Научная статья

УДК 004.912
DOI 10.25205/1818-7900-2021-19-4-50-66

Разработка чат-ботов для поддержки поиска по контенту веб-сайтов на основе тематических и жанровых характеристик

Владислав Дмитриевич Рублев [1]
Елена Анатольевна Сидорова [2]

[1] Новосибирский государственный университет
Новосибирск, Россия

[2] Институт систем информатики им. А. П. Ершова
Сибирского отделения Российской академии наук
Новосибирск, Россия

[1] v.rublev@g.nsu.ru, https://orcid.org/0000-0001-8236-0461
[2] lsidorova@iis.nsk.su, https://orcid.org/0000-0001-8731-3058

Аннотация

Рассматривается подход к созданию интеллектуальных помощников в виде чат-ботов, поддерживающих информационный поиск на основе предварительной жанровой и тематической кластеризации контента веб-сайтов. Решаются задачи поиска необходимой информации и обеспечения информационной поддержки пользователя, организации обратной связи для улучшения качества поиска. Особенностью подхода является использование жанровых моделей, разрабатываемых для заданного типа ресурса (образовательный, информационный и т. п.), на основе которых осуществляется жанровая структуризация контента конкретного сайта. Полученные жанровые структуры позволяют более точно определять границы тематических кластеров, относящиеся к теме поискового запроса пользователя. Для обеспечения обратной связи с пользователем разработан простой сценарий, позволяющий не просто уточнить запрос, но и неявно получить информацию о том, что именно не устроило пользователя в результирующей выдаче. Проведено экспериментальное исследование на платформе Telegram, полученные результаты сравнивались с поисковой системой Яндекс.

Ключевые слова

интеллектуальный помощник, поисковая система, информационный поиск, жанровая модель сайта, жанровая сегментация, тематическая кластеризация

Для цитирования

Рублев В. Д., Сидорова Е. А. Разработка чат-ботов для поддержки поиска по контенту веб-сайтов на основе тематических и жанровых характеристик // Вестник НГУ. Серия: Информационные технологии. 2021. Т. 19, № 4. С. 50–66. DOI 10.25205/1818-7900-2021-19-4-50-66

Development of Chatbots to Support Web Site Content Search Based on Thematic and Genre Characteristics

Vladislav D. Rublev [1], **Elena A. Sidorova** [2]

[1] Novosibirsk State University
Novosibirsk, Russian Federation

[2] A. P. Ershov Institute of Informatics Systems
of the Siberian Branch of the Russian Academy of Sciences
Novosibirsk, Russian Federation

[1] v.rublev@g.nsu.ru, https://orcid.org/0000-0001-8236-0461
[2] lsidorova@iis.nsk.su, https://orcid.org/0000-0001-8731-3058

Abstract

The paper considers an approach to creating intelligent assistants in the form of chatbots that support information search based on preliminary genre and thematic clustering of website content. The tasks of finding the necessary information and providing information support to the user, organizing feedback to improve the quality of the search are being solved. A feature of the approach is the use of genre models developed for a given type of resource (educational, informational, etc.), on the basis of which genre structuring of the content of a particular site is carried out. The resulting genre structures allow you to more accurately determine the boundaries of thematic clusters related to the topic of the user's search query. To provide feedback to the user, a simple script has been developed that allows not only to clarify the request, but also to implicitly get information about what exactly did not suit the user in the resulting output. An experimental study was conducted on the Telegram platform, the results were compared with the Yandex search engine.

Keywords

intelligent assistant, search engine, information search, genre model of the site, genre segmentation, thematic clustering

For citation

Rublev V. D., Sidorova E. A. Development of Chatbots to Support Web Site Content Search Based on Thematic and Genre Characteristics. *Vestnik NSU. Series: Information Technologies*, 2021, vol. 19, no. 4, p. 50–66. (in Russ.) DOI 10.25205/1818-7900-2021-19-4-50-66

Введение

Бурное развитие информационных технологий, в частности развитие сети Интернет, породило большое количество электронной информации. Простота и доступность создания и распространения данных привели к тому, что появился огромный поток нужной и ненужной информации. В настоящее время развивается информационный поиск, так как из большого количества данных необходимо находить только ту информацию, в которой заинтересован пользователь. Одним из способов улучшения качества поиска является использование методов анализа контента интернет-источников на основе знаний.

Для поиска с учетом семантики запроса существуют специальные поисковые системы. Функционал данных продуктов варьируется, начиная от переранжирования результатов поиска других поисковых систем и до полнотекстового поиска информации в проиндексированных текстах по ключевым запросам пользователей с учетом морфологических особенностей, синтаксиса и семантики слов. Также системы могут выполнять поиск с учетом синонимов и родственных слов, которые сгруппированы в соответствии с их семантическим значением. Так, система «Нигма» [1, с. 70–73] представляет собой метапоисковую систему, обеспечивающую поиск текстовой информации с учетом смысла запроса, заданного на естественном языке. Данный сервис также поддерживает формирование списка документов, разделенного на несколько кластеров, чтобы пользователь мог уточнить, в каком кластере продолжить поиск, тем самым улучшив качество поиска. Также существуют поисковые системы, основанные на тематическом кластерном анализе, например «Carrot2» [2]. Данная система предлагает два специализированных алгоритма кластеризации: Lingo – алгоритм,

основанный на сингулярном разложении, и STC – метод суффиксных деревьев. Поиск в системе «Yippy» [1, с. 74–77] основан на IBM Watson – суперкомпьютере, у которого есть технологии обработки естественного языка и который способен анализировать сложные, неструктурированные данные и даже понимать профессиональный сленг. «AskNet» [1] состоит из двух подсистем, позволяющих как поиск информации в Интернете, так и поиск информации на компьютерах пользователей в корпоративной сети. Система отличается от других поисковых систем тем, что на запрос пользователя она выдает не только ссылки на документы и ресурсы, а еще и текстовую информацию, являющуюся ответом на вопрос пользователя. «Hakia» [3] содержит свою лингвистическую базу данных, в которой слова подразделяются на различные «смыслы», которые они передают. Она извлекает все возможные запросы, относящиеся к контенту (используя свою базу данных), и они становятся путями к исходному документу. И далее независимо ранжирует контент на основе дополнительного анализа предложений. Также для определения релевантности использует достоверность и возраст контента.

Несмотря на то что множество из приведенных систем учитывают семантические особенности текста, они не используют предварительную тематическую кластеризацию и не учитывают жанровые особенности индексируемых веб-ресурсов. Также следует отметить, что большинство таких систем ориентировано на работу с текстами на английском языке.

