Разработка математической модели и программного комплекса для автоматизации научных исследований в области анализа новостей финансовой отрасли

В.С. Гаврилов, С.А. Корчагин

Финансовый университет при Правительстве Российской Федерации, Москва

Аннотация: Статья посвящена разработке математической модели и программного комплекса, предназначенных для автоматизации научных исследований в области анализа новостей финансовой отрасли. Авторы предлагают подход, основанный на использовании методов теории графов для выявления наиболее значимых научных гипотез, используемых методов, а также полученных качественных и количественных результатов научного сообщества в данной области. Предложенная модель и программный комплекс позволяют автоматизировать процесс научных исследований, что способствует более эффективному ее анализу. Результаты исследований могут быть полезны как для профессиональных участников финансовых рынков, так и для академического сообщества, поскольку выявление наиболее цитируемых и фундаментальных работ служит отправной точкой любой научной работы.
Ключевые слова: программный комплекс, моделирование, теория графов, новостные потоки, фондовый рынок РФ, акции, граф цитирований.

Введение

Ообъектом настоящего исследования являются новостные потоки, как основной фактор волатильности фондового рынка [1-3].

Для подбора релевантной литературы необходима теоретическая база, которая позволила бы систематизировать различные понятия и теории. После этого требуется выявить и рассмотреть различные подходы к проблеме количественной и качественной оценки влияния новостных и тематических новостных потоков на российский фондовый рынок [4].

Актуальность работы заключается в том, что «ручной» анализ существующих источников новостей невозможен в силу большой трудоемкости и больших затрат времени. Более того, как правило, одни и те же теории и методологии используются у разных авторов на разных этапах исследований. Средства автоматизации и комплексы программ помогают систематизировать и сделать акцент на наиболее значимых и авторитетных

работах в предметной области [4]. Помимо этого, существует достаточно мало методов автоматизации, призванных подходить к отбору наилучших источников литературы более системно, дабы не упустить наиболее важные поставленные и проверенные исследовательские гипотезы, используемые методы и полученные результаты.

Таким образом, *целью* данного исследования является разработка метода и программного комплекса, позволяющая автоматизировать процесс выбора релевантной научной литературы для последующего обзора некоторых ключевых работ авторов, специализирующихся на изучении влияния новостей на фондовый рынок, а также смежных с данной тематикой сфер, таких, как: обработка естественного языка (NLP) и эконометрические модели временных рядов семейства GARCH с помощью методов теории графов.

Исследование релевантной научной литературы подразумевает чтение большого количества научных исследований [4]. Для полного видения картины необходима систематизация всех изученных подходов. Субъективные критерии выбора литературы, не вполне соответствуют строгому научному подходу. Более оптимальным из возможных решений оказалось составление графа взаимных цитирований авторов, которые так или иначе имели отношение к изучаемой области. Граф, в отличие от линейного списка, позволяет увидеть наиболее успешные идеи (области наибольшей плотности цитирований, кластеры), а также проследить дальнейшие направления для их изучения [5-7]. Более того, с помощью графа цитирования можно предсказать некоторые тенденции развития области (используя динамику количества публикаций).

Метод оптимального выбора научных исследований в области анализа новостей финансовой отрасли.

Метод автоматизированного оптимального выбора научных исследований в области анализа новостей финансовой отрасли основан на теории графов. В разработанном методе предлагается следующий граф: множества (V, E) – где V (анг. vertex - вершина) – множество вершин, а E (анг. edge – ребро) – множество упорядоченных пар вершин, обозначающих связи между ними (множество ребер).

В настоящей работе принимаются следующие условия. Граф рассматривается и как ориентированный (для анализа цитирования ключевых источников), и как неориентированный (для изучения основных показателей графа). Все ребра имеют равный вес. Исключены петли и множественные связи. Петли (цитирование автором самого себя - самоцитирования) происходили лишь в случае цитирования группы соавторов, в составе которых находился цитирующий. Группы соавторов воспринимались, как совокупности участников. Например, если группа {I, II} процитировали группу {III, IV, V}, в графе добавлялось 5 вершин и 6 ребер (в случае если никаких из этих ребер и вершин ранее не встречалось). Авторы кодировались порядковым номером, вершины нумеровались.

