

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG**

HUỲNH ĐỨC THUẬN

**ỨNG DỤNG KHAI PHÁ DỮ LIỆU
XÂY DỰNG HỆ THỐNG PHÂN TÍCH
HOẠT ĐỘNG ĐẦU TƯ TRONG
THỊ TRƯỜNG CHỨNG KHOÁN**

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

**Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01**

ĐÀ NẴNG, NĂM 2010

MỞ ĐẦU

1. LÝ DO CHỌN ĐỀ TÀI

Khám phá tri thức (KPTT) hay khai phá dữ liệu (KPDL) trong cơ sở dữ liệu (CSDL) đang là một xu hướng quan trọng của nền công nghệ thông tin (CNTT) thế giới. KPTT có khả năng ứng dụng vào rất nhiều lớp bài toán thực tế khác nhau. Lĩnh vực tài chính nói chung và thị trường chứng khoán (TTCK) nói riêng lưu trữ một khối lượng dữ liệu khổng lồ, bao gồm thông tin các mã cổ phiếu, thông tin giao dịch và khối lượng giao dịch ròng, và thông tin dữ liệu về khách hàng... Ứng dụng sinh luật kết hợp từ KPDL để phát hiện ra quy luật 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, có hình thức và quy mô giao dịch phù hợp nhằm đạt được giá trị gia tăng hiệu quả. Tuy nhiên, trong bối cảnh hiện nay việc đầu tư vào TTCK hiện nay ở Việt Nam có rất nhiều khó khăn: lượng thông tin nhiều và không hợp nhất, sự chuyển biến khó đoán trước của diễn biến TTCK, các phần mềm trợ giúp hiện tại chưa phù hợp với môi trường TTCK tại Việt Nam... Đó là những khó khăn cần trợ giúp cho nhà đầu tư trong phân tích hoạt động đầu tư phù hợp trong TTCK.

2. MỤC TIÊU NGHIÊN CỨU

Xuất phát từ lý do đó tôi đã thực hiện đề tài: "**Ứng dụng khai phá dữ liệu xây dựng hệ thống phân tích hoạt động đầu tư trong thị trường chứng khoán**". Mục tiêu của đề tài là đề xuất giải pháp ứng dụng KPDL để xây dựng hệ thống trợ giúp nhà đầu tư trong công tác phân tích hoạt động đầu tư cổ phiếu hợp lý trong TTCK sao cho mang lại hiệu quả kinh tế trong điều kiện có thể.

Nhiệm vụ đầu tiên của đề tài là đánh giá được tính khả thi của chức năng phân tích chứng tỏ rằng các cổ phiếu trong TTCK thay đổi theo qui luật. Nhiệm vụ thứ hai là xem xét các lý thuyết, thuật toán phù hợp để áp dụng mô hình phân tích hoạt động đầu tư phù hợp trong điều kiện có thể.

3. ĐỐI TƯỢNG NGHIÊN CỨU

Phân tích hoạt động đầu tư trong TTCK là một nội dung rất khó vì tính biến động, không ổn định và khối lượng thông tin, dữ liệu trên thị trường ngày càng nhiều. Trước đây đã có một số luận văn đề cập đến KPDL nhưng chỉ ứng dụng trên các đối tượng đơn giản hơn như trợ giúp kinh doanh, trợ giúp phân loại văn bản... Với đề tài này việc thu thập dữ liệu cũng như xử lý được chúng để đưa ra những thông tin hữu ích nhất mang tính phức tạp và nhập nhằng.

4. PHƯƠNG PHÁP NGHIÊN CỨU

Để thực hiện luận văn tôi tiến hành nghiên cứu lý thuyết về KPDL ??? và ứng dụng thực tế tại các sàn giao dịch chứng khoán.

5. BỐ CỤC LUẬN VĂN

Bố cục của luận văn bao gồm những phần như sau : phần mở đầu trình bày lý do chọn đề tài, mục đích ý nghĩa và mục tiêu nhiệm vụ trong đề tài.

Trong chương một, luận văn tập trung giới thiệu TTCK và nhiệm vụ phân tích hoạt động đầu tư cổ phiếu, trong chương này ta tập trung tìm hiểu rõ về TTCK ở Việt Nam, các thông tin cần được sử dụng trong TTCK phục vụ cho mục đích, nhiệm vụ của đề tài.

Chương hai tập trung vào các phương thức dự báo cho TTCK: trong chương này ta tìm hiểu về luật kết hợp và thuật toán Apriori nhằm giải quyết các vấn đề khi tiến hành phân tích trong TTCK đã tìm hiểu ở chương một bằng KPDL.

Với những thực tiễn và khoa học được nêu ra trong chương một và hai, tôi xây dựng hệ thống ứng dụng trong chương ba. Đó là hệ thống phân tích và dự đoán bằng luật kết hợp của KPDL: trong chương này ta ứng dụng những giải quyết ở chương hai để xây dựng phần mềm tư vấn cho nhà đầu tư.

