Моделирование и анализ данных
2023. Том 13. № 4. С. 45–58
DOI: https://doi.org/10.17759/mda.2023130403
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2023. Vol. 13, no. 4, pp. 45–58
DOI: https://doi.org/10.17759/mda.2023130403
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 004.822

Установление сходства текстовых документов

Хорошилов А.А.*
Московский авиационный институт
(национальный исследовательский университет) (МАИ)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-4885-3232
e-mail: khoroshilov@mail.ru

Кан А.В.**
ФГБУ «НИЦ «Институт имени Н.Е. Жуковского»
г. Москва, Российская Федерация
e-mail: kanav@nrczh.ru

Евдокимова Е.А.***
Федеральный исследовательский центр «Информатика и управление»
Российской Академии Наук (ИПИ РАН)
г. Москва, Российская Федерация
e-mail: evdokimovaekan@mail.ru

Пицхелаури С.Г.****
Московский авиационный институт
(национальный исследовательский университет) (МАИ)
г. Москва, Российская Федерация
e-mail: sofyauptuns@gmail.com

В настоящей статье рассматривается метод оценки сходства текстов, который основан на анализе сравнения предложений из различных текстов. Преимущества метода состоят в том, что учитывается покрытие предложения-эталона предложением из сравниваемого текста, общая оценка информационной значимости слов предложения-эталона в предложении сравниваемого текста, сходство синтаксических структур предложений, совпадение семантических значений и связей. Применение этого метода проиллюстрировано на примере решения задачи нахождения сходства двух текстов.

Ключевые слова: сходство текстов, сравнение текстов, словоупотребления, естественный язык.

Для цитаты:
Хорошилов А.А., Кан А.В., Евдокимова Е.А., Пицхелаури С.Г. Установление сходства текстовых документов // Моделирование и анализ данных. 2023. Том 13. № 4. С. 45–58. DOI: https://doi.org/10.17759/mda.2023130403

***Хорошилов Александр Алексеевич,** доктор технических наук, ведущий научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российская академия наук (ИПИ РАН), профессор кафедры Московского авиационного института (МАИ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-4885-3232, e-mail: khoroshilov@mail.ru

****Кан Анна Владимировна,** кандидат технических наук, доцент МАИ, начальник аналитического отдела ФГБУ «НИЦ «Институт имени Н.Е. Жуковского», г. Москва, Российская Федерация, e-mail: kanav@nrczh.ru

*****Евдокимова Екатерина Андреевна,** математик 1 категории, Федеральный исследовательский центр «Информатика и управление» Российская Академия Наук, г. Москва, Российская Федерация, e-mail: evdokimovaekan@mail.ru

******Пицхелаури Софья Георгиевна,** студент магистратуры, институт «Информационные технологии и прикладная математика» Московсковский авиационный институт (национального исследовательского университета) (МАИ), г. Москва, Российская Федерация, e-mail: sofyauptuns@gmail.com

1. ВВЕДЕНИЕ

Анализ сходства текстов на сегодняшний день является актуальной и значимой проблемой. С развитием цифровых технологий объем текстовой информации становится все больше и продолжает стремительно расти. В связи с этим увеличивается необходимость в инструментах, способных эффективно сравнивать и анализировать тексты в различных контекстах.

Актуальность данной проблемы обусловлена широким спектром ее практических применений. В образовании и научных исследованиях анализ сходства текстов становится незаменимым в выявлении плагиата, обеспечивая тем самым честность и достоверность научных работ. В области информационной безопасности этот вид анализа помогает в борьбе со спамом, фейковыми новостями и проверке подлинности документов. Кроме того, важную роль играет его применение в машинном обучении, анализе социальных медиа и других областях обработки естественного языка (ЕЯ).

2. ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ

Методы и способы сравнения текстов могут быть разными. Например, самый простой способ – подсчитать количество общих слов в обоих текстах. В этом методе каждый текст рассматривается как «мешок слов» (bag-of-words), порядок следования слов в предложении игнорируется, учитывается только факт их наличия [1]. Так как этот метод не учитывает порядок слов и не способен уловить семантическую связь между ними, он может привести к сильному завышению показателей сходства. Таким образом, при использовании метода подсчета вхождений слов важно также учитывать его ограничения в контексте конкретной задачи.

Метод TF-IDF (Term Frequency-Inverse Document Frequency) [4; 10] используется для оценки важности слова в документе относительно коллекции документов. Он учитывает, насколько слово часто встречается в конкретном документе (чем больше

встречается, тем оно важнее) и уменьшает значимость слова, если оно встречается часто во всех документах коллекции. Этот метод помогает выделять ключевые слова и термины в документе, позволяя лучше понять его содержание и семантику. Минусом метода TF-IDF является недостаточный учет специфики контекста предложений, то есть метод не учитывает контекстуальные зависимости между словами. Он также не улавливает семантическую связь между словами, поскольку основывается на статистике встречаемости слов. И, наконец, метод TF-IDF имеет недостаток в отношении обработки синонимов и слов с похожими значениями, что может снизить точность оценки в некоторых случаях.

