КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

УДК 004.896

Е. А. Сучкова, аспирант
ИжГТУ имени М. Т. Калашникова

ПРИМЕНЕНИЕ МЕТОДОВ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ ДЛЯ ПОИСКА И ОЦЕНКИ ПОТЕНЦИАЛЬНЫХ КОНТРАГЕНТОВ

В статье рассмотрены методы и алгоритмы, применимые к задаче информационного поиска и извлечения знаний для поддержки принятия решения по выбору поставщика. На основе лингвистических правил и регулярных выражений разработана система поиска данных о потенциальных контрагентах по интернет-ресурсам, предназначенная для первичной оценки благонадежности поставщиков. Анализируется и оценивается эффективность разработанной методики и системы.

Ключевые слова: регулярные выражения, информационный поиск, оценка, контрагенты.

Анализ и нормализация естественно-языковых текстов является сложной задачей, требующей применения отдельных автоматизированных систем [1, 2], алгоритмов, разработки онтологий, адаптивных методик. Интеллектуальные технологии в системах поддержки принятия решений (СППР) ориентированы на повышение качества принимаемых руководителями управленческих решений, эффективность которых зависит от объема и содержания анализируемых данных [3]. Надежность контрагентов зависит от множества эндогенных и экзогенных факторов [4], анализ информации по которым ведет к снижению контрактационных рисков и повышению эффективности деятельности предприятия. В области закупок множество данных о предприятиях можно найти на интернет-ресурсах в виде отзывов, прайс-листов, информационных буклетов, списков недобросовестных поставщиков, но эти данные не стандартизированы и не структурированы, представлены в виде текстов, иногда таблиц и графиков, поэтому для извлечения знаний из многочисленной информации, которую можно найти в интернете по продуктам, поставщикам, ценам, необходимо использовать алгоритмы компьютерной лингвистики, направленные на извлечение численных и оценочных значений из открытых данных интернет-ресурсов.

Алгоритмы компьютерной лингвистики могут использоваться на различных этапах работы лица, принимающего решение (ЛПР). Они позволяют решать задачи информационного поиска и извлечения информации. Среди задач информационного поиска наиболее актуальны в СППР задачи поиска по ключевым словам документов и информации, индексирование больших информационных объемов, классификация текстов по категориям и рубрикам, кластеризации документов. Задачи извлечения информации (Data mining) направлены на распознавание и извлечение такой значимой информации, как сущности, их характеристики, связи и отношения между сущностями.

Источниками информации для руководителей могут быть не только общедоступные интернет-ресурсы, но и документы и данные из корпоративной сети. Поиск по всем доступным источникам позволяет увеличить объем потенциально доступных данных, но не позволяет гарантировать достоверность источников и качество получаемых данных. Другой вариант заключается в определении набора источников, по которым будет производиться поиск. В этом варианте есть несколько положительных характеристик, среди которых возможность обеспечить в качестве источников информации рецензируемые, предоставляющие достоверную информацию источники, в основном это государственные или корпоративные ресурсы, потенциальная возможность разработки отдельных программных агентов интеграции, настраиваемых под структуру каждого ресурса, возможность мониторинга изменений на детерминированном списке источников.

Исследуем возможности применения методов компьютерной лингвистики для поиска информации о контрагентах с целью оценки их надежности, добросовестности, качества продукции, возможных негативных последствий заключения контракта с определенным исполнителем. Есть два подхода к организации извлечения знаний из документов на естественном языке: с использованием машинного обучения (извлечение знаний на основе признаков или ядра) или на базе формализованных в виде правил знаний (методы, основанные на сопоставлении образцов, фразовые образцы). Эти методы отличаются полнотой извлечения и точностью, причем, как правило, чем больше полнота извлечения, тем ниже точность, и наоборот. При извлечении данных из веб-ресурсов выделяются три подхода: анализ DOM-структуры html-документа, анализа страницы после рендеринга с помощью методов компьютерного зрения, алгоритмы сравнения страниц с поиском различий. Анализ DOM-структуры в свою очередь может осуществляться путем анализа построенного дерева с помощью построения xpath-выражений, разбора документа как текстового массива, преобразования к XML и анализа получившейся структуры и использования регулярных выражений.

