Мера TF-IDF и оценка близости смысловому эталону заголовка и аннотации научной статьи без перифразирования

Михайлов Д. В., Емельянов Г. М.

Новгородский государственный университет
имени Ярослава Мудрого

Всероссийская конференция с международным участием
«Математические методы распознавания образов» (ММРО-19),

26–29 ноября 2019 г.

г. Москва

Текст научной работы и рациональная передача смысла

Оптимальная (эталонная) передача смысла

Обеспечивается набором единиц текста и их связей, *необходимым и достаточным* для представления единицы знаний.

Требования к решению

1. Сортировка источников информации по степени отражения наиболее существенных понятий изучаемой предметной области при максимальной компактности и безызбыточности изложения.
2. Эксперт не должен перифразировать 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. Отсутствие априорных ограничений на природу связей единиц текста.

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

- Построение иерархических тематических моделей крупных конференций [Стрижов В. В., 2014].
- Обраружение парафраз нейронной сетью, обучаемой на результатах синтаксического разбора исходных фраз [Huang E., 2011].
- Подготовка размеченных текстовых корпусов для обучения системы автоматического перефразирования [проект ParaPhraser].
- ParaPlag: корпус для выявления перефразированных текстовых заимствований на русском языке.
- Оценка смысловой близости предложений синтаксическим разбором и определением редакционного расстояния между полученными деревьями зависимостей [HSE School of Linguistics, 2016].

Основные проблемы:

- не предусматривается качественный анализ языковых выразительных средств, значимых для выбора лучших вариантов парафраз;
- качество обучения системы распознавания парафраз.

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* $\Leftrightarrow$ *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 $H_1, \dots H_r$, 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.

Второй вариант оценки:

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

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*.

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

Результат: статьи с максимальным значением оценки (6) по коллекциям

ММРО-15, Статистическая теория обучения	
Автор(ы)	*Воронцов К. В., Махина Г. А.*
Заголовок статьи	*Принцип максимизации зазора для монотонного классификатора ближайшего соседа*
Максимум оценки близости эталону для заголовка достигается относительно документа	*Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15*
Оценка близости эталону заголовка статьи:	0,0729
СКО близости эталону по всем фразам аннотации и заголовку:	0,0252
Значение оценки (6):	0,0711
Значение оценки (7):	0,0711
ММРО-15, Математическая теория и методы классификации	
Автор(ы)	*Генрихов И. Е., Дюкова Е. В.*
Заголовок статьи	*Полные решающие деревья в задачах классификации по прецедентам*
Максимум оценки близости эталону для заголовка достигается относительно документа	*Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15*
Оценка близости эталону заголовка статьи:	0,1253
СКО близости эталону по всем фразам аннотации и заголовку:	0,0489
Значение оценки (6):	0,1194
Значение оценки (7):	0,1194

Результат: статьи с максимальным значением оценки (6) по коллекциям

ММРО-14, Методы и модели распознавания и прогнозирования	
Автор(ы)	*Баринова О. В., Ветров Д. П.*
Заголовок статьи	*Оценки обобщающей способности бустинга с вероятностными входами*
Максимум оценки близости эталону для заголовка достигается относительно документа	*Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15*
Оценка близости эталону заголовка статьи:	0,1359
СКО близости эталону по всем фразам аннотации и заголовку:	0,0498
Значение оценки (6):	0,1295
Значение оценки (7):	0,1295
ИОИ-9, Математическая теория и методы классификации	
Автор(ы)	*Двоенко С. Д., Пшеничный Д. О.*
Заголовок статьи	*Об устранении отрицательных собственных значений матриц парных сравнений*
Максимум оценки близости эталону для заголовка достигается относительно документа	*Двоенко С. Д., Пшеничный Д. О. Метрическая коррекция матриц парных сравнений // ММРО-16*
Оценка близости эталону заголовка статьи:	0,0952
СКО близости эталону по всем фразам аннотации и заголовку:	0,0353
Значение оценки (6):	0,0920
Значение оценки (7):	0,0920

