

Оценка близости смысловому эталону без поиска перифраз и иерархия тематических текстов

Михайлов Д. В., Емельянов Г. М.

Новгородский государственный университет
имени Ярослава Мудрого

13-я Международная конференция
«Интеллектуализация обработки информации» (ИОИ-2020),

8–11 декабря 2020 г.

г. Москва

Требования к решению

- 1 Иерархизация 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 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.

- 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 (де-факто) выделение и анализ взаимосвязей смысловых эталонов отдельных текстовых документов для оценивания их взаимной сложности.

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

При этом элементы X принадлежат одному кластеру, если

$$\begin{cases} |\text{mc}(X) - \text{first}(X)| < \frac{\text{mc}(X)}{4} \\ |\text{mc}(X) - \text{last}(X)| < \frac{\text{mc}(X)}{4} \end{cases} . \quad (3)$$

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.

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:

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

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

и, соответственно,

$$val_3 = |H_1 \setminus H_{r/2} \setminus H_r| / \text{len}(X), \quad (6)$$

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

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

Первый вариант оценки:

$$N_1(\mathbf{T_s}, D) = \frac{\max_{d \in D} (val_1(Ts_1, d) \cdot val_2(Ts_1, d) \cdot val_3(Ts_1, d))}{\sigma(\max_{d \in D} (val_1(Ts_i, d) \cdot val_2(Ts_i, d) \cdot val_3(Ts_i, d)), Ts_i \in \mathbf{T_s}) + 1}. \quad (7)$$

Здесь:

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

первое слагаемое в знаменателе — СКО значения близости эталону по всем $Ts_i \in \mathbf{T_s}$.

Замечания

- оценка (7) зависит от подбора корпуса D экспертом;
- введённая оценка не подразумевает сортировку фраз $Ts_i \in \mathbf{T_s}$ 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:

$$N_2(\mathbf{T}s, D) = \frac{\max_{d \in D} (val_1(Ts_{\max}, d) \cdot val_2(Ts_{\max}, d) \cdot val_3(Ts_{\max}, d))}{\sigma(\max_{d \in D} (val_1(Ts_i, d) \cdot val_2(Ts_i, d) \cdot val_3(Ts_i, d)), Ts_i \in \mathbf{T}s) + 1}, \quad (8)$$

где $Ts_{\max} \in \mathbf{T}s$ — фраза, по которой получен максимум близости эталону.

Утверждение 2

Максимальный итоговый рейтинг по коллекции получает статья с наибольшим значением оценки (7), попадающим в один кластер со значением оценки (8) для той же статьи.

Замечания

- корректное применение *Утверждения 2* предполагает отнесение к одному кластеру значений оценки (7) для статьи с максимальным итоговым рейтингом и максимального значения оценки (7) по коллекции, из которой ведётся отбор;
- в случае отсутствия в коллекции статьи, удовлетворяющей данному требованию, *максимальный итоговый рейтинг* получает статья с наибольшим значением оценки (7) по анализируемой коллекции;
- поскольку заголовок и фразы аннотации (по определению) несут некий единый смысловой образ, то допустима мена местами оценок (7) и (8) в *Утверждении 2*.

Вход: S ; // последовательность текстов исходной коллекции,
// отсортированная по убыванию оценки (7)

Выход: S_{res} ; // результат её ранжирования применением *Утверждения 2*

```
1:  $S_{res} := \emptyset$ ;  
2: пока  $S \neq \emptyset$   
3:    $Flag := false$ ;  
4:   для всех  $Ts \in S$   
5:      $Tmp := \{N_1(\text{first}(S), D), N_1(Ts, D), N_2(\text{first}(S), D)\}$ ;  
6:     отсортировать  $Tmp$  по убыванию;  
7:     если  $good(Tmp) = true$  то  
8:        $Flag := true$ ;  
9:        $S_{res} := S_{res} \odot \{Ts\}$ ; //  $\odot$  — операция конкатенации  
10:       $S := S \setminus \{Ts\}$ ;  
11:      выход из цикла {для}  
12:    конец если  
13:  конец для  
14:  если  $Flag = false$  то  
15:     $S_{res} := S_{res} \odot \{\text{first}(S)\}$ ;  
16:     $S := S \setminus \{\text{first}(S)\}$ ;  
17:  конец если  
18: конец пока
```

Здесь:

$good$ — функция, выдающая $true/false$ в зависимости от выполнения условия (3);

$first$ — функция, возвращающая первый элемент заданной последовательности.

