Анализ неструктурированных данных с целью получения дополнительной информации при оценке кредитоспособности юридических лиц

Е.А. Макарова
Факультет информационных технологий
Брянский государственный технический университет
Брянск, Россия
e-mail: lennymbear@gmail.com

Д.Г. Лагерев
Факультет информационных технологий
Брянский государственный технический университет
Брянск, Россия
e-mail: LagerevDG@yandex.ru

Аннотация

В данной статье рассмотрены различные методики оценки кредитоспособности юридических лиц. Проанализированы достоинства и недостатки различных методик оценки кредитоспособности, используемых для оценки физических и юридических лиц. Сделано предположение о возможности разрешения спорных ситуаций, возникающих при использовании традиционных подходов, с помощью дополнения модели оценки методами анализа неструктурированных данных из открытых источников. Проведен анализ существующих методик обработки текстов на естественном языке, которые можно использовать для решения задачи оценки репутации и кредитоспособности. На основе проведенного исследования авторами предлагается архитектура программного комплекса для поддержки принятия управленческих решений в сфере работы с юридическими лицами.

1. Введение

В последние годы в Российской Федерации наблюдается интенсивный рост рынка кредитования и, в частности, сектора кредитования юридических лиц. Помимо увеличения темпов развития малого и среднего бизнеса, это так же приводит к увеличению кредитных рисков, которые принимает на себя банковская система страны, в целом, и отдельные банки, в частности.

В 2017 году Цетнробанк РФ начал реформы, призванные поддержать и упорядочить деятельность лизинговых компаний в стране.

Однако, рынок лизинга развивается и без дополнительных мер регулирования: в 2016 году стоимость переданных в лизинг активов выросла на 36% и достигла 742 млрд руб., что, по оценкам участников рынка, составляет уже 0,86% ВВП России. [13]

2. Анализ проблемы оценки кредитоспособности

В вопросах оценки кредитоспособности и репутации определенного юридического или физического лица обычно применяются следующие методы, использующие в своей основе математический аппарат:

1) Анализ кредитной истории
2) Оценка кредитоспособности заемщика по уровню финансового состояния
3) Скоринговые модели

Рассмотрим внимательнее третий пункт.

Сущность скоринга заключается в определении совокупного кредитного балла заемщика в результате его оценки по ряду критериев. Критерии имеют различные удельные веса и впоследствии агрегируются в интегральный показатель – совокупный кредитный балл. Интегральный показатель сравнивается с определенным числовым порогом, который представляет собой «линию безубыточности» для учреждения. Одобрение кредита получают заявители, интегральный показатель которых выше этой линии. Величина одобренного кредитного лимита в подобных системах чаще всего определяется, исходя из уровня доходов заемщика. [2].

При этом скоринг в лизинге решает не только задачу оценки интегрального показателя для сравнения с линией безубыточности, но и учитывает стоимость передаваемого по лизингу имущества и взаимоотношения с его владельцем на протяжении срока договора. Также важна оценка разности между планируемым снижением

Труды Шестой всероссийской научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 28-31 мая, Уфа-Ставрополь, Россия, 2018

стоимости имущества и предполагаемой стоимостью его продажи в течение всего периода действия договора лизинга. Скоринг в лизинге должен учитывать не только вероятность возникновения проблем с лизинговыми платежами (просрочки или прекращения уплаты), но и ожидаемую стоимость перепродажи возвращенного имущества (прибыль или убыток от реализации актива). Так же стоит отдельно обрабатывать ситуации, связанные с факторами задержки платежей по причинам, не зависящим от текущего пользователя имущества, но влияющими на всю сферу его деятельности. Например, учитывать засушливые периоды при рассмотрении процессов лизинга крупной сельскохозяйственной техники и т.д. Данные ситуации могут рассматриваться для корректировки интегрального показателя с точки зрения оценки добросовестности при рассмотрении кредитной истории.

