полезных моделей РФ 10.10 2011, срок действия патента истекает 25 февраля 2021 г.

11.Прохода И.А., Морозова Е.П. «Лечебно-профилактический препарат из трутневых личинок, обладающий иммуномодулирующим действием» № 2473355, заявка № 2011150581, приоритет изобретения 12 декабря 2011 г, зарегистрирован в Государственном реестре изобретений РФ 27.01.2013, срок действия патента истекает 12 декабря 2031г.

12. Prokhoda I.A., Eliseeva,E.V., Katunina, N.P Quality Management of the Apiproduct from the Drone Larvae / IOP Conference Series: Earth and Environmental Science, electronic resource. IOP Publishing Ltd, 2019

13. .Prokhoda I.A., Stratienko, E.N., Katunina, N.P., Kukhareva, O.V., Tseeva, F. N.Creating Functional Foodstuffs from High-Technological Larval Raw Materials / IOP Conference Series: Earth and Environmental Science, electronic resource. IOP Publishing Ltd, 2019

14. .Prokhoda I.A., Eliseeva,E.V., Polesskaya O.P Management of the life cycle of the innovation apiprodacts from drone larvae and its introductions in the food industry / IOP Conference Series: Earth and Environmental Science, electronic resource. IOP Publishing Ltd, 2019

УДК 004.6
ГРНТИ 20.53.19

О НАХОЖДЕНИИ НЕСХОДСТВА МЕЖДУ ТЕМАТИКАМИ СТАТЕЙ

Решетников Александр Дмитриевич
Аспирант кафедры Вычислительной Математики и Прикладных Информационных Технологий,
ФГБОУ ВО «Воронежский государственный университет»,
Россия, г. Воронеж
Леденева Татьяна Михайловна
Профессор, доктор технических наук,
заведующая кафедрой Вычислительной Математики и Прикладных Информационных Технологий
факультета Прикладной Математики, Информатики и Механики
ФГБОУ ВО «Воронежский государственный университет»,
Россия, г. Воронеж

ABOUT FINDING DIFFERENCES BETWEEN ARTICLE TOPICS

Reshetnikov A.D.
Postgraduate student,
Department of Computational Mathematics and Applied Information Technologies,
Faculty of Applied Mathematics, Informatics and Mechanics,
Voronezh State University.
Ledeneva T.M.
Doctor of Technical Science, Professor,
Department of Computational Mathematics and Applied Information Technologies,
Faculty of Applied Mathematics, Informatics and Mechanics,
Voronezh State University.

АННОТАЦИЯ

Данная работа посвящена проблеме поиска статей по интересующей пользователя тематике. Основной проблемой можно назвать тот факт, что существуют разные виды алгоритмов, некоторое из которых учитываю семантическую нагрузку текста, а некоторые предназначены для синтаксического анализа. При подборе схожих публикаций, исследователя интересует семантическое подобие. В предложенном подходе мы остановимся на обработке ключевых слов, поскольку авторы этих статьи стараются вынести в эту секцию термины, отражающие идею своей публикации. Подготовив словари для целевых тематик и проведя предварительную обработку текстов, можно получить меру сходства/несходства между двумя статьями. Используя полученную оценку, можно набрать выборку, основанную на близости к оригинальной работе. Результаты данного подхода продемонстрированы на модельном примере.

ABSTRACT

This work is devoted to the problem of finding articles on topics of interest to the user. The main problem can be called the fact that there are different types of algorithms, some of which take into account the semantic load of the text, and some are intended for syntactic analyze. When selecting similar publications, the researcher

is usually interested in semantic similarity. In the proposed approach, we will focus on the processing of keywords, since the authors of these articles try to put into this section terms that reflect the idea of their publication. Having prepared dictionaries for target topics and pre-processing the texts, you can get a measure of similarity / dissimilarity between the two articles. Using the resulting estimate, you can dial a sample based on proximity to the original work. The results of this approach are demonstrated on a model example.

Ключевые слова: мера несходства, косинусное сходство, публикации.

Key words: measure of dissimilarity, cosine similarity, publications.

