Лексико-семантическая модель кластеризации текстов на основе гибридной семантики и графового сходства

В.Н. Калинин, Т.Р. Владимирова, Н.С. Курдюков, Д.О. Жуков

Аннотация **- В работе предлагается лексико-семантическая модель кластеризации текстов, объединяющая лексическое TF-IDF представление, семантические компоненты локального и глобального уровней и гибридную меру междокументного сходства. В качестве семантических компонентов рассматриваются модель ICAN, отражающая локальные ассоциативные связи внутри документа, и модель PCAN, формирующая глобальное корпусное семантическое пространство на основе статистики совместной встречаемости слов. Для кластеризации используется итеративный алгоритм ICIC, учитывающий как близость документа к центроиду кластера, так и его сходство с ближайшими документами внутри кластера. Экспериментальная проверка проведена на двух сбалансированных корпусах новостных текстов объёмом 7200 и 4800 документов. Выполнено сравнение базовых алгоритмов кластеризации, гибридных моделей TF-IDF с ICAN и TF-IDF с PCAN, а также режимов работы алгоритма ICIC. Показано, что использование гибридной семантики повышает качество кластеризации по метрикам точности, ARI и NMI, а наилучшие результаты достигаются при сочетании PCAN с ICIC. Полученные результаты подтверждают эффективность объединения лексических, семантических и структурных характеристик в единой модели кластеризации текстов.**

Ключевые слова: кластеризация текстов, лексико-семантическая модель, гибридная семантика, графовое сходство, TF-IDF, ICAN, PCAN, ICIC.

I. Введение

Кластеризация текстов остаётся одной из ключевых задач интеллектуальной обработки данных, поскольку позволяет автоматически выявлять тематически однородные группы документов без предварительной разметки. Практическая эффективность кластеризации определяется не только выбором алгоритма разбиения, но и качеством признакового представления текста, а также используемой мерой междокументного сходства.

Классические методы, основанные на TF-IDF и линейном снижении размерности, включая SVD и LSA, хорошо отражают лексическую близость документов, однако в ограниченной степени учитывают семантические связи между словами, синонимию, вариативность терминологического описания и общие контекстные характеристики текста. В результате даже использование более сложных алгоритмов кластеризации при фиксированном признаковом описании нередко приводит лишь к ограниченному приросту качества.

Одним из перспективных направлений развития методов кластеризации является построение гибридных моделей, объединяющих лексические признаки и семантические компоненты, извлекаемые из статистики совместной встречаемости слов и графовых представлений текста. Такой подход позволяет сочетать устойчивость TF-IDF к шуму и редким словам с возможностью учитывать смысловую близость терминов, проявляющуюся через контекстные связи.

Дополнительные возможности открывает представление документов в виде графа сходства, позволяющего учитывать не только глобальную близость объектов, но и локальную структуру их взаимосвязей. В отличие от чисто векторных схем, графовая постановка создаёт основу для более гибкого задания междокументного сходства и последующего итеративного уточнения кластерных назначений с учётом как центроидных характеристик, так и локального окружения документов.

Таким образом, актуальной является задача разработки модели кластеризации, объединяющей лексическое представление текста, семантические графовые характеристики и алгоритмические процедуры, ориентированные на структуру попарного сходства документов.

В данной работе предлагается лексико-семантическая модель кластеризации текстов на основе гибридной семантики и графового сходства. В модели базовое TF-IDF-представление документов дополняется семантической компонентой, формируемой в двух вариантах: как локальная графовая характеристика документа (ICAN) и как глобальная семантика корпуса (PCAN). На этой основе задаётся гибридная мера сходства, объединяющая лексическую и семантическую составляющие, а кластеризация реализуется с использованием итеративного алгоритма (ICIC),

Статья получена 1 марта 2026.

В. Н. Калинин, Институт радиоэлектроники и информатики, РТУ МИРЭА, Москва, Россия (e-mail: kalinin_v@mirea.ru).

Т. Р. Владимирова, Научно-учебный центр «Космические системы и комплексы», РТУ МИРЭА, Москва, Россия (e-mail: vladimirova_t@mirea.ru).

Н. С. Курдюков, Институт информационных технологий, РТУ МИРЭА, Москва, Россия (e-mail: nskurdyukov@gmail.com).

Д. О. Жуков, Институт радиоэлектроники и информатики, РТУ МИРЭА, Москва, Россия (e-mail: zhukov_do@mirea.ru).

основанного на обновляемых прототипах и комбинированном скоринге близости. Такое построение позволяет объединить преимущества лексических, семантических и графовых подходов в рамках единой модели кластеризации.

II. Литературный обзор

Кластеризация текстов традиционно опирается на модели векторных представлений документов и меры близости в пространстве признаков. Базовыми методами для таких моделей выступают взвешивания термов, где идея специфичности терма и её статистическая интерпретация формируют основу IDF-подходов [1], а систематизация и практические схемы взвешивания (в т.ч. TF-IDF) показали устойчивую эффективность в задачах сравнения текстов [2].

Для снижения размерности и частичного учёта скрытой структуры частоты встречаемости термов применяется латентно-семантический анализ (LSA, LSI) через SVD матрицы «терм–документ», позволяющий уменьшать влияние синонимов и шум за счёт латентных факторов [3]. В современных постановках указанные принципы остаются базовыми при построении признакового описания и последующей кластеризации коллекций документов [4].

В работах [4-6] отмечается, что ограничения лексических моделей связаны с тем, что семантическая близость текстов не сводится к простому совпадению словоформ [4]. Дистрибутивный подход предполагает извлечение семантики из статистики контекстов, а меры взаимной информации и PMI-подобные метрики позволяют оценивать силу ассоциаций «слово–контекст» и выявлять устойчивые зависимости [5]. Как показано в обзоре векторных семантических моделей, матрицы совместной встречаемости и их факторизации выступают универсальным механизмом построения смысловых пространств [6].

В работе [7] рассмотрена предиктивная модель распределённых представлений слов (word2vec, SGNS) и показана её эффективность при обучении на больших корпусах. В работе [8] предложена интерпретация SGNS как неявной факторизации матрицы сдвинутого PMI, что сближает нейросетевые и частотные («count-based») подходы. В [9] уточняется, что качество эмбеддингов в значительной степени определяется настройками построения распределений (сдвиг, сглаживание, выбор контекстов), сопоставимыми по влиянию с архитектурными решениями.

В обзорной работе по кластеризации данных [10] авторы систематизируют методы по типу взаимодействия модулей и подчёркивают, что качество группирования всё чаще определяется геометрией и устойчивостью представлений. Для сценариев коротких и потоковых текстов в работе [11] авторы рассматривают, что ключевыми проблемами становятся разреженность, динамика тем и устойчивость к шуму, что требует более информативных мер сходства и адаптивных процедур обновления кластеров.

Существенная часть современных исследований опирается на контекстные эмбеддинги как базовое представление текстового контента для последующей кластеризации. В исследовании [12] авторы показывают, что BERT-представления при корректном выборе способа извлечения признаков и нормализации часто превосходят TF-IDF в широком наборе экспериментальных настроек, а в [13] предложена модель BERTopic, которая трактует тематическое моделирование как задачу кластеризации эмбеддингов документов и формирует интерпретируемые описания тем через class-based TF-IDF, демонстрируя практическую связку эмбеддингов с кластерами.

Кроме того, в работах [12, 14-15] исследуется применение эмбеддингов больших языковых моделей (LLM) для кластеризации и обсуждаются компромиссы между качеством и вычислительной стоимостью. А авторы в работе [15] предлагают гибридные схемы, объединяющие предварительно обученные языковые модели с алгоритмами кластеризации для повышения устойчивости и снижения влияния шума.

