Метод автоматизированного извлечения адресов из неструктурированных текстов

А.В. Комарова, А.А. Менщиков, А.В. Полев, Ю.А. Гатчин

***Аннотация*: в данной статье представлен анализ методов поиска неструктурированной информации в сети Интернет. Авторы сосредотачиваются на вопросе извлечения из текста информации, содержащей почтовые адреса и географические ориентиры. Акцент сделан на двух основных методиках: анализе шаблонов и статистическом анализе с использованием машинного обучения. В работе описаны преимущества применения технологий автоматизированного поиска для «Умных городов» и инициативы открытых данных, которые становятся очень популярными на сегодняшний день, в том числе и в России. Также авторами было разработано программное обеспечение для сбора и извлечения информации из текста. Применяемые методы могут использоваться в качестве основы для систем анализа информации на сайтах объявлений, сайтах курьерских служб, а также в рамках построения семантических веб-ресурсов и систем управления знаниями.**

***Ключевые слова*: анализ текстов, информационный поиск, геопарсинг, извлечение информации, обработка естественных языков, умный город, обработка данных.**

I. Введение

На сегодняшний день данные, которые представлены в сети Интернет, очень слабо структурированы. Фактически отсутствует какая-либо стандартизация, либо иной принцип упорядочивания данных в сети. Каждая организация решает данную проблему удобным ей способом, что приводит к появлению большого числа противоречащих друг другу стандартов. Помимо этого, мало кто задумывается не только о структуре, но и о пригодности данных к автоматизированной обработке программными средствами [1].

Сегодня вопросы автоматизированной обработки больших объемов информации и структурирования представления данных в сети возникают повсеместно. Это связано как с развитием информационных ресурсов, так и с появлением совершенно новых инструментов в общественной сфере. Одним из основных потребителей данной информации являются информационные сервисы, представляющие инициативы открытых данных и «Умных городов» [2]. Именно Город сегодня становится основой общественного развития, точкой роста современной экономики. Ни для кого не секрет, что тематика «Умных городов» с каждым годом набирает все большую и большую популярность. Многие исследователи высказываются о том, что «Smart City» или «Умный город» - это город будущего, и такие масштабные проекты требуют совершенно иного подхода к организации хранения и обработки информации в Интернете [3].

В наш век информационных технологий и массовых коммуникаций колоссальный объем информации, обрушивающийся на современного человека за единицу времени, требует своевременной её обработки. Для того, чтобы в большом массиве данных отыскать необходимые сведения, пользователю нужно затратить много времени и сил, и не всегда это приводит к предполагаемому результату. Возникающие проблемы информационного поиска требуют решения двух категорий задач [4]:

1. Разработка и внедрение новых принципов представления информации;
2. Модификация существующей информации в вид, пригодный для автоматизированной обработки.

Задача информационного поиска включает в себя процессы сбора, обработки и передачи полученной информации лицам, в ней заинтересованным [5]. В более точном смысле, данный процесс можно определить как совокупность следующих этапов: определение и формулировка информационного запроса, выявление информационных источников, извлечение релевантной информации, оценка полученных результатов поиска [6].

При изучении алгоритмов поиска особый интерес вызывают механизмы выявления неструктурированной информации. Такая проблема часто возникает при решении задач автоматизированного определения адресов организаций и помещений на основе анализа контента, собранного в автоматическом режиме с веб-ресурсов [7]. В рамках данной статьи мы фокусируем внимание на вопросе извлечения из текста информации, содержащей почтовые адреса и географические ориентиры.

Менщиков Александр Алексеевич. Университет ИТМО. Факультет информационной безопасности и компьютерных технологий, кафедра проектирования и безопасности компьютерных систем, аспирант.
e-mail: menshikov@corp.ifmo.ru.

Комарова Антонина Владиславовна. Университет ИТМО. Факультет информационной безопасности и компьютерных технологий, кафедра проектирования и безопасности компьютерных систем, аспирант.
e-mail: piter-ton@mail.ru.