Одной из проблем большинства скоринговых моделей является плохая классификационная способность в области граничного балла. В данной области ошибка модели может достигать 50%. Поскольку в данной области сконцентрированы средние заемщики, качество модели становиться весьма важным, так как человеку весьма сложно самостоятельно классифицировать таких заемщиков. Значительно улучшить качество скоринговой модели может добавление дополнительной входной информации.

Источниками необходимой информации могут служить различные открытые базы нормативно-правовых актов, «черных списков», государственных реестров по различным отраслям и т.д.

Кроме того, объекты лизинговых операций могут быть возвращены в компанию, предоставившую их, и оценка рисков в данной ситуации становится процессом, непрерывным в течении всего времени, прописанного в договоре. Данное обстоятельство требует от компаний, предоставляющих объекты, наличия крупного штата финансовых аналитиков для проработки постоянно увеличивающихся объемов информации, их агрегации и принятия решений по прекращению/продлению договора. Одним из способов сокращения расходов на оценку рисков в данной ситуации является автоматизация части функционала аналитиков с помощью различных методик сбора и обработки информации с помощью программных средств.

Так же стоит рассматривать источники неструктурированной информации, такие как: различные Интернет-СМИ, социальные сети, открытые реестры документов и нормативно-правовых актов. Актуальность привлечения данных источников определяется высокими темпами роста объемов слабоструктурированной текстовой информации в свободном доступе сети Интернет. Однако, увеличивается сложность и трудоемкость их обработки в режиме реального времени.

По данным ФОМ, осенью 2016 года 67% совершеннолетнего населения страны пользовались интернетом хотя бы раз в месяц, а почти 60 миллионов человек выходили в сеть ежедневно. [3]

В России более 4,5 тысячи активных Интернет-СМИ. За год их число выросло во всех регионах страны. Почти четверть активных изданий расположены в Москве. [3]

Различные банковские и управляющие системы в данный момент активно внедряют интеллектуальные системы принятия решений и автоматизированной обработки информации. [14] Исследования в области автоматизированной обработки естественного языка в последнее время так же существенно продвинулись вперед.

В данной работе предлагается методика, основанная на дополнении традиционных скоринговых моделей результатами анализа неструктурированных данных. Методики обработки естественного языка, которые могут быть использованы для решения подобных задач, подробнее описаны в следующих пунктах.

3. Научно-технический задел, необходимый при решении данной задачи

Задача обработки структурированных массивов данных в последние годы широко прорабатывалось, применительно к различным сферам социально-экономической жизни. Однако, немногие российские производители интеллектуальных систем привлекают к анализу источники неструктурированных данных. В то же время, процессы, связанные с обработкой естественного языка, получили новый уровень развития, благодаря применению технологий глубокого обучения при решении различных задач компьютерной лингвистики [15].

Рассмотрим научно-технический задел, который позволит выполнить разработку готового программного продукта для применения различных методик обработки естественного языка. При обработке неструктурированной информации имеют основное значение два связных процесса:

1) Извлечение и анализ информации
2) Генерация связного текста на её основе

На этапе извлечения информации перед разработчиками стоит задача анализа разнородных источников данных, приведения их к унифицированному виду и предобработка для дальнейшего анализа. К рассмотрению предлагаются такие источники, как: реестры информации о юридических лицах (реквизиты, официальные документы, упоминания в различных базах данных, связанных с финансами, арбитражными делами и т.д.) и информация из

средств массовой информации, открыто размещенных в сети Интернет. Последние будут анализироваться не только с точки зрения получения информации и поиска связных организаций/физических лиц для дальнейшего анализа, но и с целью оценки репутации данной организации с точки зрения жителей региона, в котором она функционирует.

Для поиска релевантной информации необходимо воссоздать информационную модель, максимально близко описывающую информационные сущности и связи между ними, относящиеся к объекту анализа. Другими словами, создать приближенную онтологию.

