

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

ĐẶNG VĂN THUẬN

**PHÂN TÍCH HOẠT ĐỘNG ĐẦU TƯ
 TRONG THỊ TRƯỜNG CHỨNG KHOÁN
 ỨNG DỤNG THUẬT TOÁN
 APRIORI FP-GROWTH
 TRONG KHAI PHÁ DỮ LIỆU**

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01

ĐÀ NẴNG, NĂM 2010

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học
 PGS.TS. PHAN HUY KHÁNH

Phản biện 1: PGS.TS ĐOÀN VĂN BAN

Phản biện 2: TS NGUYỄN TẤN KHÔI

Luận văn sẽ được bảo vệ tại Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 14 tháng 10 năm 2010

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin-học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng

MỞ ĐẦU

1. LÝ DO CHỌN ĐỀ TÀI

Hiện nay việc đầu tư vào TTCK hiện nay ở Việt Nam có rất nhiều biến động khó khăn: CSDL lớn, các phần mềm trợ giúp hiện tại chưa phù hợp với TTCK tại Việt Nam.... Do đó việc ứng dụng kỹ thuật KPDL để phát hiện tìm ra quy luật về sự biến động hữu ích ẩn chứa trong khối lượng dữ liệu khổng lồ đó sẽ mang lại cho các nhà đầu tư nhiều cơ hội để chọn lựa loại cổ phiếu cần đầu tư và đúng thời điểm, đúng khối lượng giao dịch nhằm đạt được giá trị gia tăng hiệu quả trong đầu tư chứng khoán. Xuất phát từ lý do đó tôi thực hiện đề tài: *"Phân tích hoạt động đầu tư trong thị trường chứng khoán ứng dụng thuật toán Apriori FP-Growth trong khai phá dữ liệu"*.

2. MỤC ĐÍCH NGHIÊN CỨU

Mục đích của đề tài là phân tích hoạt động đầu tư trong thị trường chứng khoán đề xuất giải pháp ứng dụng KPDL để xây dựng hệ thống trợ giúp nhà đầu tư đưa ra những quyết định đầu tư cổ phiếu hợp lý trong TTCK mang lại hiệu quả kinh tế trong đầu tư.

3. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

Đối tượng nghiên cứu thông tin cổ phiếu tham gia giao dịch ở các sàn giao dịch và tâm lý nhà đầu chứng khoán tại TTCK Việt Nam; nghiên cứu lý thuyết về Datamining để tìm ra giải pháp thực tế xây dựng ứng dụng.

Phạm vi nghiên cứu là nghiên cứu sự biến động của TTCK Việt Nam; trên cơ sở giao dịch của các cổ phiếu ứng dụng luật kết hợp

KPDL tìm ra thông tin hữu ích trợ giúp nhà đầu tư thực hiện chiến lược đầu tư chứng khoán mang lại hiệu quả.

4. PHƯƠNG PHÁP NGHIÊN CỨU

Phân tích tình hình đặc điểm hoạt động đầu tư trong TTCK Việt Nam, tìm hiểu lý thuyết đầu tư và tâm lý nhà đầu tư, thu thập, thống kê dữ liệu giao dịch tại các sàn; trên cơ sở nghiên cứu lý thuyết về khai phá dữ liệu và ứng dụng thuật toán Apriori FP-growth. Từ đó đề xuất giải pháp xây dựng hệ thống ứng dụng phân tích trợ giúp nhà đầu tư chứng khoán.

5. Ý NGHĨA KHOA HỌC VÀ THỰC TIỄN CỦA ĐỀ TÀI

Về tính khoa học, hiện nay các sàn giao dịch chứng khoán chưa xây dựng và triển khai kỹ thuật KPDL trong đầu tư chứng khoán để xây dựng hệ thống phân tích trợ giúp nhà đầu tư; Vì lẽ đó, tính thực tiễn của đề tài là xây dựng những chức năng phân tích hiệu quả trợ giúp cho nhà đầu tư; góp phần khai thác tốt việc đầu tư vào cổ phiếu trên TTCK nhằm mục đích cuối cùng là tăng lợi nhuận, giảm thiểu rủi ro cho nhà đầu tư.

6. BỐ CỤC CỦA LUẬN VĂN

Bố cục của luận văn bao gồm: phần mở đầu nêu lý do chọn đề tài, mục đích, đối tượng và phạm vi, phương pháp nghiên cứu, ý nghĩa khoa học và thực tiễn của đề tài. Gồm có ba Chương và phần kết luận nêu kết quả đạt được và định hướng phát triển của đề tài.

CHƯƠNG 1 TÌM HIỂU VÀ PHÂN TÍCH THỊ TRƯỜNG CHỨNG KHOÁN

1.1 GIỚI THIỆU VỀ THỊ TRƯỜNG CHỨNG KHOÁN

1.1.1 Giới thiệu tổng quan

TTCK đóng vai trò quan trọng trong nền kinh tế hiện đại, TTCK chỉ là nơi diễn ra các hoạt động trao đổi, mua bán, chuyển nhượng các loại cổ phiếu chứng khoán; qua đó thay đổi chủ thể nắm giữ chứng khoán. TTCK là một hướng đầu tư mới hấp dẫn mang lại lợi nhuận cao nhưng cũng có rất nhiều rủi ro.