Полев Александр Васильевич. ИП Полев Александр Васильевич
e-mail: apolevki09@gmail.com.

Гатчин Юрий Арменакович. Университет ИТМО. Факультет информационной безопасности и компьютерных технологий, кафедра проектирования и безопасности компьютерных систем, профессор.
e-mail: gatchin@mail.ifmo.ru.

II. Актуальность

Ввиду переноса сервисов и различных услуг в сеть Интернет появляется необходимость искать и находить, причем находить автоматизировано разного рода информацию. Также возникает вопрос достоверности данных, вопрос перевода исторически сложившихся неструктурированных данных в данные структурированные. Решение задачи поиска и разбора почтовых адресов и ориентиров является одной из подзадач информационного поиска и может быть применено для решения смежных задач после успешной апробации [7]. Решение данной задачи также интересно бизнесу, в том числе, в рамках инициативы развития «Умных городов», для работы различных географических картографических сервисов. Система может быть задействована на разных сайтах сдачи и съема жилья, анализа стоимости недвижимости, например, в определённых районах.

Согласно программе «Цифровая экономика РФ» к 2025 году в России планируется создать 50 «Умных городов» [8]. Программа развития предполагает несколько направлений. Среди них есть такие, как информационная безопасность и «Умный город». Проект данной программы был создан Минкомсвязью России и одобрен президентом Владимиром Путиным. Все созданные государственными органами документы и данные планируется перенести в единое централизованное облачное хранилище, а доли услуг, предоставляемых органами государственной власти, в электронном виде составит 80%.

Из всего вышесказанного становится чётко ясно, что создание системы поиска и автоматизированного сбора информации с электронных ресурсов - очень актуальная тема и в данный исторический момент и в будущем.

Технологии обработки больших данных (Big Data), а также автоматизированного поиска информации будут иметь большое значение для «Умных городов»: для осуществления прозрачности в расчётах ЖКХ, для анализа количества трафика на дорогах, для поиска информации о свободных парковочных местах, для поиска своевременной информации о движении общественного транспорта, для получения данных с датчиков наполнения мусорных баков, для оптимизации полива парков и освещения улиц и дорог, для контроля уровня загрязнения в городе, а также уровня радиации, для быстрого получения медицинской помощи и т.д. [9]

На сегодняшний день необходимы компании и стартапы, занимающиеся разработкой и развитием концепции «Умных городов», занимающиеся «Умными технологиями». Для «Умных городов» очень важно собирать и обрабатывать информацию в реальном времени.

III. Извлечение адресов

Существует два основных подхода к решению задачи поиска неструктурированной информации в сети Интернет. Первый подход основан на анализе шаблонов, полностью описывающих грамматики, извлекаемые из текста. Суть данного метода состоит в анализе свойств и структуры отдельных единиц текста (слова, словосочетания, предложения), это позволяет легко определить тематику и контекст при условии наличия структуры. Данный подход позволяет получать высокие показатели качества, но требует исчерпывающей базы данных, что не всегда возможно при работе с адресами, форма которых широко варьируется [10, 11].

Второй подход базируется на машинном обучении, что позволяет работать с текстами произвольной формы и содержания, однако, порождает большое число ошибок и ложных интерпретаций. Для каждой единицы текста формируется набор характеристик, каждая из которых влияет на то, является ли текст соответствующим тематике или нет [12].

Комбинация данных подходов с учетом постоянного итерационного исправления алгоритма и ручной верификации исключительных случаев позволила авторам минимизировать ошибки и повысить точность обнаружения адресов для произвольных текстов [10-12].

