DOI: 10.15514/ISPRAS-2024-36(3)-14

Автоматизация подготовки ответов на требования налоговых органов с использованием обучения со слабым контролем

А.Д. Сосновиков, ORCID: 0009-0004-1447-532X <*sosnovikov.artur@gmail.com*>
Д.Ю. Турдаков, ORCID: 0000-0001-8745-0984 <*turdakov@ispras.ru*>

Институт системного программирования РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.

Аннотация. В данной статье рассматривается применение методов обучения со слабым контролем для автоматизации обработки налоговых требований в банковском секторе. В последние годы наблюдается значительный рост интереса к автоматизации процессов с использованием методов машинного обучения и искусственного интеллекта в финансовом секторе, что обусловлено стремлением повысить эффективность, точность и качество обслуживания клиентов. Предыдущие исследования в области автоматизации финансовых процессов часто ограничивались применением классических подходов машинного обучения, требующих больших объёмов качественно размеченных данных. Однако в условиях банковского дела, особенно в таких специфических задачах, как обработка требований налоговых органов, разметка данных сталкивается с заметными трудностями из-за необходимости привлечения высококвалифицированных специалистов и соблюдения конфиденциальности. В этом контексте наша работа направлена на заполнение пробела в исследованиях путём применения обучения со слабым контролем — подхода, позволяющего использовать неточные, противоречивые или неполные данные для обучения моделей. Это особенно актуально для банковской сферы, где данные быстро устаревают и часто имеют ограниченный доступ из-за нормативных ограничений. Методологически для реализации идеи обучения со слабым контролем мы использовали фреймворк Snorkel для создания обучающего набора данных с использованием разметочных функций, разработанных в сотрудничестве с экспертами банка Точка. Это позволило существенно снизить зависимость от трудоёмкого процесса ручной разметки данных и использовать большие объёмы неразмеченных документов. Результаты исследования показали, что подходы, основанные на слабом контроле, могут значительно повысить эффективность обработки налоговых требований, предоставляя модели, способные с высокой точностью классифицировать и интерпретировать различные типы налоговых документов. Особенно важно, что применение слабого контроля позволяет учесть необходимость постоянного обновления данных и законодательства, что делает его предпочтительным для динамически изменяющихся условий финансового сектора. Использование обучения со слабым контролем в автоматизации ответов на налоговые требования не только улучшает качество обработки данных, но и способствует уменьшению нагрузки на специалистов, повышая общую эффективность финансовых операций. Эти выводы могут оказать влияние на дальнейшее применение машинного обучения в финансовом секторе, с учетом важности инновационных подходов в условиях ограниченных данных.

Ключевые слова: обучение со слабым контролем; финансовый сектор.

Для цитирования: Сосновиков А.Д., Турдаков Д.Ю. Автоматизация подготовки ответов на требования налоговых органов с использованием обучения со слабым контролем. Труды ИСП РАН, том 36, вып. 3, 2024 г., стр. 203–212. DOI: 10.15514/ISPRAS–2024–36(3)–14.

Automating the Process of Responding to a Tax Request Using Weak Supervision

A.D. Sosnovikov, ORCID: 0009-0004-1447-532X <*sosnovikov.artur@gmail.com*>
D.Y. Turdakov, ORCID: 0000-0001-8745-0984 <*turdakov@ispras.ru*>

Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.

Abstract. This paper examines the application of weak observation techniques to automate tax claim processing in the banking sector. Interest in process automation using machine learning and artificial intelligence techniques in the financial sector has grown significantly in recent years, driven by the desire to improve efficiency, accuracy and customer service. Previous research in financial process automation has often relied on traditional machine learning approaches that require large amounts of well-analyzed data. However, in the banking industry, and especially in specific tasks such as processing tax authority claims, data annotation faces significant challenges due to the need for highly skilled professionals and privacy issues. Our work therefore aims to fill the research gap by applying weak observation, a technique that allows the use of imprecise, inconsistent or incomplete data to train models. This is particularly relevant for the banking sector, where data become outdated quickly and often have limited access due to regulatory constraints. Methodologically, to implement the idea of weak supervision, we used the Snorkel framework to create a training dataset using markup functions developed together with Bank Point experts. This allowed us to significantly reduce the dependence on the time-consuming process of manual data markup and to use large volumes of unlabeled documents. The results of the study showed that weak control approaches can significantly improve the efficiency of tax claims processing by creating models capable of classifying and interpreting different types of tax documents with high accuracy. In addition, the application of weak control can accommodate the need for constant updating of data and legislation, making it preferable for the dynamically changing environment of the financial sector. Using weak supervision to automate responses to tax claims not only improves the quality of data processing, but also helps to reduce the workload of specialists, improving the overall efficiency of financial operations. These results could have an impact on the future application of machine learning in the financial sector, given the importance of innovative approaches in data-constrained environments.

