Санкт-Петербургский Государственный Университет
Математико-механический факультет

Кафедра информатики

Зернов Алексей Викторович

Разработка системы автоматического анализа новостных публикаций на финансовом рынке

Бакалаврская работа

Научный руководитель:
к. ф.-м. н., доцент Григорьев Д. А.

Рецензент:
д. т. н., декан Мусаев А. А.

Санкт-Петербург
2017

SAINT-PETERSBURG STATE UNIVERSITY
Faculty of Mathematics and Mechanics

Computer Science Department

Zernov Alexey Viktorovich

Development of the automatic analysis system of a financial market's news publications

Bachelor's Thesis

Scientific supervisor:
Sc. C., associate professor Grigoryev D. A.

Reviewer:
Sc. D., dean Musaev A. A.

Saint-Petersburg
2017

Оглавление

Введение

1. Финансовый рынок

В данном разделе будет представлен краткий обзор основных терминов, связанных с самим финансовым рынком, его структурой и основными участниками. Более подробная информация может быть получена в [3].

1.1. Определение

В более общем виде **финансовый рынок** — совокупность экономических связей его участников, касающихся создания, поддержания и обращения капитала. Финансовый рынок является довольно абстрактным термином, и под ним часто подразумеваются более конкретные: рынок купонных и бескупонных облигаций, рынок акций (или фондовый рынок) или валютный рынок. Не смотря на выделение составляющих, каждая из них является частью единого механизма, в котором финансы перемещаются между каждым из конкретных рынков.

Каждый из финансовых рынков является рынком посредников между начальными владельцами финансов и их конечными пользователями. Если рынок основывается на финансах как на капитале, он называется фондовым рынком, и именно в этой роли выступает как составная часть всего финансового рынка.

В России финансовые рынки имеют следующие критерии, влияющие на их деятельность:

- Инвестиции в экономику страны
- Международные рынки, влиние тенденций глобализации
- Современные компьютерные технологии
- Уровень комьютерной и информационной развитости участников рынков

1.2. Структура

Финансовый рынок может быть:

- Первичным или вторичным
- Организованным или неорганизованным
- Биржевым или внебиржевым
- Традиционным или компьютеризированным
- Кассовым или срочным

Первичный рынок обеспечивает выход ценных бумаг в оборот, это своеобразное «производство» ценных бумаг. На **вторичном рынке** в обороте находятся уже выпущенные ранее ценные бумаги. Вторичный рынок представляет из себя совокупность всех операций с данными ценными бумагами, в результате которых они переходят от одних владельцев к другим.

Организованный рынок отличается от **неорганизованного рынка** тем, что в первом имеются единые для всех участников рынка правила, за соблюдением которых следят организаторы. Во неорганизованном рынке соблюдение единых правил для всех участников рынка не гарантируется.

Биржевой рынок — такой рынок, на котором в качестве инструмента торговли используется аукцион. Руководителем же является некоторый специалист, например, NYSE[1] или AMEX[2]. На **внебиржевых рынках** торги организуются при помощи электронных систем.

Срочный рынок чаще всего подразумевает отложенное исполнение сделки, в отличие от **кассового рынка**, когда сделки исполняются сразу. Обычно традиционные ценные бумаги (акции, облигации) идут в оборот на кассовом рынках, а контракты на производные инструменты рынка ценных бумаг — на срочных.

[1]New York Stock Exchange — Нью-Йоркская фондовая биржа

[2]American Stock Exchange - Американская фондовая биржа

1.3. Участники

Участники рынка ценных бумаг — это физические лица или компании, которые продают или приобретают ценные бумаги, обеспечивают их оборот или расчеты по ним.

Основными участниками рынка выступают **эмитенты**, выпускающие акции или облигации, с помощью которых привлекают финансирование, а также размещающие свободные на данный момент денежные средства. Эмитентами могут быть государство, субъекты государства или коммерческие предприятия. Целью эмитентов на первичном рынке является размещение запланированного транша по максимальной цене.

