Информатика. Экономика. Управление//
Informatics. Economics. Management

2023; 2(1)
http://oajiem.com/

УДК: 004.891
DOI: https://doi.org/10.47813/2782-5280-2023-2-1-0101-0123

EDN: KJUXWO

Способы решения проблемы документального тематического поиска

П. В. Юрченко

Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича, Санкт-Петербург, Россия

Аннотация. В современном мире количество информации несоизмеримо растет. Вследствие этого тематический документальный поиск становится трудоемким процессом. Такой поиск подразумевает поиск документов, содержащих координированную информацию в заданном тематическом сегменте. Библиотеки являются хранилищами знаний, поэтому они могут предоставить сервис для осуществления упомянутого поиска. Рост объёмов библиографической информации обусловливает необходимость цифровизации библиотечной деятельности. В статье рассмотрены технологии обработки естественного языка и интеллектуального анализ текстов. Внедрение этих технологий в библиотечную сферу является перспективным для решения проблемы. Найдены и проанализированы системы, реализующие цифровые технологии для тематического поиска. К таким системам относятся платформа Yewno и модуль Discovery Znanium. Эти продукты действительно делают процесс документального поиска более удобным, но не являются полноценным решением поставленной проблемы. В результате статьи было предложено 2 способа решения проблемы. Первый способ заключается в создании алгоритма обработки запроса читателя с учётом его индивидуальных характеристик. Второй способ предполагает создание системы, выполняющей функции библиографа в рамках справочно-библиографического обслуживания. Таким образом, предложенные способы решения проблемы основываются на интерпретации традиционной библиотечной деятельности в цифровую сферу. По этой причине было исследовано текущее состояние развития библиотек и классическая схема библиотечного обслуживания.

Ключевые слова: тематический поиск, библиотеки, библиографическое обслуживание, цифровизация, экспертные системы, интеллектуальный анализ, машинное обучение.

Для цитирования: Юрченко, П. (2023). Способы решения проблемы документального тематического поиска. Информатика. Экономика. Управление - Informatics. Economics. Management, 2(1), 0101–0123. https://doi.org/10.47813/2782-5280-2023-2-1-0101-0123

Ways to solve the problem of documentary thematic search

P. V. Yurchenko

The Bonch-Bruevich Saint-Petersburg State University of Telecommunication, Saint-Petersburg, Russia

Abstract. In today's world, the amount of information is growing exponentially. As a result, thematic documentary search becomes a time-consuming process. Such a search involves searching for documents containing coordinated information in a given thematic segment. Libraries are repositories of knowledge; therefore, they can provide a service to carry out this search. The growth in the volume of bibliographic information necessitates the digitalization of library activities. The article considers technologies for natural language processing and text mining. The integration of these technologies in the library sector is promising for solving the problem. The study found and analyzed systems that implement digital technologies for thematic search. These systems are the Yewno platform and the Discovery Znanium module. These products do make the process of documentary search more convenient, but they are not a complete solution to the problem. As a result of the article, 2 methods to solve the problem were proposed. The first method is to create an algorithm for processing the reader's request, taking into account its individual characteristics. The second method involves the creation of a system that performs the functions of a bibliographer within the framework of reference and bibliographic services. Thus, the proposed ways of solving the problem are based on the interpretation of traditional library activities in the digital sphere. For this reason, the article investigated the current state of library development and the classical scheme of library services.

Keywords: thematic search, libraries, bibliographic services, digitalization, expert systems, text mining, machine learning.

For citation: Yurchenko, P. (2023). Ways to solve the problem of documentary thematic search. Informatics. Economics. Management, 2(1), 0101–0123. https://doi.org/10.47813/2782-5280-2023-2-1-0101-0123

ВВЕДЕНИЕ

В настоящее время наблюдается лавинообразное увеличение цифровой информации. Согласно опубликованному в январе 2021 г. международной компанией Statista отчёту, объём данных/информации, созданных, переведённых в цифровой вид, скопированных и потреблённых во всём мире в 2020 г. оценивается примерно в 74 зеттабайта [1].

В связи с этим люди тратят большое количество времени на документальный тематический поиск, то есть на поиск документов, содержащих координированную информацию в заданном тематическом сегменте. При этом качество упомянутого поиска не всегда удовлетворительно [2].

Так происходит по нескольким причинам. Во-первых, поиск обычно проводится преимущественно по каким-либо ключевым словам, в то время как семантическая составляющая документов остается доступной преимущественно только для человека [3]. Во-вторых, искомая информация часто находится на стыке смежных областей [2]. В-третьих, одновременно с информацией о предмете поиска желательно получать множество сопутствующей информации, например, сведения о ретроспективе, перспективе, взаимосвязях найденных информационных объектов и т.д. [2].

Несмотря на наличие мощных поисковых систем для информационных ресурсов Интернета или для специализированных баз данных (БД), процесс поиска продолжает оставаться трудоемким и слабо поддерживается программно и методологически [2].

Как следствие, пользователи вынуждены применять в поисковых запросах множество сочетаний ключевых понятий, уточняя их в ходе анализа промежуточных результатов поиска [2]. В итоге в распоряжении будет большой объем данных (тысячи документов), в той или иной степени релевантных сформулированным запросам [2]. При этом все найденные документы подробно не рассматриваются (в большинстве случаев просматривается не более 2–3 страниц результатов поиска) [2].

Эта проблема особенно актуальна в образовательной и научной деятельности. Так, в книге «Реорганизация знаний» профессор Ким Вельтман отмечает, что в настоящее время «учёный, занимающийся научной деятельностью, тратит 90% своего времени на поиск документов, 5% на их изучение и всего 5% на науку» [4].