Введём обозначения:

$\mathbf{H}_1(Ts_i)$, $\mathbf{H}_{r/2}(Ts_i)$ и $\mathbf{H}_r(Ts_i)$ — множества слов кластеров H_1 , $H_{r/2}$ и H_r , соответственно, для фразы $Ts_i \in \mathbf{T}s$ относительно документа $d \in D$, по которому получен максимум близости эталону, $\mathbf{T}s \in S_{res}$;

$$\mathbf{H}_1(\mathbf{T}s) = \bigcup_{Ts_i \in \mathbf{T}s} \mathbf{H}_1(Ts_i);$$

$\mathbf{H}_{\bar{Z}}(Ts_i)$ — множество слов фразы Ts_i с ненулевыми значениями TF-IDF относительно того же документа d ;

$$\mathbf{H}_{\bar{Z}}(\mathbf{T}s) = \bigcup_{Ts_i \in \mathbf{T}s} (\mathbf{H}_{\bar{Z}}(Ts_i) \setminus \mathbf{H}_1(Ts_i)).$$

Пусть $\mathbf{T}s_i$ и $\mathbf{T}s_j$ — тексты из входящих в S_{res} , причём $i > j$, то есть рейтинг статьи, отвечающей группе фраз $\mathbf{T}s_i$, выше, чем по $\mathbf{T}s_j$.

Основная гипотеза

Мера, в которой текст $\mathbf{T}s_j$ дополняется по смыслу текстом $\mathbf{T}s_i$, соответствует величине $|(\mathbf{H}_{\bar{Z}}(\mathbf{T}s_j) \setminus \mathbf{H}_1(\mathbf{T}s_j)) \cap \mathbf{H}_1(\mathbf{T}s_i)|$.

Сама дополняемость текста $\mathbf{T}s_j$ текстом $\mathbf{T}s_i$ определяется как

$$K_1(\mathbf{T}s_j, \mathbf{T}s_i) = \frac{|(\mathbf{H}_{\bar{Z}}(\mathbf{T}s_j) \setminus \mathbf{H}_1(\mathbf{T}s_j)) \cap \mathbf{H}_1(\mathbf{T}s_i)|}{|\mathbf{H}_1(\mathbf{T}s_i)|}. \quad (9)$$

Пусть

$\mathbf{Kw}(\mathbf{Ts}_i)$ — множество ключевых сочетаний слов, отвечающих условию *Утверждения 1* и найденных для \mathbf{Ts}_i ;

$\mathbf{H}_{\mathbf{Kw}}(\mathbf{Ts}_i)$ — множество слов в составе указанных сочетаний.

Введём в рассмотрение $\mathbf{Kw}'(\mathbf{Ts}_j, \mathbf{Ts}_i) \subset \mathbf{Kw}(\mathbf{Ts}_i)$, куда войдут сочетания слов множества $\mathbf{H}_{\bar{Z}}(Ts_{jk}) \setminus \mathbf{H}_1(\mathbf{Ts}_j)$ по каждой фразе $Ts_{jk} \in \mathbf{Ts}_j$, причём для каждого сочетания минимум одно слово должно принадлежать $\mathbf{H}_1(\mathbf{Ts}_i)$.

С учётом искомых сочетаний слов оценка (9) принимает следующий вид:

$$K_2(\mathbf{Ts}_j, \mathbf{Ts}_i) = \frac{|\mathbf{Kw}'(\mathbf{Ts}_j, \mathbf{Ts}_i)| + |((\mathbf{H}_{\bar{Z}}(\mathbf{Ts}_j) \setminus \mathbf{H}_1(\mathbf{Ts}_j)) \cap \mathbf{H}_1(\mathbf{Ts}_i)) \setminus \mathbf{H}_{\mathbf{Kw}'}(\mathbf{Ts}_j, \mathbf{Ts}_i)|}{|\mathbf{Kw}(\mathbf{Ts}_i)| + |\mathbf{H}_1(\mathbf{Ts}_i) \setminus \mathbf{H}_{\mathbf{Kw}}(\mathbf{Ts}_i)|}, \quad (10)$$

где $\mathbf{H}_{\mathbf{Kw}'}(\mathbf{Ts}_j, \mathbf{Ts}_i)$ — множество слов в составе сочетаний из $\mathbf{Kw}'(\mathbf{Ts}_j, \mathbf{Ts}_i)$.