В качестве отправной точки построения графа является поиск в Google Scholar с использованием поиска по ключевым словам. Вершинами графа являются авторы, ребрами – цитирования. Автор включался в выборку только в том случае, если в результате построений можно было прийти к выводу о том, что его работы привносят в изучаемую область некоторую новизну, а общие идеи связаны с моделированием влияния новостей на финансовый рынок.

Модель графа.

Математическая модель графа цитирований используется для анализа и визуализации связей между научными работами. Она позволяет исследователям изучать структуру цитирований, выявлять ключевые работы, определять влиятельных авторов и тенденции развития научных областей. Такая модель помогает улучшить понимание научной деятельности и её взаимосвязей. Суть исследования состоит в построении матрицы графа на основе реальных данных о взаимных цитированиях авторов с последующим ее анализом и получением полезной информации для автоматизации научного исследования.

Матрица смежности графа цитирований может быть записана в виде:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \quad (1)$$

Для неориентированного графа:

$$a_{ij} = \begin{cases} 1, \text{если существует цитата между авторами } i, j \\ 0, \text{иначе} \end{cases} \quad (2)$$

В случае с ориентированным графом:

$$a_{ij} = \begin{cases} 1, \text{если } i \text{ цитирует } j \\ 0, \text{иначе} \end{cases} \quad (3)$$

Дополнительно исключаются самоцитирования (петли графа):

$$a_{ii} = 0 \ \ \forall i = 1, \dots, n \qquad (4)$$

Учитывая контекст исследования, исключение самоцитирований в графе имеет несколько преимуществ: упрощение анализа, уменьшение избыточности, чистота данных (самоцитирования могут привести к искажению результатов анализа или интерпретации данных).

Основные характеристики

- Число цитирований:

$$CI = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \qquad (5)$$

- Плотность:

$$D = \frac{CI}{n^2} \qquad (6)$$

Смысловая интерпретация плотности графа цитирований заключается в интенсивности связей: чем выше плотность графа цитирований, тем интенсивнее связи между узлами. Также плотность указывает на влиятельность или активность: плотный граф цитирований может свидетельствовать о более активной или влиятельной сети узлов авторов или публикаций, что может быть важным при оценке значимости конкретных элементов графа. Кроме того, плотность указывает на потенциальную сложность: плотные графы цитирований могут быть более сложными для

анализа из-за большего количества связей, которые необходимо учитывать при изучении структуры и взаимосвязей в графе.

Разработка программного комплекса и построение графа цитирований

Рис. 1. – Алгоритм последовательности выбора ключевых источников

Алгоритм последовательности действий при разработке программного комплекса и построения графа цитирований представлен на схеме (1). Первым этапом определяется ключевые слова и словосочетания, по которым будет проводиться поиск. В данном случае это “**news in finance**”. На втором этапе происходит сбор цитирований по самым цитируемым источникам в данной области. Последующие манипуляции с данными, включая их

предобработку, происходят на языке программирования Python и фреймворков pandas, numpy и igraph. В результате анализа и валидации теоретической модели графа цитирований на реальных данных, который изображен на (2), выбираются наиболее валидные источники научной литературы (их авторы представлены в таблице 2), которые и будут служить отправной точкой нового исследования.

Ориентированная версия этого же графа представлена на (3). Всего удалось собрать данные о 83 публикациях и 86 авторах, а также 161 взаимное цитирование, что обеспечивает графу плотность порядка 4.7%. В среднем у каждого автора 3.88 цитирований, максимальное число цитирований составляет 34 и принадлежит Tetlock-у (индекс 0). Мода цитирований у авторов равняется 2. Число авторов с количеством цитирований, равным моде, составляет 32. Иллюстрации графов цитирования представлены на (2) и (3), соответственно.