Từ những kết quả đạt được, phần cuối của luận văn nêu ra những phép đo tính hiệu quả của nghiên cứu, đưa ra đánh giá trên các kết quả đạt được, những hạn chế và đề xuất hướng nghiên cứu tiếp theo.

CHƯƠNG 1 : TÌM HIỂU THỊ TRƯỜNG CHỨNG KHOÁN VÀ HOẠT ĐỘNG ĐẦU TƯ

1.1 TÌM HIỂU VỀ TTCK

1.1.1 Đặc điểm TTCK

TTCK phong phú về lĩnh vực đầu tư, đa dạng về chủng loại hàng hóa và phức tạp về các qui luật đầu tư; là nơi mua bán các chứng khoán và thường được thực hiện chủ yếu tại sở giao dịch chứng khoán, một phần ở các công ty môi giới.

1.1.2 TTCK 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. Việc nghiên cứu và xây dựng một hệ thống phân tích và dự đoán (nhiệm vụ tư vấn) cho TTCK là quan trọng và cấp thiết cho các nhà đầu tư và nhà hoạch định chính sách vĩ mô. TTCK Việt Nam hiện tại gồm hai sàn giao dịch: HOSE và HASTC.

1.1.3 Những rủi ro gặp phải của nhà đầu tư

Các rủi ro thường gặp của nhà đầu tư: 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.

1.2 TÌM HIỂU PHƯƠNG PHÁP VÀ MÔ HÌNH PHÂN TÍCH HOẠT ĐỘNG ĐẦU TƯ

1.2.1 Tìm hiểu các phương pháp phân tích hoạt động đầu tư

Các phương pháp phân tích hiện nay chủ yếu dựa vào bốn cách chính: dựa vào các phân tích kỹ thuật để đưa ra tư vấn, dựa

vào các phân tích cơ sở để đưa ra tư vấn, dựa vào phương pháp dự báo chuỗi thời gian quá khứ và dựa vào phương pháp máy học

Trong phạm vi nghiên cứu và ứng dụng của luận văn sẽ tập trung vào phương pháp sử dụng tập dữ liệu mẫu và xem xét sự thay đổi của nó theo thời gian để đưa ra các phân tích và dự đoán

1.2.2 Mô hình hệ thống phân tích-dự đoán TTCK

Thu thập dữ liệu

Đây là quá trình lấy dữ liệu từ các nguồn internet, báo chí, thông cáo...

Phân tích ý nghĩa chỉ số

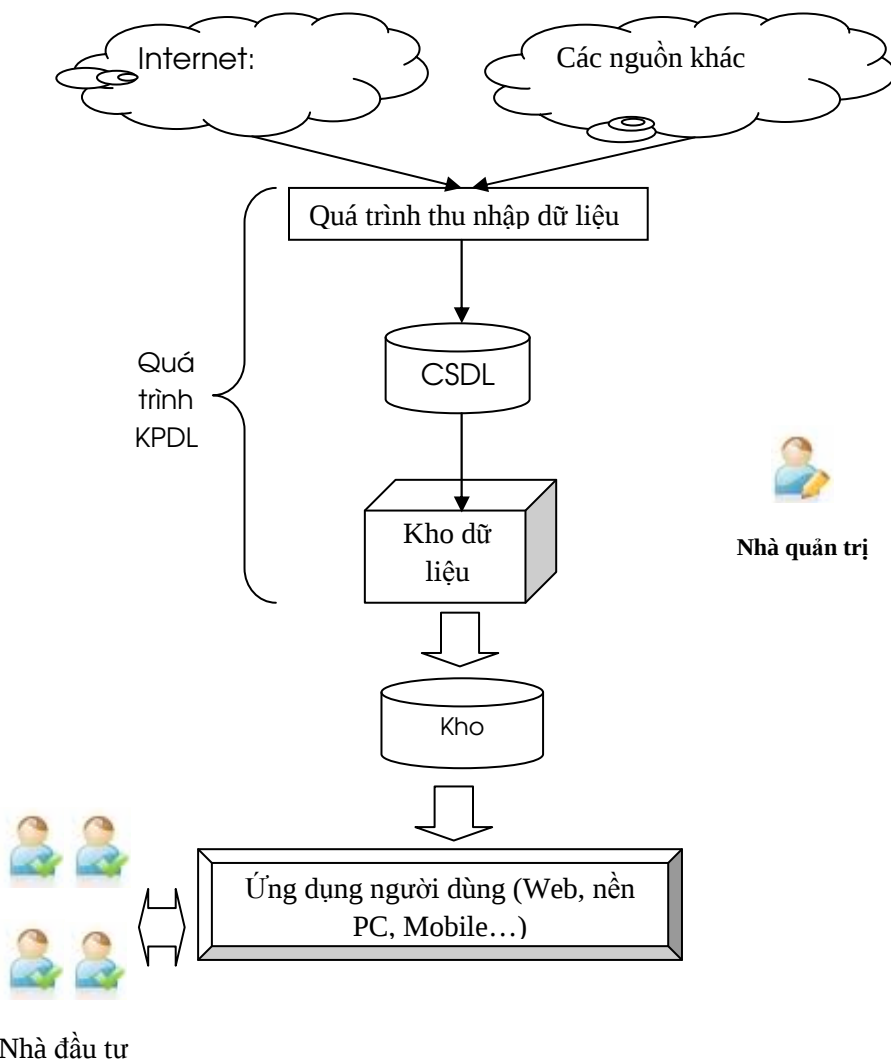
Dữ liệu sau khi được thu thập và chuyển đổi phù hợp sẽ được tiến hành phân tích và đưa ra các dự đoán.

