

دانشگاه پیام نور
دانشکده فنی و مهندسی

پایان نامه

برای دریافت درجه کارشناسی ارشد
رشته مهندسی کامپیوتر - گرایش نرم افزار
گروه مهندسی کامپیوتر و فناوری اطلاعات

ارائه یک الگوریتم به منظور کشف قوانین وابستگی در تغییرات قیمت سهام (مطالعه موردی: بورس تهران)

نگارش:

مجتبی علاء

استاد راهنما:

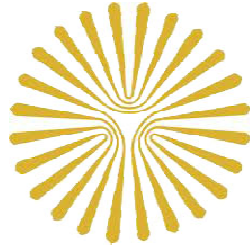
دکتر محمد صنیعی آباده

استاد مشاور:

دکتر داود کریم زادگان مقدم

اردیبهشت ۱۳۹۱

سورة الاحقاف



دانشگاه پیام نور
دانشکده فنی و مهندسی
دانشگاه پیام نور مرکز تهران

پایان نامه
برای دریافت درجه کارشناسی ارشد
رشته مهندسی کامپیوتر - گرایش نرم افزار
گروه مهندسی کامپیوتر و فناوری اطلاعات

ارائه یک الگوریتم به منظور کشف قوانین وابستگی در تغییرات قیمت سهام (مطالعه موردی: بورس تهران)

نگارش:
مجتبی علاء

استاد راهنما:
دکتر محمد صنیعی آباده

استاد مشاور:
دکتر داود کریم زادگان مقدم

تقديم به:

پدر و مادر عزيزم

تشکر و قدردانی:

در انجام این تحقیق، بر خود واجب می‌دانم از زحمات استاد ارجمند جناب آقای دکتر محمد صنیعی آباده، که مسئولیت راهنمایی پایان‌نامه را بر عهده داشتند، جناب آقای دکتر داود کریم زادگان مقدم، استاد محترم مشاور و نیز جناب آقای دکتر رضا عسگری مقدم، که با دقت و حوصله فراوان داوری این پایان‌نامه را به انجام رساندند، صمیمانه تشکر کنم. همچنین باید سپاس فراوان خود را تقدیم دوست ارجمند جناب آقای حمید محمدی نمایم که با حمایت‌های خود، زمینه هر چه بهتر شدن این تحقیق را فراهم ساخت.

چکیده

کاوش قوانین وابستگی یک تکنیک مهم در داده کاوی است که در آن یافتن الگوهای تکراری، یک مرحله بسیار زمان بر و پرهزینه به شمار می رود. در این تحقیق یک روش جدید برای یافتن الگوهای تکراری بر روی داده هایی که مقادیر آن ها به شکل نوسانی (افزایش، کاهش یا بدون تغییر) هستند مانند داده های تغییرات قیمت سهام در بورس اوراق بهادار ارائه می شود. در این روش، فقط تعداد کمی از تکرار مجموعه عناصر از پایگاه داده جستجو می شود و تکرار بقیه مجموعه عناصر به کمک تکرار مجموعه عناصر بدست آمده و روابط پیدا شده بین تکرار آن ها محاسبه می شود. بدین ترتیب نیاز به کاوش کل الگوهای تکراری از پایگاه داده نیست. این امر باعث صرف زمان بسیار کمتری برای یافتن الگوهای تکراری و در نهایت یافتن قوانین وابستگی می شود. با توجه به نکات یاد شده، در این تحقیق با پیاده سازی الگوریتم FP-Math (مخفف Frequent Patterns-Math) کارایی آن بر روی داده های واقعی بورس تهران بررسی و نشان داده می شود. همچنین نتایج آزمایشات نشان می دهد الگوریتم FP-Math در مقایسه با الگوریتم های پیشین مانند الگوریتم Apriori، که یکی از مشهورترین و پر کاربردترین آن ها در این زمینه است، تقریباً سه برابر سریع تر اجرا می شود. در مجموع، بهره گیری از الگوریتم پیشنهادی توانست الگوهای تکراری و قوانین وابستگی جالبی را از داده های بورس اوراق بهادار تهران استخراج نماید.

