УДК 621.9-114

ПОДХОДЫ К ИЗВЛЕЧЕНИЮ ИНФОРМАЦИИ ИЗ СЛАБОСТРУКТУРИРОВАННЫХ ДАННЫХ

О. А. Молокович

o.molokovich@ya.ru

ФГБОУ ВО «Уфимский государственный авиационный технический университет» (УГАТУ)

Аннотация. В статье описаны источники больших данных, разделение их на типы и приведен обзор методов обработки слабоструктурированных данных, таких, как распознавание именованных сущностей и извлечение отношений.

Ключевые слова: слабоструктурированные данные; big data; data mining; natural language processing.

ВВЕДЕНИЕ

На протяжении последних десятилетий наблюдается беспрецедентный рост объема создаваемых и копируемых данных различной природы. Для описания этого феномена исследовательская и консалтинговая компания Gartner ввела понятие «Big Data». Большие данные – это информационные ресурсы большого объема, скорости и разнообразия, требующие рентабельных, инновационных форм обработки информации для улучшения понимания и принятия решений [1]. Источники больших данных грубо можно разделить на три категории: всемирная паутина и социальные сети, данные, генерируемые машинами (оборудование, датчики, интернет вещей), транзакционные данные (отражение различных операций, например, оплаты или отгрузки). Данные из этих источников могут быть структурированными, неструктурированными и слабоструктурированными.

Структурированные данные – описываются некоторой моделью, определяющей их обработку. Ярким примером этого типа служат реляционные базы данных.

Неструктурированные данные – не соответствуют какой-либо модели, либо она не задана явно. Их обработка алгоритмом напрямую затруднена, для анализа требуется предварительная обработка [2] Примеры неструктурированных данных: изображения, видео, аудио, текстовые и pdf-файлы.

Слабоструктурированные данные – не соответствуют реляционной модели данных, однако имеют некоторую структуру и элементы, помогающие ее распознать, что значительно упрощает обработку этих данных в отличие от неструктурированных данных. Примеры: xml-файлы, веб-страницы, пакеты TCP/IP, JSON-файлы.

Обработка слабоструктурированных и неструктурированных данных из различных источников может улучшить наше понимание происходящих процессов практически в любой области. Например, анализ данных в социальных сетях может помочь выявить тенденции в поведении людей, их потребности или получить обратную связь по услугам и товарам в бизнесе. Обработка и анализ сообщений системных журналов и трассировки промышленных систем способствует повышению надежности [3, 4, 5, 6]. Одной из тенденций в области данных и аналитики на 2020 Gartner указала использование структурированного и неструктурированного контента для решения самых сложных проблем общества, таких как, изменение климата, профилактику заболеваний и защиту дикой природы [7]. Таким образом, извлечение информации из слабоструктурированных и неструктурированных данных является важной задачей в современном мире.

Результаты исследований, приведенные в статье, получены в рамках выполнения грантов РФФИ 18-07-00193, 19-07-00709 и государственного задания № FEUE-2020-0007.

ПОДХОДЫ К ИЗВЛЕЧЕНИЮ ИНФОРМАЦИИ ИЗ СЛАБОСТРУКТУРИРОВАННЫХ ДАННЫХ

Одним из подходов к извлечению информации является распознавание именованных сущностей (NER - Named Entity Recognition). К основным типам сущностей относятся люди, организации, локации и даты. Может показаться, что для распознавания сущности достаточно выполнить поиск совпадения со значениями из некоторого словаря, но это не так, поскольку невозможно перечислить все имена собственные и варианты их написания, а также выдергивать сущность из контекста. Рассмотрим 2 высказывания «Мустай Карим родился 20 октября 1919 года» и «Главное событие отечественной навигации 2020 года — теплоход «Мустай Карим». В первом случае речь идет о поэте, во втором о теплоходе.

В литературе встречаются две группы методов NER: методы на основе правил и статистические подходы к обучению. Методы первой группы требуют создания множества правил, состоящих из паттерна и действия. Паттерн – это регулярное выражение, определенное для свойств токена (последовательности символов в документе). Действие может помечать один или несколько токенов. Далее текст проверяется на соответствие правилам. Набор правил можно задавать вручную или автоматически. Методы второй группы решают задачу распознавания именованных сущностей как задачу маркировки последовательностей. Последовательность – это предложение, слово в нем – наблюдение. Предсказание метки для наблюдения может ставиться не только исходя из значения самого наблюдения, но и других рядом стоящих наблюдений рассматриваемой последовательности, таким образом, текст делится на фрагменты с использованием BIO-нотации (beginning - inside - outside) [8]. Пример разбиения и его последовательность меток NER:

Мустай	Карим	–	народный	поэт	Башкирской	АССР.
B-PER	I-PER	0	0	0	B-LOC	I-LOC

Другой подход к извлечению информации из слабоструктурированных и неструктурированных данных - извлечение отношений (RE - Relation Extraction). RE - это задача обнаружения и характеристики семантических отношений между сущностями в тексте. Пример извлечения отношения:

Мустай Карим – народный поэт Башкирской АССР.
Народный поэт чего (Мустай Карим, Башкирская АССР)

Одним из подходов к решению проблемы основан на интерпретации RE как задачи классификации типа отношения между двумя сущностями из одного предложения (Feature-based Classification). Другим подходом являются ядерные методы [9, 10], цель которых – определение степени схожести между объектами на основе отображения исходных нелинейных признаков в пространство более высокой размерности исходного, в котором они становятся разделяемыми.