Представленность слов фразы $Ts_{jk} \in \mathbf{T}s_j$ в кластерах $\{H_1, H_{r/2}, H_r\} := Cl$:

$$N(Ts_{jk}, Cl(Ts_{jk})) = \frac{\sqrt{\sum_{m \in \{1, r/2, r\}} (|\mathbf{H}_m(Ts_{jk})| / \text{len}(Ts_{jk}))^2}}{\sigma(|\mathbf{H}_m(Ts_{jk})| / \text{len}(Ts_{jk})) + 1}, \quad (11)$$

где $\text{len}(Ts_{jk})$ — число слов во фразе Ts_{jk} .

Замечания

- если число полученных по TF-IDF кластеров меньше двух, то величины $|\mathbf{H}_{r/2}(Ts_{jk})|$ и $|\mathbf{H}_r(Ts_{jk})|$ принимаются равными нулю;
- при ровно двух кластерах по TF-IDF нулевым считается значение $|\mathbf{H}_r(Ts_{jk})|$.

Определим дополнение эталона текста $\mathbf{T}s_j$ эталоном для $\mathbf{T}s_i$ введением

$$\begin{aligned} \mathbf{H}'_1(Ts_{jk}, \mathbf{T}s_i) &= \mathbf{H}_1(Ts_{jk}) \cup ((\mathbf{H}_{\bar{Z}}(Ts_{jk}) \setminus \mathbf{H}_1(Ts_{jk})) \cap \mathbf{H}_1(\mathbf{T}s_i)); \\ \mathbf{H}'_{r/2}(Ts_{jk}, \mathbf{T}s_i) &= \mathbf{H}_{r/2}(Ts_{jk}) \setminus ((\mathbf{H}_{\bar{Z}}(Ts_{jk}) \setminus \mathbf{H}_1(Ts_{jk})) \cap \mathbf{H}_1(\mathbf{T}s_i)); \\ \mathbf{H}'_r(Ts_{jk}, \mathbf{T}s_i) &= \mathbf{H}_r(Ts_{jk}) \setminus ((\mathbf{H}_{\bar{Z}}(Ts_{jk}) \setminus \mathbf{H}_1(Ts_{jk})) \cap \mathbf{H}_1(\mathbf{T}s_i)). \end{aligned}$$

в оценку (11), которая при этом примет следующий вид:

$$N'(Ts_{jk}, Cl(Ts_{jk}), \mathbf{T}s_i) = \frac{\sqrt{\sum_{m \in \{1, r/2, r\}} (|\mathbf{H}'_m(Ts_{jk}, \mathbf{T}s_i)| / \text{len}(Ts_{jk}))^2}}{\sigma(|\mathbf{H}'_m(Ts_{jk}, \mathbf{T}s_i)| / \text{len}(Ts_{jk})) + 1}. \quad (12)$$

$$N_3(\mathbf{T}s_j) = \frac{N(Ts_{j1}, Cl(Ts_{j1}))}{\sigma(\{N(Ts_{jk}, Cl(Ts_{jk})) : Ts_{jk} \in \mathbf{T}s_j\}) + 1}, \quad (13)$$

$$N_4(\mathbf{T}s_j) = \frac{\max[\{N(Ts_{jk}, Cl(Ts_{jk})) : Ts_{jk} \in \mathbf{T}s_j\}]}{\sigma(\{N(Ts_{jk}, Cl(Ts_{jk})) : Ts_{jk} \in \mathbf{T}s_j\}) + 1}, \quad (14)$$

$$N'_3(\mathbf{T}s_j, \mathbf{T}s_i) = \frac{N'(Ts_{j1}, Cl(Ts_{j1}), \mathbf{T}s_i)}{\sigma(\{N'(Ts_{jk}, Cl(Ts_{jk}), \mathbf{T}s_i) : Ts_{jk} \in \mathbf{T}s_j\}) + 1}, \quad (15)$$

$$N'_4(\mathbf{T}s_j, \mathbf{T}s_i) = \frac{\max[\{N'(Ts_{jk}, Cl(Ts_{jk}), \mathbf{T}s_i) : Ts_{jk} \in \mathbf{T}s_j\}]}{\sigma(\{N'(Ts_{jk}, Cl(Ts_{jk}), \mathbf{T}s_i) : Ts_{jk} \in \mathbf{T}s_j\}) + 1}. \quad (16)$$

Утверждение 3

Критерием выбора вышестоящего текста $\mathbf{T}s_i$ для заданного текста $\mathbf{T}s_j$ в формируемой иерархии является неубывание значений оценок (15) и (16) по отношению к соответствующим им оценкам (13) и (14) при максимизации оценок (9) и (10).