Cung cấp thông tin tư vấn cho nhà đầu tư

Dữ liệu sau khi được phân tích dự báo sẽ được cung cấp cho nhà đầu tư thông qua các giao diện thân thiện

Tóm lại, mục đích chính của luận văn có thể được tóm tắt như sau: cho $t_i \{i = 1, 2, \dots, n\}$ là giá trị của cổ phiếu S trong các ngày thứ 1, 2, ..., n , chúng ta xác định được diễn biến cổ phiếu S trong các ngày $n + 1, n + 2, n + 3$

Quá trình trên được mô tả trong hình 1.1 dưới đây.



Hình 1.1. Mô hình hệ thống phân tích và dự đoán TTCK

1.3 CÁC THÔNG TIN LIÊN QUAN ĐẾN TƯ VẤN TRONG TTCK

1.3.1 Lí thuyết đầu tư

Giới đầu tư dựa vào hai lí thuyết chính: Firm Foundation và Castle in the Air. Dự theo những lí thuyết này chúng ta sẽ xác định được các thị trường định hình, hay nói cách khác là cách các nhà đầu tư nghĩ và phản ứng trước những thay đổi của chỉ số và làn sóng đầu tư.

1.3.2 Dữ liệu trong TTCK

Dữ liệu bao gồm các thông tin trên Web, thông tin niêm yết của chính công ty tham gia TTCK. Ngoài ra nhà đầu tư còn dựa vào loại dữ liệu kĩ thuật, dữ liệu sơ cấp và dữ liệu thứ cấp.

1.4 PHÂN TÍCH TRONG TTCK

1.4.1 Xác định nhiệm vụ phân tích hoạt động đầu tư

Nhiệm vụ tư vấn có hai mục đích chính. Đó là phân tích: dựa trên tất cả dữ liệu quá khứ, hiện tại để đưa ra các phân tích trên những chỉ số sẵn có, chẳng hạn: giá trị cổ phiếu đang tăng, nhà đầu tư đã không còn đầu tư vào cổ phiếu này...những phân tích này dựa trên số liệu thực tế nêu lại hiện trạng cho một loại cổ phiếu cho trước. Từ những phân tích đó, 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 kèm theo.

1.4.2 Khả năng phân tích hoạt động đầu tư trong TTCK

Khả năng tư vấn trong TTCK theo các học thuyết là khó theo EMH.

1.4.3 Phương thức phân tích hoạt động đầu tư

Chúng ta phân loại những kỹ thuật này như sau: phương pháp phân tích kỹ thuật, phương pháp phân tích cơ sở, phương pháp dự báo chuỗi thời gian quá khứ và phương pháp máy học. Tiêu chuẩn cho việc phân loại là loại công cụ và loại dữ liệu mà mỗi phương pháp được sử dụng để dự báo thị trường.

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, cổ phiếu và giao dịch. 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 cổ phiếu bằng các kỹ thuật KPDL.

CHƯƠNG 2 : TÌM HIỂU KHAI PHÁ DỮ LIỆU VÀ THUẬT TOÁN SINH LUẬT KẾT HỢP

2.1 MỞ ĐẦU

Trong chương hai, tôi đi sâu vào các phương pháp, kỹ thuật tư vấn thực tế trong thị trường chứng khoán, qua đó sử dụng các kiến thức của KPDL vào để phân tích và dự đoán các kết quả của TTCK.

2.2 KHAI PHÁ DỮ LIỆU (KPDL)

2.2.1 Các khái niệm cơ bản

Khi lưu trữ các dữ liệu khổng lồ thì chúng ta thấy rằng chắc chắn chúng phải chứa những giá trị nhất định nào đó. Tuy nhiên, theo thống kê thì chỉ có một lượng nhỏ của những dữ liệu này (khoảng từ 5% đến 10%) là luôn được phân tích, số còn lại họ không biết sẽ phải làm gì hoặc có thể làm gì với chúng nhưng họ vẫn tiếp tục thu thập rất tốn kém với ý nghĩ lo sợ rằng sẽ có cái gì đó quan trọng đã bị bỏ qua sau này có lúc cần đến nó. Mặt khác, trong môi trường cạnh tranh, người ta ngày càng cần có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định và ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Từ thực tế đó đã làm phát triển một khuynh hướng kỹ thuật mới đó là kỹ thuật phát hiện tri thức và khai phá dữ liệu.

2.2.2 Mục tiêu của khai phá dữ liệu

Mục tiêu chính của KPDL là lấy được những thông tin hữu ích từ lượng dữ liệu khổng lồ.

