

Xây dựng Giải thuật Cải tiến cho Bài toán Khai thác Tập Phổ biến và Ứng dụng

Nguyễn Thành Trung

Khoa Khoa học Máy tính, Đại học Công nghệ Thông tin,
Đại học Quốc gia TP. HCM, Việt Nam
{nguyen_thanh_trung_key@yahoo.com.vn}

Tóm tắt. Khai thác tập phổ biến là một vấn đề cơ bản trong khai thác dữ liệu và khám phá tri thức. Các tập phổ biến được khám phá có thể được sử dụng như đầu vào cho việc phân tích luật kết hợp, khai thác mẫu tuần tự, nhận diện cụm, phân lớp, ... Tuy nhiên, quá trình khai thác tập phổ biến trong các tập dữ liệu có kích cỡ lớn là một tác vụ cực kỳ tốn thời gian, đặc biệt là trên các cơ sở dữ liệu có nhiều biến động (các thao tác thêm, xóa, sửa diễn ra thường xuyên). Vượt qua trở ngại này là mục đích nghiên cứu chính của luận văn tiến sĩ mà nghiên cứu sinh phải hoàn thành. Báo cáo này nhằm trình bày chi tiết mục tiêu nghiên cứu, tổng quan tình hình nghiên cứu trên Thế giới, kế hoạch thực hiện, cũng như các kết quả nghiên cứu hiện đã đạt được của nghiên cứu sinh.

Từ khoá: khai thác dữ liệu, khai thác tập phổ biến, khám phá tri thức.

1 Đặt Vấn đề

1.1 Giới thiệu

Các tập phổ biến là các tập mục, các chuỗi con hay các cấu trúc con xuất hiện trong một tập dữ liệu với tần suất không nhỏ hơn một ngưỡng do người dùng xác định. Thí dụ, một tập hợp các mục như sữa và bánh mì xuất hiện cùng với nhau một cách thường xuyên trong một tập dữ liệu giao tác, là một *tập mục phổ biến*. Một chuỗi con, như việc đầu tiên mua một máy tính cá nhân, sau đó mua một máy quay phim kỹ thuật số và tiếp sau là mua một thẻ nhớ, nếu điều này diễn ra một cách thường xuyên trong một cơ sở dữ liệu ghi nhận việc mua hàng, là một *mẫu liên tục (phổ biến)*. Một *cấu trúc con* có thể dùng để chỉ các dạng cấu trúc khác nhau, như các đồ thị con, các cây con, hay các dàn con, những cấu trúc này có thể kết hợp với các tập mục hay các chuỗi con. Nếu một cấu trúc con xuất hiện thường xuyên trong một cơ sở dữ liệu đồ thị, nó được gọi là một *mẫu cấu trúc (phổ biến)*. Việc tìm kiếm các tập phổ biến đóng một vai trò thiết yếu trong việc khai thác sự kết hợp, sự tương quan, và nhiều mối quan hệ thú vị khác giữa các dữ liệu. Hơn nữa, nó còn giúp ích trong việc chỉ mục, phân lớp, gom cụm dữ liệu và các tác vụ khai thác dữ liệu khác. Ngày nay, nó còn được phát triển và ứng dụng vào trong một lĩnh vực nghiên cứu mới, mạng xã hội. Vì

vậy, khai thác tập phổ biến đã trở thành một tác vụ khai thác dữ liệu quan trọng và một chủ đề được tập trung trong nghiên cứu khai thác dữ liệu.

Khai thác tập phổ biến được đề xuất lần đầu tiên bởi Agrawal và các cộng sự (1993) cho việc phân tích giỏ thị trường trong hình thức khai thác luật kết hợp. Hình thức này phân tích thói quen mua hàng của khách hàng bằng cách tìm ra sự kết hợp giữa các tập mục khác nhau mà khách hàng chọn mua trong “giỏ mua hàng” của họ. Thí dụ, nếu khách hàng định mua sữa thì khả năng họ cũng sẽ mua thức ăn làm từ ngũ cốc (và các loại ngũ cốc) trong cùng chuyến đi mua hàng ở siêu thị là như thế nào? Những thông tin như vậy có thể giúp tăng doanh số bán bằng cách giúp các nhà bán lẻ thực hiện các chiến dịch marketing có lựa chọn và sắp xếp không gian các kệ hàng của họ.

1.2 Những Thách thức

Từ sự đề xuất đầu tiên về tác vụ khai thác tập phổ biến mới này và các giải thuật khai thác hiệu quả đi kèm của nó, đã có hàng trăm công trình nghiên cứu tiếp nối trên nhiều loại ứng dụng và mở rộng khác nhau, trải dài từ các hệ phương pháp luận khai thác dữ liệu có thể phát triển được cho tới việc xử lý một sự khác biệt lớn các loại dữ liệu, các tác vụ khai thác mở rộng đa dạng, và sự phong phú của các ứng dụng mới. Trải qua hai thập niên của nhiều nghiên cứu quan trọng và thành công, hiện nay, trên Thế giới, các phương pháp tìm tập phổ biến được chia thành các nhóm sau:

- Các phương pháp sinh ứng viên: phương pháp Apriori do Agrawal đề xuất và các giải thuật dựa Apriori như: AprioriTID, AprioriHybrid, DIC, DHP, PHP, ...
- Các phương pháp không sinh ứng viên: phương pháp của J. Han dựa trên cây FP để khai thác các tập phổ biến, hay các phương pháp như Lcm, DCI, ...
- Phương pháp sử dụng định dạng dữ liệu dọc: phương pháp của Zaki dựa trên cây IT và phân giao của các Tidset để tính độ phổ biến.
- Các phương pháp song song hoá các thuật toán đã có nhằm tăng cường hiệu suất và giảm thời gian thực thi.

