Мера TF-IDF, согласование смысловых эталонов и взаимная релевантность документов тематического корпуса

Михайлов Д. В., Емельянов Г. М.

Новгородский государственный университет
имени Ярослава Мудрого

Всероссийская конференция с международным участием
«Математические методы распознавания образов» (ММРО-2021),

7–10 декабря 2021 г.

г. Москва

Текст научной работы и рациональная (эталонная) передача смысла

Требования к решению

1. Иерархизация источников информации по степени отражения наиболее существенных понятий изучаемой предметной области при максимальной компактности и безызбыточности изложения.
2. Эксперт не должен перефразировать текст для поиска семантически эквивалентных языковых форм описания единицы знаний.
3. Выделение набора единиц текста и их связей, отвечающих эталонному варианту описания представляемого фрагмента знаний.
4. В иерархии документов эталон вышестоящего должен доопределять эталон непосредственно связанного с ним нижестоящего.

Эталонной передаче смысла отвечает набор единиц текста и их связей, *необходимый и достаточный* для представления единицы знаний.

Аннотация и заголовок научной работы

1. Отражают основное содержание и наиболее значимые из полученных авторами результатов без излишних методологических деталей.
2. Заголовок отображает название описываемого метода, модели, алгоритма, а также теоретическую основу предлагаемых решений.

Мера TF-IDF

Согласно классическому определению, данная мера есть произведение TF-меры (отношения числа вхождений слова к общему числу слов документа) и инверсии частоты встречаемости в документах корпуса (IDF).

TF-мера оценивает важность слова t_i в пределах отдельного документа d и определяется как

$$\mathrm{tf}\,(t_i, d) = \frac{n_i}{\sum_k n_k}, \tag{1}$$

где n_i — число вхождений слова t_i в документ d,
а в знаменателе — общее число слов в документе.

IDF (inverse document frequency) — обратная частота документа, является единственной для каждого уникального слова в корпусе D и равна

$$\mathrm{idf}\,(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right), \tag{2}$$

где в числителе представлено общее число документов корпуса,
а $|D_i \subset D|$ есть число документов, где t_i встретилось хотя бы раз.

Интерпретация TF-IDF для сочетаний слов: значение числителя в (1) отождествляется с числом одновременных вхождений всех слов сочетания во фразы отдельного $d \in D$; при подсчёте значения в знаменателе (1) раздельно учитываются случаи совместной встречаемости слов сочетания и встречаемость без одновременного вхождения во фразу.

Пусть

D — исходное текстовое множество (корпус).

X — упорядоченная по убыванию последовательность $\mathrm{tf}\,(t_i, d) \cdot \mathrm{idf}\,(t_i, D)$ для всех слов t_i исходной фразы относительно документа $d \in D$.

F — последовательность кластеров $H_1, \ldots H_r$, на которые разбивается X алгоритмом, содержательно близким алгоритмам класса FOREL.

Центром масс H_i, $\mathrm{mc}(H_i)$, возьмём среднее арифметическое всех $x_j \in H_i$.

При этом элементы X *принадлежат одному кластеру*, если

$$\begin{cases} \left|\,\mathrm{mc}(X) - \mathrm{first}(X)\,\right| < \dfrac{\mathrm{mc}(X)}{4} \\ \left|\,\mathrm{mc}(X) - \mathrm{last}(X)\,\right| < \dfrac{\mathrm{mc}(X)}{4} \end{cases}. \tag{3}$$

Наибольший интерес для оценки близости фразы смысловому эталону представляют слова кластеров:

$H_1\,(X)$ — *слова-термины* исходной фразы, *наиболее уникальные* для d;

$H_{r/2}\,(X)$ — *общая лексика*, обеспечивающая *синонимические перифразы*, и *термины-синонимы*;

$H_r\,(X)$ — *слова-термины*, *преобладающие* в корпусе.

Оценивание близости фразы смысловому эталону

Основные эмпирические соображения

- как можно более выраженное разделение слов на общую лексику и термины;
- слова в кластерах $H_1, \ldots H_r$, формируемых по TF-IDF слов фразы относительно некоторого $d \in D$, должны быть распределены более или менее равномерно;
- число получившихся кластеров на последовательности X должно быть как можно ближе к трём при максимуме значений TF-IDF для слов кластера H_1.

