Д. С. Бухаров, *канд. техн. наук, филиал ОАО «СО ЕЭС» «Региональное диспетчерское управление энергосистемы Иркутской области», г. Иркутск, bukharovds@gmail.com*

О поиске эквивалентных текстов

В статье описан подход к формированию поискового множества, используемого при определении эквивалентов текста. Задача такого вида возникает при поиске дубликатов текста, определении авторства и возможного плагиата, организации библиотечного поиска, а также при создании поисковых систем Интернета. В подходе, представленном в статье, учитывается ряд особенностей: частотность слов, пунктуация, морфемная структура слов, регистр букв и артефакты текста (специфические цифро-буквенные сочетания). Разработанная программа протестирована на наборе данных, в число которых включены как оригиналы текстов, так и их специальным образом модифицированные варианты. В результате проведенного эксперимента определены слабые стороны подхода. Приведены варианты по улучшению разработанного программного средства и схема взаимодействия модулей разработанной программы после модификации.

Ключевые слова: поиск эквивалента, поисковое множество, сравнение текстов, библиотечный поиск, поиск плагиата.

Введение

Актуальные задачи поиска эквивалентных текстов в настоящее время — «библиотечный поиск» и «поиск плагиата». Под поиском эквивалентного текста подразумевается определение текста, максимально подобного некоторому запросу или другому тексту. Формализация такой задачи — сложный процесс, так как сравнение выполняется на естественном языке, который содержит в себе множество аспектов, слабо поддающихся математико-алгоритмическому описанию.

Для поиска эквивалентных текстов разрабатываются различные подходы, обеспечивающие эффективное решение в рамках поставленной задачи. Перспективное направление — построение универсального метода решения, однако сейчас это проблематично в силу слабой формализации «поисковой» задачи. В работах, посвященных поиску подобных текстов, особое внимание уделяется морфологическим, лексическим, семантическим и синтаксическим особенностям сравниваемых экземпляров текста.

В работах [1; 2] применяется наиболее популярный подход к поиску дубликатов текстов, основанный на построении шинглов — специальных наборов слов, составляющих в некотором смысле «маску» текста, используемую в процессе сравнения.

М. Г. Крейнесом [3] дано описание технологии поиска «ключи от текста», позволяющей определить множество слов, наиболее сильно связанных по смыслу. Такое семантическое объединение позволяет построить «смысловой портрет» текста, существенно влияющий на результат поиска.

В статье А. П. Колосова [4] представлено применение концептуальных графов в поисковых системах: между парой слов устанавливается отношение, несущее смысловую нагрузку. Данный подход схож с объектным подходом [5], позволяющим учесть семантическую составляющую не одной фразы или предложения, а некоторого блока текста.

В работе [6] приведен обзор компьютерных программ сравнения текстов (Лингвоанализатор, Атрибутор, СМАЛТ, Стилеанализатор, Авторовед), позволяющих прово-