Tuy nhiên, vẫn còn đó nhiều thách thức được đặt ra hiện tại. Trong đó, thách thức lớn nhất là việc xây dựng một giải thuật tăng cường nhằm tìm kiếm các tập phổ biến trên một cơ sở dữ liệu nhiều biến động (các thao tác thêm, xóa, sửa diễn ra thường xuyên). Khi biến động xảy ra, các tập phổ biến đã tìm được được tự động cập nhật mà không cần chạy lại giải thuật từ đầu.

Ngoài ra, mặc dù có một số thuật toán hiệu quả nhưng cơ sở toán học và quá trình cài đặt của chúng lại khá phức tạp và nặng nề, cộng vào đó là hạn chế của bộ nhớ vật lý trên các máy tính cá nhân.

Bên cạnh đó, đặc thù của thị trường kinh doanh ở Việt Nam cũng đặt ra các thách thức cần giải quyết:

- Phần lớn các doanh nghiệp có quy mô vừa và nhỏ: trang trí nội thất, kinh doanh ăn uống, đào tạo tin học, ngoại ngữ, doanh nghiệp xe máy, ... Số chủng loại mặt hàng mà họ kinh doanh khoảng 200 đến 1000. Trường hợp lên đến 1000 mặt hàng, thường là các siêu thị mua sắm, đa dạng các chủng loại như: ăn uống, kim khí điện máy, trang điểm, nội trợ, ...

- Các hoá đơn mua bán có một nguyên tắc chung là số chủng loại mặt hàng bán ra khoảng 20 (mẫu mã hoá đơn theo quy định của Nhà nước).
- Thực tế cho thấy các luật chỉ ra khuynh hướng mua sắm của thị trường một cách tốt nhất được rút ra từ khối lượng hoá đơn có thời gian tích lũy từ 6 tháng đến 1 năm trước đó. Vì nhu cầu thị trường hiện nay cũng như mẫu mã, chức năng công dụng của hàng hoá thay đổi rất nhanh và liên tục, nên không thể dùng các luật của các năm trước để áp dụng cho hiện tại.
- Không thể áp dụng các luật từ các doanh nghiệp khác vì xu hướng của thị trường ở các khu vực khác nhau là khác nhau (ví dụ: ở Quận 3, người tiêu dùng thích mua xe máy của hãng Honda, còn ở Quận Tân Phú thì người tiêu dùng lại thích xe của hãng Yamaha).
- Các doanh nghiệp bắt buộc phải thay đổi thường xuyên ngưỡng hỗ trợ tối thiểu để tìm ra các luật chấp nhận được dựa trên số lượng người mua.
- Do đặc thù quản lý của các doanh nghiệp Việt Nam, các thao tác thêm, xoá, sửa tác động thường xuyên lên cơ sở dữ liệu.
- Nhu cầu tích lũy kết quả ngay sau mỗi thao tác trên hoá đơn để có thể tham khảo các luật bất kỳ lúc nào.

1.3 Mục tiêu Nghiên cứu

Với những thách thức được đặt ra, từ trong và ngoài nước, mục tiêu nghiên cứu của luận văn là nhằm vào việc xây dựng một không gian toán học chặt chẽ với đầy đủ các định nghĩa, mệnh đề, hệ quả, định lý, thuật toán, chứng minh ... để áp dụng vào việc giải bài toán khai thác tập phổ biến với mục tiêu đặt ra là giải thuật được xây dựng mới phải có tính dễ hiểu, có độ phức tạp thấp, giải quyết được những thách thức hiện đang còn tồn tại, bổ sung thêm những ý tưởng mới (như độ đo về độ khách quan của tập phổ biến) để từ đó dễ cài đặt và đưa vào ứng dụng đạt được hiệu quả cao nhất.

Nói tiếp những kết quả trên, một yêu cầu đặt ra là giải thuật tăng cường này phải có thể được song song hoá theo chiều dọc hoặc chiều ngang của dữ liệu đầu vào nhằm hỗ trợ tốt nhất cho kỹ thuật lập trình cũng như hệ thống máy tính song song hoá.

Một khả năng mở ra là có thể áp dụng thuật toán cải tiến này vào bài toán rút gọn thuộc tính của hệ thống thông tin nhằm tìm ra những thuộc tính đặc trưng và quan trọng nhất, cũng như loại bỏ đi các thuộc tính dư thừa, không cần thiết, đây là bài toán hiện tại cũng đang đặt ra nhiều thách thức cho giới nghiên cứu trong lĩnh vực Lý thuyết Tập thô do Pawlak khởi xướng vào năm 2003.

2 Tổng quan Tình hình Nghiên cứu trên Thế giới

Vào năm 1993, Agrawal và các cộng sự đã khởi xướng vấn đề khai thác tập phổ biến. Vào thời điểm đó, không biết Agrawal có tiên đoán trước được tương lai hay không, nhưng trải qua 2 thập kỷ hình thành và phát triển, khai thác tập phổ biến đã phát triển mạnh mẽ từ lý thuyết đến ứng dụng, có tác động sâu rộng vào hầu khắp các lĩnh vực trong cuộc sống, không chỉ trong lĩnh vực công nghệ thông tin và kinh tế mà còn

trong hoá học, sinh học và thậm chí các lĩnh vực khoa học xã hội như phân tích tâm lý tội phạm.

Phần này thực hiện một sự miêu tả tổng quan về các phương thức, sự mở rộng và các ứng dụng khai thác tập phổ biến, và được tổ chức thành 4 giai đoạn sau:

- Giai đoạn 1993 đến 1997: Các hệ phương pháp luận nền tảng
- Giai đoạn 1998 đến 2004: Cải tiến và phát triển
- Giai đoạn 2003 đến 2007: Các ứng dụng
- Giai đoạn 2008 đến 2012: Các xu hướng nghiên cứu mới

2.1 Giai đoạn 1993-1997: Các Hệ Phương pháp luận Nền tảng

Sau đây là các định nghĩa của Agrawal để chuẩn bị cho việc nghiên cứu của ông: Cho $I = \{i_1, i_2, \dots, i_n\}$ là tập hợp của tất cả các mục. Một tập-mục- k α , bao gồm k phần tử từ I , là phổ biến nếu α xuất hiện trong một cơ sở dữ liệu giao tác D không ít hơn $\theta|D|$ lần, với θ là một ngưỡng hỗ trợ tối thiểu do người dùng xác định (trong báo cáo này gọi là min_sup), và $|D|$ là tổng số lượng các giao tác trong D .