2.2.3 Các bước chính của khám phá tri thức

Gom dữ liệu (Gathering)

Tập hợp dữ liệu là bước đầu tiên trong quá trình KPDL. Đây là bước được khai thác trong một CSDL, một kho dữ liệu và thậm chí các dữ liệu từ các nguồn ứng dụng Web.

Trích lọc dữ liệu (Selection)

Ở giai đoạn này dữ liệu được lựa chọn hoặc phân chia theo một số tiêu chuẩn nào đó, ví dụ chọn tất cả những người có tuổi đời từ hai lăm đến ba lăm và có trình độ đại học.

Làm sạch, tiền xử lý và chuẩn bị trước dữ liệu (Cleansing, Pre-processing and Preparation)

Giai đoạn thứ ba này là giai đoạn hay bị sao lãng, nhưng thực tế nó là một bước rất quan trọng trong quá trình KPDL. Một số lỗi thường mắc phải trong khi gom dữ liệu là tính không đủ chặt chẽ, logic. Vì vậy, dữ liệu thường chứa các giá trị vô nghĩa và không có khả năng kết nối dữ liệu. Ví dụ: tuổi = sáu trăm bảy mươi ba. Giai đoạn này sẽ tiến hành xử lý những dạng dữ liệu không chặt chẽ nói trên. Những dữ liệu dạng này được xem như thông tin dư thừa, không có giá trị. Bởi vậy, đây là một quá trình rất quan trọng vì dữ liệu này nếu không được “làm sạch - tiền xử lý - chuẩn bị trước” thì sẽ gây nên những kết quả sai lệch nghiêm trọng.

Chuyển đổi dữ liệu (Transformation)

Tiếp theo là giai đoạn chuyển đổi dữ liệu, dữ liệu đưa ra có thể sử dụng và điều khiển được bởi việc tổ chức lại nó. Dữ liệu đã được chuyển đổi phù hợp với mục đích khai thác.

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

Đây là bước mang tính tư duy trong KPDL. Ở giai đoạn này nhiều thuật toán khác nhau đã được sử dụng để trích ra các mẫu từ dữ liệu. Thuật toán thường dùng là nguyên tắc phân loại, nguyên tắc kết hợp hoặc các mô hình dữ liệu tuần tự, v.v.

Đánh giá kết quả mẫu (Evaluation of Result)

Đây là giai đoạn cuối trong quá trình KPDL. Ở giai đoạn này, các mẫu dữ liệu được chiết xuất ra bởi phần mềm KPDL. Không phải bất cứ mẫu dữ liệu nào cũng đều hữu ích, đôi khi nó còn bị sai lệch. Vì vậy, cần phải ưu tiên những tiêu chuẩn đánh giá để chiết xuất ra các tri thức cần chiết xuất ra.

Trên đây là sáu giai đoạn trong quá trình KPDL, trong đó giai đoạn 5 là giai đoạn được quan tâm nhiều nhất hay còn gọi đó là KPDL.

2.2.4 Phát hiện vấn đề trong KPDL

Đây là một quá trình mang tính định tính với mục đích xác định được lĩnh vực yêu cầu phát hiện tri thức và xây dựng bài toán tổng kết.

2.2.5 Các hướng tiếp cận KPDL

Các hướng tiếp cận của KPDL có thể được phân chia theo chức năng hay lớp các bài toán khác nhau. Sau đây là một số hướng tiếp cận chính.

Hướng tiếp cận phổ biến là phân lớp và dự đoán, Một trong những hướng tiếp cận hiệu quả là sử dụng luật kết hợp, Một trong

những hướng tiếp cận dễ hình dung là khai phá chuỗi theo thời gian, Một hướng tiếp cận khó thực hiện là phân cụm

Một trong những hướng tiếp cận hiệu quả là sử dụng luật kết hợp (association rules): là dạng luật biểu diễn tri thức ở dạng khá đơn giản Phương pháp này nhằm phát hiện ra các luật kết hợp giữa các thành phần dữ liệu trong CSDL. Mẫu đầu ra của giải thuật KPDL là tập luật kết hợp tìm được.

2.2.6 Nhiệm vụ của KPDL

Những nhiệm vụ cơ bản nhất của khai phá dữ liệu là: phân cụm, phân loại, phân nhóm, phân lớp ; khai phá luật kết hợp; lập mô hình dự báo; phân tích đối tượng ngoài cuộc; phân tích sự tiến hóa.

2.2.7 Các kỹ thuật KPDL

Quá trình KPDL là quá trình phát hiện mẫu trong đó giải thuật KPDL tìm kiếm các mẫu đáng quan tâm theo dạng xác định như các luật, cây phân lớp, hồi quy, phân nhóm,... Các phương pháp phổ biến ở đây thường là phương pháp quy nạp, cây quyết định và luật, khai phá luật kết hợp, các phương pháp phân lớp và hồi quy phi tuyến, phân nhóm và phân đoạn, các phương pháp dựa trên mẫu, KPDL văn bản và mạng neuron.

