УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2021 Том 21 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2021 Vol. 21 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2021-21-5-709-719
УДК 004.822

Автоматическое построение дерева диалога по неразмеченным текстовым корпусам на русском языке

Евгения Александровна Фельдина[1], Олеся Владимировна Махныткина[2]✉

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] feldinazhenja@gmail.com, https://orcid.org/0000-0001-6208-691X
[2] makhnytkina@itmo.ru✉, https://orcid.org/0000-0002-8992-9654

Аннотация

Предмет исследования. В работе предложен метод автоматического определения структуры дерева и ключевых тематик узлов в процессе построения дерева диалога по неразмеченным текстовым корпусам. Построение дерева диалога является одной из трудоемких задач при создании автоматической диалоговой системы и в большинстве случаев производится на основе ручной разметки, что занимает достаточно много времени и ресурсов. **Метод.** Разработанный метод иерархической кластеризации диалогов учитывает семантическую близость сообщений, позволяет выделять различное количество узлов на каждом уровне иерархии и ограничивать дерево диалогов в ширину и глубину. Алгоритм построения аннотаций узлов дерева диалога учитывает иерархию тем за счет построения тематических цепочек. В основе метода лежит комплексное использование методов обработки естественного языка (токенизация, лемматизация, частеречная разметка, построение векторных представлений слов и др.), анализа главных компонент для снижения размерности и методов кластерного анализа. **Основные результаты.** Эксперименты по построению структуры дерева диалога и аннотированию узлов показали большие возможности предложенного метода для построения автоматического дерева диалога. Точность распознавания на примере эталонного дерева диалога, содержащего 13 узлов на первом, 381 узел на втором и 299 узлов на третьем уровнях составила 0,8, 0,7 и 0,5 соответственно. **Практическая значимость.** Автоматическое построение деревьев диалога может быть востребовано при разработке диалоговых систем и повышения качества решения задачи генерации ответов на вопросы пользователей.

Ключевые слова

дерево диалога, диалоговая система, машинное обучение, кластерный анализ, обработка естественного языка

Ссылка для цитирования: Фельдина Е.А., Махныткина О.В. Автоматическое построение дерева диалога по неразмеченным текстовым корпусам на русском языке // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 5. С. 709–719. doi: 10.17586/2226-1494-2021-21-5-709-719

Automatic construction of the dialog tree based on unmarked text corpora in Russian

Evgeniya A. Feldina[1], Olesia V. Makhnytkina[2]✉

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] feldinazhenja@gmail.com, https://orcid.org/0000-0001-6208-691X
[2] makhnytkina@itmo.ru✉, https://orcid.org/0000-0002-8992-9654

Abstract

In this paper, we propose a method for automatically determining the structure of the tree and the key topics of nodes in the process of building a dialog tree based on unmarked text corpora. Building a dialog tree is one of the time-consuming tasks when creating an automatic dialog system and in most cases is performed on the basis of manual markup, which takes a lot of time and resources. The method of hierarchical clustering of dialogs takes into account the semantic proximity of messages, allows one to allocate a different number of nodes at each level of the hierarchy and limit the dialog tree in width and depth. The algorithm for constructing annotations of nodes of the dialog tree takes

into account the hierarchy of topics by building thematic chains. The method is based on the complex use of natural language processing methods (tokenization, lemmatization, part-of-speech tagging, word embeddings, etc.), analysis of the main components to reduce the dimension and methods of cluster analysis. Experiments on constructing the structure of the dialog tree and annotating nodes have shown the great possibilities of the proposed method for constructing an automatic dialog tree. The recognition accuracy on the example of the reference dialog tree containing 13 nodes at the first level, 381 nodes at the second level and 299 nodes at the third level was 0.8, 0.7 and 0.5, respectively. Automatic construction of dialog trees can be in demand when developing automatic dialog systems and for improving the quality of generating answers to user questions.

Keywords
dialog tree, dialog system, machine learning, cluster analysis, natural language processing

For citation: Feldina E.A., Makhnytkina O.V. Automatic construction of the dialog tree based on unmarked text corpora in Russian. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 5, pp. 709–719 (in Russian). doi: 10.17586/2226-1494-2021-21-5-709-719

Введение

В настоящее время построение автоматических диалоговых систем на естественном языке востребовано не только в традиционных областях применения (например, повышение уровня автоматизации обслуживания клиентов контактных центров, служб поддержки пользователей), но и в сравнительно новых кейсах (например, разработка голосовых ассистентов, виртуальных помощников, «умных колонок», интерактивных роботов). По типу использования выделяют диалоговые системы предметно-ориентированные и общего назначения. Примером диалоговых систем общего назначения являются голосовые помощники, которые используются для удаленного и быстрого выполнения рутинных задач, поиска информации в интернете, в качестве ведущего в настольных играх и другие. Предметно-ориентированные диалоговые системы ограничены в наборе тем, и как правило, содержат темы домена и оффтоп, использующиеся для начала и окончания диалога. Цель таких систем — решение пользовательских проблем: консультация по конкретным услугам и товарам компании, разбор ситуации/проблемы конкретного клиента. Для предметно-ориентированных диалоговых систем предъявляются высокие требования к качеству, производительности и поддержке в отличие от систем общего назначения, так как ошибки системы могут привести к оттоку клиентов и повлиять на репутацию компании. Для создания автоматической диалоговой системы необходимо определить тематики, по которым будет отвечать виртуальный консультант. Структура тем предметной области имеет иерархию и может быть представлена в виде дерева. В большинстве случаев выделение тематик производится на основе ручной разметки, что занимает много времени, ресурсов и требует формирование методики проставления меток и обучение группы, осуществляющей разметку.

Для решения описанных проблем необходимо создать инструмент, позволяющий сформировать структуру тематик на основе пользовательских диалогов. Основные требования к такому инструменту — независимость от предметной области, время выполнения задачи и легкость в использовании.

В целом построение дерева диалога можно декомпозировать на решение двух подзадач — выявление групп схожих сообщений (кластеров) и аннотирование полученных тематических кластеров.

Существует множество исследований, посвященных кластеризации текстовых сообщений (документов). Значительная часть исследований посвящена неиерархической (плоской) кластеризации [1, 2], при этом в большинстве работ используются классические методы: *k*-ближайших соседей (*k*-nearest neighbors algorithm, *k*-NN) [3, 4] и основанная на плотности пространственная кластеризация для приложений с шумами DBSCAN (Density-based spatial clustering of applications with noise) [5, 6]. Методы неиерархической кластеризации позволяют выделять группы сообщений для отдельных уровней дерева диалога, но не дают построить дерево диалога. В некоторых работах [7–9] для кластеризации текстовых документов применены методы иерархической кластеризации, которые последовательно объединяют текстовые сообщения в группы и получают дендрограмму (дерево вложенных кластеров), имеющую, как правило, значительно большее количество узлов, чем дерево диалога.

