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

Текст научной работы и рациональная (эталонная) передача смысла

Требования к решению

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

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

Аннотация и заголовок научной работы

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

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 (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

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

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

При этом элементы X *принадлежат одному кластеру*, если

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

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

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

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

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

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:

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

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

и, соответственно,

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

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

Замечания

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

Оценка близости эталону для группы фраз: заголовок + аннотация

Пусть

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

Первый вариант оценки:

$$\mathrm{N}_1(\mathbf{Ts}, D) = \frac{\max\limits_{d \in D}\left(val_1(Ts_1, d) \cdot val_2(Ts_1, d) \cdot val_3(Ts_1, d)\right)}{\sigma\left(\max\limits_{d \in D}\left(val_1(Ts_i, d) \cdot val_2(Ts_i, d) \cdot val_3(Ts_i, d)\right), Ts_i \in \mathbf{Ts}\right) + 1}. \quad (7)$$

Здесь:

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

Замечания

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

$$\mathrm{N}_2\left(\mathbf{Ts}, D\right) = \frac{\max\limits_{d \in D}\left(val_1\left(Ts_{\max}, d\right) \cdot val_2\left(Ts_{\max}, d\right) \cdot val_3\left(Ts_{\max}, d\right)\right)}{\sigma\left(\max\limits_{d \in D}\left(val_1\left(Ts_i, d\right) \cdot val_2\left(Ts_i, d\right) \cdot val_3\left(Ts_i, d\right)\right), Ts_i \in \mathbf{Ts}\right) + 1}, \quad (8)$$

где $Ts_{\max} \in \mathbf{Ts}$ — 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*.

Ранжирование текстов исходной коллекции

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

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

1: $S_{res} := \varnothing$;
2: **пока** $S \neq \varnothing$
3: $Flag := \text{false}$;
4: **для всех** $\mathbf{Ts} \in S$
5: $Tmp := \{\mathrm{N}_1(\text{first}(S), D), \mathrm{N}_1(\mathbf{Ts}, D), \mathrm{N}_2(\text{first}(S), D)\}$;
6: отсортировать Tmp по убыванию;
7: **если** $\text{good}(Tmp) = \text{true}$ **то**
8: $Flag := \text{true}$;
9: $S_{res} := S_{res} \odot \{\mathbf{Ts}\}$; // $\odot$ — операция конкатенации
10: $S := S \setminus \{\mathbf{Ts}\}$;
11: **выход из цикла** *{для}*
12: **конец если**
13: **конец для**
14: **если** $Flag = \text{false}$ **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 $Ts_i \in \mathbf{Ts}$ относительно документа $d \in D$, по которому получен максимум близости эталону, $\mathbf{Ts} \in S_{res}$;

$\mathbf{H_1}(\mathbf{Ts}) = \bigcup_{Ts_i \in \mathbf{Ts}} \mathbf{H}_1(Ts_i)$;

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

$\mathbf{H}_{\bar{\mathbf{Z}}}(\mathbf{Ts}) = \bigcup_{Ts_i \in \mathbf{Ts}} \left(\mathbf{H}_{\overline{\mathbf{Z}}}(Ts_i) \setminus \mathbf{H}_1(Ts_i)\right)$.

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

Основная гипотеза

Мера, в которой текст $\mathbf{Ts}_j$ дополняется по смыслу текстом $\mathbf{Ts}_i$, соответствует величине $\left|\left(\mathbf{H}_{\bar{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)\right|$.