Направление графовых и спектральных методов представляет особый интерес, поскольку качество кластеризации в этих подходах в значительной степени определяется процедурой построения и последующей обработки матрицы сходства (графа близости), включая выбор меры близости, стратегию разрежения и нормировки [16-18]. В работе [16] предлагается масштабируемая модификация spectral clustering с «self-guiding» и блочно-диагональной аппроксимацией, направленная на снижение вычислительных затрат при сохранении полезной информации между итерациями, а в [18] вводится нейросетевая модель с усилением структурных эмбеддингов, что отражает общий тренд с использованием явной топологии и структуры сходства. Таким образом, современные методы кластеризации текстов развиваются в направлениях лексических, семантических, эмбеддинговых и графовых моделей. Однако задача их объединения в единой гибридной схеме остаётся актуальной. Особый интерес представляют подходы, совместно учитывающие лексическое представление, локальную и глобальную семантику, а также графовую структуру сходства документов.

III. Методология исследования

A. Сбор и предобработка данных

В качестве исходных данных используется корпус текстовых документов, представленных в формате JSONL. Каждая запись содержит идентификатор документа, исходный текст, его нормализованную версию, а также тематическую метку, применяемую на этапе оценки качества кластеризации.

Структура записи включает следующие поля:

1) id – уникальный идентификатор документа;
2) text – исходный текст;
3) text_norm – нормализованный текст после предобработки;
4) ts – временная метка сбора;
5) url – ссылка на источник;
6) label – тематическая метка;
7) word_count – количество слов в тексте.

Корпус сформирован на основе двух датасетов, собранных из открытых источников Российских новостных порталов с использованием специализированных парсеров.

Первый датасет состоит из 7200 текстов и содержит 12 тематик: искусство, информационные технологии, литература, маркетинг, медиа, научная деятельность, образование, политика, праздники, промышленность, спорт и туризм.

Второй датасет состоит из 4800 текстов и содержит 8 тематик: экономика, ИТ и инновации, образование, политика, промышленность, путешествия и туризм, сельское хозяйство и здравоохранение.

Оба датасета сбалансированы по тематикам и длине текстов. В каждой тематической группе представлено по 200 коротких, 200 средних и 200 длинных документов. Такой способ формирования корпуса обеспечивает сопоставимость экспериментальных результатов и снижает влияние разбросов как по классам, так и по длине текстов. Длина документов задаётся по количеству слов:

1) короткие – до 300 слов;
2) средние – от 300 до 1000 слов;
3) длинные – от 1000 до 3000 слов.

Предобработка текстов включает приведение символов к нижнему регистру, токенизацию, удаление служебных слов и формирование нормализованного текстового представления, используемого на последующих этапах векторизации и кластеризации.

Тематические метки при построении признакового описания документов не используются и применяются исключительно на этапе вычисления внешних метрик качества кластеризации, что позволяет избежать методологических искажений в постановке эксперимента.

B. Лексико-семантическая модель представления текстовых документов

После этапа предобработки каждый документ корпуса преобразуется в последовательность нормализованных токенов. Далее для представления документов используется лексико-семантическая модель, объединяющая две взаимодополняющие компоненты, лексическую, основанную на вхождении термов в документ, и семантическую, отражающую связи между словами либо на уровне отдельного документа, либо на уровне корпуса. Применение гибридной модели позволяет одновременно учитывать как явную лексическую близость текстов, так и скрытые смысловые связи, не сводящиеся к совпадению словоформ (термов).

Пусть корпус документов задан множеством (1).

$$D = \{d_1, d_2, \dots, d_N\} \quad (1)$$

где каждый документ d_i после предобработки представлен последовательностью токенов (2).

$$d_i = [t_1, t_2, \dots, t_{n_i}] \quad (2)$$

Тогда для каждого документа строится гибридное представление, включающее лексическую компоненту x_i и семантическую компоненту s_i. Итоговое представление используется при вычислении меры сходства между документами и последующей кластеризации.

C. Лексическая компонента модели

В качестве базового лексического представления используется модель TF-IDF, отражающая значимость терма в конкретном документе с учётом его распространённости по корпусу. Для терма w в документе d_i значение TF-IDF определяется как (3).

$$\mathrm{tfidf}(w, d_i) = \mathrm{tf}(w, d_i) \cdot \mathrm{idf}(w) \quad (3)$$

где $\mathrm{tf}(w, d_i)$ – частота терма w в документе d_i,

$\mathrm{idf}(w)$ – обратная документная частота (4).

$$\mathrm{idf}(w) = \log \frac{N}{\mathrm{df}(w)+1} \quad (4)$$

где N – число документов в корпусе,

$\mathrm{df}(w)$ – число документов, содержащих термин w.

Для каждого документа формируется вектор (5).

$$\mathbf{x}_i \in \mathbb{R}^{|V|} \quad (5)$$

где V – общий словарь корпуса. Далее для уменьшения размерности и подавления шума применяется линейное понижение размерности с использованием SVD, после чего выполняется нормировка (6):

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2} \quad (6)$$

Лексическая компонента обеспечивает устойчивое и интерпретируемое описание документа, однако в ограниченной степени учитывает смысловые связи между терминами. Для компенсации этого ограничения в модель вводится семантическая компонента.

D. Семантическая компонента на основе ICAN [19]

В качестве одного из вариантов семантического представления можно использовать модель ICAN (Incremental Construction of an Associative Network) [19], ориентированный на построение локальной ассоциативной структуры внутри отдельного документа.

Для документа d_i строится взвешенный граф совместной встречаемости терминов, в котором вершинами выступают уникальные термы документа, а веса рёбер отражают силу ассоциативной связи между ними [19, 20].

Пусть локальный словарь документа равен (7).

$$\mathcal{W}_{d_i} = \{w : w \in d_i\} \quad (7)$$

Тогда документу сопоставляется матрица весов (8).

$$M^{(i)} \in \mathbb{R}^{|\mathcal{W}_{d_i}| \times |\mathcal{W}_{d_i}|} \quad (8)$$

где элемент $M_{xy}^{(i)}$ задаёт силу связи между словами x и y. Связи формируются в пределах скользящего окна длины W. Для текущего слова $x = t_k$ контекст задаётся множеством (9).

$$\mathcal{N}(x) = \{t_j : |j-k| \le h, j \ne k, h = \lfloor W/2 \rfloor\} \quad (9)$$

При обнаружении совместной встречаемости термов x и y вес прямой связи обновляется по правилу насыщения (10):

$$M_{xy}^{(i)} = \begin{cases} 0.5, & M_{xy}^{(i)} = 0 \\ M_{xy}^{(i)} + \frac{1 - M_{xy}^{(i)}}{2}, & M_{xy}^{(i)} > 0 \end{cases} \quad (10)$$

Таким образом, повторные совместные появления одних и тех же термов увеличивают вес связи, но с уменьшающимся приращением, что предотвращает неограниченный рост значений [19, 20].

Кроме прямых связей, в модели учитывается транзитивное усиление ассоциаций. Если слово x связано со словом y, а y связано с z, то формируется связь между

x и z по правилу (11).

$$M_{xz}^{(i)} = \begin{cases} M_{xy}^{(i)} M_{yz}^{(i)}, & M_{xz}^{(i)} = 0 \\ M_{xz}^{(i)} + \gamma\left(1 - M_{xz}^{(i)}\right) M_{xy}^{(i)} M_{yz}^{(i)}, & M_{xz}^{(i)} > 0 \end{cases} \quad (11)$$

где γ – коэффициент масштабирования транзитивного усиления.