2.2.8 Ứng dụng của KPDL

KPDL là một lĩnh vực được quan tâm và ứng dụng rộng rãi. Một số ứng dụng điển hình trong KPDL có thể liệt kê: phân tích dữ liệu và hỗ trợ ra quyết định; điều trị y học; phát hiện văn bản; tin sinh học; tài chính và TTCK; bảo hiểm...

2.2.9 Những tồn tại trong KPDL

Các tồn tại cần phải giải quyết trong KPDL: dữ liệu lớn; kích thước lớn; dữ liệu động; các trường dữ liệu không phù hợp; các giá trị bị thiếu; các trường dữ liệu bị thiếu; quá phù hợp; khả năng biểu đạt mẫu; sự tương tác với người sử dụng các tri thức sẵn có

2.3 KHAI PHÁ LUẬT KẾT HỢP

2.3.1 Tìm hiểu luật kết hợp

Luật kết hợp là dạng luật khá đơn giản nhưng lại mang khá nhiều ý nghĩa. Thông tin mà dạng luật này đem lại là rất đáng kể và hỗ trợ không nhỏ trong quá trình ra quyết định. Tìm kiếm được các luật kết hợp quý hiếm và mang nhiều thông tin từ CSDL tác nghiệp là một trong những hướng tiếp cận chính của lĩnh vực khai thác dữ liệu.

2.3.2 Định nghĩa

Cho $I = \{I_1, I_2, \dots, I_m\}$ là tập hợp của m tính chất riêng biệt. Giả sử D là CSDL, với các bản ghi chứa một tập con T các tính chất (có thể coi như T là tập con của I), các bản ghi đều có chỉ số riêng. Một luật kết hợp là một mệnh đề kéo theo có dạng $X \Rightarrow Y$, trong đó X, Y cũng là tập con của I , thỏa mãn điều kiện : X giao $Y = \text{trống}$. Các tập hợp X và Y được gọi là các tập mục (theo tiếng Anh là itemset).

2.3.3 CSDL giao dịch

CSDL GIAO DỊCH (Transaction DB) là một hệ CSDL dùng cho mục đích khai phá dữ liệu, được hình thành từ các nguồn dữ liệu gốc được chuyển đổi theo mục đích nào đó của người sử dụng

(ở đây là được chuyển đổi từ CSDL quan hệ các cổ phiếu được lấy từ nhiều nguồn khác nhau).

2.3.4 Giải thuật chuyển đổi CSDL

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 bằng luật kết hợp

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

Lĩnh vực khai thác luật kết hợp cho đến nay đã được nghiên cứu và phát triển theo nhiều hướng khác nhau: luật kết hợp nhị phân là hướng nghiên cứu đầu tiên của luật kết hợp, luật kết hợp có thuộc tính số và thuộc tính hạng mục, luật kết hợp tiếp cận theo hướng tập thô, luật kết hợp nhiều mức, luật kết hợp mờ, luật kết hợp với thuộc tính được đánh trọng số, luật kết hợp song song. Bên cạnh những nghiên cứu về các biến thể của luật kết hợp, các nhà nghiên cứu còn chú trọng đề xuất những thuật toán nhằm tăng tốc quá trình tìm kiếm tập phổ biến từ CSDL.

2.3.6 Bài toán luật kết hợp

Khái niệm: 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. Một giao dịch $T \in D$ hỗ trợ một tập $X \subseteq I$ nếu nó chứa tất cả các item của X.

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 thoả 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 chung là nếu gọi ABCD và AB là các tập mục phổ biến, thì chúng ta có thể xác định luật nếu AB.

2.3.7 Quy trình khai thác luật kết hợp

Bước một: Tìm tất cả các tập phổ biến (theo ngưỡng minsup)

Bước hai: Tạo ra các luật từ các tập phổ biến Đối với mỗi tập 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 LKH cần tìm nếu: $\text{conf}(A \Rightarrow (S - A)) = \text{supp}(S) / \text{supp}(A) \geq \text{minconf}$

2.3.8 Một số tính chất liên quan đến các hạng mục phổ biến:

Với tập mục phổ biến, có 3 tính chất sau:

Tính chất 1 (Độ hỗ trợ của tập con): Với A và B là tập các mục, nếu $A \subseteq B$ thì $\text{supp}(A) \geq \text{supp}(B)$. Điều này là rõ ràng vì tất cả các giao tác của D hỗ trợ B thì cũng hỗ trợ A.

Tính chất 2: Một tập chứa một tập không phổ biến thì cũng là tập không phổ biến. Nếu một mục trong B không có độ hỗ trợ tối thiểu trên D nghĩa là $\text{supp}(B) < \text{minsup}$ thì một tập con A của B sẽ không phải là một tập phổ biến vì $\text{support}(B) \leq \text{support}(A) < \text{minsup}$ (theo tính chất 1)

Tính chất 3: Các tập con của tập phổ biến cũng là tập phổ biến

Nếu mục B là mục phổ biến trên D , nghĩa là $\text{support}(B) \geq \text{minsup}$ thì mọi tập con A của B là tập phổ biến trên D vì $\text{support}(A) \geq \text{support}(B) > \text{minsup}$.