Онтология – это концептуальная схема, отображающая некоторую область знаний. В данной задаче «точкой отсчета» для создания онтологии будет являться юридическое лицо, либо же сфера какой-либо деятельности в регионе, в целом (например, ООО «ТрансАгроЗапад» или «сельское хозяйство в Орловской области»).

На данный момент наиболее исследованной подзадачей является распознавание именованных сущностей (NER: Named Entity Recognition). Наиболее сложна и требует дальнейшего изучения подзадача выявления событий и связей между сущностями, она позволяет отвечать на вопросы о том, что произошло, кто это сделал, когда, где, как и почему [4]

Для создания схемы связей потребуется извлечь следующую информацию:

- ключевые понятия конкретного процесса/отрасли;
- связи между понятиями (входит в, используется в/с) и т.д.

Текущие наработки в области разпознавания сущностей и связей между ними:

1. Графематический анализатор, входящий в состав проекта АОТ [5], позволяет распознавать ФИО, целые числа, имена файлов и электронные адреса.
2. Язык LSPL и поддерживающий его программный комплекс, созданные специально для обработки русскоязычных текстов и учитывают его специфику. [6]
3. Модуль Stanford OpenIE служит для выявления отношений в неразмеченных текстах любых предметных областей. [7]
4. Инструменты от компании «Яндекс» - Томита-парсер и др.

4. Методики обработки естественного языка, имеющие применение при решении описываемой задачи.

Ниже представлен список некоторых широко исследуемых задач в области обработки естественного языка. У некоторых из этих задач есть прямые применения в вопросах принятия управленческих решений, связанных с областью кредитования и лизинга, в то время как другие обычно служат скорее подзадачами, которые используются, чтобы помочь в решении больших задач.

Многие из этих задач относятся к широкой задаче извлечения информации. Так же многие из них подойдут для анализа таких специфических источников информации, как Интернет-СМИ. Более того, анализ новостных текстов – одно из первых прикладных исследование в области извлечения информации. Данные исследования были проведены ещё в 80х годах в США для решения задачи поиска неочевидных связей и закономерностей среди событий, описываемых в англоязычной прессе. [7]

В современных условиях, когда объемы информации заметно возросли, для решения этой задачи требуются новые методы. Рассмотрим некоторые методы автоматизированной обработки естественного языка, которые в данных момент широко используются для различных прикладных задач.

1. Автоматическое резюмирование (аннотирование): Генерация удобочитаемого резюме (сокращения) текста. Часто используется, чтобы предоставить аннотацию текста известного типа, такие как статьи в финансовом разделе газеты, сводка новостей, биржевые данные и т.д.

2. Разрешение кореференции. Задача состоит в том, чтобы, учитывая предложение или больший кусок текста, определить, какие слова («упоминания») относятся к тем же самым объектам («предприятия, юр. лица»). Например, в предложении, таком как «Данная организация вложила крупную сумму в реконструкцию театра», «крупная сумма» является относящимся выражением, и отношения между сущностями, которые будут определены, являются фактом, что упоминаемая сумма относится к действиям, связанным с театром (а не некоторой другой структуры, которая могла бы также быть упомянута).

3. Анализ беседы: Эта рубрика включает много связанных задач. Одна задача определяет структуру беседы связанного текста, т.е. природу отношений беседы между предложениями (например, разработка, объяснение, контраст). Другая возможная задача признает и классифицирует речевые акты в куске текста (например, да - никакой вопрос, вопрос о содержании, заявление, утверждение, и т.д.). Так же эта задача часто тесно связана с задачей распознавания текста по аудио и видео-источникам.

В контексте данной задачи, анализ беседы может быть полезен при анализе таких видов публикаций, как стенография интервью. Кроме того, использовав соответствующие инструменты для анализа речи, возможно так же дополнение источников информации с помощью видео-интервью, размещенных в открытых источниках, репортажей и т.д. Однако, внедрение данного анализа потребует намного больших вычислительных мощностей и хранилищ данных, по сравнению с задачей анализа текстовой информации.