Инвестор — лицо, заинтересованное во вложении капитала в ценные бумаги. Их целью является как можно более выгодная покупка ценных бумаг максимально перспективных компаний.

2. Интеллектуальный анализ текста

В настоящее время можно заметить увеличение роли компьютеров в жизни каждого человека. Инорфмация хранится преимущественно в цифровом виде, что значительно упрощает поиск или работу с ней. Но не смотря на это, многие данные все равно остаются довольно трудными для анализа, не смотря на оцифрованный вид, из-за чего можно подразделить из на следующие формы:

- Структурированные данные
- Частично структурированные данные
- Неструктурированные данные

Хорошим примером **структурированных данных** могут являться базы данных. **Частично структурированные данные** — это электронные письма, разнообразные файлы на языках разметок (HTML, XML и другие).

Если работа со структурированными или частично структурированными данными достаточно детерминированная, то **неструктурированные данные** представляют наибольший интерес в этом вопросе. Около 80% корпоративных данных находится именно в неструктурированном формате, в котором сложно проводить поиск или извлекать необходимую информацию. Для этого нужны специфические методы и алгоритмы обработки. И поскольку самая популярная форма хранения информации — это текст, интеллектуальный анализ текста (text mining) является более важным процессом, нежели интеллектуальный анализ данных (data mining).

Интеллектуальный анализ текста стоит на пересечении дисциплин, включая в себя: обработку web-данных, информационный поиск, компьютерную лингвистику и обработку естественного языка.

2.1. Процесс интеллектуального анализа текста

Концепция интеллектуального анализа текста представлена в [2]. В интеллектуальном анализе текста можно выделить два основных этапа (Рис. 1):

- Фильтрация текста
- Извлечение знаний

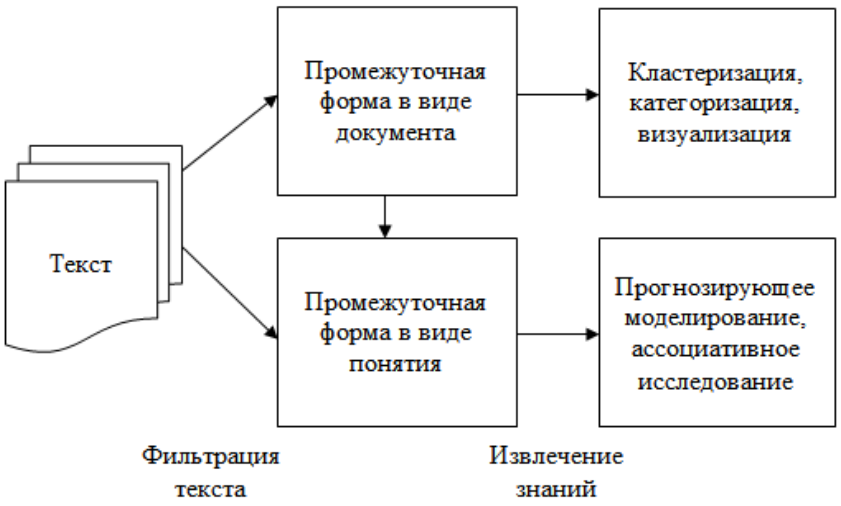

Рис. 1: Общий процесс интеллектуального анализа текста

Фильтрация (или очистка) преобразует исходный текстовый документ в некоторое промежуточное представление. **Извлечение знаний**, в свою очередь, получает знания или некоторые шаблоны из уже промежуточного представления. Промежуточное представление может быть как структурированным, так и частично структурированным. Промежуточное представление может быть как новым текстовым документом, так понятием, в котором составляющие являются данными или наборами данных из какой-либо предметной области.

Анализ промежуточного представления документов выдает образцы и связи между всеми документами. Примеры задач: кластеризация, визуализация и категоризация документов.

Анализ промежуточного представления понятий выдает образцы и связи между объектами или другими понятиями. Примеры задач: прогнозирующее моделирование и ассоциативное исследование.

