Мера TF-IDF, сила связи слов и ключевые сочетания для безызбыточной передачи единицы знаний

Стрещук В. А., Кузнецов П. А., Михайлов Д. В.

Новгородский государственный университет
имени Ярослава Мудрого

XXIII Международная научно-техническая конференция

«Медико-экологические информационные технологии – 2020»

20–22 мая 2020 г.

г. Курск

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.

Наиболее близкие исследования и проекты

- Вероятностное тематическое моделирование и разведочный информационный поиск [Воронцов К. В., 2019].
- Построение иерархических тематических моделей крупных конференций [Стрижов В. В., 2014].
- Квантильный подход к оцениванию когнитивной сложности текста [Еремеев М. А., 2019].
- Тезаурусное представление онтологии предметной области анализа изображений [ВЦ РАН, тезаурус «Чёрный квадрат»].
- Подготовка размеченных текстовых корпусов для обучения системы автоматического перефразирования [проект ParaPhraser].

Основные проблемы:

- не предусматривается качественный анализ языковых выразительных средств, значимых для выбора лучших вариантов парафраз;
- требуется выделение и анализ взаимосвязей смысловых эталонов отдельных текстовых документов.

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.

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

Классификация слов исходной фразы по значению TF-IDF: базовые предположения

1. Наиболее уникальные слова в документе (с наибольшими значениями TF*IDF) будут относиться к терминам его предметной области.
2. Наличие синонимов у слова-термина ведёт к снижению значения TF относительно документа в случае, когда синонимы встречаются в этом же документе.
3. Термины, преобладающие в корпусе, а также слова общей лексики будут иметь значения IDF, близкие к нулю.
4. Слова-синонимы, уникальные для отдельных документов корпуса, будут иметь более высокие значения IDF.

   Пример — слова общей лексики, задающие конверсивные замены: *«приводить ⇔ являться следствием»*.

Утверждение 1

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

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 кластера H_i возьмём среднее арифметическое всех $x_j \in H_i$.

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.

Выбор оценки 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:

$$val_1 = -1/\log_{10}\left(\Sigma_{H_1}\right), \tag{3}$$

$$val_2 = 10^{-\sigma\left(\left|H_i,\, i=\left\{1,\, r/2,\, r\right\}\right|\right)}, \tag{4}$$

и, соответственно,

$$val_3 = \left|H_1 \setminus H_{r/2} \setminus H_r\right| / \mathrm{len}\,(X), \tag{5}$$

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 $\left|H_r\right|$.

Оценка близости эталону для группы фраз: заголовок + аннотация

Пусть

Ts — группа фраз, первая из которых — заголовок научной статьи, а остальные представляют аннотацию.

Первый вариант оценки:

$$\mathrm{N}_1(\mathrm{Ts}, D) = \frac{\max\limits_{d \in D}\left(val_1(Ts_1, d) \cdot val_2(Ts_1, d) \cdot val_3(Ts_1, d)\right)}{\sigma\left(\max\limits_{d \in D}\left(val_1(Ts_i, d) \cdot val_2(Ts_i, d) \cdot val_3(Ts_i, d)\right), Ts_i \in \mathrm{Ts}\right) + 1}. \quad (6)$$

Здесь:

в числителе — оценка близости эталону заголовка статьи (Ts_1);
первое слагаемое в знаменателе — СКО значения близости эталону по всем $Ts_i \in \mathrm{Ts}$.

Замечания

- оценка (6) зависит от подбора корпуса D экспертом;
- введённая оценка не подразумевает сортировку фраз $Ts_i \in \mathrm{Ts}$ по близости эталону и содержательно соответствует порядку отбора статей, начиная с анализа заголовка;
- априорное предположение о максимальной близости эталону именно заголовка статьи на практике выполняется не всегда.

Оценка близости эталону для группы фраз: заголовок + аннотация

Второй вариант оценки:

$$\mathrm{N}_2(\mathrm{Ts}, D) = \frac{\max\limits_{d \in D}\left(val_1\left(Ts_{\max}, d\right) \cdot val_2\left(Ts_{\max}, d\right) \cdot val_3\left(Ts_{\max}, d\right)\right)}{\sigma\Big(\max\limits_{d \in D}\left(val_1\left(Ts_i, d\right) \cdot val_2\left(Ts_i, d\right) \cdot val_3\left(Ts_i, d\right)\right), Ts_i \in \mathrm{Ts}\Big) + 1}, \quad (7)$$

где $Ts_{\max} \in \mathrm{Ts}$ — фраза, по которой получен максимум близости эталону.