Введение

На сегодняшний день поиск сходства между текстами имеет большое практическое применение. Одним из вариантов является обнаружение и классификация научных статей по заданным тематикам. Научное сообщество постоянно развивается и благодаря развитию технологий мы можем получить доступ к огромному количеству материалов, не выходя из дома. Так, научная электронная библиотека eLIBRARY.RU содержит более 29 млн научных статей и публикаций и это число постоянно растет. Ситуации, когда необходим поиск статей по тематике не являются редкостью, поскольку огромное количество материалов требует чрезмерно длительного и неэффективного анализа. Ведь действительно, зная, какие тексты максимально близки к искомой тематике можно сэкономить ценнейший ресурс – время.

Для сравнения текстов в настоящее время существует множество различных подходов, которые основаны на семантическом или синтаксическом сходстве [1]. Неотъемлемым атрибутом каждой статьи являются ключевые слова. Автор старается вынести в данный пункт высокоуровневое описание, тематику публикации. Таким образом, сравнивая наборы этих ключевых слов у разных статей, можно говорить о их сходстве/несходстве между собой. Для решения данной проблемы воспользуемся предложенным ниже алгоритмом.

1. Понятие меры несходства и ее свойства

В первую очередь, понятие меры несходства между двумя объектами определим следующим образом: пусть a и b два объекта из E. Тогда *мера несходства* $d(a,b)$ это функция, которая удовлетворяет следующим условиям [2]:

1) $d(a,b) = d(b,a)$;
2) $d(a,a) = d(b,b) > d(a,b)$ для $\forall\, a \neq b$;
3) $d(a,a) = 0$ для $\forall\, a \in E$.

Метрикой назовем меру несходства, которая также удовлетворяет свойствам

4) $d(a,b) = 0 \Rightarrow a = b$;
5) $d(a,b) \leq d(a,c) + d(c,b)$ для $\forall\, a,b,c \in E$.

Ультраметрикой называется метрика, которая удовлетворяет следующему свойству

6) $d(a,b) \leq Max\{d(a,c), d(c,b)\}$ для $\forall\, a,b,c \in E$.

Матрицей расстояний для множества объектов $a_1, \dots, a_m$ в E назовем матрицу D с элементами $d(a_i, a_j), i,j = 1, \dots, m$. Если элементы этой матрицы монотонно увеличиваются по мере удаления от диагонали (по столбцам и строкам), то она называется матрицей Робинсона.

2. Постановка задачи

Пусть имеется некое U – универсальное множество слов, причем в это множество включены только те слова, которые используются в словарях. Под словарем в данной статье будем понимать вектор $S = (s_1, \dots, s_m)$, где m – количество элементов в словаре, а $s = (value, weight)$ или $s = (value)$, который характеризует определённую тему. Данный вектор может быть, как упорядочен по не возрастанию (иметь веса для каждого из элементов), так и быть неупорядоченным (без весов). Нужно отметить, что словари могут иметь различный уровень значимости.

В качестве входных данных будем использовать вектора, содержащие ключевые слова из статей. Перед непосредственным анализом, необходимо выполнить предобработку этих входных данных, которая включает удаление синонимов, путем приведения слова к той форме, которая находится в универсальном множестве слов (например: если в векторе “lsa” или “семантический анализ”, а в U “латентно-семантический анализ”). Также необходимо исключить слова, не принадлежащие ни одному из словарей.

Следующий этап заключается в построении вектора D, элементами которого является частоты появления слов из вектора ключевых слов в словарях, либо суммы вероятностей в том случае, когда элементы словаря имеют весовые коэффициенты и упорядочены.

Последним шагом является вычисление расстояния между двумя векторами, которое будет характеризовать степень сходства/несходства между текстами. В данной статье мы будем пользоваться косинусным сходством.

Косинусное сходство – это мера сходства между двумя векторами предгильбертового пространства, которая используется для измерения косинуса угла между ними [3]. Косинусное сходство двух документов изменяется в диапазоне от 0 до 1.

$$S_{cos}(A,B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2}\sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

3. Иллюстративный пример

Для того, чтобы использовать вышеописанный подход на практике, составим выборку из десяти статей, которые будут относиться к одной из двух основных тем, по пять статей из каждой. В первую очередь, выполним построение словарей и

проведем предварительную обработку входных данных.