Методы информационного поиска сегодня используются в разных приложениях, и одно из самых популярных приложений – мессенджер со встроенным интеллектуальным помощником, или чат-ботом. Существует 6 основных методов построения чат-ботов [2]: методы, основанные на правилах, поиске, генеративный подход, ансамблевые методы, обоснованное обучение и интерактивное обучение. Системы, основанные на правилах [4], обучаются на основе предопределенной иерархии правил, которые определяют, как преобразовать вводимые пользователем данные в ответ или действие. Основанные на поиске методы [5] применяются сегодня в большинстве чат-ботов. Такие системы работают с помощью ориентированных графов и обучены предоставлять наилучший возможный ответ из своей базы данных заранее определенных ответов. Вместо того чтобы использовать заранее определенные ответы, разговорный чат-бот, использующий генеративные методы [6], получает большое количество данных (реальных диалогов) и обучается генерировать новый диалог, который на них похож. Современные разговорные чат-боты, которые могут говорить на любую тему, были созданы с помощью ансамблевых методов [7], которые в зависимости от контекста используют некоторую комбинацию подходов на основе правил, поиска и генеративного подхода. Интеллектуальный помощник с использованием обоснованного обучения [8], анализируя вводимый пользователем запрос, генерирует нейронную сеть, которая настраивается для этого конкретного запроса и задачи. Такой интеллектуальный помощник лучше «обоснован» благодаря его способности учиться и использовать представления знаний реального мира. Интерактивное машинное обучение – это алгоритмы и интеллектуальные структуры пользовательского интерфейса, которые упрощают машинное обучение благодаря взаимодействию с человеком. Эта разработка позволяет компьютерам учиться у людей, «взаимодействуя» с ними на естественном языке и «наблюдая» за ними.

В данной работе предлагается подход, который интегрирует разные методы поиска, основанные на кластерном анализе, использует жанровые модели и опирается на интернет-жанр сайта. И для апробации предложенного подхода реализуется чат-бот для поиска по сайтам образовательных учреждений г. Новосибирска.

Для анализа, построения жанровой модели сайта и проведения экспериментов был составлен корпус сайтов общеобразовательных учреждений размером 9 760 текстов, который содержит контент 208 сайтов.

[1] Официальный сайт вопросно-ответной поисковой системы AskNet. URL: http://asknet.ru/.
[2] Technical Approaches for Building Conversational AI. URL: https://www.topbots.com/building-conversational-ai/.

Модель представления сайта определенного жанра

В данном подходе поиск базируется на предварительной индексации сайта на основе его жанровой структуры.

Каждый сайт обладает чертами, которые определяются спецификой сферы деятельности и формируются благодаря сходству тематики, композиции и стиля, что соответствует классическому определению речевого жанра, сформулированному М. М. Бахтиным [9]. Жанр – это типовая модель построения речевого целого. Для каждого типа сайта может быть определена жанровая модель, представляющая его «типическую воспроизводимую жанровую форму». Каждая жанровая модель представляет собой общую структуру сайтов, принадлежащих этой модели. Таким образом, контент каждого сайта может быть разбит на жанровые фрагменты, представляющие некоторые аспекты содержания. На основе анализа корпуса текстов разработана жанровая модель сайта образовательного учреждения, верхний уровень которой представлен в табл. 1.

Таблица 1

Спектр жанров фрагментов текста в зависимости от жанра сайта

Table 1

The range of genres of text fragments depending on the genre of the site

	Жанр сайта: образовательное учреждение	
	высшее	среднее
Жанр фрагмента	Описание научного учреждения	Прием в школу
	Описание факультетов	Родителям
	Описание кампуса	Учителям
	Поступающим	Итоговое сочинение
	Обучающимся	ГИА
	Выпускникам	
	Сотрудникам	
	Аспирантам	
	Новостная лента	
	Комментарий	
	Описание мероприятия	

Контент (содержательная часть) сайта представляет собой последовательность текстовых блоков. Для определения жанра этих блоков используется набор жанровых маркеров и язык маркеров, который позволяет указать термины, их комбинацию или перечисление. Для описания жанровой модели используется язык, предложенный в работе [10], который позволяет на основе жанровых маркеров составлять описание аспектов содержания любого жанрового блока. Жанровая модель содержит наборы маркеров, каждый набор описывает некоторый аспект содержания, для каждого жанра описываются типичные аспекты содержания и указывается, в какой части html-разметки их искать.

Шаблон для описания аспектов содержания:

Аспект содержания: [“Маркер 1”][“Маркер 2”][“Маркер 3”][“Маркер 4”, “Маркер 5”].

Шаблон для описания жанров:

“**Идентификатор жанра**”: [<Аспект содержания 1, тег>] [<Аспект содержания 2, тег > <Аспект содержания 3, тег >].

Примеры аспектов содержания:

_поступающим*: ["поступать"]["абитуриент"]["прием"]["приемная комиссия"]["приемная кампания"]["правило приема"]["рейтинговый список"]["стоимость обучения"]*

_уровниОбразов*: ["бакалавриат"]["специалитет"]["магистратура"]["аспирантура"]*

Пример описания жанра страницы:

*"**поступающим**":* [*<_поступающим, text>*][*<_поступающим, all><_уровниОбразов, all>*].

Разработанная жанровая модель используется для выделения жанровых фрагментов в тексте.

Тематическая кластеризация текстов

После сегментации текстового контента сайта осуществляется его тематическая кластеризация – выделение тематических кластеров, по которым в дальнейшем будет осуществляться поиск.

Кластеризация (кластерный анализ) – это задача группирования множества объектов в группы, называемые кластерами, чтобы внутри каждой группы оказались «похожие» элементы, существенно отличающиеся от элементов других групп. Предметом данного исследования являются методы кластеризации текстов.

Для кластеризации контента и анализа запроса пользователя необходим словарь, поэтому в системе извлечения предметной лексики KLAN [11] был создан словарь терминов. Эта система поддерживает основные этапы анализа текста: синтаксический, семантический и морфологический. Словарь, созданный на основе корпуса сайтов общеобразовательных учреждений, содержит однословные и многословные термины, его размер 17 500 терминов, для которых собрана статистика.

По способу организации кластеров алгоритмы кластеризации можно классифицировать на плоские и иерархические. Был использован алгоритм k-средних [12], вследствие того что обладает высокой скоростью обучения и агломеративной иерархической кластеризации [13], потому что позволяет хорошо интерпретировать результат.

Каждый документ представлен вектором в пространстве R^n, где n – размерность словаря. Для получения такого представления используется статистическая мера tf-idf [14], которая показывает частоту встречаемости термина в документе. Расстояние между векторами можно вычислять по различным метрикам, были опробованы соответственно косинусное сходство и Евклидова метрика:

$$\rho(a,b) = \frac{\sum_{k=1}^{n} a_k \times b_k}{\sqrt{\sum_{k=1}^{n} a_k^2} \times \sqrt{\sum_{k=1}^{n} b_k^2}},$$

$$\rho(a,b) = \sqrt{\sum_{k=1}^{n} (a_k - b_k)^2}.$$

Косинусное сходство статистически хуже работает на собранном корпусе текстов, поэтому в данном исследовании используется Евклидова метрика.