Для получения описаний узлов дерева диалога решается задача аннотирования тематических кластеров. Для аннотирования отдельных текстов используются генеративные [10] и сверточные нейронные сети [11], наборы правил [12]. В последние годы появились работы по многодокументному аннотированию (Multi-document Summarization) [13]. Однако такие методы не учитывают специфику задачи, а именно, представление текстов в виде диалогов пользователя и оператора.

Методы тематического моделирования осуществляют кластеризацию текстов с учетом их семантической информации и выделяют самые значимые слова, которые служат описанием узлов дерева диалога. Среди методов тематического моделирования, используемых для разработки диалоговых систем, можно выделить латентно-семантический анализ (LSA) [14] и скрытое распределение Дирихле (LDA) [15]. Рассмотренные методы позволяют решать отдельные подзадачи построения дерева диалога, однако требуется нахождение комплексного решения, которое бы позволяло определять структуру дерева на основе анализа диалогов пользователей и операторов и давать описание узлов дерева.

Решение задачи построения дерева диалога для корпусов диалогов на английском языке представлено в работе [16]. Авторы на основе использования автоматически построенного дерева диалога добились повышения оценки качества BLEU (BiLingual Evaluation Understudy) в задаче генерации ответов на 15 %, при

этом оценка точности автоматически построенного дерева диалога в работе не представлена.

В настоящей работе решена задача автоматического построения дерева диалога по неразмеченным текстовым корпусам на русском языке. Определен уровень точности построения дерева диалога, на который можно будет ориентироваться в дальнейших исследованиях, так на данный момент отсутствуют исследования, с которыми возможно сравнение.

Методика сбора, разметки корпуса диалогов для оценки качества автоматически построенного дерева диалогов

Датасеты для построения автоматических диалоговых систем собираются из диалогов клиентов с операторами контакт-центров в текстовых и голосовых каналах обслуживания.

Каждый диалог состоит из вопросов клиента и ответов или уточняющих вопросов оператора. Один вопрос клиента и соответствующий ответ/уточняющий вопрос оператора образуют шаг диалога. Для построения автоматической диалоговой системы необходимо разметить диалоги, т. е. отнести вопросы клиента к темам.

Структура тем предметной области имеет иерархию и представлена в виде дерева. Корень дерева – предметная область, например телеком или банк. На первом уровне дерева находятся объекты-сущности предметной области, например: интернет, тарифы, услуги. Каждый объект-сущность имеет характеристики и действия, которые формируют последующие уровни дерева. Пример структуры дерева диалога представлен на рис. 1.

Ручная разметка данных требует человеческих ресурсов, хорошо знающих предметную область, и занимает порядка двух месяцев у заказчиков среднего объема предметной области. На практике разметку осуществляет аналитик, хорошо знакомый с предметной областью. Методика разметки корпуса диалогов включает несколько этапов. На первом этапе осуществляется определение главного шага диалога. Шаг — это вопрос клиента и ответ или уточняющий вопрос оператора. Среднее количество шагов в текстовых каналах составляет два-три шага, в голосовых каналах — пять-шесть. На втором этапе шагу присваивается тема и при необходимости подтема. Тема и подтема формируют дерево диалога. Далее процесс повторяется для других диалогов, только при определении темы аналитик проверяет, есть ли в формируемом дереве тема и подтема, к которой относится текущий шаг. Заключительным этапом является ревью получившегося дерева главным аналитиком. Ревью представляет из себя выборочную проверку вопросов из всех узлов дерева диалога. Схематично методика разметки изображена на рис. 2.

Пример реплик клиентов и разметка представлены в табл. 1.

На основе методики сформирован корпус, состоящий из 178 044 пользовательских диалогов по тематике Телеком. В табл. 2 представлено количество диалогов в зависимости от количества шагов, где шаг — это вопрос клиента и ответ оператора.

Рис. 1. Структура дерева диалога

Fig. 1. Structure of the dialog tree

Таблица 1. Пример разметки диалога

Table 1. Example of a dialog markup

Идентификатор диалога	Реплика клиента	Тематика
1234567890	как связаться с оператором?	6/6.1
1234567810	отключить мобильный интернет	7/7.1/7.1-2