Keywords: weak supervison; financial sector**.**

For citation: Sosnovikov A.D., Turdakov D.Yu. Automating the process of responding to a tax request using weak supervision. *Trudy ISP RAN/Proc. ISP RAS*, vol. 36, issue 3, 2024. pp. 203-212 (in Russian). DOI: 10.15514/ISPRAS-2024-36(3)-14.

1. Введение

В последние годы в финансовом секторе произошёл значительный сдвиг в сторону автоматизации, обусловленный потребностью в повышении эффективности, точности и улучшении качества обслуживания клиентов. Одна из новых задач в этой области — автоматизация обработки налоговых требований, которая является критически важной и сложной задачей, требующей знания нормативных требований и предоставления точных и своевременных ответов клиентам. В этой статье рассматривается новое применение обучения со слабым контролем для автоматизации процесса обработки налоговых требований в банковском секторе, что представляет собой существенное отклонение от традиционных подходов, которые в значительной степени опираются на системы, основанные на правилах, или создание классификаторов на основе машинного обучения.

Процесс получения размеченных наборов данных для обучения моделей машинного обучения, особенно в таких узкоспециализированных областях, как банковское дело и соблюдение налогового законодательства, связан с рядом серьёзных трудностей. Во-первых, одной из ключевых проблем является потребность в привлечении экспертов, специализирующихся на точном аннотировании данных, однако такие специалисты зачастую перегружены основными обязанностями и не имеют возможности тратить время на аннотирование наборов данных.

Во-вторых, конфиденциальность финансовой и личной информации требует строгого соблюдения законов и правил конфиденциальности, таких как GDPR в Европе, которые могут ограничивать доступ к данным или требовать их анонимности, потенциально снижая их полезность. Более того, изменчивость налогового законодательства обусловливает быстрое устаревание наборов данных, что требует их постоянного обновления. Наконец, часто наблюдается дефицит общедоступных данных в данной сфере, что затрудняет получение репрезентативных выборок для обучения и тестирования. В итоге совокупность этих проблем делает процесс сбора данных для обучения моделей машинного обучения в банковской и налоговой сферах особенно сложным.

Метод обучения со слабым контролем (weak supervision) представляет собой подход в области машинного обучения, при котором используются неточные, противоречивые или неполные данные для обучения моделей. Этот подход позволяет эффективно использовать большие объемы данных, даже если лишь часть из них обладает качественными метками.

Для нашего исследования в рамках автоматизации процесса ответа на налоговые требования в АО Точка, где имелась ограниченная выборка размеченных данных, получение новых размеченных данных являлось дорогостоящим и трудоемким процессом. Привлечение бухгалтеров второй линии поддержки для этой задачи было невозможным из-за их высокой рабочей загруженности. Однако, путем совместной работы с экспертами была разработана базовая версия системы правил, которая, несмотря на свою неидеальность, позволила провести аннотацию значительного объема неразмеченных документов.

2. Обзор релевантных работ

Автоматизация финансовых процессов с помощью машинного обучения вызывает растущий интерес: несколько исследований подчеркивают потенциал повышения эффективности, точности и удовлетворенности клиентов. Модели машинного обучения и автоматизация процессов с их помощью нашли широкое применение в таких задачах как: алгоритмическая торговля [1, 2], детекция фрода [3], чат-боты поддержки клиентов, в задачах кибербезопасности [4] и т.д.

Что касается соблюдения нормативных требований, недавние исследования показали, что использование алгоритмов обработки естественного языка (NLP) для интерпретации нормативных документов и реагирования на них имеют высокий потенциал для оптимизации сложных административных задач [5]. Однако все описанные методы требуют большого числа размеченных примеров для обучения моделей, что редко осуществимо в финансовой сфере.