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

Độ hỗ trợ các vector chỉ báo nhị phân

Cho $X1 \subset D$, độ hỗ trợ của $vB(X1)$ biểu diễn $\text{supB}(vB(X1))$ được định nghĩa:

$$\text{supB}(vB(X1)) = \{o \in O \mid \forall d \in X1, \chi(o, d) = 1\}$$

$$\text{Để thấy rằng: } \text{card}(\text{supB}(vB(X1))) = \text{card}(\rho B(X1))$$

Tính $\text{card}(\rho B(S))$ (lực lượng của tập hợp): Cho $S = \{s1, s2, \dots, sk\}$ là tập con của D . Trong đó s_j là bộ chỉ báo của SB , $j = 1 \div k$. Mỗi s_j tương ứng với vector chỉ báo nhị phân $vB(\{s_j\})$. Các yếu tố của $\rho B(S)$ được tính bằng:

$$\text{card}(\rho B(S)) = \text{card}(\text{supB}(vB\{s1\}) \cap \dots \cap \text{supB}(vB\{sk\}))$$

2.4 THUẬT TOÁN SINH LUẬT KẾT HỢP

2.4.1 Thuật toán AIS

Thuật toán do Agrwal đề nghị năm 1993. Thuật toán này chú trọng khai phá luật kết hợp có dạng $X \rightarrow Y$, với Y là tập hợp chỉ bao gồm 1 tính chất (tập hợp một phần tử). Thuật toán tìm cách xây dựng dần dần các tập ứng cử viên cho tập mục phổ biến. Với cách đánh số thứ tự từ điển cho từng tính chất, việc bổ sung phần tử cho tập ứng cử viên tránh được trùng lặp, do vậy tiết kiệm tối đa thời gian tính toán.

2.4.2 Thuật toán SETM

Thuật toán do Houtsma đề nghị năm 1995. Thuật toán này cũng sử dụng kỹ thuật bổ sung dần dần từng phần tử (từ tập hợp 1 phần tử) nhằm tìm kiếm các tập hợp ứng cử viên. Một cải tiến đáng kể là Thuật toán đề nghị lưu lại cả ID của giao dịch cùng với tập hợp ứng cử viên. Agrawal đã chỉ ra, Thuật toán này không những không có phương án quản lý bộ nhớ mà nó còn giả định nhét toàn bộ tập hợp ứng cử viên của bước trước vào bộ nhớ để bước sau tiện bề sử dụng.

2.4.3 Thuật toán Apriori-Tid

Thuật toán được tĩa bớt những tập ứng cử viên có tập con không phổ biến trước khi tính độ hỗ trợ. Thuật toán Apriori tính tất cả các tập ứng cử của tập k trong một lần duyệt CSDL. Apriori dựa vào cấu trúc cây băm. Tìm kiếm đi xuống trên cấu trúc cây mỗi khi ta chạm lá, ta tìm được một tập ứng cử viên có tiền tố chung được bao gồm trong giao dịch. Sau đó các tập ứng cử này được tìm trong giao dịch đã được ánh xạ trước đó. Trong trường hợp tìm thấy biến đếm được tăng lên 1.

2.4.4 Thuật toán Apriori mở rộng để sinh ra luật kết hợp

Sau khi các tập mục phổ biến từ các tác vụ trong CSDL đã được tìm thấy, nó có thể sinh ra các luật kết hợp mạnh, ở đó luật kết hợp mạnh (strong association rule) là luật thoả mãn cả hai độ hỗ trợ cực tiểu và độ tin cậy cực tiểu. Điều đó có thể thực hiện bằng việc sử dụng tính độ tin cậy của luật, ta nhắc lại: độ tin cậy của luật $X \rightarrow Y$ là: $\text{conf}(X \rightarrow Y) = P(Y/X) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)}$,

ở đó $\text{sup}(X \cup Y)$ là độ hỗ trợ của $X \cup Y$ và $\text{sup}(X)$ là độ hỗ trợ của X .

Có thể coi tỷ số trên là tỷ số giữa: số các tác vụ chứa $X \cup Y$ và số các tác vụ chứa X . Dựa trên biểu thức tính toán đó, các luật kết hợp có thể được sinh như sau: 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 thoả mãn, tức là độ hỗ trợ của luật chính là $\text{sup}(l)$.

CHƯƠNG 3 : XÂY DỰNG HỆ THỐNG TRỢ GIÚP PHÂN TÍCH HOẠT ĐỘNG ĐẦU TƯ

3.1 PHÁT BIỂU BÀI TOÁN

Với số lượng giao dịch hàng ngày tăng, bình quân 80,650,490 lượt/ ngày. CSDL của giao dịch ngày càng tăng. Với mỗi ngày thay đổi, mỗi loại cổ phiếu sẽ tăng thêm 1 dòng trong CSDL, ngoài ra các thông tin khác cũng tăng thêm 1 dòng/1 ngày. Lượng CSDL mỗi lần phân tích có thể chia theo khoảng thời gian (1 tuần, 1 tháng, 3 tháng, 6 tháng, 12 tháng, 24 tháng, 36 tháng). Tất cả CSDL này hoàn toàn được truy xuất.