Документы в составе корпуса D сортируются по убыванию произведения оценок:

$$val_1 = -1 / \log_{10} \left(\Sigma_{H_1} \right), \tag{4}$$

$$val_2 = 10^{-\sigma\left(\left| H_i,\, i = \left\{ 1,\, r/2,\, r \right\} \right| \right)}, \tag{5}$$

и, соответственно,

$$val_3 = \left| H_1 \setminus H_{r/2} \setminus H_r \right| / \mathrm{len}\,(X), \tag{6}$$

где Σ_{H_1} есть сумма величин TF-IDF слов, отнесённых к кластеру H_1 относительно $d \in D$;
$\sigma\left(\left| H_i,\, i = \left\{ 1,\, r/2,\, r \right\} \right| \right)$ — СКО числа элементов в кластере из списка $\left\{ H_1,\, H_{r/2},\, H_r \right\}$;
$\mathrm{len}\,(X)$ — длина последовательности X.

Замечания

- в случае $\Sigma_{H_1} = 0$ значение val_1 принимается равным нулю;
- если число полученных по TF-IDF кластеров меньше двух, то величины $\left| H_{r/2} \right|$ и $\left| H_r \right|$ принимаются равными нулю;
- при ровно двух кластерах по TF-IDF нулевым считается значение $\left| H_r \right|$.

Согласованность фрагментов знаний

Суть проблемы

Для каждой фразы максимум близости эталону достигается относительно своего документа корпуса и, как следствие, требуется оценить взаимную релевантность таких документов по разным фразам анализируемого текста.

Необходимо оценить на предмет отнесения к единому образу

Ключевые сочетания слов из задающих смысловые образы отдельных фраз, которые будут здесь выделяться относительно разных документов корпуса.

Наиболее близкие (в т. ч. «связанные») исследования и проекты

- Распознавание смысловых сверхфразовых единств на уровне глубинного синтаксиса [Емельянов Г. М., 2003].
- Вероятностное тематическое моделирование и разведочный информационный поиск [Воронцов К. В., 2020].

Основные особенности рассматриваемого круга задач

- формальные смысловые образы аналогов сверхфразовых единств задаются неявно;
- их выделение в тексте должно основываться на сопоставлении результатов классификации слов каждой его фразы по TF-IDF относительно разных $d \in D$;
- помимо терминов, выделяемые при этом образы должны учитывать языковые выразительные средства, определяющие лучший вариант среди возможных перифраз;
- основная трудность — относительно разных $d \in D$ одни и те же слова по-разному разделяются на общую лексику и термины.

Пусть $\mathbf{Ts}$ — группа фраз, первая из которых — заголовок научной статьи, а остальные представляют аннотацию.

Введём в рассмотрение для каждой $Ts_i \in \mathbf{Ts}$ вектор TF-IDF её слов:

$$\overrightarrow{Ts}_{ij} = \left(v_1, \ldots, v_{len(Ts_i)}\right), \tag{7}$$

получаемый относительно документа $d_j \in D$, $len\left(Ts_i\right)$ — длина Ts_i в словах.

Пусть $\overrightarrow{Ts}_{i,\max(D,i)}$ — вектор вида (7) для $d_{\max(i)} \in D$, относительно которого достигнут максимум произведения оценок (4), (5) и (6) по фразе $Ts_i \in \mathbf{Ts}$.

Обозначим последовательность векторов (7) по Ts_i для документов $d_j \in D$: $d_j \neq d_{\max(i)}$, сортируемую по убыванию расстояния до $\overrightarrow{Ts}_{i,\max(D,i)}$, как $\mathbb{T}_i$.

Разобьём $\mathbb{T}_i$ на кластеры $H_1, \ldots, H_{r(\mathbb{T}_i)}$, где $H_{r(\mathbb{T}_i)}$ по определению будет отвечать документам с наименьшим расстоянием до документа $d_{\max(i)}$.

Определение 1

Классификацию слов фразы $Ts_i \in \mathbf{Ts}$ по значению TF-IDF, выполненную относительно некоторого $d_j \in D$, будем считать *сопоставимой* с аналогичной классификацией относительно $d_{\max(i)}$ при выполнении одного из двух условий:

- $d_j \in H_{r(\mathbb{T}_i)}$;
- $\exists\, Ts_j \in \mathbf{Ts}$: $Ts_j \neq Ts_i$, $d_j = d_{\max(j)}$, при этом $\exists\, d_k \in D$: $d_k \neq d_j$, $d_k \neq d_{\max(i)}$, причём d_k одновременно относится и к $H_{r(\mathbb{T}_i)}$, и к $H_{r(\mathbb{T}_j)}$.