1.1.2 Đặc điểm thị trường chứng khoán Việt Nam

TKCK Việt Nam ra đời mới hơn 10 năm nhưng đã có những ảnh hưởng to lớn đến nền kinh tế quốc gia.

1.1.2.1 Bối cảnh ra đời và quá trình hình thành TTCK Việt Nam

Sau nhiều năm chuẩn bị và chờ đợi, ngày 11-7-1998 Chính phủ đã ký Nghị định số 48/CP ban hành về chứng khoán và TTCK chính thức khai sinh cho TTCK Việt Nam ra đời.

Chỉ số VNIndex: VNIndex là ký hiệu của chỉ số chứng khoán Việt Nam. VNIndex xây dựng căn cứ vào giá trị thị trường của tất cả các cổ phiếu được niêm yết.

Chỉ số VNIndex = (Giá trị thị trường hiện hành / Giá trị thị trường cơ sở) x 100

Trung tâm giao dịch chứng khoán TP.Hồ Chí Minh ngày 20/07/2000 và thực hiện phiên giao dịch đầu tiên vào ngày 28/07/2000. Ở thời điểm lúc bấy giờ, chỉ có hai doanh nghiệp niêm

yết hai loại cổ phiếu (REE và SAM) với số vốn hai trăm bảy mươi tỷ đồng và một số ít trái phiếu Chính phủ được niêm yết giao dịch.

Ngày 8/3/2005 Trung tâm giao dịch chứng khoán Hà Nội (TTGDCK HN) chính thức đi vào hoạt động. Giai đoạn tình ngủ dân xuất hiện từ năm 2005 khi tỷ lệ nắm giữ của nhà đầu tư nước ngoài được nâng từ 30% lên 49% (trừ lĩnh vực ngân hàng).

1.1.2.2 Diễn biến thực tế TTCK Việt Nam

Hiện nay tình hình TTCK Việt Nam trên các sàn cũng chưa ổn định, tâm lý chung vẫn là dựa vào số đông quan hệ ảo cung cầu để đầu tư.

1.1.3 Khó khăn

Rủi ro do tính thanh khoản thấp

Rủi ro từ thông tin

Rủi ro từ các quy định và chất lượng dịch vụ của sàn giao dịch

Rủi ro từ các chấn động thị trường

1.1.4 Thuận lợi

TTCK nước ta còn non trẻ Nhà nước có nhiều chính sách ưu đãi thu hút các Nhà đầu tư như: các ưu đãi về thuế; vốn vay để đầu tư và cơ chế giao dịch thông thoáng. Nguồn vốn nhân rộng trong nhân dân rất nhiều.

1.2 PHƯƠNG PHÁP PHÂN TÍCH THỊ TRƯỜNG CHỨNG KHOÁN

Phương pháp nghiên cứu, biểu đồ mô tả dữ liệu thống kê của TTCK.

Phương pháp dựa trên xem xét giá trị thực chất của một cổ

Phương pháp dự báo chuỗi thời gian quá khứ.

Phương pháp máy học

1.3 TƯ VẤN TRONG THỊ TRƯỜNG CHỨNG KHOÁN

1.3.1 Lí thuyết đầu tư chứng khoán

Hai lí thuyết chính là Firm Foundation[17] và Castle in the Air[12].

1.3.2 Các nguồn dữ liệu trong chứng khoán

Từ các Sàn giao dịch, các trang web quảng cáo. Hiện nay có hai trang web đăng tải đầy đủ và chính xác các thông tin cổ phiếu là: <http://chungkhoan.com.vn> và <http://cophieu68.com>.

Trang web của công ty niêm yết giá cổ phiếu của họ.

Giá trị cổ phiếu niêm yết tại các sàn bao gồm: giá mở cửa vào lúc 8:30 sáng hàng ngày (bằng với giá đóng cửa của ngày hôm trước), giá đóng cửa vào lúc 11:00 cuối ngày, giá cao nhất và giá thấp nhất do quá trình giao dịch ngày.

1.4 MÔ HÌNH HỆ THỐNG PHÂN TÍCH TRỢ GIÚP THỊ TRƯỜNG CHỨNG KHOÁN

Mô tả hệ thống phân tích TTCK

1.4.1 Thu thập dữ liệu

Dữ liệu được thu thập từ hai nguồn: nguồn một là từ web thông qua dịch vụ của web server, nguồn hai là từ các server CSDL của các công ty.

1.4.2 Phân tích ý nghĩa chỉ số

Phân tích đưa ra các dự đoán để có thể có được những kết quả về xu hướng tăng giá của các loại cổ phiếu khác nhau trong tương lai. Đây chính là nhiệm vụ chính của luận văn.