Метод k-средних – это итеративный алгоритм (рис. 1), который основан на минимизации суммарного квадратичного отклонения точек кластеров от центров этих кластеров.

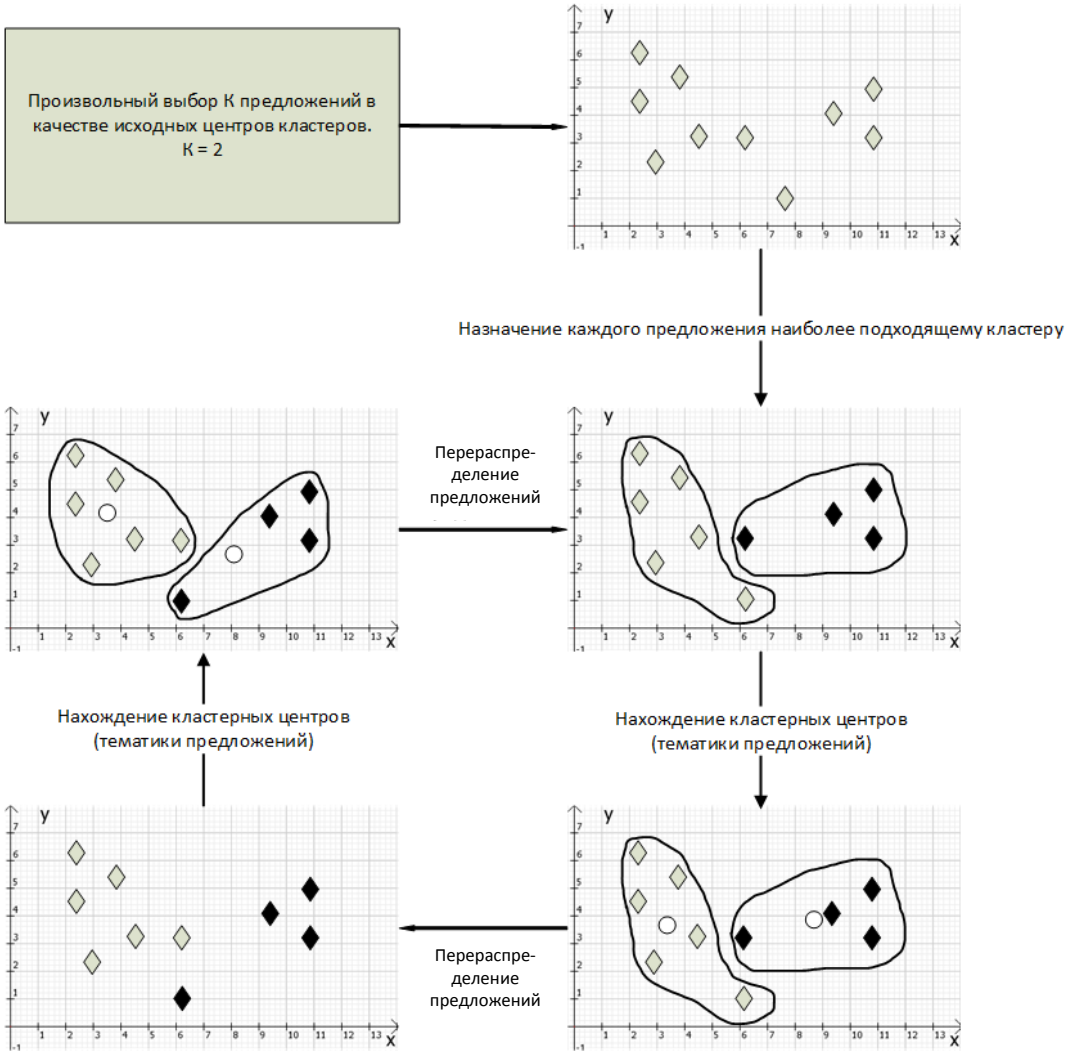

Рис. 1. Метод k-средних
Fig. 1. K-means clustering

Центроид (центр кластера) – вектор, элементы которого представляют средние значения, вычисленные по всем предложениям (представленным в векторном виде) из кластера. Из множества входных данных выбирается *k* предложений, которые будут начальными центрами кластеров (каждый кластер соответствует своей теме), для каждого предложения определяется ближайший центр кластера и на каждой итерации перевычисляется центроид каждого кластера (S_j), с учетом координат новых предложений:

$$\mu_j = \frac{1}{|S_j|} \sum_{x^{(j)} \in S_j} x^{(j)}.$$

Далее предложения снова разбиваются на кластеры в соответствии с тем, какой из новых центроидов оказался ближе. Алгоритм выполняется до тех пор, пока на каком-то шаге не произойдет изменения внутрикластерного расстояния или не будет достигнут порог по коли-

честву итераций. Таким образом, центроид является векторным представлением тематики предложений, входящих в кластер.

Агломеративные алгоритмы работают по принципу «снизу вверх» (рис. 2): в начале работы помещают каждое предложение в отдельный кластер, а затем происходит объединение двух ближайших во все более крупные, пока все кластеры не сольются в один или не будет найдено необходимое число кластеров. Таким образом строится дерево разбиений.

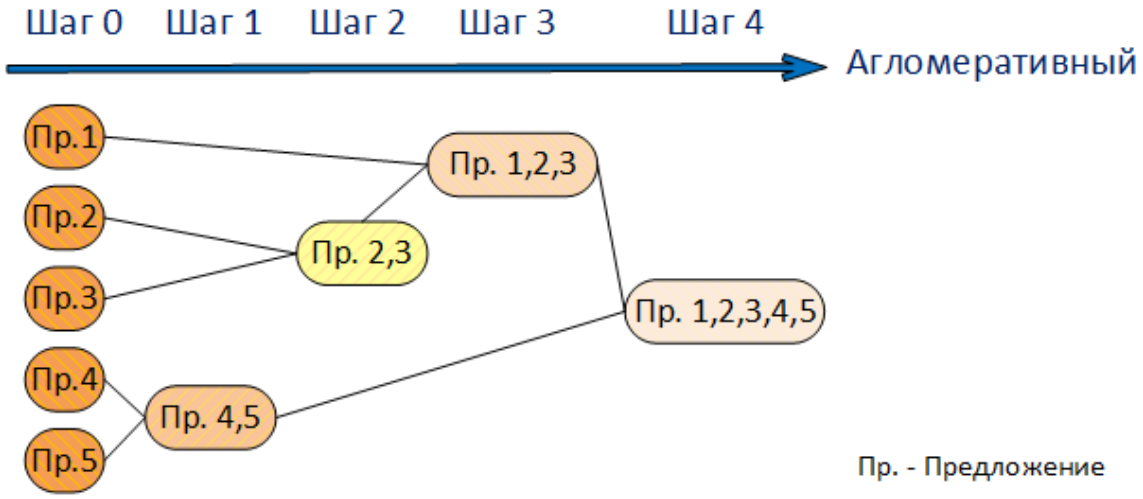

Рис. 2. Иерархическая кластеризация
Fig. 2. Hierarchical clustering

Для вычисления расстояний между кластерами чаще всего пользуются следующими методами:

1) метод одиночной связи – расстояние между двумя кластерами определяется как минимальное расстояние между элементами этих кластеров:

$$\min\{\rho(a,b):\ a\in A,\ b\in B\};$$

2) метод полной связи – расстояние между двумя кластерами определяется как максимальное расстояние между элементами этих кластеров:

$$\max\{\rho(a,b):\ a\in A,\ b\in B\};$$

3) метод средней связи – расстояние между кластерами определяется как среднее расстояние между элементами этих кластеров:

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B}\rho(a,b);$$

4) центроидный метод – расстояние между кластерами полагается равным расстоянию между центроидами кластеров:

$$\|\mu_A-\mu_B\|.$$

В работе используется метод средней связи, потому что он лучше работает на собранном корпусе текстов.

Существует большое количество мер для оценки качества кластеризации и проведено множество их сравнений [15–17]. Одной из таких мер является “Silhouette score” [18], которая статистически значимо показывает результаты лучше, чем остальные меры. Вычисляется отдельно для каждого объекта по формуле

$$s = \frac{b - a}{\max\{a, b\}},$$

где a – среднее расстояние от выбранного объекта до объектов из его кластера, b – среднее расстояние от выбранного объекта до объектов ближайшего кластера (не содержащего выбранный объект). Для оценки качества кластеризации вычисляется среднее значение “Silhouette score” по всем объектам. С помощью этой меры была проведена оценка методов кластеризации, и результаты получились неоднозначные, так как исход сильно зависел от входных данных.

Поиск по сайту

Поиск информации, необходимой пользователю, осуществляется на основе предварительного индексирования, построенного с помощью жанрового анализа и тематической кластеризации.

На рис. 3 представлена архитектура поисковой системы, которая включает два основных модуля: подсистема предварительной обработки и поисковая подсистема.

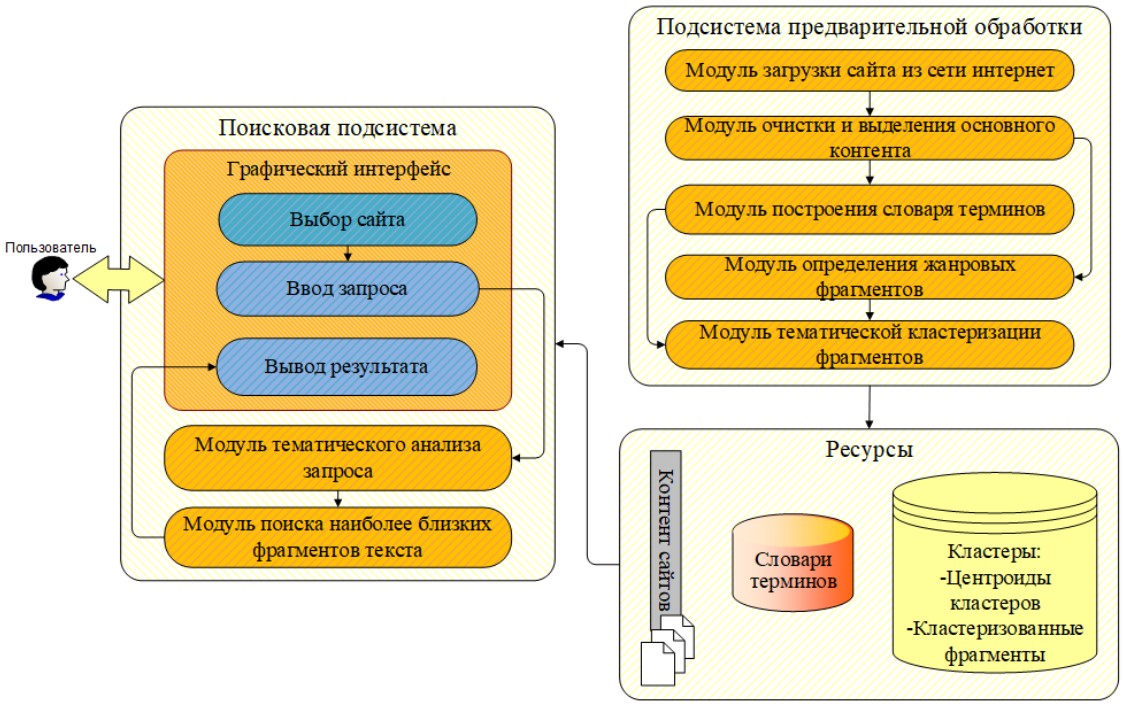

Рис. 3. Архитектура поисковой системы
Fig. 3. Search Engine architecture

Предварительная обработка сайта делается заранее и состоит из следующих этапов:

1) загрузка веб-сайта из сети Интернет – рекурсивно находятся все ссылки на страницы сайта и скачиваютсяж

2) очистка и выделение основного контента – на всех страницах сайта находятся и удаляются данные, которые не являются основным контентом (футер, выпадающее меню и т. д.), также удаляются все html-теги, кроме разрешенных: h1, h2, h3, h4, h5, h6, a, b, ul, ol, li;

3) генерация словаря сайта – полученные файлы с основным контентом сайта загружаются в словарную систему, где в автоматическом режиме находятся все термины и для них собирается статистика;

4) предварительная обработка текста – производится лемматизация, удаление стоп-слов и слов, которые встречаются слишком редко;

5) жанровая классификация – на базе разработанной жанровой модели находится количество жанров на странице, если их несколько, то страница разбивается на фрагменты от заголовка до заголовка, и определяются жанры этих фрагментов. Фрагменты одного жанра, идущие подряд объединяются в один, а фрагменты, жанр которых определить не удалось, выбрасываются из рассмотрения;

6) тематическая кластеризация – выбирается один из доступных алгоритмов кластеризации, текст представляется в векторном виде с помощью меры tf-idf, определяется количество кластеров, на которое будет разбиваться текст в зависимости от его жанра и объема. Производится кластеризация выбранным методом на заданное количество кластеров;

7) сохранение результатов – сохраняются результат кластеризации, текстовая коллекция и словарь сайта.

В результате индексации сайт Новосибирского государственного университета содержит 153 страницы, 998 кластеров и 4 661 предложение. Сайт Московского государственного университета имени М. В. Ломоносова состоит из 123 страниц, 858 кластеров и 2 512 предложений.