4. Поиск именованных сущностей: Задача состоит в том, чтобы, учитывая поток текста, определить, какие сущности в тексте имеют отношения к именам собственным, таким как люди или места, и что тип каждого такого имени (например, организация, банк, населенный пункт). Несмотря на то, что капитализация может помочь в признании названий предприятий на алфавитных языках, таких как английский, русский, немецкий, испанский, эта информация не может помочь в определении сущностей, связанных с предприятием, и часто капитализация бывает неточной или недостаточной. Например, первое слово предложения также использовано для определенной цели (обозначение красной строки), и название предприятия часто охватывают несколько слов, только некоторые из которых могут быть использованы ц целях задачи. Кроме того, у многих других языков в других языковых семьях (например, китайский, турецкий, арабский язык) капитализация отсутствует, и даже языки с капитализацией могут не всегда использовать её, чтобы отличить имена.

5.Тематическое моделирование – способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов.

Применение тематического моделирования позволяет решить такие дополнительные задачи, как:

1) ранжировать документы по степени релевантности заданной теме (другими словами, провести тематический поиск)
2) определить, как тематики новостных потоков изменялись со временем
3) определить тематику различных сущностей (ссылок), связанных с документами

Чаще всего в подходах машинного обучения в задаче классификации документов с несколькими темами (тегами) используются методы дискриминантного моделирования, таких как метод опорных векторов. Недостатком этих подходов является то, что производительность быстро снижается, так как общее количество тем и количество тем, охваченных в одном документе увеличиваются. Эта проблема хорошо заметна на текстах обиходных (заметки, газетные статьи и т.д.), которые планируется использовать в данном исследовании. Тимоти Н. Рубин предлагает для быстрой классификации подобных текстов использовать различные модели, построенные на основе метода латентного размещения Дирихле, чью эффективность демонстрируют исследования, представленные в его работах[9]. Согласно приведенным данным, данный метод дает большую производительность, по сравнению с остальными. В процессе решения данной задачи вполне закономерно столкнуться с увеличением количества входных параметров для поиска необходимых материалов (редкий контекст, конкретизация города, периода поиска информации и т.д.)

Метод латентного размещения Дирихле (latent Dirichlet allocation, LDA) предложен Дэвидом Блеем в 2003 году. В этом методе устранены основные недостатки PLSA – вероятностного латентного семантического анализа. Метод LDA основан на той же вероятностной модели, что и PLSA при некоторых дополнительных предположениях [16]. Для идентификации параметров модели по коллекции документов применяется сэмплирование Гиббса, вариационный байесовский вывод или метод Expectation-Propagation.

Схема тематического моделирования коллекции документов представлена на рис. 1.

6. Анализ тональностей: класс методов контент-анализа в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов (мнений) по отношению к объектам, речь о которых идёт в тексте. Тональность – это эмоциональное отношение автора некоторого высказывания (поста, текста, статьи) к некоторому объекту (в контексте данной работы – предприятия, юридического лица и т.д.), выраженное в тексте. Эмоциональная составляющая, выраженная на уровне лексемы или лексико-синтаксического шаблона, называется лексической тональностью (или лексическим сентиментом). Тональность всего текста часто определяется как функция (в самом примитивном варианте: сумму векторов) лексических тональностей составляющих его единиц (предложений) и правил их сочетания.

Результатом анализа тональности обычно является процентное соотношение сообщений с разным эмоциональным окрасом (см. рис. 2).