Сами d_j и $d_{\max(i)}$ назовём *взаимно релевантными по TF-IDF*.

Утверждение 1

Значение TF-IDF ключевого сочетания слов должно быть не ниже минимального из значений указанной меры по его отдельным словам.

Утверждение 2

Выделяемые ключевые сочетания с наибольшей вероятностью будут определять единый смысловой образ текста $\mathbf{Ts}$, если они:

- идентифицируются как таковые относительно документа $d_{\max} \in D$, по которому максимум произведения оценок (4), (5) и (6) достигался по наибольшему числу фраз в составе текстов анализируемой коллекции;
- выделяются в некоторой фразе $Ts_i \in \mathbf{Ts}$ и идентифицируются как таковые относительно некоторого документа $d_{\max(i)} \in D$, причём вышеупомянутый документ $d_{\max}$ будет относиться к кластеру $H_{r(\mathbb{T}_i)}$.

Утверждение 3

При прочих равных условиях из ключевых сочетаний, идентифицируемых относительно единственных документов, не равных $d_{\max}$, меньшее преимущество будет у того, по документу которого $\left|H_{r(\mathbb{T}_i)}\right| > \left|D\right|/2$.

Введём в рассмотрение граф, где вершины соответствуют тем $d \in D$, относительно которых достигается максимум произведения оценок (4), (5) и (6) минимум по одной $Ts_i \in \mathbf{Ts}$, а каждое ребро соединяет вершины для пары взаимно релевантных по TF-IDF документов (далее — *граф релевантности, ГР*).

Утверждение 4

При объединении графов релевантности всех текстов коллекции вышестоящий текст $\mathbf{Ts}_i$ и непосредственно связанный с ним нижестоящий текст $\mathbf{Ts}_j$ в формируемой иерархии должны иметь свои графы релевантности подграфами некоторой компоненты связности (КС) объединённого релевантности графа (ОГР) по коллекции.

При прочих равных условиях при выборе вышестоящего для заданного $\mathbf{Ts}_j$ в формируемой иерархии предпочтение отдаётся тексту $\mathbf{Ts}_i$, который отвечает условию *Утверждения 4*.

Пусть S — последовательность текстов анализируемой коллекции;
$\bigcup_S \mathbf{Ts}$ — объединённое множество фраз по всем текстам $\mathbf{Ts}$ в составе S.

Тогда значимость документа $d \in D$ для формирования графа релевантности текста $\mathbf{Ts}$ может из геометрических соображений быть оценена как

$$\mathrm{N}(d) = \frac{\left|D\right| - \min_d\left(\left|H_{r(\mathbb{T}_i)} : Ts_i \in \bigcup_S \mathbf{Ts}\right|\right)}{\sigma\left(\left|H_j \in \left\{H_1, \ldots, H_{r(\mathbb{T}_i)}\right\}_d : Ts_i \in \bigcup_S \mathbf{Ts}\right|\right) + 1}. \qquad (8)$$

Исходные данные эксперимента: состав текстового корпуса

- 3 статьи в журнале «Таврический вестник информатики и математики» (ТВИМ);
- 2 статьи в сборниках трудов 8-й и 9-й международных конференций «Интеллектуализация обработки информации» 2010 и 2012 гг.;
- 1 статья в сборнике трудов 15-й Всероссийской конференции «Математические методы распознавания образов» (ММРО, 2011 г.);
- материалы тезисов двух докладов на 13-й Всероссийской конференции «Математические методы распознавания образов» (2007 г.);
- материалы тезисов четырнадцати докладов на 16-й Всероссийской конференции «Математические методы распознавания образов» (2013 г.);
- материалы тезисов двух докладов на международной конференции «Интеллектуализация обработки информации» (ИОИ) 2014 г.;
- материалы одного научного отчёта (Михайлов Д. В., 2003 г.).

Примечание

Число слов в документах корпуса здесь варьировалось от 218 до 6298, число фраз — от 9 до 587.

Исходные данные эксперимента: тематика работ в составе корпуса

- математические методы обучения по прецедентам (К. В. Воронцов, М. Ю. Хачай, Е. В. Дюкова, Н. Г. Загоруйко, Ю. Ю. Дюличева, И. Е. Генрихов, А. А. Ивахненко);
- модели и методы распознавания и прогнозирования (В. В. Моттль, О. С. Середин, А. И. Татарчук, П. А. Турков, М. А. Суворов, А. И. Майсурадзе);
- интеллектуальный анализ экспериментальных данных (С. Д. Двоенко, Н. И. Боровых);
- обработка, анализ, классификация и распознавание изображений (А. Л. Жизняков, К. В. Жукова, И. А. Рейер, Д. М. Мурашов, Н. Г. Федотов, В. Ю. Мартьянов, М. В. Харинов).