В настоящее время в России функционирует множество государственных программ и фондов

поддержки малого предпринимательства, что создает благоприятную структуру для выбора в качестве контрагента малых предприятий. Но при выборе малого предприятия следует предпринять дополнительные шаги для проверки благонадежности контрагента, для этого необходимо использовать как официальные источники (сайты Федеральной налоговой службы – ФНС, Росстата, Картотеки Высшего арбитражного суда, сайта zakupki.gov.ru, сайта судебных приставов, сайтов лицензирующих органов, сайта Федеральной службы по интеллектуальной собственности (Роспатента), сайта Генеральной прокуратуры, Единого федерального реестра сведений о банкротстве) для проверки соответствия предоставляемых контрагентом сведений о юридическом лице (ЮЛ), так и дополнительный поиск по открытым интернет-источникам и внутренним базам предприятия для получения дополнительной информации. После получения выписки из ЕГРЮЛ, копий регистрационных документов, свидетельств, лицензий необходимо проверить всю полученную информацию. Большинство этих проверок можно автоматизировать, реализовав информационную систему, выполняющую выделение атрибутов потенциальной компании контрагента из имеющихся данных и отправляющую запросы к существующим сервисам с дальнейшим анализом результатов запроса с применением методов компьютерной лингвистики. Поскольку информация на таких ресурсах является частично структурированной, лингвистический анализатор реализуется на основе описательных лингвистических правил и регулярных выражений, что позволит достичь высокой точности и низкой вычислительной сложности, малого времени выполнения запросов. Основные критерии оценки контрагента, информацию по которым можно найти на сайте ФНС, а также URL и список необходимых параметров для запросов представлены в табл. 1.

Помимо сайта ФНС, информацию о контрагентах, такую как наличие в списках недобросовестных поставщиков, юридическую информацию о компании, отзывы, можно найти на таких ресурсах, как сайт государственных закупок, федеральной антимонопольной службы, и других, критерии оценки контрагентов, URL и параметры запросов для получения информации представлены в табл. 2.

Таблица 1. **Источники данных для проверки контрагента через сайт ФНС**

№	Критерий оценки контрагента	Автоматизация получения данных
1	Наличие контрагента в ЕГРЮЛ	https://egrul.nalog.ru запрос с параметрами: ОГРН/ИНН, наименование (для ЮЛ), ОГРНИП/ИНН, ФИО и регион места жительства
2	Отсутствие директора контрагента, главного бухгалтера и других ответственных лиц в списке дисквалифицированных лиц	https://service.nalog.ru/disqualified.do запрос с параметрами: фамилия, имя, отчество, дата рождения, наименование организации, ИНН организации
3	Наличие сведений о контрагенте в сведениях о документах для государственной регистрации	https://service.nalog.ru/uwsfind.do запрос с параметрами: ОГРН/ОГНИП, наименование ЮЛ/ФИО ИП, форма поданного документа, ИФНС, период, в который был подан документ
4	Отсутствие в исполнительных органах контрагента дисквалифицированных лиц	https://service.nalog.ru/disfind.do запрос с параметрами: наименование юридического лица, ОГРН
5	Проверка адреса, указанного при государственной регистрации	https://service.nalog.ru/addrfind.do запрос с параметрами: регион, район, город, населенный пункт, улица, дом
6	Проверка отсутствия представителя контрагента в списке лиц, в судебном порядке лишенных права участия в организации	https://service.nalog.ru/svl.do запрос с параметрами: ОГРН, ИНН организации
7	Сведения о контрагентах, связь с которыми по адресу (месту нахождения) ЕГРЮЛ отсутствует	https://service.nalog.ru/baddr.do запрос с параметрами: ОГРН, ИНН, наименование
8	Сведения о юридических лицах, имеющих задолженность по уплате налогов и/или не представляющих налоговую отчетность более года	https://service.nalog.ru/zd.do запрос с параметрами: ИНН