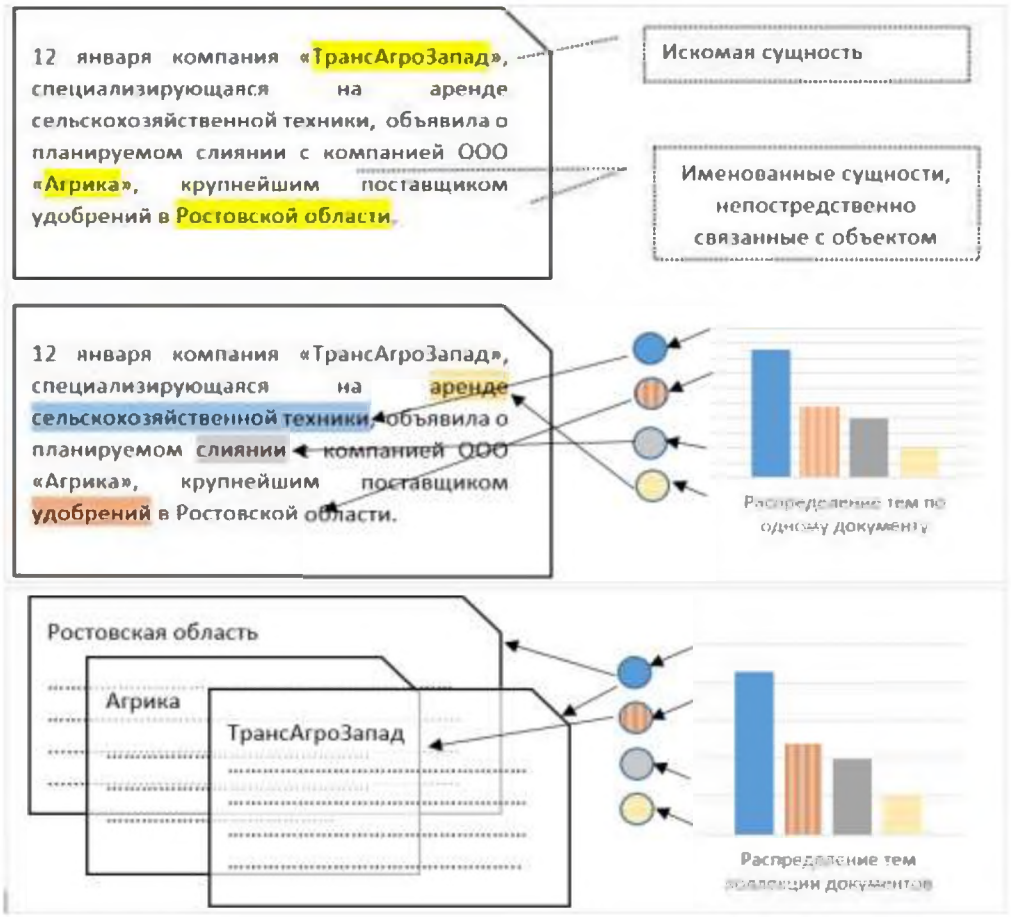

Рис. 1. Пример поиска связанных именованных сущностей и тематического анализа коллекции документов

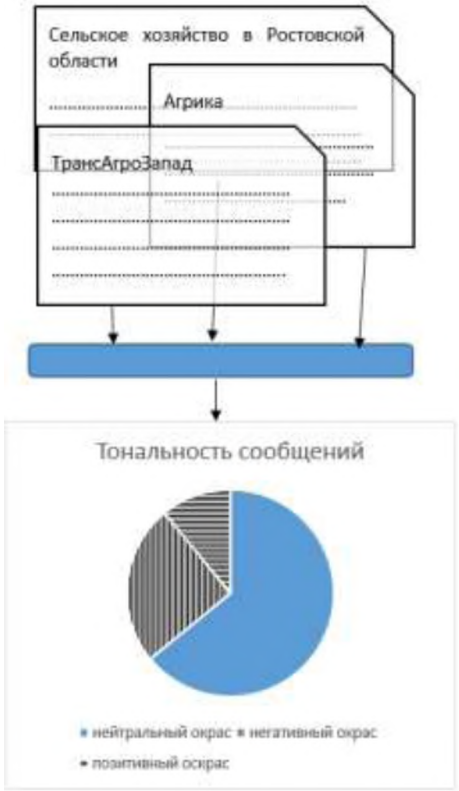

Рис. 2. Анализ тональностей коллекции документов

7. Анализ формальных понятий хорошо зарекомендовал себя в анализе неструктурированных данных. Например, для выявления (почти) дубликатов по большим коллекциям веб-документов [10, 11] и анализа текстов полицейских отчетов [12].