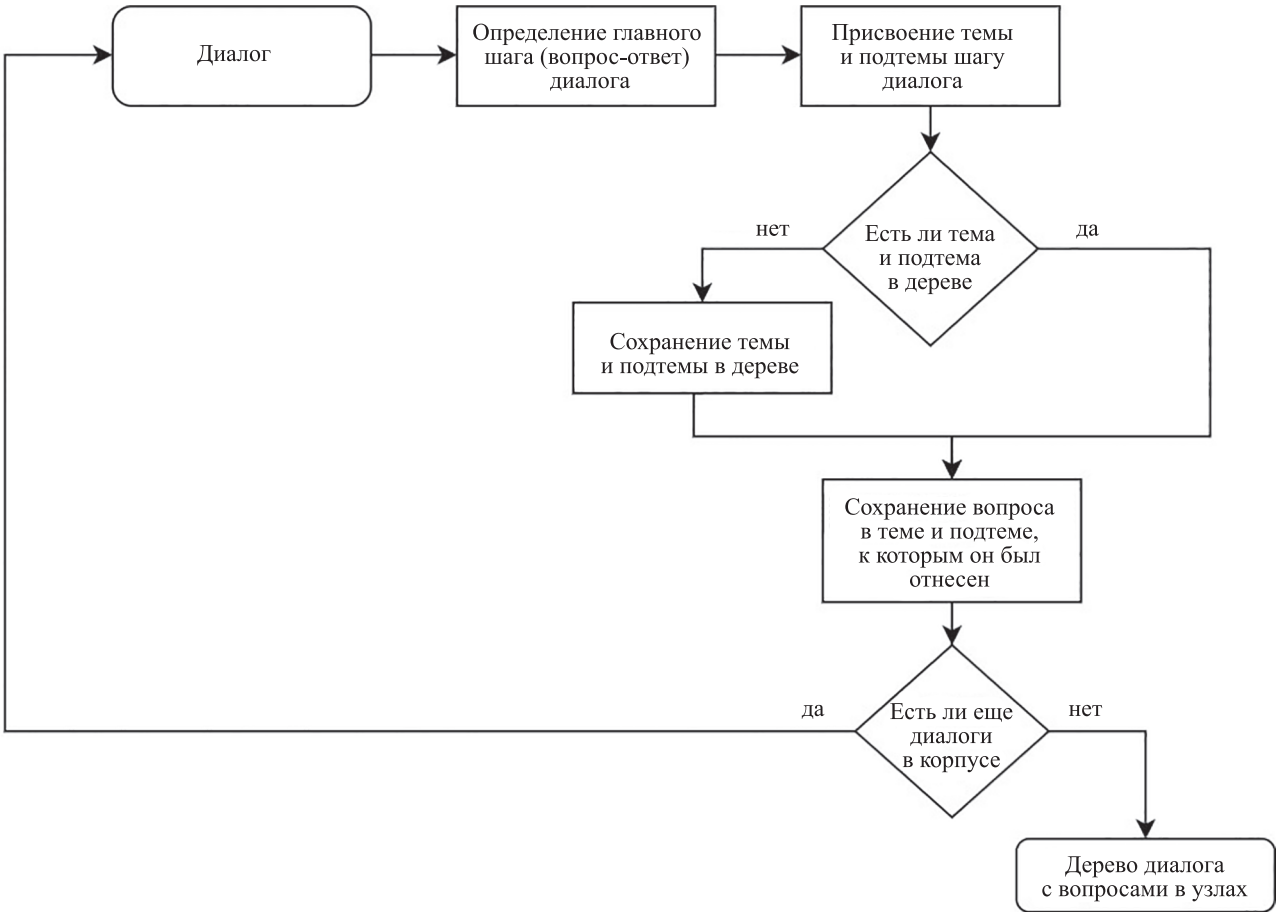

Рис. 2. Методика разметки

Fig. 2. Markup methodology

Таблица 2. Количество диалогов в зависимости от количества шагов

Table 2. Number of dialogs depending on the number of steps

Количество шагов в диалоге	Количество диалогов
1	140 493
2	14 561
3	2090
4	377
5	99
6	14
7	8

Эталонное дерево диалога содержит узлы на уровнях: 13 — на первом, 381 — на втором и 299 — на третьем.

Метод автоматического построения дерева диалогов

Разработанный метод автоматического построения дерева диалога по неразмеченным текстовым корпусам включает три этапа.

Первый этап — предварительная обработка текстовых данных. Этап включает формирование наборов данных из реплик клиентов и получение векторных представлений.

Второй этап — построение структуры дерева диалога. На данном этапе осуществляется кластеризация реплик клиентов, вычисление внутренней метрики качества кластеризации и принимается решение о необходимости деления реплик кластера на подкластеры. Также на данном этапе: определяется оптимальное количество вопросов клиентов для построения дерева диалога; проводятся эксперименты по построению структуры дерева диалога для различного количества вопросов клиентов; устанавливается на основе усредненной оценки качества кластеризации оптимальное количество вопросов пользователя, которое используется в дальнейших экспериментах.

Третий этап — аннотирование тематических кластеров. Формируются наименования узлов дерева диалога с учетом иерархии тем за счет построения тематических цепочек, учитывающих семантические отношения тем.

Общая схема построения дерева диалога представлена на рис. 3.

Первый и второй этапы автоматического построения дерева диалогов позволяют сформировать структуру дерева диалога. Схема метода автоматического построения структуры дерева диалогов представлена на рис. 4.

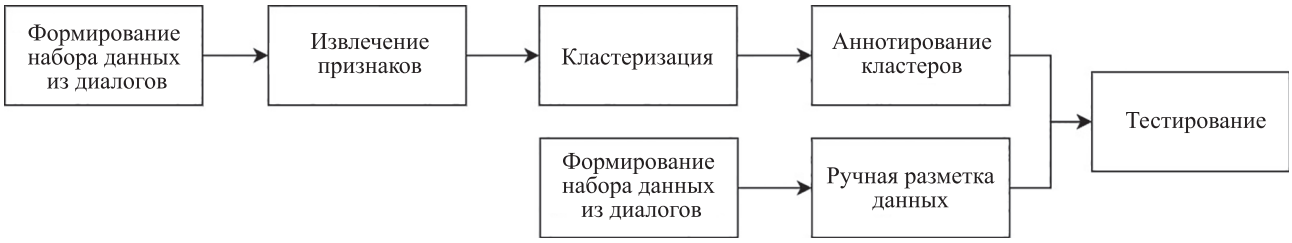

Рис. 3. Схема построения дерева диалога
Fig. 3. Diagram for building a dialog tree