Результат: статьи с максимальным значением оценки (7) по коллекциям

ММРО-15, Статистическая теория обучения			
Автор(ы)	*Воронцов К. В., Махина Г. А.*		
Заголовок статьи	*Принцип максимизации зазора для монотонного классификатора ближайшего соседа*		
Фраза, максимально близкая эталону	*Принцип максимизации зазора для монотонного классификатора ближайшего соседа*		
Максимум оценки близости фразы эталону достигается относительно документа	*Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15*		
Оценка близости эталону заголовка статьи:		0,0729	
СКО близости эталону по всем фразам аннотации и заголовку:		0,0252	
Значение оценки (7):	0,0711	Значение оценки (6):	0,0711

ММРО-15, Математическая теория и методы классификации			
Автор(ы)	*Генрихов И. Е., Дюкова Е. В.*		
Заголовок статьи	*Полные решающие деревья в задачах классификации по прецедентам*		
Фраза, максимально близкая эталону	*Полные решающие деревья в задачах классификации по прецедентам*		
Максимум оценки близости фразы эталону достигается относительно документа	*Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15*		
Оценка близости эталону заголовка статьи:		0,1253	
СКО близости эталону по всем фразам аннотации и заголовку:		0,0489	
Значение оценки (7):	0,1194	Значение оценки (6):	0,1194

Результат: статьи с максимальным значением оценки (7) по коллекциям

ММРО-14, Методы и модели распознавания и прогнозирования			
Автор(ы)	*Мельников Д. И., Стрижов В. В., Андреева Е. Ю., Эденхартер Г.*		
Заголовок статьи	*Выбор опорного множества при построении устойчивых интегральных индикаторов*		
Фраза, максимально близкая эталону	*Объекты описаны в линейных шкалах*		
Максимум оценки близости фразы эталону достигается относительно документа	*Абрамов В. И., Середин О. С., Сулимова В. В., Моттль В. В. Эквивалентность потенциальных функций и линейных пространств представления объектов произвольной природы // ИОИ-8*		
Оценка близости эталону заголовка статьи:	0,0137		
СКО близости эталону по всем фразам аннотации и заголовку:	0,0639		
Значение оценки (7):	0,1426	Значение оценки (6):	0,0129

ИОИ-9, Математическая теория и методы классификации			
Автор(ы)	*Животовский Н. К., Воронцов К. В.*		
Заголовок статьи	*Критерии точности комбинаторных оценок обобщающей способности*		
Фраза, максимально близкая эталону	*Комбинаторная теория переобучения даёт точные оценки вероятности переобучения для некоторых нетривиальных семейств алгоритмов классификации*		
Максимум оценки близости фразы эталону достигается относительно документа	*Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15*		
Оценка близости эталону заголовка статьи:	0,0634		
СКО близости эталону по всем фразам аннотации и заголовку:	0,0578		
Значение оценки (7):	0,1336	Значение оценки (6):	0,0600

Статьи с максимальным итоговым рейтингом по коллекциям

Относительно оценки (6)

ММРО-15, Статистическая теория обучения			
Автор(ы)	*Воронцов К. В., Махина Г. А.*		
Значение оценки (6):	0,0711/0,0711	Значение оценки (7):	0,0711/0,0711[1]
ММРО-15, Математическая теория и методы классификации			
Автор(ы)	*Генрихов И. Е., Дюкова Е. В.*		
Значение оценки (6):	0,1194/0,1194	Значение оценки (7):	0,1194/0,1194
ММРО-14, Методы и модели распознавания и прогнозирования			
Автор(ы)	*Баринова О. В., Ветров Д. П.*		
Значение оценки (6):	0,1295/0,1295	Значение оценки (7):	0,1295/0,1426
ИОИ-9, Математическая теория и методы классификации			
Автор(ы)	*Двоенко С. Д., Пшеничный Д. О.*		
Значение оценки (6):	0,0920/0,0920	Значение оценки (7):	0,0920/0,1336

Не получила максимального итогового рейтинга

ММРО-14, Методы и модели распознавания и прогнозирования			
Автор(ы)	*Мельников Д. И., Стрижов В. В., Андреева Е. Ю., Эденхартер Г.*		
Значение оценки (6):	0,0129/0,1295	Значение оценки (7):	0,1426/0,1426

Относительно оценки (7)