Таблица 2. **Источники данных для проверки контрагента через другие ресурсы**

№	Критерий оценки контрагента	Автоматизация получения данных
1	Отсутствие ЮЛ в списке лиц, в отношении которых налоговые органы приняли решение о предстоящем исключении из ЕГРЮЛ	Сведения, опубликованные в журнале «Вестник государственной регистрации» о принятых регистрирующими органами решениях о предстоящем исключении недействующих юридических лиц из Единого государственного реестра юридических лиц http://www.vestnik-gosreg.ru/publ/fz83 запрос с параметрами: ОГРН/ИНН
2	Сообщения юридических лиц, опубликованные в журнале «Вестник государственной регистрации»	Сообщения юридических лиц, опубликованные в журнале «Вестник государственной регистрации» http://www.vestnik-gosreg.ru/publ/vgr/ запрос с параметрами: ОГРН/ИНН

Окончание табл. 2

№	Критерий оценки контрагента	Автоматизация получения данных
3	Проверка основной информации о контрагенте	Единый федеральный реестр сведений о фактах деятельности юридических лиц http://www.fedresurs.ru/companies/IsSearching запрос с параметрами: наименование, адрес, ОГРН/ИНН
4	Проверка отсутствия контрагента в реестре недобросовестных поставщиков	Реестр недобросовестных поставщиков сайта Федеральной Антимонопольной службы http://rnp.fas.gov.ru запрос с параметрами: наименование/ФИО, ИНН
5	Проверка отсутствия контрагента в реестре недобросовестных поставщиков	Реестр недобросовестных поставщиков сайта государственных закупок http://zakupki.gov.ru/epz/dishonestsupplier/quicksearch/search.html запрос с параметрами: наименование, ФИО, ИНН
6	Проверка отзывов о компании	Общероссийская сеть бизнес-порталов http://www.rosfirm.ru/ запрос с параметрами: наименование, ИНН, или др.

Программные агенты реализованы на языка Java с использованием технологий selenium, htmlunit, xpath. Параметры внешних ресурсов и xpath-выражения для поиска на страницах хранятся в базе данных, что позволяет быстро и без перекомпиляции приложения исправлять алгоритм поиска элементов в случае изменения стороннего сайта, с которого ведется поиск информации. Правила обработки данных реализованы на стороне приложения в виде методов на языке Java. Для возможности интеграции системы с другими информационными продуктами реализован REST API, с помощью которого можно отправлять в формате json запросы и получать ответ с результатом проверки. Пример запроса:

{"authorization":
"eyJ1c2VybmFtZSI6ImlwZXJ0b3ZAbWFpbC5ydSIsD
QogInRva2VuIjoiOTMyN2JjOWUtNWQ2Yy00ZDRjL
TljZWEtOTRjZThiMDAxMWFhIn0=",
"company": {
"ИНН":1831083343,
"ОГРН": 1021801143374}}

Схема взаимодействия программных модулей представлена на рис. 1.

Интерфейс системы позволяет осуществить поиск компании и анализ ее благонадежности согласно разработанным правилам и алгоритмам. Интерфейс системы представлен на рис. 2.