Статистические подходы зависят не только от обучающей выборки и программной реализации, но и от используемого языка. Существующие исследования преимущественно касаются английского языка. Так, зарубежные исследователи предлагают использовать извлечение информации на основе статистики с выделением ключевых слов из текста и тегирования с помощью условных случайных полей. Ручное обучение на базе корпуса из 400 сайтов дало результаты точности порядка 0,89 и F-меры 0,74 [13]. Подход на основе шаблонов и пороговых значений совпадения с шаблоном позволил получить точность порядка 0,745 и F-меру 0,734 при извлечении адресов из досок объявлений сайта Yahoo [14, 15]. Использование всеобъемлющих справочников позволило исследователям из Германии достичь полноты 0,95 [16], однако, точность распознавания в статье не приводится. К тому же, появление подобного справочника относительно затруднено применительно к России. Использование гибридного подхода позволяет получить значения F-меры порядка 0,73 и 0,93 в зависимости от того, учитывать ли полноту распознаваемости адреса [17]. Исследователи из Китая смогли добиться точности распознавания адресов в 0,90, используя статистический подход и учитывая особенности языка [18].

IV. Источники данных

Для обучения и тестирования системы были сформированы наборы данных на базе нескольких сайтов объявлений об аренде недвижимости. Был осуществлен автоматизированный сбор таких объявлений. Данные были разделены на две части – для тех сайтов, где объявления формировались в виде обычного неформатированного текста и для тех, где помимо неформатированного текста были представлены отдельные поля для ввода адреса: город, улица, номер дома, номер корпуса и строения.

Данные с известными параметрами использовались

для обучения и верификации модели, а также дополнительно были просмотрены вручную и очищены от испорченных записей. Данные на основе чистого теста были вручную классифицированы. В результате было получено 960 записей, размеченных вручную и 4178 записей с классификацией, взятой из форматированных полей сайта.

Для улучшения результатов обнаружения ключевых слов, использовались справочники адресов, для их формирования пришлось осуществить парсинг государственных ресурсов, крупных картографических сервисов, а также изучить слова в размеченных полях сайтов, имеющих фиксированную структуру (некоторые сайты недвижимости с полями «город», «улица»).

Данные были разделены на несколько частей для формирования обучающей и тестовой выборки, а также нескольких небольших проверочных выборок для оценки влияния изменений состава признаков и кросс-валидации.

Для оценки результатов использовались численные оценки качества.

$$P = \frac{T_P}{T_P + F_P} \quad (1)$$

$$R = \frac{T_P}{T_P + F_N} \quad (2)$$

Точность определения адреса P вычисляется как отношение числа верных классификаций к общему числу классификаций является ли слово заданным элементом адреса (правильных и ошибочных).

Полнота определения адреса R вычисляется как доля верных классификаций элемента адреса относительно общего количества элементов данного класса.

Одновременно изучались точность и полнота как в среднем по разным типам элементов (улица, дом, ориентир), так и внутри данных групп.

На основе данных метрик вычислялась объединённая метрика F, задаваемая следующим образом.

$$F = 2 * \frac{P \times R}{P + R} \quad (3)$$

V. Процесс извлечения адресов

В каждой стране и регионе есть свои особенности формирования почтовых адресов, что накладывает определенную вариативность на модель. Алгоритм извлечения перебирает несколько вариантов и оценивает наилучший на основе простейших весовых характеристик и с помощью эксперта. Авторами был сформирован набор шаблонов, выделяющих блоки текста, содержащие якорные слова и характерные для почтового адреса выражения. Данная фраза извлекалась из текста с запасом в несколько слов сначала и после конца выражения. Для повышения точности обнаружения извлекались несколько блоков текста со сдвигом по «скользящему окну» с шагом в одно слово. Это позволило захватить лишние фразы, которые не относились к самому содержанию адреса, но давали понимание о контексте. Например, слова «квартира», «проживание».

VI. Процесс очистки текста

На данном этапе из текста удаляются нерелевантные специальные символы, скобки, элементы выравнивания или HTML разметки. Дополнительно на данном этапе происходила модификация текста с учётом морфологии, фраза разбивалась на отдельные слова и выражения и при помощи Томита-парсера Яндекса слова переводились в начальную форму пригодную для дальнейшего поиска в словаре.