Так как библиотеки являются хранилищами знаний, то именно они могут предоставить сервис для осуществления документального тематического поиска. Качество информации гарантировано за счёт предоставления документов из собственных и партнерских библиотечных фондов [5]. На данном этапе развития информационного общества нельзя далее воспринимать библиотеки просто как хранилища книг, журналов, карточных каталогов [6].

НЫНЕШНЕЕ СОСТОЯНИЕ БИБЛИОТЕК

В традиционной схеме работы библиотеки функцию документального поиска выполняют библиографы. По словам О. П. Коршунова «возникает потребность в специальных посредниках между документами и потребителями, содействующих наиболее рациональному и эффективному использованию накопленных обществом

гигантских документных ресурсов» [7]. Коршунов отмечает, что к числу таких посредников относится библиография [7].

С распространением информационных технологий происходят качественные изменения форм, методов библиотечно-информационных услуг, способов коммуникации между библиотекарем и пользователем [8]. На данный момент библиотеки переживают период информатизации, в области которой основным направлением их работы является комплексная автоматизация всех библиотечных процессов [6]. Под автоматизацией библиотеки, по определению профессора Я. Л. Шрайберга, следует понимать все процессы, связанные с установкой компьютеров на рабочие столы сотрудников и читательские места и позволяющие поэтапно освободить сотрудников библиотеки от рутинной работы, а читателям — создать эффективный и комфортный сервис в поиске и получении изданий из фондов [9]. Для решения этих задач применяются системы автоматизации библиотек (САБ), например, ИРБИС, МегаПро, МАРК-SQL и др.

Библиографическое обслуживание автоматизируется созданием электронной библиотеки, с одной стороны, и появлением виртуальных форм этого обслуживания — с другой.

Электронная библиотека предоставляет читателям доступ к полнотекстовым электронным документам и информации о них. В этом случае имеет место быть поисковая система, включающая в себя несколько видов поиска в зависимости от используемой САБ: простой поиск, расширенный поиск, поиск по словарям и др. Работа с такой поисковой системой, как было сказано выше, может быть кропотливой.

Виртуальная справочная служба представляет собой сервис, позволяющий читателю общаться с библиографом в удаленном режиме. Основная функция службы — выполнение разовых запросов читателей, в число которых входит запрос подбора литературы по конкретной теме. Однако этот подбор ограничен определенным библиотекой количеством источников и выполняется несколько дней, что является достаточно времязатратным способом получения нужных документов.

Таким образом, количественный рост объёмов библиографической информации обусловливает необходимость перехода к следующему этапу развития библиотек — этапу цифровизации. Главное отличие цифровизации от информатизации состоит в том, что она предполагает использование независимых цифровых систем с аналитическими и

прогностическими функциями [8]. Появляется необходимость в цифровом инструментарии сбора, анализа, обработки и распространения библиографической информации [10]. Цифровые технологии обладают потенциалом для улучшения процессов библиотечной деятельности.

СУЩЕСТВУЮЩИЕ ТЕХНОЛОГИИ, ПОДХОДЯЩИЕ ДЛЯ РЕШЕНИЯ ПРОБЛЕМЫ

Перспективным является внедрение в библиотечную сферу интеллектуальных самообучаемых экспертных систем, работа которых построена на большом массиве полнотекстовых первичных документов [11].

Такая система включает в себя обработку естественного языка и интеллектуальный анализ текстов, основанный на методах машинного обучения [12]. Эти технологии целесообразно использовать совместно для автоматической классификации и обнаружения шаблонов в электронных документах, а также для формирования их семантического представления [12]. Построение структуры данных, которая может представлять документы, и построение классификатора, который можно использовать для определения метки класса документа с высокой точностью, являются ключевыми моментами в классификации текста [12].

Представление (индексирование) документов — один из методов обработки естественного языка, который является важным аспектом в классификации документов и обозначает отображение документа в компактную форму его содержимого [12]. Представление документа состоит в том, что он составлен из совместного набора терминов, имеющих различные закономерности появления [12]. Текстовый документ обычно представляется в виде вектора весов терминов (признаков слов) из набора терминов (словаря), где каждый термин встречается по крайней мере один раз в определенном минимальном количестве документов [12]:

$$\overrightarrow{d_j} = \langle \omega_{1j}, \ldots, \omega_{|T|j} \rangle,$$

где T — словарь, и $0 \leq \omega_{kj} \leq 1$ определяет значимость терма t_k, в документе d_j [13].

Для классификации текста основной проблемой является высокая размерность пространства признаков [12]. От данной размерности напрямую зависит вычислительная сложность различных методов классификации [13]. К тому же, множество признаков

содержит нерелевантные и даже шумовые признаки для задачи классификации текста, которые могут резко снизить точность классификации [12]. По этим двум причинам в таких задачах часто прибегают к сокращению числа используемых термов [13]. Существуют два подхода для уменьшения размерности пространства признаков: выбор признаков (FS) и извлечение признаков (FE) [12].

Основная идея выбора признаков (FS) заключается в выборе подмножества функций из исходных документов [12]. FS выполняется путем сохранения слов с наивысшей ценностью в соответствии с заранее определенной метрикой важности слова [12]. Существует множество показателей оценки признаков, в которое входят: информационный прирост (IG), частота терминов, хи-квадрат и др. [12,13]. На практике применяется подход, при котором частота термина/обратная частота документа (TF-IDF) обычно используется для взвешивания каждого слова в текстовом документе в соответствии с тем, насколько оно уникально [12]. Другими словами, подход TF-IDF учитывает релевантность слов, текстовых документов и конкретных категорий [12].