1.4.3 Thông tin tư vấn cho nhà đầu tư

Từ các thông tin dữ liệu quá khứ được trích lọc tại các Sàn giao dịch cổ phiếu chứng khoán xây dựng ứng dụng phân tích xác định được diễn biến cổ phiếu trong các ngày tiếp theo có xu hướng biến động tăng giảm.

1.5 TƯ VẤN TRONG THỊ TRƯỜNG CHỨNG KHOÁN

1.5.1 Khả năng tư vấn trong thị trường chứng khoán

Để phân tích TTCK dựa dữ liệu quá khứ làm nền tảng phân tích.

1.5.2 Xác định nhiệm vụ tư vấn

Hệ thống tư vấn sẽ đưa ra các dự đoán những cổ phiếu nào có khả năng tăng trong lần giao dịch kế tiếp dựa trên luật kết hợp và thuật toán.

1.5.3 Phương pháp phân tích và tư vấn

Gồm các Phương pháp phân tích kỹ thuật; phân tích cơ sở; dự báo chuỗi thời gian quá khứ và phương pháp máy học. Mỗi phương pháp

có một lợi thế nhất định, chính vì thế tùy theo từng yêu cầu cụ thể ta chọn phương pháp phù hợp với thực tế.

1.6 TÓM TẮT

Các nội dung trong chương này tập trung giới thiệu về TTCK tại Việt Nam, các đặc điểm về giao dịch cũng như những thông tin cơ bản về TTCK. Từ những phân tích ban đầu về TTCK, ta đưa ra được nhiệm vụ chính của luận văn, nhiệm vụ của phân tích và dự đoán về xu hướng tăng giảm của cổ phiếu bằng các kỹ thuật KPDL trong chương hai.

CHƯƠNG 2 TỔNG QUAN VỀ KHAI PHÁ TRI THỨC VÀ KHAI PHÁ DỮ LIỆU

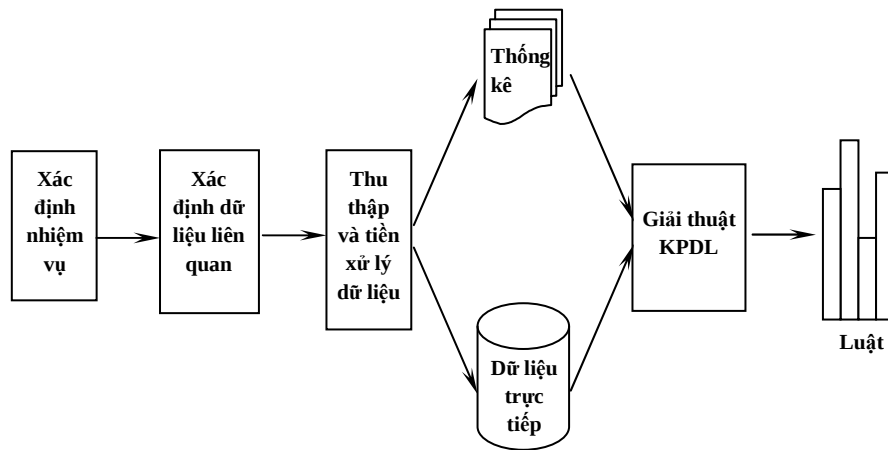
2.1 KHAI PHÁ DỮ LIỆU

2.1.1 Các khái niệm cơ bản

Định nghĩa 1 của William J Frawley, Gregory Piatetsky-Shapiro, và Christopher J Matheus, năm 1991: “Knowledge discovery in databases, also known KPDL, is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.” (Khám phá kiến thức trong các CSDL, hay là KPDL, là 1 tiến trình tìm kiếm những giá trị mới lạ, có khả năng hữu ích trong dữ liệu).

Định nghĩa 2 của Marcel Holshemier và Arno Siebes, năm 1994: “KPDL is the search for relationships and global patterns that exist in large databases but are ‘hidden’ among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database.” (KPDL là quá trình tìm kiếm đối với những thành phần có mối liên hệ và tổng quát tồn tại trong 1 CSDL lớn nhưng đã bị giấu đi trong 1 lượng lớn dữ liệu, như là mối quan hệ giữ dữ liệu của các bệnh nhân và thuốc điều trị của họ. Mối quan hệ này biểu diễn những tri thức có giá trị về CSDL và các đối tượng trong CSDL, và nếu CSDL này phản ánh trung thực thế giới thực được ghi chép vào CSDL).

2.1.2 Các bước khai phá tri thức



Hình 2.1: Sơ đồ mô tả quá trình KPD

Bước 1: Gom dữ liệu (Gathering) và trích lọc dữ liệu (Selection)

Bước 2: Tiền xử lý dữ liệu (Cleansing, Pre-processing and Preparation)

Bước 3 : Chuyển đổi dữ liệu (Transformation)

Bước 4: Phát hiện và trích mẫu dữ liệu (Pattern Extraction and Discovery)

Bước 5: Đánh giá kết quả mẫu (Evaluation of Result)

2.1.3 Các phương pháp khai phá dữ liệu

Quá trình KPD là quá trình phát hiện mẫu trong đó giải thuật KPD tìm kiếm các mẫu đáng quan tâm theo dạng xác định như các luật, cây phân lớp, hồi quy, phân nhóm,...

