Использование нечетких онтологий в задачах кластеризации библиографической информации

А.А. Дырночкин
Ульяновский государственный технический университет
Ульяновск, Россия
dyrno4kin@gmail.com

В.С. Мошкин
Ульяновский государственный технический университет
Ульяновск, Россия
postforvadim@ya.ru

***Аннотация* — В данной статье представлен подход к кластеризации библиографической информации с использованием нечеткой онтологии. В качестве алгоритма векторизации библиографических данных использовалась модифицированная модель TF-IDF с учетом значений функций принадлежности (идентичности) между терминами предметной области. Также в работе представлены результаты экспериментов с данными по статьям из научной библиотеки elibrary.**

Ключевые слова— анализ текста, наукометрия, нечеткая онтология, кластеризация

1. Введение

Наукометрические данные в настоящее время являются ценной информацией для исследователей при решении задач оценки научной активности и формирования научных групп. Однако, наукометрические данные можно рассмотреть как ресурс для исследования, автоматизация анализа которого позволит эффективно расходовать ресурсы на выполнение данных задач.

Библиографическая информация является текстом и имеет слабую формализацию и структуру, вследствие чего алгоритмы извлечения, предобработки и дальнейший анализ данных ресурсов должны проводиться с использованием подходов NLP [5].

При выполнении запроса к документу обычной практикой является расширение набора понятий, уже присутствующих в запросе, другими понятиями, которые могут быть получены из онтологии. Как правило, при заданном концепте в запрос можно добавить его родительские и дочерние элементы, а затем выполнить поиск в документе.

Возможное использование нечеткой онтологии состоит в расширении запросов, помимо дочерних и родительских, экземплярами понятий, которые в определенной степени удовлетворяют запросу. Это особенно актуально при извлечении библиографической информации, где один и тот же объект может быть представлен различными терминами.

Поэтому подход, описанный в работе [1], был модифицирован с учетом возможности расширения запросов с применением нечетких онтологий.

2. Построение нечеткой онтологии предметной области

На рисунке 1 в общем виде представлен фрагмент нечеткой онтологии сложной проблемной области, формируемой из различных источников.

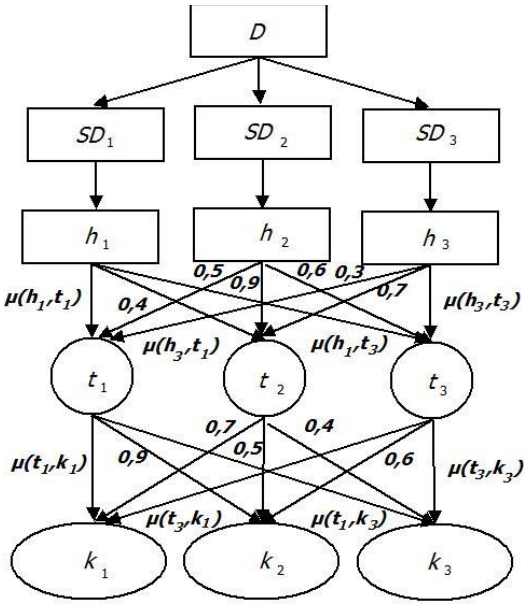

Рис. 1. Фрагмент нечеткой онтологии

Верхний (первый) уровень онтологии предполагает комплексное описание ПрО *D*, полученное путем слияния знаний из разных источников (экспертов, коллекций текстов). На втором уровне *D* декомпозируется на подобласти *SD*, при этом каждой подобласти соответствует определенный источник информации (h_i).

Между базовыми понятиями ПрО (t_i, где $i = [1..n]$) и лингвистическими переменными (k_j , где $j = [1 .. m]$), находящимися на четвертом и пятом уровнях соответственно, устанавливаются семантические отношения, и каждому отношению присваивается значение функции принадлежности $0 \le \mu(t, k) \le 1$, где $t \in T_i$, $k \in K_j$.

Помимо этого, в процессе построения нечеткой онтологии ПрО производится соотнесение категорий t_i ПрО с источниками h_i, из которых данные категории могли быть извлечены. Подобным образом формируется набор ключевых слов для каждого источника знаний, а также для каждой подобласти комплексной ПрО, и устанавливаются значения степеней принадлежности между ними, выражаемые числами из интервала [0, 1] ($\mu(h, t)$) [7].

3. Алгоритм кластеризации библиографических данных с использованием нечеткой онтологии

В рамках данного исследования был разработан алгоритм извлечения и анализа библиографической информации из информационной системы elibrary. Данный алгоритм включает следующие этапы:

1. ***Извлечение текстовых библиографических данных.*** Данные могут извлекаться как путем парсинга веб-страниц, так и с использованием собственного API

портала. Загружается следующая информация: авторы, аннотация, год издания, ключевые слова и. др. [2].

2. ***Предобработка загруженных статей.*** На рисунке 2 представлено описание схемы предобработки статей [3].