Метод векторного сходства [11] представляет собой способ оценки схожести между двумя текстами путем представления каждого текста в виде вектора в многомерном пространстве. Этот метод использует косинусное сходство для определения степени схожести между векторами текстов: чем ближе векторы в многомерном пространстве, тем выше их схожесть. Векторное сходство учитывает семантическую структуру текста, позволяя сравнивать не только конкретные слова, но и их контекст и значение в предложении, что делает его более гибким и точным способом оценки сходства текстов. Недостатки векторного сходства включают ограничения в работе с разреженными данными и большими текстовыми корпусами, поскольку он требует хранения и работы с векторами большой размерности. Также эта методика не всегда улавливает смысловую связь между словами, так как не всегда способна различать семантически близкие, но формально отличающиеся фразы. И, в конечном итоге, векторное сходство может столкнуться с проблемой переобучения при работе с небольшими объемами данных или при неоптимальном подборе параметров модели.

Сравнение текстов с использованием нейронных сетей включает преобразование каждого текста в числовое векторное представление, которое затем подается на вход нейронной сети для вычисления степени их сходства. Эмбеддинги слов или другие методы преобразования текста в числовые векторы позволяют учесть семантическую близость слов и контекст текста. Нейронная сеть может использовать различные архитектуры такие как: Siamese нейронные сети или архитектуры с использованием сверточных и рекуррентных слоев [6]. Обученная нейронная сеть затем может использоваться для сравнения новых текстов – предсказывать их сходство или различие на основе прошлых вычислений. При правильной настройке и обучении нейронные сети способны улавливать сложные зависимости между текстами, учитывая семантическую и контекстуальную информацию, что делает их мощным инструментом для сравнения текстовых данных. К недостаткам нейронных сетей можно отнести то, что они требуют большого объема данных для обучения, собирать которые может быть сложно и затратно. Также использование нейронных сетей для сравнения текстов требует значительных вычислительных ресурсов. Нейронные сети могут быть чувствительны к шуму или неправильной разметке данных.

Также следует рассмотреть модель фразеологического концептуального анализа текстов на естественном языке [7; 8]. Данная модель предназначена для выявления

фразеологических единиц и анализа их концептуальной структуры в текстах. Она использует методы лингвистического анализа и компьютерной обработки текстов для идентификации и интерпретации фразеологизмов и их семантических связей. Модель фразеологического концептуального анализа текстов может быть использована для установления сходства текстовых документов путем анализа фразеологических единиц, их концептуальных связей и использования семантических моделей для определения степени сходства между текстами на основе общих фразеологических концептов. Таким образом, фразеологический анализ может быть важным инструментом для улучшения процессов сравнения и классификации текстовых документов на основе их семантического содержания.

В настоящей статье предлагается метод сходства текстов, основанный на сравнении предложений эталонного текста и сравниваемого с ним. Преимущество данного метода заключается в том, что он учитывает различные аспекты, такие как покрытие предложения-эталона предложением из сравниваемого текста, общую информационную значимость слов предложения-эталона в предложении сравниваемого текста, сходство синтаксических структур предложений, совпадение семантических значений и связей для оценки их сходства.

3. ПОСТАНОВКА ЗАДАЧИ

Будем использовать следующие определения.

- Лемма – начальная, словарная форма слова. В русском языке для существительных и прилагательных это форма именительного падежа единственного числа, для глаголов и глагольных форм – форма инфинитива.
- Лексема – совокупность всех значений и грамматических форм слова. Например, словарь, словарем, словарю – это формы одной и той же лексемы, по соглашению пишущейся как «словарь».
- Словоупотребление – элемент множества контекстов, в которых может использоваться рассматриваемое слово.
- Словоформа – форма слова, которая формируется в конкретном контекстном окружении.
- Синтаксема – минимальная синтаксическая единица. Синтаксемам приписываются семантические значения, а сами синтаксемы связываются с другими синтаксемами семантическими отношениями [5].