Cây quyết định và luật

Khai phá dữ liệu văn bản

Mạng neuron

Phân nhóm và phân đoạn

Khai phá luật kết hợp

Các phương pháp phân lớp và hồi quy phi tuyến

Các phương pháp dựa trên mẫu

2.1.4 Hướng tiếp cận trong khai phá dữ liệu

Các hướng tiếp cận của KPD có thể được phân chia theo chức năng hay lớp các bài toán khác nhau.

Hướng tiếp cận phổ biến là phân lớp và dự đoán (classification & prediction); cây quyết định; mạng nơ ron; một trong những hướng tiếp cận dễ hình dung là khai phá chuỗi theo thời gian (sequential/temporal patterns): tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và TTCK vì nó có tính dự báo cao.

2.1.5 Các dạng dữ liệu dùng để khai phá

CSDL quan hệ (relational databases), CSDL đa chiều (multidimensional structures, data warehouses), CSDL dạng giao

dịch (transactional databases), CSDL quan hệ - hướng đối tượng (object-relational databases), Dữ liệu không gian và thời gian (spatial and temporal data), Dữ liệu chuỗi thời gian (time-series data), CSDL đa phương tiện (multimedia databases) như âm thanh (audio), hình ảnh (image), phim ảnh (video) Dữ liệu Text và Web (text database & www)...

2.1.6 Khai phá luật kết hợp và ứng dụng

Luật kết hợp là một biểu thức có dạng: $X \Rightarrow Y$, trong đó X và Y là tập các trường gọi là item. Ý nghĩa của các luật kết hợp khá dễ nhận thấy: Cho trước một CSDL có D là tập các giao tác - trong đó mỗi giao tác $T \in D$ là tập các item - khi đó $X \Rightarrow Y$ diễn đạt ý nghĩa rằng bất cứ khi nào giao tác T có chứa X thì chắc chắn T có chứa Y. Độ tin cậy của luật (rule confidence) có thể được hiểu như xác suất điều kiện $P(Y \subseteq T | X \subseteq T)$. Được ứng dụng KPDL trong các Lĩnh vực: Y học, Ngân hàng, TTCK...

2.2 KHAI PHÁ LUẬT KẾT HỢP

Vấn đề khám phá luật kết hợp[2][5][6][7] được phát biểu như sau: Cho trước tỉ lệ hỗ trợ θ và độ tin cậy β . Đánh số tất cả các luật trong D có các giá trị tỉ lệ hỗ trợ và tin cậy lớn hơn θ và β tương ứng.

Giả thiết D là CSDL giao dịch và với $\theta = 30\%$, $\beta = 60\%$. Vấn đề phát hiện luật kết hợp được thực hiện như sau: Liệt kê, đếm tất cả những qui luật chỉ ra sự xuất hiện một số các mục sẽ kéo theo một số mục khác, chỉ xét những qui luật mà tỉ lệ hỗ trợ lớn hơn 30% và độ tin cậy lớn hơn 60%.

2.2.1 Cơ sở dữ liệu giao dịch

Khái niệm CSDL giao dịch

Lấy ví dụ giỏ tăng trưởng giá cổ phiếu như sau:

Giỏ 1: {MCV, DRC, TRC}

Giỏ 2: {AMM, SHB, VCB}

...

Giỏ n: {HAG, DIG, DRC}

Ta có định nghĩa: Hạng mục(Item) : mặt hàng trong giỏ hay một thuộc tính; Giao dịch(Transaction) : tập các hạng mục được mua trong một giỏ hàng (có TID – mã giao dịch); CSDL giao dịch: tập các giao dịch;

Như vậy thông tin từng loại cổ phiếu được giao dịch tại TTCK cũng là giao dịch.

2.2.2 Giải thuật chuyển đổi cơ sở dữ liệu

Để đơn giản hơn cho các giải thuật khai phá luật kết hợp chúng ta có thể xây dựng giải thuật cho phép chuyển đổi từ một CSDL dạng quan hệ truyền thống sang CSDL giao dịch để trợ giúp cho quá trình KPDL tình hình sử dụng các loại hình dịch vụ bằng luật kết hợp[phụ lục 1].

2.2.3 Một số hướng tiếp cận trong khai phá luật kết hợp

Một số hướng: Luật kết hợp nhị phân; Luật kết hợp có thuộc tính số và thuộc tính hạng mục; Luật kết hợp tiếp cận theo hướng tập thô; Tìm kiếm luật kết hợp dựa trên lý thuyết tập thô; Luật kết hợp với

thuộc tính được đánh trọng số; Luật kết hợp song song; Luật kết hợp mờ

2.2.4 Luật kết hợp

2.2.4.1 Một số khái niệm cơ bản

Cho một tập $I = \{I_1, I_2, \dots, I_m\}$ các tập m mục, một giao dịch T được định nghĩa như một tập con của các khoản mục trong I ($T \subseteq I$).