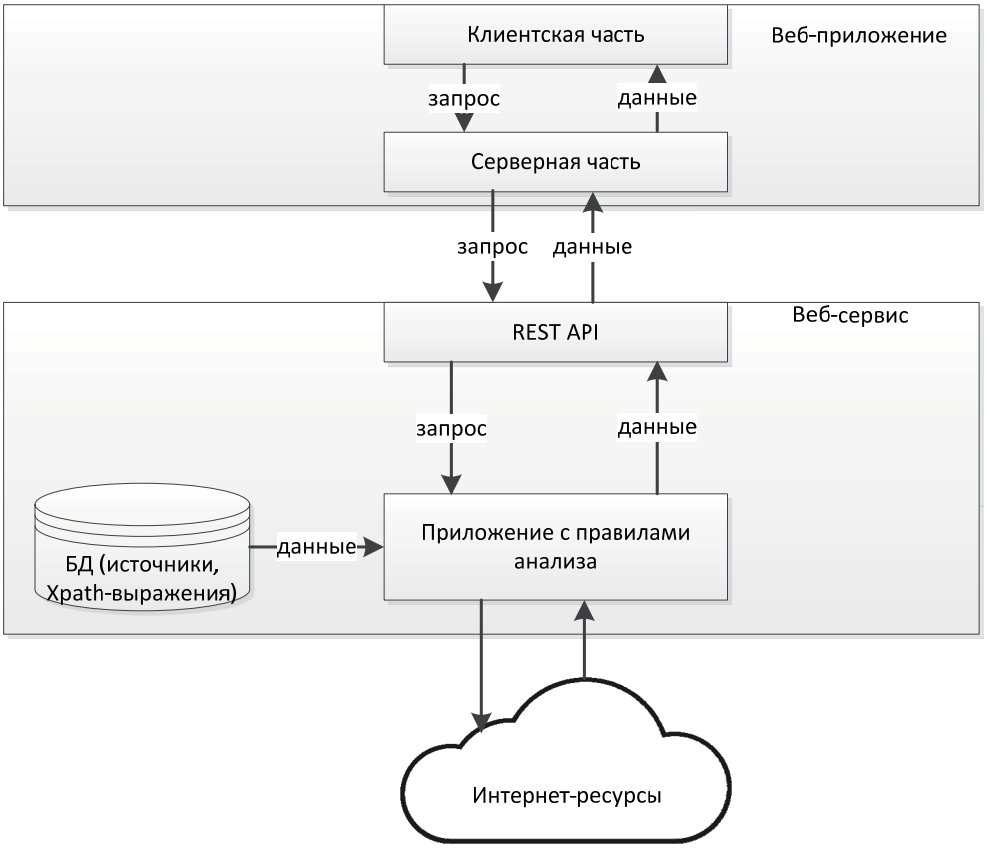
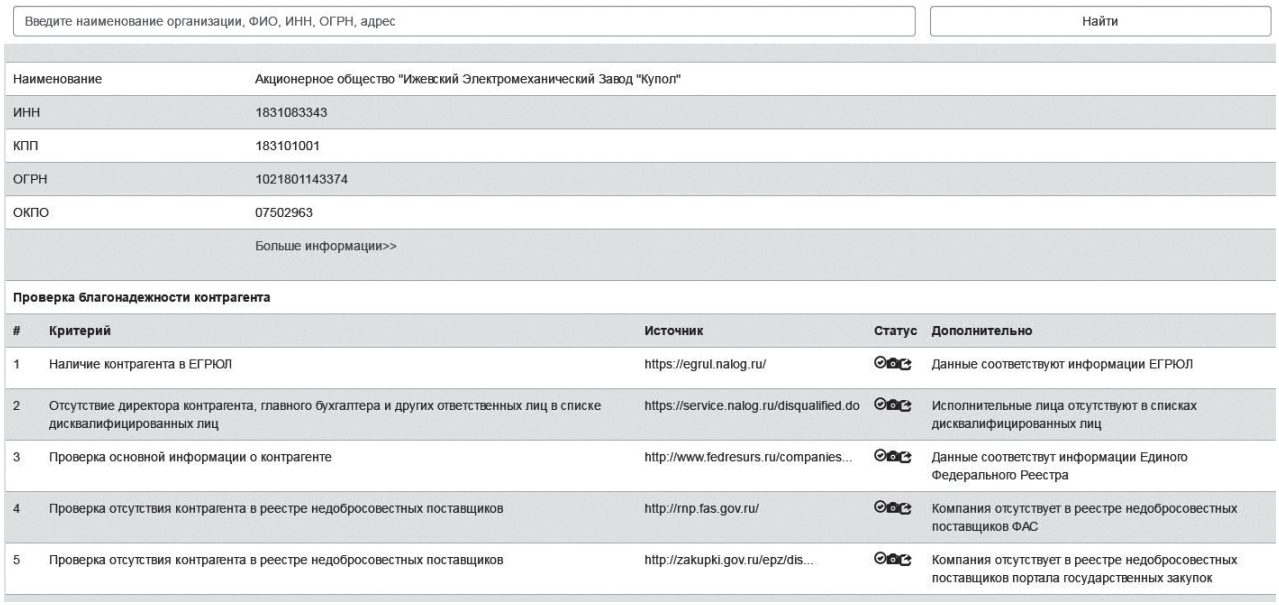