Các dữ liệu được thu thập về sẽ được phân tích, định dạng và chứa trong data warehouse, là loại dữ liệu được sử dụng để khai phá. Sau giai đoạn khai phá, ta sử dụng thuật toán Apriori để đưa ra các mẫu phân tích dùng cho dự đoán.

Các dữ liệu dự đoán sẽ bao gồm chỉ số cổ phiếu các ngày kết tiếp (ngày T+1, T+2 và T+3), các khả năng mua/bán/chuyển nhượng (gọi chung là giao dịch) hàng ngày, dự đoán các khả năng sẽ diễn ra của các cổ phiếu.

Phân tích cơ bản về chương trình

Giai đoạn tiền xử lý: Giai đoạn này nhằm thiết lập các đối tượng dữ liệu từ dữ liệu trong CSDL. Dữ liệu được tiền xử lý đưa về dạng text, các thuộc tính (chính là các item) được ánh xạ bởi các số tự nhiên (tức là đánh số thứ tự các thuộc tính từ 1 đến hết). Mỗi dòng (bản ghi) được mô tả thành một dòng. Tập hợp phổ biến và luật kết hợp dựa trên các số thứ tự này, kết quả được ánh xạ ngược trở lại tên các mục.

Giai đoạn khai phá: Đây là quá trình thực hiện các thuật toán (Apriori-Tid, Apriori mở rộng) áp dụng đối với dữ liệu cung cấp sau giai đoạn tiền xử lý

Các kết quả đạt được

Các kết quả phân tích ra sẽ được thể hệ qua một bản báo cáo, trên đó có các kết quả dự kiến hiển thị trong report tư vấn (cho một nhóm cổ phiếu được dự báo cùng tăng trong ngày).

3.2 CÁC NGUỒN DỮ LIỆU

Như đã phân tích ở trên các nguồn dữ liệu ở được thu thập từ nhiều nguồn khác nhau. Các thông tin của cổ phiếu được lấy trực tiếp từ sàn giao dịch chứng khoán quốc gia. Tổ chức dữ liệu vật lí chương trình gồm 2 thư mục chính: RawData chứa các file csv là dữ liệu gốc được lấy về, XMLData chứa các dữ liệu được xử lí và chứa trong kho trí thức nhằm phục vụ mục đích phân tích và dự đoán

3.3 KỊCH BẢN HỆ THỐNG

Một ví dụ của kịch bản hệ thống như sau:

Bước một: nhà đầu tư lựa chọn 4 ngày giao dịch (tương ứng có 4 dòng giao dịch trong CSDL), các cổ phiếu quan tâm bao gồm: VNE, HRC, MCV và KLS. Hệ thống sẽ phản hồi các thông tin của cổ phiếu và biểu đồ kĩ thuật.

CSDL có các dòng như sau: Giao dịch T1: VNE, HRC, MCV; Giao dịch T2: HRC, MCV, KLS; Giao dịch T3: HRC; Giao dịch T4: VNE, HRC.

Bước hai: đây là bước chạy của thuật toán. Nhà đầu tư lựa chọn mức độ tăng của cổ phiếu là tăng khá, độ hỗ trợ 40%.

Đầu tiên hệ thống quét trên tập 1-mục để tìm ra độ hỗ trợ tương ứng của chúng: {VNE}: độ hỗ trợ = $2/4 = 50\%$; {HRC}: độ hỗ trợ = $4/4 = 100\%$; {MCV}: độ hỗ trợ = $2/4 = 50\%$; {KLS}: độ hỗ trợ = $1/4 = 25\%$.

Hệ thống loại ra cổ phiếu KLS có độ hỗ trợ 25%. Khi đó danh mục tập mục 1-mục thỏa mãn tập mục phổ biến chỉ còn VNE, HRC và MCV.

Hệ thống sinh ra tập 2-mục và tính toán độ hỗ trợ tương ứng: {VNE, HRC}: độ hỗ trợ = $2/4 = 50\%$; {VNE, MCV}: độ hỗ trợ = $1/4 = 25\%$; {HRC, MCV}: độ hỗ trợ = $2/4 = 50\%$.

Hệ thống loại ra nhóm cổ phiếu {VNE, MCV} có độ hỗ trợ 25%. Khi đó danh mục tập mục 2-mục thỏa mãn tập mục phổ biến chỉ còn {VNE, HRC} và {HRC, MCV}.

Khi xét đến tập 3-mục {VNE, HRC, MCV} hệ thống tính toán độ hỗ trợ chỉ đạt 25% nên loại bộ 3 cổ phiếu này ra khỏi tập mục phổ biến.