واژه های کلیدی: داده کاوی، کاوش قوانین وابستگی، الگوهای تکراری، بورس.

فهرست مطالب

صفحه	عنوان
۱	فصل ۱ مقدمه
۲	۱-۱. مقدمه
۴	۲-۱. تعریف مسأله و سوالات اصلی تحقیق
۶	۳-۱. فرضیه‌ها
۶	۴-۱. اهداف تحقیق
۷	۵-۱. روش تحقیق
۷	۶-۱. مراحل انجام تحقیق
۸	۷-۱. دست آوردهای این تحقیق
۸	۸-۱. ساختار تحقیق
۱۰	فصل ۲ مروری بر منابع مطالعاتی
۱۱	۱-۲. مقدمه
۱۲	۲-۲. الگوریتم‌های کاوش الگوهای تکراری و قوانین وابستگی
۱۴	۳-۲. استفاده از قوانین وابستگی در بازار سرمایه
۱۷	۱-۳-۲. الگوریتم‌های قوانین وابستگی مختص داده‌های بورس
۱۸	۲-۳-۲. نگارش‌های گذشته الگوریتم پیشنهادی
۱۸	۴-۲. نتیجه‌گیری
۲۰	فصل ۳ داده‌کاوی و کاوش قوانین وابستگی
۲۱	۱-۳. مقدمه
۲۲	۲-۳. داده‌کاوی
۲۳	۱-۲-۳. تکنیک‌های داده‌کاوی
۲۵	۲-۲-۳. انتخاب تکنیک‌های داده‌کاوی
۲۵	۳-۲-۳. مراحل اصلی داده‌کاوی

۲۶ ۴-۲-۳ محدودیت‌های داده‌کاوی
۲۷ ۳-۳ کاوش قوانین وابستگی
۲۸ ۱-۳-۳ مجموعه عنصر
۲۸ ۲-۳-۳ مجموعه عناصر تکراری
۲۹ ۳-۳-۳ پشتیبانی
۳۰ ۴-۳-۳ اطمینان
۳۰ ۵-۳-۳ تولید قوانین وابستگی
۳۱ ۴-۳ داده‌هایی با نوع مقدار نوسانی
۳۳ ۵-۳ تعاریف به شکل ریاضی
۳۴ ۶-۳ جمع‌بندی

فصل ۴ الگوریتم FP-Math

۳۵	
۳۶ ۱-۴ مقدمه
۳۷ ۲-۴ مفهوم اصلی
۳۸ ۱-۲-۴ تولید مجموعه عناصر کاندید
۴۰ ۲-۲-۴ تولید تمام مجموعه عناصر ممکن
۴۱ ۳-۲-۴ یافتن تکرار مجموعه عناصر
۴۶ ۳-۴ الگوریتم FP-Math
۴۷ ۱-۳-۴ مرحله اول: تولید مجموعه عناصر کاندید
۴۸ ۲-۳-۴ مرحله دوم: هرس کردن مجموعه عناصر غیر تکراری
۵۰ ۳-۳-۴ مرحله سوم: شمارش برخی از مجموعه عناصر از طریق پایگاه داده
۵۱ ۴-۳-۴ مرحله چهارم: محاسبه بقیه مجموعه عناصر، بدون نیاز به پایگاه داده
۵۲ ۴-۴ جمع‌بندی

فصل ۵ ارزیابی عملکرد

۵۳	
۵۴ ۱-۵ مقدمه
۵۵ ۲-۵ پارامترهای ارزیابی
۵۵ ۳-۵ نتایج ارزیابی
۵۶ ۱-۳-۵ مقایسه حجم فایل تراکنش‌ها
۵۶ ۲-۳-۵ تأثیر افزایش تعداد تراکنش‌ها
۵۷ ۳-۳-۵ تأثیر افزایش طول تراکنش‌ها
۵۸ ۴-۳-۵ تأثیر افزایش حداقل پشتیبانی
۶۱ ۴-۵ الگوهای تکراری و قوانین استخراج شده از داده‌های بورس تهران
۶۵ ۵-۵ جمع‌بندی