Gọi D là CSDL của n giao dịch và mỗi giao dịch được đánh nhãn với một định danh duy nhất. Nói rằng, một giao dịch $T \in D$ hỗ trợ một tập $X \subseteq I$ nếu nó chứa tất cả các item của X .

Điều này nghĩa là $X \subseteq T$, trong một số trường hợp người ta dùng ký hiệu $T(X)$ để chỉ tập các giao dịch hỗ trợ cho X . Kí hiệu $\text{support}(X)$ (hoặc $\text{sup}(X)$, $s(X)$) là tỷ lệ phần trăm của các giao dịch hỗ trợ X trên tổng các giao dịch trong D , nghĩa là:

$$\text{sup}(X) = \frac{|\{T \in D \mid X \subseteq T\}|}{D} \quad (2.1)$$

Độ hỗ trợ tối thiểu minsup là một giá trị cho trước bởi người sử dụng. Nếu tập mục X có $\text{sup}(X) \geq \text{minsup}$ thì ta nói X là một tập các mục phổ biến. Một luật kết hợp có dạng $R: X \Rightarrow Y$, trong đó X, Y là tập các mục, $X, Y \subseteq I$ và $X \cap Y = \emptyset$. X được gọi là tiên đề và Y được gọi là hệ quả của luật.

Luật $X \Rightarrow Y$ tồn tại một độ tin cậy c . Độ tin cậy c được định nghĩa là khả năng giao dịch T hỗ trợ X thì cũng hỗ trợ Y .

Ta có công thức tính độ tin cậy c như sau:

$$\text{conf}(X \Rightarrow Y) = p(Y \subseteq I \mid X \subseteq I) = \frac{p(Y \subseteq T \wedge X \subseteq T)}{p(X \subseteq T)} = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)} \quad (2.2)$$

2.2.4.2 Một số tính chất của luật kết hợp

Các tính chất về độ hỗ trợ của tập mục

Tính chất 1: Cho A và B là hai tập mục, nếu $A \subseteq B$ thì $\text{sup}(A) \geq \text{sup}(B)$.

Tính chất 2: Cho A và B là hai tập mục và $A \subseteq B$, nếu A là tập mục không phổ biến thì B cũng không là tập mục phổ biến.

Tính chất 3: Cho B là tập mục phổ biến, nếu $A \subseteq B$ thì A cũng là tập mục phổ biến.

Các tính chất về cơ bản của luật kết hợp

Tính chất 1: Nếu có $A \rightarrow C$ và $B \rightarrow C$ trong D thì $A \cup B \rightarrow C$ có thể là chưa chắc đúng.

Tính chất 2: Nếu $A \cup B \rightarrow C$ thì $A \rightarrow C$ và $B \rightarrow C$ chưa chắc đúng.

Tính chất 3: Nếu $A \rightarrow B$ và $B \rightarrow C$, chúng ta không thể suy ra $A \rightarrow C$.

Tính chất 4: Nếu $A \rightarrow (L - A)$ không thỏa mãn độ tin cậy cực tiểu thì luật $B \rightarrow (L - B)$ cũng không thỏa mãn, với các tập mục L, A, B và $B \subseteq A \subseteq L$.

2.2.4.3 Một số bài toán cơ bản về luật Kết hợp

Bài toán 1: Tìm tất cả các tập mục mà có độ hỗ trợ lớn hơn độ hỗ trợ tối thiểu do người dùng xác định. Các tập mục thỏa mãn độ hỗ trợ tối thiểu được gọi là các tập mục phổ biến.

Bài toán 2: Dùng các tập mục phổ biến để sinh ra các luật mong muốn. Ý tưởng là nếu gọi ABCD và AB là các tập mục phổ biến, thì chúng ta có thể xác định luật nếu $AB \Rightarrow CD$ giữ lại với tỷ lệ độ tin cậy:

$$conf = \frac{\sup(ABCD)}{\sup(AB)} \quad (2.4)$$

Nếu $conf \geq minconf$ thì luật được giữ lại (luật này sẽ thỏa mãn độ hỗ trợ tối thiểu vì ABCD là phổ biến).

2.2.4.4 Các khai thác luật kết hợp

B1: Tìm tất cả các tập mục phổ biến (theo ngưỡng minsup)

B2: Tạo ra các luật từ các tập mục phổ biến đối với mỗi tập mục phổ biến S, tạo ra tất cả các tập con khác rỗng của S.

Đối với mỗi tập con khác rỗng A của S thì luật $A \Rightarrow (S - A)$ là luật kết hợp cần tìm nếu:

$$conf(A \Rightarrow (S - A)) = \frac{\sup(S)}{\sup(A)} \geq minconf$$

Hầu hết các thuật toán khai phá luật kết hợp thường chia thành hai pha:

Pha 1: Tìm tất cả các tập mục phổ biến từ CSDL tức là tìm tất cả các tập mục X thỏa mãn $s(X) \geq minsup$.