Утверждение 2

Максимальный итоговый рейтинг по коллекции получает статья с наибольшим значением оценки (6), попадающим в один кластер со значением оценки (7) для той же статьи.

Замечания

- корректное применение *Утверждения 2* предполагает отнесение к одному кластеру значений оценки (6) для статьи с максимальным итоговым рейтингом и максимального значения оценки (6) по коллекции, из которой ведётся отбор;
- в случае отсутствия в коллекции статьи, удовлетворяющей данному требованию, *максимальный итоговый рейтинг* получает статья с наибольшим значением оценки (6) по анализируемой коллекции;
- поскольку заголовок и фразы аннотации (по определению) несут некий единый смысловой образ, то допустима мена местами оценок (6) и (7) в *Утверждении 2*.

Выбор оценки силы связи слов

Дистрибутивно-статистический метод [Москович В. А., 1971] построения тезаурусов — сила связи совместно встречающихся во фразе слов:

$$K_{AB} = \frac{k}{a + b - k}, \tag{8}$$

где a — число фраз текста, которые содержат слово A, b — слово B, k — A и B одновременно.

Замечание

Оценка (8) вычисляется только в том случае, если значение TF-IDF минимум одного из слов пары принадлежит либо $H_1(X)$, либо $H_{r/2}(X)$.

Пусть $L(d)$ есть последовательность *пар слов* (A, B) исходной фразы, упорядоченная по *убыванию* оценки (8) относительно некоторого документа $d \in D$.

Определение 1

Биграммы (A_1, B_1) и (A_2, B_2) войдут в одну n-грамму $T \subseteq L(d)$, если

$$\big((A_1 = A_2) \vee (B_1 = B_2) \vee (A_1 = B_2) \vee (B_1 = A_2)\big) = \text{true}.$$

Значимость n-граммы T для оценки ранга документа d относительно D

$$\mathrm{N}(T,d) = \frac{\sqrt{\sum_{i=1}^{\mathrm{len}(T)} \left[S_i\,(d)\right]^2}}{\sigma\left(S_i\,(d)\right) + 1}, \tag{9}$$

где $S_i\,(d)$ — сила связи слов i-й биграммы относительно d;

$\sigma\left(S_i\,(d)\right)$ — среднеквадратическое отклонение указанной величины;

$\mathrm{len}\,(T)$ — длина n-граммы T (в биграммах).

Замечание

Будем вычислять оценку (9) только для n-грамм с ненулевым значением меры TF-IDF относительно документа d с наибольшим значением произведения оценок (3), (4) и (5) по заданному корпусу D.

Разобьём множество найденных n-грамм на кластеры по значению оценки (9).

Утверждение 3

«Ключевые» n-граммы не попадут в кластер наименьших значений оценки (9).

Будем считать ключевым сочетание, отвечающее *Утверждению 1* либо *3*.

Исходные данные эксперимента: состав текстового корпуса

- 3 статьи в журнале «Таврический вестник информатики и математики»;
- 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 (3)–(7) — 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 $d \in D$, относительно которых достигался максимум произведения оценок (3), (4) и (5).

№, i	Документ
1	Воронцов К. В. Обзор современных исследований по проблеме качества обучения алгоритмов // Таврический вестник информатики и математики. 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	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 2. Документы анализируемой коллекции, из которых производилось выделение ключевых сочетаний слов.

№, j	Документ
1	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения
2	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа
3	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов
4	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля
5	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов
6	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска

Таблица 3. Найденные сочетания слов[1].