فصل ۶ جمع‌بندی و پیشنهادات

۶۶

۶-۱. مقدمه ۶۷

۶-۲. یافته‌های تحقیق ۶۸

۶-۳. نوآوری تحقیق ۶۹

۶-۴. پیشنهادات ۶۹

۷۱

مراجع

۷۴

واژه‌نامه

فهرست شکل‌ها

عنوان صفحه

فصل ۱ مقدمه

- شکل ۱-۱. عوامل تأثیر گذار بر روی قیمت سهام شرکتهای بورس اوراق بهادار تهران ۲
- شکل ۱-۲. تغییرات قیمت سهام بانک پارسیان در یک دوره زمانی سه ماهه بر اساس نمودار کندل استیک ۳

فصل ۲ مروری بر منابع مطالعاتی

فصل ۳ داده‌کاوی و کاوش قوانین وابستگی

- شکل ۳-۱. مجموعه عناصر کاندید ۳۱

فصل ۴ الگوریتم FP-Math

- شکل ۴-۱. تولید مجموعه عناصر کاندید ۳۹
- شکل ۴-۲. یک مجموعه عنصر کاندید ۲ تایی و ۸ مجموعه عنصر تولید شده توسط آن ۴۰
- شکل ۴-۳. یک مجموعه عنصر کاندید k تایی و ۸ مجموعه عنصر تولید شده توسط آن ۴۱
- شکل ۴-۴. روابط بین مجموعه عناصر ۲ تایی ۴۳
- شکل ۴-۵. روابط بین مجموعه عناصر k تایی ۴۴
- شکل ۴-۶. الگوریتم FP-Math ۴۷
- شکل ۴-۷. تابع کاندید ۴۸
- شکل ۴-۸. تابع هرس ۵۰
- شکل ۴-۹. تابع محاسبه ۵۱

فصل ۵ ارزیابی عملکرد

- شکل ۵-۱. دیتاست در مقابل حجم فایل ۵۶
- شکل ۵-۲. تعداد تراکنش‌ها در مقابل زمان اجراء با پشتیبانی ۲۰٪ ۵۷
- شکل ۵-۳. تراکنش‌ها در مقابل زمان اجراء، با حداقل پشتیبانی ۲۰٪ ۵۸
- شکل ۵-۴. حداقل پشتیبانی در مقابل زمان اجراء، دیتاست ۴ ۵۹
- شکل ۵-۵. حداقل پشتیبانی در مقابل زمان اجراء، دیتاست ۴ ۶۰
- شکل ۵-۶. حداقل پشتیبانی در مقابل زمان اجراء، دیتاست ۳ ۶۰

فصل ۶ جمع بندی و پیشنهادها

فهرست جداول

فصل ۱ مقدمه

فصل ۲ مروری بر منابع مطالعاتی

فصل ۳ داده‌کاوی و کاوش قوانین وابستگی

- جدول ۳-۱. پایگاه داده بورس در ۴ روز معاملاتی برای شرکت‌های ایران خودرو، داروسازی جابر ابن حیان و پتروشیمی اصفهان ۳۱
- جدول ۳-۲. پایگاه داده تراکشی ۳۲

فصل ۴ الگوریتم FP-Math

فصل ۵ ارزیابی عملکرد

- جدول ۵-۱. تنظیمات پارامترها برای ۵ دیتاست ۵۵
- جدول ۵-۲. برخی از الگوهای تکراری تولید شده با حداقل پشتیبانی ۶ درصد ۶۱
- جدول ۵-۳. برخی از قوانین وابستگی استخراج شده از داده‌های واقعی بورس با حداقل اطمینان ۳۰ درصد ۶۳

فصل ۶ جمع بندی و پیشنهادها

فهرست علائم اختصاری

AR	Association Rules	قوانین وابستگی
DM	Data Mining	داده کاوی
DB	Database	پایگاه داده
FP	Frequent Patterns	الگوهای تکراری
KDD	Knowledge Discovery from Database	استخراج دانش از پایگاه داده
TID	Transaction Identify	شناسه تراکنش