Для поиска ответов на запросы пользователя необходимо:

1) произвести анализ запроса пользователя – при помощи словаря сайта строится вектор запроса, состоящий из нулей и единиц, где единица – слово содержится в запросе, нуль – иначе;

2) найти и выдать пользователю фрагменты, наиболее релевантные запросу, – с помощью Евклидовой метрики находятся расстояния от вектора запроса до центров кластеров, берется n ближайших кластеров и определяется, каким текстам из коллекции они соответствуют.

Информационный поиск, представленный выше, используется для создания интеллектуальных помощников, реализованных в виде чат-ботов.

Информационный чат-бот

Информационный чат-бот – программное обеспечение, которое имитирует диалог с пользователем на естественном языке и используется для оперативного поиска информации по заданной тематике. Чат-боты позволяют общаться с помощью сообщений на сайтах, в мессенджерах или мобильных приложениях.

Для повышения качества информационной поддержки пользователя разработан сценарий взаимодействия с чат-ботом (рис. 4).

Сценарий состоит из следующих этапов. Пользователь задает поисковый запрос чат-боту. Чат-бот отправляет результат поиска и предлагает пользователю уточнить запрос (если пользователя не устраивает результат) или начать новый поиск. Если пользователь хочет уточнить запрос, то чат-бот предлагает ему один из трех параметров поиска (поиск с исключением предыдущих результатов, расширение или сужение выборки поиска). Вместе с параметром поиска пользователю предлагается уточнить запрос (написать новый или исправить предыдущий). Далее чат-бот возвращается в начало сценария и отвечает пользователю на вопрос с учетом выбранного параметра поиска. Благодаря цикличности сценария удается добиться итеративного улучшения качества поиска.

Для апробации подхода в качестве интерфейса был выбран мессенджер Telegram в силу своей популярности, простоты и бесплатного доступа к API.

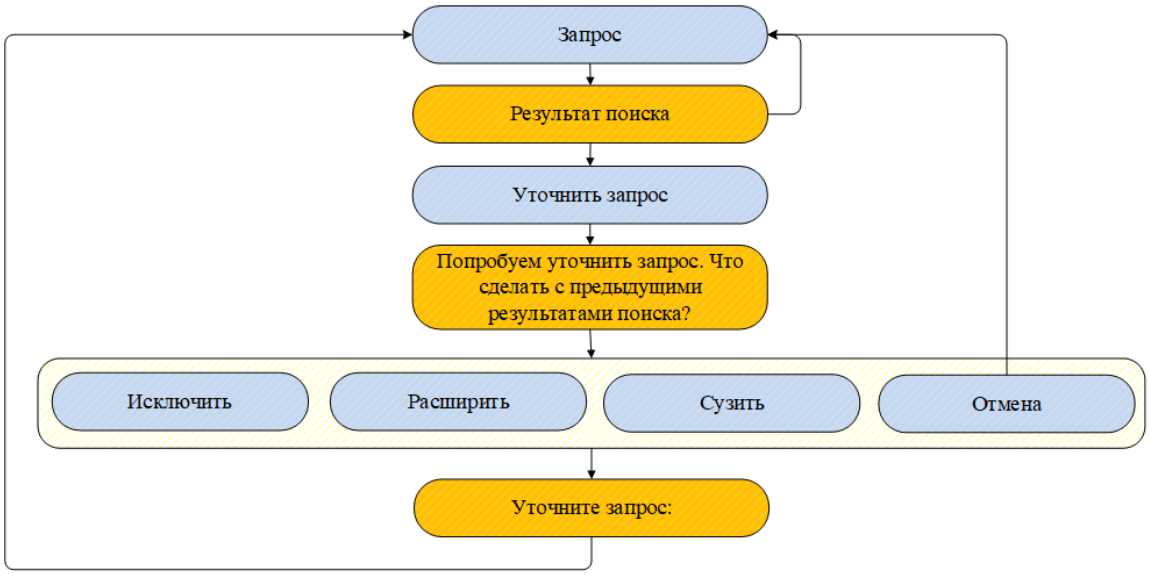

Рис. 4. Сценарий взаимодействия пользователя с чат-ботом
Fig. 4. Scenario of user interaction with a chatbot

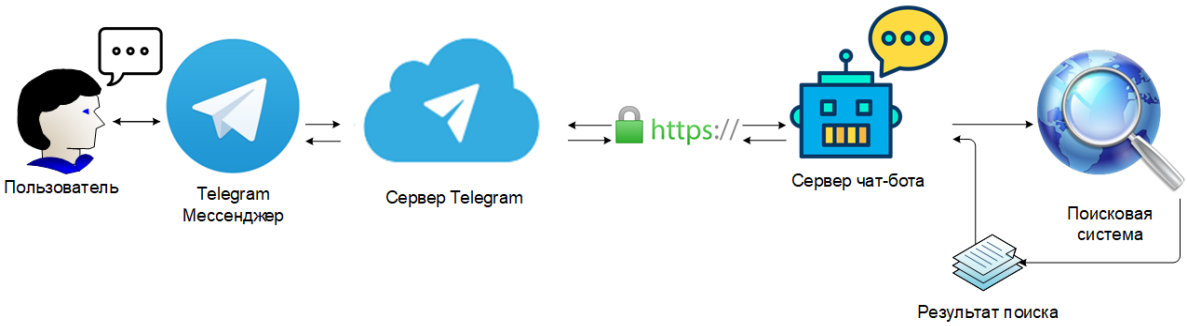

Рис. 5. Схема работы интеллектуального помощника на базе мессенджера Telegram
Fig. 5. The scheme of the intelligent assistant based on the Telegram messenger

Боты – специальные аккаунты в Telegram, созданные для того, чтобы автоматически обрабатывать и отправлять сообщения. Пользователи могут взаимодействовать с ботами при помощи сообщений, отправляемых через обычные или групповые чаты. Для реализации чат-ботов Telegram предоставляет Telegram Bot API [3]. Логика бота контролируется при помощи HTTPS запросов к этому API. Сообщения и запросы, отправленные пользователями чат-боту, передаются программному обеспечению, работающему на сервере (рис. 5). Всё шифрование и связь с Telegram API промежуточный сервер Telegram обрабатывает самостоятельно. Все запросы к этому серверу должны быть следующего вида:

https://api.telegram.org/bot<token>/METHOD_NAME?Param1=<p1>&ParamN=<pn>,

[3] Telegram Bot API. URL: https://core.telegram.org/bots/api/.

где

<token> – уникальный идентификатор чат-бота,

METHOD_NAME – название метода, который необходимо вызвать у бота,

Param1, ParamN – набор параметров вызываемого метода (может отсутствовать),

p1, pn – значения параметров.

