МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ ПОЛІТЕХНІЧНИЙ УНІВЕРСИТЕТ

ІНСТИТУТ КОМП'ЮТЕРНИХ СИСТЕМ

МАТЕРІАЛИ ДЕВ'ЯТОЇ МІЖНАРОДНОЇ НАУКОВОЇ КОНФЕРЕНЦІЇ СТУДЕНТІВ ТА МОЛОДИХ ВЧЕНИХ

присвячена 55-річчю
ІНСТИТУТУ КОМП'ЮТЕРНИХ СИСТЕМ

" Сучасні інформаційні технології 2019 "

" Modern Information Technology 2019 "

23-24 травня

Одеса
«Екологія»
2019

УДК 004.912

ПОВЫШЕНИЕ СКОРОСТИ ПОИСКА ДЕФИНИЦИЙ ТЕРМИНОВ В ИНТЕРНЕТЕ

Иванченко М. Е., Мельникова В. Е.
к.т.н., профессор. ИКС Кунгурцев А. Б.
Одесский Национальный Политехнический Университет, УКРАИНА

АННОТАЦИЯ. В работе рассматривается проблема поиска дефиниций терминов в интернете и подчеркивается важность быстрого и эффективного их поиска. Присутствует анализ характеристик и минусов уже имеющихся подобных продуктов. Разработан и описан алгоритм работы разрабатываемого программного продукта.

Введение. Поиск определенных терминов является задачей, которая часто встречается в разных областях деятельности. Частным случаем этой задачи является определение дефиниций терминов для словарей предметных областей [1] , используемых во всех программных проектах, выполняемых под заказ. Определение дефиниций терминов – трудоемкий процесс, который выполняется экспертом в определенной предметной области (ПО). Поэтому исследования, направленные на автоматизацию поиска толкований весьма актуальны.

Цель работы. Сократить время поиска толкований терминов, приблизительно на 30%, за счет автоматизированной фильтрации текстов, основанной на анализе вхождении в них терминов из рассматриваемой предметной области.

Основная часть работы. В качестве вариантов решений проанализирована возможность использования нескольких популярных сайтов [2,3,4].

- Сайт «Википедия» - данный ресурс при поиске формирует большую статью, содержащую много информации, не относящейся к толкованию искомого термина.

- Сайт «Академик» - содержит краткую выборку по толкованию терминов, но не содержит фильтра по предметным областям, хотя обладает необходимым для этого базисом.

- Сайт «Словопедия» – отличается очень слабой дружелюбностью интерфейса пользователя, а также не решает проблемы, указанные для Википедии и Академика.

Исходя из приведенного анализа, принято решение учитывать ПО определяемого термина, фильтровать информацию, полученную в результате поиска толкования, создать интуитивно понятный интерфейс. Ниже приведен соответствующий алгоритм.

1. Из исходного текста Т1 выделяются термины [1].

2. Термин поступает в систему. Если система уже содержит словари предметных областей, то производится поиск толкования термина в этих словарях. Если толкование найдено в некотором словаре СПО1, то происходит переход к пункту 3, иначе – к пункту 4.

3. Если 50% терминов (уточняется экспертом) из Т1 также найдены в СПО1, то найденное толкование представляется как результат поиска. Алгоритм поиска завершается.

4. Формируется запрос на сайты-источники: Академик [2] и Wikipedia [4]. Выполняется парсинг полученной информации. Если результат Р1 принадлежит сайту-источнику Wikipedia, происходит переход к пункту 5, если - сайту Академик или Словопедия - к пункту 6.

5. Из Р1 формируется массив структур ключ-значение, где ключ - название словаря, а значение - толкование. Далее анализируется массив данных структур. Если же ключ или значение содержат в тексте входящие в ПО корни, то данный ключ/значение добавляется в конечный результат и происходит переход к пункту 6, в противном случае структура удаляется

6. Выполняется преобразование результирующего массива структур пар ключ-значение в результат Т2. Полученный результат Т2 разбивается на приложения. Анализируются приложения. Если в приложении нет слов-ключей, указывающих на толкование, то данное приложение удаляется из результата Т2, В противном случае, происходит переход к пункту 7.

7. Завершение процесса формирования толкования.

Выводы. Реализация приведенного алгоритма позволила сократить время поиска толкований в среднем в 2 раза при сокращении объема предлагаемого текста до 60%. Данное решение может быть использовано для реализации различных задач поиска толкований при наличии определенного набора терминов (ключевых слов), определяющих ПО.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Kungurtsev O. Development of information technology of term extraction from documents in natural language / O. Kungurtsev, S. Zinovatnaya, Ia. Potochniak, M. Kutasevych // Eastern-European Journal of Enterprise Technologies. Vol 6, No 2 (96) (2018). pp. 44-51.

2. Академик.URL:https://dic.academic.ru(дата обращения: 17.04.2019);

3. Википедия.URL:https://ru.wikipedia.org(дата обращения: 17.04.2019)

4. Словопедия.URL:http://www.slovopedia.com(дата обращения: 17.04.2019).