Пусть имеется два текста c и a. Приведем представление текстовой информации. Пусть D – некоторое универсальное множество лемм, B – множество словоформ всех лексем ЕЯ, $A=\{a\}$ – множество текстов (где a – произвольный текст). Текст содержит конечное множество словоупотреблений $R^a=\{r_i\}$ и конечное множество меток $F=\{f_i\}$. SR – множество видов синтаксических связей. Определим разбиение множества словоупотреблений на предложения – множество $S^a=\{s_i\}$, $S^a \subset 2^{R^a}$, где 2^{R^a} – булеан множества R^a. Пусть $v\left(r^a\right)$ – числовая функция, определяющая вес словоупотребления в тексте. Roles – конечное множество категориально-

семантических значений синтаксем. δ^a – бинарное отношение на множестве словоупотреблений R^a и D, которое ставит каждому словоупотреблению в соответствие его нормальную форму: $\forall r \in R^a \ \exists d \in D : <r, d> \in \delta^a$. Допускается, что одно словоупотребление можно нормализовать разными способами и, соответственно, получать разные нормальные формы для одного и того же словоупотребления. ψ^a – бинарное отношение на множестве R^a и множестве B. Каждое словоупотребление имеет единственную форму $\left(\forall r \in R^a \ \exists g \in B : <r, g> \in \psi^a\right) \& \left(\exists g' \in B : <r, g'> \in \psi^a\right) \to g = g'$, т.е. отношение ψ^a функционально по определению. θ^a – бинарное отношение, соотносящее словоупотреблению определенную метку (например, гипертекстовой разметки). Каждому словоупотреблению соответствует единственная метка: $\left(\forall r \in R^a \ \exists f \in F : <r, f> \in \theta^a\right) \& \left(\exists f' \in F : <r, f'> \in \theta^a\right) \to f = f'$, т.е. отношение θ^a функционально по определению. $\Sigma^a \subseteq R^a \times R^a$ – бинарное отношение, которое определяет всевозможные синтаксические связи между словоупотреблениями, согласно работе [2]. Так как в этой работе синтаксические структуры рассматриваются в виде деревьев, введенное определение корректно. Будем считать, что в паре r_i, r_j первый элемент – $r_i \in R^a$ – соответствует главному словоупотреблению (ГС), а второй элемент – $r_j \in R^a$ – зависимому (подчиненному) словоупотреблению. $\Omega^a \subseteq R^a \times R^a$ – бинарное отношение, представляющее семантически связанные словоупотребления в тексте. $SemRoles^a$ – бинарное отношение, которое ставит в соответствие словоупотреблениям текста семантические значения синтаксем. Таким образом, каждое словоупотребление может иметь 0 и более семантических значений в тексте.

Требуется разработать метод оценки сходства текстов. В результате работы метода требуется получить числовое значение от нуля до единицы, которое будет отражать сходство текстов.

4. МЕТОД ОЦЕНКИ СХОДСТВА ТЕКСТОВ

Пусть имеется текст-эталон $\varepsilon \in A$ и сравниваемый с ним текст $\tau \in A$. Чтобы получить оценку сходства текстов, будем сравнивать их по множествам предложений S^ε ($s^\varepsilon \in S^\varepsilon$) и S^τ ($s^\tau \in S^\tau$). Если тексты имеют длину менее 3 предложений, то не имеет смысла сравнивать их по предложениям, можно сравнить их целиком.

Для сопоставления предложений будем использовать множество $N\left(s^\varepsilon, s^\tau\right) = \{<r^\varepsilon, r^\tau> \in R^\varepsilon \times R^\tau \mid r^\varepsilon \in s^\varepsilon, \exists r^\tau \in s^\tau, \exists d \in D : <r^\varepsilon, d> \in \delta^\varepsilon \ \& \ <r^\tau, d> \in \delta^\tau\}$ пар словоупотреблений, которые будем называть соответственными.

Для оценки сходства предложений s^ε и s^τ будем использовать нижеописанные критерии.

1. Для расчета покрытия предложения-эталона предложением сопоставляемого текста введем формулу:

$$I_1\left(s^\varepsilon, s^\tau\right) = \sum_{\left\langle r^\varepsilon, r^\tau \right\rangle \in N\left(s^\varepsilon, s^\tau\right)} v\left(r^\varepsilon\right).$$

Как функция определения весов $v\left(r^{\varepsilon}\right)$ может применяться TF-IDF [4; 10] или характеристика тематической значимости [11].

Введем ограничение:

$$\sum_{r^{\varepsilon}\in R^{\varepsilon}} v\left(r^{\varepsilon}\right)=1. \quad (1)$$

2. Для определения общей оценки информационной значимости слов предложения-эталона в предложении сравниваемого текста используется формула:

$$I_2\left(s^{\varepsilon},s^{\tau}\right)=\sum_{\left\langle r^{\varepsilon},r^{\tau}\right\rangle\in N\left(s^{\varepsilon},s^{\tau}\right)} f\left(r^{\varepsilon},r^{\tau}\right)v\left(r^{\varepsilon}\right)v'\left(r^{\tau}\right).$$

Здесь $f\left(r^{\varepsilon},r^{\tau}\right)$ – это «штраф» за несовпадение форм словоупотреблений r^{ε},r^{τ}:

$$f(r^{\varepsilon},r^{\tau})=\begin{cases}1, \exists g\in B: <r^{\varepsilon},g>\in \psi^{\varepsilon} \text{ and } <r^{\tau},g>\in \psi^{\tau},\\ f_0, \quad \text{в противном случае,}\end{cases}$$

где $0\le f_0\le 1$ – параметр метода. Как функцию определения весов $v'\left(r^{\tau}\right)$ можно выбрать классическую оценку term frequency (TF) [4; 10]. Дополнительное условие на $v'\left(r^{\tau}\right)$, следующее из ограничения (1) можно записать следующим образом: $0\le v'\left(r^{\tau}\right)\le 1$.