Nguyên lý Apriori, giải thuật Apriori và các mở rộng của nó

Agrawal và Srikant (1994) đã quan sát một thuộc tính *đồng hướng xuống*, được gọi là Apriori, được phát biểu như sau: *Một tập-mục- k là phổ biến chỉ khi các tập mục con của nó là phổ biến.* Đây là điều cốt lõi của giải thuật Apriori.

Từ khi giải thuật Apriori được đề xuất, có nhiều nghiên cứu rộng rãi về sự cải thiện và mở rộng của Apriori, thí dụ như: kỹ thuật băm (Park và các cộng sự 1995), kỹ thuật phân hoạch (Savasere và các cộng sự 1995), phương pháp lấy mẫu (Toivonen 1996), khai thác tăng cường (Cheung và các cộng sự 1996), cách đếm tập mục động (Brin và các cộng sự 1997), khai thác phân bố và song song (Park và các cộng sự 1995; Agrawal và Shafer 1996; Cheung và các cộng sự 1996; Zaki và các cộng sự 1997).

Khai thác các tập mục phổ biến sử dụng định dạng dữ liệu dọc

Phương thức Apriori khai thác các tập phổ biến theo *định dạng dữ liệu ngang*, ngoài ra việc khai thác cũng có thể được thực hiện với dữ liệu được trình bày theo *định dạng dọc*. Một công trình khai thác các tập mục phổ biến với định dạng dữ liệu dọc đầu tiên là (Holsheimer và các cộng sự 1995).

Khai thác các luật kết hợp đa cấp độ, đa kích cỡ và có định lượng

Dữ liệu được lưu trữ với số lượng rất lớn và thời gian dài, thường tính theo năm. Do đó, khi khai thác, không phải lúc nào cũng phải sử dụng hết toàn bộ số lượng dữ liệu hiện có, mà cần phải xác định thời điểm lấy, số lượng cần lấy cho phù hợp.

Các tập phổ biến ở thời điểm bắt đầu chủ yếu là dựa trên các mục rời rạc, như là tên mục, loại sản phẩm, và vị trí. Tuy nhiên, các nhà nghiên cứu đã tìm ra các tập phổ biến cho các thuộc tính số, như là lương, tuổi, và điểm.

Khai thác các mẫu liên tục

Có nhiều ứng dụng liên quan đến khai thác các mẫu liên tục, như là chuỗi mua hàng của khách hàng, các chuỗi click chuột trên Web, và các chuỗi sinh học. *Khai thác mẫu liên tục* được giới thiệu lần đầu tiên bởi Agrawal và Srikant (1995). Giải thuật GSP đã được đề xuất bởi Srikant và Agrawal (1996). Năm 1997, Mannila và các cộng sự đề xuất xét các episode phổ biến.

Khai thác các mẫu có cấu trúc: đồ thị, cây, và dàn

Nhiều ứng dụng khoa học và thương mại cần các mẫu phức tạp như: các cây, các dàn, và các đồ thị. Như một cấu trúc dữ liệu phổ biến, đồ thị đã trở nên ngày càng quan trọng trong các ứng dụng rộng lớn bao gồm hoá-tin-học, sinh-tin-học, thị lực của máy tính, chỉ mục video, phục hồi văn bản, và phân tích Web. Nghiên cứu đầu tiên về phương thức khai thác mẫu đồ thị phổ biến: Holder và các cộng sự (1994).

Các độ đo của luật kết hợp

Khái niệm về luật kết hợp đã được giới thiệu cùng với tập phổ biến bởi (Agrawal và các cộng sự 1993). Dựa trên định nghĩa của luật kết hợp, hầu hết các nghiên cứu đều nhận quá trình khai thác tập phổ biến như là bước đầu tiên và thiết yếu trong việc khai thác luật kết hợp. Một luật kết hợp có dạng $\alpha \Rightarrow \beta$, *độ hỗ trợ* cùng với *độ tin cậy* là hai độ đo độ quan tâm của luật. Tuy nhiên, không phải tất cả các luật kết hợp được tạo ra đều đáng quan tâm, đặc biệt là khi việc khai thác ở một ngưỡng hỗ trợ thấp hay khai thác các tập dài. Điều này dẫn đến các luật tương quan có dạng $\alpha \Rightarrow \beta$ [*độ hỗ trợ, độ tin cậy, độ tương quan*]. Một trong những nghiên cứu đầu tiên là: Piatetsky-Shapiro 1991; Brin và các cộng sự (1997).

2.2 Giai đoạn 1998-2004: Cải tiến và Phát triển

Cải tiến và phát triển Nguyên lý Apriori

Sarawagi và các cộng sự (1998) đã kết hợp việc khai thác với các hệ thống cơ sở dữ liệu quan hệ. Ngoài ra còn có sự đóng góp của Geerts và các cộng sự (2001).

Khai thác các tập mục phổ biến không cần tạo ứng viên

Một phương pháp khai thác tập phổ biến mới được Han và các cộng sự (2000) phát minh, phương thức FP-growth (tạm dịch là “phát triển mẫu”) có thể khai thác tập hợp đầy đủ các tập mục phổ biến mà không cần tạo ra ứng viên. Tiếp theo, từ 2001-2003, đã có nhiều phương thức thay thế và mở rộng đối với cách tiếp cận phát triển mẫu.

