dòng trong bảng tiêu đề gồm hai trường là *tên_hoạt_động* và *nút_liên_kết* trỏ đến nút đầu tiên bên dưới nút gốc chứa hoạt động tương ứng này. Mỗi nút trong cây ngữ cảnh (ngoại trừ nút gốc) bao gồm ba thuộc tính: *tên_hoạt_động*: xác định hoạt động được biểu diễn trong các nút đại diện cho mỗi nhánh của cây; *số_vết*: số lượng các vết đi qua nút này; *nút_liên_kết*: con trỏ trỏ tới nút con của nó hoặc null nếu nút là nút lá.

4.1.4 Xây dựng cây ngữ cảnh

Ý tưởng xây dựng cây ngữ cảnh là ánh xạ các vết có cùng tiền tố vào cùng một nhánh của cây. Thuật toán xây dựng cây ngữ cảnh được mô tả như sau:

Thuật toán 1: ContextTreeConstruction(L)

Đầu vào: Nhật ký sự kiện L

Đầu ra: Cây ngữ cảnh T tương ứng.

Thuật toán: Thuật toán gồm 3 bước:

1. Tạo một nút gốc cho cây ngữ cảnh với nhãn "root". Khi đó cây T=root
2. *Foreach* vết t *in* L *do*
   Xét t=a|q trong đó a là hoạt động đầu tiên và q là chuỗi các hoạt động còn lại của t
   Gọi thủ tục *insert_activity*(a|q,T) chèn vết t=a|q vào cây T
   EndFor
3. Tạo bảng tiêu đề và cập nhật nút_liên_kết trỏ tới nút con trực tiếp của nút gốc.

Thuật toán 2: insert_activity(a|q,T)

Đầu vào: - Cây ngữ cảnh T
- Vết t có dạng t=a|q trong đó a là hoạt động đầu tiên và q là chuỗi các hoạt động còn lại của t.

Đầu ra: Cây ngữ cảnh T được cập nhật thêm vết t.

Thuật toán:

1. *If* T có nút con N thỏa mãn N.tên_hoạt_động=a *then*
   N.số_vết = N.số_vết + 1
   Else
   Tạo một nút mới N với N.số_vết = 1

Tạo một nút_liên_kết trỏ từ T tới N
EndIf

2. *If* q khác rỗng *then*
   Gọi đệ quy thủ tục *insert_activity*(q,N)
   EndIf

4.1.5 Xác định ngữ cảnh vết

Thuật toán xác định ngữ cảnh của một vết trên cây ngữ cảnh:

Thuật toán 3: ContextDetection(a|q,T,context)

Đầu vào: - Cây ngữ cảnh T
- Một vết có dạng a|q trong đó a là hoạt động đầu tiên và q là chuỗi các hoạt động còn lại của vết

Đầu ra: Ngữ cảnh của vết.

Thuật toán:

1. *If* T=root *then*
   context = ∅;
   Xác định nút N được trỏ bởi nút_liên_kết từ bảng tiêu đề tại dòng tương ứng với hoạt động a;
   Else
   Tìm nút con N của T thỏa mãn N.tên_hoạt_động=a
   EndIf
2. *If* N.số_vết > 1 *then*
   context = context|a; *//Phép toán nối chuỗi*
   If q khác rỗng *then*
   Gọi đệ quy ContextDetection(q,N,context);
   EndIf
   EndIf

4.2 Giải pháp phân cụm vết mới dựa theo ngữ cảnh

4.2.1 Ý tưởng đề xuất

Thuật toán phân cụm vết tương ứng với giải pháp này được luận án đặt tên là **ContextTracClus**. Thuật toán bao gồm 2 pha: i) Xác định ngữ cảnh vết và xây dựng các cụm; ii) Điều chỉnh cụm.

Pha đầu tiên, *Xác định ngữ cảnh vết và Xây dựng các cụm*, bao gồm hai bước. *Bước 1* xây dựng cây ngữ cảnh. *Bước 2* duyệt từng vết qua cây ngữ cảnh nhằm xác định ngữ cảnh của vết và gán vết vào cụm tương ứng với ngữ cảnh này.

