

АВТОМАТИЗИРОВАННОЕ ИЗВЛЕЧЕНИЕ АДРЕСОВ ИЗ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВ

А.А. Менищikov, А.В. Комарова, Ю.А. Гатчин

Университет ИТМО

Санкт-Петербург

Как известно, поиск информации включает в себя сбор, обработку и передачу полученной информации заинтересованным лицам. Более конкретно, данный процесс состоит из следующих этапов: определение и формулировка информационного запроса, выявление информационных источников, извлечение информации, оценка полученных результатов поиска. В данной статье мы сосредотачиваемся на вопросе извлечения из текста информации, содержащей почтовые адреса и географические ориентиры. Данная проблема возникает при решении задач автоматизированного определения адресов организаций и помещений на основе анализа контента, собранного в автоматическом режиме с веб-ресурсов [1].

Существует два основных подхода к решению задачи поиска неструктурированной информации в сети Интернет [2]. Первый подход основан на анализе шаблонов, полностью описывающих грамматики, извлекаемые из текста [1, 3]. Данный подход позволяет получать высокие показатели качества, но требует исчерпывающей базы данных, что не всегда возможно при работе с адресами, форма которых широко варьируется [4, 5]. Второй подход базируется на машинном обучении, что позволяет работать с текстами произвольной формы и содержания, однако, порождает большое число ошибок и ложных интерпретаций [2, 6, 7]. 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.

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. В итоге была зафиксирована средняя точность классификации на уровне 92%, а F1 метрика на контрольной выборке составила 0,95. Скорость обработки текстов напрямую зависит от их структуры и объема, но является приемлемой, позволяя обрабатывать десятки тысяч текстов за несколько часов. 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.

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.

ЛИТЕРАТУРА

1. 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. <http://dl.acm.org/citation.cfm?doid=2494188.2494193> (дата обращения: 15.04.2017).
2. 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» - RCDL'2009, Петрозаводск, Россия, 2009. С. 237-246. http://rcdl.ru/doc/2009/237_246_Section07-1.pdf (дата обращения 14.04.2017).
3. 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.
https://www.researchgate.net/publication/267422107_On_Chinese_Postal_Address_and_Associated_Information_Extraction (дата обращения 15.04.2017).
4. 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. <http://dl.acm.org/citation.cfm?id=2910172> (дата обращения 16.04.2017).
5. Zheyuan Yu. High accuracy postal address extraction from web pages // Masters Abstracts International. Vol. 45. No. 05. 2007.
6. 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. https://link.springer.com/chapter/10.1007/978-3-540-78849-2_41 (дата обращения 15.04.2017).
7. Pasternack J. and Roth D. Extracting Article Text from The Web With Maximum Subsequence Segmentation // WWW 2009. Pp. 971-980.
http://www.academia.edu/2661588/Extracting_article_text_from_the_web_with_maximum_subsequence_segmentation (дата обращения 14.04.2017).