Предположим, что текст синтаксически связный. Тогда имеет смысл рассматривать следующие критерии:

3. Для оценки сходства предложения-эталона и предложения сравниваемого текста на основе совпадения синтаксических структур введем формулу:

$$I_3(s^{\varepsilon},s^{\tau})=\frac{\sum_{<r^{\varepsilon},r^{\tau}>\in N_{Syn}(s^{\varepsilon},s^{\tau})} v(r^{\varepsilon})}{\sum_{r^{\varepsilon}\in\{r\in R^{\varepsilon}|\exists r'\in R^{\varepsilon}:<r,r'>\in\Sigma^{\varepsilon}\}} v(r^{\varepsilon})},$$

где $N_{Syn}(s^{\varepsilon},s^{\tau})=\{<r^{\varepsilon},r^{\tau}>\in N(s^{\varepsilon},s^{\tau})|\exists\tilde{r}^{\varepsilon}\in R^{\varepsilon},\exists\tilde{r}^{\tau}\in R^{\tau},\exists z\in SR:<\tilde{r}^{\varepsilon},\tilde{r}^{\tau}>\in N(s^{\varepsilon},s^{\tau})$ $\&<r^{\varepsilon},\tilde{r}^{\varepsilon}>\in\Sigma_z^{\varepsilon}\&<r^{\tau},\tilde{r}^{\tau}>\in\Sigma_z^{\tau}\}$ представляет собой множество пар соответственных словоупотреблений в эталонном предложении s^{ε} и сопоставляемом предложении s^{τ}, для которых совпадают (по нормальным формам лексем) главные $<r^{\varepsilon},r^{\tau}>\in N\left(s^{\varepsilon},s^{\tau}\right)$ и зависимые $<\tilde{r}^{\varepsilon},\tilde{r}^{\tau}>\in N\left(s^{\varepsilon},s^{\tau}\right)$ слова, а сами словоупотребления связаны в контексте эталонного и сопоставляемого предложения однотипными синтаксическими связями: $<r^{\varepsilon},\tilde{r}^{\varepsilon}>\in\Sigma_z^{\varepsilon}\&<r^{\tau},\tilde{r}^{\tau}>\in\Sigma_z^{\tau}$.. Знаменатель формулы – это совокупный вес словоупотреблений, которые являются главными элементами в синтаксически связанных парах словоупотреблений в эталонном тексте.

4. Выделим группу слов с семантическими значениями в предложении эталона, для которых в сопоставляемом тексте существуют соответствующие слова с теми же семантическими значениями:

$$\rho(s^{\varepsilon}, s^{\tau}) = \{< r^{\varepsilon}, a > \in SemRoles^{\varepsilon} \mid r^{\varepsilon} \in s^{\varepsilon} \& a \in Roles \ \& \ \exists r^{\tau} \in s^{\tau} : < r^{\varepsilon}, r^{\tau} > \in N(s^{\varepsilon}, s^{\tau}) \ \& < r^{\tau}, a > \in SemRoles^{\tau}\},$$

Тогда сходство предложения-эталона s^{ε} и предложения s^{τ} сопоставляемого текста на основе совпадения семантических значений определяется формулой:

$$I_4\left(s^{\varepsilon}, s^{\tau}\right) = \frac{\left|\rho\left(s^{\varepsilon}, s^{\tau}\right)\right|}{\left|WR^{\varepsilon}\right|},$$

где WR^{ε} – множество всех словоупотреблений, имеющих семантические значения.

Числитель в формуле выражает количество совпавших семантических значений у словоупотреблений в предложении эталона и в предложении сопоставляемого текста. Знаменатель формулы задает условие нормировки на 1 по всем словоупотреблениям, имеющим семантические значения в тексте эталона: $0 \le I_4\left(s^{\varepsilon}, s^{\tau}\right) \le 1$.