Pha thứ hai, *Điều chỉnh cụm*, xử lý trường hợp khi các cụm nhỏ được tạo ra. Nếu kích thước cụm, tức là số lượng các vết trong cụm, nhỏ hơn ngưỡng kích thước cụm tối thiểu cho trước thì cụm này sẽ được ghép vào cụm gần với nó nhất.

4.2.2 Thuật toán phân cụm vết mới - ContextTracClus

Thuật toán 4: ContextTracClus

Đầu vào: - Nhật ký sự kiện *L*

- Ngưỡng kích thước cụm tối thiểu *mcs*

Đầu ra: Tập các cụm vết *C*.

Thuật toán:

Pha 1: Xác định ngữ cảnh vết và Xây dựng các cụm

```
1. C={};
2. T=ContextTreeConstruction(L);
3. Foreach vết t in L do
       ContextDetection(t,T,context);
       If context == null then
          Tạo một cụm mới c; // c không có nhãn
          Thêm vết t vào cụm c; //Cụm c chỉ gồm 1 vết
          C=C∪c;
       Else  //tìm được ngữ cảnh context
          If C chưa có cụm được gán nhãn context then
             Tạo cụm mới c gán nhãn context;
             Thêm vết t vào cụm c;
              C=C∪c;
          Else
              Thêm vết t vào cụm được gán nhãn context;
          EndIf
       EndIf
   EndFor
```

Pha 2: Điều chỉnh cụm

```
   Foreach cụm c in C do
   If size(c) < mcs then
         Ghép c vào cụm gần nó nhất trong C;
   EndIf
   EndFor
```

4.3 Khung mô hình ứng dụng thuật toán ContextTracClus trong phân cụm vết

4.3.1 Mô hình ứng dụng

Để áp dụng thuật toán ContextTracClus trong bài toán phân cụm vết nói riêng và bài toán phát hiện mô hình quy trình nói chung, luận án đề xuất một khung ứng dụng bao gồm 5 bước: Tiền xử lý; Xác định ngữ cảnh vết và Xây dựng các cụm; Điều chỉnh cụm; Phát hiện mô hình; Đánh giá mô hình.

4.3.2 Thực nghiệm

Trong kịch bản thực nghiệm của thuật toán K-means, DBSCAN luận án sử dụng véc-tơ nhị phân tương ứng với các biểu diễn k-gram (k=1,2,3) của các vết.

Bảng 4.1 Kết quả thực nghiệm giữa các thuật toán phân cụm và ContextTracClus

	Lfull		**prAm6**		**prHm6**	
	Fitness	*Precision*	*Fitness*	*Precision*	*Fitness*	*Precision*
Kịch bản 1: Thuật toán K-means						
1-gram	**0.991**	0.754	0.968	0.809	0.902	0.66
2-gram	0.951	0.958	0.968	0.809	0.902	0.66
3-gram	0.955	0.962	0.968	0.809	0.902	0.66
Kịch bản 2: Thuật toán DBSCAN						
1-gram	0.993	0.952	0.970	0.844	0.945	0.904
2-gram	0.982	0.949	0.975	0.526	0.945	0.904
3-gram	0.982	0.949	0.975	0.526	0.945	0.904
Kịch bản 3: Thuật toán ContextTracClus						
	0.982	**1**	**0.975**	**0.904**	**0.922**	**0.673**

Kết quả thực nghiệm cho thấy thuật toán ContextTracClus có hiệu quả cao khi so sánh với thuật toán phân cụm truyền thống đặc biệt là K-means trên hai độ đo Fitness và Precision, cũng như khả năng tự động phát hiện số cụm phù hợp; độ phức tạp và thời gian tính toán cũng được giảm đáng kể do thuật toán chỉ sử dụng hai vòng lặp (một vòng duyệt qua tất cả các vết khi xây dựng cụm và một vòng duyệt qua tất cả các cụm khi điều chỉnh cụm) thay vì sử dụng vòng lặp hội tụ như những thuật toán khác. Ngoài ra thuật toán sử dụng trực tiếp tệp NKSK đầu vào mà không phải qua bước biểu diễn lại vết.