Словари:

Словарь 1: Тематика – “сравнение текстов”. Вес словаря возьмем равный 0.6.

$S_1 =$ [(сравнение текстов,0.18);(текстовая близость,0.15);(семантика,0.14);

(мера сходства,0.11);(text mining,0.09);(текстовые пассажи,0.08);

(латентно-семантический анализ,0.07);(текстовый документ,0.06);

(стемминг,0.05);(естественные языки,0.04);(обработка данных,0.03)]

Словарь 2: Тематика – “ нечеткие системы”. Вес словаря возьмем равный 0.3.

$S_2 =$ [(база знаний,0.20);(экспертная система,0.17);(нечеткие правила,0.15);

(нечеткий регулятор,0.12);(полнота и непротиворечивость базы нечетких правил,0.11);

(оптимизация баз знаний экспертных систем,0.10);(нечеткая система управления,0.08);(обработка данных,0.07)]

Словарь 3: Тематика – “ обработка данных”. Вес словаря возьмем равный 0.1.

$S_3 =$ [(обработка данных,0.38);(классификация,0.24);

(интеллектуальный анализ данных,0.17);(text mining,0.13);

(гибридный алгоритм,0.06);(матрицы,0.02)]

Наборы ключевых слов после обработки представлены ниже, в скобках, указано, что именно было сделано.

[классификация, текстовый документ, мера сходства] (рубрицирование → deleted);

[экспертная система, нечеткие правила] (экспертное решение → экспертная система; эффективность инвестиционных проектов → deleted);

[нечеткая система управления, полнота и непротиворечивость базы нечетких правил] (иерархическая система лингвистических правил → deleted);

[интеллектуальный поиск, латентно-семантический анализ, text mining, классификация] (интеллектуальный поиск → интеллектуальный анализ данных; семантический анализ → латентно-семантический анализ; методы классификации → классификация);

[интеллектуальный анализ данных, нечеткие правила, база знаний, экспертная система] (сложный объект → deleted; нечетко-продукционное правило → нечеткие правила; нечеткая нейронная сеть → deleted; диагностика → deleted);

[база знаний, экспертная система, нечеткие правила, оптимизация баз знаний экспертных систем] (редукция нечетких правил → нечеткие правила);

[нечеткая система управления, гибридный алгоритм, нечеткий регулятор, нечеткие правила] (автоматизированные системы управления → нечеткая система управления; классический регулятор → deleted; база нечетких правил → нечеткие правила);

[латентно-семантический анализ, стемминг] (латентно-семантический анализ текста → латентно-семантический анализ; сингулярное разложение → deleted);

[текстовая близость, сравнение текстов, текстовые пассажи] (семантические классы → deleted; представление семантических схем → deleted);

[латентно-семантический анализ, обработка данных, естественные языки, матрицы, семантика] (LSA → латентно-семантический анализ).

Множество всех слов представлено следующим списком:

U = [сравнение текстов; текстовая близость; семантика; мера сходства;

текстовые пассажи; латентно-семантический анализ; текстовый документ;

стемминг; естественные языки; обработка данных; база знаний;

экспертная система; нечеткие правила; нечеткий регулятор;

полнота и непротиворечивость базы нечетких правил;

оптимизация баз знаний экспертных систем; нечеткая система управления;

классификация; интеллектуальный анализ данных; text mining;

гибридный алгоритм; матрицы]

Построим вектора D для текстов 1 и 2 с учетов весов и без.