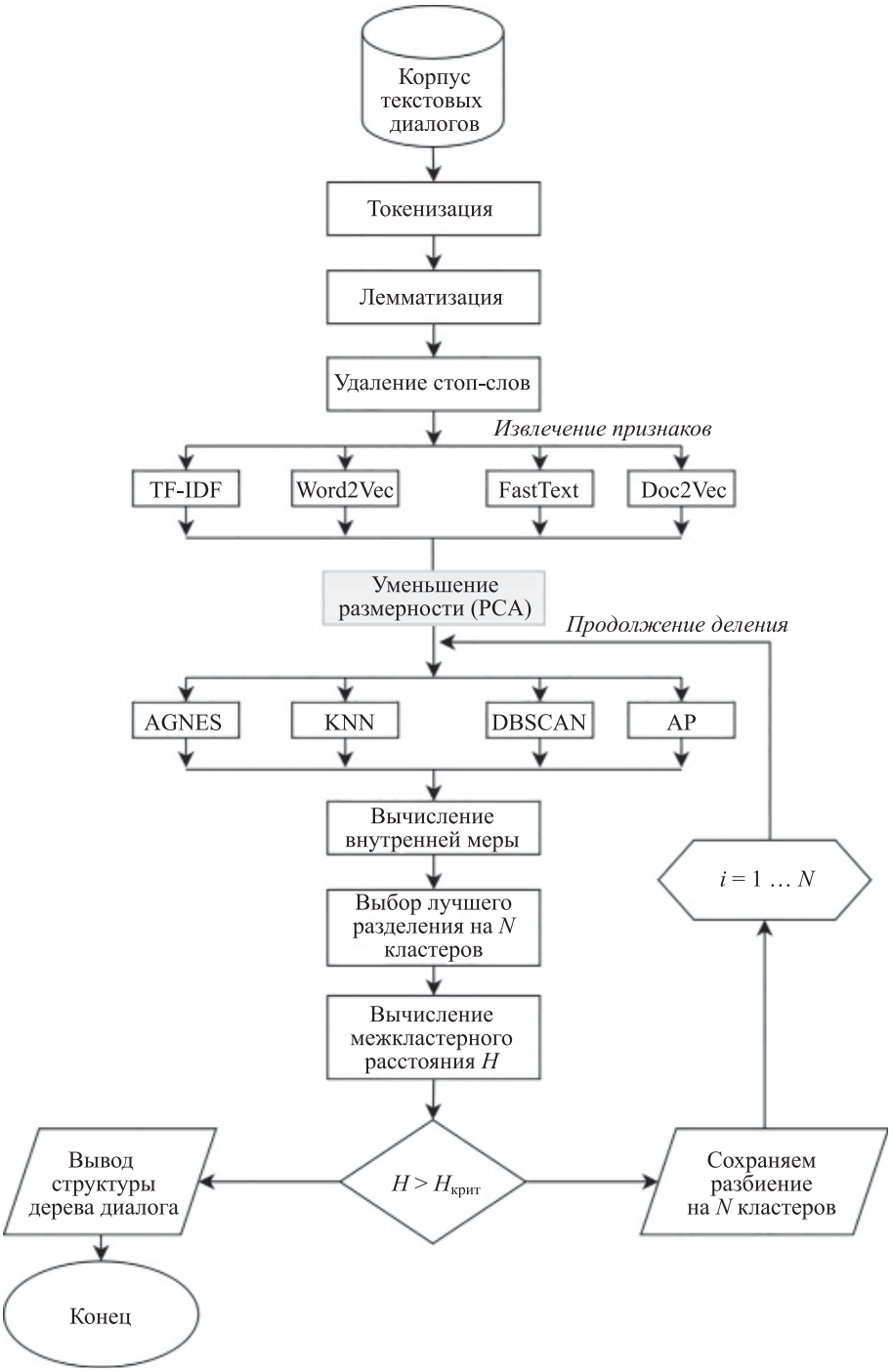

Рис. 4. Схема метода автоматического построения структуры дерева диалогов
Fig. 4. Scheme for the automatic construction of the dialog tree structure

Диалоги, как правило, состоят из большого количества шагов. Определим, информация в каких шагах представляет основную ценность для составления структуры дерева. Для поиска оптимального количества шагов на первом этапе корпус разбивается на N выборок, каждая из которых содержит от 1 до N реплик клиента соответственно. Для каждой из выборок осуществляется полный цикл экспериментов по построению дерева диалога.

Перед извлечением признаков выполняется предварительная обработка корпуса, которая состоит из следующих этапов.

Токенизация — процесс разделения текста на слова. В русском языке к разделителям относят пробел и знаки пунктуации (точка, запятая, вопросительный и восклицательный знаки, точка с запятой).

Лемматизация — процесс приведения слов в начальную форму. В русском языке, например, словарными формами являются существительные именительного падежа единственного числа; прилагательные мужского рода именительного падежа единственного числа; инфинитивные глаголы и выделение морфологических признаков (частеречная разметка).

Удаление стоп-слов — процесс исключения часто используемых слов, которые не вносят никакой дополнительной информации в текст. Слова типа «в», «на», «а» не несут никакой ценности и только добавляют шум в данные.

Рассмотрим методы извлечения признаков.

Мешок слов с мерой TF–IDF. TF–IDF (Term Frequency–Inverse Document Frequency — частота слова–обратная частота документа) нормализует частоту лексем в документе с учетом содержимого в остальном корпусе. Метод придает больше веса терминам, релевантным для конкретного экземпляра. Оценка TF–IDF вычисляется на уровне лексем, поэтому релевантность лексемы в документе измеряется масштабированной частотой появления лексемы в документе, нормализованной обратной масштабированной частотой появления лексемы во всем корпусе.

Word2Vec для построения векторных представлений слов. Word2Vec имеет два подхода к представлению векторов слов: CBOW (Continuous Bag of Words) и skip-gram. Задача метода CBOW — предсказание слова на основании близлежащих слов. У skip-gram обратная задача — предсказание набора близлежащих слов на основании одного слова. Оба метода используют в качестве алгоритмов классификации искусственные нейронные сети. Первоначально каждое слово в словаре — случайный N-мерный вектор. Во время обучения алгоритм формирует оптимальный вектор для каждого слова с помощью метода CBOW или skip-gram. Поскольку реплики для кластеризации имеют различную длину, то в качестве входных данных для кластеризации можно использовать усредненный вектор, полученный из всех векторов слов.

Doc2Vec представляет собой два метода: distributed memory (DM, распределенная память) и distributed bag of words (DBOW, распределенный мешок слов). Метод DM прогнозирует слово по известным предшествующим словам и вектору абзаца. Несмотря на то, что контекст перемещается по тексту, вектор абзаца не перемещается (отсюда название «распределенная память») и позволяет учесть порядок слов. DBOW прогнозирует случайные группы слов в абзаце только на основании вектора абзаца.

FastText, как и Word2Vec, имеет два подхода к представлению векторов слов: CBOW и skip-gram. Однако вместо подачи отдельных слов в нейронную сеть, FastText разбивает слова на несколько n-грамм (подслов).

Уменьшение размерности позволяет извлечь самую важную информацию из полученного на этапе извлечения признаков пространства и уменьшить время вычислений и потребление памяти. В качестве метода уменьшения размерности выбран анализ основных компонент (PCA).

Кластеризация текстовых документов осуществлена с использованием методов машинного обучения: агломеративная кластеризация (AGNES, Agglomerative Nesting); метод K-ближайших соседей (K-Nearest Neighbors method, KNN); пространственная кластеризация, основанная на плотности для приложений с шумами (DBSCAN); распространение похожести (Affinity Propagation, AP).

Агломеративный метод кластеризации относится к методам иерархической кластеризации. Объекты последовательно объединяются в кластер, исходя из вычисленного расстояния между ними.

Метод K-ближайших соседей разбивает множество элементов векторного пространства на заранее известное число кластеров k. На каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

DBSCAN — пространственная кластеризация, основанная на плотности. Для данного алгоритма не требуется задавать число кластеров. Основная идея: внутри каждого кластера наблюдается типичная плотность точек (объектов), которая заметно выше, чем плотность снаружи кластера, а также плотность в областях с шумом ниже плотности любого из кластеров. Иными словами, для каждой точки кластера ее соседство заданного радиуса должно содержать не менее некоторого числа точек, это число точек задается пороговым значением.

Affinity Propagation создает кластеры, отправляя сообщения между парами образцов до схождения. Затем набор данных описывается с использованием небольшого количества образцов, которые идентифицируются как наиболее репрезентативные для других образцов. Сообщения, отправляемые между парами, представляют пригодность одного образца быть образцом другого, который обновляется в ответ на значения из других пар. Это обновление происходит итеративно до сходимости, после чего выбираются окончательные образцы и, следовательно, дается окончательная кластеризация.