Pha 2: Sinh các luật tin cậy từ các tập mục phổ biến đã tìm thấy ở pha 1.

Nếu X là một tập luật phổ biến thì tập luật kết hợp được sinh từ X có dạng :

$$Y \xrightarrow{c} X \setminus Y, \text{ trong đó:}$$

Y là tập con khác rỗng của X.

$X \setminus Y$ là hiệu của hai tập hợp X và Y. c là độ tin cậy của luật thỏa mãn $c \geq minconf$.

2.2.5 Phát hiện luật kết hợp trên hệ thông tin nhị phân

2.2.5.1 Hệ thông tin nhị phân

Cho các tập $O = \{o_1, o_2, \dots, o_n\}$ là một tập hữu hạn gồm n đối tượng, $D = \{d_1, d_2, \dots, d_m\}$ là một tập hữu hạn gồm m chỉ báo, $B = \{0, 1\}$

Hệ thông tin nhị phân được định nghĩa là $SB = (O, D, B, \chi)$ trong đó χ là ánh xạ $\chi: O \times D \rightarrow B$, $\chi(o, d) = 1$ nếu đối tượng o có chỉ báo d và $\chi(o, d) = 0$ nếu ngược lại.

2.2.5.2 Tập chỉ báo phổ biến nhị phân

Cho hệ thông tin nhị phân $SB = (O, D, B, \chi)$ và một ngưỡng $\theta \in (0, 1)$. Cho $S \subseteq D$, S là tập chỉ báo phổ biến nhị phân với ngưỡng θ nếu $\text{card}(\rho_B(S)) \geq \theta * \text{card}(O)$

Cho LB là một tập gồm tất cả các tập chỉ báo phổ biến nhị phân đã phát hiện từ SB, chúng có thuộc tính như sau: $\forall S \in LB, T \subset S$ thì $T \in LB$.

Trong đó LB_h là tập con của LB nếu $X \in LB_h$ thì $\text{card}(X) = h$ (với h là số nguyên dương).

2.2.5.3 Các luật kết hợp phổ biến nhị phân và hệ số tin cậy

Cho hệ thông tin nhị phân $SB = (O, D, B, \chi)$ và một ngưỡng $\theta \in (0, 1)$. Cho L là một phần tử của LB , X và Y là hai tập con của L , trong đó:

$$L = X \cup Y, X \neq \{\}, Y \neq \{\} \text{ và } X \cap Y = \{\}$$

Chúng ta xác định các luật kết hợp nhị phân giữa tập chỉ số X và tập chỉ số Y là một ánh xạ thông tin: $X \rightarrow Y$. Hệ số tin cậy của luật này được biểu diễn là:

$$CF_B(X \rightarrow Y) = \frac{\text{card}(\rho_B(X) \cup \rho_B(Y))}{\text{card}(\rho_B(X))} \quad (2.5)$$

Gọi RB, β là tập tất cả các luật kết hợp phổ biến nhị phân được phát hiện từ SB . Trong đó $CFB(r) \geq \beta, \forall r \in RB, \beta$

2.2.5.4 Các vector chỉ báo nhị phân và các phép toán

Cho hệ thông tin nhị phân $SB = (O, D, B, \chi)$ trong đó $O = \{o_1, o_2, \dots, o_n\}$ là một tập hữu hạn gồm n đối tượng, $D = \{d_1, d_2, \dots, d_m\}$ là một tập hữu hạn gồm m chỉ báo.

Vector chỉ báo nhị phân: $vB(X) = \{X_1, X_2, \dots, X_n\}$ trong đó: $X \subset D$ là một vector với n thành phần, mỗi thành phần X_j chiếm một giá trị trong B . Cho VSB là tập tất cả các vector chỉ báo nhị phân của SB , nếu $\text{card}(X) = 1$ thì X là bộ chỉ báo của SB và $X_j = \chi(o, X)$

2.2.5.5 Độ hỗ trợ các vector chỉ báo nhị phân

Cho $X_1 \subset D$, độ hỗ trợ của $vB(X_1)$ biểu diễn $\text{supB}(vB(X_1))$ được định nghĩa:

$$\text{supB}(vB(X_1)) = \{o \in O \mid \forall d \in X_1, \chi(o, d) = 1\} \quad (2.6)$$

2.2.6 Thuật toán phát hiện tập chỉ báo và luật kết hợp

Thuật toán Apriori-Tid có hai pha [phụ lục 2].

2.3 THUẬT TOÁN APRIORI FP-GROWTH

2.3.1 Thuật toán Apriori cơ bản

Apriori là một thuật giải được do Rakesh Agrawal, Tomasz Imielinski, Arun Swami đề xuất lần đầu vào năm 1993. Thuật toán tìm giao dịch t có độ hỗ trợ và độ tin cậy thỏa mãn lớn hơn một giá trị ngưỡng. được trình bày ở [phụ lục 2] và [phụ lục 3].

