УДК 004.49

Ю. В. Алферов, С. А. Бородин*

МЕТОД ФОРМАЛИЗАЦИИ НЕЧЕТКИХ КОЛЛОКАЦИЙ ТЕРМОВ В ТЕКСТАХ НА ОСНОВЕ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ

Анализ текстовой информации является важнейшей отраслью современной *IT*-индустрии. Выявление скрытых семантических структур позволяет не только качественнее осуществлять информационный поиск, кластеризацию и классификацию текстовых коллекций, но и более эффективно проектировать архитектуру программных компонентов искусственного интеллекта в рамках группы *IT*-технологий *Big Data*.

Подход, основанный на частотной оценке термов, на данный момент, является важнейшим атрибутом систем анализа текстовой информации. Вместе с тем, все чаще в литературе [1] встречается концепция дополнения частотной векторно-пространственной модели текстовых документов оценкой семантической значимости коллокаций – наборов термов с заданным взаимным расположением.

Одним из направлений развития исследований коллокаций стали так называемые нечеткие коллокации [2]. Они представляют собой группы термов, расстояние между которыми формализуется посредством функции принадлежности. Под расстоянием между термами в коллокации понимается число слов в тексте между ними. Классический взгляд на коллокации предполагает, что термы ее составляющие появляются непосредственно рядом друг с другом.

Вместе с тем, было отмечено [3], что наличие некоторого количества термов между формирующими коллокацию не приводит к обязательной утрате семантической значимости последней. Таким образом, возникли следующие вопросы:

- какой диапазон расстояний считать приемлемым;
- меняется ли семантическая значимость с изменением расстояния в рамках выбранного диапазона?

Ответом стал подход, формализующий расстояние с помощью функции принадлежности. Естественность этого подхода заключается в том, что, по сути, мы задаем конкретное расстояние между термами, но формализуем его посредством нечеткого числа. Форма нечеткого числа и определяет как диапазон допустимых значений расстояний между термами, так и колебания семантической значимости в зависимости от изменения расстояния между термами.

* Работа представлена в отборочном туре программы У.М.Н.И.К. 2016 г. в рамках Одиннадцатой межвузовской научной студенческой конференции Ассоциации «Объединенный университет им. В. И. Вернадского» «Проблемы техногенной безопасности и устойчивого развития» и выполнена под руководством канд. техн. наук, ст. преподавателя ФГБОУ ВО «ТГТУ» Д. В. Полякова.

Исследование нечетких коллокаций позволило построить обобщенную векторно-пространственную модель текстовой коллекции [4]. Показано, что в рамках этой модели текстовые документы могут быть формализованы векторами, элементами которых являются как частотные характеристики термов, так и их аналоги для нечетких коллокаций. При определенных параметрах элементы, соответствующие термам, принимали классический вид *tf-idf*. Данная оценка семантической значимости является стандартной и зарекомендовала себя, как отмечается в литературе [1, 4], частотной характеристикой термов.

При *SVD*-разложении матрицы, формализующей текстовую коллекцию и состоящей из значений *tf-idf* для термов и ее аналогов для коллокаций, нормированные значения сингулярных чисел представляют собой оценку семантической значимости каждого объекта. Этот подход позволяет сравнивать семантическую значимость коллокаций и термов, что повышает адекватность предлагаемой оценки.

Был разработан метод выявления нечетких коллокаций с треугольными и трапециевидными нечеткими числами для формализации расстояния. Вследствие чего спроектирован и поставлен вычислительный эксперимент.

Для проведения эксперимента была выбрана текстовая коллекция, состоящая из подборки номеров Журнала «Радио» (Издательство журнала «Радио») за период с 1949 по 1994 года. Объем исследуемой текстовой коллекции составил 453 документа, включающих в себя 13012 терма. Важно отметить, что под термами здесь понимаются лемманизированые, т.е. приведенные к единой форме слова.

В результате были получены коллокации, представленные в табл. 1.

Важно отметить, что выражения, представленные в столбце «Функция» данной таблицы, отражают лишь непараллельную оси абсцисс часть функции принадлежности. Так, если обозначить эту часть $f(x)$, соответствующая ей функция принадлежности будет иметь вид:

$$\mu(x) = \max\{0, \min\{1, f(x)\}\}.$$

Важно отметить, что выявленные в ходе вычислительных экспериментов семантически значимые коллокации не являются побочным эффектом семантической значимости термов. Так, выявленные в том же вычислительном эксперименте семантически значимые термы представлены в табл. 2.