Ответ на все запросы приходит в виде json-объекта, в котором присутствует булево поле *ok*, которое истинно в случае успешного запроса, и результат его выполнения можно увидеть в поле «result».


```
{
"ok":true,
"result":
{
"message_id":594,
"from": {"id":457981543,
          "is bot":true,
          first_name":"RublevBot",
          "username":"RubleffBot",
          "language_code": "ru"},
"chat":{"id":254438520,
         "first_name":"Vladislav",
         "last name":"Rublev",
         "username":"spac1k",
         "type": "private"},
"date":1617858474,
"text":"Привет"
}
}
```


Если сделать запрос с методом *send_message* (отправить сообщение), то поле «result» будет содержать информацию о номере сообщения, отправителе, получателе, дате запроса, а также текст отправленного сообщения.

В случае ошибки (*ok: false*) в поле *description* будет указана причина ошибки, а в поле *error_code* указан код ошибки:


```
{
"ok":false,
"error_code":400,
"description":"Bad Request: message text is empty"
}
```


Разработанный поисковый чат-бот на запрос пользователя производит информационный поиск и отправляет результат в виде трех сообщений с фрагментами текста и гиперссылками на источник (рис. 6). Если пользователя не устраивают результаты поиска, то он может уточнить запрос (нажав кнопку «Уточнить запрос»). Далее пользователю необходимо выбрать один из трех параметров поиска: «Исключить предыдущий результат», «Расширить выборку», «Сузить выборку (искать в найденном)».

После выбора параметра пользователь уточняет сам запрос (рис. 7) и в ответ получает результат повторного поиска с параметром.

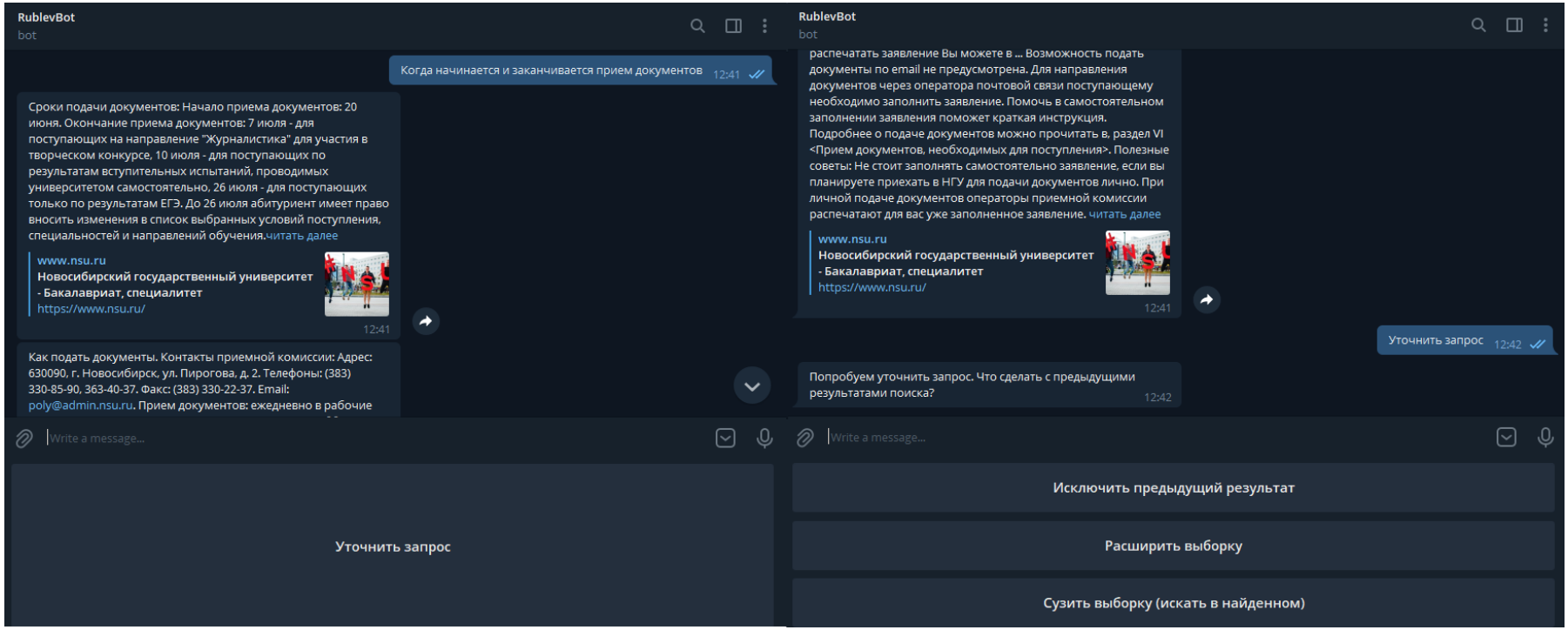

Рис. 6. Пример работы чат-бота после запроса пользователя
Fig. 6. An example of a chatbot working after a user request