Cải tiến và phát triển Khai thác các tập mục phổ biến sử dụng định dạng dữ liệu dọc

Phát triển phương pháp khảo sát định dạng dữ liệu dọc, vào năm 2000, Zaki đã đề xuất giải thuật Eclat.

Khai thác các tập mục phổ biến đóng và tối đại

Khai thác các tập mục phổ biến đóng được đề xuất lần đầu tiên bởi Pasquier và các cộng sự (1999). Đến năm 2003, Goethals và Zaki đã có báo cáo khẳng định về việc khai thác tập mục đóng đạt được hiệu quả tốt hơn.

Khai thác các tập tối đại được nghiên cứu lần đầu tiên bởi Bayardo (1998), với giải thuật MaxMiner.

Khai thác các tập dữ liệu có kích cỡ lớn và khai thác các mẫu không lồ

Trong y học, với phương pháp quét phổ các chuỗi gen để chẩn đoán ung thư đã tạo ra các tập dữ liệu sinh học có kích cỡ lớn. Đây là một thách thức lớn cho các giải thuật khai thác tập mục phổ biến (đóng) hiện tại. Năm 2003, Pan và các cộng sự lần đầu tiên đã đề xuất CARPENTER để giải quyết thách thức trên.

Cải tiến và phát triển Khai thác các mẫu liên tục

Năm 2001, Zaki đã phát triển một phương thức khai thác mẫu liên tục dựa-vào-định-dạng-đứng gọi là SPADE. Cũng trong năm 2001, Pei và các cộng sự đã giới thiệu một cách tiếp cận phát-triển-mẫu đối với việc khai thác mẫu liên tục, có tên gọi là PrefixSpan. Ngoài ra, các nhà khoa học còn phát triển khái niệm các chuỗi con đóng. Và nghiên cứu đầu tiên về vấn đề này do Yan và các cộng sự thực hiện vào năm 2003.

Cải tiến và phát triển Khai thác các mẫu có cấu trúc: đồ thị, cây, và dàn

Ở giai đoạn này, đã hình thành hai cách tiếp cận cơ bản đối với bài toán khai thác cấu trúc con phổ biến: một cách tiếp cận dựa-trên-Apriori và một cách tiếp cận phát-triển-mẫu.

Khai thác dựa-trên-sự-ràng-buộc

Vì không phải tất cả các tập phổ biến đều đáng quan tâm nên người dùng có quyền đưa các ràng buộc của mình vào nhằm giảm bớt việc hình thành các tập không như ý muốn.

Khai thác các mẫu nén hay xấp xỉ

Để giảm tập hợp khổng lồ các tập phổ biến được tạo ra trong quá trình khai thác dữ liệu trong khi vẫn duy trì chất lượng cao của chúng, các nghiên cứu đã tập trung vào việc khai thác một tập hợp các tập phổ biến nén hoặc xấp xỉ. Tổng quát, việc nén mẫu có thể được chia thành hai loại: nén không mất thông tin và nén bị mất thông tin, đây nói về mặt thông tin mà tập kết quả chứa, so sánh với toàn thể tập hợp các tập phổ biến.

Cải tiến và phát triển Các độ đo của luật kết hợp

Ngoài các nghiên cứu được phát triển ở trên đây, các nghiên cứu cũng hướng đến việc khai thác các tập gây ngạc nhiên hoặc đáng quan tâm so với các tri thức đã biết trước của người sử dụng.

Cuối cùng ở giai đoạn này là các tác động mạnh mẽ của tập phổ biến đến việc phân tích dữ liệu và các ứng dụng khai thác như: phân lớp dựa-trên-tập-phổ-biến, phân tích gom cụm dựa-trên-tập-phổ-biến, tính toán khối lập phương, khai thác mối liên hệ và phân tích sự khác biệt.

2.3 Giai đoạn 2003-2007: Các Ứng dụng

Có rất nhiều ứng dụng liên quan đến tập phổ biến, báo cáo này chỉ xin trình bày các ứng dụng đặc biệt quan trọng

Chỉ mục và tìm kiếm sự tương đồng của dữ liệu có cấu trúc phức tạp

Các đối tượng phức tạp như chuỗi giao tác, các ghi nhận sự kiện, các Protein và hình ảnh được sử dụng rộng rãi trong nhiều lĩnh vực. Việc tìm kiếm hiệu quả các đối tượng này trở thành một vấn đề cấp thiết đối với nhiều ứng dụng.

Khai thác dữ liệu đa truyền thông và dữ liệu về không-gian-thời-gian

Dữ liệu về không gian như là bản đồ, hay dữ liệu hình ảnh y khoa, và dữ liệu bố cục chip điện tử VLSI. Dữ liệu về không-gian-thời-gian như là động lực học thời tiết, các đối tượng di chuyển, hay các phát triển vùng. Dữ liệu đa truyền thông là âm thanh, video, hình ảnh, đồ thị, lời nói, văn bản, tài liệu, và dữ liệu siêu văn bản.

Khai thác luồng dữ liệu

Khối lượng các luồng dữ liệu rất khủng khiếp và có thể là vô tận, thường được tạo ra bởi các hệ thống giám sát thời-gian-thực, các mạng truyền thông, lưu lượng Internet, các giao tác trực tuyến trong thị trường tài chính hay ngành công nghiệp bán lẻ, các lưới năng lượng điện, các quá trình sản xuất công nghiệp, các thí nghiệm khoa học kỹ thuật, các cảm biến từ xa, và các môi trường động lực khác.

Khai thác web

Có ba loại khai thác web khác nhau: khai thác nội dung web, khai thác cấu trúc web, và khai thác cách sử dụng web.

Khai thác lỗi phần mềm và bộ nhớ đệm hệ thống

Khai thác tập phổ biến đã đóng một vai trò rất quan trọng trong sự phân tích và phát hiện lỗi phần mềm.