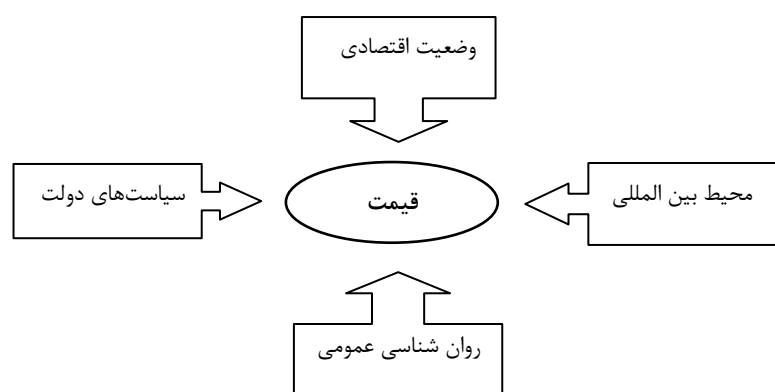
فصل ۱

مقدمه

۱-۱. مقدمه

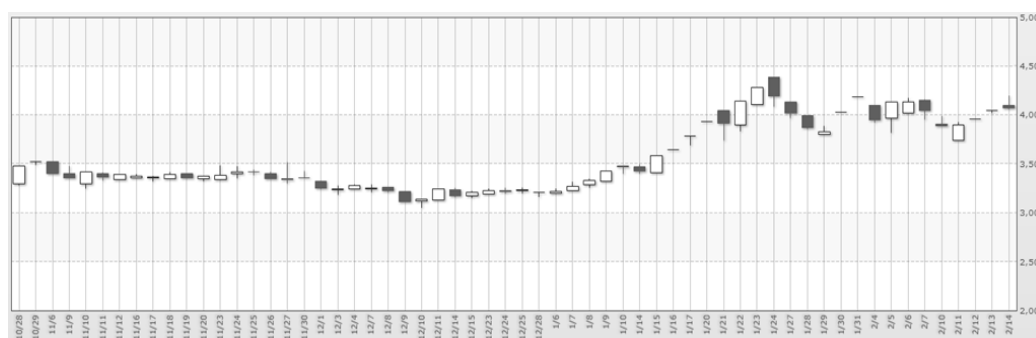
بورس اوراق بهادار تهران، بازار رسمی و سازمان یافته سرمایه است که در آن اوراق بهادار پذیرفته شده از قبیل سهام شرکت‌ها و اوراق مشارکت، تحت قوانین و مقررات خاصی داد و ستد می‌شوند. رسالت این بازار، جذب سرمایه‌های اندک و سرگردان عموم و هدایت آن به سمت فعالیت‌هایی است که متضمن منفعت عامه هستند.

قیمت سهام تحت تأثیر عوامل جالب زیادی مانند محیط بین‌المللی، سیاست‌های دولت، وضعیت آب و هوا، وضعیت اقتصادی، روان‌شناسی عمومی و حتی شایعات دچار تغییر و نوسان (مثبت و منفی) می‌شود. شکل ۱-۱ به خوبی بیانگر تأثیرگذاری عوامل مختلف بر روی قیمت سهام است.



شکل ۱-۱. عوامل تأثیرگذار بر روی قیمت سهام شرکت‌های بورس اوراق بهادار تهران

در زمینه سرمایه‌گذاری در بازار بورس، سرمایه‌گذاران در یک تحلیل تکنیکی، معتقد هستند که همه اطلاعات یک سهام در سابقه قیمت آن نهفته است و به عبارتی روند قیمت سهام تکرار شونده است به طوری که بررسی نوسانات قیمت در گذشته، آینده بازار را نشان می‌دهد؛ بنابراین اطلاعات ارزشمندی در داده‌های تاریخی بازارهای بورس اوراق بهادار از جمله بازار بورس تهران وجود دارد که تلاش برای استخراج این اطلاعات و استفاده از آن در جهت کمک به تصمیم‌گیری سرمایه‌گذاران حائز اهمیت فراوان می‌باشد. شکل ۱-۲ نمونه‌ای از روند قیمت سهام را در بازار بورس تهران در یک دوره زمانی سه ماهه نشان می‌دهد.