Извлечение признаков (FE) заключается в применении методов кластеризации термов и латентно-семантического индексирования, в результате которых образуются (извлекаются) новые признаки [13]. При кластеризации признаков происходит объединение в группы термов с высокой попарной семантической близостью, представления этих групп или их центроиды используются в качестве признаков для уменьшения размерности пространства [13].

Ещё одним методом обработки естественного языка является семантический анализ [12]. Семантический анализ — это процесс лингвистического разбора предложений и абзацев на ключевые понятия, глаголы и имена собственные [12]. Используя технологию, основанную на статистике, эти слова затем сравниваются с таксономией (категориями) и группируются в соответствии с релевантностью [12].

Семантический анализ может применяться при автоматической классификации текста наряду с представлением документов или вместе с ним. В процессе анализа исходный текст сопоставляется с базовыми семантическими шаблонами, в результате чего формируются семантические зависимости, связывающие части анализируемого предложения [14]. Синтаксические шаблоны сборки именных групп применяются для извлечения терминов-словосочетаний [14], что может быть использовано при формировании векторного представления документа.

В отличии от представления документов семантический анализ является более сложной технологией в области интеллектуального анализа текста. При правильном внедрении этого подхода будет улучшена классификация и процесс поиска информации [12].

Интеллектуальный анализ текстов включает в себя такие подходы к машинному обучению, как байесовский классификатор, дерево решений, K-ближайших соседей (k-NN), машины опорных векторов (SVM), нейронные сети, метод Rocchio, нечеткая корреляция, генетические алгоритмы и т. д. [12]. Различные исследования показали, что SVM превосходит другие алгоритмы классификации [12].

Машины опорных векторов (SVM) являются одним из методов дискриминационной классификации, основанном на принципе минимизации структурных рисков из теории вычислительного обучения [12]. SVM нуждается как в положительном, так и в отрицательном обучающем наборе, что необычно для других методов классификации [12]. Эти наборы необходимы SVM для поиска поверхности принятия решений — гиперплоскости [12]. Гиперплоскость, у которой минимальное расстояние до ближайших примеров максимально, разделяет пространство признаков на две части: положительные примеры в одной и отрицательные в другой [13]. Признаки документа, которые находятся ближе всего к поверхности принятия решения, называются опорным вектором [12]. Гиперплоскость и опорные векторы изображены на рис.1 [12].

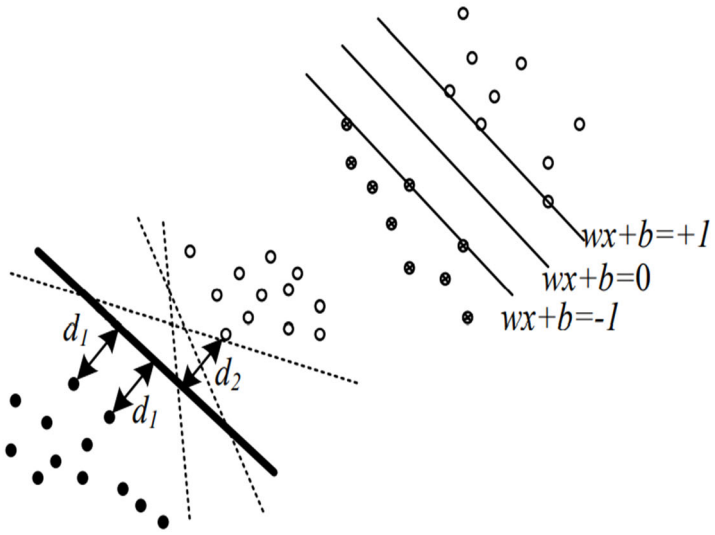

Рисунок 1. Оптимальное разделение гиперплоскости, гиперплоскостей и опорных векторов.

Figure 1. Optimal separation of hyperplane, hyperplanes and support vectors.

SVM может обрабатывать документы с большим пространством ввода и отбрасывать большинство нерелевантных функций [12]. Его способность к обучению может быть независимой от размерности пространства признаков [12]. Основным недостатком SVM является его относительно сложные алгоритмы обучения и категоризации, а также большие затраты времени и памяти на этапе обучения и классификации [12].

Стоит отметить, что в контексте объединения нескольких классификаторов для категоризации текста ряд исследователей показали, что объединение различных классификаторов может повысить точность классификации [12].

СУЩЕСТВУЮЩИЕ СИСТЕМЫ, ПРИМЕНЯЮЩИЕ ЦИФРОВЫЕ ТЕХНОЛОГИИ ДЛЯ ДОКУМЕНТАЛЬНОГО ПОИСКА

На текущий момент в библиотечной области существуют решения, реализующие цифровые технологии. Внимания заслуживают платформа Yewno и модуль Discovery Znanium.

Yewno — инструмент для академических и публичных библиотек [15], позволяющий трансформировать данные в информацию, а информацию — в знания [16]. Инструмент создан одноименной компанией со штаб-квартирой в Редвуд-Сити, штат Калифорния, и офисами в Лондоне и Нью-Йорке [17]. У компании есть многочисленные партнерские отношения с ведущими исследовательскими университетами, издателями, финансовым сектором и агрегаторами контента по всему миру [17].

Yewno использует машинное обучение, когнитивную науку, нейронные сети и компьютерную лингвистику для анализа контента с целью извлечения понятий и выявления закономерностей и взаимосвязей, что позволяет эффективно анализировать большие объемы информации [16,17].