Разбиение на слова

Текст после предварительной очистки разбивается на слова (токены) в соответствии с регулярным выражением [!,?;\S]+. Дополнительно фильтруются излишние токены из черного списка. Это помогает максимизировать скорость работы системы за счёт перебора только наиболее релевантных вариантов. Каждому токену сопоставляется тип (числовой, строковый). Также, каждый токен проверяется на то, является ли он географическим ориентиром, названием улицы или специальным якорем. Якорь – это слово, которое отсылает к появлению в пределах нескольких следующих слов искомой информации. Например, якорями являются слова «ул.», «Д.», «корп» и прочие.

Система позволяет варьировать размер окна, который указывает на количество слов в сегменте поиска.

В таблице представлены несколько наиболее эффективных якорей и их типы.

Таб. 1. Типы «якорей»

Тип якоря	Якорь
Улица	«Ул», «г.», «район»
Ориентир	«Площадь», «метро», «станция», «проспект»
Номер дома	«д.», «дом»
Дополнительный модификатор	«этаж», «кв.», «стр», «строение», «корп»

Каждый якорь дополнительно модифицируется при поиске с учётом регистронезависимости, падежа, рода, и числа.

Помимо текстовых модификаторов, существуют шаблонные модификаторы, которые учитывают варианты сокращённого последовательного написания адреса: например, «д. 24А/5».

Схематичная работа системы извлечения информации может быть представлена следующим рисунком.

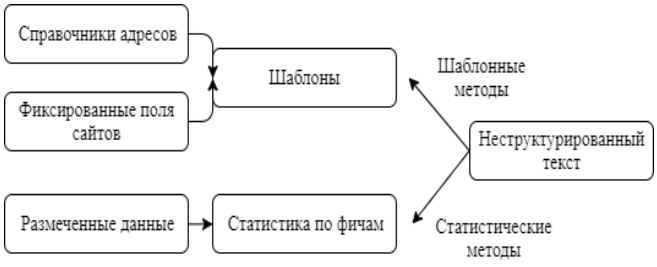

Рисунок 1. Схема используемых методов

VII. Реализация

Была написана программа на языке Python, которая состоит из трёх модулей. Модуль формирования словарей включает в себя функционал добавления новых текстовых якорей, ключевых слов, регулярных выражений и географических ориентиров, а также связей между ними. Его задачей является обновление базы данных и пересчёт моделей после внесения изменений и обновлений. Модуль классификации получает на вход неструктурированный текст и извлекает из него адреса и географические ориентиры. Модуль предполагает настройку уровня чувствительности, при необходимости можно отключить любые эвристические методы и оставить только поиск с использованием регулярных выражений и якорных слов. Модуль тестирования отвечает за сбор информации и поочерёдную проверку каждого текста из массива данных. В модуль встроен краулер, осуществляющий сбор данных с сайтов объявлений, а также генератор отчёта по результатам анализа, который позволяет обнаруживать ошибки и трудные случаи классификации для внесения их в словарь.

В качестве классификатора относится ли ключевое выражение к адресу был выбран наивный байесовский классификатор. Его простота реализации и скорость работы стали решающими при выборе, изучение других алгоритмов является дополнительной задачей для дальнейших исследований.

Для определения относится ли текстовый элемент к тому или иному классу использовался метод «один против всех». Были построены несколько классификаторов по числу возможных компонент адреса и оценена принадлежность слова или группы слов к каждому из них.

Для обучения классификатора были выбраны следующие основные признаки:

– количество распознанных якорей и расстояние до них от опорного слова;

– количество числовых элементов и расстояние до них от опорного слова;

– количество слов в блоке;

– соотношение типов якорей в блоке (например, улицы, дома, ориентиры);

– количество слов в предложении;

– порядок распознанных слов (например, улица после дома или номер дома после улицы);

– расстояния Левенштейна опорного слова до каждого из якорей выше определённого порога;

На первом этапе был сформирован перечень целевых веб-ресурсов для сбора объявлений и текстов, содержащих почтовые адреса. Сбор информации осуществлялся в несколько потоков.

