К. Ю. Володько
(*ГрГУ им. Я. Купалы, Гродно*)

ОСОБЕННОСТИ ПРОЕКТИРОВАНИЯ СИСТЕМЫ ПОИСКА ПО НЕСТРУКТУРИРОВАННЫМ ДАННЫМ

В отчете IDC «Расплата за невозможность обнаружить информацию» («The High Cost of Not Finding Information», 2003) было отмечено, что на средних предприятиях прямые убытки, вызванные потерей времени из-за неудобства работы с информацией, в пересчете на одного работающего оцениваются в 2,5–3,5 тыс. долл. [1].

Таким образом главная задача системы состоит в том, чтобы превратить неструктурированные данные в цифровую интеллектуальную, связанную информацию, чтобы ее можно было быстро искать и находить, она была надежной и точной.

На текущий момент основной идеей, применяемой при поиске по таким данным, является получения уровня похожести строк, на основе их сравнения. Для поиска по тексту, нахождения опечаток, такой вариант является удачным. Уровень схожести между «инструмент» и «иснтрумент» будет несомненно высоким. Однако, недостатком является то, что такая система не увидит разницы между «45м» и «45т», что может быть критичным для промышленных производств, где необходимо быстро найти подходящую деталь по нужным критериям в соответствующей системе ERP.

Применение интеллектуальных методов обработки информации, а именно типизации и классификации, перед загрузкой данных позволяет повысить эффективность процесса производства и сократить время поиска информации. Основой классификации является разделение описания детали на токены и определение типа токена из таких вариантов как: код, число + единицы измерения или только текст. Это помогает обеспечить больший уровень уверенности в том, что «45м» и «50м» являются схожими, чем «45м» и «45т». Для выбора инструментов автоматизации были выбраны основные требования: интеграция с существующими системами, цена и возможность решения всех выявленных проблем.

Литература

1 Аналитика неструктурированных данных / Открытые системы. СУБД [Электронный ресурс]. – 2012. – № 06. – Режим доступа: https://www.osp.ru/os/2012/06/13017038. – Дата доступа: 14.01.2021.