Chương 5. Ứng dụng học sâu để sinh biểu diễn vết

5.1 Mạng nơ-ron học sâu trong biểu diễn vết

5.1.1 Mạng nơ-ron học sâu

Mạng nơ-ron học sâu là một thuật toán học máy được phát triển dựa trên mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN), cho phép máy tính có thể "học" ở các mức độ trừu tượng khác nhau. Ý tưởng cơ bản của các mạng nơ-ron là bắt chước các hoạt động của não người bằng cách sử dụng một số lượng lớn các nơ-ron kết nối với nhau để xử lý thông tin.

5.1.2 Biểu diễn vết cô đọng dựa trên mạng nơ-ron học sâu DNN

Một trong những mục đích của mạng nơ-ron học sâu là huấn luyện giá trị đầu vào X thành biểu diễn trung gian cô đọng (giá trị ẩn H) với biểu diễn mới và tốt hơn để dự đoán chính xác giá trị đầu ra Y. Trong nội dung chương 5, luận án đề xuất một phương pháp biểu diễn vết mới sử dụng kỹ thuật học có giám sát trong mạng nơ-ron học sâu DNN nhằm cải thiện hiệu quả của phương pháp biểu diễn vết trong nhật ký sự kiện. Thay vì sử dụng biểu diễn vết ban đầu, một *biểu diễn vết cô đọng* (*Compact Trace Representation*) sẽ được sử dụng để phân cụm.

5.2 Ứng dụng mô hình CBOW để sinh biểu diễn vết

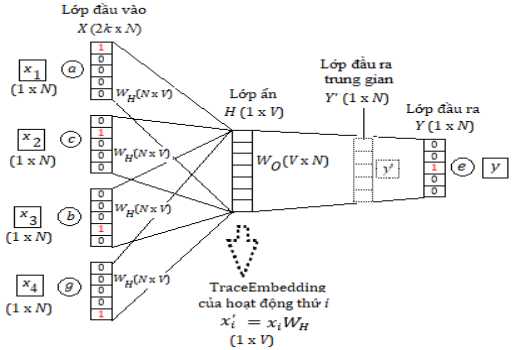
5.2.1 Giới thiệu về mô hình CBOW

CBOW là một mô hình tiên tiến trong phương pháp *nhúng từ* (*word embedding*), là phương pháp ánh xạ mỗi từ vào một không gian số thực nhiều chiều, hay một cách đơn giản là biểu diễn một văn bản dưới dạng một véc-tơ số và cho phép người dùng có thể khai thác mối quan hệ tiềm ẩn giữa các từ trong đoạn văn bản đó.

5.2.2 Phương pháp biểu diễn vết TraceEmbedding dựa trên mô hình CBOW

Theo cách thức hoạt động của mô hình word embedding CBOW, các ma trận trọng số lớp đầu vào - ẩn W_H được huấn luyện một cách tốt nhất để biến đổi các từ x thuộc lớp đầu vào

từ các giá trị ban đầu là các số nhị phân đơn giản thành các giá trị thực chứa đựng thông tin của các từ xung quanh nó với mục đích giúp mô hình dự đoán chính xác giá trị đầu ra Y. Như vậy thông tin của từ x sau khi biến đổi phong phú hơn rất nhiều. Đây chính là động lực để luận án đề xuất một phương pháp biểu diễn vết mới có tên là TraceEmbedding sử dụng mô hình CBOW nhằm nâng cao chất lượng bài toán biểu diễn vết. Mô hình huấn luyện được thiết kế như sau (Hình 5.4)



Hình 5.4. Mô hình CBOW trong biểu diễn vết

Ma trận trọng số lớp ẩn W_H sau khi huấn luyện xong sẽ được dùng để tạo *trace embedding* của một hoạt động. Gọi x_i là véc-tơ one-hot của hoạt động x ta có *trace embedding* tương ứng w của x được tính theo công thức sau:

$$w = x_i W_H \tag{5.6}$$

5.3 Mô hình LSTM trong bài toán biểu diễn vết

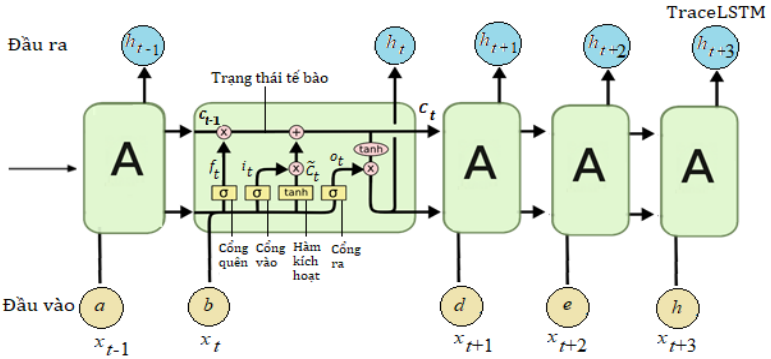
5.3.1 Giới thiệu về mô hình LSTM

Bộ nhớ dài-ngắn hạn (Long Short Term Memory - LSTM) là một dạng đặc biệt của mạng nơ-ron hồi quy RNN (Recurrent Neural Network), nó có khả năng khắc phục lỗi không học được các phụ thuộc xa của RNN.

5.3.2 Phương pháp biểu diễn vết TraceLSTM dựa trên mô hình LSTM

Trong mô hình Trace LSTM luận án thiết kế mỗi hoạt động hd trong một vết là đầu vào x của mô hình, hoạt động thứ t tương ứng với đầu vào x_t, ta có $x_t = embedding(hd_t)$ với kích thước $(1 \times N)$. Giá trị ẩn h_{t+1} đầu ra của trạng thái thứ

t trong quá trình huấn luyện cho kết quả là một véc-tơ kích thước $(1 \times V)$ chứa thông tin của t hoạt động trước. N và V là hai giá trị do người dùng thiết lập, trong quá trình thực nghiệm luận án sẽ thay đổi các giá trị này để tìm ra các tham số tốt nhất. Hình 5.7 dưới đây mô tả một phần của mô hình Trace LSTM huấn luyện vết $v = \langle abdeh \rangle$



Hình 5.7 Mô hình LSTM trong biểu diễn vết

5.4 Khung mô hình ứng dụng học sâu biểu diễn vết

5.4.1 Khung mô hình ứng dụng

Khung mô hình thực nghiệm được luận án đề xuất bao gồm 5 bước: Tiền xử lý vết; Áp dụng mô hình; Phân cụm vết; Phát hiện mô hình và Đánh giá mô hình.

5.4.2 Kết quả thực nghiệm

Để đánh giá hiệu quả các phương pháp biểu diễn vết sử dụng các mô hình học sâu đối với giải pháp phân cụm vết nâng cao chất lượng các mô hình quy trình, luận án tiến hành thực nghiệm theo bốn kịch bản được mô tả trong bảng 5.2 như sau:

Bảng 5.2 Kết quả thực nghiệm chất lượng mô hình quy trình sử dụng học sâu

Nhật ký sự kiện	Phương pháp biểu diễn vết	Độ đo			
		Dim	Time	Fitness	Precision
Lfull	*Kịch bản1: Các phương pháp biểu diễn vết truyền thống*				
	Túi hoạt động	**8**	**0.1s**	0.991	0.754
	k-grams	23	1.9s	0.955	0.962