ММРО-14, Методы и модели распознавания и прогнозирования			
Автор(ы)	*Баринова О. В., Ветров Д. П.*		
ИОИ-9, Математическая теория и методы классификации			
Автор(ы)	*Животовский Н. К., Воронцов К. В.*		
Значение оценки (6):	0,0600/0,0920	Значение оценки (7):	0,1336/0,1336

[1] Число после дроби — максимум оценки по коллекции

Ключевые сочетания слов для статей с максимальным значением оценки (6)

ММРО-15, Статистическая теория обучения	
Автор(ы)	*Воронцов К. В., Махина Г. А.*
Слова из кластеров наибольших значений TF-IDF по отдельным фразам	*монотонный, сосед, близкий, скользящий, контроль, обобщать, способность*
Образуемые из них сочетания, отвечающие условию *Утверждения 1*	*ближайший сосед, скользящий контроль, обобщающая способность*
в т. ч. со словами «серединных» кластеров	*разделяющая поверхность*

ММРО-15, Математическая теория и методы классификации	
Автор(ы)	*Генрихов И. Е., Дюкова Е. В.*
Слова из кластеров наибольших значений TF-IDF по отдельным фразам	*классификация, полный, решающий, дерево, прецедент, процедура, описание, обзор, дать*
Образуемые из них сочетания, отвечающие условию *Утверждения 1*	*решающее дерево*
в т. ч. со словами «серединных» кластеров	*распознающая процедура*

ММРО-14, Методы и модели распознавания и прогнозирования	
Автор(ы)	*Баринова О. В., Ветров Д. П.*
Слова из кластеров наибольших значений TF-IDF по отдельным фразам	*обобщать, способность, композиция, ошибка, вероятность, классификация, выборка, верхний, бустинг*
Образуемые из них сочетания, отвечающие условию *Утверждения 1*	*обобщающая способность*

ИОИ-9, Математическая теория и методы классификации	
Автор(ы)	*Двоенко С. Д., Пшеничный Д. О.*
Слова из кластеров наибольших значений TF-IDF по отдельным фразам	*парный, матрица, численный, измерение, корректный, обработка, следовать*
Образуемые из них сочетания, отвечающие условию *Утверждения 1*	*матрица парных*
в т. ч. со словами «серединных» кластеров	*матрица парных сравнений*

Ключевые сочетания слов для статей с максимальным значением оценки (7)

ММРО-14, Методы и модели распознавания и прогнозирования	
Автор(ы)	*Мельников Д. И., Стрижов В. В., Андреева Е. Ю., Эденхартер Г.*
Слова из кластеров наибольших значений TF-IDF по отдельным фразам	*опорный, описание, линейный, помощь, учитель, основный, предложить, получение*
Образуемые из них сочетания, отвечающие условию *Утверждения 1*	
в том числе со словами «серединных» кластеров	*выбор опорного, описаны (в) линейных*
Оценка близости эталону заголовка статьи:	0,0137
Максимум оценки близости фразы эталону:	0,1517
СКО близости эталону по всем фразам аннотации и заголовку:	0,0639

ИОИ-9, Математическая теория и методы классификации	
Автор(ы)	*Животовский Н. К., Воронцов К. В.*
Слова из кластеров наибольших значений TF-IDF по отдельным фразам	*обобщать, комбинаторный, вероятность, семейство*
Образуемые из них сочетания, отвечающие условию *Утверждения 1*	
в том числе со словами «серединных» кластеров	
Оценка близости эталону заголовка статьи:	0,0634
Максимум оценки близости фразы эталону:	0,1413
СКО близости эталону по всем фразам аннотации и заголовку:	0,0578