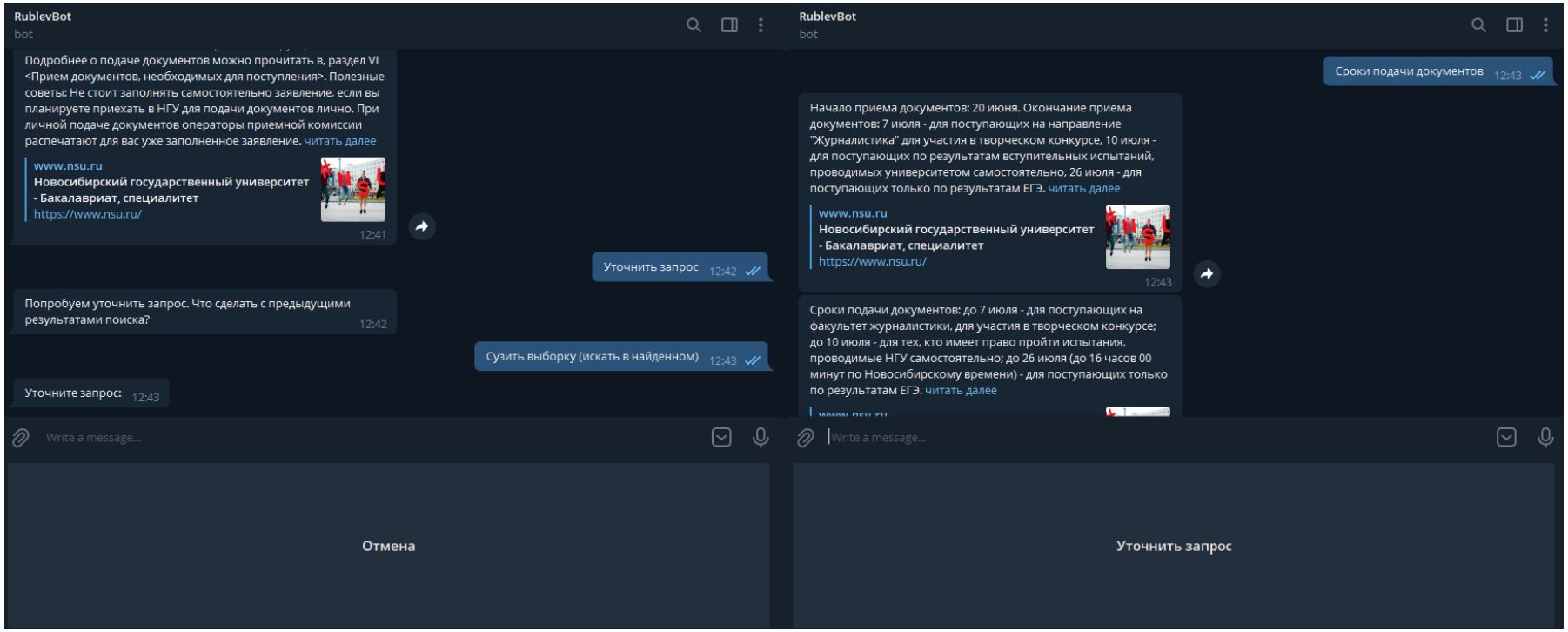

Рис. 7. Пример работы чат-бота во время уточнения запроса
Fig. 7. An example of a chatbot working during request refinement

Экспериментальное исследование

Качество работы чат-бота при поиске оценивалось в сравнении с поисковой системой компании «Яндекс». Поиск осуществлялся по сайту Новосибирского государственного университета. В качестве поисковых запросов были взяты часто задаваемые вопросы (19 вопросов) с сайта Московского государственного технического университета им. Н. Э. Баумана [4].

Пусть rel – множество всех релевантных документов, det – множество найденных документов. Для оценки качества использовались стандартные меры для оценки информационного поиска [19] – полнота, точность, F-мера и оригинальность.

1. Полнота определяет, насколько хорошо система находит нужные пользователю документы, представляет собой отношение найденных релевантных документов к общему количеству релевантных документов:

$$R = \frac{\mathrm{rel} \cap \mathrm{det}}{\mathrm{rel}}.$$

2. Точность определяет способность системы выдавать пользователю только релевантные документы, вычисляется как отношение найденных релевантных документов к общему количеству найденных документов:

$$P = \frac{\mathrm{rel} \cap \mathrm{det}}{\mathrm{det}}.$$

3. F-мера представляет собой взвешенное гармоническое среднее полноты и точности, позволяет придать различный вес полноте и точности, если необходимо отдать приоритет одной из этих метрик:

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}, \ \alpha \in [0,1].$$

При α =1/2 получается сбалансированная F-мера и вычисляется она по следующей формуле:

$$F = \frac{2PR}{P+R}.$$

4. Оригинальность определяет количество различных результатов поиска (документы с разным контентом).

Средние значения этих показателей были вычислены по всем вопросам и первым 3-м результатам выдачи поисковых систем. Оценка этих значений проводилась вручную [20]. Полученные результаты отражены в табл. 2.

Таблица 2

Оценка качества поиска, %

Table 2

Evaluation of search quality, %

	Полнота	Точность	F-мера	Оригинальность
«Яндекс»	82,16	91,83	86,73	71,84
Чат-бот	83,37	92,75	87,81	80,92

[4] Официальный сайт МГТУ им. Н. Э. Баумана. Ответы на часто задаваемые вопросы. URL: https://bmstu.ru/abitur/general/qanda/.

При сравнении результатов поиска, полученных разработанным интеллектуальным помощником и системой компании «Яндекс», можно сделать вывод, что помощник немного превосходит систему «Яндекс» для поиска по конкретным сайтам. При этом обе поисковые системы предоставляют возможность найти ответы на основные вопросы пользователя, но не гарантируют получения всей необходимой информации. Но в отличие от поисковика «Яндекс» чат-боту благодаря «диалогу» с пользователем удается добиться итеративного улучшения качества поиска.

Заключение

В работе представлен подход к созданию информационных помощников для поиска в контенте веб-сайтов на основе жанровой модели и предварительной тематической кластеризации текстового контента. Предлагаемый подход позволяет найти необходимую информацию, организовать обратную связь с пользователем и обеспечить итеративное улучшение результатов поиска. Особенностью подхода является использование жанровой информации о сайте, на основе которой осуществляется жанровая сегментация, которая позволяет более точно структурировать контент. Над множеством жанровых сегментов осуществляется тематическая кластеризация, целью которой является выделение и группировка фрагментов текста, относящихся к одной области. Дальнейший поиск осуществляется стандартными методами. Результаты экспериментов показывают, что добавление таких особенностей улучшает результаты поиска, и, следовательно, предложенный подход можно применять для улучшения качества поиска, например, в метапоисковых системах (смешивания и переранжирования результатов поиска других поисковых систем).

Разработанная система хорошо масштабируется, в частности созданные ресурсы применимы для произвольных образовательных сайтов, а для того чтобы настроить систему на другие типы сайтов, достаточно написать новую жанровую модель и проиндексировать заданные сайты нового типа (для этого в системе разработан независимый модуль индексации).

Список литературы

1. **Кутовенко А.** Профессиональный поиск в интернете. СПб.: Питер, 2011. 252 с.
2. **Stanislaw Osinski, Dawid Weiss**. Carrot2 Project. In: Carrot2 – Open Source Search Results Clustering Engine. URL: http://project.carrot2.org/.
3. **Radhakrishnan Arun.** Hakia’s Semantic Search : The Answer to Poor Keyword Based Relevancy. *Search Engine Journal*. URL: https://www.searchenginejournal.com/hakias-semantic-search-the-answer-to-poor-keyword-based-relevancy/5246/.
4. **Nimavat K., Champaneria T.** Chatbots: an overview of types, architecture, tools and future possibilities. *Int. J. Sci. Res. Dev.*, 2017, pp. 1019–1024.
5. **Wu Y., Wu W., Xing C., Zhou M., Li Z.** Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. In: ArXiv:11612.01627, 2017.
6. **Kapočiūtė-Dzikienė J.** A Domain-Specific Generative Chatbot Trained from Little Data. *Applied Sciences*, 2020, vol. 10, p. 2221.
7. **Heriberto Cuayáhuitl, Donghyeon Lee, Seonghan Ryu, Yongjin Cho, Sungja Choi, Satish Indurthi, Seunghak Yu, Hyungtak Choi, Inchul Hwang, Jihie Kim.** Ensemble-based deep reinforcement learning for chatbots. *Neurocomputing*, 2019, vol. 366, pp. 118–130.
8. **Kim Sihyung, Kwon Oh-Woog, Kim Harksoo.** Knowledge-Grounded Chatbot Based on Dual Wasserstein Generative Adversarial Networks with Effective Attention Mechanisms. *Applied Sciences*, 2020, vol. 10.
9. **Бахтин М. М.** Проблема речевых жанров // Эстетика словесного творчества. М.: Искусство, 1986. С. 250–296.