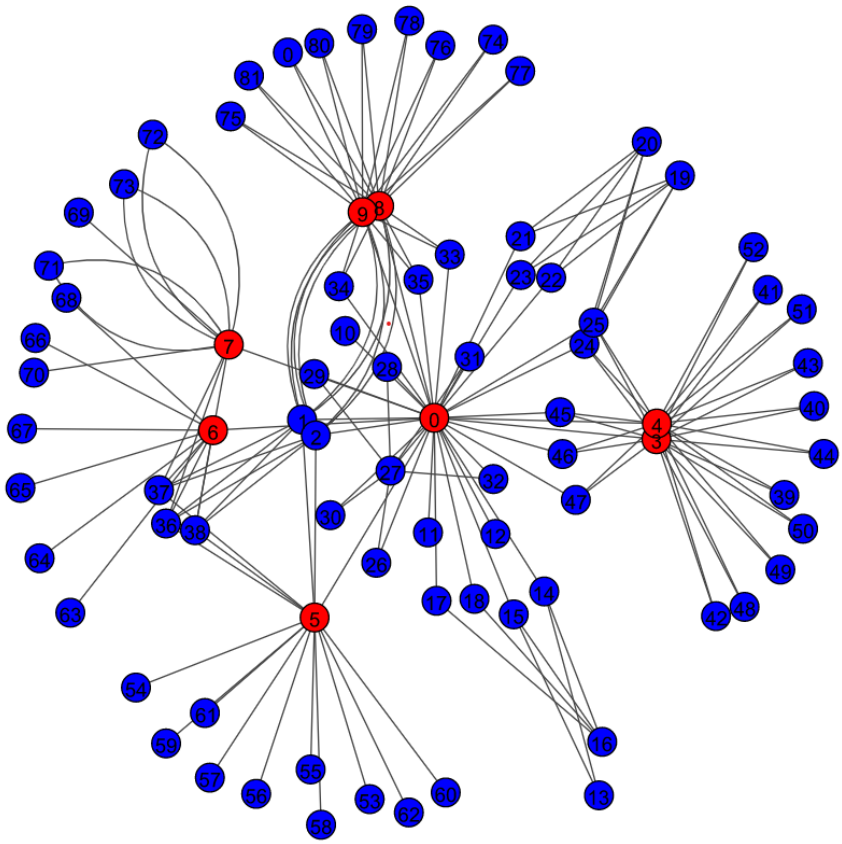

Рис. 2. – Неориентированный граф цитирований

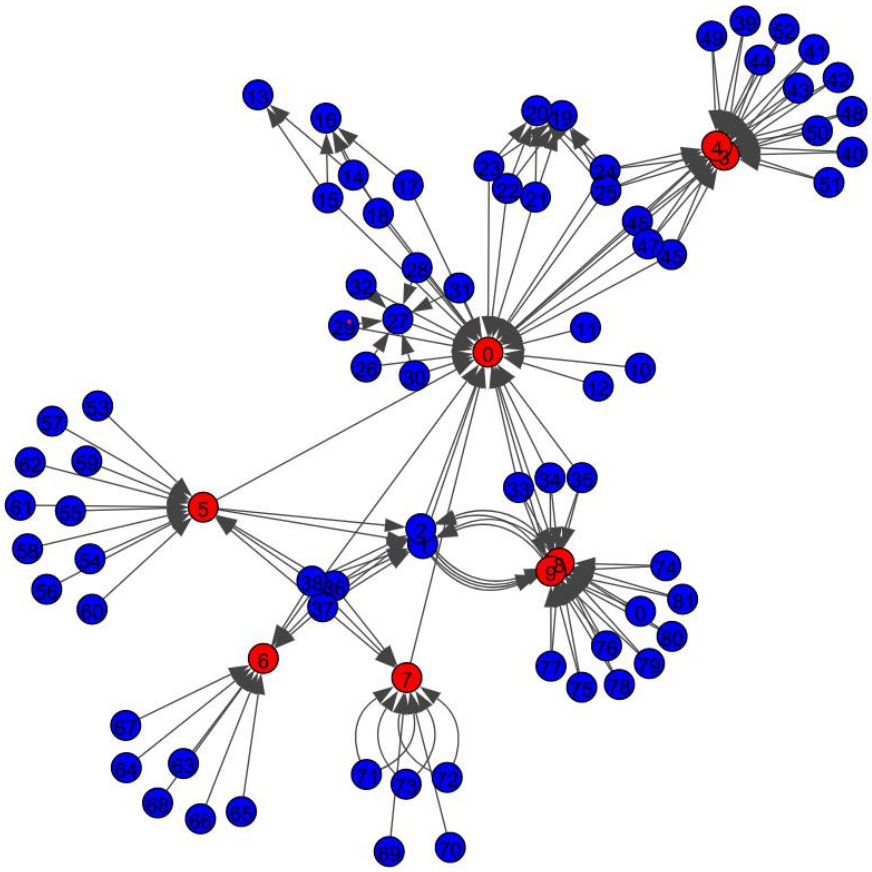

Рис. 3. – Ориентированный граф цитирований

Для систематизации вышеизложенного составлена сводная таблица 1 по проведенному подсчету характеристик графа.