Yewno работает с полным текстом любого ресурса и не зависит от определенных человеком метаданных [16]. Используя технологии искусственного интеллекта, программа автоматически извлекает понятия и различает отношения между ними, дополняя документы метаданными [16]. Yewno позволяет анализировать любое количество текста из разных источников, а также семантически изучать полный текст на основе связывания похожих понятий в разных типах документов [16]. Затем Yewno

представляет графический интерфейс этих отношений, что позволяет исследовать темы в контексте, создавать новые ассоциации между темами и предметами [16].

На рис.2 представлен пример, демонстрирующий технологию Yewno [18]. Yewno начинает с того, что дает исследователю краткий обзор темы в левой части экрана, и графическое резюме всех тем, которые были извлечены из ресурсов по этой теме — в правой [18]. Статьи, используемые для раскрытия резюме, доступны для прочтения [16].

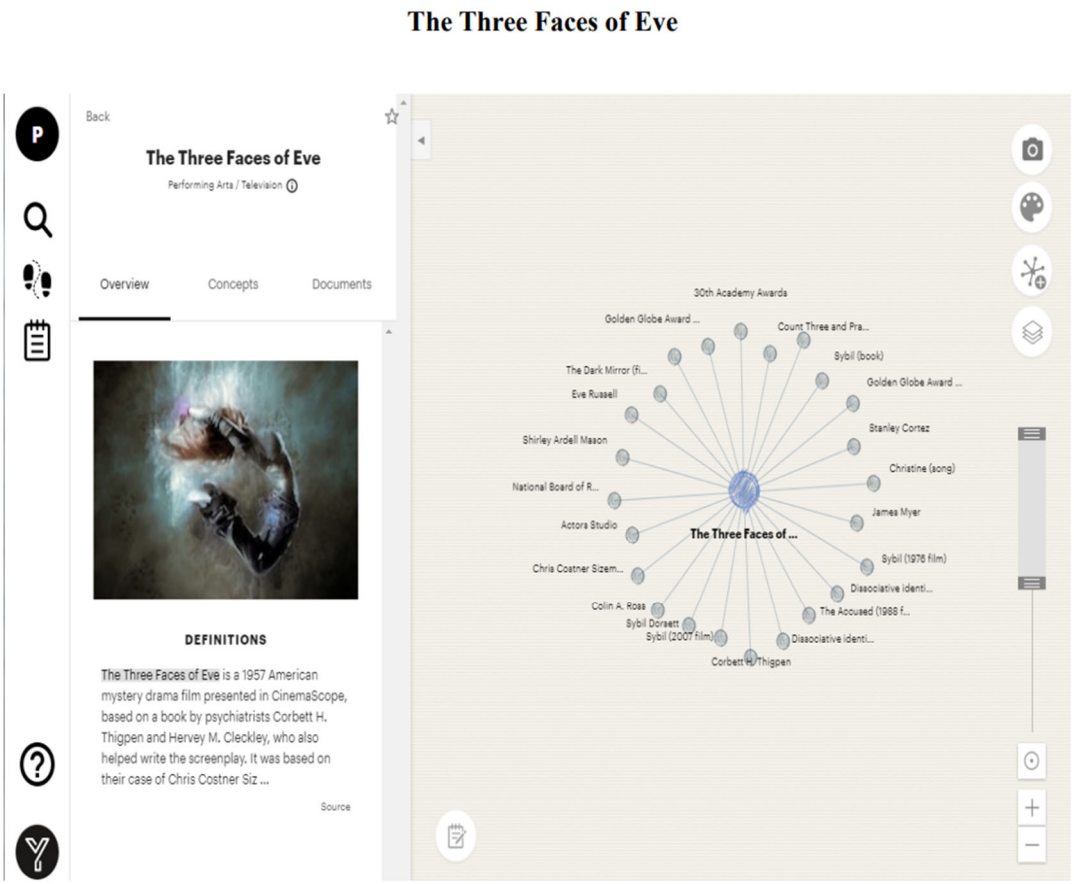

Рисунок 2. Исследование фильма 1957 года «Три лица Евы» в Yewno.

Figure 2. The study of the 1957 film "Three Faces of Eve" in Yewno.

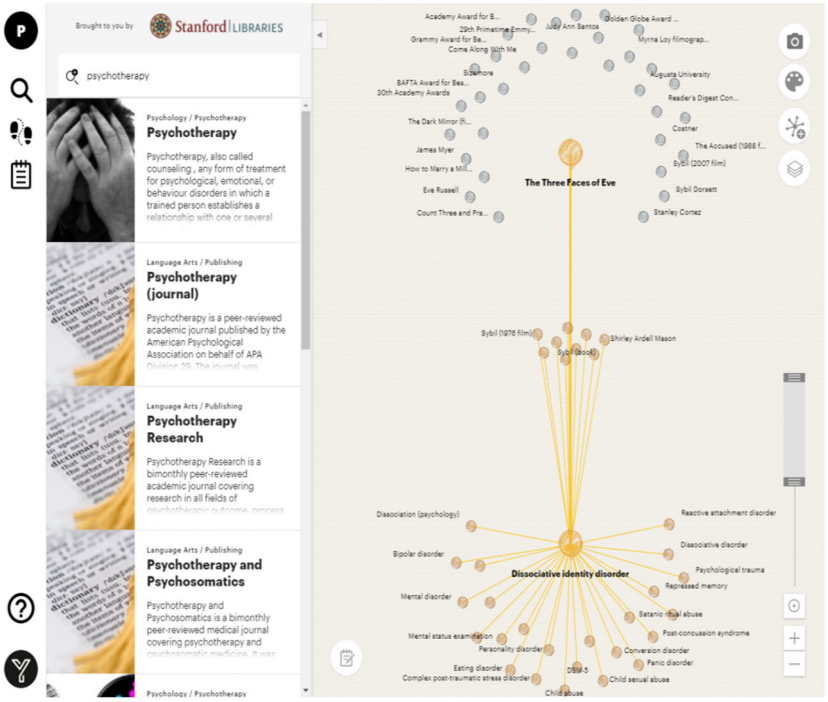

Рисунок 3. Более глубокий уровень предметного анализа фильма «Три лица Евы» в Yewno.