№	j	i	Сочетание слов	Оценка (9)
1	1	1	обобщающая способность	0,863636363636
2	2	1	разделяющая поверхность	0,333333333333
3	2	1	обобщающая способность	0,863636363636
4	2	1	ближайший сосед	0,666666666667
5	2	1	скользящий контроль	0,958333333333
6	2	2	монотонный классификатор ближайшего соседа	0,601128719509
7	2	2	ближайший сосед	0,700000000000
8	3	1	полный скользящий контроль	0,680552311539
9	3	1	скользящий контроль	0,958333333333
10	3	1	(обучение) классификатора (по) выборке	0,018518518519
11	3	1	скользящий контроль	0,958333333333
12	4	1	скользящий контроль	0,958333333333
13	5	2	учитывать эффект	0,285714285714
14	6	2	минимизация эмпирического риска	0,769230769231
15	6	3	комбинаторная теория	0,250000000000
16	6	3	обобщающая способность	0,750000000000

[1] Строки для сочетаний, не отвечающих *Утверждению 1*, выделены более тёмным фоном.

Таблица 4. Кластеры по значению оценки (9) для сочетаний слов из *Таблицы 3*.

i	№ кластера	Сочетание слов	Оценка (9)
1	0	скользящий контроль	0,958333333333
		обобщающая способность	0,863636363636
		полный скользящий контроль	0,680552311539
		ближайший сосед	0,666666666667
1	1	разделяющая поверхность	0,333333333333
1	2	(обучение) классификатора (по) выборке	0,018518518519
2	0	минимизация эмпирического риска	0,769230769231
		ближайший сосед	0,700000000000
		монотонный классификатор ближайшего соседа	0,601128719509
2	1	учитывать эффект	0,285714285714
3	0	обобщающая способность	0,750000000000
3	1	комбинаторная теория	0,250000000000

Выделенные цветом примеры сочетаний слов:

- — сочетание признано экспертом как допустимое и относится к ключевым;
- — сочетание не признано как значимое для передачи единицы знаний.

Таблица 5. Кластеры для сочетаний слов: материал *Таблицы 2* дополнен документами коллекции «Математическая теория и методы классификации»[2], $i = 1$.

№ кластера	Сочетание слов	№ п/п	Оценка (9)
0	скользящий контроль	1	0,958333333333
	обобщающая способность	2	0,863636363636
1	нейронная сеть	3	0,750000000000
	полный скользящий контроль	4	0,680552311539
	ближайший сосед	5	0,666666666667
	минимизация риска	6	0,611111111111
	независимо (и) случайно	7	0,500000000000
	опорный вектор	8	0,500000000000
2	обучающая выборка	9	0,416666666667
	заданный обчающей выборкой	10	0,377592871278
	решающее правило	11	0,357142857143
	обучающая выборка (в) системе	12	0,355672566871
	разделяющая поверхность	13	0,333333333333
	линейное решающее правило	14	0,32257427855
3	линейный классификатор	15	0,111111111111
	сложная зависимость	16	0,080000000000
4	алгоритм оптимизации	17	0,0434782608696
5	обучающий материал	18	0,0277777777778
	расширение семейства	19	0,027027027027
	(обучение) классификатора (по) выборке	20	0,018518518519

[2] Строки для сочетаний, не отвечающих *Утверждению 1*, выделены более тёмным фоном.

Выводы

1. *Применение классификатора* на основе оценки (9) *как дополнения* классификации на основе *Утверждения 1*, *в значительной мере зависит от* документа, относительно которого вычисляется оценка (9).
2. Для *повышения точности* выделения ключевых сочетаний слов здесь *требуется* дополнительная статистика по документам, относительно которых *достигался максимум* близости эталону по *различным* фразам *разных* документов анализируемой коллекции.
3. *Более точное* выделение ключевых сочетаний слов на основе меры TF-IDF *обусловлено* вычислением IDF-меры для сочетания *относительно всего* заданного текстового корпуса.
4. Выделение *ключевых сочетаний* слов на основе *Утверждения 1* представляется *более целесообразным* с позиции основной задачи — поиска наиболее *значимых составляющих* образа исходной фразы.
5. Представляет интерес *исследование связи* между
   - *распределениями* частот встречаемости слов в кластерах наибольших значений TF-IDF по фразам разных текстов анализируемой коллекции;
   - *случаями* достижения максимума произведения оценок (3), (4) и (5) относительно конкретных документов заданного текстового корпуса.