Сама дополняемость текста $\mathbf{Ts}_j$ текстом $\mathbf{Ts}_i$ определяется как

$$K_1(\mathbf{Ts}_j, \mathbf{Ts}_i) = \frac{\left|\left(\mathbf{H}_{\bar{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)\right|}{\left|\mathbf{H_1}(\mathbf{Ts}_i)\right|}. \quad (9)$$

Пусть

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

$\mathbf{H_{Kw}}\,(\mathbf{Ts}_i)$ — 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 $\mathbf{H}_{\overline{\mathbf{Z}}}(Ts_{jk}) \setminus \mathbf{H_1}\,(\mathbf{Ts}_j)$ 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:

$$K_2(\mathbf{Ts}_j, \mathbf{Ts}_i) = \\ = \frac{\left|\mathbf{Kw'}(\mathbf{Ts}_j, \mathbf{Ts}_i)\right| + \left|\left(\left(\mathbf{H}_{\overline{\mathbf{Z}}}\,(\mathbf{Ts}_j) \setminus \mathbf{H_1}\,(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}\,(\mathbf{Ts}_i)\right) \setminus \mathbf{H_{Kw'}}\,(\mathbf{Ts}_j, \mathbf{Ts}_i)\right|}{\left|\mathbf{Kw}(\mathbf{Ts}_i)\right| + \left|\mathbf{H_1}\,(\mathbf{Ts}_i) \setminus \mathbf{H_{Kw}}\,(\mathbf{Ts}_i)\right|}, \quad (10)$$

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

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

$$\mathrm{N}\Big(Ts_{jk}, Cl\left(Ts_{jk}\right)\Big) = \frac{\sqrt{\sum_{\mathrm{m}\in\{1,\mathrm{r}/2,\mathrm{r}\}}\Big(\left|\mathbf{H}_{\mathrm{m}}\left(Ts_{jk}\right)\right|/\mathrm{len}\left(Ts_{jk}\right)\Big)^2}}{\sigma\Big(\left|\mathbf{H}_{\mathrm{m}}\left(Ts_{jk}\right)\right|/\mathrm{len}\left(Ts_{jk}\right)\Big)+1}, \qquad (11)$$

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

Замечания

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

Определим дополнение эталона текста $\mathbf{Ts}_j$ эталоном для $\mathbf{Ts}_i$ введением

$$\mathbf{H}'_1\left(Ts_{jk}, \mathbf{Ts}_i\right) = \mathbf{H}_1\left(Ts_{jk}\right) \cup \left(\left(\mathbf{H}_{\overline{\mathrm{Z}}}\left(Ts_{jk}\right) \setminus \mathbf{H}_1\left(Ts_{jk}\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)\right);$$
$$\mathbf{H}'_{\mathrm{r}/2}\left(Ts_{jk}, \mathbf{Ts}_i\right) = \mathbf{H}_{\mathrm{r}/2}\left(Ts_{jk}\right) \setminus \left(\left(\mathbf{H}_{\overline{\mathrm{Z}}}\left(Ts_{jk}\right) \setminus \mathbf{H}_1\left(Ts_{jk}\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)\right);$$
$$\mathbf{H}'_{\mathrm{r}}\left(Ts_{jk}, \mathbf{Ts}_i\right) = \mathbf{H}_{\mathrm{r}}\left(Ts_{jk}\right) \setminus \left(\left(\mathbf{H}_{\overline{\mathrm{Z}}}\left(Ts_{jk}\right) \setminus \mathbf{H}_1\left(Ts_{jk}\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)\right).$$

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

$$\mathrm{N}'\Big(Ts_{jk}, Cl\left(Ts_{jk}\right), \mathbf{Ts}_i\Big) = \frac{\sqrt{\sum_{\mathrm{m}\in\{1,\mathrm{r}/2,\mathrm{r}\}}\Big(\left|\mathbf{H}'_{\mathrm{m}}\left(Ts_{jk}, \mathbf{Ts}_i\right)\right|/\mathrm{len}\left(Ts_{jk}\right)\Big)^2}}{\sigma\Big(\left|\mathbf{H}'_{\mathrm{m}}\left(Ts_{jk}, \mathbf{Ts}_i\right)\right|/\mathrm{len}\left(Ts_{jk}\right)\Big)+1}, \qquad (12)$$