Концепция слабого контроля, при которой модели обучаются на зашумленных, ограниченных или неточно маркированных данных, получила распространение как экономически эффективная альтернатива традиционным методам контролируемого обучения. Методы обучения со слабым контролем применимы и показывают успехи как в задачах бинарной классификации [6], так и в задачах с более сложным пространством целевых меток [7].

Ратнер и др. [6] представили Snorkel, структуру для генерации обучающих данных с использованием слабого контроля, что значительно снижает потребность в больших аннотированных наборах данных [8]. Snorkel представляет собой инновационный подход в области машинного обучения, основанный на принципе обучения со слабым контролем, который позволяет эффективно использовать большие объемы данных с неточными, скоррелированными или неполными метками. Процесс построения модели включает два ключевых этапа: первый — генеративная модель [9], представляющая собой факторный граф, которая создает вероятностные метки на основе набора правил или эвристик, разработанных экспертами; второй — дискриминативная модель, которая использует эти метки созданные генеративной моделью для традиционного обучения с учителем модели и

извлечения ценной информации из данных, которые ранее были непригодны для использования в обычных алгоритмах обучения.

Этот подход применялся в различных областях, но его применение в финансовом секторе, особенно для процессов, связанных с налогообложением, остается недостаточно изученным.

3 Метод

Для автоматизации процесса ответа на требования налоговой, первоочередной задачей является определение темы документа. Экспертами было выделено 20 потенциальных тем требований. На начальном этапе мы сосредотачивались на задаче определения связи требования налоговой с вопросами имущества. Для этого мы решали задачу бинарной классификации, где документы, связанные с имуществом, были отнесены к положительному классу. На тот момент у нас был небольшой набор из 79 размеченных и значительный объем неразмеченных документов, что считается важным условием успешной реализации методов обучения со слабым контролем.

После чего мы провели анализ размеченной выборки и разработали разметочные функции. Сначала мы сформировали правила на основе анализа текстов размеченной выборки и совместно с экспертами-бухгалтерами уточнили и расширили эти правила для разметочных функций.

Применяя разметочные функции с использованием Snorkel к неразмеченным данным, мы отобрали 140 документов, которые генеративная модель Snorkel уверенно классифицировала как документы положительного класса, и передали их для дальнейшей разметки. После этого мы провели еще одну итерацию доработки разметочных функций на основе полученной выборки.

Далее мы случайным образом выбрали 270 документов и попросили экспертов разметить их для создания тестовой выборки. На этой выборке мы измеряли качество применяемых нами методов.

4. Экспериментальное исследование

4.1 Набор данных

В рамках данного исследования были использованы три основных набора данных (табл. 1):

- **Обучающая выборка (train)**: Этот набор данных состоит из 12,640 неразмеченных документов, которые включают требования налоговой, отправленные клиентам Точка Банка в период с января 2022 года по февраль 2024 года. Документы служат сырыми данными для обучения модели с использованием методов обучения со слабым контролем.
- **Выборка для разработки (dev)**: В эту выборку входит 219 размеченных документов, которые использовались для формирования и проверки эвристических правил, основанных на регулярных выражениях. Эта выборка содержит примеры всех типов требований, присутствующих в данных. Она была сформирована итеративно, не является случайной выборкой из общей совокупности. Доля положительного класса составляет 65%.
- **Тестовая выборка (test)**: Этот набор включает в себя 270 размеченных документов, которые были случайно выбраны из общей совокупности всех требований налоговой. Выборка используется для оценки эффективности разработанных моделей. Процент положительного класса в тестовой выборке составляет 7.4%.

Табл. 1. Характеристик наборов данных, используемых для обучения и оценки качества моделей.
Table 1. Characteristics of data sets used for training and assessing the quality of models.

Тип выборки	Количество документов	Статус разметки	Доля положительного класса	Примечания
Обучающая выборка	12,640	Неразмеченные	Не применимо	Данные за период с января 2022 г. по февраль 2024 г.
Выборка для разработки	219	Размеченные	65%	Итеративно сформирована, все типы требований
Тестовая выборка	270	Размеченные	7.4%	Случайная выборка из генеральной совокупности

4.2 Используемые модели и их гиперпараметры

В рамках исследования были применены две основные модели машинного обучения: логистическая регрессия и градиентный бустинг. Ниже представлены детали моделей и их гиперпараметры.