Для оценки качества кластеризации были рассмотрены метрики внешние (индекс Rand, индекс Adjusted Rand, однородность, полнота, V-мера) и внутренние (Силуэт (Silhouette), индекс Calinski–Harabasz, индекс Дэвиса–Болдуина (Davies–Bouldin Index)).

Для принятия решения о дальнейшем разбиении кластера на подкластеры вычислено межкластерное расстояние, и выполнено сравнение с критическим значением.

За счет вариативности некоторых блоков метод позволяет осуществлять выбор наилучшей комбинации алгоритмов на каждом этапе для выбора лучшего разделения на кластеры и может быть использован для построения дерева диалогов на данных из различных предметных областей.

На третьем этапе автоматического построения дерева диалога осуществлено составление тематических цепочек для формирования смысловых аннотаций узлов дерева диалога. Для извлечения кандидатов на ключевые тематики кластеров использован ансамбль подходов — частотный анализ, LSA, LDA. Схема алгоритма построения тематических цепочек для формирования смысловых аннотаций узлов дерева диалога представлена на рис. 5.

Результаты экспериментов

На основе корпуса диалогов сформировано три выборки. Первая выборка включает только первые репли-

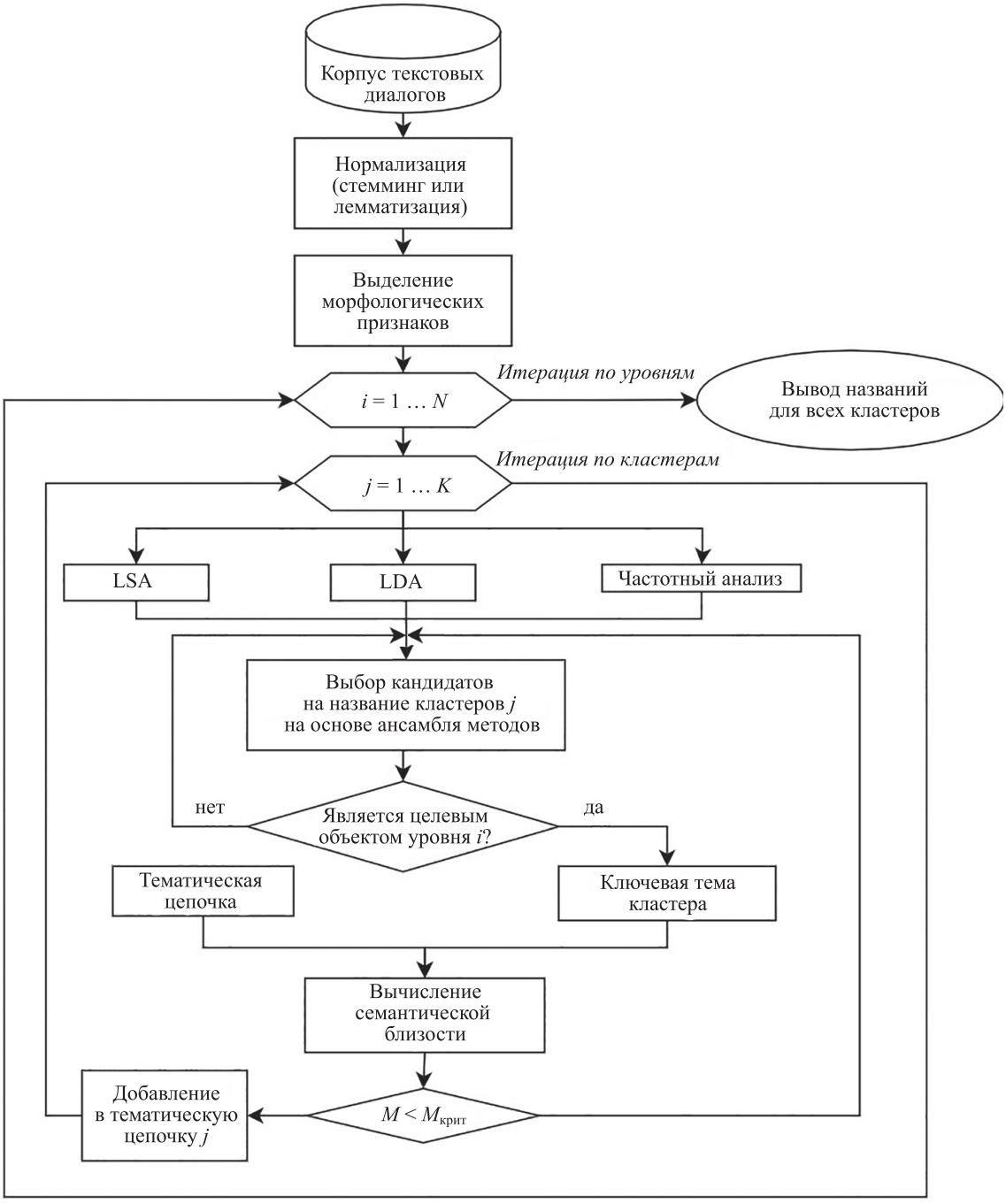

Рис. 5. Алгоритм построения тематических цепочек для формирования смысловых аннотаций узлов дерева диалога

Fig. 5. Algorithm for constructing thematic chains for the formation of semantic annotations of the nodes of the dialog tree

Таблица 3. Значения метрики Силуэт для различных методов кластеризации и количества пользовательских вопросов
Table 3. The values of the Silhouette measure for various clustering methods and the number of user questions

Количество вопросов клиента	Кластеризация			
	Агломеративная		*K*-means	
	TF-IDF	FastText	TF-IDF	FastText
1	0,612	0,450	0,580	0,465
2	0,356	0,349	0,415	0,381
3	0,348	0,304	0,319	0,339

ки пользователя. Вторая выборка состоит из склеенных первых двух реплик пользователя в порядке: первая реплика, а затем вторая. Третья выборка состоит из последовательно склеенных первых трех и более реплик пользователя. В результате экспериментов определено, что использование только первых реплик клиента показывает более высокие значения внутренних метрик. В табл. 3 представлен фрагмент таблицы с результатами экспериментов, включающий самые высокие значения метрики Силуэт [17].

Сравнение полученного дерева диалогов с эталонной структурой тем, осуществленной в результате ручной разметки, выполнено на основании модификации алгоритма неточного сравнения деревьев поиском в глубину на базе семантической близости слов. В результате экспериментов установлено, что индекс Дэвиса–Болдуина точнее других внутренних метрик согласуется с оценками внешних метрик. Лучшие результаты кластеризации получены с использованием мешка слов с мерой TF-IDF для извлечения признаков и метода *K*-средних для кластеризации. На рис. 6 и 7 показаны фрагменты сравнения построенных деревьев диалога в результате автоматической и эталонной ручной разметок.