Выбор вышестоящего документа в формируемой иерархии

$$\mathrm{N}_3\left(\mathbf{Ts}_j\right) = \frac{\mathrm{N}\left(Ts_{j1}, Cl\left(Ts_{j1}\right)\right)}{\sigma\left(\left\{\mathrm{N}\left(Ts_{jk}, Cl\left(Ts_{jk}\right)\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right) + 1}, \tag{13}$$

$$\mathrm{N}_4\left(\mathbf{Ts}_j\right) = \frac{\max\left[\left\{\mathrm{N}\left(Ts_{jk}, Cl\left(Ts_{jk}\right)\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right]}{\sigma\left(\left\{\mathrm{N}\left(Ts_{jk}, Cl\left(Ts_{jk}\right)\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right) + 1}, \tag{14}$$

$$\mathrm{N}'_3\left(\mathbf{Ts}_j, \mathbf{Ts}_i\right) = \frac{\mathrm{N}'\left(Ts_{j1}, Cl\left(Ts_{j1}\right), \mathbf{Ts}_i\right)}{\sigma\left(\left\{\mathrm{N}'\left(Ts_{jk}, Cl\left(Ts_{jk}\right), \mathbf{Ts}_i\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right) + 1}, \tag{15}$$

$$\mathrm{N}'_4\left(\mathbf{Ts}_j, \mathbf{Ts}_i\right) = \frac{\max\left[\left\{\mathrm{N}'\left(Ts_{jk}, Cl\left(Ts_{jk}\right), \mathbf{Ts}_i\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right]}{\sigma\left(\left\{\mathrm{N}'\left(Ts_{jk}, Cl\left(Ts_{jk}\right), \mathbf{Ts}_i\right) : Ts_{jk} \in \mathbf{Ts}_j\right\}\right) + 1}, \tag{16}$$

Утверждение 3

Критерием выбора вышестоящего текста $\mathbf{Ts}_i$ для заданного текста $\mathbf{Ts}_j$ 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.

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 (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 радемахеровская 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 радемахеровская 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
$\left(\mathbf{H}_{\neg \mathbf{z}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*обобщать, монотонный, способность*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*контроль, скользящий, выборка*
$\mathbf{H_1}(\mathbf{Ts}_i)$	*контроль, монотонный, скользящий, обобщать, способность, близкий, сосед*
$j = 6, i = 1$	Оценка (9) 0,28571429
$\left(\mathbf{H}_{\neg \mathbf{z}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*обобщать, способность*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*комбинаторный, семейство, обучать, завышенность*
$j = 9, i = 1$	Оценка (9) 0,14285714
$\left(\mathbf{H}_{\neg \mathbf{z}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*монотонный*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*связность, переобучение*
$j = 4, i = 3$	Оценка (9) 1,00000000
$\left(\mathbf{H}_{\neg \mathbf{z}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*эмпирический*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*минимизация, обобщать, способность, риск, комбинаторный*
$\mathbf{H_1}(\mathbf{Ts}_i)$	*эмпирический*

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

Продолжение таблицы 3.

$j = 6, i = 4$	Оценка (9) 0,40000000
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*обобщать, способность*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*комбинаторный, семейство, обучать, завышенность*
$\mathbf{H_1}(\mathbf{Ts}_i)$	*минимизация, обобщать, способность, риск, комбинаторный*
$j = 8, i = 4$	Оценка (9) 0,40000000
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*минимизация, риск*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*семейство, статистический, способность, комбинаторный, обобщать, вероятность*
$j = 9, i = 4$	Оценка (9) 0,20000000
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*комбинаторный*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*связность, переобучение*
$j = 9, i = 5$	Оценка (9) 0,20000000
$\left(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{Ts}_j) \setminus \mathbf{H_1}(\mathbf{Ts}_j)\right) \cap \mathbf{H_1}(\mathbf{Ts}_i)$	*комбинаторный*
$\mathbf{H_1}(\mathbf{Ts}_j)$	*связность, переобучение*
$\mathbf{H_1}(\mathbf{Ts}_i)$	*комбинаторный, связность, контроль, скользящий, выборка*