Замечание

Назовём далее первые слагаемые в знаменателях формул (13) и (14) как СКО оценки (11), а в знаменателях формул (15) и (16) — как СКО оценки (12), соответственно.

- 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 (4)–(8) — 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. Ранжирование статей согласно алгоритму на *Слайде 10* относительно оценки (7).

№	Автор (ы) и заголовок статьи	Оценка (7)	Оценка (8)
1	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	0,07112036	0,07112036
2	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	0,05185727	0,05185727
3	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	0,05169631	0,05169631
4	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	0,03992817	0,03992817
5	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	0,02178213	0,02178213
6	Каневский Д. Ю. Переобучение и комбинаторная радемаховская сложность в задачах восстановления регрессии	0,01969541	0,01969541
7	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	0,01851287	0,01851287
8	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	0,01731464	0,01731464
9	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	0,01591723	0,01591723
10	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	0,00285329	0,03573024

Таблица 2. Ранжирование статей согласно алгоритму на Слайде 10 относительно оценки (8).

№	Автор (ы) и заголовок статьи	Оценка (8)	Оценка (7)
1	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	0,07112036	0,07112036
2	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	0,05185727	0,05185727
3	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	0,05169631	0,05169631
4	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	0,03992817	0,03992817
5	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	0,03573024	0,00285329
6	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	0,02178213	0,02178213
7	Каневский Д. Ю. Переобучение и комбинаторная радема-хервская сложность в задачах восстановления регрессии	0,01969541	0,01969541
8	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	0,01851287	0,01851287
9	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	0,01731464	0,01731464
10	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	0,01591723	0,01591723

Таблица 3. Дополняемость текстов по смыслу без учёта ключевых сочетаний слов ¹.

$j = 2, i = 1$	Оценка (9)	0,42857143
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>обобщать, монотонный, способность</i>	
$H_1(Ts_j)$	<i>контроль, скользящий, выборка</i>	
$H_1(Ts_i)$	<i>контроль, монотонный, скользящий, обобщать, способность, близкий, сосед</i>	
$j = 6, i = 1$	Оценка (9)	0,28571429
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>обобщать, способность</i>	
$H_1(Ts_j)$	<i>комбинаторный, семейство, обучать, завышенность</i>	
$j = 9, i = 1$	Оценка (9)	0,14285714
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>монотонный</i>	
$H_1(Ts_j)$	<i>связность, переобучение</i>	
$j = 4, i = 3$	Оценка (9)	1,00000000
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>эмпирический</i>	
$H_1(Ts_j)$	<i>минимизация, обобщать, способность, риск, комбинаторный</i>	
$H_1(Ts_i)$	<i>эмпирический</i>	

¹ Здесь и далее i и j — порядковые номера документов по Таблице 1.

Продолжение таблицы 3.

$j = 6, i = 4$	Оценка (9)	0,40000000
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>обобщать, способность</i>	
$H_1(Ts_j)$	<i>комбинаторный, семейство, обучать, завышенность</i>	
$H_1(Ts_i)$	<i>минимизация, обобщать, способность, риск, комбинаторный</i>	
$j = 8, i = 4$	Оценка (9)	0,40000000
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>минимизация, риск</i>	
$H_1(Ts_j)$	<i>семейство, статистический, способность, комбинаторный, обобщать, вероятность</i>	
$j = 9, i = 4$	Оценка (9)	0,20000000
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>комбинаторный</i>	
$H_1(Ts_j)$	<i>связность, переобучение</i>	
$j = 9, i = 5$	Оценка (9)	0,20000000
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>комбинаторный</i>	
$H_1(Ts_j)$	<i>связность, переобучение</i>	
$H_1(Ts_i)$	<i>комбинаторный, связность, контроль, скользящий, выборка</i>	

Здесь и далее в Таблицах 5 и 6 графы для связей, отвечающих условию Утверждения 3, выделены зелёным цветом; для частично отвечающих данному условию — жёлтым.