Промежуточное представление в виде документа может быть преобразовано в промежуточное представление в виде понятия путем выделения релевантной информации, которая отношесится к необходимым объектам из какой-либо предметной области. Отсюда вытекает то, что промежуточное представление чаще не зависит от конкретное предметной области. К примеру, новостные потоки при фильтрации текста преобразуются в промежуточные представления в виде документов, соответствующим определенным статьям. Затем, в зависимости от поставленных задач визуализации или навигации, каждый документ (статья) проходит обработку знаний. Для извлечения же знаний в опрделеннной предметной области промежуточное представление в виде документа может быть преобразовано в промежуточное представление в виде понятия в соответствии с необходимыми требованиями. К примеру, можно извлечь информацию, касающуюся определенного товара или услуги из промежуточного представления в виде документа и сформировать базу данных товаров или услуг для предоставления знаний о них.

Ниже будут рассмотрены шаги, выполняемые при интеллектуальном анализе текста.

2.1.1. Предварительная обработка текста

Предварительная обработка включает в себя:

1. Токенизацию
2. Удаление «стоп-слов»
3. Определение происхождения слов

Токенизация. Сначала текст разделяется на отдельные слова, освобождаясь от пробелов и знаков препинания.

Удаление «стоп-слов». На этом этапе происходит избавление от «ненужных» конструкций текста. Это могут быть HTML или XML теги, предлоги, артикли и прочее.

Определение происхождения слов представляет из себя выявление корней определенных слов. Преимущественно встречаются два типа происхождения: флективный и деривационный.

2.1.2. Преобразование текста

Текстовый документ состоит из слов и информации об их происхождении. Два основных подхода представления документа: «мешок слов» («bag-of-words») и векторные пространства слов.

2.1.3. Поиск признаков

Под признаками можно понимать переменные. То есть в результате этого шага отбирается подмножество наиболее значимых признаков для их дальнейшего применения при построении моделей. Убираются, например, признаки, которые избыточны или не несут никакой информации.

2.1.4. Методы анализа текста

На данном шаге начинается построение модели с использованием разных методов, таких как кластеризация, классификация, информационный поиск и других. Данные методы распознавания данных также подходят и для интеллектуального анализа текста.

2.1.5. Интерпретация и оценка

На последнем шаге (в зависимости от того, что требуется) проводится анализ результатов.

2.2. Области применения интеллектуального анализа текста

Как уже упоминалось выше, интеллектуальный анализ текста стоит на пересечении разных дисциплин, включая в себя: извлечение информации, информационный поиск, обработку естественного языка и

интеллектуальный анализ данных. Ниже будет подробнее рассказано о каждой из областей.

2.2.1. Извлечение информации

В процессе извлечения информации автоматически извлекается структурированная информация из неструктурированных данных. С помощью распознавания образов данная система определяет, например, где имена людей, где названия компаний, а где местоположение. То есть в документах происходит поиск предопределенных последовательностей. Подобное решение позволяет получить элементы, подходящие для использования в базах данных для дальнейшего хранения, анализа или обработки.

2.2.2. Информационный поиск

Наиболее известными системами информационного поиска являются поисковые системы Google. В данной задаче используются методы, используемые для хранения, представления и доступа к информации, которая преимущественно представлена в виде текстовых документов (а также новостных лент или книг), которые могут быть получены по запросу пользователя. Это своего рода расширение поиска по документам, позволяющее сужать набор документов, имеющих отношение к запросу пользователя. Эти системы значительно сокращают время, необходимое для поиска необходимой информации.

2.2.3. Обработка естественного языка

Данная задача представляет из себя самую активную проблему в области искусственного интеллекта. Цель: исследовать естественный язык так, чтобы у компьютеров была возможность понимать языки, подобные тем, что используют для общения люди. Обработка естественного языка включает в себя распознавание и генерацию, которые отвечают за такие способности компьютера как «читать» и «говорить» на

естественном языке соответственно. Подобные системы включают в себя проверку грамматики, лексические, синтаксические и семантические анализаторы.