	Lặp cực đại	50	2s	0.950	**1**
	Đồ thị khoảng cách	43	1.9s	0.992	**1**
	Kịch bản2: Biểu diễn vết sử dụng mạng nơ-ron học sâu DNN				
	Compact trace	50	17s	0.99995	0.794
	Kịch bản 3: Biểu diễn vết sử dụng mô hình CBOW				
	Trace embedding	80	60s	0.99995	0.822
	Kịch bản 4: Biểu diễn vết sử dụng mô hình LSTM				
	Trace LSTM	30	50m	**1**	0.980
prAm6	*Kịch bản1: Các phương pháp biểu diễn vết truyền thống*				
	Túi hoạt động	317	**0.3s**	0.968	0.809
	k-grams	2467	76s	0.968	0.809
	Lặp cực đại	9493	8h	0.968	0.332
	Đồ thị khoảng cách	1927	93s	0.968	0.809
	Kịch bản2: Biểu diễn vết sử dụng mạng nơ-ron học sâu DNN				
	Compact trace	**30**	43s	0.973	0.911
	Kịch bản 3: Biểu diễn vết sử dụng mô hình CBOW				
	Trace embedding	40	30m	0.973	0.911
	Kịch bản 4: Biểu diễn vết sử dụng mô hình LSTM				
	Trace LSTM	70	90m	**0.974**	**0.920**
prHm6	*Kịch bản1: Các phương pháp biểu diễn vết truyền thống*				
	Túi hoạt động	321	**0.2s**	0.902	0.660
	k-grams	730	9.8s	0.902	0.660
	Lặp cực đại	592	59s	0.897	0.730
	Đồ thị khoảng cách	1841	54s	0.902	0.660
	Kịch bản2: Biểu diễn vết sử dụng mạng nơ-ron học sâu DNN				
	Compact trace	**30**	37s	0.902	0.762
	Kịch bản 3: Biểu diễn vết sử dụng mô hình CBOW				
	Trace embedding	50	22m	0.922	0.730

	Kịch bản 4: Biểu diễn vết sử dụng mô hình LSTM				
	Trace LSTM	60	93m	**0.961**	**0.789**

Các kết quả thử nghiệm cho thấy, các phương pháp biểu diễn vết học sâu đã mang lại những hiệu quả hơn hẳn các phương pháp biểu diễn vết truyền thống với mức độ tối ưu giảm dần về chất lượng mô hình quy trình sinh ra gồm: Biểu diễn vết sử dụng mô hình LSTM > Biểu diễn vết sử dụng mô hình CBOW ≥ Biểu diễn vết sử dụng mô hình DNN > Các phương pháp biểu diễn vết truyền thống..

Kết luận: Những kết quả chính của luận án

Thứ nhất, luận án đề xuất một mô hình biểu diễn vết dựa trên đồ thị khoảng cách (một mô hình biểu diễn văn bản hiệu quả). Trong đó, mỗi vết được ánh xạ như là một đoạn văn bản, mỗi hoạt động trong vết được ánh xạ tương đương một từ trong đoạn văn bản. Trong cách biểu diễn này, các hoạt động tương ứng với các đỉnh của đồ thị, các cung là sự kết nối mối quan hệ giữa hai hoạt động. Đây là một biểu diễn trung gian tự nhiên có thể mô tả thông tin về thứ tự và khoảng cách bậc k giữa các hoạt động trong một vết, làm tăng tính linh hoạt, hiệu quả và cung cấp một giải pháp biểu diễn vết phong phú hơn cho bài toán phân cụm vết. Đồ khoảng cách có thể mô tả mối quan hệ với một khoảng cách k nhất định giữa các hoạt động trong một vết, do đó có thể nắm bắt cấu trúc tổng thể của nhật ký sự kiện từ đó giúp cải thiện chất lượng phân cụm vết. Mô hình biểu diễn vết dựa trên đồ thị khoảng cách đặc biệt phát huy hiệu quả đối với các tập nhật ký sự kiện có sự lặp lại của các hoạt động trong một vết.

Thứ hai, luận án đã đề xuất một độ đo mới - Độ đo trọng số vết chuẩn hóa **NTW** – cho phép xác định khoảng cách hay mức độ tương tự giữa các vết trong một nhật ký sự kiện sử dụng ý tưởng của độ đo Google. Dựa trên tần suất xuất hiện riêng lẻ và tần suất xuất hiện đồng thời của các cặp hoạt động trên toàn bộ nhật ký sự kiện, độ đo NTW cung cấp một phương pháp tính khoảng cách toàn cục giữa các vết trên bối cảnh so sánh chúng với tất cả các vết khác trong nhật ký sự

kiện. So với các phương pháp tính toán khoảng cách truyền thống chỉ xem xét mối quan hệ cục bộ giữa hai vết và bỏ qua sự hiện diện của tất cả các vết còn lại, NTW đã được chứng minh có tính hiệu quả cao hơn, đóng góp một phương pháp đo khoảng cách vết mới nâng cao chất lượng bài toán phân cụm vết trong lĩnh vực khai phá quy trình.