Figure 3. A deeper level of subject analysis of the film "Three Faces of Eve" in Yewno.

Как показано на рис.3, при расширении полосы в правом нижнем углу экрана, осуществляется переход к более глубокому уровню предметного анализа [18]. У исследователя также есть возможность добавить вторую тему исследования, посмотреть, как эти две темы связаны друг с другом, и экспортировать отчет в необходимом формате [16].

Важной особенностью Yewno является возможность сопоставления базы данных библиотеки с базой знаний Yewno, при этом виджеты Yewno могут быть предоставлены для работы с другими ресурсами учреждения [15].

Таким образом, Yewno — действительно эффективный инструмент для тематического поиска и анализа предметной области, который возможно интегрировать в конкретную библиотеку. Тем не менее данный продукт может быть недоступен для русскоязычной части пользователей по двум причинам. Первая причина — большинство ресурсов Yewno представлены на английском языке. Вторая причина заключается в принадлежности продукта американской компании, которая в качестве санкций может запретить его пользование русским читателям.

Ещё одним недостатком системы является сложность использования. Тематические графы могут быть громоздкими и трудными для восприятия. При углублении в конкретную тему программа предоставляет большое количество информации, в которой не всегда легко ориентироваться.

Теперь рассмотрим сервис от отечественной электронно-библиотечной системы (ЭБС) Znanium. Модуль Discovery Znanium — совместная разработка ИНФРА-М и Института системного анализа РАН [19], в основе которой лежат идеи и технологии системы Exactus [20].

К таким технологиям относятся: алгоритм построения инвертированного поискового индекса (ИПИ) коллекций документов и алгоритм поиска информации по запросу, основанный на методе многокритериальной оценки сходства текстов [21]. Разработке этих алгоритмов посвящена диссертация Соченкова И. В.

В задаче многокритериальной оценки сходства текстов текст описывается множествами различных признаков:

- внетекстовых (теговая разметка, веса вхождений слов);
- лексико-морфологических (нормальные формы лексем, формы словоупотреблений);
- синтаксических (словосочетания, представленные именными и глагольными группами);
- семантических (значения синтаксем и семантические связи) [21].

При выполнении поиска запрос пользователя на естественном языке рассматривается в качестве эталонного текста [21]. С применением методов лингвистического анализа к поисковому запросу формируется представление текста запроса [21]. По нормальным формам словоупотреблений выполняется выборка данных из специальной структуры данных, содержащей информацию о словоупотреблениях в текстах коллекции документов, в которой осуществляется поиск [21]. Эта структура данных содержит информацию о текстах документов в соответствии с описанным выше представлением текста [21].

Предложенное представление текстовой информации значительно отличается от наиболее распространённого в системах информационного поиска представления текстовой информации в виде векторов признаков [21]. Это означает, что классические алгоритмы работы с текстами, как с векторами признаков, и известные методы

индексирования текстовой информации не применимы для реализации метода оценки сходства текстов [21]. Поэтому данный метод требует разработки специализированных структур данных и алгоритмов для эффективного представления и обработки данных [21].

Для реализации метода были разработаны ИПИ, отличительной особенностью которого является возможность поиска с учётом синтаксической и семантической информации, а также структуры данных, необходимые для представления информации о текстовых документах [21].

Разработанный метод оценки сходства текстов позволяет проводить многокритериальное сопоставление текстов с учётом различных лингвистических признаков [21]. Критерии оценки сходства предложений текстов опираются на лексико-морфологическую информацию, а также синтаксическую и семантические структуры предложений [21].

Целью создания Discovery Znanium является предоставление исследователям возможности находить контент по интересующей их теме или дисциплине за минимальное время [22]. Этому способствуют интеллектуальный поиск, основанный на алгоритмах искусственного интеллекта, и большой объём контента: в Discovery свыше 14,5 млн документов [22].

Модуль включает в себя:

- поиск документов в выбранных репозиториях;
- поиск заимствований в проверяемом тексте;
- поиск документов, тематически похожих на заданный текст или файл;
- анализ публикационной активности по заданной теме в заданном диапазоне времени;
- анализ качества научного уровня проверяемого текста [19].

При интеллектуальном поиске пользователь может читать документы в репозиториях по указанным адресам, а также получить информацию о документе [19]. Для каждого документа автоматически формируются резюме и облако ключевых слов, которые можно масштабировать. Резюме документа представлено на рис. 4, облако ключевых слов — рис.5.