2.3.2 Sinh các luật kết hợp từ tập mục phổ biến

Với mỗi tập mục phổ biến l , sinh ra tất cả các tập con không rỗng của l

Với mỗi tập con không rỗng a của l , ta có luật $a \rightarrow (l-a)$

Nếu $\frac{\text{sup}(l)}{\text{sup}(a)} \geq \text{minconf}$ ở đó minconf là ngưỡng độ tin cậy cực tiểu

Vì các luật được sinh ra từ các tập mục phổ biến nên độ hỗ trợ của luật đã được thỏa mãn, tức là độ hỗ trợ của luật chính là $\text{sup}(l)$.

2.3.3 Thuật toán FP-Growth

Thuật toán xây dựng cây FP_Tree :

Input: cơ sở dữ liệu giao dịch D và ngưỡng độ hỗ trợ minsup

Output: cây mẫu Phổ biến FP_Tree .

Method:

Bước 1: Duyệt qua cơ sở dữ liệu D để đếm số lần xuất hiện của các mục trong giao tác và xác định mục Phổ biến và độ hỗ trợ của

chúng, sắp xếp các mục Phổ biến giảm dần theo độ hỗ trợ, ta được danh sách các mục sắp xếp L.

Bước 2: Xây dựng FP_Tree. Đầu tiên tạo nút gốc, sau đó với mỗi giao tác t chọn và sắp xếp các mục Phổ biến theo thứ tự trong danh sách L, thực hiện thêm vào cây FP_Tree bằng cách gọi hàm insert_tree(p|T), thay đổi trường count cho phù hợp.

Thủ tục thêm các mục Phổ biến vào cây FP_Tree:

Procedure Insert_Tree(string[p|P], Tree T)

// Trong đó p là mục đầu tiên của dãy và P là phần còn lại của dãy

{If cây T có nút con N mà N.Item_name = p Then N.count++

Else

Tạo nút mới N;

N.Item_name:=p;N.count:=1

Thay đổi nút liên kết cho p;

If $p \neq \emptyset$ then

Insert_Tree(p,N):

}

Ta tiến hành xây dựng cây mẫu Phổ biến :

Khai thác tập mục Phổ biến

Thuật toán Xây dựng các mẫu Phổ biến từ cây FP-Growth

Input: Cây FP-Growth của CSDL D, ngưỡng minsup

Output: Tập phổ biến F

Method: Call FP-Growth(Tree, null)

Procedue FP Growth(tree, α)

{

F = $\emptyset$

If Tree chỉ chứa một đường dẫn đơn P then {

For mỗi tổ hợp β của các nút trong đường dẫn P Do {

Sinh mẫu $p = \alpha \cup \beta$ với support_count(p) = độ hỗ trợ nhỏ nhất của các nút trong β

F=F $\cup$ p

}

Else

For mỗi a_i trong header của tree Do

{ Sinh mẫu $\beta = a_i \cup \alpha$;

Support_count(β) = a_i .support_count

F= F $\cup$ β

Xây dựng cơ sở mẫu có điều kiện của β

Xây dựng FP_tree có điều kiện Tree $_{\beta}$ của β

If Tree $_{\beta}$!= $\emptyset$ then

Call FP_Growth(Tree $_{\beta}$, β)

}

CHƯƠNG 3 XÂY DỰNG HỆ THỐNG TRỢ GIÚP PHÂN TÍCH VÀ DỰ ĐOÁN DÙNG LUẬT KẾT HỢP

3.1 KHẢO SÁT TÌNH HÌNH THỰC TẾ TẠI CÁC SÀN GIAO DỊCH CHỨNG KHOÁN

Sàn giao dịch chứng khoán TP. HCM luôn diễn ra sôi động, là nơi niêm yết tất cả các cổ phiếu giao dịch trên thị trường; được xem là trung tâm giao dịch của chứng khoán Việt Nam. Vì vậy tôi chọn sàn giao dịch chứng khoán TP. HCM để tìm hiểu, thu thập tài liệu để xây dựng chương trình ứng dụng.

3.2 PHÁT BIỂU BÀI TOÁN

TTCK là nơi diễn ra các giao dịch mua/bán/chuyển nhượng trên các cổ phiếu được niêm yết. Số lượng các cổ phiếu tham gia niêm yết trong ngày, các cổ phiếu tham gia giao dịch; các cổ phiếu giao dịch trong cùng một ngày được định danh cùng một định danh TID để phân tích và được lưu trữ trong CSDL giao dịch. từ giá đóng cửa các cổ phiếu giao dịch trong các ngày xác định giá trị tăng giảm các cổ phiếu là dữ liệu đầu vào của chương trình.