При решении данной задачи, анализ формальных понятий потребуется для исключения статей из Интернет-СМИ, которые являются перефразированными копиями других статей, опубликованных примерно в одно время по одной тематике.

Все вышеперечисленные методы могут найти хорошее применение для анализа открытых источников информации, таких как СМИ и социальные сети, при описании планируемой архитектуры программного комплекса будут рассмотрено, в каких сочетаниях и комбинациях их планируется опробовать для решения данной задачи.

При создании программного комплекса будут использоваться основные положения теории реляционных баз данных и методы объектно-ориентированного проектирования и программирования.

5. Архитектура разрабатываемого программного комплекса.

Архитектура программного продукта изображена на рис. 3. Предполагается, что в случае постоянного мониторинга (который, как было рассмотрено выше, необходим в случаях лизинга), инструменты для выгрузки и предобработки материалов должны работать в фоновом режиме с индивидуальным расписанием проверок обновлений по искомой информации, исходя из потребностей и наличия вычислительных ресурсов.

Инструменты для выгрузки и каталогизации необходимых материалов (новостных потоков) должны выполнять следующий функционал:

1) загрузка документов в базу данных;
2) предобработка документов;
3) мониторинг появления новых материалов по теме.

Кроме того, выгрузка документов не должна чрезмерно нагружать сервера новостных изданий во избежание блокировок с их стороны. При эксплуатации в реальных системах потребуется проанализировать наличие лицензий или авторских прав на новостной контент и, в случае запрета их использования в коммерческих целях, прийти к дополнительным соглашениям с владельцами информационных ресурсов.

Сервис анализа информации состоит из нескольких модулей, которые в режиме реального времени (при обновлении документов в базе), либо же по запросу конкретного отчета, проводят соответствующий анализ. Это может быть как классический скоринг, созданный на основе анкет, заполняемых аналитиками, так и различные методы, основанные на автоматизированной обработке данных и предоставлении отчетов по ним.

Рис 3. Архитектура продукта

Процесс оценки информации из структурированных источников планируется построить на взаимодействии с API официальных реестров и автоматизированной оценки содержащейся в них информации по ряду критериев для получения общей оценки благонадежности изучаемого объекта.

Процесс семантического анализа, в свою очередь, должен состоять из модулей извлечения связей и сущностей (и, соответственно, построения онтологии искомой темы) и тематического моделирования как конкретного документа, так и коллекций документов, составленных с помощью перекрёстных ссылок и связей, обнаруженных на предыдущих этапах. В рамках тематического моделирования важно найти не только текущее распределение по темам, но и изменение тем, связанных с объектом, в искомом промежутке времени, что позволит определить тренд настроений и событий, связанных с конкретной организацией.

Инструменты для формирования отчетов, которые должен будет изучить аналитик при наличии спорных ситуаций или в решениях, где требуется особая внимательность, должны так же включать методы автоматического реферирования.

В процессе создания программного комплекса необходимо отдельно проверять каждый алгоритм, используемый при анализе.

После окончания работ по тестированию корректности и производительности каждого отдельно взятого алгоритма, необходимо провести апробацию методологии в целом, путем сравнения экспертной оценки с оценкой, полученной программным комплексом. Эксперимент для апробации методологии должен быть так же проведен на поиск релевантных и адекватных задаче материалов и должен проводиться с привлечением эксперта в предметной области (информации о ряде компаний, юридических лиц и т. д., исследуемых в рамках эксперимента).