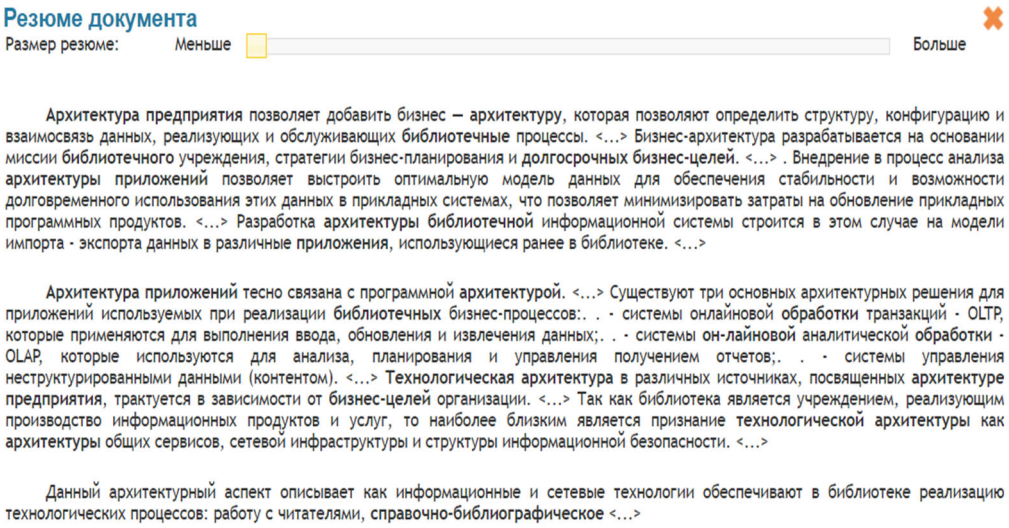

Рисунок 4. Резюме документа.

Figure 4. Synopsis of the document.

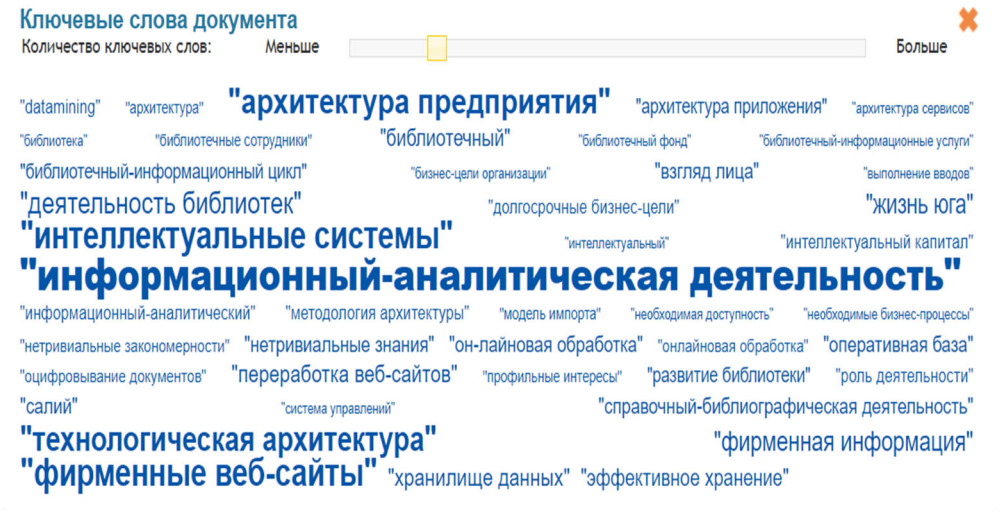

Рисунок 5. Облако ключевых слов документа.

Figure 5. Document keyword cloud.

У пользователя есть возможность получить HTML-копию документа с подсветкой по тексту слов поискового запроса и найти документы, содержательно похожие на заданный документ [19]. При этом для каждого объекта в результатах поиска будет указываться процент сходства с исходным документом, как показано на рис.6.

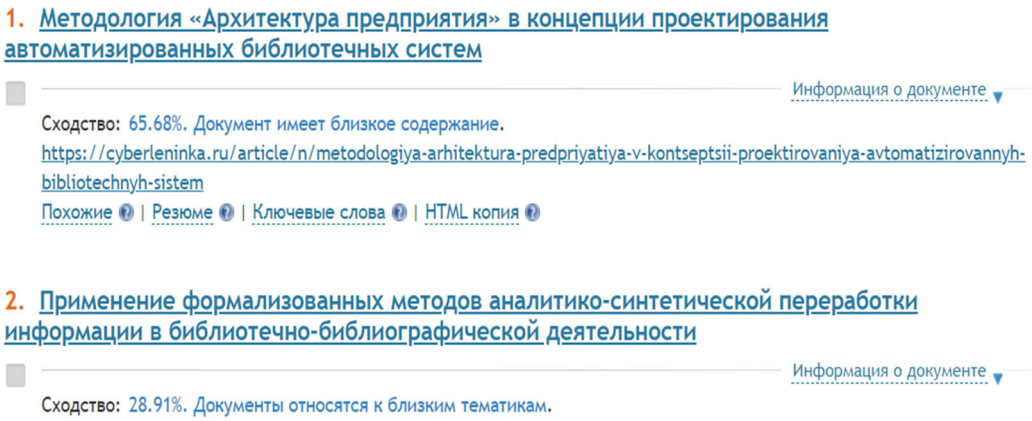

Рисунок 6. Поиск документов, тематически похожих на заданный документ.

Figure 6. Search for documents that are thematically similar to a given document.

После исследования данного сервиса можно сделать следующие выводы. Discovery Znanium на самой деле облегчает тематический поиск документов. Резюме документа позволяет быстро определить его полезность для пользователя, что значительно сокращает время знакомства с результатами поиска. Для каждого документа с помощью интеллектуального алгоритма выделены ключевые слова, которые могут быть не очевидны. Они позволяют пользователю найти дополнительные материалы, а также документы по своей тематике в смежных областях. Найдя подходящий документ, пользователь может найти семантически похожие документы, что также автоматизирует тематический поиск.