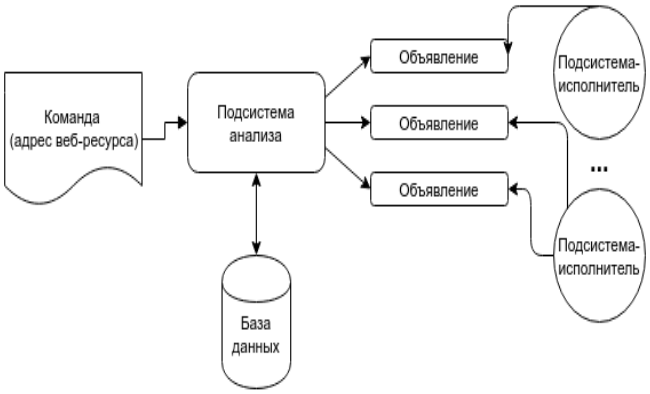

Рисунок 2. Схема системы сбора данных

Каждая независимая программа-исполнитель извлекала задание с адресом ресурса из очереди, осуществляла очистку текста, разбиение текста на токены и передавала задачу через систему очередей модулю анализатора, который непосредственно осуществлял извлечение информации, соответствие шаблонам и подсчёт характеристик для статистического анализа. Результаты помещались в базу данных, с которой непосредственно работал оператор системы.

Рис. 3. Схема извлечения информации

VIII. Результаты

Авторами был загружен набор текстов пользователей сайтов объявлений, связанных с недвижимостью. Многопоточным парсером в автоматизированном режиме было собрано 20 тысяч сообщений. Набор из 960 постов был размечен вручную для обучения дерева классификаций и составления фильтров. Данные также были разделены на две части – для тех сайтов, где объявления формировались в виде обычного неформатированного текста и для тех, где помимо неформатированного текста были представлены отдельные поля для ввода адреса: город, улица, номер дома, номер корпуса и строения. Данные с известными параметрами использовались для обучения и верификации модели, а также дополнительно были просмотрены вручную и очищены от испорченных записей. Изучались исключительные и типовые форматы. Система была протестирована на нескольких выборках для проверки ошибок и итерационного уточнения модели. В результате из полного набора было извлечено 880 валидных адресов. Также, осуществлялись контрольные проверки на случайных выборках из корпуса текстов.