2.4 Giai đoạn 2008-2012: Các Xu hướng Nghiên cứu Mới

Các giải thuật mới

Trong giai đoạn này, chủ đề được nghiên cứu rộng rãi và tập trung nhất trong khai thác tập phổ biến vẫn là các phương thức khai thác có thể phát triển được.

Các tập phổ biến xấp xỉ

Xu hướng hiện tại cũng đang xem các tập phổ biến xấp xỉ là lựa chọn tốt nhất trong nhiều ứng dụng, thí dụ, trong việc phân tích các chuỗi DNA hay Protein.

Các phương thức khai thác dựa trên mẫu

Như đã trình bày, phương thức khai thác phát triển mẫu đã thể hiện rõ thế mạnh của nó trong việc giảm số lần quét tập dữ liệu. Do đó nhiều nghiên cứu mới cũng đã cố gắng phát triển hơn nữa các phương thức khai thác dựa-trên-mẫu.

Ngữ nghĩa và ngữ cảnh của tập phổ biến

Nhiều nghiên cứu phát triển các kỹ thuật để diễn giải và hiểu sâu về các mẫu, chẳng hạn, sự chú giải về ngữ nghĩa cho các tập phổ biến, và sự phân tích ngữ cảnh của các tập phổ biến. Ngữ nghĩa của một tập phổ biến bao gồm các thông tin sâu hơn: ý nghĩa của mẫu là gì; các mẫu đồng nghĩa là gì; và các giao tác đặc thù mà mẫu này đang tồn tại trong đó là gì? Một phân tích ngữ cảnh của các tập phổ biến có thể giúp trả lời các câu hỏi như “tại sao tập này phổ biến?”. Một trả lời thí dụ có thể là “tập này phổ biến vì nó xảy ra quá nhiều trong suốt thời gian từ T1 đến T2”.

Các ứng dụng

Các ứng dụng thường tạo ra các kết quả nghiên cứu mới và mang lại sự hiểu biết sâu sắc về điểm mạnh, yếu của một giải pháp hiện tại. Xu hướng của các ứng dụng hiện nay chủ yếu tập trung vào ngành sinh-tin-học, mạng xã hội và phát hiện sự lây lan của virus trong hệ thống mạng máy tính.

3 Kế hoạch Nghiên cứu

3.1 Cơ sở Lý thuyết

Dựa trên các hệ phương pháp luận về Khai phá Dữ liệu và Lý thuyết Tập thô để nắm bắt kỹ cơ sở lý thuyết từ đó nhận định chính xác được bản chất của những khó khăn, thách thức đang đặt ra, đồng thời kết hợp việc nghiên cứu các tài liệu tham khảo có liên quan đến đề tài nghiên cứu trong lĩnh vực Tin học và Toán học để lần lượt thực hiện các bước:

- Tiến hành nghiên cứu cơ sở lý thuyết, nắm bắt những vấn đề cần phải cải thiện và có thể cải thiện được.
- Tiến hành phân tích và đề ra các hướng giải quyết. Dự kiến dựa trên không gian toán học của các chuỗi bit để tận dụng được thế mạnh về khả năng xử lý cực kỳ hiệu quả trên các dây bit nhị phân của hệ thống máy tính. Từ đó, có thể giảm được độ phức tạp của giải thuật và tăng hiệu suất của hệ thống.
- Viết các chương trình máy tính cho từng giải pháp.

- Thử nghiệm để loại bỏ những hướng giải quyết không hiệu quả cũng như không đạt được mục tiêu đề ra.
- Bổ sung những ý tưởng mới nhằm phát huy tốt hơn hiệu quả đạt được và đi đến mục tiêu cuối cùng.

3.2 Dữ liệu Thực nghiệm

Để đảm bảo tính khách quan và tính khoa học, dữ liệu thực nghiệm, một mặt, sẽ được lấy từ các hoá đơn kinh doanh của các doanh nghiệp vừa và nhỏ trong thực tế như các trung tâm đào tạo ngoại ngữ, tin học, các doanh nghiệp về ẩm thực, kinh doanh xe máy, ... Mặt khác dữ liệu thực nghiệm cũng sẽ được lấy từ các nhóm nghiên cứu có uy tín trên Thế giới, một thí dụ: nhóm nghiên cứu IBM Almaden Quest của tập đoàn IBM.

Từ các dữ liệu thu được cũng cần có các chương trình máy tính làm nhiệm vụ tiền xử lý và xử lý dữ liệu nhằm quy chuẩn chúng để chuẩn bị cho giai đoạn chạy thực nghiệm.

Dữ liệu thực tế và học thuật sẽ cùng được thực nghiệm để so sánh các kết quả thu được nhằm tạo cơ sở đánh giá hiệu quả của giải thuật mới.

3.3 Thiết bị Thực nghiệm

Dựa trên dữ liệu thực nghiệm, các giải thuật được đề xuất cần được cài đặt trên 2 máy tính cá nhân có cấu hình như sau: Intel(R) Core(TM) i3-2100 CPU @ 3.10GHz (4 CPUs), ~3.1GHz; và bộ nhớ RAM 4096MB; hệ điều hành Windows 7 Ultimate 64-bit (6.1, Build 7601) Service Pack 1; ngôn ngữ lập trình C#.NET.

4 Các Kết quả Nghiên cứu Đã Đạt được

Bài báo [50] đã đề xuất được một mô hình toán học trên cơ sở các chuỗi nhị phân với các định nghĩa, mệnh đề, hệ quả, định lý, chứng minh. Từ đó, bài báo cũng đã đề xuất được một giải thuật cải tiến cho bài toán Khai thác tập phổ biến (kèm theo đó là việc chứng minh tính đúng đắn của giải thuật) phần nào có thể giải quyết những thách thức hiện đang tồn tại trên Thế giới, đặc biệt khi áp dụng giải thuật, không cần phải “quét” cơ sở dữ liệu nhiều lần. Hội nghị có chỉ mục ISTP (ISI Thomson Proceedings) – IEEE.