Рис. 1. Схема работы программных модулей


Введите наименование организации, ФИО, ИНН, ОГРН, адрес

Найти

Наименование	Акционерное общество "Ижевский Электромеханический Завод "Купол"
ИНН	1831083343
КПП	183101001
ОГРН	1021801143374
ОКПО	07502963
	Больше информации>>

Проверка благонадежности контрагента

#	Критерий	Источник	Статус	Дополнительно
1	Наличие контрагента в ЕГРЮЛ	https://egrul.nalog.ru/		Данные соответствуют информации ЕГРЮЛ
2	Отсутствие директора контрагента, главного бухгалтера и других ответственных лиц в списке дисквалифицированных лиц	https://service.nalog.ru/disqualified.do		Исполнительные лица отсутствуют в списках дисквалифицированных лиц
3	Проверка основной информации о контрагенте	http://www.fedresurs.ru/companies...		Данные соответствуют информации Единого Федерального Реестра
4	Проверка отсутствия контрагента в реестре недобросовестных поставщиков	http://rnp.fas.gov.ru/		Компания отсутствует в реестре недобросовестных поставщиков ФАС
5	Проверка отсутствия контрагента в реестре недобросовестных поставщиков	http://zakupki.gov.ru/epz/dis...		Компания отсутствует в реестре недобросовестных поставщиков портала государственных закупок

Рис. 2. Интерфейс системы

Таким образом, разработанная система, принцип действия которой основан на лингвистических правилах и регулярных выражениях, позволяет с высокой точностью осуществлять извлечение данных из естественно-языковых частично-структурированных текстов интернет-ресурсов. Высокая скорость и точность проверок осуществляется за счет агентного подхода к разработке архитектуры приложения и использования обоснованных методов компьютерной лингвистики.

Библиографические ссылки

1. *Мокроусов М. Н.* Автоматизированная система нормализации естественно-языковых текстов // Интеллектуальные системы в производстве. – 2015. – № 3 (27). – С. 93–96.

2. *Сучкова Е. А., Лялин В. Е.* Проблемы выбора поставщика – критерии, инструменты и методы оценки // Математические модели и информационные технологии в организации производства. – 2012. – № 2 (25). – С. 39–48.

3. *Давлетова Р. С., Файзуллин Р. В.* Моделирование зависимости состояния нефтедобывающего предприятия от эндогенных и экзогенных факторов // Проблемы экономики и управления нефтегазовым комплексом. – 2013. – № 3. – С. 33–37.

4. *Большакова Е. И., Жеребцова Ю. А.* Эксперименты по извлечению информации из аналитических текстов финансовых обзоров. // Сайт Всероссийской объединенной конференции «Интернет и современное общество». [Электронный ресурс]. – URL: http://conf.infosoc.ru/ 2012/ matherials/BOOK1/27BolshakovaZherebtsova.pdf, свободный.

* * *

E. A. Suchkova, Post-graduate, Kalashnikov ISTU

Application of computational linguistics for search and evaluation of potential contractors

The paper is devoted to methods and algorithms which can be applied to information search and data mining for decision support in supplier selection. Based on linguistic patterns and regular expressions, the author developed a system for potential contractors data retrieval in the Internet intended for the initial reliability assessment. The author analyzes and evaluates the effectiveness of the developed methods and system.

Keywords: regular expressions, information retrieval, evaluation, contractors.

Получено: 16.05.16