После обработки очередного окна W веса связей подвергаются затуханию (12):

$$M_{xy}^{(i)} = \delta M_{xy}^{(i)} \quad (12)$$

где $\delta \in (0,1)$ – коэффициент затухания.

Слабые связи отсекаются по достижению нижней границы порога ε (13):

$$M_{xy}^{(i)} = \begin{cases} 0, & \delta M_{xy}^{(i)} < \varepsilon \\ \delta M_{xy}^{(i)}, & \delta M_{xy}^{(i)} \geq \varepsilon \end{cases} \quad (13)$$

После построения графа для каждой вершины $u \in \mathcal{W}_{d_i}$ вычисляется взвешенная степень (14).

$$k_u = \sum_v M_{uv}^{(i)} + \sum_{v \neq u} M_{vu}^{(i)} \quad (14)$$

На основе этих значений формируется семантический вектор документа (15):

$$\mathbf{s}_i^{\text{ICAN}} \in \mathbb{R}^{|V|} \quad (15)$$

В котором каждой координате общего словаря соответствует степень соответствующего терма. Далее выполняется нормировка (16):

$$\hat{\mathbf{s}}_i^{\text{ICAN}} = \frac{\mathbf{s}_i^{\text{ICAN}}}{\|\mathbf{s}_i^{\text{ICAN}}\|_2} \quad (16)$$

Таким образом, ICAN [19] задаёт локальную семантику документа как структуру ассоциативных связей между термами, возникающих в его контексте.

E. Семантическая компонента на основе PCAN

Несмотря на то, что модель ICAN позволяет эффективно учитывать локальные ассоциативные связи между термами в пределах отдельного документа, её семантическое представление ограничено внутридокументным контекстом.

В связи с этим перспективно рассмотреть вариант семантического представления PCAN (Pointwise Co-occurrence and Association Norms), ориентированный на извлечение глобальной семантики из статистики совместной встречаемости слов во всём корпусе.

В отличие от ICAN, где граф строится отдельно для каждого документа, PCAN формирует единое корпусное семантическое пространство.

На первом этапе строится глобальная матрица совместной встречаемости (17):

$$C \in \mathbb{R}^{|V| \times |V|} \quad (17)$$

где элемент $C[w,c]$ равен числу появлений слова c в окне вокруг слова w в документах корпуса.

На основе матрицы C вычисляется положительная матрица взаимной информации с контекстным сглаживанием.

Пусть $f(w) = \sum_c C[w,c]$ – частота слова w, а сглаженное распределение контекстов задаётся как (18):

$$f(c) = f(c)^{\alpha}, \qquad T_{\alpha} = \sum_c f(c)^{\alpha} \quad (18)$$

где $\alpha \in (0,1]$ – параметр сглаживания контекстного распределения.

Тогда элемент матрицы PPMI (Positive Pointwise Mutual Information) определяется формулой (19):

$$\text{PPMI}_{\alpha}(w,c) = \max\left(0,\ \log\frac{C[w,c] \cdot T_{\alpha}}{f(w)\, f(c)^{\alpha}} - \text{s}\right) \quad (19)$$

где s – дополнительный параметр сдвига.

Полученная матрица подвергается усечённому сингулярному разложению (20):

$$M \approx U \Sigma V^{\top} \quad (20)$$

где M – матрица PPMI, U – матрица левых сингулярных векторов, Σ – диагональная матрица сингулярных значений. Векторы слов формируются как (21):

$$W = U \Sigma^{p} \quad (21)$$

где $p \in [0,1]$ – показатель степенного взвешивания сингулярных значений.

Далее строки матрицы W нормируются.

Для перехода от словарных векторов к документным используется SIF-агрегация.

Пусть вероятность слова в корпусе равна (22):

$$p(w) = \frac{f(w)}{\sum_u f(u)} \quad (22)$$

Тогда для документа d_i его семантический вектор вычисляется как взвешенное среднее словарных векторов (23):

$$\mathbf{s}_i^{\text{PCAN}} = \frac{\sum_{w \in d_i} \frac{a}{i a + p(w)} w_w}{\sum_{w \in d_i} \frac{a}{i a + p(w)}} \quad (23)$$

где $a > 0$ – параметр сглаживания, а w_w – вектор слова w, полученный из матрицы W.

Для уменьшения влияния общекорпусных компонент после формирования матрицы документных векторов удаляются первые главные компоненты. Если $\mathbf{u}_1, \ldots, \mathbf{u}_k$ – первые главные компоненты, то откорректированный вектор документа имеет вид (24):

$$\tilde{\mathbf{s}}_i^{\text{PCAN}} = \mathbf{s}_i^{\text{PCAN}} - \sum_{j=1}^{k} \left(\mathbf{s}_i^{\text{PCAN}} \cdot \mathbf{u}_j\right) \mathbf{u}_j \quad (24)$$

После этого выполняется L2-нормировка (25):

$$\hat{\mathbf{s}}_i^{\text{PCAN}} = \frac{\tilde{\mathbf{s}}_i^{\text{PCAN}}}{\|\tilde{\mathbf{s}}_i^{\text{PCAN}}\|_2} \quad (25)$$

Таким образом, модель PCAN описывает документ через его положение в глобальном корпусном семантическом пространстве, формируемом на основе статистики совместной встречаемости слов.

F. Гибридная мера сходства документов

Поскольку лексическая и семантическая компоненты отражают различные аспекты содержания текста, в модели используется гибридная мера сходства, объединяющая оба источника информации. Для пары документов d_i и d_j лексическая близость определяется через косинусное сходство TF-IDF-векторов (26):

$$S_{\text{lex}}(d_i, d_j) = \cos(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) = \frac{\hat{\mathbf{x}}_i \cdot \hat{\mathbf{x}}_j}{\|\hat{\mathbf{x}}_i\|_2 \|\hat{\mathbf{x}}_j\|_2} \quad (26)$$

Аналогично, семантическая близость вычисляется как косинусное сходство семантических векторов (27):

$$S_{\text{sem}}(d_i, d_j) = \cos(\hat{\mathbf{s}}_i, \hat{\mathbf{s}}_j) \quad (27)$$

В зависимости от выбранного режима вектор $\hat{\mathbf{s}}_i$ может соответствовать либо ICAN, либо PCAN представлению документа.

Итоговая гибридная мера сходства задаётся линейным смешиванием (28):

$$S_{\text{hyb}}(d_i, d_j) = (1 - \lambda)\, S_{\text{lex}}(d_i, d_j) + \lambda\, S_{\text{sem}}(d_i, d_j) \quad (28)$$

где $\lambda \in [0,1]$ – коэффициент баланса между лексической и семантической составляющими.

При $\lambda = 0$ модель сводится к чисто лексическому сравнению документов, а при $\lambda = 1$ к семантическому.

В итоге, рассматриваемая лексико-семантическая модель объединяет интерпретируемость TF-IDF представления и или локальную графовую семантику ICAN, или глобальную корпусную семантику PCAN в единой схеме вычисления междокументного сходства.

Такое построение создаёт основу для последующего применения алгоритмов кластеризации, использующих как значения сходства между документами, так и структуру графа близости.

IV. АЛГОРИТМ КЛАСТЕРИЗАЦИИ ICIC

Для кластеризации документов предложен итеративный алгоритм на базе графовой семантики (Iterative Centroid-Interaction Clustering). Основная идея данного алгоритма состоит в том, что каждый кластер описывается не только текущим центроидом в пространстве признаков, но и множеством уже отнесённых к нему документов.