Bài báo [51] đã đề xuất được một khái niệm mới, độ đo độ khách quan của một tập phổ biến, đây là tỉ lệ phần trăm giữa những chủ thể tham gia hình thành nên tập phổ biến đó với tất cả các chủ thể hình thành nên cơ sở dữ liệu. Cụ thể hơn, độ đo độ khách quan giúp các doanh nghiệp có thể hiểu rõ rằng: một tập phổ biến thỏa một ngưỡng hỗ trợ cho trước có được hình thành bởi nhu cầu mua sắm của đa số khách hàng hay chỉ được tạo ra từ một số ít khách hàng có hành vi mua sắm với số lượng mặt hàng cực lớn. Từ đó, doanh nghiệp có thêm thông tin để quyết định có nên theo những quy luật được hình thành từ một tập phổ biến cụ thể nào đó hay không. Cùng với khái niệm mới này, bài báo cũng đề xuất một giải thuật để tìm ra độ đo độ khách

quan của các tập phổ biến. Hội nghị có ISSN, và sau đó, bài báo đã được lựa chọn hiệu chỉnh, mở rộng để đăng trong sách có nhan đề “Intelligent Automation and Systems Engineering” được xuất bản bởi Springer.

Trong quá trình nghiên cứu mở rộng, thử nghiệm các hướng tiếp cận khác nhau để giải quyết mục tiêu của luận văn đề ra, bài báo [52] được hình thành. Hội nghị có ISSN.

Trong quá trình tìm hiểu về Lý thuyết Tập thô nhằm giải quyết mục tiêu nghiên cứu của luận văn đã đề ra (như đã đề cập trong phần trên), bài báo [53] đã đề xuất được một mô hình toán học và giải thuật để góp phần giải quyết những thách thức hiện có của bài toán Rút gọn thuộc tính trong một hệ thống thông tin do Zdzislaw Pawlak khởi xướng. Đây chính là nền tảng, tiền đề cho việc áp dụng thuật toán cải tiến tìm được cho bài toán Khai thác tập phổ biến vào bài toán Rút gọn thuộc tính. Hội nghị có chỉ mục ISTP – Springer. Hiện tại, bài báo đã nhận được lời mời hiệu chỉnh và mở rộng để đăng trên tạp chí “Journal of Theoretical and Applied Computer Science” được xuất bản bởi Polish Academy of Sciences.

Phát huy kết quả đạt được từ bài báo [53], bài báo [54] sử dụng giải thuật đã có kết hợp với các xác suất hậu nghiệm để đi sâu phân tích đặc trưng của khách hàng nhằm tìm kiếm các ý tưởng mới phục vụ cho luận văn từ các kết quả mở rộng này. Hội nghị có chỉ mục ISTP – IEEE.

Một mô hình toán học và giải thuật đã được đề xuất nhằm tạo ra một phương pháp tiếp cận mới cho bài toán Khai thác mẫu liên tục (phổ biến) trong bài báo [55]. Đây là kết quả thu được trong quá trình nghiên cứu giải thuật cải tiến cho bài toán Khai thác tập phổ biến. Hội nghị có chỉ mục ISTP – Springer.

Bài báo [56] đã hoàn thiện giải thuật cải tiến cho bài toán Khai thác tập phổ biến đồng thời khắc phục được những thách thức mang tính đặc thù của thị trường kinh doanh ở Việt Nam và có thể áp dụng cho cơ sở dữ liệu có nhiều biến động (các thao tác, thêm, xóa, sửa diễn ra thường xuyên). Hội nghị có chỉ mục CPCI-S. Bài báo sẽ được xuất bản bởi Springer Lecture Notes Information Technology.

5 Kết luận và Công việc Tương lai

Báo cáo đã trình bày chi tiết về mục tiêu nghiên cứu, cũng như tổng quan tình hình nghiên cứu trên Thế giới thành 4 giai đoạn chủ yếu. Bên cạnh đó, kế hoạch thực hiện, và các kết quả nghiên cứu hiện đã đạt được cũng được đề cập trong báo cáo. Tiếp nối các kết quả đã có, hiện có 2 nghiên cứu đang được thực hiện và dự kiến sẽ thu được kết quả trong tương lai gần:

- Từ giải thuật cải tiến cho bài toán Khai thác tập phổ biến đã có, tiếp tục xây dựng các mô hình và giải thuật cần có cho việc thực thi giải thuật trên các hệ thống máy tính song song hóa, đặc biệt việc thực thi song song có thể theo cả chiều dọc hoặc chiều ngang của dữ liệu đầu vào.
- Ứng dụng giải thuật cải tiến cho bài toán Khai thác tập phổ biến đã có vào bài toán Rút gọn thuộc tính trong Lý thuyết Tập thô.