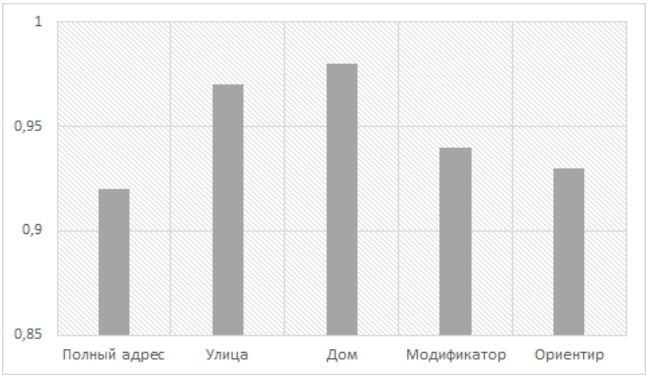

Рис. 4. Точность классификации компонент адреса

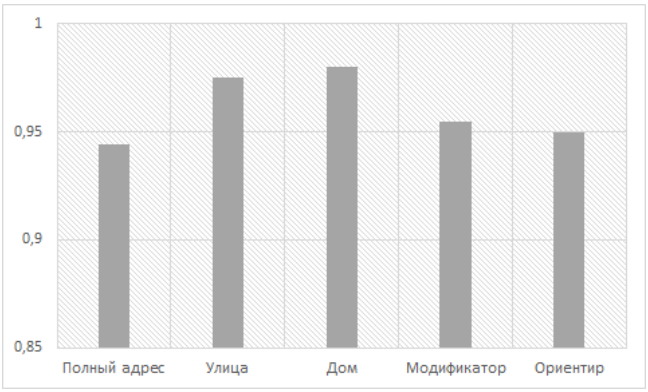

Рис. 5. F-мера классификации компонент адреса

В итоге была зафиксирована средняя точность классификации на уровне 92%, а F1 мера на контрольной выборке составила порядка 0,95. Результаты сильно зависят от используемых словарей адресов и ориентиров, поскольку многие адреса на большом числе сайтов используют валидацию адресов и семантическую разметку страницы. Для исключения таких упрощённых вариантов было решено взять большую часть информации с сайтов и форумов, предоставляющих обычный неструктурированный текст, а также проанализировать объявления в социальных сетях. Также, на результаты влияют географические и языковые особенности региона и структура аудитории веб-ресурса.

Скорость обработки текстов напрямую зависит от их структуры и объема, но является приемлемой, позволяя обрабатывать десятки тысяч текстов за несколько часов. Работа системы была автоматизирована, реализована программа, осуществляющая сбор, обработку и извлечение требуемых данных с возможностью оценки результатов.

IX. Выводы

Результаты исследования показали допустимую скорость работы при обработке больших массивов неструктурированных данных. Также, подход демонстрирует высокие показатели точности и полноты извлечения, что позволяет использовать его при решении практических задач информационного поиска. В силу того, что проблема автоматизированного поиска почтовых адресов и географических ориентиров из неструктурированных текстов в Интернете на сегодняшний день очень актуальна и важна, то дальнейшая работа в данном направлении будет способствовать развитию этой области и может послужить хорошим базисом для будущих исследований. Разработанное программное обеспечение и применяемые методы могут использоваться в качестве основы для систем анализа информации на сайтах объявлений, сайтах курьерских служб, а также в рамках построения семантических веб-ресурсов и систем управления знаниями.

Библиография

[1] Хорошевский В. Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1) // Искусственный интеллект и принятие решений. – 2008. – №1. – С. 80-97.

[2] Долгих Е. И., Антонов Е. В., Ерлич В. А. Умные города: перспективы развития в России // Урбанистика и рынок недвижимости. – 2015. – № 1. – С. 50–61.

[3] Hollands, R. G. Will the real smart city please stand up? Intelligent, progressive or entrepreneurial? // City. – 2008. – №12(3). – P. 303-320.

[4] Schmidt, Sebastian, et al. Extraction of address data from unstructured text using free knowledge resources. - Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies. – ACM. – 2013. – Article №7. URL: http://dl.acm.org/citation.cfm?doid=2494188.2494193 (дата обращения: 19.10.2017).

[5] Алексеев С. С., Морозов В. В., Симаков К. В. Методы машинного обучения в задачах извлечения информации из текстов по эталону // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – 2009. – С. 237-246. – URL: http://rcdl.ru/doc/2009/237_246_Section07-1.pdf (дата обращения 19.10.2017).

[6] Chang, Chia-Hui, Chia-Yi Huang and Yueng-Sheng Su. On Chinese Postal Address and Associated Information Extraction // The 26th Annual Conference of the Japanese Society for Artificial Intelligence. – 2012. – Pp. 1-7. – URL: https://www.researchgate.net/publication/267422107_On_Chinese_Postal_Address_and_Associated_Information_Extraction (дата обращения 19.10.2017).

[7] Nesi, Paolo, Gianni Pantaleo, and Marco Tenti. Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering // Engineering Applications of Artificial Intelligence 51. – 2016. – Pp. 202-211. – URL: http://dl.acm.org/citation.cfm?id=2910172 (дата обращения 19.10.2017).