Соподчинённые слова кластеров наибольших значений TF-IDF

ММРО-15, Статистическая теория обучения, Воронцов К. В., Махина Г. А.	
ближайший сосед	*«Принцип максимизации зазора для монотонного классификатора ближайшего соседа»*
скользящий контроль	*«Получены точные оценки полного скользящего контроля для монотонных классификаторов, основанных на принципе ближайшего соседа»*
обобщающая способность	*«Показано, что наилучшей обобщающей способностью обладает монотонный классификатор, в котором разделяющая поверхность проходит посередине зазора между классами»*
в том числе со словами «серединных» кластеров	
монотонный классификатор	*«Принцип максимизации зазора для монотонного классификатора ближайшего соседа»*
ММРО-15, Математическая теория и методы классификации, Генрихов И. Е., Дюкова Е. В.	
полный	*«Полные решающие деревья в задачах классификации по прецедентам»*
(полное) решающее дерево	*«В докладе представлены результаты, полученные авторами, разработки алгоритмов классификации на основе полных решающих деревьев»*
дан обзор	*«Дан обзор основных результатов, полученных авторами ранее в данной области»*
в том числе со словами «серединных» кластеров	
распознающая процедура	*«Построены модели распознающих процедур, нацеленные на решение задач с неполными данными (с пропусками в признаковых описаниях объектов) и с неравномерным распределением обучающих объектов по классам»*

Соподчинённые слова кластеров наибольших значений TF-IDF

ММРО-14, Баринова О. В., Ветров Д. П.	
обобщающая способность	*«Оценки обобщающей способности бустинга с вероятностными входами»*
в том числе со словами «серединных» кластеров	
ошибка классификации	*«В данной работе предлагается новая верхняя оценка ошибки классификации для композиций простых классификаторов, основанная на сведении бинарной задачи классификации с перекрывающимися распределениями классов к задаче классификации с неперекрывающимися классами»*
ММРО-14, Мельников Д. И., Стрижов В. В., Андреева Е. Ю., Эденхартер Г.	
в том числе со словами «серединных» кластеров	
выбор опорного	*«Выбор опорного множества при построении устойчивых интегральных индикаторов»*
описание объекта	*«Исследуется задача построения интегрального индикатора множества объектов, устойчивого к выбросам в описаниях объектов»*
описаны (в) линейных	*«Объекты описаны в линейных шкалах»*
ИОИ-9, Двоенко С. Д., Пшеничный Д. О.	
матрица парных	*«В интеллектуальном анализе данных результаты экспериментов часто представлены в виде матриц парных сравнений элементов анализируемого множества между собой»*
в том числе со словами «серединных» кластеров	
матрица парных сравнений	*«Об устранении отрицательных собственных значений матриц парных сравнений»*

Соподчинённые слова кластеров наибольших значений TF-IDF

ИОИ-9, Животовский Н. К., Воронцов К. В.	
в том числе со словами «серединных» кластеров	
точность комбинаторных	*«Критерии точности комбинаторных оценок обобщающей способности»*
комбинаторная теория	*«Комбинаторная теория переобучения даёт точные оценки вероятности переобучения для некоторых нетривиальных семейств алгоритмов классификации»*

Замечания

- поскольку заголовок и фразы аннотации несут единый смысловой образ, то вполне допустимо анализировать вхождение слов, отнесенных к кластеру наибольших значений TF-IDF по одной фразе, в связи слов относительно других фраз;
- в настоящей работе набор вышеуказанных связей отождествляется с ключевым сочетанием, если ему соответствует связный подграф дерева синтаксического разбора фразы и минимум одно сочетание слов отвечает условию *Утверждения 1*.

Некоторые технические детали

Для выявления связей слов был задействован *MaltParser* — программный инструмент синтаксического разбора фраз естественных языков и работы с деревьями зависимостей.

Выводы

1. Основной *результат* настоящей работы — *метод* оценки близости текста смысловому эталону относительно тематического текстового корпуса.
2. *Эффективность* метода может быть *оценена* разбиением коллекции на классы по значению оценки близости эталону и отношением числа текстов класса наибольших значений к общему числу текстов коллекции.
3. Предложенный метод даёт *минимум трёхкратное* сокращение *числа документов (научных статей)*, с которыми следует ознакомиться в первую очередь при изучении заданной предметной области.
4. Транзитивность синтаксического отношения на последовательности соподчинённых слов *требует изучения* динамики изменения TF-IDF при переходе от отдельных слов к L-граммам (по К. Шеннону).
5. Расположение слов класса наибольших значений TF-IDF по соседству во фразе — *необходимое, но не достаточное* условие отнесения к ключевым сочетаниям, определяющим смысловой образ текста.
6. Представляет интерес *поиск* ключевых сочетаний слов в аннотациях и заголовках *как основа* назначения в спорных случаях итогового рейтинга и иерархизации статей по значимости для изучения предметной области.