Благодаря этому решение о принадлежности документа к кластеру принимается на основе сходства с центроидом кластера и на основе сходства с наиболее близкими документами внутри кластера. Такой подход позволяет совместить устойчивость центроидной схемы с большей чувствительностью к локальной структуре данных.

Пусть корпус состоит из N документов, которые требуется разбить на заранее заданное число K кластеров. Для каждого документа d_i заданы:

1) лексический вектор $\hat{\mathbf{x}}_i$;

2) семантический вектор $\hat{\mathbf{s}}_i$.

При этом $\hat{\mathbf{s}}_i$ может соответствовать либо представлению ICAN, либо представлению PCAN в зависимости от выбранного варианта семантической компоненты. Тогда текущее разбиение документов на итерации t задаётся множеством кластеров (29):

$$C_1^{(t)}, C_2^{(t)}, \dots, C_K^{(t)} \tag{29}$$

A. Инициализация кластеров

Для устойчивого старта итерационного процесса начальное разбиение формируется с использованием алгоритма K-means в лексическом пространстве TF-IDF. Это позволяет получить начальные группы документов, уже обладающие базовой лексической согласованностью. Начальная разметка определяется по формуле (30):

$$\mathbf{L}^{(0)} = \mathrm{KMeans}\left(\hat{X}, K\right) \tag{30}$$

где $\hat{X}$ – матрица лексических векторов документов, а K – число кластеров.

После инициализации для каждого кластера вычисляются центроиды в лексическом и семантическом пространствах. Для кластера c на итерации t лексический центроид определяется формулой (31):

$$\boldsymbol{\mu}_c^x = \frac{1}{\left|C_c^{(t)}\right|} \sum_{j \in C_c^{(t)}} \hat{\mathbf{x}}_j \tag{31}$$

а семантический центроид – формулой (32):

$$\boldsymbol{\mu}_c^s = \frac{1}{\left|C_c^{(t)}\right|} \sum_{j \in C_c^{(t)}} \hat{\mathbf{s}}_j \tag{32}$$

После вычисления центроиды нормируются (33):

$$\hat{\boldsymbol{\mu}}_c^x = \frac{\boldsymbol{\mu}_c^x}{\|\boldsymbol{\mu}_c^x\|_2}, \qquad \hat{\boldsymbol{\mu}}_c^s = \frac{\boldsymbol{\mu}_c^s}{\|\boldsymbol{\mu}_c^s\|_2} \tag{33}$$

B. Функция близости документа к кластеру

В алгоритме ICIC близость документа к кластеру определяется не одной, а двумя компонентами.

Первая из них отражает сходство документа с центроидом кластера.

Для документа d_i и кластера c лексическая центроидная близость равна (34):

$$\mathrm{CentSim}_x(i, c) = \hat{\mathbf{x}}_i \cdot \hat{\boldsymbol{\mu}}_c^x \tag{34}$$

Семантическая центроидная близость вычисляется по формуле (35):

$$\mathrm{CentSim}_s(i, c) = \hat{\mathbf{s}}_i \cdot \hat{\boldsymbol{\mu}}_c^s \tag{35}$$

Вторая компонента учитывает локальную структуру кластера и задаётся через среднее значение по k наибольшим сходствам документа с элементами данного кластера.

Для лексического пространства эта величина определяется по формуле (36):

$$\mathrm{MaxSim}_x^{@k}(i, c) = \frac{1}{k} \sum\nolimits_{\text{top-}k} (\hat{\mathbf{x}}_i \cdot \hat{\mathbf{x}}_j : j \in C_c^{(t)}) \tag{36}$$

А для семантического пространства данная величина определяется по формуле (37):

$$\mathrm{MaxSim}_s^{@k}(i, c) = \frac{1}{k} \sum\nolimits_{\text{top-}k} (\hat{\mathbf{s}}_i \cdot \hat{\mathbf{s}}_j : j \in C_c^{(t)}) \tag{37}$$

Таким образом, для каждого кластера вычисляются две лексические и две семантические оценки. Они объединяются в два промежуточных скора. Лексический скор задаётся формулой (38):

$$\mathrm{TF}(i, c) = \alpha\, \mathrm{MaxSim}_x^{@k}(i, c) + (1 - \alpha)\, \mathrm{CentSim}_x(i, c) \tag{38}$$

где $\alpha \in [0,1]$ регулирует вклад локального соседского сходства и центроидной близости в лексическом пространстве.

Аналогично семантический скор определяется по формуле (39):

$$\mathrm{SEM}(i, c) = \beta\, \mathrm{MaxSim}_s^{@k}(i, c) + (1 - \beta)\, \mathrm{CentSim}_s(i, c) \tag{39}$$

где $\beta \in [0,1]$ задаёт баланс между локальной и центроидной оценкой в семантическом пространстве.

Итоговая гибридная функция близости документа d_i к кластеру c имеет вид (40)

$$S(i, c) = (1 - \lambda)\, \mathrm{TF}(i, c) + \lambda\, \mathrm{SEM}(i, c) \tag{40}$$

где $\lambda \in [0,1]$ определяет соотношение между лексической и семантической составляющими.

При $\lambda = 0$ алгоритм фактически работает только в лексическом пространстве, а при $\lambda = 1$ – только в семантическом. Промежуточные значения позволяют реализовать гибридный режим кластеризации.

C. Итерационное переназначение документов

После вычисления значений $S(i, c)$ для всех документов и всех кластеров выполняется шаг переназначения. Каждый документ относится к тому кластеру, для которого значение гибридного скора максимально (41):

$$L_i^{(t+1)} = \arg \max_{c \in \{1, \dots, K\}} S(i, c) \tag{41}$$

Затем по новому разбиению пересчитываются центроиды кластеров, и процедура повторяется. Таким образом, алгоритм работает по итерационной схеме:

1) вычисление центроидов кластеров;

2) оценка близости каждого документа ко всем кластерам;

3) переназначение документов;

4) повторение до сходимости.

Для оценки изменения разбиения между соседними итерациями используется доля документов, изменивших кластер (42):

$$\Delta^{(t)} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[L_i^{(t+1)} \neq L_i^{(t)}\right] \quad (42)$$

Итерационный процесс завершается при достижении условия в формуле (43).

$$\Delta^{(t)} \leq \text{tol} \quad (43)$$

где tol – заданный порог сходимости, либо если достигнуто максимальное число итераций.

D. Выбор итогового разбиения

Так как результат кластеризации может зависеть от начальной инициализации, алгоритм ICIC допускает многократный запуск с различными начальными разбиениями.

Для каждого запуска вычисляется качество полученного решения как сумма скорингов документов относительно назначенных им кластеров (44):

$$J(\mathbf{L}) = \sum_{i=1}^{N} S(i, L_i) \quad (44)$$

Итоговым считается то разбиение, для которого значение функционала $J(\mathbf{L})$ максимально.

E. Особенности алгоритма ICIC

В отличие от классического K-means, где решение о принадлежности объекта к кластеру определяется только расстоянием до центроида, алгоритм ICIC учитывает также локальную внутрикластерную структуру через MaxSim.

Это особенно важно для текстовых данных, где тематически близкие документы могут образовывать неравномерные и частично перекрывающиеся группы. Дополнительным преимуществом является возможность гибкого комбинирования лексической и семантической информации с помощью параметра λ.