2.2.4. Интеллектуальный анализ данных

Данные задачи относятся к поиску знаний или релевантной информации в большом объеме данных. Система пытается обраружить правила (статистически) и образцы (автоматически) от данных. Подобные системы имеют возможность предсказания, основываясь на «опыте», полученном в результате исследования.

3. Обзор существующих инструментов

В данном разделе будут рассмотрены основные инструменты, представленные в виде библиотек или отдельных сервисов. Внимание уделено в основном инструментам, работающим с русским языком.

3.1. Natural Language Toolkit

NTLK[3] является пакетом библиотек и программ для разработки программ на Python, работающих с естественным языком. Сопровождается обширной документацией, а также книгой[4], объясняющей основные концепции проблем, для решения которых предназначен данный пакет. NTLK — свободное программное обеспечение, то есть доступное бесплатно.

Данный пакет подходит для таких областей как компьютерная лингвистика, эмпирическая лингвистика, когнитивистика, искусственный интеллект, информационный поиск и машинное обучение. NTLK используется преимущественно в качестве учебного пособия, индивидуального обучения или прототипирования и создания систем, ориентированных на научно-исследовательскую деятельность.

Изначально пакет предназначен для англоязычных текстов, но имеется возможность обучения классификаторов для остальных языков.

3.2. pymorphy2

Pymorphy2[5][1] написан на языке Python и имеет следующие возможности:

- Приведение слова к нормальной форме
- Ставить слово в нужную форму
- Возвращать грамматическую информацию о слове

[3] http://www.nltk.org

[4] http://www.nltk.org/book/

[5] https://pymorphy2.readthedocs.io/en/latest/index.html

Распространяется pymorphy2 под лицензией MIT[6], если используется в научной работе.

3.3. Томита-парсер

Томита-парсер[7] способен извлекать структурированные данные из текстов на естественном языке. Как и почти во всех инструментах, рассматриваемых в данном разделе, Томита-парсер ориентирован преимущественно на русскоязычные тексты. В нем используются контекстно-свободные грамматики и словари ключевых слов. Код проекта[8] находится в свободном доступе.

3.4. Яндекс.Спеллер

Яндекс.Спеллер[9] выполняет задачу проверки орфографии в текстах на английском, русском и украинском языках. Для этого используется орфографический словарь. К тому же, предоставлен набор API методов для реализации данной проверки разработчиками сайтов или приложений.

3.5. OntosMiner

OntosMiner[10] является решением компании Eventos[11], занимающейся в большей степени разработкой продуктов в области лингвистического анализа текстовой информации, кластеризацией и классификацией информации. Конкретно OntosMiner является целой комплексной системой, дающей возможность распозавания связей между сущностями в текстах на естественной языке. Также, она позволяет определеять общую тональность текста.

[6] https://opensource.org/licenses/MIT
[7] https://tech.yandex.ru/tomita/
[8] https://github.com/yandex/tomita-parser/
[9] https://tech.yandex.ru/speller/
[10] http://my-eventos.com/solution/ontosminer/
[11] http://my-eventos.com/solution/ontosminer/

4. Теоретическая информация

5. Практическая часть

Заключение

Список литературы

[1] Korobov Mikhail. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts / Ed. by Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko et al. — Springer International Publishing, 2015. — Vol. 542 of Communications in Computer and Information Science. — P. 320–332. — URL: http://dx.doi.org/10.1007/978-3-319-26123-2_31.

[2] Sumathy K. L., Chidambaram M. Article: Text Mining: Concepts, Applications, Tools and Issues — An Overview // International Journal of Computer Applications. — 2013. — October. — Vol. 80, no. 4. — P. 29–32. — Full text available.

[3] V.P. Romanov. Information technology modeling of financial markets - (Applied Information Technology) / Informatsionnye tekhnologii modelirovaniya finansovykh rynkov - ("Prikladnye informatsionnye tekhnologii"). — Finansy i statistika, 2010. — ISBN: 5279034444. — URL: https://www.amazon.com/Information-technology-modeling-financial-markets/dp/5279034444?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=5279034444.