10. **Кононенко И. С., Сидорова Е. А.** Жанровые аспекты классификации веб-сайтов // Программная инженерия. 2015. № 8. С. 32–40.
11. **Сидорова Е. А.** Комплексный подход к исследованию лексических характеристик текста // Вестник СибГУТИ. 2019. № 3. С. 80–88.
12. **MacQueen J. B.** Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 1967, pp. 281–297.
13. **Guo J., Hartung S., Komusiewicz C. et al.** Exact algorithms and experiments for hierarchical tree clustering. In: Proceedings of the TwentyFourth AAAI Conference on Artificial Intelligence (AAAI-10), 2010, pp. 1–6.
14. **Manwar A., Mahalle H., Chinchkhede K. et al.** A vector space model for information retrieval: a matlab approach. *Indian Journal of Computer Science and Engineering*, 2012, no. 3, pp. 222–230.
15. **Erendira Rendon, Itzel Abundez, Alejandra Arizmendi et al.** Internal versus external cluster validation indexes. *International Journal of Computers and Communications*, 2011, vol. 5, no. 1, pp. 27–34.
16. **Yanchi Liu, Zhongmou Li, Hui Xiong et al.** Understanding of internal clustering validation measures. In: IEEE International Conference on Data Mining, 2010, pp. 911–916. DOI 10.1109/tsmcb.2012.2220543
17. **Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza et al.** An extensive comparative study of cluster validity indices. *Pattern Recognition*, 2013, vol. 46, no. 1, pp. 243–256. DOI 10.1016/j.patcog.2012.07.021
18. **Rousseeuw Peter J.** Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 1987, vol. 20, pp. 53–65. DOI 10.1016/0377-0427(87)90125-7
19. **Sirotkin P. F.** On Search Engine Evaluation Metrics. In: ArXiv:abs/1302.2318, 2013, pp. 24–26.
20. **Белозеров В. Н.** Эффективность систем Яндекс и Гугл для поиска учебного материала // Вестник МГУКИ. 2015. № 1. С. 208–213.

References

1. **Kutovenko A.** Professional internet search. St. Petersbug, Peter, 2011, 252 p. (in Russ.)
2. **Stanislaw Osinski, Dawid Weiss**. Carrot2 Project. In: Carrot2 – Open Source Search Results Clustering Engine. URL: http://project.carrot2.org/.
3. **Radhakrishnan Arun.** Hakia's Semantic Search : The Answer to Poor Keyword Based Relevancy. *Search Engine Journal*. URL: https://www.searchenginejournal.com/hakias-semantic-search-the-answer-to-poor-keyword-based-relevancy/5246/.
4. **Nimavat K., Champaneria T.** Chatbots: an overview of types, architecture, tools and future possibilities. *Int. J. Sci. Res. Dev.*, 2017, pp. 1019–1024.
5. **Wu Y., Wu W., Xing C., Zhou M., Li Z.** Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. In: ArXiv:11612.01627, 2017.
6. **Kapočiūtė-Dzikienė J.** A Domain-Specific Generative Chatbot Trained from Little Data. *Applied Sciences*, 2020, vol. 10, p. 2221.
7. **Heriberto Cuayáhuitl, Donghyeon Lee, Seonghan Ryu, Yongjin Cho, Sungja Choi, Satish Indurthi, Seunghak Yu, Hyungtak Choi, Inchul Hwang, Jihie Kim.** Ensemble-based deep reinforcement learning for chatbots. *Neurocomputing*, 2019, vol. 366, pp. 118–130.
8. **Kim Sihyung, Kwon Oh-Woog, Kim Harksoo.** Knowledge-Grounded Chatbot Based on Dual Wasserstein Generative Adversarial Networks with Effective Attention Mechanisms. *Applied Sciences*, 2020, vol. 10.

9. **Bahtin M. M.** The problem of speech genres. In: Estetika slovesnogo tvorchestva [Aesthetics of Verbal Creation]. Moscow, Iskusstvo, 1986, pp. 250–296. (in Russ.)
10. **Kononenko I. S., Sidorova E. A.** Genre aspects of website classification. Software Engineering, 2015, no. 8, pp. 32–40.
11. **Sidorova E. A.** A comprehensive approach to the study of lexical characteristics of the text. *Vestnik SibSUTI*, 2019, no. 3, pp. 80–88.
12. **MacQueen J. B.** Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 1967, pp. 281–297.
13. **Guo J., Hartung S., Komusiewicz C. et al.** Exact algorithms and experiments for hierarchical tree clustering. In: Proceedings of the TwentyFourth AAAI Conference on Artificial Intelligence (AAAI-10), 2010, pp. 1–6.
14. **Manwar A., Mahalle H., Chinchkhede K. et al.** A vector space model for information retrieval: a matlab approach. *Indian Journal of Computer Science and Engineering*, 2012, no. 3, pp. 222–230.
15. **Erendira Rendon, Itzel Abundez, Alejandra Arizmendi et al.** Internal versus external cluster validation indexes. *International Journal of Computers and Communications*, 2011, vol. 5, no. 1, pp. 27–34.
16. **Yanchi Liu, Zhongmou Li, Hui Xiong et al.** Understanding of internal clustering validation measures. In: IEEE International Conference on Data Mining, 2010, pp. 911–916. DOI 10.1109/tsmcb.2012.2220543
17. **Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza et al.** An extensive comparative study of cluster validity indices. *Pattern Recognition*, 2013, vol. 46, no. 1, pp. 243–256. DOI 10.1016/j.patcog.2012.07.021
18. **Rousseeuw Peter J.** Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 1987, vol. 20, pp. 53–65. DOI 10.1016/0377-0427(87)90125-7
19. **Sirotkin P. F.** On Search Engine Evaluation Metrics. In: ArXiv:abs/1302.2318, 2013, pp. 24–26.
20. **Belozerov V. N.** The efficiency of the engines Yandex and Google to search for educational material. *Vestnik MGIK*, 2015. no. 1, pp. 208–213.

Информация об авторах

Владислав Дмитриевич Рублев, студент магистратуры
Елена Анатольевна Сидорова, кандидат физико-математических наук, старший научный сотрудник

Information about the Authors

Vladislav D. Rublev, Master's Student
Elena A. Sidorova, Candidate of Sciences (Physics and Mathematics), Senior Researcher

Статья поступила в редакцию 01.08.2021;
одобрена после рецензирования 01.10.2021; *принята к публикации 01.12.2021*
The article was submitted 01.08.2021;
approved after reviewing 01.10.2021; *accepted for publication 01.12.2021*