Однако у этого модуля есть существенный недостаток. В первую очередь он ориентирован на учёных. Поиск в данной системе производится по научным публикациям. Для того, чтобы эффективно пользоваться интеллектуальным (в т.ч. семантическим) поиском, нужно владеть языком запросов, а также уметь пользоваться различными фильтрами. Следовательно, модуль полезен только узкой категории людей и не ориентирован на массового пользователя. Для субъектов образовательной сферы, например для студентов и преподавателей, модуль будет малополезен.

Исходя из рассмотрения инструмента Yewno и модуля Discovery Znanium можно сделать вывод, что проблема тематического поиска для массового пользователя остается нерешённой.

РЕШЕНИЕ ПОСТАВЛЕННОЙ ПРОБЛЕМЫ

Рассмотрим схему традиционного библиотечного взаимодействия: читатель – библиотека – библиотекарь – результат [23]. Читатель приходит со своим запросом в библиотеку, обращается к библиотекарю, получает результат (выполненная услуга, информационный ресурс — печатный и/или электронный) [23].

При выполнении тематического запроса от читателя особую важность приобретает первый этап — прием запроса [24]. Разговорить читателя, помочь ему сформулировать тему — непростая задача, которую ежедневно решают библиографы [24]. Кроме точно сформулированной темы, необходимо выяснить степень общеобразовательной и специальной подготовки читателя (студент, аспирант и т.д.), цель запроса (курсовая работа, диссертация и т.д.), установить, какие виды документов интересуют читателя (книги, статьи из периодических изданий и т.д.), хронологические рамки и язык представления информации [24].

В условиях модернизации библиотечных процессов схема взаимодействия библиотеки и читателя меняется: пользователь – проактивная библиотека – результат (электронные информационные ресурсы, выполненная услуга) [23]. Пользователь приходит в виртуальное пространство библиотеки со своим запросом, самостоятельно находит необходимый ему ресурс или обращается к цифровому сервису библиотеки и с его помощью получает необходимый результат [23].

В этой системе электронная библиотека получает сведения о пользователе из множества доступных источников, анализирует его персональные данные, историю поиска, запросы и предлагает информацию с опережением [23]. Информация носит персонифицированный характер и является отражением индивидуально-личностных характеристик пользователя [23].

Исходя из вышесказанного можно сформулировать 2 способа решения проблемы тематического поиска.

Первый способ заключается в обработке запроса читателя с учётом его индивидуальных характеристик, таких как уровень подготовки, специальность, история поиска и др. В таком случае необходимо предусмотреть методы хранения и использования личных данных пользователя.

Второй способ предполагает создание системы, выполняющей функции библиографа в рамках справочно-библиографического обслуживания. Такое

обслуживание состоит в приёме запроса потребителя на информацию и выдаче ему ответа на запрос [7]. Основной задачей системы является уточнение запроса пользователя в удобном для него формате. В ходе приёма запроса система должна выяснить у читателя его цели, мотивы, интересы и т.д. В основе системы может лежать алгоритм выполнения тематического запроса библиографом, изображенный на рисунке 7 [24].

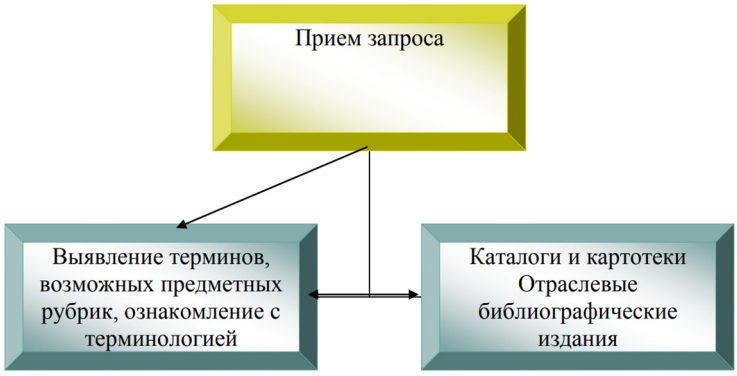

Рисунок 7. Алгоритм выполнения тематического запроса библиографом.

Figure 7. The algorithm for performing a thematic query by a bibliographer.

ЗАКЛЮЧЕНИЕ

В статье рассмотрена проблема документального тематического поиска. Актуальность проблемы обосновывается всё увеличивающимся объёмом цифровых данных и неспособностью поисковых систем предоставить эффективный инструмент получения релевантной тематической информации. Библиотеки могут предоставить сервис поиска качественной тематической информации, так как содержат большой объём документов.

В статье дан обзор существующих технологий для решения проблемы, а именно самообучаемых экспертных систем на основе машинного обучения. Были найдены и проанализированы существующие продукты, реализующие подобные системы. Эти продукты действительно упрощают и автоматизируют документальный тематический поиск, однако не являются полноценным решением данной проблемы.

Результатом статьи являются предложенные способы решения поставленной проблемы. К ним относятся:

- создание алгоритма обработки запроса читателя с учётом его индивидуальных характеристик;
- создание системы, выполняющей функции библиографа в рамках справочно-библиографического обслуживания.

Упомянутые способы решения проблемы основываются на интерпретации традиционной библиотечной (в т.ч. библиографической) деятельности в цифровую сферу. По этой причине было исследовано текущее состояние развития библиотек и классическая схема библиотечного обслуживания.