Các dữ liệu được thu thập về sẽ được phân tích, định dạng và chứa trong kho, là loại dữ liệu được sử dụng để khai phá. Sau giai đoạn khai phá, ta sử dụng thuật toán Apriori FP-Growth để phân tích tìm ra các cổ phiếu thường xuyên tăng giá, nhóm cổ phiếu tăng giá và qua hệ giữa chúng; từ đó nhà đầu tư sử dụng cho việc dự đoán để ra quyết định đầu tư cổ phiếu.

3.3 PHÂN TÍCH BÀI TOÁN

TTCK số giao dịch hàng ngày rất nhiều làm cho dung lượng CSDL tăng theo; chính vì thế dữ liệu phân tích rất lớn. Nếu dùng các phương pháp thống kê cổ điển thì sẽ không thể kết xuất ra được tri thức. Do vậy cần dùng các kỹ thuật KPD – sử dụng luật kết hợp.

3.4 CÀI ĐẶT CHƯƠNG TRÌNH

3.4.1 Ngôn ngữ và môi trường lập trình

Để hoàn thành hệ thống phân tích TTCK tư vấn cho nhà đầu tư mới, tôi sử dụng ngôn ngữ lập trình C# trong bộ Visual Studio 2008 (cũng chính là môi trường lập trình cho phần mềm); Hệ CSDL SQL2005 chạy trên hệ điều hành Windows XP.

3.4.2 Xây dựng chương trình

Dựa trên cơ sở phân tích bài toán thực tế của TTCK chương trình được xây dựng trên nguyên tắc nguồn dữ liệu được cập nhật vào CSDL; sau đó được chuyển đổi sang CSDL giao dịch, hình thành kho dữ liệu. Tiến hành khai phá, sử dụng thuật toán Apriori và FP Growth cài đặt trên chương trình phân tích trích lọc tìm ra các tập mục phổ biến và luật kết hợp hình thành cơ sở tri thức tư vấn cho nhà đầu tư.

3.4.3 Giới thiệu chức năng chương trình

Hệ thống bao gồm các giao diện cập nhật CSDL, phân tích và đưa ra kết quả trợ giúp nhà đầu tư.

3.5 ĐÁNH GIÁ KẾT QUẢ

Chương trình đã cài đặt hoàn thành bước đầu đã mang lại kết quả; dùng thuật toán FP-Growth để phân tích giải quyết bài toán .

KẾT LUẬN

1. Kết quả đạt được

Với cách tiếp cận dựa trên những đề xuất đã có trong lĩnh vực nghiên cứu về KPDL, bản luận văn này là một sự tổng hợp được những kiến thức lí thuyết cơ bản trong KPDL nói chung và khai phá dữ liệu bằng luật kết hợp nói riêng cùng với một vài đề xuất mới. Sau đây là những điểm chính mà luận văn đã tập trung giải quyết:

Luận văn đã tìm hiểu và phân tích rõ nét về TTCK; trình bày tổng quan và các nét đặc trưng nhất trong lĩnh vực KPDL bao gồm các vấn đề cần khám phá tri thức, các hướng tiếp cận và nghiên cứu tiêu biểu, trong đó phát hiện luật kết hợp ứng dụng trong việc xây dựng hệ thống phân tích trợ giúp nhà đầu tư trong thị trường chứng khoán, trợ giúp nhà đầu tư nhỏ lẻ trong việc phân tích và chọn lựa việc đầu tư vào cổ phiếu, đây là một phương pháp khám phá tri thức quan trọng trong KPDL có nhiều ý nghĩa trong khoa học cũng như trong thực tiễn. Đây là chủ đề trọng tâm cho nội dung của luận văn.

Trong quá trình thực hiện luận văn, tôi đã cố gắng tập trung tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên, với thời gian và trình độ có hạn nên không tránh khỏi những hạn chế và thiếu sót. Tôi rất mong được sự nhận xét và góp ý của các thầy cô giáo và bạn bè, đồng nghiệp và những người cùng quan tâm để hoàn thiện hơn kết quả nghiên cứu của mình.

2. Định hướng phát triển

KPDL bằng luật kết hợp được xây dựng ứng dụng rộng rãi trong nhiều lĩnh vực như: TTCK, Ngân hàng, Kinh tế, Y học... mang lại

nhiều kết quả đáng tin cậy; tìm ra các luật có ý nghĩa thực tế ẩn chứa trong các CSDL lớn. Chính vì vậy tôi tiếp tục hoàn thiện các chứng năng và nâng cao chất lượng chương trình ứng dụng trong thời gian sắp đến để có thể áp dụng vào thực tế; người dùng thực sự tin tưởng sử dụng cho việc dự báo thông tin cổ phiếu; giúp nhà đầu tư chọn lựa phương án và ra quyết định đầu tư vào cổ phiếu để mang lại lợi nhuận.

Hướng giải quyết nhiệm vụ này ở chức năng cập nhật dữ liệu của chương trình được cải tiến trích lọc dữ liệu online; Sử dụng thuật toán xử lí song song trong luật kết hợp để tối ưu việc phân tích để giảm thiểu thời gian chạy chương trình. Ứng dụng luật kết hợp mờ cho kết quả dự báo chính xác hơn.