Логистическая регрессия: для реализации использовалась библиотека `sklearn.linear_model.LogisticRegression`. Выбор гиперпараметров следующий: тип регуляризации `penalty='l1'`, коэффициент регуляризации `C=5` (для обучения на выборке для разработки) и `C=0.1` (для обучения на слабой разметке), алгоритм оптимизации `solver='liblinear'`. Все остальные гиперпараметры были оставлены по умолчанию.

При использовании логистической регрессии текст документов признаки были векторизованы с использованием обучающей выборки и tfidf

Градиентный бустинг: применялась реализация градиентного бустинга из библиотеки `CatBoostClassifier`. Конфигурация гиперпараметров включала в себя: количество деревьев `n_estimators=1200`, темп обучения `learning_rate=0.01`, глубина деревьев `depth=5`, и количество раундов для ранней остановки `early_stopping_rounds=50`. В качестве валидационной выборки для ранней остановки использовалась выборка для разработки (dev).

При использовании бустинга текстовые признаки обрабатывались встроенными алгоритмами `CatBoostClassifier`, гиперпараметры, которого оставлены по умолчанию.

Для слабой разметки с помощью библиотеки Snorkel был составлен набор разметочных функций (табл. 2), возвращающих одну из меток: NEGATIVE или POSITIVE в случае выполнения правила и ABSTAIN в противном случае.

Все функции разметки имеют следующую структуру:

```
@labeling_function()
def <Название функции>(x):
    x = x.text.lower()
    if <Условие>:
        return <Метка>
    return ABSTAIN
```

4.3 Результаты экспериментов

В ходе экспериментов мы сравнили несколько моделей, а также подходы обучения со слабым контролем и классическое обучение с учителем. Ниже приведены метрики, которые были оценены на тестовой выборке документов (табл. 3).

Табл. 2. Разметочные функции (правила) для Snorkel
Table 2. Snorkel labeling functions

Условие функции разметки	Метка при выполнении правила	Текстовый комментарий
`'страховых взносов' in x`	NEGATIVE	Помечает текст как NEGATIVE, если он содержит упоминание о страховых взносах.
`'%' in x and '6%' not in x and '15%' not in x and '20%' not in x`	NEGATIVE	Помечает текст как NEGATIVE, если он содержит нестандартные проценты налогов.
`'земельн' in x`	POSITIVE	Помечает текст как POSITIVE, если он содержит упоминание о земле.
`'отчужд' in x`	POSITIVE	Помечает текст как POSITIVE, если он содержит упоминание об отчуждении имущества.
`'имущество' in x`	POSITIVE	Помечает текст как POSITIVE, если он содержит упоминание о каком-либо имуществе.
`re.search(r'\d{2}:\d{2}:\d{6,7}:\d{2,4}', x)`	POSITIVE	Помечает текст как POSITIVE, если он содержит кадастровый номер.
`'квартир' in x`	POSITIVE	Помечает текст как POSITIVE, если он содержит упоминание о квартирах.
`' дом' in x`	POSITIVE	Помечает текст как POSITIVE, если он содержит упоминание о домах.
`'земельный' in x and '(земельн' not in x \ or re.search(r'земельных участков', x) or re.search('земельного', x) or 'садовод' in x \ or 'земельные объекты (участки)' in x or re.search('земельные[ ]{0,5}объекты[ ]{0,5}\(участки\)', x`	POSITIVE	Помечает текст как POSITIVE, если он содержит упоминание о земельных объектах.
`' дом' in x or 'строени' in x\ or 'помещен' in x or 'квартир' in x or re.search(r'\d{2}:\d{2}:\d{6,7}:\d{2,4}', x) \ or re.search(r"(отчуждени)*(недвижим\w{2,3} имуществ\w{1,2})", x): if 'зарегистрировано имуществ' not in x and not re.search(r'земельных участков', x) \ and not re.search('земельного', x)`	POSITIVE	Помечает текст как POSITIVE, если он содержит упоминание о домах или строениях.
`' тс ' in x or 'марка' in x and 'гос.номер' in x`	POSITIVE	Помечает текст как POSITIVE, если он содержит упоминание о транспортных средствах.
`re.search(r'[а-я]\d{3}[а-я]{2}[ ]{0,1}\d{1,3}', x) \ or re.search(r'марк? тс', x) or re.search(r'транспортно\w{1,2} средств[оа] марк?', x) \ or re.search(r'транспортное средство [А-Я]+ [А-Я]+', bs)`	POSITIVE	Помечает текст как POSITIVE, если он содержит упоминание о транспортных средствах.
`re.search('сумма доходов[ а-я]{0,20}[\d]{5,15}', x)`	NEGATIVE	Помечает текст как NEGATIVE, если он содержит упоминание о доходах.
`re.search('сумм[ а]{0,3} расход\w{1,2}', x)`	NEGATIVE	Помечает текст как NEGATIVE, если он содержит упоминание о расходах.