С вычислительной точки зрения наиболее затратной частью алгоритма ICIC является учёт локальной внутрикластерной структуры через компоненту MaxSim. Если n – число документов, K – число кластеров, T – число итераций, а d_{lex} и d_{sem} – размерности лексического и семантического пространств соответственно, то пересчёт центроидов на одной итерации требует $O\big(n(d_{\text{lex}} + d_{\text{sem}})\big)$, а вычисление центроидной близости всех документов ко всем кластерам – $O\big(nK(d_{\text{lex}} + d_{\text{sem}})\big)$.

При реализации компоненты MaxSim для каждого документа требуется сравнение с элементами кластеров, что в сумме по всем кластерам даёт $O\big(n^2(d_{\text{lex}} + d_{\text{sem}})\big)$ операций на итерацию. Таким образом, общая трудоёмкость алгоритма ICIC может быть оценена как $O\left(T\big(nK(d_{\text{lex}} + d_{\text{sem}}) + n^2(d_{\text{lex}} + d_{\text{sem}})\big)\right)$.

По вычислительной стоимости ICIC превосходит классический KMeans, однако этот рост затрат компенсируется возможностью учитывать локальную структуру сходства документов и повышать качество итогового разбиения.

Таким образом, алгоритм Iterative Centroid-Interaction Clustering реализует гибридную итеративную кластеризацию, в которой решение о назначении документа в кластер определяется одновременно его лексической близостью, семантическим сходством, положением относительно центроида и локальными связями с наиболее близкими документами внутри кластера.

V. МЕТРИКИ ОЦЕНКИ КАЧЕСТВА

Для оценки качества кластеризации в работе используются метрики, позволяющие сопоставить полученное разбиение документов с эталонной тематической структурой корпуса.

В качестве основных метрик применяются Adjusted Rand Index (ARI) и Normalized Mutual Information (NMI).

Метрика ARI оценивает степень согласованности между кластерным и эталонным разбиениями с поправкой на случайные совпадения.

Метрика NMI основана на взаимной информации между двумя разбиениями и показывает, насколько полученная кластерная структура соответствует исходной тематической разметке.

Также для оценки кластеризации используется точность кластеризации, отражающая долю документов, для которых кластерное назначение совпадает с эталонной меткой после оптимального сопоставления кластеров и классов. Поскольку индексы кластеров не связаны с индексами эталонных классов, перед вычислением точности кластеризации выполняется процедура их наилучшего взаимного соответствия.

VI. ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ

A. Постановка экспериментального исследования

Экспериментальное исследование включает в себя несколько последовательных этапов.

На первом этапе проводится сравнение базовых известных алгоритмов кластеризации в лексическом пространстве TF-IDF:

1) KMeans;
2) Agglomerative Clustering;
3) Spectral Clustering;
4) Affinity Propagation;
5) DBSCAN;
6) HDBSCAN.

После чего, исследуется влияние гибридизации TF-IDF представления с семантическими компонентами ICAN и PCAN. В рамках данного эксперимента варьируется коэффициент λ, а для ICAN дополнительно анализируется влияние размера скользящего окна.

Далее проводится оценка предложенного алгоритма ICIC в сочетании с ICAN и PCAN представлениями, а также проводится исследование вклада комбинированного режима по сравнению с вариантом, использующим только центроидную близость.

Во всех экспериментах число кластеров K задавалось априори и принималось равным числу тематик в соответствующем датасете. Такая постановка соответствует контролируемому сценарию оценки качества кластеризации, в котором фиксированная мощность разбиения позволяет сопоставлять влияние признакового описания документов и алгоритма кластеризации без дополнительной неопределённости,

связанной с автоматическим выбором числа кластеров.

При этом тематические метки не используются ни при построении признаковых представлений, ни при выполнении кластеризации, а служат только для внешней оценки качества полученного разбиения.

B. Сравнение методов

Таблица 1 – Сравнение базовых алгоритмов кластеризации в TF-IDF-пространстве. Датасет №1, 7200 текстов, 12 тематик

Алгоритм	Лучшие параметры	k	Accuracy	ARI	NMI
KMeans	–	12	0,60	0,40	0,63
Agglomerative	–	12	0,57	0,40	0,60
Spectral	n_neighbors=30	12	0,48	0,30	0,56
AffinityProp	damping=0,9	487	0,08	0,10	0,55
DBSCAN	eps=0,7; min_samples=5	235	0,26	0,10	0,49
HDBSCAN	min_cl_size=20; min_samples=5	28	0,38	0,10	0,50

Таблица 2 – Сравнение базовых алгоритмов кластеризации в TF-IDF-пространстве. Датасет №2, 4800 текстов, 8 тематик

Алгоритм	Лучшие параметры	k	Accuracy	ARI	NMI
KMeans	–	8	0,68	0,45	0,57
Agglomerative	–	8	0,75	0,53	0,62
Spectral	n_neighbors=30	8	0,73	0,55	0,68
AffinityProp	damping=0,9	505	0,05	0,03	0,45
DBSCAN	eps=0,7; min_samples=5	87	0,16	0,00	0,24
HDBSCAN	min_cl_size=20; min_samples=5	26	0,30	0,04	0,36

Результаты, приведённые в таблицах 1 и 2, показывают, что на базовом TF-IDF представлении наилучшие результаты демонстрируют алгоритмы KMeans, Agglomerative Clustering и Spectral Clustering.

Для датасета №1 лучшим оказался KMeans с точностью 0,60, ARI 0,40 и NMI 0,63.

Для датасета №2 наилучшие показатели продемонстрировали Agglomerative Clustering и Spectral Clustering. Точность кластеризации достигла 0,75 и 0,73, а ARI – 0,53 и 0,55 соответственно.

Методы Affinity Propagation, DBSCAN и HDBSCAN показали существенно более слабые результаты, что указывает на их ограниченную устойчивость в условиях многотемных текстовых корпусов.

Исходя из полученных результатов в таблице 3, добавление локальной семантической компоненты ICAN в составе гибридной модели с TF-IDF и ICAN не приводит к резкому росту качества по сравнению с базовым TF-IDF представлением.

Наиболее высокие значения точности кластеризации достигаются для Agglomerative Clustering при $\lambda = 0{,}3$ и скользящих окнах 5 или 11, где точность кластеризации составляет 0,74.

Наилучшие значения ARI и NMI в этой группе достигаются Spectral Clustering при $\lambda = 0{,}3$ и скользящем окне 5, ARI $= 0{,}52$, NMI $= 0{,}67$.

При увеличении вклада семантической компоненты до $\lambda = 0{,}5$ и $\lambda = 0{,}7$ точность кластеризации в большинстве случаев и алгоритмах снижается, что указывает на целесообразность умеренного вклада ICAN в гибридную меру сходства.