[8] Распоряжение Правительства РФ от 28.07.2017 N 1632-р Об утверждении программы "Цифровая экономика Российской Федерации". - URL: http://static.government.ru/media/files/9gFM4FHj4PsB79I5v7yLVuPgu4bvR7M0.pdf (дата обращения: 19.10.2017).

[9] Добрынин А. П, Черных К. Ю., Куприяновский В. П., Куприяновский П. В., Синягов С. А. Цифровая экономика - различные пути к эффективному применению технологий (BIM, PLM, CAD, IOT, Smart City, BIG DATA и другие) // International Journal of Open Information Technologies. – 2016. – №1. – URL: http://cyberleninka.ru/article/n/tsifrovaya-ekonomika-razlichnye-puti-k-effektivnomu-primeneniyu-tehnologiy-bim-plm-cad-iot-smart-city-big-data-i-drugie (дата обращения: 19.10.2017).

[10] Zheyuan Yu. High accuracy postal address extraction from web pages // Masters Abstracts International. – 2007. – Vol. 45. – No. 05.

[11] Asadi S., Yang G., Zhou X., Shi Y., Zhai B., Jiang W. Pattern-Based Extraction of Addresses from Web Page Content // APWeb. – 2008. – Pp. 407-418. – URL: https://link.springer.com/chapter/10.1007/978-3-540-78849-2_41 (дата обращения 19.10.2017).

[12] Pasternack J. and Roth D. Extracting Article Text from The Web With Maximum Subsequence Segmentation // WWW. – 2009. – Pp. 971-980. – URL:

http://www.academia.edu/2661588/Extracting_article_text_from_the_web_with_maximum_subsequence_segmentation (дата обращения 19.10.2017).

[13] B. Loos and C. Biemann. Supporting Web-based Address Extraction with Unsupervised Tagging. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker // Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization. – 2008. – Pp. 577–584.

[14] S. Asadi, G. Yang, X. Zhou, Y. Shi, B. Zhai, and W.-R. Jiang. Pattern-Based Extraction of Addresses from Web Page Content. In Y. Zhang, G. Yu, E. Bertino, and G. Xu // Progress in WWW Research and Development. – 2008. – Vol. 4976 of Lecture Notes in Computer Science. – Pp. 407–418.

[15] D. Ahlers and S. Boll. Retrieving Address-based Locations from the Web // Proceedings of the 2nd international workshop on Geographic information retrieval, ACM . – 2008. – Pp. 27–34.

[16] Schmidt, S., Manschitz, S., Rensing, C., and Steinmetz, R. Extraction of address data from unstructured text using free knowledge resources // Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies. – P. 7.

[17] W. Cai, S. Wang, and Q. Jiang. Address extraction: Extraction of location-based information from the web // Web Technologies Research and Development. – 2005. – Vol. 3399 of Lecture Notes in Computer Science. – Pp. 925–937.

[18] Chang, Chia-Hui, Chia-Yi Huang, and Yueng-Sheng Su. On chinese postal address and associated information extraction. // The 26th Annual Conference of the Japanese Society for Artificial Intelligence. – 2012.

Method of Automated Address Data Extraction from Unstructured Text

A.V. Komarova, A.A. Menshchikov, A.V. Polev, Y.A. Gatchin

Abstract **– This article presents a method of automated address data extraction from unstructured text on the Internet. The authors focus on the issue of extracting information from the text containing postal addresses and geographical landmarks. The emphasis is on two main techniques: template analysis and statistical analysis with the use of machine learning. The paper describes the advantages of using automated search technologies for Smart Cities and for open data initiatives that are becoming very popular today. In addition, the authors developed software for collecting and retrieving information from the text. The method can be used as a basis for information analysis system on real estate web resources, as well as in semantic web resources and knowledge management systems building.**

Key words: **Text analysis, information search, geoparsing, information gathering, natural language processing, Smart City, data processing.**