Исходные данные эксперимента: коллекции для отбора статей

- сборник трудов конференции «Интеллектуализация обработки информации» 2012 г., раздел «Математическая теория и методы классификации» (14 статей);
- сборник трудов 14-й Всероссийской конференции «Математические методы распознавания образов» (2009 г.), раздел «Методы и модели распознавания и прогнозирования» (35 статей);
- сборник трудов 15-й Всероссийской конференции «Математические методы распознавания образов», разделы «Математическая теория и методы классификации» (18 статей) и «Статистическая теория обучения» (10 статей).

Некоторые технические детали

- Вычисление оценок (4)–(6) — без учёта предлогов и союзов.
- Извлечение текста из PDF-файла — с помощью функций классов *pdfinterp*, *converter*, *layout* и *pdfpage* в составе пакета *PDFMiner*.
- В целях корректности распознавания все формулы из анализируемых документов переводились экспертом вручную в формат, близкий используемому в LaTeX.
- Для выделения границ предложений в тексте по знакам препинания был задействован метод *sent_tokenize()* класса *tokenize* из входящих в *NLTK*.
- Приведение слов к начальной форме — с помощью *PyMorphy2*.
- При более одном варианте разбора слова для определения его начальной формы берётся ближайший выдаваемому n-граммным теггером в составе *nltk4russian*.

Программная реализация на Python 2.7 и результаты экспериментов

Таблица 1. Документы анализируемой коллекции.

№	Автор (ы) и заголовок статьи	$len\,(Ts_i)$ и $d_{\max(i)}$ по отдельным $Ts_i \in \mathbf{Ts}$
1	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	$7(2), 12(1), 13(1), 9(1)$ [1]
2	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	$6(1), 10(1), 18(1)$
3	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	$7(1), 14(1)$
4	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	$7(2), 14(6), 8(2)$
5	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	$8(1), 15(1), 11(2)$
6	Каневский Д. Ю. Переобучение и комбинаторная радемахеровская сложность в задачах восстановления регрессии	$7(1), 20(1), 10(3), 18(1), 9(1)$
7	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	$7(1), 16(1), 9(1), 5(4)$
8	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	$7(1), 25(2), 14(1)$
9	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	$8(2), 13(2), 10(2)$
10	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	$10(5), 18(1), 15(2), 11(1), 18(1), 10(2)$

[1] Длина фразы в словах и далее в скобках — порядковый номер для $d_{\max(i)}$ по *Таблице 2.*

Таблица 2. Документы $d \in D$, относительно которых достигался максимум произведения оценок (4), (5) и (6) как минимум по одной фразе.

№	Автор(ы), название и выходные данные работы
1	Воронцов К. В. Обзор современных исследований по проблеме качества обучения алгоритмов // ТВИМ. 2004. № 1. С. 5–24.
2	Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // 15-я Всерос. конф. «Математические методы распознавания образов» (ММРО-15): Тез. докл. М., 2011. С. 40–43.
3	Дюличева Ю. Ю. Стратегии редукции решающих деревьев (обзор) // Таврический вестник информатики и математики. 2002. № 1. С. 10–16.
4	Дюкова Е. В., Песков Н. В. Об алгоритме классификации на основе полного решающего дерева // 13-я Всерос. конф. «Математические методы распознавания образов» (ММРО-13): Тез. докл. М., 2007. С. 125–126.
5	Дюличева Ю. Ю. О программной реализации и апробации алгоритма DFBSA синтеза эмпирического решающего леса // ТВИМ. 2003. № 2. С. 35–44.
6	Ишкина Ш. Х., Ивахненко А. А. Комбинаторные оценки переобучения пороговых решающих правил // 16-я Всерос. конф. «Математические методы распознавания образов» (ММРО-16): Тез. докл. М., 2013. С. 23.

Таблица 3. Документы из *Таблицы 2* и число фраз по коллекции с достигнутым максимумом произведения оценок (4), (5) и (6).

Номер документа по *Таблице 2*	1	2	3	4	5	6
Число фраз по коллекции	22	10	1	1	1	1
Минимальная (максимальная) длина фразы	6(20)	7(25)	10(10)	5(5)	10(10)	14(14)

Таблица 4. Компоненты связности объединённого графа релевантности по $\mathbf{Ts}_j$ и $\mathbf{Ts}_i$ [2].