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
$\left(\mathbf{H}_{\neg \mathbf{Z}}\left(\mathbf{Ts}_j\right) \setminus \mathbf{H_1}\left(\mathbf{Ts}_j\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*комбинаторный*
$\mathbf{H_1}\left(\mathbf{Ts}_j\right)$	*связность, переобучение*
$\mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*комбинаторный, семейство, обучать, завышенность*
$j = 8, i = 7$	Оценка (9) 0,33333333
$\left(\mathbf{H}_{\neg \mathbf{Z}}\left(\mathbf{Ts}_j\right) \setminus \mathbf{H_1}\left(\mathbf{Ts}_j\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*риск*
$\mathbf{H_1}\left(\mathbf{Ts}_j\right)$	*семейство, статистический, способность, комбинаторный, обобщать, вероятность*
$\mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*достоинство, риск, эмпирический*
$j = 9, i = 8$	Оценка (9) 0,33333333
$\left(\mathbf{H}_{\neg \mathbf{Z}}\left(\mathbf{Ts}_j\right) \setminus \mathbf{H_1}\left(\mathbf{Ts}_j\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*комбинаторный, вероятность*
$\mathbf{H_1}\left(\mathbf{Ts}_j\right)$	*связность, переобучение*
$\mathbf{H_1}\left(\mathbf{Ts}_i\right)$	*семейство, статистический, способность, комбинаторный, обобщать, вероятность*