5. Для оценки сходства предложений на основе совпадения семантических связей введем множество $SemR_r^{\varepsilon} = \{a \in Roles | \exists r' \in R^{\tau}, \exists x \in \mathbb{R} : < r, r' > \in \Omega_x^{\tau} \ \& < r', a > \in \Omega_x^{\tau} \ \& < r', a > \in SemRoles^{\tau}\}$ значений синтаксем, которые связаны в тексте с различными семантическими связями. На основе этих связей и определяется сходство между s^{ε} и s^{τ}.

$$I_5(s^{\varepsilon}, s^{\tau}) = \frac{\sum_{<r^{\varepsilon}, r^{\tau}> \in N(s^{\varepsilon}, s^{\tau})} |SemR_{r^{\varepsilon}}^{\varepsilon} \cap SemR_{r^{\tau}}^{\tau}|}{|WR^{\varepsilon}|}.$$

Общая оценка сходства предложения эталона и сравниваемого предложения определяется суммой критериев, учитывая их взвешенное значение.

$$Sim\left(s^{\varepsilon}, s^{\tau}\right) = \sum_{n=1}^{5} \alpha_n I_n\left(s^{\varepsilon}, s^{\tau}\right), \sum_{n=1}^{5} \alpha_n = 1.$$

Из всех предложений в сравниваемом тексте выбираются наиболее подходящие к предложению-эталону с учетом максимизации оценки:

$$J\left(s^{\varepsilon}, \tau\right) = \max_{s^{\tau} \in S^{\tau}} \left\{Sim\left(s^{\varepsilon}, s^{\tau}\right)\right\}.$$

Учитывая все вышеперечисленные величины, определим общую оценку сходства текста-эталона и сравниваемого текста:

$$I\left(\varepsilon, \tau\right) = \sum_{s^{\tau} \in S^{\tau}} J\left(s^{\varepsilon}, \tau\right). \qquad (2)$$

5. ПРИМЕР

Рассмотрим два текста, приведенных в табл. 1.

Таблица 1

Текст-эталон и сравниваемый текст

Текст-эталон	Сравниваемый текст
Совет директоров Центробанка РФ 27 октября принял решение поднять ключевую ставку сразу на 200 базисных пунктов – с 13 до 15 % годовых, говорится в сообщении на сайте регулятора. Это четвертое подряд повышение уровня показателя.	Совет директоров Банка России поднял ключевую ставку сразу на 200 б.п. – до 15 % годовых, говорится в сообщении регулятора. Это четвертое подряд повышение – к ужесточению политики ЦБ перешел в июле 2023 года и один раз даже повышал ставку на внеплановом заседании.
Известия, Экономика, 27 октября 2023, 13:31	*РБК, Финансы, 27 октября 2023, 13:30*

Для наглядного примера взяты небольшие тексты, поэтому не имеет смысла сравнивать их по предложениям, сравним их целиком. Тогда метод завершится на этапе подсчета взвешенной суммы всех критериев.

В качестве функции определения весов $v\left(r^{\varepsilon}\right)$ будем применять TF-IDF, которая рассчитывается следующим образом:

$$TF\left(\varepsilon,\tau\right)=\frac{n_{\varepsilon}}{\sum_{\tau}n_{\tau}},$$

где n_{ε} – число вхождений наименования понятия t в документ; $\Sigma_{\tau}n_{\tau}$ – общее число наименований понятий в данном документе.

$$IDF\left(\varepsilon,\ D\right)=\log\frac{\left|D\right|}{\left|\{d_i\ \in\ D|\ \varepsilon\ \in\ d_i\}\right|},$$

где $|D|$– число документов в коллекции (в нашем случае 20);
$|\{d_i\ \in\ D|\,\varepsilon\ \in\ d_i\}|$ – число документов из коллекции D, в которых встречается ε (когда $n_{\varepsilon}\neq 0$).

Мера TF-IDF является произведением двух сомножителей:

$$TF-IDF\left(\varepsilon,\ \tau,D\right)=\ TF\left(\varepsilon,\ \tau\right)\times\ IDF\left(\varepsilon,\ D\right)$$

Рассчитаем покрытие предложения-эталона предложением сопоставляемого текста. В табл. 2 показан расчет TF-IDF первых 10 слов документа.

Таблица 2

Расчет TF-IDF

	встречаемость в документах	**IDF**	**количество повторений слова в тексте-эталоне**	**TF текста-эталона**	**TF-IDF**
совет	10	0,30103	1	0,03030	0,00912
директоров	10	0,30103	1	0,03030	0,00912
Центробанка	11	0,25964	1	0,03030	0,00787
РФ	6	0,52288	1	0,03030	0,01585
27	5	0,60206	1	0,03030	0,01824
октября	7	0,45593	1	0,03030	0,01382
принял	11	0,25964	1	0,03030	0,00787
решение	12	0,22185	1	0,03030	0,00672
поднять	12	0,22185	1	0,03030	0,00672
ключевую	13	0,18709	1	0,03030	0,00567
…					

Тогда величина первого критерия

$$I_1\left(s^{\varepsilon}, s^{\tau}\right) = \sum_{\left\langle r^{\varepsilon}, r^{\tau}\right\rangle \in N\left(s^{\varepsilon}, s^{\tau}\right)} v\left(r^{\varepsilon}\right) = 0,234.$$

Рассчитаем общую оценку информационной значимости слов предложения-эталона в предложении сравниваемого текста. Параметр метода f_0 положим равным 0.5. Как функцию определения весов $v'\left(r^{\tau}\right)$ выберем TF, рассчитанную по формуле (3). В табл. 3 показаны результаты расчета оценки информационной значимости первых 10 слов документа.

