УДК 004.91

Обзор методов выделения иерархии, структурирования и семантического поиска в документах

Подморин Д.О (ИТМО)
Научный руководитель – аспирант, инженер Ковальчук М.А. (ИТМО)

Введение. В современном мире объем электронного документооборота растет в геометрической прогрессии. Бесконтрольный рост электронных документов, например, в рабочей почте может приводить к долгому поиску необходимой информации или ее потере, что ведет к понижению эффективности при решении задач. Сложности также возникают при работе с отчетами и спецификациями. Структурная сложность этих документов может затруднить процесс анализа и поиска необходимых данных. Нередко ценная информация разбросана по различным отчетам и спецификациям, что делает ее обнаружение трудоемким и неэффективным. Помимо этого, размер документов также растет и их анализ становится сложной задачей для человека.

Для решения этой проблемы необходимо создать систему, способную работать с большими текстовыми документами, находить в них структуру и иерархию. Такая система способна дать представление об общем содержании и стиле документа, а также выделить необходимые структуры [1] и осуществлять семантический поиск. При создании такой системы возникает ряд задач, которые необходимо решить:

- Скорость работы;
- Сложность сохранения внимания на большом объеме данных;
- Работа с разнородными текстами.

Создание такой системы позволит существенно сократить время поиска информации и повысить эффективность компании.

Основная часть. В рамках решения проблемы анализа больших текстовых документов существует два подхода, которые стоит рассмотреть:

1. Неконтролируемое извлечение данных с применением графического интерфейса на основе программы SIMX TextConverter. Данный подход позволяет найти текстовые сущности с похожим форматированием(паттерны), обнаружить их во встречающемся потоке данных и выявить иерархическую структуру документа. При работе с финансовыми документами такой подход позволяет получить эффективность 95% [2].
2. Синтаксический анализ представления дискурса. Данный подход использует трансформерную модель(кодер-декодер) и представление текстового документа в древовидной структуре с шестью типами узлов. Узлы представляют собой различный компоненты, используемые для моделирования семантической структуры документа. По метрикам exa-f1(полное соответствие шаблону) и part-f1(частичное соответствие шаблону) модель показала значения 69.45 и 75.89 соответственно [3].

Выводы. Проблема снижения эффективности структурирования и семантического поиска наблюдается при увеличении объема текстового документа. Создание системы, сохраняющей эффективность на больших электронных документах, позволит уменьшить затраты на поиск информации.

Список использованных источников:

1. L. Wei, D. Hu, W. Zhou, X. Tang, X. Zhang, X. Wang, J. Han, S. Hu, Hierarchical interaction networks with rethinking mechanism for document-level sentiment analysis, in: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Springer, 2020

2. Vladimir Bernstein, Andrei Afanassenkov. Unsupervised Data Extraction from Computer-generated Documents with Single Line Formatting. arXiv:2007.07082 (2020).
3. Jiangming Liu, Shay B. Cohen, Mirella Lapata (2019). Discourse Representation Parsing for Sentences and Documents.

Подморин Д.О. (автор) Подпись _____

Ковальчук М.А. (научный руководитель) Подпись _____