Замечания

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

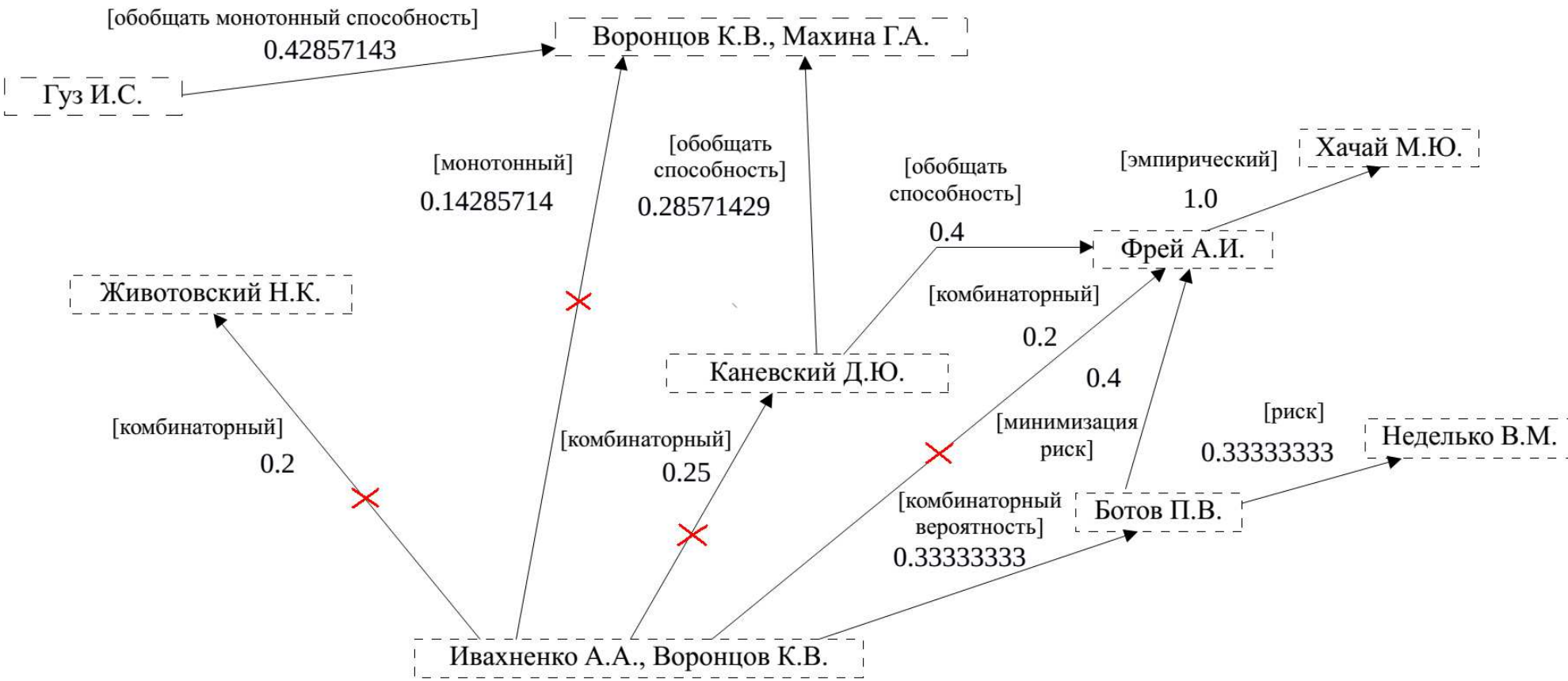

Рис. 1. Иерархизация документов без учёта ключевых сочетаний слов [2, 3].

[2] В квадратных скобках по каждой связи указаны слова из $\left(\mathbf{H}_{\overline{\mathbf{Z}}}\left(\mathbf{Ts}_j\right) \setminus \mathbf{H_1}\left(\mathbf{Ts}_j\right)\right) \cap \mathbf{H_1}\left(\mathbf{Ts}_i\right)$.

[3] Дуги для связей, не отвечающих условию *Утверждения 3*, выделены значком «×».

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

$j = 2, i = 1$	Оценка (10)	0,40000000
$\mathbf{Kw}\,(\mathbf{Ts}_i)$	*ближайший сосед, скользящий контроль, обобщающая способность, разделяющая поверхность*	
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	*обобщающая способность*	
$j = 6, i = 1$	Оценка (10)	0,20000000
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	*обобщающая способность*	
$j = 9, i = 1$	Оценка (10)	0,20000000
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	—	
$j = 6, i = 4$	Оценка (10)	0,16666667
$\mathbf{Kw}\,(\mathbf{Ts}_i)$	*обобщающая способность, комбинаторная теория, минимизация эмпирического риска*	
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	*обобщающая способность*	
$j = 8, i = 4$	Оценка (10)	0,33333333
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	—	
$j = 9, i = 4$	Оценка (10)	0,16666667
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	—	
$j = 9, i = 5$	Оценка (10)	0,16666667
$\mathbf{Kw}\,(\mathbf{Ts}_i)$	*скользящий контроль*	
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	—	
$j = 9, i = 8$	Оценка (10)	0,40000000
$\mathbf{Kw}\,(\mathbf{Ts}_i)$	*обобщающая способность*	
$\mathbf{Kw}'\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	—	

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

j	$\mathrm{N}_3\,(\mathbf{Ts}_j)$	$\mathrm{N}_4\,(\mathbf{Ts}_j)$	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.

$j \to i$	$\mathrm{N}'_3\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	$\mathrm{N}'_4\,(\mathbf{Ts}_j, \mathbf{Ts}_i)$	СКО оценки (12)
$2 \to 1$	0,528411029776	0,537246561748	0,0387975635841
$6 \to 1$	0,365859769250	0,508510844834	0,0974995421573
$9 \to 1$	0,346022664877	0,346022664877	0,0234374100578
$4 \to 3$	0,457707643226	0,554500713867	0,0458362979168
$6 \to 4$	0,365859769250	0,508510844834	0,0974995421573
$8 \to 4$	0,457175036510	0,457175036510	0,0470546921663
$9 \to 4$	0,346022664877	0,346022664877	0,0234374100578
$9 \to 5$	0,346022664877	0,346022664877	0,0234374100578
$9 \to 6$	0,346022664877	0,346022664877	0,0234374100578
$8 \to 7$	0,454613088142	0,454613088142	0,0529553143232
$9 \to 8$	0,341968227624	0,376375190389	0,0355714693817