Окончание таблицы 3.

$j = 9, i = 6$	Оценка (9)	0,25000000
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>комбинаторный</i>	
$H_1(Ts_j)$	<i>связность, переобучение</i>	
$H_1(Ts_i)$	<i>комбинаторный, семейство, обучать, завышенность</i>	
$j = 8, i = 7$	Оценка (9)	0,33333333
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>риск</i>	
$H_1(Ts_j)$	<i>семейство, статистический, способность, комбинаторный, обобщать, вероятность</i>	
$H_1(Ts_i)$	<i>достоинство, риск, эмпирический</i>	
$j = 9, i = 8$	Оценка (9)	0,33333333
$(H_{\bar{z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$	<i>комбинаторный, вероятность</i>	
$H_1(Ts_j)$	<i>связность, переобучение</i>	
$H_1(Ts_i)$	<i>семейство, статистический, способность, комбинаторный, обобщать, вероятность</i>	

Замечания

- связь документов исключается из рассмотрения, если значения оценок (9) и (10) одновременно равны нулю;
- оценка (10) вычисляется только тогда, когда $|Kw(Ts_i)| > 0$;
- при $|Kw(Ts_i)| = 0$ связь не рассматривается при нулевом значении оценки (9).

Таблица 4. Дополняемость текстов по смыслу с учётом ключевых сочетаний слов.

$j = 2, i = 1$	Оценка (10)	0,40000000
$Kw(Ts_i)$	<i>ближайший сосед, скользящий контроль, обобщающая способность, разделяющая поверхность</i>	
$Kw'(Ts_j, Ts_i)$	<i>обобщающая способность</i>	
$j = 6, i = 1$	Оценка (10)	0,20000000
$Kw'(Ts_j, Ts_i)$	<i>обобщающая способность</i>	
$j = 9, i = 1$	Оценка (10)	0,20000000
$Kw'(Ts_j, Ts_i)$	—	
$j = 6, i = 4$	Оценка (10)	0,16666667
$Kw(Ts_i)$	<i>обобщающая способность, комбинаторная теория, минимизация эмпирического риска</i>	
$Kw'(Ts_j, Ts_i)$	<i>обобщающая способность</i>	
$j = 8, i = 4$	Оценка (10)	0,33333333
$Kw'(Ts_j, Ts_i)$	—	
$j = 9, i = 4$	Оценка (10)	0,16666667
$Kw'(Ts_j, Ts_i)$	—	
$j = 9, i = 5$	Оценка (10)	0,16666667
$Kw(Ts_i)$	<i>скользящий контроль</i>	
$Kw'(Ts_j, Ts_i)$	—	
$j = 9, i = 8$	Оценка (10)	0,40000000
$Kw(Ts_i)$	<i>обобщающая способность</i>	
$Kw'(Ts_j, Ts_i)$	—	

Таблица 5. Оценивание представленности слов в трёх наиболее значимых для эталона кластерах.

j	$N_3(Ts_j)$	$N_4(Ts_j)$	СКО оценки (11)
2	0,442059165587	0,502119662613	0,048678362277
4	0,376446598212	0,529404830202	0,066634500491
6	0,362818302898	0,504283491203	0,106699761390
8	0,452293583860	0,452293583860	0,058355201581
9	0,346816072806	0,346816072806	0,021096101739

Таблица 6. Оценивание представленности слов в трёх наиболее значимых для эталона кластерах с учётом связей документов.

$j \rightarrow i$	$N'_3(Ts_j, Ts_i)$	$N'_4(Ts_j, Ts_i)$	СКО оценки (12)
2 \rightarrow 1	0,528411029776	0,537246561748	0,0387975635841
6 \rightarrow 1	0,365859769250	0,508510844834	0,0974995421573
9 \rightarrow 1	0,346022664877	0,346022664877	0,0234374100578
4 \rightarrow 3	0,457707643226	0,554500713867	0,0458362979168
6 \rightarrow 4	0,365859769250	0,508510844834	0,0974995421573
8 \rightarrow 4	0,457175036510	0,457175036510	0,0470546921663
9 \rightarrow 4	0,346022664877	0,346022664877	0,0234374100578
9 \rightarrow 5	0,346022664877	0,346022664877	0,0234374100578
9 \rightarrow 6	0,346022664877	0,346022664877	0,0234374100578
8 \rightarrow 7	0,454613088142	0,454613088142	0,0529553143232
9 \rightarrow 8	0,341968227624	0,376375190389	0,0355714693817