Таблица 3

Расчет оценки информационной значимости слов

	TF-IDF	**TF сравниваемого текста**	**оценка информационной значимости слов**
совет	0,00912	0,025	0,000114027
директоров	0,00912	0,025	0,000114027
Центробанка	0,00787	0	0
РФ	0,01585	0	0
27	0,01824	0	0
октября	0,01382	0	0
принял	0,00787	0	0
решение	0,00672	0	0
поднять	0,00672	0,025	0,000084034
ключевую	0,00567	0,025	0,000070866
…			

Получим, что значение второго критерия равняется

$$I_2\left(s^{\varepsilon}, s^{\tau}\right) = \sum_{\left\langle r^{\varepsilon}, r^{\tau}\right\rangle \in N\left(s^{\varepsilon}, s^{\tau}\right)} f\left(r^{\varepsilon}, r^{\tau}\right) v\left(r^{\varepsilon}\right) v'\left(r^{\tau}\right) = 0,002.$$

Для подсчета $I_3\left(s^{\varepsilon}, s^{\tau}\right)$ выделим синтаксические структуры текстов. Цветом покажем главные слова: желтым – совпадающие в двух текстах, зеленым – все остальные (табл. 4).

Таблица 4

Выделение синтаксических структур текстов

Совет директоров Центробанка РФ 27 _октября принял решение поднять ключевую ставку сразу на 200 базисных пунктов – с 13 до 15 % годовых, говорится в сообщении на сайте регулятора. Это четвертое подряд повышение уровня показателя.	Совет директоров Банка России поднял ключевую ставку сразу на 200 б.п. – до 15 % годовых, говорится в сообщении регулятора. Это четвертое подряд повышение – к ужесточению политики ЦБ перешел в июле 2023 года и один раз даже повышал ставку на внеплановом заседании.

Тогда значение третьего критерия будет равно

$$I_3(s^{\varepsilon}, s^{\tau}) = \frac{\sum_{<r^{\varepsilon}, r^{\tau}> \in N_{Syn}(s^{\varepsilon}, s^{\tau})} v(r^{\varepsilon})}{\sum_{r^{\varepsilon} \in \{r \in R^{\varepsilon} \mid \exists r' \in R^{\varepsilon}: <r, r'> \in \Sigma^{\varepsilon}\}} v(r^{\varepsilon})} = \frac{0,034}{0,096} = 0,354.$$

Посчитаем сходство текста-эталона и сравниваемого текста на основе совпадения семантических значений предложений:

$$I_4\left(s^{\varepsilon}, s^{\tau}\right) = \frac{\left|\rho\left(s^{\varepsilon}, s^{\tau}\right)\right|}{\left|WR^{\varepsilon}\right|} = \frac{18}{28} = 0,643.$$

Проведем оценку сходства текстов на основе совпадения семантических связей в их предложениях:

$$I_5\left(s^{\varepsilon}, s^{\tau}\right) = \frac{\sum_{\left\langle r^{\varepsilon}, r^{\tau}\right\rangle \in N\left(s^{\varepsilon}, s^{\tau}\right)} \left|SemR^{\varepsilon}_{r^{\varepsilon}} \cap SemR^{\tau}_{r^{\tau}}\right|}{\left|WR^{\varepsilon}\right|} = \frac{15}{28} = 0,536.$$

Для параметров метода α_n, где $n \in \{1,2,3,4,5\}$ выберем следующие значения $\alpha_1 = 0,2$; $\alpha_2 = 0,05$; $\alpha_3 = 0,1$; $\alpha_4 = 0,4$; $\alpha_5 = 0,25$. Тогда общая оценка сходства текстов будет равна:

$$Sim(\varepsilon, \tau) = 0,234 \cdot 0,2 + 0,002 \cdot 0,05 + 0,354 \cdot 0,1 + 0,643 \cdot 0,4 + 0,536 \cdot 0,25 = 0,474.$$

Так как сравниваемые тексты содержат не более 3 предложений, то показатель (2) не рассчитывается.

Значит, тексты сходятся с $Sim(\varepsilon, \tau) = 0,474$. То есть сравниваемый текст повторяет эталонный примерно в половину.