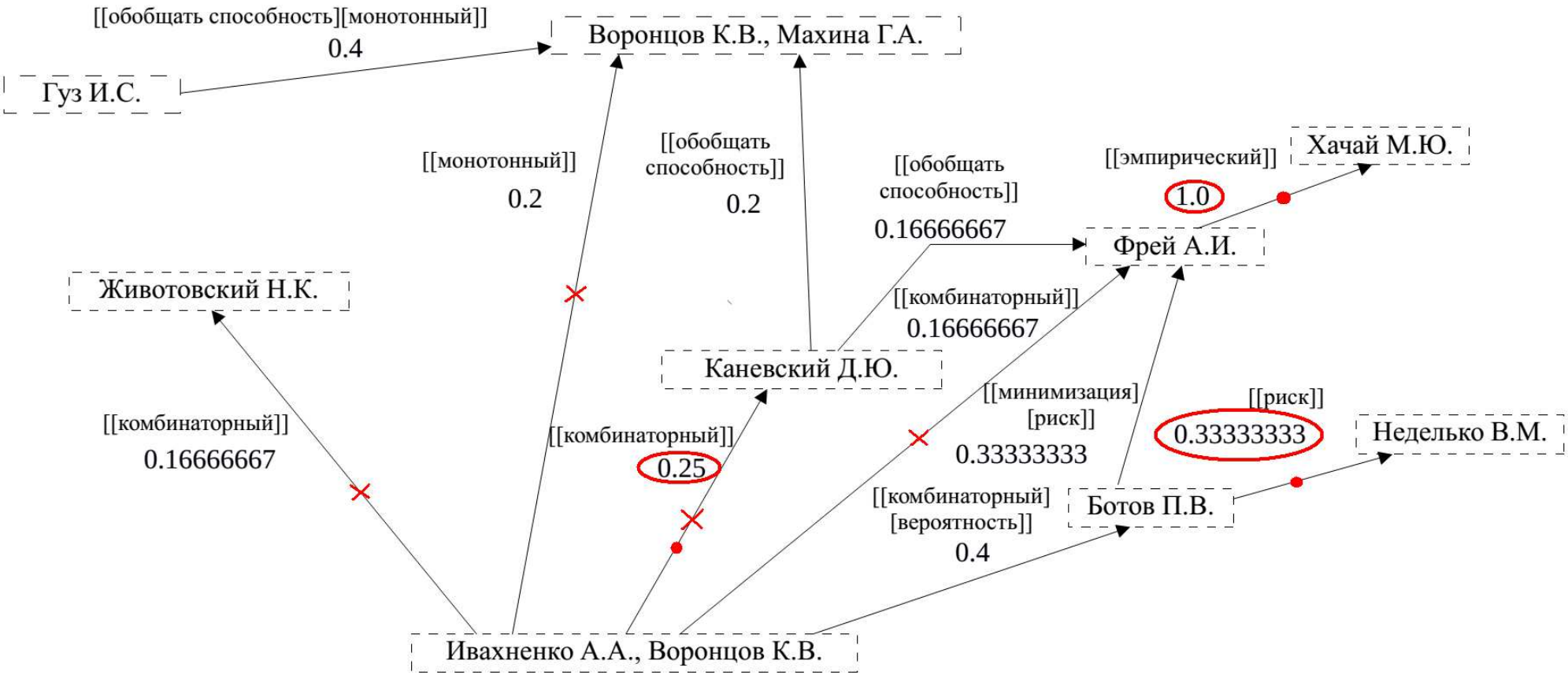

Рис. 2. Иерархизация документов с учётом ключевых сочетаний слов [4, 5].

[4] При $|\mathbf{Kw}(\mathbf{Ts}_i)| = 0$ дуги выделяются значком «●», а на них указаны значения оценки (9).

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

Выводы

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.