شکل ۱-۲. تغییرات قیمت سهام بانک پارسیان در یک دوره زمانی سه ماهه بر اساس نمودار کندل استیک^۱

یکی از مهم‌ترین تکنیک‌ها در داده‌کاوی که تحقیقات گسترده‌ای را در بازار سرمایه به خصوص بورس اوراق بهادار در سال‌های اخیر به خود معطوف نموده، کاوش قوانین وابستگی است. اما با وجود پیشرفت‌های صورت گرفته در این تکنیک، کاوش الگوهای تکراری در آن به دلیل جستجوی مجموعه عناصر تکراری، بسیار زمان‌بر و پرهزینه می‌باشد.

با توجه به نقطه ضعف یاد شده، هدف از انجام این تحقیق ارائه یک الگوریتم سریع برای یافتن قوانین وابستگی در داده‌های بورس اوراق بهادار می‌باشد. جهت رسیدن به این هدف، در

^۱ Candlestick

مرحله کاوش الگوهای تکراری، در گام نخست با استفاده از ویژگی‌های داده‌های بورس، روابط جالبی بین تکرار مجموعه عناصر پیدا می‌شود که در گاه بعدی با به کار گیری آن کشف الگوهای تکراری سریع‌تر میسر خواهد گردید.

به منظور دستیابی به ساختاری منسجم و مناسب برای انجام تحقیق، در ادامه این فصل به بیان مهم‌ترین اصول و پاسخ‌گویی به سوالات اصلی پژوهش، پرداخته خواهد شد.

۲-۱. تعریف مسأله و سوالات اصلی تحقیق

داده‌کاوی^۱ فرآیندی است که با نگرشی نو، به مسئله استخراج اطلاعات ناشناخته و بالقوه مفید از پایگاه داده‌ها می‌پردازد (هان^۲ و کمبر^۳، ۲۰۰۶). طی دهه‌های گذشته الگوریتم‌ها و تکنیک‌های مختلفی برای بدست آوردن دانش مبتنی بر قاعده در داده‌کاوی توسعه یافته‌اند. یکی از مطرح‌ترین این روش‌ها، **کاوش قوانین وابستگی**^۴ است که در آن روابط تکراری بین اشیاء مختلف با یکدیگر (الگوهای تکراری)، کشف می‌شود.

استفاده از قوانین وابستگی در تحلیل داده‌های بازار سرمایه می‌تواند برای یافتن وابستگی بین تغییرات قیمت سهام شرکت‌های بورسی و نشان دادن قواعد و الگوهای موجود در آن‌ها به منظور پیشنهاد سبد سهام و گروه‌بندی مجموعه‌های سرمایه‌گذاری سهام مورد استفاده قرار گیرد.

هر چند الگوریتم‌های متنوع زیادی وجود دارند که به کشف الگوهای تکراری و قوانین

¹ Data Mining

² Han, J.

³ Kamber, M.

⁴ Association Rule Mining

وابستگی در پایگاه داده‌های مختلف می‌پردازند، ولی با توجه به تعداد زیاد شرکت‌های پذیرفته شده در بورس و متراکم بودن پایگاه داده‌های بورس، کشف قوانین وابستگی در داده‌های بورس در مرحله یافتن الگوهای تکراری زمان بسیار زیادی را صرف می‌کند. به این دلیل و با در نظر گرفتن اهمیت تحلیل داده‌های بازار سرمایه، نیاز به یک الگوریتم سریع مختص این نوع داده‌ها احساس می‌شود.