Компоненты связности [3]	Связи, $j \to i$ [4]
$\{1, 2\}$	$2 \to 1,\ \ 9 \to 1,\ \ 9 \to 5,\ \ 9 \to 8$
$\{1, 2, 3\}$	$6 \to 1$
$\{1\}, \{2, 6\}$	$4 \to 3$
$\{2, 6\}, \{1, 3\}$	$6 \to 4$
$\{1, 2, 6\}$	$8 \to 4$
$\{2, 6\}$	$9 \to 4$
$\{1, 3\}, \{2\}$	$9 \to 6$
$\{1, 2, 4\}$	$8 \to 7$

В *Таблице 4* представлены связи $j \to i$, у которых значения дополняемости соответствующих им текстов отличны от нуля. Непосредственно дополняемость текста $\mathbf{Ts}_j$ текстом $\mathbf{Ts}_i$ относительно их смысловых эталонов определяется долей слов кластеров наибольших значений TF-IDF фраз текста $\mathbf{Ts}_i$, не входящих в кластеры наибольших значений указанной меры по фразам текста $\mathbf{Ts}_j$, но, тем не менее, имеющих относительно тех же фраз ненулевые значения TF-IDF.

Число выполняемых шагов поиска наименьшей компоненты связности в ОГР по коллекции может служить оценкой силы связи: чем меньше шагов, тем сильнее связь.

[2] Строки для связей, где ОГР состоит из нескольких КС, выделены *более тёмным фоном*.

[3] представлены списками вершин, а каждая вершина — номером по *Таблице 2*.

[4] от нижестоящего к вышестоящему, где i и j — номера документов по *Таблице 1*.

Таблица 5. Значимость документов $d \in D$ при подборе пары взаимно релевантных.

Номер по Таблице 2	Оценка (8)	Номер кластера
1	3,22591590939	1
2	2,89575917000	2
3	2,30983870471	
5	2,09090909091	
6	2,09090909091	
4	1,00000000000	3

Основная гипотеза

Если разбить документы, представляемые *Таблицей 2*, на кластеры по значению оценки (8), то при прочих равных условиях наименьший приоритет будет у связи с тем текстом $\mathbf{Ts}_i$, у которого максимум близости эталону минимум для одной фразы достигается относительно некоторого $d \in D$, относимого к кластеру наименьших значений оценки (8).

Пример: связи $8 \to 4$ и $8 \to 7$.

В документе с порядковым номером 7 по *Таблице 1*:

7 Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации $7(1), 16(1), 9(1), 5(4)$

по 4-й фразе наибольшая близость эталону достигнута относительно документа № 4 из *Таблицы 2*, отнесённого (как видно из *Таблицы 5*) к кластеру наименьших значений оценки (8).

Следовательно, при прочих равных условиях предпочтение отдаётся связи $8 \to 4$.

Для сравнения: здесь имеем $\left|H_{r(\mathbb{T}_i)}\right| = 24$ при $|D| = 25$, что не гарантирует более высокие значения TF-IDF ключевых терминов в родительском документе по сравнению с дочерним в формируемой иерархии и, следовательно, навигацию по коллекции с постепенным фокусированием внимания на подтемах.

Заключение

1. Основной *результат* настоящей работы — *методика* анализа взаимной релевантности документов тематического корпуса, относительно которых оценивается близость текста смысловому эталону.
2. В *Определении 1* не сказано о возможной мене местами документов, анализируемых на взаимную релевантность по TF-IDF, с сохранением выполнимости первого из его условий. *В реальном тексте* вероятность существования пары отличных друг от друга фраз с сопоставимыми классификациями по TF-IDF *существенно зависит* от длины текста.
3. Открытая проблема — качество кластеризации документов корпуса D по величине значимости для формирования ГР текста. Представляет интерес *изучение распределения частот* встречаемости $d \in D$ в кластере наименьших значений оценки (8) по разным коллекциям одной тематики на основе *квантилей эмпирических распределений* указанных частот.
4. Саму оценку (8) следует рассматривать как основу выделения необходимого и достаточного набора документов $d \in D$ для оценки близости эталону как отдельных фраз, так и текстов анализируемой коллекции.
5. Для определения «эталонного» диапазона кластеров значений оценки (8) наиболее перспективным представляется поиск документов $d \in D$, *наименее часто меняющих* местоположение в кластерах по значению оценки (8) *при переходе между коллекциями* внутри одной тематики.