Таблица № 1

Сводная таблица основных характеристик графа

Характеристика	Значение характеристики
Число вершин (публикаций)	83
Число ребер (цитирований)	161
Плотность графа	0.047
Средняя степень	3.88
Максимальная степень	34
Мода степени	2
Число вершин, имеющих степень 2	32

Расшифровка вершин графа представлена в таблице:

Таблица № 2

№	Автор
0	P.Tetlock
1	T Loughran
2	B McDonald
3	Mitra L
4	Mitra G
5	Dzielinski M.
6	Rieger M
7	Talpsepp T
8	Ahern K. R.
9	Sosyura D

Из плотности распределения явным образом отображается мода. Распределение имеет форму ассиметричного Хи-квадрат распределения, и это говорит о том, что имеется возможность построения вероятностной модели с помощью оценки параметров плотности распределения графа:

$$p = p(x, \theta)$$

Подобные графы (их структура) и их распределения сильно зависят от природы данных, на которых они построены. Вводятся и другие характеристики, которые имеют некоторые эмпирические правила. Так, например, коэффициент ассортативности, который по модулю не больше 1, для социальных процессов ближе к 1, а для графа, построенного на данных, связанных с биологическими или химическими процессами, ближе к -1. У данного же графа коэффициент ассортативности равен -0.4186, что ближе ко второму варианту.

Из рисунка 2 видно, что есть вершины (отмечены красным), которые являются центральными (имеют наибольшие степени). Данным вершинам соответствуют авторы работ [1-3, 5-6].

На основании построения графа и полученных результатов изучения взаимной цитируемости научных публикаций в области изучения влияния новостных потоков и СМИ на финансовые (в частности, фондовые) рынки, перейдем к анализу "центральных" авторов и их работ, имеющих наибольшее количество цитирований.

Рис. 3. – Распределение степеней графа

На графике распределения степеней графа по оси абцисс лежит характеристика вершины – степень, а по оси ординат – число вершин, имеющих данную степень. На рисунке (4) видно, что точкой максимума плотности эмпирического распределения является степень, равная двум, далее есть тяжелый хвост, о чем свидетельствует смещение среднего вправо до 3.88. Мода равна медиане. По графику эмпирической плотности также наблюдаются тяжелые выбросы, соответствующие степеням 5 и 17. Если на большой выборке такая тенденция сохранится, то распределение не будет являться унимодальным. Кроме этого, есть выброс со степенью 34, который говорит о том, что есть небольшое количество наиболее цитируемых, и, вероятно, фундаментальных в своей области работ. В целом, форма распределения может быть откалибрована под некоторую теоретическую модель. Например, хи-квадрат или степенное распределение. Говоря о том, как найти самые цитируемые работы, нужно идти справа налево и выбирать

вершины, которым соответствует наибольшая степень. Это служит основным принципом подбора научной литературы, который и использовался в настоящей работе (см. табл. 2).

Заключение

В данной статье была представлена разработанная математическая модель и программный комплекс, предназначенные для оптимального выбора научных исследований в области анализа новостей финансовой отрасли. Результаты показали, что использование данного подхода позволяет эффективно выявлять ключевые гипотезы, методы и тренды в научной литературе, а также проводить более точный анализ влияния новостей на финансовый рынок. Настоящее исследование имеет потенциал значительно улучшить процессы принятия решений в инвестиционной сфере и повысить эффективность стратегий управления портфелем. Дальнейшие исследования и развитие данной модели и программного комплекса могут привести к новым инновационным методам анализа финансовых данных, способствуя улучшению прогнозирования и оптимизации инвестиционных решений. Дополнительно, данная модель и программный комплекс могут быть адаптированы для других областей научного знания, где анализ новостей играет ключевую роль, таких, как маркетинг, политика или научные исследования. Также важно отметить, что разработка данной математической модели подчеркивает значимость взаимодействия между финансовыми знаниями и современными методами анализа данных, что открывает новые перспективы для дальнейших исследований в этой области. В целом, данная работа представляет собой важный шаг к более глубокому пониманию влияния новостей на финансовую отрасль и может послужить основой для развития инновационных подходов к анализу данных в будущем.

