

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

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

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 безызыточности 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 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.

- Построение иерархических тематических моделей крупных конференций [[Стрижов В. В., 2014](#)].
- Обнаружение 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

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

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

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

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 отдельно 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 $TF \cdot IDF$) 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 $\text{tf}(t_i, d) \cdot \text{idf}(t_i, D)$ для всех слов t_i исходной фразы относительно документа $d \in D$.

F — последовательность кластеров H_1, \dots, 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.

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:

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

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

и, соответственно,

$$val_3 = |H_1 \setminus H_{r/2} \setminus H_r|/\text{len}(X), \quad (5)$$

где Σ_{H_1} есть сумма величин TF-IDF слов, отнесённых к кластеру H_1 относительно $d \in D$;
 $\sigma(|H_i, i=\{1, r/2, r\}|)$ — СКО числа элементов в кластере из списка $\{H_1, H_{r/2}, H_r\}$;
 $\text{len}(X)$ — 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 $|H_{r/2}|$ и $|H_r|$ принимаются равными нулю;
- при ровно двух кластерах по TF-IDF нулевым считается значение $|H_r|$.

Пусть

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

Первый вариант оценки:

$$N_1(Ts, D) = \frac{\max_{d \in D} (val_1(Ts_1, d) \cdot val_2(Ts_1, d) \cdot val_3(Ts_1, d))}{\sigma(\max_{d \in D} (val_1(Ts_i, d) \cdot val_2(Ts_i, d) \cdot val_3(Ts_i, d)), Ts_i \in Ts) + 1}. \quad (6)$$

Здесь:

в числителе — оценка близости эталону заголовка статьи (Ts_1);

первое слагаемое в знаменателе — СКО значения близости эталону по всем $Ts_i \in Ts$.

Замечания

- оценка (6) зависит от подбора корпуса D экспертом;
- введённая оценка не подразумевает сортировку фраз $Ts_i \in Ts$ 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:

$$N_2(Ts, D) = \frac{\max_{d \in D} (val_1(Ts_{\max}, d) \cdot val_2(Ts_{\max}, d) \cdot val_3(Ts_{\max}, d))}{\sigma\left(\max_{d \in D} (val_1(Ts_i, d) \cdot val_2(Ts_i, d) \cdot val_3(Ts_i, d)), Ts_i \in Ts\right) + 1}, \quad (7)$$

где $Ts_{\max} \in Ts$ — 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*.

- 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.

- 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).

- 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 *PyMorphology2*.
- 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

ММРО-15, Статистическая теория обучения	
Автор(ы)	<i>Воронцов К. В., Махина Г. А.</i>
Заголовок статьи	<i>Принцип максимизации зазора для монотонного классификатора ближайшего соседа</i>
Максимум оценки близости эталону для заголовка достигается относительно документа	<i>Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15</i>
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)	<i>Генрихов И. Е., Дюкова Е. В.</i>
Заголовок статьи	<i>Полные решающие деревья в задачах классификации по прецедентам</i>
Максимум оценки близости эталону для заголовка достигается относительно документа	<i>Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15</i>
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

ММРО-14, Методы и модели распознавания и прогнозирования	
Автор(ы)	<i>Барина О. В., Ветров Д. П.</i>
Заголовок статьи	<i>Оценки обобщающей способности бустинга с вероятностными входами</i>
Максимум оценки близости эталону для заголовка достигается относительно документа	<i>Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15</i>
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)	<i>Двоенко С. Д., Пшеничный Д. О.</i>
Заголовок статьи	<i>Об устранении отрицательных собственных значений матриц парных сравнений</i>
Максимум оценки близости эталону для заголовка достигается относительно документа	<i>Двоенко С. Д., Пшеничный Д. О. Метрическая коррекция матриц парных сравнений // ММРО-16</i>
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

ММРО-15, Статистическая теория обучения			
Автор(ы)	<i>Воронцов К. В., Машина Г. А.</i>		
Заголовок статьи	<i>Принцип максимизации зазора для монотонного классификатора ближайшего соседа</i>		
Фраза, максимально близкая эталону	<i>Принцип максимизации зазора для монотонного классификатора ближайшего соседа</i>		
Максимум оценки близости фразы эталону достигается относительно документа	<i>Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15</i>		
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)	<i>Генрихов И. Е., Дюкова Е. В.</i>		
Заголовок статьи	<i>Полные решающие деревья в задачах классификации по прецедентам</i>		
Фраза, максимально близкая эталону	<i>Полные решающие деревья в задачах классификации по прецедентам</i>		
Максимум оценки близости фразы эталону достигается относительно документа	<i>Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15</i>		
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)	<i>Мельников Д. И., Стрижов В. В., Андреева Е. Ю., Эденхарттер Г.</i>		
Заголовок статьи	<i>Выбор опорного множества при построении устойчивых интегральных индикаторов</i>		
Фраза, максимально близкая эталону	<i>Объекты описаны в линейных шкалах</i>		
Максимум оценки близости фразы эталону достигается относительно документа	<i>Абрамов В. И., Середин О. С., Сулимова В. В., Моттль В. В. Эквивалентность потенциальных функций и линейных пространств представления объектов произвольной природы // ИОИ-8</i>		
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)	<i>Животовский Н. К., Воронцов К. В.</i>		
Заголовок статьи	<i>Критерии точности комбинаторных оценок обобщающей способности</i>		
Фраза, максимально близкая эталону	<i>Комбинаторная теория переобучения даёт точные оценки вероятности переобучения для некоторых нетривиальных семейств алгоритмов классификации</i>		
Максимум оценки близости фразы эталону достигается относительно документа	<i>Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15</i>		
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)