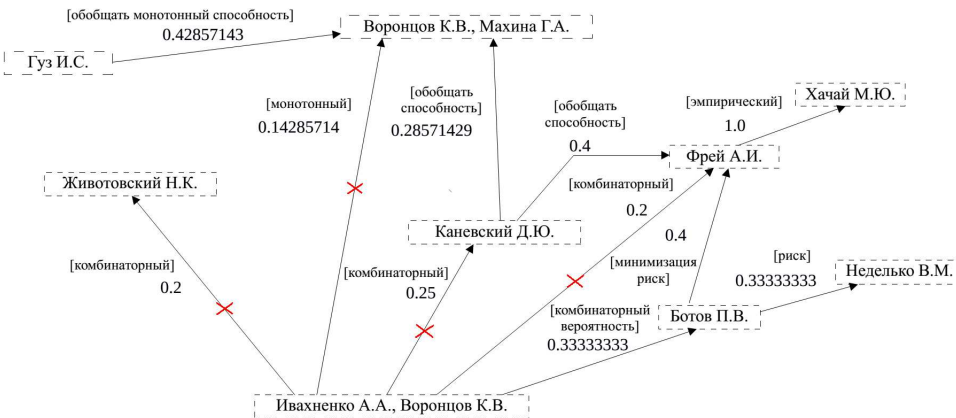


Рис. 1. Иерархизация документов без учёта ключевых сочетаний слов ^{2, 3}.

² В квадратных скобках по каждой связи указаны слова из $(H_{\frac{1}{Z}}(Ts_j) \setminus H_1(Ts_j)) \cap H_1(Ts_i)$.

³ Дуги для связей, не отвечающих условию Утверждения 3, выделены значком «X».

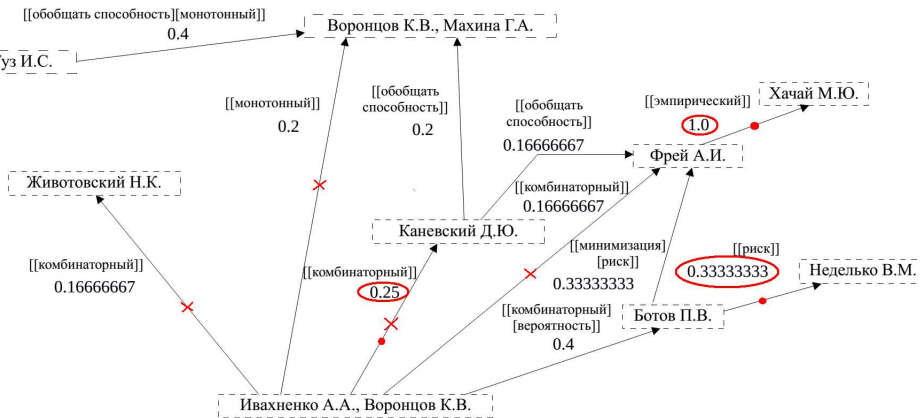


Рис. 2. Иерархизация документов с учётом ключевых сочетаний слов^{4, 5}.

⁴ При $|Kw(Ts_i)| = 0$ дуги выделяются значком «•», а на них указаны значения оценки (9).

⁵ Дуги для связей, не отвечающих условию Утверждения 3, выделены значком «×».

- 1 Основной *результат* настоящей работы — *методика* иерархизации текстов предметно-ограниченного естественного языка на основе оценок близости тематического текста смысловому эталону.
- 2 *Эффективность* решения может быть *оценена* по числу и виду компонент связности графа, полученного из графа связей между документами коллекции путём замены ориентированных рёбер неориентированными.
- 3 После удаления из исходного графа связей, не отвечающих условию *Утверждения 3*, *подграф* для максимальной компоненты связности *в числе вершин* с максимальной степенью *будет содержать* вершину статьи с максимальным итоговым рейтингом по коллекции.
- 4 За счёт статей, не отражаемых максимальной компонентой связности, *минимум на 20%* сокращается *число документов*, рассматриваемых в первую очередь при изучении заданной предметной области.
- 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 (4), (5) и (6) относительно конкретных документов заданного текстового корпуса.