В результате сформирована таблица сравнения тем дерева ручной разметки с деревом, построенным автоматически (табл. 4).

Сравнение дерева, полученного автоматически, с деревом, построенным в результате ручной разметки, осуществлено на основе расчета метрики accuracy для каждого уровня кластеризации: первый уровень — 0,8; второй уровень — 0,7; третий уровень — 0,5.

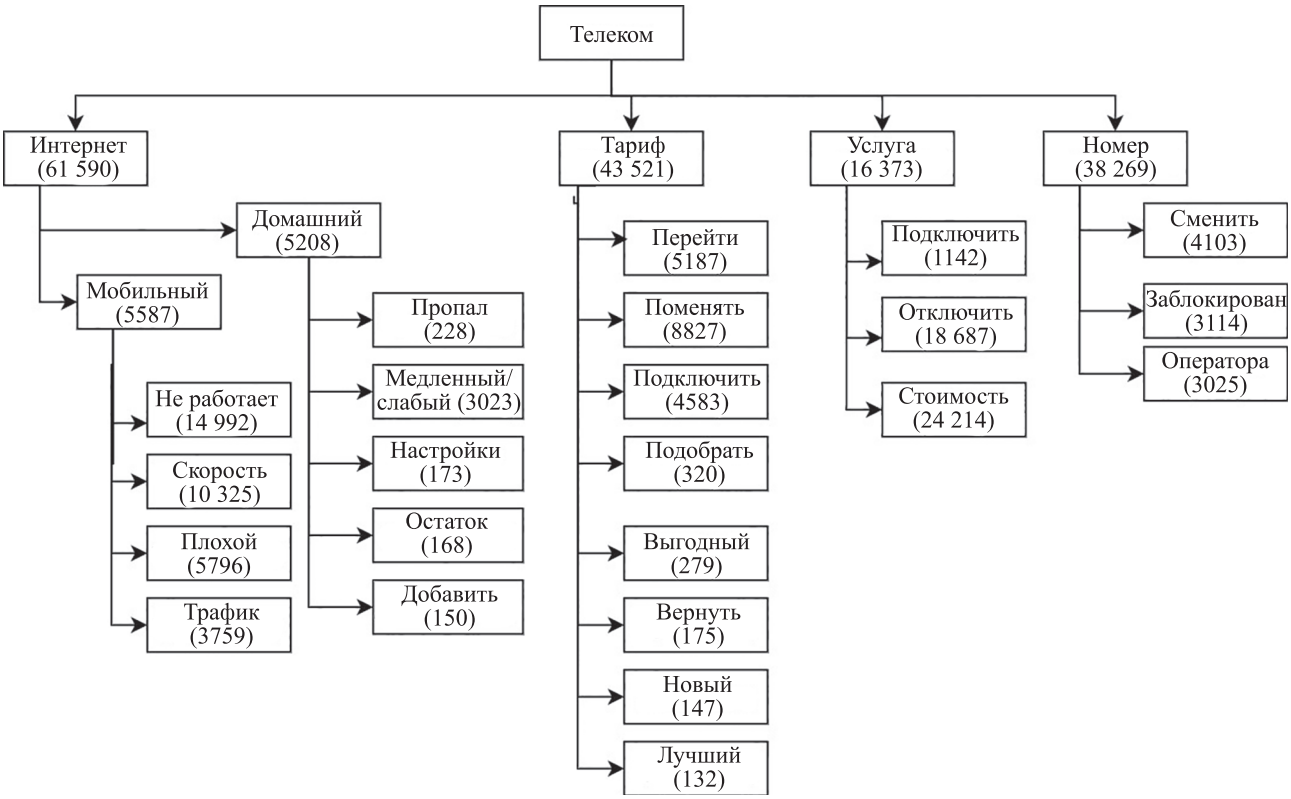

Рис. 6. Структура дерева, полученного путем применения автоматических алгоритмов
Fig. 6. Structure of the tree obtained by using automatic algorithms