6. ЗАКЛЮЧЕНИЕ

В настоящей статье представлен метод оценки схожести текстов, который определяет соответствие между текстами на основе сравнения предложений. Для определения близости текстов используются различные критерии, такие как: покрытие предложения из эталонного текста предложением из сравниваемого текста, оценка информационной значимости слов, сравнение синтаксических структур, семантических значений и семантических связей. Общая оценка схожести предложений определяется с использованием взвешенной суммы этих критериев, вклад в каждый из которых вносят параметры метода.

Литература

1. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. – М.: Изд-во НИУ ВШЭ, 2017. – 269 с.
2. *Сокирко А.Ю.* Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) / Дисс канд.т.н. // [Электронный ресурс] URL: http://www.aot.ru/docs/sokirko/sokirko-candid-1.html (дата обращения 23.10.2023)
3. *Соченков И.В.* Метод сравнения текстов для решения поисково- аналитических задач // Искусственный интеллект и принятие решений. М.: ИСА РАН, 2013, No2, с.95–106.
4. *Маннинг К., Рагхаван П., Шютце Х.* Введение в информационный поиск. – Вильямс, 2011. – ISBN 978-5-8459-1623-5
5. *Осипов Г.С., Тихомиров И.А., Смирнов И.В.* «Семантический поиск в сети интернет средствами поисковой машины Exactus».// Труды одиннадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2008. – Т. 3. – М. : ЛЕНАНД, 2008. – С. 323–328
6. *Пекунов В.В.* Извлечение информации из нейронных сетей прямого распространения в виде простых алгебраических моделей // Информационные технологии. 17. Т. 23. № 1. С. 76
7. *Хорошилов Ал-др А., Кан А.В., Ковернинский И.В., Ревина В.Д., Хорошилов А.А.* Автоматическое извлечение фактогрфической информации из научно-технических текстов авиационной отрасли // сб. «Информационные и телекоммуникационные технологии», № 43, 2019, стр. 71–78.
8. *Хорошилов Ал-др А., Мусабаев Р.Р., Козловская Я.Д., Никитин Ю.В., Хорошилов Алей А.* Автоматическое выявление и классификация информационных событий в текстах СМИ // Научно-техническая информация. Серия 2: Информационные процессы и системы. ВИНИТИ РАН. 2020. №7. С. 27–38. ISSN: 0548-0027. DOI: 10.36535/0548-0027-2020-07-4.
9. *Мбайкоджи Э., Драль А.А., Соченков И.В.* Метод автоматической классификации коротких текстовых сообщений // Информационные технологии и вычислительные системы. М.: ИСА РАН No3, 2012. С. 93–102.

10. *Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze.* Introduction to Information Retrieval. Cambridge University Press, 2008
11. *Rafael C. Gonzalez, Richard E.* Woods. Digital Image Processing, Prentice Hall. – 2002. – 793 p.
12. *Zipf, G.K.* Selected studies of the principle of relative frequencies of language / Cambridge, Massachusetts: Harvard Unive, 1932.

Establishing Similarities between Text Documents

Alexander A. Khoroshilov*
Moscow Aviation Institute (national research university) (MAI), Moscow, Russia
ORCID: https://orcid.org/0000-0003-4885-3232
e-mail: khoroshilov@mail.ru

Anna V. Kan**
FSBI «National Research Center» Institute
named after N.E. Zhukovsky, Moscow, Russia
e-mail: kanav@nrczh.ru

Ekaterina A. Evdokimova **
Federal Research Center "Informatics and Management"
of the Russian Academy of Sciences (IPI RAS), Moscow, Russia
e-mail: evdokimovaekan@mail.ru

Sofya G. Pitskhelauri ***
Moscow Aviation Institute (national research university) (MAI), Moscow, Russian
e-mail: sofyauptuns@gmail.com

This article discusses a method for assessing the similarity of texts, which is based on the analysis of comparison of sentences from different texts. The advantages of the method are that it takes into account the coverage of the standard sentence by a sentence from the compared text, the general assessment of the informational significance of the words of the standard sentence in the sentence of the compared text, the similarity of the syntactic structures of sentences, the coincidence of semantic meanings and connections. The application of this method is illustrated by the example of solving the problem of finding the similarities between two texts.

Keywords: similarity of texts, comparison of texts, word usage, natural language.

For citation:
Khoroshilov A.A., Kan A.V., Evdokimova E.A., Pitskhelauri S.G. Establishing Similarity between Text Documents. *Modelirovanie i analiz dannykh = Modelling and Data Analysis,* 2023. Vol. 13, no. 4, pp. 45–58. DOI: https://doi.org/10.17759/mda.2023130403 (In Russ., abstr. in Engl.).