$$D_1 = (2/3; 0; 1/3);\ D_2 = (0; 1; 0)$$

$$D_{1w} = (0.17; 0; 0.24);\ D_{2w} = (0; 0.32; 0)$$

Вектора с весами необходимо нормировать с учетом значимости словарей

$$D_{1w_{norm}} = (\frac{0.17*0.6}{0.17*0.6+0.24*0.1}; 0; \frac{0.24*0.1}{0.17*0.6+0.24*0.1} = (0.809524; 0; 0.190476);$$

$$D_{2w_{norm}} = (0; 1; 0)$$

Последним шагом является расчёт косинусного сходства

$$S_{cos}(D_1, D_2) = \frac{\frac{2}{3} * 0 + 0 * 1 + \frac{1}{3} * 0}{\sqrt{\frac{4}{9} + 0 + \frac{1}{9}} * \sqrt{0 + 1 + 0}} = 0$$

$$S_{cos}\left(D_{1w_{norm}}, D_{2w_{norm}}\right) = \frac{0.809524 * 0 + 0 * 1 + 0.190476 * 0}{\sqrt{0.809524^2 + 0 + 0.190476^2} * \sqrt{0 + 1 + 0}} = 0$$

В таблицах 1 и 2 представлены расчёты по вышеописанному алгоритму. В скобках римскими цифрами указано, к какой тематике относится текст с точки зрения эксперта, где I – "сравнение текстов", а II – " нечеткие системы".

Таблица 1

Косинусное сходство без учета весов элементов словарей

	1(I)	2(II)	3(II)	4(I)	5(II)	6(II)	7(II)	8(I)	9(I)	10(I)
1(I)	x	0.0	0.0	0.9966	0.0091	0.0	0.0091	0.9965	0.9965	0.9923
2(II)	0.0	x	1.0	0.0	0.9938	1.0	0.9938	0.0	0.0	0.1236
3(II)	0.0		x	0.0	0.9938	1.0	0.9938	0.0	0.0	0.1236
4(I)	0.9966			x	0.0181	0.0	0.0181	0.9863	0.9863	0.9889
5(II)	0.0091				x	0.9938	0.9999	0.0	0.0	0.1319
6(II)	0.0					x	0.9938	0.0	0.0	0.1236
7(II)	0.0091						x	0.0	0.0	0.1319
8(I)	0.9965							x	1.0	0.9889
9(I)	0.9965								x	0.9889
10(I)	0.9923									x

Таблица 2

Косинусное сходство с учетом весов элементов словарей

	1(I)	2(II)	3(II)	4(I)	5(II)	6(II)	7(II)	8(I)	9(I)	10(I)
1(I)	x	0.0	0.0	0.9906	0.0248	0.0	0.013	0.9734	0.9734	0.9926
2(II)	0.0	x	1.0	0.0	0.9941	1.0	0.9983	0.0	0.0	0.1207
3(II)	0.0		x	0.0	0.9941	1.0	0.9983	0.0	0.0	0.1207
4(I)	0.9906			x	0.0389	0.0	0.0205	0.9330	0.9330	0.9837
5(II)	0.0248				x	0.9941	0.9986	0.0	0.0	0.1449
6(II)	0.0					x	0.9983	0.0	0.0	0.1207
7(II)	0.013						x	0.0	0.0	0.1336
8(I)	0.9734							x	1.0	0.9656
9(I)	0.9734								x	0.9656
10(I)	0.9926									x

Заключение

В данной работе был предложен алгоритм получения оценки сходства статей на основе наборов ключевых слов. Расчеты представлены с учетов весов элементов в словарях и без. Используя полученную оценку, можно набрать выборку из статей, основанную на близости к оригинальной работе. Кроме того, основываясь на данной оценке можно классифицировать тексты.

Список литературы

1. Решетников А.Д. О подходах для определения меры несходства в текстовых данных / А.Д. Решетников // Вестник Воронежского Института Высоких Технологий. – 2019. – №3. – С.35–38 [Reshetnikov A. D. About approaches to determining the measure of dissimilarity in text data // The Bulletin of the Voronezh Institute of High Technologies. – 2019. – №3. – p.35–38 (in Russ.)]

2. Billard, L Symbolic data analysis / L. Billard, E. Diday. – 1st ed. – Wiley, 2006. – 330 p.

3. Deza M. Encyclopedia of Distances / M. Deza, E. Deza. – 3rd ed. – Springer, 2014. – 733p.