در این پایان‌نامه، ضمن بررسی ویژگی‌های داده‌های بورس، با تمرکز بر روی روابط بین تکرار الگوها در این نوع داده‌ها، روش سریعی برای کشف الگوهای تکراری پیدا خواهد شد که با استفاده از آن، یک الگوریتم بسیار سریع مختص داده‌های با نوع مقدار نوسانی، مانند داده‌های بورس، ارائه می‌شود. همچنین با توجه به قابلیت قیاس بین سبد سهام^۱ و سبد بازار، از الگوریتم پیشنهادی برای پیدا کردن وابستگی‌ها در بین داده‌های بورس تهران استفاده خواهد شد. الگوهای تکراری در بین چند شرکت فعال تر^۲ در صنایع مختلف در بورس تهران کشف خواهد شد که ارتباط بین تغییرات قیمت سهام آن‌ها را نشان می‌دهد. در نتیجه، برای سرمایه‌گذاران اطلاعات مفیدی درباره رفتار بازار سهام فراهم خواهد شد که به کمک این اطلاعات می‌توانند سیاست سرمایه‌گذاری‌هایشان را به صورت دقیق‌تری مشخص کنند. در نهایت از داده‌های واقعی بورس تهران جهت بررسی عملکرد این الگوریتم استفاده خواهد شد.

بر این اساس، پرسش پژوهش پیش رو این است که «آیا تکنیک قوانین وابستگی با استفاده از الگوریتم پیشنهادی می‌تواند به نحو مؤثرتری الگوهای تکراری را از تغییرات قیمت سهام در داده‌های بورس تهران استخراج کند؟».

^۱ Portfolio

^۲ ۵۰ شرکت به عنوان شرکت‌های فعال تر در هر دوره ۳ ماه توسط بورس اوراق بهادار تهران معرفی می‌شود.

۳-۱. فرضیه‌ها

- پایگاه داده بورس، متراکم^۱ است و چون نمادها به ازای هر روز معاملاتی دارای یک تراکنش هستند، شامل حجم بزرگی از اطلاعات می‌باشد.
- برای کاوش قوانین وابستگی، از نوع مقدار داده به صورت تغییرات (افزایش، کاهش یا بدون تغییر) قیمت سهام استفاده می‌شود.
- محدودیت حافظه وجود ندارد.
- در دوره مورد بررسی پارامترهای اقتصادی بورس اوراق بهادار تهران نباید تحت تأثیر رخدادهای کاملاً ویژه و بحرانی قرار گیرد.

۴-۱. اهداف تحقیق

هدف اصلی پژوهش حاضر آن است که برای کشف قوانین وابستگی، یک الگوریتم جدید با سرعت بسیار بالاتر، مختص داده‌های بورس، ارائه دهد. با استفاده از این الگوریتم، کشف وابستگی‌ها در داده‌های بورس، به طور مؤثرتری انجام خواهد شد.

اهداف فرعی از این پژوهش نیز عبارتند از:

- کشف الگوهای تکراری و قوانین وابستگی در تغییرات قیمت سهام چند شرکت فعال تر در بورس اوراق بهادار تهران؛

^۱ متراکم بودن پایگاه داده بورس به این معنی می‌باشد که تمام نمادهای در تمام تراکنش‌ها حتماً دارای یکی از سه مقدار افزایش، کاهش و یا بدون تغییر می‌باشند.

- تسهیل سرمایه‌گذاری در بورس اوراق بهادار؛
- ایجاد اطمینان خاطر به سرمایه‌گذاران در پیش‌بینی قیمت سهام؛

۱-۵. روش تحقیق

این پژوهش به روش تحلیلی- تطبیقی صورت خواهد گرفت و در آن سعی خواهد شد به دو صورت مطالعه تاریخی (مطالعه طرح‌ها و روش‌های مختلف موجود در راستای موضوع تحقیق) و تطبیقی (مقایسه با روش‌های مرتبط موجود) سوالات و فرضیات تحقیق مورد ارزیابی و آزمون قرار گیرد.

۱-۶. مراحل انجام تحقیق

- شناسایی و بیان کامل مسئله
- ارزیابی روش‌های موجود
- تشریح مسائل مربوط به کاوش الگوهای تکراری بر روی داده‌های با نوع مقدار نوسانی
- ارائه یک روش جدید برای کشف قوانین وابستگی در داده‌های بورس و پیاده‌سازی الگوریتم آن
- ارزیابی عملکرد الگوریتم جدید از طریق مقایسه با الگوریتم‌های پیشین
- استخراج الگوها و قوانین وابستگی از داده‌های تاریخی بازار بورس توسط الگوریتم جدید