6. Заключение

- Сделан вывод, что объемы открытых неструктурированных данных и развитие способов их обработки, позволяют говорить о новом источнике для анализа данных, который поможет расширить существующие модели анализа репутации и кредитоспособности.
- На данном этапе развития технологии ещё невозможна полная замена аналитика на интеллектуальную систему для принятия управленческих решений, но ряд рутинных задач, таких как поиск, очистка и предварительный анализ текстовых данных, уже сейчас возможно реализовать в программном комплексе.
- Использование неструктурированных данных для скоринга позволяет улучшить классификационную способность модели в области граничного балла, избежав ошибок и уменьшив общую стоимость финансово-аналитических работ, автоматизировав часть функционала.

Список используемых источников

1. О типичных банковских рисках: Указание оперативного характера Банка России № 70-Т от 23.06.2004. Доступ из справ.-правовой системы «КонсультантПлюс».
2. Зобова Е. В., Самойлова С. С. Управление кредитным риском в коммерческих банках // Социально-экономические явления и процессы. Тамбов, 2012. № 12 (046). С. 74-81.
3. Развитие интернета в регионах России. [Электронный ресурс] – URL:https://yandex.ru/company/researches/2016/ya_internet_regions_2016 (дата обращения: 01.03.2018)
4. Wiil U. Counterterrorism and Open Source Intelligence. Lecture Notes in Social Networks. Springer, 2011.
5. Сокирко А.В. Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции Диалог'2004 / Под ред. И.М.Кобозевой, А.С. Нариьяни, В.П. Селегея. М.: Наука, 2004. с. 559-564.
6. Большакова Е.И., Носков А.А. Программные средства анализа текста на основе лексико-синтаксических шаблонов языка LSPL // Программные системы и инструменты: Тематический сборник, № 11 / Под ред. Королева Л.Н. – М.: МАКС Пресс, 2010, с. 61-73.
7. Angeli G. et al. Manning. Leveraging Linguistic Structure For Open Domain Information Extraction. In Proceedings of the Association of Computational Linguistics (ACL), 2015.
8. Злобина Н.В. Управленческие решения: учебное пособие. – Тамбов: Изд-во ТГТУ, 2007. – 80 с.
9. Rubin T. N., Chambers A., Smyth P., and Steyvers M. (2012), Statistical topic models for multi-label document classification, Machine Learning, Vol. 88, No. 1-2, pp. 157-208.
10. Игнатов Д.И., Кузнецов С.О. О поиске сходства Интернет-документов с помощью частых замкнутых множеств признаков // Труды 10-й национальной конференции по искусственному интеллекту с международным участием (КИИ'06). – М.:Физматлит, 2006, Т.2, стр.249-258
11. Кузнецов С.О., Игнатов Д.И., Объедков С.А., Самохин М.В. Порождение кластеров документов дубликатов: подход, основанный на поиске частых замкнутых множеств признаков. Интернет-математика 2005. Автоматическая обработка веб-данных. Москва: "Яndex", 2005, стр. 302-319
12. Jonas Poelmans, Paul Elzinga, Stijn Viaene, Guido Dedene: A Case of Using Formal Concept Analysis in Combination with Emergent Self Organizing Maps for Detecting Domestic Violence.ICDM 2009: 247-260
13 Остаточное сопротивление – // "Лизинг". – 2017-09-11 – С. 15. [Электронный ресурс] – URL:https://www.kommersant.ru/doc/3460553 (Дата обращения: 01.03.2018)
14. Глава Сбербанка: интеллектуальная система управления ежедневно ставит 300 тыс. Задач – URL:http://www.banki.ru/news/lenta/?id=9763272 (дата обращения: 01.03.2018)
15. Ruder, S. (2017). An Overview of Multi-Task Learning in Deep Neural Networks. In arXiv preprint [Электронный ресурс] – URL: https://arxiv.org/abs/1706.05098 (дата обращения: 01.03.2018)
16. David M. Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation // Journal of Machine Learning Research (3) 2003 pp. 993-1022.