Поскольку вышеописанные подходы требуют большое количество размеченных данных, существует подход к обучению со слабой разметкой данных (weakly supervised learning) [11]. Идея подхода заключается в применении слабых меток (детерминированные функции, эвристические правила, выходные данные других классификаторов, дистанционное наблюдение [12]), ограничения, распределение вероятности, инвариантности для снижения человеческих трудозатрат на разметку данных.

Отсутствие обучающих выборок и большие затраты на их формирование для автоматического построения модели являются серьезным вызовом в обработке неструктурированных данных. В связи с этим усилия исследователей сосредоточены на разработке методов извлечения информации без учителя (Unsupervised Information Extraction) [13]. В основе этого подхода лежит поиск ассоциированных пар сущностей (идентификация отношений) и их дальнейшая кластеризация (характеризация отношений).

ЗАКЛЮЧЕНИЕ

В данной статье приведены подходы к извлечению информации из слабоструктурированных данных. В рамках этих подходов существует немало методов, однако пока не существует решения, конкурирующего в понимании текста человеком. Сегодня задача извлечения информации актуальна как никогда. Ее решение осложнено отсутствием обучающих данных. Хотя усилия ученных и специалистов в данной области сосредоточены на поиске выхода из положения, подготовка обучающих выборок является не менее сложной задачей.

СПИСОК ЛИТЕРАТУРЫ

1. Gartner Glossary [Электронный ресурс]. — Режим доступа: URL: https://www.gartner.com/en/information-technology/glossary/big-data (20.02.2021).

2. Maria Nancy, R Maheswari, "A review on unstructured data in medical data" // JOURNAL of CRITICAL REVIEWS 7: 2020, 1

3. Shubham Jain, Amy de Buitléir, Enda Fallon, "A Review of Unstructured Data Analysis and Parsing Methods" // Proceedings of the IEEE International Conference on Emerging Smart Computing and Informatics (IEEE – ESCI 2020) Scopus, Web of Science Journal Publication

4. W. Xu, L. Huang, A. Fox, D. Patterson, and M. I. Jordan, "Detecting large-scale system problems by mining console logs" // Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles.ACM:,2009, 117–132.

5. Q. Fu, J.-G. Lou, Y. Wang, J. Li, "Execution anomaly detection in distributed systems through unstructured log analysis" // Proceedings of the 2009 Ninth IEEE International Conference on Data Mining. IEEE, 2009.

6. S. He, J. Zhu, P. He, M. R. Lyu, "Experience report: System log analysis for anomaly detection" // Proceedings of the 2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE). IEEE, 2016.

7. Gartner Top 10 Trends in Data and Analytics for 2020 [Электронный ресурс]. — Режим доступа: URL: https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020 (25.02.2021).

8. Lance A. Ramshaw, Mitch P. Marcus, "Text chunking using transformation-based learning" // Proceedings of the 3rd Workship on Very Large Corpora, 1995, 82–94.

9. D. Zelenko, C. Aone, A. Richardella, "Kernel methods for relation extraction" // Proceedings of the Conference on Empirical Methods in Nat. Language Processing (EMNLP), Philadelphia, 2002, 7.

10. R. Bunescu, R. Mooney, "A shortest path dependency kernel for relation extraction" // Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, 1995, 724–731.

11. A. Ratner, S. Bach, P. Varma, C. Ré, "Weak Supervision: The New Programming Paradigm for Machine Learning" [Электронный ресурс]. — Режим доступа: URL: https://dawn.cs.stanford.edu/2017/07/16/weak-supervision (25.02.2021).

12. M. Mintz, S. Bills, R. Snow, D. Jurafsky, "Distant supervision for relation extraction without labeled data" // Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, 1003–1011.

13. B. Hachey, "Multi-document Summarisation Using Generic Relation Extraction" // Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 2009 Conference on Empirical Methods in Natural Language Processing, 2009.

14. Грант С. Ингерсолл, Томас С. Мортон, Эндрю Л. Фэррис Обработка неструктурированных текстов. Поиск, организация и манипулирование. / Пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2015.

15. Charu C. Aggarwal, ChengXiang Zhai Mining Text Data. / Charu C. Aggarwal, ChengXiang Zhai – N.Y., United States: Springer-Verlag New York, 2012.

ОБ АВТОРЕ

МОЛОКОВИЧ Ольга Александровна, аспирант 1-го курса ФИРТ.

METADATA

Title: Approaches to extracting information from semi-structured data.

Author: O. A. Molokovich

Affiliation: Ufa State Aviation Technical University (UGATU), Russia.

Email: o.molokovich@ya.ru

Language: Russian.

Source: Molodezhnyj Vestnik UGATU (scientific journal of Ufa State Aviation Technical University), no. 2 (25), pp. 64-66, 2021. ISSN 2225-9309 (Print).

Abstract: The article describes the sources of big data, their division into types and provides an overview of methods for processing semi-structured data, such as recognizing named entities and extracting relations.

Key words: semi-structured data; big data; data mining; natural language processing.

About author:

MOLOKOVICH, Olga Alexandrovna, postgraduate student 1 year, Ufa state aviation technical University.