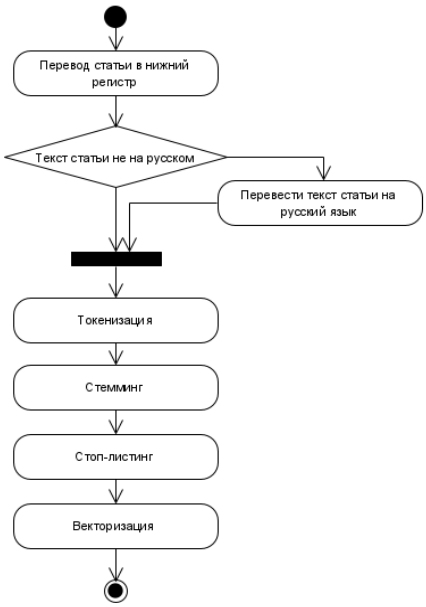

Рис. 2. Схема предобработки библиографической информации

3. ***Векторизация предобработанных статей***. На данном этапе происходит векторизация предобработанных документов методом TF-IDF [6] с учетом нечетких данных в разработанной FuzzyOWL-онтологии в соответствии со следующей моделью:

$$tf\_idf(w,d,D) = tf(w,d) \times idf(w,D) \times \mu(w,v) \quad (1)$$

где

$$tf(w,d) = \log(1 + f(w,d)) \quad (2)$$

где *f(w, d)* — частота слова w в документе d

$$idf(w,D) = \log \frac{N}{f(w,D)} \quad (3)$$

где N – количество документов в наборе данных;

f(w, D) – частотой слова w во всем наборе данных [7]

μ – значение функции принадлежности (идентичности) слов *w* и *v* в нечеткой онтологии, $\mu(w,v) \subseteq [0,1]$.

4. ***Кластеризация векторизованных текстов методом k-средних.*** Алгоритм кластеризации K-средних вычисляет центроиды и выполняет итерацию, пока не найдет оптимальный центроид [6]. В этом алгоритме точки данных назначаются кластеру таким образом, чтобы сумма квадратов расстояния между точками данных и центроидом была минимальной. Минимальное суммарное отклонение рассчитывается по формуле 4[3].

$$\min\left[\sum_{i=1}^{k}\sum_{x(j)\in Si}\left\|x^{(j)} - u_i\right\|^2\right] \quad (4)$$

где u_i - центроид для кластера S_i

Для проведения кластеризации была загружена библиографическая информация по сотрудникам УлГТУ, порядка 14 тысяч статей в формате: название и полное описание статей для проведения экспериментов. В результате кластеризации корпус статей был разделен на 12 кластеров. На рисунке 3 представлены результаты кластеризации.

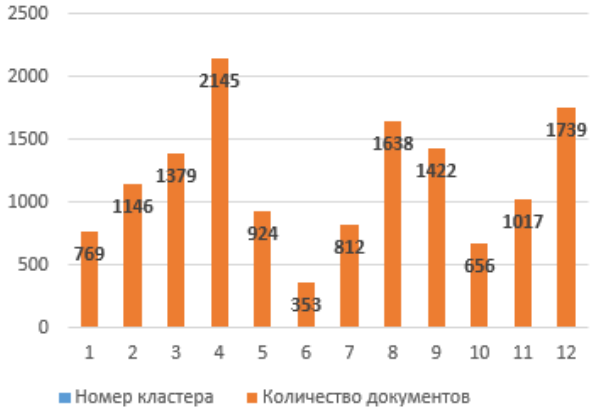

Рис. 3. Результаты кластеризации

4. Заключение

Таким образом, разработанная в рамках данного исследования программная система позволила формировать рекомендации при составлении научных групп с учетом возможной нечеткости в определении терминов, которые могут относиться к различным предметным областям. Достигнуто эта возможность была посредством разработки нечеткой предметной FuzzyOWL-онтологии, хранящей возможные нечеткие термины, извлекаемые из наукометрических баз таких, как elibrary, и расширения запросов при поиске данных.

Литература

[1] Dyrnochkin, A. Approach to extraction and clustering bibliographic information / A. Dyrnochkin, V. Moshkin // 2022 VIII International Conference on Information Technology and Nanotechnology (ITNT). – 2022. – P. 1-4. doi: 10.1109/ITNT55410.2022.98485

[2] Ингерсолл, Г. Обработка неструктурированных текстов. Поиск, организация и манипулирование / Г. Ингерсолл, Т. Мортон, Э. Фэррис. – Москва : ДМК Пресс, 2015. – 414 с.

[3] Кравченко, Ю. А. Векторизация текста с использованием методов интеллектуального анализа данных / Ю. А. Кравченко, А. М. Мансур, М. Ж. Хуссайн // Известия Южного федерального университета. Технические науки. – 2021. – Т. 2, № 219. – С. 154-167.

[4] Пархоменко, П. А. Обзор и экспериментальное сравнение методов кластеризации текстов / П. А. Пархоменко, А. А. Григорьев, Н. А. Астраханцев // Труды Института системного программирования РАН. – 2017. – Т. 29, №. 2. – С. 161-200.

[5] Трубников, В. С. Проектирование системы сбора, анализа и визуализации наукометрических данных / В. С. Трубников, К. А. Туральчук // Проблемы современной науки и образования. – 2015. – Т. 6, №36.

[6] Юферев, В. И. Векторизация текстов на основе word-embedding моделей с использованием кластеризации / В. И. Юферев, Н. А. Разин // Моделирование и анализ информационных систем. – 2021. – Т. 28, №. 3. – С. 292-311.

[7] Ярушкина, Н.Г. Применение способа интеграции нечетких временных рядов и нечѐтких онтологий в задачах диагностики технических систем / Н.Г. Ярушкина, В.С. Мошкин, Г.Р. Ишмуратова, И.А. Андреев, И.А. Мошкина // Онтология проектирования. – 2018. – Т.8, №4(30). – С.594-604. DOI: 10.18287/2223-9537-2018-8-4-594-604.