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

Не получила максимального итогового рейтинга

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

Относительно оценки (7)

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

¹ Число после дроби — максимум оценки по коллекции

ММО-15, Статистическая теория обучения	
<p>Автор(ы) Слова из кластеров наибольших значений TF-IDF по отдельным фразам Образуемые из них сочетания, отвечающие условию <i>Утверждения 1</i> в т. ч. со словами «серединных» кластеров</p>	<p><i>Воронцов К. В., Мазина Г. А.</i> <i>монотонный, сосед, близкий, скользящий, контроль, обобщать, способность ближайший сосед, скользящий контроль, обобщающая способность разделяющая поверхность</i></p>
ММО-15, Математическая теория и методы классификации	
<p>Автор(ы) Слова из кластеров наибольших значений TF-IDF по отдельным фразам Образуемые из них сочетания, отвечающие условию <i>Утверждения 1</i> в т. ч. со словами «серединных» кластеров</p>	<p><i>Генрихов И. Е., Дюкова Е. В.</i> <i>классификация, полный, решающий, дерево, прецедент, процедура, описание, обзор, дать решающее дерево</i> <i>распознающая процедура</i></p>
ММО-14, Методы и модели распознавания и прогнозирования	
<p>Автор(ы) Слова из кластеров наибольших значений TF-IDF по отдельным фразам Образуемые из них сочетания, отвечающие условию <i>Утверждения 1</i></p>	<p><i>Баринова О. В., Ветров Д. П.</i> <i>обобщать, способность, композиция, ошибка, вероятность, классификация, выборка, верхний, бустинг обобщающая способность</i></p>
ИОИ-9, Математическая теория и методы классификации	
<p>Автор(ы) Слова из кластеров наибольших значений TF-IDF по отдельным фразам Образуемые из них сочетания, отвечающие условию <i>Утверждения 1</i> в т. ч. со словами «серединных» кластеров</p>	<p><i>Двоенко С. Д., Пшеничный Д. О.</i> <i>парный, матрица, численный, измерение, корректный, обработка, следовать матрица парных</i> <i>матрица парных сравнений</i></p>

ММРО-14, Методы и модели распознавания и прогнозирования	
<p>Автор(ы)</p> <p>Слова из кластеров наибольших значений TF-IDF по отдельным фразам</p> <p>Образуемые из них сочетания, отвечающие условию <i>Утверждения 1</i> в том числе со словами «серединых» кластеров</p>	<p><i>Мельников Д. И., Стрижов В. В., Андреева Е. Ю., Эденхартер Г.</i></p> <p><i>опорный, описание, линейный, помощь, учитель, основной, предложить, получение</i></p> <p><i>выбор опорного, описаны (в) линейных</i></p>
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	
<p>Автор(ы)</p> <p>Слова из кластеров наибольших значений TF-IDF по отдельным фразам</p> <p>Образуемые из них сочетания, отвечающие условию <i>Утверждения 1</i> в том числе со словами «серединых» кластеров</p>	<p><i>Животовский Н. К., Воронцов К. В.</i></p> <p><i>обобщать, комбинаторный, вероятность, семейство</i></p>
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

ММРО-15, Статистическая теория обучения, Воронцов К. В., Махина Г. А.	
ближайший сосед скользящий контроль обобщающая способность	<p>«Принцип максимизации зазора для монотонного классификатора ближайшего соседа»</p> <p>«Получены точные оценки полного скользящего контроля для монотонных классификаторов, основанных на принципе ближайшего соседа»</p> <p>«Показано, что наилучшей обобщающей способностью обладает монотонный классификатор, в котором разделяющая поверхность проходит посередине зазора между классами»</p>
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 (полное) решающее дерево дан обзор	<p>«Полные решающие деревья в задачах классификации по прецедентам»</p> <p>«В докладе представлены результаты, полученные авторами, разработки алгоритмов классификации на основе полных решающих деревьев»</p> <p>«Дан обзор основных результатов, полученных авторами ранее в данной области»</p>
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»

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

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 заголовков 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.

- 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.