***Alexander A. Khoroshilov,** Doctor of Technical Sciences, Leading Researcher, Federal Research Center "Informatics and Management" Russian Academy of Sciences (IPI RAS), Professor, Department of the Moscow Aviation Institute (MAI), Moscow, Russia, ORCID: https://orcid.org/0000-0003-4885-3232, e-mail: khoroshilov@mail.ru

****Anna V. Kan,** Candidate of Technical Sciences, Associate Professor, Moscow Aviation Institute, Head of the Analytical Department, Federal State Budgetary Institution «National Research Center» Institute named after N.E. Zhukovsky, e-mail: kanav@nrczh.ru

***Ekaterina A. Evdokimova,** 1st Category Mathematician, Federal Research Center «Informatics and Management» Russian Academy of Sciences (IPI RAS), Moscow, Russia, e-mail: evdokimovaekan@mail.ru

****Sofya G. Pitskhelauri,** Master's Student at the Institute of Information Technologies and Applied Mathematics, Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia, e-mail: sofyauptuns@gmail.com

References

1. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannykh: ucheb. posobie [Automatic natural language text processing and data analysis: tutorial] / Bol'shakova E.I., Vorontsov K.V., Efremova N.E., Klyshinskii E.S., Lukashevich N.V., Sapin A.S. – M.: Izd-vo NIU VShE, 2017. – 269 p.
2. A.Yu. Sokirko. Semanticheskie slovari v avtomaticheskoĭ obrabotke teksta (po materialam sistemy DIALING) [Semantic dictionaries in automatic text processing (based on the materials of the DARLING system)] / Diss kand.t.n. // [Elektronnyi resurs] URL: http://www.aot.ru/docs/sokirko/sokirko-candid-1.html (date of application 23.10.2023)
3. I.V. Sochenkov. Metod sravneniya tekstov dlya resheniya poiskovo- analiticheskikh zadach // Iskusstvennyi intellekt i prinyatie reshenиĭ [Text comparison method for solving search and analytical problems // Artificial intelligence and decision making]. M.: ISA RAN, 2013, No2, p. 95–106.
4. Manning K., Ragkhavan P., Shyuttse Kh. Vvedenie v informatsionnyi poisk. – Vil'yams [Introduction to information retrieval. – Williams], 2011. – ISBN 978-5-8459-1623-5
5. Osipov G.S., Tikhomirov I.A., Smirnov I.V. «Semanticheskiĭ poisk v seti internet sredstvami poiskovoĭ mashiny Exactus» [Semantic search on the Internet using the Exactus search engine]. Trudy odinnadtsatoĭ natsional'noĭ konferentsii po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2008. – T. 3. – M.: LENAND, 2008. – S. 323–328
6. Pekunov V.V. Izvlechenie informatsii iz neironnykh setei pryamogo rasprostraneniya v vide prostykh algebraicheskikh modelei // Informatsionnye tekhnologii. 17. T. 23. № 1. S. 76
7. Khoroshilov Al-dr A., Kan A.V. Koverninskii I.V., Revina V.D., Khoroshilov A.A. Avtomaticheskoe izvlechenie faktogrficheskoi informatsii iz nauchno-tekhnicheskikh tekstov aviatsionnoi otrasli // sb. «Informatsionnye i telekommunikatsionnye tekhnologii» [Automatic extraction of factual information from scientific and technical texts of the aviation industry // Sat. "Information and telecommunication technologies"], № 43, 2019, str. 71–78.
8. Khoroshilov Al-dr A., Musabaev R.R., Kozlovskaya Ya.D., Nikitin Yu.V., Khoroshilov Al-ei A. Avtomaticheskoe vyyavlenie i klassifikatsiya informatsionnykh sobytii v tekstakh SMI // Nauchno-tekhnicheskaya informatsiya [Automatic detection and classification of information events in media texts // Scientific and technical information]. Seriya 2: Informatsionnye protsessy i sistemy. VINITI RAN. 2020. №7. S. 27–38. ISSN: 0548-0027. DOI: 10.36535/0548-0027-2020-07-4.
9. Mbaĭkodzhi E., Dral' A.A., Sochenkov I.V. Metod avtomaticheskoi klassifikatsii korotkikh tekstovykh soobshcheniĭ // Informatsionnye tekhnologii i vychislitel'nye sistemy [Method for automatic classification of short text messages // Information technologies and computing systems.]. M.: ISA RAN No3, 2012. S. 93-102.
10. Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze. Introduction to Information Retrieval. Cambridge University Press, 2008
11. Rafael C. Gonzalez, Richard E. Woods. Digital Image Processing, Prentice Hall. – 2002. – 793 p.
12. Zipf, G.K. Selected studies of the principle of relative frequencies of language / Cambridge, Massachusetts: Harvard Unive, 1932.

Получена 20.11.2023 Received 20.11.2023
Принята в печать 06.12.2023 Accepted 06.12.2023