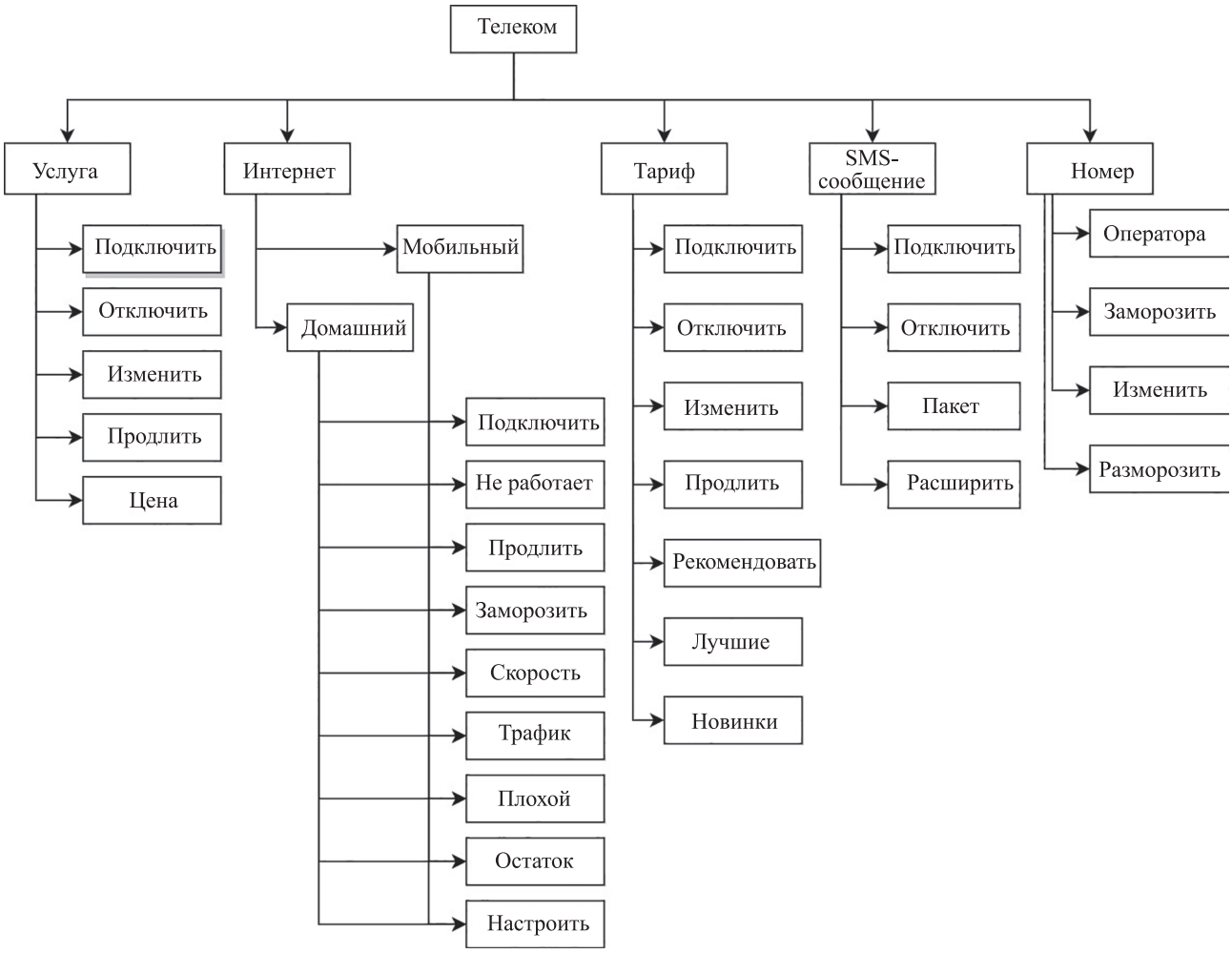

Рис. 7. Структура дерева, полученного в результате ручной разметки

Fig. 7. Structure of the tree resulting from the manual markup

Таблица 4. Сравнение тем дерева ручной разметки с деревом, построенным автоматически

Table 4. Comparing manual markup tree topics with an automatically constructed tree

Автоматическая разметка	Ручная разметка
Услуга	Услуга
Интернет	Интернет
…	…
(NULL)	SMS-сообщение
Подключить	Подключить
Стоимость	Цена
Мобильный	Мобильный
Сменить	Изменить
…	…

Заключение

В работе предложен принципиально новый подход к автоматическому построению дерева диалога по неразмеченным корпусам на русском языке на основе методов обработки естественного языка и кластерного анализа. До настоящего времени известные решения ограничены использованием структуры сайтов компаний для определения предварительной структуры дерева и ручной разметкой диалогов аналитиками, хорошо знакомых с предметной областью. В научных исследованиях встречаются решения отдельных подзадач задачи автоматического построения дерева диалога, таких как кластеризация текстовых сообщений или аннотирование текстовых документов, однако в целом решение такой задачи практически не встречается. В процессе решения задачи разработан метод иерархической кластеризации диалогов, учитывающий семантическую близость сообщений. Отличие предлагаемого метода от существующих методов заключается в том, что разбиение на кластеры не является бинарным и ограничено условиями деления в ширину и глубину. Предложен алгоритм построения аннотаций узлов дерева диалога с учетом иерархии тем за счет построения тематических цепочек, учитывающих семантические отношения тем. Для экспериментальной проверки сформирован корпус пользовательских диалогов по тематике Телеком. Корпус состоит из 178 044 пользовательских диалогов. Эталонное дерево, построенное аналитиком на основе диалогов, содержит узлы: 13 на первом уровне, 381 на

втором уровне и 299 на третьем уровне. Для данного корпуса диалогов автоматически сформировано дерево диалогов на основе предложенного метода.

Разработанный метод позволяет автоматически определять структуру дерева и ключевые тематики узлов, что ранее не осуществлялось для русскоязычных текстовых корпусов. Автоматическое построение деревьев диалога может быть востребовано при разработке автоматических диалоговых систем и повышения качество решения задачи генерации ответов на вопросы пользователей.

Литература

1. Yin J., Wang J. A text clustering algorithm using an online clustering scheme for initialization // Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1995–2004. https://doi.org/10.1145/2939672.2939841
2. Svadas T., Jha J. Document cluster mining on text documents // International Journal of Computer Science and Mobile Computing. 2015. V. 4. N 6. P. 778–782.
3. Kim H., Kim H.K., Cho S. Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling // Expert Systems with Applications. 2020. V. 150. P. 113288. https://doi.org/10.1016/j.eswa.2020.113288
4. Abasi A., Khader A., Al-Betar M., Naim S., Alyasseri Z.A., Makhadmeh S. A novel hybrid multi-verse optimizer with K-means for text documents clustering // Neural Computing and Applications. 2020. V. 32. N 23. P. 17703–17729. https://doi.org/10.1007/s00521-020-04945-0
5. Mohammed S.M., Jacksi K., Zeebaree S.R.M. Glove word embedding and DBSCAN algorithms for semantic document clustering // Proc. 3rd International Conference on Advanced Science and Engineering (ICOASE). 2020. P. 211–216. https://doi.org/10.1109/ICOASE51841.2020.9436540
6. Cretulescu R., Morariu D., Breazu M., Volovici D. DBSCAN algorithm for document clustering // International Journal of Advanced Statistics and IT&C for Economics and Life Sciences. 2019. V. 9. N 1. P. 58–66. https://doi.org/10.2478/ijasitels-2019-0007
7. Kotouza M.T., Psomopoulos F., Mitkas P. A dockerized framework for hierarchical frequency-based document clustering on cloud computing infrastructures // Journal of Cloud Computing. 2020. V. 9. N 1. P. 1–17. https://doi.org/10.1186/s13677-019-0150-y
8. Popat S.K., Deshmukh P.B., Metre V.A. Hierarchical document clustering based on cosine similarity measure // Proc. 1st International Conference on Intelligent Systems and Information Management (ICISIM). 2017. P. 153–159. https://doi.org/10.1109/ICISIM.2017.8122166
9. Nagarajan R., Nair S., Puviarasan N., Aruna P. Document clustering using agglomerative hierarchical clustering approach (AHDC) and proposed TSG keyword extraction method // IJRET: International Journal of Research in Engineering and Technology. 2016. V. 5. N 11. P. 118–124. https://doi.org/10.15623/ijret.2016.0511023
10. Rekabdar B., Mousas C., Gupta B. Generative adversarial network with policy gradient for text summarization // Proc. 13th IEEE International Conference on Semantic Computing (ICSC). 2019. P. 204–207. https://doi.org/10.1109/ICOSC.2019.8665583
11. Zhang Y., Li D., Wang Y., Fang Y., Xiao W. Abstract text summarization with a convolutional Seq2seq model // Applied Sciences. 2019. V. 9. N 8. P. 1665. https://doi.org/10.3390/app9081665
12. Jindal S.G., Kaur A. Automatic keyword and sentence-based text summarization for software bug reports // IEEE Access. 2020. V. 8. P. 65352–65370. https://doi.org/10.1109/ACCESS.2020.2985222
13. Varalakshmi K.P.N., Kallimani J.S. Survey on extractive text summarization methods with multi-document datasets // Proc. 7th International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2018. P. 2113–2119. https://doi.org/10.1109/ICACCI.2018.8554768
14. Thomas N. An e-business chatbot using AIML and LSA // Proc. 5th International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2016. P. 2740–2742. https://doi.org/10.1109/ICACCI.2016.7732476
15. Touimi Y.B., Hadioui A., Faddouli N.E., Bennani S. Intelligent Chatbot-LDA recommender system // International Journal of Emerging Technologies in Learning. 2020. V. 15. N 20. P. 4–20. https://doi.org/10.3991/ijet.v15i20.15657