Vậy tập mục phổ biến bây giờ có 5 dòng trong CSDL: {VNE}: độ hỗ trợ = $2/4 = 50\%$; {HRC}: độ hỗ trợ = $4/4 = 100\%$; {MCV}: độ hỗ trợ = $2/4 = 50\%$; {VNE, HRC}: độ hỗ trợ = $2/4 = 50\%$; {HRC, MCV}: độ hỗ trợ = $2/4 = 50\%$.

Bước ba: nhà đầu tư nhập vào độ tin cậy là 80%, hệ thống sinh ra các luật kết hợp tương ứng với độ tin cậy này (xem 80% như là độ tin cậy tối thiểu).

Luật 1: HRC $\rightarrow$ VNE: đạt độ tin cậy 100%

Luật 2: HRC $\rightarrow$ MCV: đạt độ tin cậy 100%

Hai luật không thỏa mãn là:

Luật 1': VNE $\rightarrow$ HRC: đạt độ tin cậy 50% < 80%

Luật 2': MCV $\rightarrow$ HRC: đạt độ tin cậy 50% < 80%

Như vậy hệ thống hoàn tất quá trình phân tích của mình sau khi tìm ra được các luật kết hợp và diễn giải cho chúng.

3.4 CÀI ĐẶT CHƯƠNG TRÌNH

Ngôn ngữ và môi trường lập trình

Để hoàn thành hệ thống phân tích hoạt động đầu tư trong TTCK, tôi sử dụng ngôn ngữ lập trình C# trong bộ Visual Studio 2008, chạy trên nền CSDL SQL2000.

Xây dựng chương trình

Chương trình được xây dựng trên nguyên tắc từ các dữ liệu gốc, sau khi thực hiện thuật toán chuyển đổi CSDL sẽ sinh ra CSDL giao dịch, hình thành kho dữ liệu. Thuật toán Apriori được áp dụng trên CSDL giao dịch này để tìm ra các tập mục phổ biến và luật kết hợp, cơ sở hình thành phân tích và dự đoán.

Với thuật toán Apriori, toàn bộ CSDL giao dịch được chuyển đổi thành các tập luật cần thiết, chính tập luật này là cơ sở để phân tích và dự đoán xu hướng cổ phiếu.

Tổng hợp của toàn bộ luật trên là các phân tích kèm theo, dựa vào những kết quả cổ phiếu hiện tại và những kết quả về độ tin cậy hạng mục và tập luật kết hợp được sinh ra, chương trình sẽ cho ra các phân tích và dự đoán đó hỗ trợ nhà đầu tư trong việc dự đoán TTCK trong tương lai.

KẾT LUẬN

1. Đánh giá

Đối với lý thuyết, khai phá tri thức bao gồm các bước: Hình thành, xác định và định nghĩa bài toán; thu thập và tiền xử lý dữ liệu; KPDL, rút ra các tri thức; sử dụng tri thức phát hiện được nhằm trợ giúp cho việc ra quyết định tư vấn trong giao dịch cổ phiếu. Về thuật toán KPTT, luận văn trình bày và minh họa một số thuật toán, luật kết hợp để phát hiện tập chỉ mục phổ biến và qui luật trong giao dịch cổ phiếu. Về mặt cài đặt thử nghiệm, luận văn giới thiệu kỹ thuật KPDL theo thuật toán Apriori áp dụng vào bài toán phân tích và dự đoán kết quả của TTCK.

2. Định hướng phát triển

Trong luận văn này tôi đã thực hiện các nhiệm vụ đã nêu ra. Với những kết quả đạt được và được sự giới thiệu của CBHD, luận văn có thể được nghiên cứu và triển khai ở diện rộng và sâu hơn.

Hiện nay các thuật toán được cung cấp thực hiện tốt nhiệm vụ của nó. Tuy nhiên thị trường chứng khoán diễn biến liên tục theo thời gian thực nên thuật toán cũng phải đáp ứng được thời gian thực này. Tôi sẽ nghiên cứu sâu các thuật toán KPDL, tiếp tục hoàn thiện và mở rộng chương trình để có thể thỏa mãn yêu cầu này. Trong đó các kỹ thuật chọn lựa để xử lý sẽ được quan tâm nhiều hơn (giống như mờ hóa thông tin) để loại bỏ các dữ liệu ít liên quan, chỉ xử lý trên lượng dữ liệu được quan tâm. Ngoài ra qua quá trình khảo sát tại sàn giao dịch chứng khoán Thăng Long nhận thấy nhà đầu tư có rất nhiều nhu cầu để có một cách thức

giao dịch thuận tiện. Ngoài ra hiện nay Chính phủ chuẩn bị đưa ra qui định cho phép giao dịch chứng khoán qua mạng Internet (vào cuối năm 2010) nên các ứng dụng nếu được xây dựng trên nhiều nền hệ thống khác nhau sẽ được sử dụng rộng rãi (ứng dụng web, ứng dụng trên điện thoại di động, giải pháp SMS...). Từ đó tôi sẽ tiếp tục nghiên cứu các công nghệ này và chuyển đổi ứng dụng cho phù hợp nhu cầu của nhà đầu tư thực hiện đúng qui định của Nhà nước.