В рамках данного исследования применялся специальный метод формирования обучающей выборки, направленный на повышение ее информативности. В результате распределение классов в обучающей выборке отличалось от исходного в генеральной совокупности. Чтобы гарантировать корректную оценку качества модели на новых данных, многоблочная кросс-валидация с объединением выборок не проводилась. Вместо этого использовались отдельные стратифицированные по классам тестовая и обучающая выборки, сохраняющие первоначальное соотношение классов.

Генеративная модель, обученная на тренировочной выборке и разметочных функциях, продемонстрировала неплохие результаты, и послужила базовой моделью для проведения сравнений. Данная модель учитывает корреляционную структуру и конфликты меток, создаваемых разметочными функциями при генерации итоговых вероятностных меток документов. Кроме того, на основе разметки, полученной от генеративной модели, были обучены дискриминативные модели.

Среди дискриминативных моделей для проведения экспериментов были выбраны **Логистическая регрессия** (Snorkel + LogReg) и **Бустинг** (Snorkel + CatBoost), обученные на слабой разметке на обучающей выборке. Пороги бинаризации и гиперпараметры моделей подбирались так, чтобы максимизировать точность на выборке для разработки, имея при этом максимально возможную полноту.

Сравнительный анализ проведен с моделью логистической регрессии LogReg (dev), обученной на выборке для разработки без использования слабой разметки, т.е. был применен классический подход обучения с учителем.

Использование аналогичного подхода с моделью CatBoost, обученной на размеченной выборке для разработки, не дало положительных результатов, поскольку модель градиентного бустинга не смогла эффективно обучиться на столь малом объеме данных.

Дообучение моделей семейства BERT для классификации документов при малом объеме размеченной выборки, как в нашей задаче, не представляется возможным. Ещё одной проблемой применения моделей семейства BERT к рассматриваемой задаче является наличие значительного количества нерелевантного по отношению к метке контекста в документах.

Табл. 3. Сравнение метрик качества моделей
Table 3. Comparison of model metrics

Модель	AUC-ROC	Средняя точность (AP)	F1-мера	Точность (precision)	Полнота (recall)
Snorkel	0.906	0.819	0.857	1.0	0.75
Snorkel + LogReg	0.976	0.903	0.824	1.0	0.7
Snorkel + CatBoost	0.991	0.938	0.919	1.0	0.85
LogReg (dev)	0.985	0.79	0.765	0.929	0.65

5. Анализ результатов

Анализ результатов экспериментов свидетельствует о том, что дискриминативная модель CatBoost, обученная на разметке, полученной от генеративной модели Snorkel, имеет самые высокие метрики качества. Это подчеркивает эффективность метода обучения со слабым контролем в условиях ограниченного объема размеченных данных.

Этот подход показал себя лучше, чем обучение на размеченной дискриминативной модели из-за недостаточного количества размеченных документов.

Однако, несмотря на высокие метрики качества, возникают две значительные проблемы. Во-первых, для расширения задачи на мультиклассификацию требуется создание значительного

количества правил, что может быть трудоемким процессом. Во-вторых, написание эффективных правил для данной задачи оказалось непростым и длительным процессом, требующим многократного итеративного анализа и вычитывания размеченных текстов на выборке для разработки. Что перекладывает нагрузку аннотирования с экспертов в данной области на специалистов по машинному обучению.