Таблица 3 – Сравнение методов для гибридной модели TF-IDF и ICAN. Датасет №2, 4800 текстов, 8 тематик

Алгоритм	k	Accuracy	ARI	NMI	λ	Окно ICAN
KMeans	8	0,69	0,46	0,55	0,3	5
Agglomerative	8	0,74	0,48	0,59	0,3	5
Spectral	8	0,72	0,52	0,67	0,3	5
KMeans	8	0,66	0,44	0,53	0,5	5
Agglomerative	8	0,71	0,50	0,59	0,5	5
Spectral	8	0,69	0,46	0,64	0,5	5
KMeans	8	0,65	0,42	0,52	0,7	5
Agglomerative	8	0,65	0,39	0,51	0,7	5
Spectral	8	0,64	0,39	0,60	0,7	5
KMeans	8	0,69	0,46	0,56	0,3	11
Agglomerative	8	0,74	0,47	0,59	0,3	11
Spectral	8	0,70	0,49	0,65	0,3	11
KMeans	8	0,64	0,41	0,51	0,5	11
Agglomerative	8	0,62	0,36	0,51	0,5	11
Spectral	8	0,62	0,47	0,61	0,5	11
KMeans	8	0,62	0,38	0,48	0,7	11
Agglomerative	8	0,64	0,40	0,53	0,7	11
Spectral	8	0,58	0,41	0,58	0,7	11
AffinityProp	472	0,05	0,03	0,45	0,3	5
DBSCAN	73	0,16	0,00	0,21	0,3	5
HDBSCAN	31	0,30	0,05	0,37	0,3	5
AffinityProp	412	0,06	0,03	0,44	0,5	5
DBSCAN	35	0,17	0,00	0,13	0,5	5
HDBSCAN	25	0,25	0,02	0,30	0,5	5
AffinityProp	356	0,06	0,04	0,45	0,7	5
DBSCAN	23	0,16	0,00	0,11	0,7	5
HDBSCAN	44	0,23	0,01	0,30	0,7	5
AffinityProp	472	0,05	0,03	0,45	0,3	11
DBSCAN	68	0,16	0,00	0,20	0,3	11
HDBSCAN	31	0,30	0,04	0,38	0,3	11
AffinityProp	398	0,06	0,04	0,45	0,5	11
DBSCAN	30	0,15	0,00	0,11	0,5	11
HDBSCAN	22	0,26	0,02	0,30	0,5	11
AffinityProp	358	0,06	0,04	0,44	0,7	11
DBSCAN	9	0,15	0,00	0,03	0,7	11
HDBSCAN	99	0,17	0,01	0,28	0,7	11

Таблицы 4 и 5 показывают, что гибридная модель TF-IDF с PCAN даёт более заметный положительный эффект, чем TF-IDF с ICAN.

Таблица 4 – Сравнение методов для гибридной модели TF-IDF и PCAN. Датасет №1, 7200 текстов, 12 тематик

Алгоритм	k	Accuracy	ARI	NMI	λ
KMeans	12	0,70	0,48	0,65	0,3
Agglomerative	12	0,63	0,38	0,59	0,3
Spectral	12	0,56	0,36	0,57	0,3
AffinityProp	421	0,09	0,07	0,54	0,3
DBSCAN	94	0,30	0,05	0,43	0,3
HDBSCAN	18	0,31	0,06	0,39	0,3
KMeans	12	0,69	0,49	0,62	0,5
Agglomerative	12	0,59	0,38	0,59	0,5
Spectral	12	0,57	0,40	0,59	0,5
AffinityProp	334	0,11	0,08	0,54	0,5
DBSCAN	72	0,27	0,05	0,39	0,5
HDBSCAN	30	0,33	0,07	0,46	0,5
KMeans	12	0,65	0,46	0,61	0,7
Agglomerative	12	0,58	0,38	0,55	0,7
Spectral	12	0,58	0,41	0,60	0,7
AffinityProp	289	0,12	0,09	0,54	0,7
DBSCAN	60	0,26	0,06	0,39	0,7
HDBSCAN	31	0,36	0,09	0,48	0,7

Таблица 5 – Сравнение методов для гибридной модели TF-IDF и PCAN. Датасет №2, 4800 текстов, 8 тематик

Алгоритм	k	Accuracy	ARI	NMI	λ
KMeans	8	0,83	0,65	0,67	0,3
Agglomerative	8	0,78	0,58	0,64	0,3
Spectral	8	0,83	0,66	0,69	0,3
AffinityProp	410	0,05	0,03	0,45	0,3
DBSCAN	183	0,21	0,02	0,35	0,3
HDBSCAN	22	0,29	0,04	0,34	0,3
KMeans	8	0,82	0,63	0,65	0,5
Agglomerative	8	0,73	0,57	0,61	0,5
Spectral	8	0,83	0,65	0,68	0,5
AffinityProp	330	0,06	0,04	0,45	0,5
DBSCAN	135	0,35	0,07	0,40	0,5
HDBSCAN	26	0,34	0,06	0,40	0,5
KMeans	8	0,82	0,63	0,65	0,7
Agglomerative	8	0,76	0,61	0,63	0,7
Spectral	8	0,82	0,63	0,67	0,7
AffinityProp	316	0,06	0,04	0,45	0,7
DBSCAN	64	0,38	0,08	0,41	0,7
HDBSCAN	24	0,35	0,07	0,40	0,7

Для датасета №1 наилучший результат достигается KMeans при $\lambda = 0{,}3$, при этом точность кластеризации достигает 0,70, ARI 0,48, NMI 0,65, что превышает соответствующие значения базового TF-IDF представления.

Для датасета №2 наилучшие результаты достигаются KMeans и Spectral Clustering, где точность кластеризации возрастает до 0,83, ARI до 0,65–0,66, а NMI до 0,67–0,69.

В целом применение PCAN демонстрирует улучшение точности кластеризации и оказывается наиболее эффективной семантической компонентой среди рассмотренных гибридных представлений.

Как видно из таблиц 6 и 7, наилучшие результаты достигаются при использовании предложенного алгоритма ICIC.

Для датасета №1 модель ICIC с PCAN показывает точность кластеризации 0,83, ARI 0,65, NMI 0,68, а ICIC с ICAN даёт сопоставимый результат.

Для датасета №2 обе версии алгоритма достигают точности кластеризации 0,85, а значения ARI и NMI возрастают до 0,69 и 0,70–0,71 соответственно.

В данном случае можно отметить, что использование специализированного итеративного алгоритма с комбинированным скорингом близости обеспечивает дополнительный выигрыш по точности кластеризации по сравнению со стандартными алгоритмами кластеризации.

Таблица 6 – Сравнение алгоритма ICIC для гибридных моделей TF-IDF и ICAN/PCAN. Датасет №1, 7200 текстов, 12 тематик

Алгоритм	k	Accuracy	ARI	NMI
ICIC и ICAN	12	0,82	0,63	0,67
ICIC и PCAN	12	0,83	0,65	0,68

Таблица 7 – Сравнение алгоритма ICIC для гибридных моделей TF-IDF и ICAN/PCAN. Датасет №2, 4800 текстов, 8 тематик

Алгоритм	k	Accuracy	ARI	NMI
ICIC и ICAN	8	0,85	0,69	0,71
ICIC и PCAN	8	0,85	0,69	0,70

Таблица 8 – Сравнение режимов работы алгоритма ICIC для гибридных моделей TF-IDF и ICAN/PCAN. Датасет №1, 7200 текстов, 12 тематик

Алгоритм	Режим работы	k	Accuracy	ARI	NMI
ICIC и ICAN	MaxSim и Centroid	12	0,82	0,63	0,67
ICIC и ICAN	Centroid only	12	0,71	0,50	0,57
ICIC и PCAN	MaxSim и Centroid	12	0,83	0,65	0,68
ICIC и PCAN	Centroid only	12	0,74	0,58	0,62

Таблица 9 – Сравнение режимов работы алгоритма ICIC для гибридных моделей TF-IDF + ICAN/PCAN. Датасет №2, 4800 текстов, 8 тематик

Алгоритм	Режим работы	k	Accuracy	ARI	NMI
ICIC и ICAN	MaxSim и Centroid	8	0,85	0,69	0,71
ICIC и ICAN	Centroid only	8	0,72	0,47	0,52
ICIC и PCAN	MaxSim и Centroid	8	0,85	0,69	0,70
ICIC и PCAN	Centroid only	8	0,69	0,54	0,60

Таблицы 8 и 9 показывают, что комбинированный режим MaxSim с Centroid существенно превосходит режим, использующий только центроидную близость.