References

1. Yin J., Wang J. A text clustering algorithm using an online clustering scheme for initialization. *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1995–2004. https://doi.org/10.1145/2939672.2939841
2. Svadas T., Jha J. Document cluster mining on text documents. *International Journal of Computer Science and Mobile Computing*, 2015, vol. 4, no. 6, pp. 778–782.
3. Kim H., Kim H.K., Cho S. Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling. *Expert Systems with Applications*, 2020, vol. 150, pp. 113288. https://doi.org/10.1016/j.eswa.2020.113288
4. Abasi A., Khader A., Al-Betar M., Naim S., Alyasseri Z.A., Makhadmeh S. A novel hybrid multi-verse optimizer with K-means for text documents clustering. *Neural Computing and Applications*, 2020, vol. 32, no. 23, pp. 17703–17729. https://doi.org/10.1007/s00521-020-04945-0
5. Mohammed S.M., Jacksi K., Zeebaree S.R.M. Glove word embedding and DBSCAN algorithms for semantic document clustering. *Proc. 3rd International Conference on Advanced Science and Engineering (ICOASE)*, 2020, pp. 211–216. https://doi.org/10.1109/ICOASE51841.2020.9436540
6. Cretulescu R., Morariu D., Breazu M., Volovici D. DBSCAN algorithm for document clustering. *International Journal of Advanced Statistics and IT&C for Economics and Life Sciences*, 2019, vol. 9, no. 1, pp. 58–66. https://doi.org/10.2478/ijasitels-2019-0007
7. Kotouza M.T., Psomopoulos F., Mitkas P. A dockerized framework for hierarchical frequency-based document clustering on cloud computing infrastructures. *Journal of Cloud Computing*, 2020, vol. 9, no. 1, pp. 1–17. https://doi.org/10.1186/s13677-019-0150-y
8. Popat S.K., Deshmukh P.B., Metre V.A. Hierarchical document clustering based on cosine similarity measure. *Proc. 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, 2017, pp. 153–159. https://doi.org/10.1109/ICISIM.2017.8122166
9. Nagarajan R., Nair S., Puviarasan N., Aruna P. Document clustering using agglomerative hierarchical clustering approach (AHDC) and proposed TSG keyword extraction method. *IJRET: International Journal of Research in Engineering and Technology*, 2016, vol. 5, no. 11, pp. 118–124. https://doi.org/10.15623/ijret.2016.0511023
10. Rekabdar B., Mousas C., Gupta B. Generative adversarial network with policy gradient for text summarization. *Proc. 13th IEEE International Conference on Semantic Computing (ICSC)*, 2019, pp. 204–207. https://doi.org/10.1109/ICOSC.2019.8665583
11. Zhang Y., Li D., Wang Y., Fang Y., Xiao W. Abstract text summarization with a convolutional Seq2seq model. *Applied Sciences*, 2019, vol. 9, no. 8, pp. 1665. https://doi.org/10.3390/app9081665
12. Jindal S.G., Kaur A. Automatic keyword and sentence-based text summarization for software bug reports. *IEEE Access*, 2020, vol. 8, pp. 65352–65370. https://doi.org/10.1109/ACCESS.2020.2985222
13. Varalakshmi K.P.N., Kallimani J.S. Survey on extractive text summarization methods with multi-document datasets. *Proc. 7th International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2018, pp. 2113–2119. https://doi.org/10.1109/ICACCI.2018.8554768
14. Thomas N. An e-business chatbot using AIML and LSA. *Proc. 5th International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 2740–2742. https://doi.org/10.1109/ICACCI.2016.7732476
15. Touimi Y.B., Hadioui A., Faddouli N.E., Bennani S. Intelligent Chatbot-LDA recommender system. *International Journal of Emerging Technologies in Learning*, 2020, vol. 15, no. 20, pp. 4–20. https://doi.org/10.3991/ijet.v15i20.15657

16. Юсупов И.Ф., Трофимова М.В., Бурцев М.С. Построение и использование диалогового графа для улучшения оценки качества в целенаправленном диалоге // Труды Московского физико-технического института (национального исследовательского университета). 2020. Т. 12. № 3(47). С. 75–86.
17. Feldina E., Makhnytkina O. Clustering approach to topic modeling in users dialogue // Advances in Intelligent Systems and Computing. 2021. V. 1251 AISC. P. 611–617. https://doi.org/10.1007/978-3-030-55187-2_44

16. Yusupov I.F., Trofimova M.V., Burtsev M.S. Unsupervised graph extraction for improvement of multi-domain task-oriented dialogue modelling. *Proceedings of Moscow Institute of Physics and Technology*, 2020, vol. 12, no. 3(47), pp. 75–86. (in Russian)
17. Feldina E., Makhnytkina O. Clustering approach to topic modeling in users dialogue. *Advances in Intelligent Systems and Computing*, 2021, vol. 1251 AISC, pp. 611–617. https://doi.org/10.1007/978-3-030-55187-2_44

Авторы

Фельдина Евгения Александровна — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57195673829, https://orcid.org/0000-0001-6208-691X, feldinazhenja@gmail.com

Махныткина Олеся Владимировна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57195435256, https://orcid.org/0000-0002-8992-9654, makhnytkina@itmo.ru

Authors

Evgeniya A. Feldina — Postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57195673829, https://orcid.org/0000-0001-6208-691X, feldinazhenja@gmail.com

Olesia V. Makhnytkina — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57195435256, https://orcid.org/0000-0002-8992-9654, makhnytkina@itmo.ru

Статья поступила в редакцию 23.07.2021
Одобрена после рецензирования 10.09.2021
Принята к печати 01.10.2021

Received 23.07.2021
Approved after reviewing 10.09.2021
Accepted 01.10.2021