Tài liệu tham khảo

1. Agrawal R, Imielinski T, Swami A (1993) *Mining association rules between sets of items in large databases*. In: Proceedings of the 1993 ACM-SIGMOD international conference on management of data (SIGMOD'93), Washington, DC, pp 207–216
2. Agrawal R, Shafer JC (1996) *Parallel mining of association rules: design, implementation, and experience*. IEEE Trans Knowl Data Eng 8:962–969
3. Agrawal R, Srikant R (1994) *Fast algorithms for mining association rules*. In: Proceedings of the 1994 international conference on very large data bases (VLDB'94), Santiago, Chile, pp 487–499
4. Agrawal R, Srikant R (1995) *Mining sequential patterns*. In: Proceedings of the 1995 international conference on data engineering (ICDE'95), Taipei, Taiwan, pp 3–14
5. Appice A., Ceci M., Malerba ATD. (2011) – *A parallel, distributed algorithm for relational frequent pattern discovery from very large datasets* – In: Intelligent Data Analysis 15 (2011) pp. 69–88.
6. Asai T, Abe K, Kawasoe S, Arimura H, Satamoto H, Arikawa S (2002) *Efficient substructure discovery from large semi-structured data*. In: Proceedings of the 2002 SIAM international conference on data mining (SDM'02), Arlington, VA, pp 158–174
7. Bahel M, Dule C (2010) *Analysis of frequent itemset generation process in apriori and RCS (reduced candidate set) algorithm*. In: Special Issue - NCICT'10 - New Horizon College, Bangalore, Volume: 02, Issue: 02, Sep - Oct 2010.
8. Bayardo RJ (1998) *Efficiently mining long patterns from databases*. In: Proceeding of the 1998 ACM-SIGMOD international conference on management of data (SIGMOD'98), Seattle, WA, pp 85–93
9. Blanchard J, Guillet F, Gras R, Briand H (2005) *Using information-theoretic measures to assess association rule interestingness*. In: Proceeding of the 2005 international conference on data mining (ICDM'05), Houston, TX, pp 66–73
10. Bonchi F, Lucchese C (2004) *On closed constrained frequent pattern mining*. In: Proceeding of the 2004 international conference on data mining (ICDM'04), Brighton, UK, pp 35–42
11. Brin S, Motwani R, Silverstein C (1997) *Beyond market basket: generalizing association rules to correlations*. In: Proceeding of the 1997 ACM-SIGMOD international conference on management of data (SIGMOD'97), Tucson, AZ, pp 265–276
12. Brin S, Motwani R, Ullman JD, Tsur S (1997) *Dynamic itemset counting and implication rules for market basket analysis*. In: Proceeding of the 1997 ACM-SIGMOD international conference on management of data (SIGMOD'97), Tucson, AZ, pp 255–264
13. Chang L, Wang T, Yang D, Luan H, Tang S (2009) *Efficient algorithms for incremental maintenance of closed sequential patterns in large databases*. In: Data & Knowledge Engineering 68 (2009) 68–106.
14. Chen X, Liu H, Chen P, Li L (2008) *A high performance algorithm for mining frequent patterns: LPS-Miner*. In: vol. 2, pp.7-11, 2008 International Symposium on Information Science and Engineering, 2008.
15. Cheng H, Yan X, Han J, Hsu C (2007) *Discriminative frequent pattern analysis for effective classification*. In: Proceeding of the 2007 international conference on data engineering (ICDE'07), Istanbul, Turkey
16. Cheung DW, Han J, Ng V, Fu A, Fu Y (1996) *A fast distributed algorithm for mining association rules*. In: Proceeding of the 1996 international conference on parallel and distributed information systems, Miami Beach, FL, pp 31–44
17. Cheung DW, Han J, Ng V, Wong CY (1996) *Maintenance of discovered association rules in large an incremental updating technique*. In: Proceeding of the 1996 international conference on data engineering (ICDE'96), New Orleans, LA, pp 106–114

18. Geerts F, Goethals B, Bussche J (2001) *A tight upper bound on the number of candidate patterns*. In: Proceeding of the 2001 international conference on data mining (ICDM'01), San Jose, CA, pp 155–162
19. Gionis A, Mannila H, Mielikäinen T, Tsaparas P (2006) *Assessing data mining results via swap randomization*. In: Proceeding of the 2006 ACM SIGKDD international conference on knowledge discovery in databases (KDD'06), Philadelphia, PA, pp 167–176
20. Goethals B, Zaki M (2003) *An introduction to workshop on frequent itemset mining implementations*. In: Proceeding of the ICDM'03 international workshop on frequent itemset mining implementations (FIMI'03), Melbourne, FL, pp 1–13
21. Han J, Pei J, Yin Y (2000) *Mining frequent patterns without candidate generation*. In: Proceeding of the 2000 ACM-SIGMOD international conference on management of data (SIGMOD'00), Dallas, TX, pp 1–12
22. Holder LB, Cook DJ, Djoko S (1994) *Substructure discovery in the subdue system*. In: Proceeding of the AAAI'94 workshop knowledge discovery in databases (KDD'94), Seattle, WA, pp 169–180
23. Holsheimer M, Kersten M, Mannila H, Toivonen H (1995) *A perspective on databases and data mining*. In Proceeding of the 1995 international conference on knowledge discovery and data mining (KDD'95), Montreal, Canada, pp 150–155
24. Jayanthi B., Duraiswamy K. (2012) – *A novel algorithm for cross level frequent pattern mining in multidatasets* – In: International Journal of Computer Applications (0975 – 8887) Volume 37– No.6, January 2012.
25. Mannila H, Toivonen H, Verkamo AI (1997) *Discovery of frequent episodes in event sequences*. Data Min Knowl Discov 1:259–289
26. Pan F, Cong G, Tung AKH, Yang J, Zaki M (2003) *CARPENTER: finding closed patterns in long biological datasets*. In: Proceeding of the 2003 ACM SIGKDD international conference on knowledge discovery and data mining (KDD'03), Washington, DC, pp 637–642
27. Park JS, Chen MS, Yu PS (1995) *An effective hash-based algorithm for mining association rules*. In: Proceeding of the 1995 ACM-SIGMOD international conference on management of data (SIGMOD'95), San Jose, CA, pp 175–186
28. Park JS, Chen MS, Yu PS (1995) *Efficient parallel mining for association rules*. In: Proceeding of the 4th international conference on information and knowledge management, Baltimore, MD, pp 31–36
29. Pasquier N, Bastide Y, Taouil R, Lakhal L (1999) *Discovering frequent closed itemsets for association rules*. In: Proceeding of the 7th international conference on database theory (ICDT'99), Jerusalem, Israel, pp 398–416
30. Patro SN., Mishra S., Khuntia P. and Bhagabati C. (2012) – *Construction of FP tree using Huffman coding* – In: IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 2, May 2012.
31. Pei J, Han J, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U, Hsu M-C (2001) *PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth*. In: Proceeding of the 2001 international conference on data engineering (ICDE'01), Heidelberg, Germany, pp 215–224
32. Piatetsky-Shapiro G (1991) *Notes of AAAI'91 workshop knowledge discovery in databases (KDD'91)*. AAAI/MIT Press, Anaheim, CA
33. Prasad KSN, Ramakrishna S (2011) *Frequent pattern mining and current state of the art*. In: International Journal of Computer Applications (0975 – 8887), Volume 26 - No.7, July 2011.
34. Raghunathan A, Murugesan K (2010) *Optimized frequent pattern mining for classified data sets*. In: ©2010 International Journal of Computer Applications (0975 - 8887) Volume 1 - No. 27.