Для датасета №1 применение комбинированного режима повышает точность кластеризации с 0,71 до 0,82 для ICIC с ICAN, и с 0,74 до 0,83 для ICIC с PCAN.

Для датасета №2 аналогичный эффект выражен ещё сильнее. Точность кластеризации возрастает с 0,72 до 0,85 для ICIC с ICAN и с 0,69 до 0,85 для ICIC с PCAN.

Существенный рост ARI и NMI подтверждает, что учёт локального внутрикластерного соседства через механизм MaxSim является важным фактором повышения качества и точности кластеризации.

VII. Обсуждение результатов и выводы

Проведённые экспериментальные исследования показали, что качество кластеризации текстов в значительной степени определяется не только выбором алгоритма разбиения, но и способом представления документов.

Базовые эксперименты в пространстве TF-IDF подтвердили, что классические методы, такие как KMeans, Agglomerative Clustering и Spectral Clustering, обеспечивают приемлемое качество на сбалансированных корпусах, однако их возможности ограничены лексическим уровнем описания текста.

В условиях тематически близких документов, различающихся терминологически, использование только лексических признаков оказывается недостаточным для устойчивого выявления содержательно однородных групп.

Результаты экспериментов с гибридными представлениями показали, что включение семантической компоненты действительно способствует повышению качества кластеризации, однако эффект зависит от характера самой семантической модели.

Локальная ассоциативная семантика ICAN, формируемая на уровне отдельного документа, в сочетании со стандартными алгоритмами кластеризации даёт ограниченный прирост и наиболее эффективно работает при умеренном вкладе в гибридную меру

сходства.

Напротив, глобальная семантическая компонента PCAN, построенная на корпусной статистике совместной встречаемости, продемонстрировала более устойчивое улучшение результатов на обоих датасетах. Это позволяет сделать вывод о том, что при решении задачи тематической кластеризации глобальные распределённые зависимости между словами оказываются более информативными, чем локальные внутридокументные ассоциации, если они используются в рамках стандартных кластеризационных схем.

Наиболее значимый результат связан с применением предложенного алгоритма ICIC. В отличие от классических центроидных методов, он учитывает не только близость документа к центру кластера, но и его сходство с наиболее близкими элементами внутри кластера. Эксперименты показали, что такой комбинированный подход обеспечивает устойчивый прирост качества по всем основным метрикам.

Вместе с тем данный выигрыш достигается ценой более высокой вычислительной сложности, поскольку процедура переназначения документов требует учёта локального внутрикластерного соседства. Это подтверждает, что для текстовых данных использование локальной структуры сходства может быть оправдано в тех случаях, когда приоритетом является качество кластеризации, а не минимизация вычислительных затрат.

Особенно показательно сравнение режимов работы ICIC с использованием схемы MaxSim и Centroid, которая заметно превосходит вариант, основанный только на центроидной близости.

Это подтверждает, что для текстовых данных важным фактором является учёт локальной структуры сходства, поскольку тематические группы документов часто имеют неоднородную внутреннюю организацию и не всегда хорошо описываются одним усреднённым прототипом.

Таким образом, результаты исследования подтверждают эффективность предложенной лексико-семантической модели кластеризации текстов на основе гибридной семантики и графового сходства.

Более того установлено, что объединение TF-IDF представления с семантическими компонентами ICAN и PCAN позволяет повысить качество междокументного сходства, а применение алгоритма ICIC обеспечивает дополнительный выигрыш за счёт учёта как центроидных, так и локально-структурных характеристик кластеров.

Наилучшие результаты в проведённой серии экспериментов были получены для конфигураций с использованием PCAN и алгоритма ICIC, хотя вариант ICIC с ICAN также продемонстрировал сопоставимо высокое качество.

К ограничениям проведённого исследования следует отнести зависимость результатов от параметров гибридизации, в частности от коэффициента λ, а также от настроек построения семантических представлений.

Для ICAN существенное влияние оказывают размер скользящего окна и параметры обновления графа, а для PCAN – размерность латентного пространства, параметры PPMI-преобразования и схема агрегации словарных векторов в документные представления.

Кроме того, исследование проводилось на двух корпусах новостных текстов, что ограничивает возможность прямого переноса выводов на документы других жанров и предметных областей без дополнительной проверки.

Следует учитывать, что эксперименты проводились при заранее известном числе кластеров K, равном числу тематик корпуса.

Перспективы дальнейшей работы связаны с расширением экспериментальной базы, тестированием модели на корпусах иной тематической и жанровой структуры, а также с дальнейшим развитием алгоритмической части.

Представляется перспективным более детальное исследование адаптивного выбора коэффициента смешивания между лексической и семантической компонентами, а также развитие графо-ориентированных процедур кластеризации на основе предвычисленных матриц сходства.

Полученные результаты подтверждают, что объединение лексических, семантических и структурных характеристик в рамках единой модели является продуктивным направлением развития методов кластеризации текстов.

Библиография

[1] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. 1972. Vol. 28, No. 1. P. 11–21. DOI: 10.1108/eb026526.

[2] Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management. 1988. Vol. 24, No. 5. P. 513–523. DOI: 10.1016/0306-4573(88)90021-0.

[3] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. Indexing by latent semantic analysis // Journal of the American Society for Information Science. 1990. Vol. 41, No. 6. P. 391–407.

[4] Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. 506 p.

[5] Church K. W., Hanks P. Word association norms, mutual information, and lexicography // Computational Linguistics. 1990. Vol. 16, No. 1. P. 22–29.

[6] Turney P. D., Pantel P. From frequency to meaning: vector space models of semantics // Journal of Artificial Intelligence Research. 2010. Vol. 37. P. 141–188.

[7] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013.

[8] Levy O., Goldberg Y. Neural word embedding as implicit matrix factorization // Advances in Neural Information Processing Systems. 2014. Vol. 27. P. 2177–2185.

[9] Levy O., Goldberg Y., Dagan I. Improving distributional similarity with lessons learned from word embeddings // Transactions of the Association for Computational Linguistics. 2015. Vol. 3. P. 211–225. DOI: 10.1162/tacl_a_00134.

[10] Zhou S., Xu H., Zheng Z., Chen J., Li Z., Bu J., Wu J., Wang X., Zhu W., Ester M. A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions // ACM Computing Surveys. 2024. DOI: 10.1145/3689036.

[11] Maden E., Karagoz P. Recent methods on short text stream clustering: A survey study // Wiley Interdisciplinary Reviews: Computational Statistics. 2023. Vol. 15, No. 6. DOI: 10.1002/wics.1610.

[12] Subakti A., Murfi H., Hariadi N. The performance of BERT as data representation of text clustering // Journal of Big Data. 2022. Vol. 9. Art. 15. DOI: 10.1186/s40537-022-00564-9.

[13] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794, 2022.

[14] Petukhova A., Matos-Carvalho J. P., Fachada N. Text Clustering with Large Language Model Embeddings. arXiv:2403.15112, 2024.