1. Выявленные в ходе вычислительного эксперимента коллокации с наибольшей семантической важностью

Терм 1	Терм 2	Функция	Значение нормированного сингулярного числа ($\cdot 10^{-6}$)
Программа	Символ	$-0{,}167x + 1{,}667$	2,602
Работа	Датчик	$-0{,}25x + 2{,}25$	2,319

Программа	Системный	$-0{,}143x+1{,}29$	2,062
Работа	Командир	$-0{,}143x+1{,}286$	1,857
Транзистор	Испытывать	$-0{,}167x+1{,}667$	1,455
Система	Информационный	$-0{,}143x+1{,}429$	1,290
Связь	Документ	$-0{,}125x+1{,}25$	1,263

2. Выявленные в ходе вычислительного эксперимента термы с наибольшей семантической важностью

Терм	Значение нормированного сингулярного числа ($\cdot 10^{-4}$)
Автоматизировать	6,238
Аппаратура	3,393
База	2,327
Блок	1,835
Бюрократ	1,333
Военный	1,252

Как видно из представленной таблицы, термы с наибольшими значениями оценок семантической значимости не входят в аналогичную группу коллокаций. Это позволяет говорить о наличии нечетких коллокаций, как семантических факторов текстового документа.

К недостаткам данного вычислительного эксперимента относится существенное ограничение на вид функций принадлежности. Они в постановке данного эксперимента могли иметь вид только линейного сплайна.

Вместе с тем, нет никаких оснований считать, что аппроксимация функции принадлежности линейным сплайном дает нечеткие коллокации с наибольшей семантической значимостью.

Таким образом, возникает важнейшая задача уточнить вид функций принадлежности и, тем самым, повысить семантическую значимость выявленных нечетких коллокаций.

Из-за крайне сложного вида функции полезности (*Fitness*) и невозможности полного перебора на множестве функций для решения поставленной задачи в качестве метода оптимизации были выбраны генетические алгоритмы.

Вместе с тем, кроме сложного вида *Fitness* в рассматриваемой задаче отличается высокой вычислительной сложностью (пройти по всем текстам и осуществить *SVD*-разложение). С другой стороны, возможно одновременное вычисление данной функции у числа нечетких коллокаций, сопоставимого с количеством документов. Это означает, что необ-

ходимо выбрать генетический алгоритм, сходящийся на рассматриваемом семействе задач за наименьшее число поколений.

Заметим, что функция, к которой стремится популяция хромосом генетического алгоритма, уже существует и чем ближе наилучшее решение некоторого поколения к оптимальной функции, тем выше значение *Fitness*.

Это означает, что если рассмотреть аналогичное семейство задач, но с простым вычислением *Fitness*, это позволит провести вычислительные эксперименты и выявить вид генетического алгоритма, решающего задачу оптимизации функции принадлежности нечеткой коллокации в среднем за наименьшее число поколений.

Список литературы

1. *Иванова*, *О. Г.* Кластеризация текстовых коллекций на основе нечеткого описания коллокаций / О. Г. Иванова и [др.] // Информация и безопасность. – Воронеж : Издательско-полиграфический центр Воронежского государственного университета. – 2011. – № 3. – С. 459 – 462.

2. *Поляков*, *Д. В.* Определение пертинентности результатов запроса с использованием нечеткой логики / Д. В. Поляков и [др.] // Приборы и системы. Управление, контроль, диагностика. – 2012. – № 3. – С. 29 – 33.

3. *Поляков*, *Д. В.* Формализация информационной потребности пользователя на основе нечеткой логики / Д. В. Поляков и [др.] // Приборы и системы. Управление, контроль, диагностика. – 2012. – № 3 – С. 47 – 50.

4. *Поляков*, *Д. В.* Оценка семантической значимости нечетких коллокаций на основе обобщенной векторно-пространственной модели текстовой коллекции / Д. В. Поляков и [др.] // Прикаспийский журнал: Управление и высокие технологии. – 2016. – № 1(33). – С. 167 – 183.

Кафедра «*Информационные системы и защита информации*»
ФГБОУ ВО «*ТГТУ*»