Литература

1. Tetlock P. Giving Content To Investor Sentiment: The Role Of Media In The Stock Market // The Journal Of Finance. 2007. №LXII. pp. 1139-1168.
2. Mitra L., Mitra G. Applications of news analytics in finance: A review // The Handbook of News analytics in finance. 2011. pp. 1-36.
3. Ahern K. R., Sosyura D. Who Writes the News? Corporate Press Releases during Merger Negotiations // The Journal of Finance Vol. LXIX, No. 1. - 2014. – pp.241-291.
4. Jackson M. O. Social and Economic Networks // Princeton University Press, 2008. 605 P.
5. Dzielinski M., Rieger M., Talpsepp T. Volatility asymmetry, news, and private investors // The Handbook of News analytics in finance. – 2011. – pp. 255-269.
6. Ager Hafez, How news events impact market sentiment // The Handbook of News analytics in finance. – 2011. – pp. 129-145.
7. Красий Н.П. О вычислении спрэда для обобщённой модели (B, S) рынка в случае скупки акций // Инженерный вестник Дона, 2012, №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2114
8. Назарько О.В., Павлов И.В., Чернов А.В. Моделирование оптимальной полосы пропускания телекоммуникационных каналов при условии гарантированной и негарантированной доставки пакетов // Инженерный дестник Дона, 2012, №1. URL: ivdon.ru/ru/magazine/archive/n1y2012/652
9. Сидоров С., Дате П., Балаш В. Использование данных новостной аналитики в GARCH моделях // Прикладная эконометрика №29(1). - 2013. – стр. 82-96.
10. Назарько О.В., Павлов И.В., Моделирование сбоев и их устранение на финансовых рынках с потоком событий, порожденным бинарным

деревом // Инженерный дестник Дона, 2013, №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2163

11. Аганин А.Д. Волатильность российского фондового индекса: нефть и санкции // Вопросы экономики. 2020. № 2. С. 86–100.
12. Гаврилов В., Иванов М., Клачкова О., Королев В., Рощина Я. Влияние тематических новостных потоков на компоненты волатильности фондового рынка России // Вестник Института экономики Российской академии наук. 2022. №2. С. 93 - 111
13. Гимранов Р.Д., Тищенко С.А., Шахмурадян М.А., Вакорин П.О., Выслоух А.А., Коматовский М.О. Граф цитирований как инструмент методологии исследования научной литературы по онтологии бизнес-процессов предприятия // Вестник московского университета. 2019. №6. с. 99-110.

References

1. Tetlock P. The Journal of Finance. 2007. №LXII. pp. 1139-1168.
2. Mitra L., Mitra G. Applications of news analytics in finance: A review. The Handbook of News analytics in finance. 2011. pp. 1-36.
3. Ahern K. R., Sosyura D. The Journal of Finance Vol. LXIX, 2014. No. 1. pp.241-291.
4. Jackson M. O. Social and Economic Networks. Princeton University Press, 2008. 605 P.
5. Dzielinski M., Rieger M., Talpsepp T. The Handbook of News analytics in finance, 2011. pp. 255-269.
6. Ager Hafez The Handbook of News analytics in finance, 2011. pp. 129-145.
7. Krasij N.P. Inzhenernyj vestnik Dona, 2012, №4 (chast' 2). URL: ivdon.ru/ru/magazine/archive/n4y2013/2114

8. Nazar'ko O.V., Pavlov I.V., CHernov A.V. Inzhenernyj vestnik Dona, 2012, №1. URL: ivdon.ru/ru/magazine/archive/n1y2012/652
9. Sidorov S., Date P., Balash V. Prikladnaya ekonometrika. 2013. №29 (1), pp.82-96.
10. Nazar'ko O.V., Pavlov I.V. Inzhenernyj Vestnik Dona, 2013, №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2163
11. Aganin A.D. Voprosy ekonomiki. 2020, № 2. pp. 86–100.
12. Gavrilov V., Ivanov M., Klachkova O., Korolev V., Roshchina YA. Vestnik Instituta ekonomiki Rossijskoj akademii nauk, 2022, №2. pp. 93-111.
13. Gimranov R.D., Tishchenko S.A., SHahmuradyan M.A., Vakorin P.O., Vyslouh A.A., Komatovskij M.O. Vestnik moskovskogo universiteta, 2019, №6. pp. 99-110.

Дата поступления: 12.01.2024
Дата публикации: 18.02.2024