[15] Xu Q., Gu H., Ji S. Text clustering based on pre-trained models and autoencoders // Frontiers in Computational Neuroscience. 2024. Vol. 17. Art. 1334436. DOI: 10.3389/fncom.2023.1334436.
[16] Guo Y., Wu G. A restarted large-scale spectral clustering with self-guiding and block diagonal representation // Pattern Recognition. 2024. Vol. 156. Art. 110746. DOI: 10.1016/j.patcog.2024.110746.
[17] Sadjadi F., Torra V., Jamshidi M. Preprocessed Spectral Clustering with Higher Connectivity for Robustness in Real-World Applications // International Journal of Computational Intelligence Systems. 2024. Vol. 17. Art. 86. DOI: 10.1007/s44196-024-00455-2.
[18] Ding S., Wu B., Xu X., Guo L., Ding L. Graph clustering network with structure embedding enhanced // Pattern Recognition. 2023. Vol. 144. Art. 109833. DOI: 10.1016/j.patcog.2023.109833.
[19] Lemaire, Benoit & Denhiere, Guy. (2004). Incremental Construction of an Associative Network from a Corpus. Proceedings of the 26th Annual Meeting of the Cognitive Science Society.
[20] Ismael Ali, Austin Melton. Semantic-Based Text Document Clustering Using Cognitive Semantic Learning and Graph Theory. Proceedings – 12 th IEEE International Conference on Semantic Computing, ICSC 2018 Vol. 2018-January, 9 April 2018, Pages 243-247 DOI: 10.1109/ICSC.2018.0004

Об авторах:
Калинин Владимир Николаевич, ассистент кафедры телекоммуникаций, Института радиоэлектроники и информатики, МИРЭА - Российский технологический университет
Владимирова Таисия Руслановна, лаборант Научно-учебного центра «Космические системы и комплексы», МИРЭА - Российский технологический университет
Курдюков Никита Сергеевич, аспирант кафедры инструментального и прикладного программного обеспечения, Института информационных технологий, МИРЭА - Российский технологический университет
Жуков Дмитрий Олегович, профессор, д.т.н., профессор кафедры телекоммуникаций, Института радиоэлектроники и информатики, МИРЭА - Российский технологический университет

Lexico-Semantic Model for Text Clustering Based on Hybrid Semantics and Graph Similarity

Vladimir N. Kalinin, Taisiya R. Vladimirova, Nikita S. Kurdyukov, and Dmitry O. Zhukov

Abstract - **This paper proposes a lexico-semantic model for text clustering that combines a lexical TF-IDF representation, semantic components at the local and global levels, and a hybrid measure of inter-document similarity. The semantic components include the ICAN model, which captures local associative relations within a document, and the PCAN model, which constructs a global corpus-based semantic space from word co-occurrence statistics. Text clustering is performed using the iterative ICIC algorithm, which takes into account both the similarity of a document to the cluster centroid and its similarity to the nearest documents within the cluster. The experimental evaluation was carried out on two balanced corpora of news texts containing 7,200 and 4,800 documents. A comparison was made between baseline clustering algorithms, hybrid TF-IDF + ICAN and TF-IDF + PCAN models, as well as different operating modes of the ICIC algorithm. The results show that the use of hybrid semantics improves clustering quality in terms of Accuracy, ARI, and NMI, while the best results are achieved by combining PCAN with ICIC. The findings confirm the effectiveness of integrating lexical, semantic, and structural characteristics within a unified text clustering model.**

Keywords: **text clustering, lexico-semantic model, hybrid semantics, graph similarity, TF-IDF, ICAN, PCAN, ICIC.**

REFERENCES

[1] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. 1972. Vol. 28, No. 1. P. 11–21. DOI: 10.1108/eb026526.

[2] Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management. 1988. Vol. 24, No. 5. P. 513–523. DOI: 10.1016/0306-4573(88)90021-0.

[3] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. Indexing by latent semantic analysis // Journal of the American Society for Information Science. 1990. Vol. 41, No. 6. P. 391–407.

[4] Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. 506 p.

[5] Church K. W., Hanks P. Word association norms, mutual information, and lexicography // Computational Linguistics. 1990. Vol. 16, No. 1. P. 22–29.

[6] Turney P. D., Pantel P. From frequency to meaning: vector space models of semantics // Journal of Artificial Intelligence Research. 2010. Vol. 37. P. 141–188.

[7] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013.

[8] Levy O., Goldberg Y. Neural word embedding as implicit matrix factorization // Advances in Neural Information Processing Systems. 2014. Vol. 27. P. 2177–2185.

[9] Levy O., Goldberg Y., Dagan I. Improving distributional similarity with lessons learned from word embeddings // Transactions of the Association for Computational Linguistics. 2015. Vol. 3. P. 211–225. DOI: 10.1162/tacl_a_00134.

[10] Zhou S., Xu H., Zheng Z., Chen J., Li Z., Bu J., Wu J., Wang X., Zhu W., Ester M. A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions // ACM Computing Surveys. 2024. DOI: 10.1145/3689036.

[11] Maden E., Karagoz P. Recent methods on short text stream clustering: A survey study // Wiley Interdisciplinary Reviews: Computational Statistics. 2023. Vol. 15, No. 6. DOI: 10.1002/wics.1610.

[12] Subakti A., Murfi H., Hariadi N. The performance of BERT as data representation of text clustering // Journal of Big Data. 2022. Vol. 9. Art. 15. DOI: 10.1186/s40537-022-00564-9.

[13] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794, 2022.

[14] Petukhova A., Matos-Carvalho J. P., Fachada N. Text Clustering with Large Language Model Embeddings. arXiv:2403.15112, 2024.

[15] Xu Q., Gu H., Ji S. Text clustering based on pre-trained models and autoencoders // Frontiers in Computational Neuroscience. 2024. Vol. 17. Art. 1334436. DOI: 10.3389/fncom.2023.1334436.

[16] Guo Y., Wu G. A restarted large-scale spectral clustering with self-guiding and block diagonal representation // Pattern Recognition. 2024. Vol. 156. Art. 110746. DOI: 10.1016/j.patcog.2024.110746.

[17] Sadjadi F., Torra V., Jamshidi M. Preprocessed Spectral Clustering with Higher Connectivity for Robustness in Real-World Applications // International Journal of Computational Intelligence Systems. 2024. Vol. 17. Art. 86. DOI: 10.1007/s44196-024-00455-2.

[18] Ding S., Wu B., Xu X., Guo L., Ding L. Graph clustering network with structure embedding enhanced // Pattern Recognition. 2023. Vol. 144. Art. 109833. DOI: 10.1016/j.patcog.2023.109833.

[19] Lemaire, Benoit & Denhiere, Guy. (2004). Incremental Construction of an Associative Network from a Corpus. Proceedings of the 26th Annual Meeting of the Cognitive Science Society.

[20] Ismael Ali, Austin Melton. Semantic-Based Text Document Clustering Using Cognitive Semantic Learning and Graph Theory. Proceedings – 12 th IEEE International Conference on Semantic Computing, ICSC 2018 Vol. 2018-January, 9 April 2018, Pages 243-247 DOI: 10.1109/ICSC.2018.0004

About the Authors:

Vladimir Nikolaevich Kalinin, Assistant Lecturer at the Department of Telecommunications, Institute of Radioelectronics and Informatics, MIREA – Russian Technological University

Taisiya Ruslanovna Vladimirova, Laboratory Assistant at the Scientific and Educational Center «Space Systems and Complexes», MIREA – Russian Technological University

Nikita Sergeevich Kurdyukov, Postgraduate Student at the Department of Instrumental and Applied Software, Institute of Information Technologies, MIREA – Russian Technological University

Dmitry Olegovich Zhukov, Professor, Doctor of Technical Sciences, Professor of the Department of Telecommunications, Institute of Radioelectronics and Informatics, MIREA – Russian Technological University