6. Заключение

Это исследование положило новаторскому применению обучения со слабым контролем для автоматизации ответов на налоговые претензии, что является важной, но сложной задачей в банковском секторе. Наш подход, основанный на разработке и сравнении нескольких решений в рамках этой парадигмы, позволил существенно понять преимущества и ограничения обучения со слабым контролем в этом контексте. Особенно заметной проблемой, с которой пришлось столкнуться, была неизбежная трудность в разработке эффективных функций разметки для задач многоклассовой классификации. Необходимость разработки обширного набора правил для охвата множества сценариев, возникающих при соблюдении налогового законодательства, значительно усложняет процесс разработки модели. Эта сложность усугубляется специализированным характером данных и тонким пониманием, необходимым для точной интерпретации запросов, связанных с налогами.

Как показывают наши результаты, разработка этих правил не всегда проста и требует глубоких знаний предметной области, которые не всегда легко доступны, а также многократной проверки системы правил на объектах обучающей выборки и выборки для разработки. Наш опыт показывает, что усилия, необходимые для написания эффективных правил для решения задач бинарной классификации и особенно для многоклассовых задач в специализированных областях, таких как соблюдение налогового законодательства, могут стать барьером на пути более широкого внедрения методов обучения со слабым контролем в индустрии.

Кроме того, наш анализ подчеркивает динамичный характер налогового законодательства и нормативных актов, что требует постоянного обновления правил и функций маркировки для поддержания актуальности и точности моделей. Этот аспект создает постоянную проблему обслуживания, которую необходимо решить, чтобы обеспечить долгосрочную жизнеспособность решений, разработанных с использованием обучения со слабым контролем.

Несмотря на эти проблемы, наше исследование также подчеркивает потенциал обучения со слабым контролем для автоматизации и повышения эффективности реагирования на налоговые претензии. Структурированно используя существующие знания, мы можем уменьшить зависимость от недостаточного объема размеченных данных и снизить некоторые затраты, связанные с традиционными подходами к обучению с учителем, при этом ощутимо повысив метрики качества моделей по сравнению с традиционным подходом обучения с учителем.

Список литературы / References

[1]. Treleaven, Philip, and Bogdan Batrinca. "Algorithmic regulation: automating financial compliance monitoring and regulation using AI and blockchain." Journal of Financial Transformation 45 (2017): 14 - 21.

[2]. Cartea, Álvaro, Sebastian Jaimungal, and José Penalva. Algorithmic and high-frequency trading. Cambridge University Press, 2015.

[3]. Wilson, H. James, and Paul R. Daugherty. "Collaborative intelligence: Humans and AI are joining forces." Harvard Business Review 96.4 (2018): 114-123.

[4]. Horowitz, Michael C., et al. Artificial intelligence and international security. Center for a New American Security., 2022.

[5]. Winter, Karolin, and Stefanie Rinderle-Ma. "Detecting constraints and their relations from regulatory documents using nlp techniques." On the Move to Meaningful Internet Systems. OTM 2018 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part I. Springer International Publishing, 2018.
[6]. Alexander J Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In VLDB, 2017
[7]. Shin, Changho, et al. "Universalizing weak supervision." arXiv preprint arXiv:2112.03865 (2021).
[8]. Ratner, Alexander J., et al. "Data programming: Creating large training sets, quickly." Advances in neural information processing systems 29 (2016).
[9]. Bach, S. H., He, B., Ratner, A., & Ré, C. (2017, July). Learning the structure of generative models without labeled data. In International Conference on Machine Learning (pp. 273-282). PMLR.

Информация об авторах / Information about authors

Артур Дмитриевич СОСНОВИКОВ – аспирант Института системного программирования с 2023 года. Сфера научных интересов: методы машинного обучения, обучение со слабым контролем.

Artur Dmitrievich SOSNOVIKOV – graduate student at the Institute of System Programming since 2023. Research interests: machine learning methods, weakly supervised learning.

Денис Юрьевич ТУРДАКОВ – кандидат физико-математических наук, заведующий отделом ИСП РАН, доцент кафедры системного программирования ф-та ВМК МГУ. Научные интересы: анализ естественного языка, извлечение информации, обработка больших данных, анализ социальных сетей.

Denis Yurievich TURDAKOV – Cand. Sci (Phys.-Math.), Head of Department at ISP RAS, associate professor of the Department of System Programming at Moscow State University. Research interests: natural language processing, information extraction, big data analysis, social network analysis.