35. Rawat SS, Rajamani L (2010) *Discovering potential user browsing behaviors using custom-built apriori algorithm*. In: International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010.
36. Romero AOC (2011) *Mining moving flock patterns in large spatio-temporal datasets using a frequent pattern mining approach*. In: Master of Science thesis, University of Twente, 2011.
37. Sadat MH., Samuel HW., Patel S., Zaïane OR. (2011) – *Fastest association rule mining algorithm predictor (FARM-AP)* – In: ProceedingC3S2E '11 Proceedings of The Fourth International Conference on Computer Science and Software Engineering, 2011.
38. Sarawagi S, Thomas S, Agrawal R (1998) *Integrating association rule mining with relational database systems: alternatives and implications*. In: Proceeding of the 1998 ACM-SIGMOD international conference on management of data (SIGMOD'98), Seattle, WA, pp 343–354
39. Savasere A, Omiecinski E, Navathe S (1995) *An efficient algorithm for mining association rules in large databases*. In: Proceeding of the 1995 international conference on very large data bases (VLDB'95), Zurich, Switzerland, pp 432–443
40. Sharma H., Garg D. (2011) – *Comparative analysis of various approaches used in frequent pattern mining* – In: International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence IJACSA pp. 141-147 August 2011.
41. Srikant R, Agrawal R (1996) *Mining sequential patterns: generalizations and performance improvements*. In: Proceeding of the 5th international conference on extending database technology (EDBT'96), Avignon, France, pp 3–17
42. Sumathi K., Kannan S., Nagarajan K. (2012) – *A new MFI mining algorithm with effective pruning mechanisms* – In: International Journal of Computer Applications (0975 – 8887) Volume 41– No.6, March 2012.
43. Toivonen H (1996) *Sampling large databases for association rules*. In: Proceeding of the 1996 international conference on very large data bases (VLDB'96), Bombay, India, pp 134–145
44. Utmal M., Chourasia S., Vishwakarma R. (2012) – *A novel approach for finding frequent item sets done by comparison based technique* – In: International Journal of Computer Applications (0975 – 8887) Volume 44– No.9, April 2012.
45. Yan X, Han J, Afshar R (2003) *CloSpan: mining closed sequential patterns in large datasets*. In: Proceeding of the 2003 SIAM international conference on data mining (SDM'03), San Francisco, CA, pp 166–177
46. Zaki MJ (2000) *Scalable algorithms for association mining*. IEEE Trans Knowl Data Eng 12:372–390
47. Zaki MJ (2001) *SPADE: an efficient algorithm for mining frequent sequences*. Mach Learn 40:31–60
48. Zaki MJ, Parthasarathy S, Ogihara M, Li W (1997) *Parallel algorithm for discovery of association rules*. Data Mining Knowl Discov, 1:343–374
49. Zheng Z, Zhao Y, Zuo Z, Cao L (2010) *An efficient GA-based algorithm for mining negative sequential patterns*. In: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, 2010, Volume 6118/2010, 262-273.
50. Thanh-Trung Nguyen – *An Improved Algorithm for Frequent Patterns Mining Problem* – 3CA2010: 2010 International Symposium on Computer, Communication, Control and Automation (May 5-7, 2010, Tainan, Taiwan)
51. Thanh-Trung Nguyen, Phi-Khu Nguyen – *The Objectivity Measurement of Frequent Patterns* – WCECS2010: The World Congress on Engineering and Computer Science 2010 (20-22 Oct 2010, San Francisco, USA)
52. Nguyễn Phi Khứ, Nguyễn Thành Trung – *Điều khiển vận hành lò hơi bằng giải thuật mạng nơron nhân tạo* – Hội nghị Cơ học Thủy Khí 2011 (21-23 tháng 07 năm 2011, TP. Vinh, Nghệ An, Việt Nam)

53. Thanh-Trung Nguyen, Viet-Long Huu Nguyen, Phi-Khu Nguyen – *A Bit-chain Based Algorithm for Problem of Attribute Reduction* – ACIIDS2012: The 4th Asian Conference on Intelligent Information and Database Systems (19-21 March, 2012, Kaohsiung, Taiwan)
54. Thanh-Trung Nguyen, Viet-Long Huu Nguyen, Phi-Khu Nguyen – *Identifying Customer Characteristics by Using Rough Set Theory with a New Algorithm and Posterior Probabilities* – ICCIS2012: The 4th International Conference on Computational and Information Sciences (17-19 August, 2012, Chongqing, China)
55. Thanh-Trung Nguyen, Phi-Khu Nguyen – *A New Approach for Problem of Sequential Pattern Mining* – ICCCI2012: The 4th International Conference on Computational Collective Intelligence Technologies and Applications (28-30 November 2012, Ho Chi Minh city, Vietnam)
56. Thanh-Trung Nguyen, Viet-Long Huu Nguyen, Phi-Khu Nguyen – *Accumulated Frequent Pattern* – ICTMF2012: The Third International Conference on Theoretical and Mathematical Foundations of Computer Science (December 1-2, 2012, Bali, Indonesia)