Thứ ba, luận án đề xuất một thuật toán phân cụm vết mới **ContextTracClus** dành riêng cho lĩnh vực khai phá quy trình, góp một phần công sức giải quyết thách thức đặt ra trong bản tuyên ngôn khai khá quy trình: "Khai phá quy trình không phải là một dạng cụ thể của khai phá dữ liệu do đó cần phải đề xuất các phương pháp và thuật toán riêng cho khai phá quy trình". Trong nghiên cứu này, luận án đã có những đóng góp mới về mặt lý thuyết gồm đề xuất hai khái niệm mới về ngữ cảnh vết và cây ngữ cảnh; đề xuất 3 thuật toán về xây dựng cây ngữ cảnh; xác định ngữ cảnh vết từ cây ngữ cảnh và cuối cùng là thuật toán phân cụm vết ContextTracClus cho phép phân cụm các vết thành tập các vết con có sự tương đồng nhau về ngữ cảnh thực hiện.

Thứ tư luận án đã nghiên cứu và đề xuất ba giải pháp ứng dụng những kết quả nghiên cứu tiên tiến trên thế giới về học sâu vào bài toán biểu diễn vết nhật ký sự kiện, gồm: giải pháp biểu diễn vết cô đọng **Compact trace** sử dụng mô hình mạng nơ-ron học sâu DNN; giải pháp biểu diễn vết **Trace embedding** sử dụng mô hình nhúng từ CBOW và giải pháp biểu diễn vết **Trace LSTM** sử dụng mô hình bộ nhớ dài-ngắn hạn LSTM. Với khả năng học và sự kết nối thông tin một cách chặt chẽ từ các mạng học sâu, các giải pháp biểu diễn vết Compact trace, Trace embedding và đặc biệt là Trace LSTM đã có những hiệu quả vượt bậc so với các phương pháp biểu diễn vết truyền thống. Những kết quả này đã góp một phần tích cực trong dòng nghiên cứu về các phương pháp biểu diễn vết, làm phong phú thêm những giải pháp biểu diễn vết hiệu quả trong lĩnh vực khai phá quy trình.

Danh mục công trình khoa học của Tác giả liên quan tới luận án

1. [BTHNhung01] Quang-Thuy Ha, Hong-Nhung Bui, and Tri-Thanh Nguyen (2016), "A trace clustering solution based on using the distance graph model", Proceeding of the 8th International Conference on Computational Collective Intelligence (ICCCI), Lecture Note of Artificial Intelligence (LNAI), Springer, pp. 313-320.
2. [BTHNhung02] Hong-Nhung Bui, Quang-Thuy Ha, and Tri-Thanh Nguyen (2018), "A Novel Similarity Measure for Trace Clustering Based on Normalized Google Distance", JP Journal of Heat and Mass Transfer, Special Volume, Issue III, Advances in Mechanical System and ICT-convergence, Pushpa publishing house, pp. 341-346.
3. [BTHNhung03] Hong-Nhung Bui, Tri-Thanh Nguyen, Thi-Cham Nguyen, and Quang-Thuy Ha (2018), "A new trace clustering algorithm based on context in process mining", Proceeding of International Joint Conference on Rough Sets (IJCRS), Lecture Note of Artificial Intelligence (LNAI), Springer, pp. 644-658.
4. [BTHNhung04] Hong-Nhung Bui, Trong-Sinh Vu, Tri-Thanh Nguyen, Thi-Cham Nguyen, and Quang-Thuy Ha (2019), "A Compact Trace Representation Using Deep Neural Networks for Process Mining", Proceeding of the 11th IEEE International Conference on Knowledge and Systems Engineering (KSE), pp. 312-316.
5. [BTHNhung05] Hong-Nhung Bui, Trong-Sinh Vu, Hien-Hanh Nguyen, Tri-Thanh Nguyen, and Quang-Thuy Ha (2020), "Exploiting CBOW and LSTM Models to Generate Trace Representation for Process Mining", Proceeding of the 12th Asian Conference on Intelligent Information and Database Systems (ACIIDS), pp. 35-46.

Hà Nội – 2020