Также важно, чтобы решение проблемы было предоставлено читателям в комфортной для них среде. В данном случае целесообразно использовать мобильные технологии. Среди тенденций преобразования глобальной информационной экономики ИФЛА называет переход к мобильному устройству как основному средству доступа к информации, следовательно, можно сделать вывод, что мобильные технологии являются перспективным инструментом предоставления доступа к информации [25].

Автором данной статьи было произведено исследование ЭБС СПбГУТ и обнаружены следующие проблемы: отсутствие мобильного приложения, трудоёмкий поиск, рассмотренный в этой статье.

В ЭБС СПбГУТ установлена САБ J-ИРБИС 2.0, которая предоставляет API для удалённого обращения к базам данных. API позволяет в авторизованном режиме выполнять поиск и расформатирование записей из БД по протоколу JSON-RPC 2.0 [26].

На данный момент времени идет разработка мобильного приложения с использованием упомянутого API. Следующим шагом после создания приложения является автоматизация тематического поиска.

СПИСОК ЛИТЕРАТУРЫ

[1] Юмашева Ю.Ю. Историческая наука, архивы, библиотеки, музеи и искусственный интеллект: что день грядущий нам готовит? Документ. Архив. История. Современность. 2021; 21: 247-279.

[2] Иванов В.К., Мескин П.И. Реализация генетического алгоритма для эффективного документального тематического поиска. Программные продукты и системы. 2014; 4(108): 118-126. https://doi.org/10.15827/0236-235X.108.118-126

[3] Костин В.В. К вопросу создания системы поддержки работы с научными

публикациями. Вестник НГУ. Серия: Информационные технологии. 2014; 4: 32-37.

[4] Арентова Т.Н. От доступа к информации – к пространству знаний и цифровой культуре: новые реалии для библиотек. Цифровое пространство библиотеки как среда интеллектуального развития личности: материалы межрегиональной науч.-практ. конф. 22–23 мая 2019 года. М-во культуры и архивного дела Сахалин. обл., Сахалин. обл. универс. науч. б-ка; ред.-сост. Д. А. Ускова. Южно-Сахалинск; 2019. 9-17.

[5] Носков М.В., Шершнева В.А., Барышев Р.А., Манушкина М.М. Информатизация образования в вузе: Актуальные вопросы развития электронных библиотек. Вестник ТГПУ. 2016; 1(166): 151-155.

[6] Кутуева, Б.К. Информатизация вузовских библиотек Кыргызстана. Рыскулбеков атындагы Кыргыз экономикалык университетинин кабарлары. Москва; 2018. 3(45); 78-80.

[7] Коршунов О.П. Библиографоведение. Общий курс. Основы теории библиографии: учебник. МГУ культуры и искусств. Москва; 2000. 149.

[8] Нещерет М.Ю. Цифровизация процессов обслуживания в библиотеках — это уже реальность. Библиосфера. 2019; 2: 19-25.

[9] Исаева С.К. Автоматизация и информатизация библиотеки им. Д.Р. Новикова. Эпоха на книжных страницах: Сборник статей (к 180-летию Белорусской государственной сельскохозяйственной академии и библиотеки им. Д.Р. Новикова). Под редакцией А.И. Малько. Белорусская государственная сельскохозяйственная академия. Горки; 2020. 16-20.

[10] Нещерет М.Ю. Цифровая библиография: библиотеки в поисках инновационных инструментов библиографической деятельности. Научные и технические библиотеки. 2021; 7: 33-50. https://doi.org/10.33186/1027-3689-2021-7-33-50

[11] Тимошенко И.В. Искусственный интеллект в библиотечных технологиях. Уже пора? Румянцевские чтения - 2019: Материалы Международной научно-практической конференции: в 3 частях, Москва, 23–24 апреля 2019 года. Издательство "Пашков дом". Москва; 2019. 153-157.

[12] Дементьев В.Е., Киреев С.Х. Выбор алгоритмов машинного обучения для классификации текстовых документов. Техника средств связи. 2022; 2(158): 22-52.

[13] Епрев А.С. Автоматическая классификация текстовых документов. МСиМ. 2010; 1(21): 65-81.

[14] Мочалова А.В. Алгоритм семантического анализа текста, основанный на базовых семантических шаблонах с удалением. Научно-технический вестник информационных технологий, механики и оптики. 2014; 5(93): 126-132.

[15] Yewno Discover: Annual Commitment 2020-2022 [электронный ресурс] URL: https://subscriptionsmanager.jisc.ac.uk/catalogue/2408 (дата обращения 15.11.2022).

[16] Редькина Н.С. Векторы развития научных библиотек: обзор ключевых докладов Всемирного конгресса ИФЛА 2019 г. Библиосфера. 2020; 2: 71-81. https://doi.org/10.20913/1815-3186-2020-2-71-81

[17] Yewno - Silverchair [электронный ресурс] URL: https://www.silverchair.com/the-silverchair-platform/silverchair-universe/yewno/ (дата обращения 15.11.2022).

[18] Schreur PhE. Yewno: Transforming Data into Information, Transforming Information into Knowledge. Paper presented at: IFLA WLIC 2019 - Athens, Greece - Libraries: dialogue for change in Session 114 - Knowledge Management with Information Technology and Big Data.

[19] Электронно-библиотечная система Znanium Использование модуля Discovery Znanium [Электронный ресурс]: руководство читателя. 2020. 25. URL: https://znanium.com/help/reader-discovery (дата обращения: 06.11.2022).

[20] TextAppliance [Электронный ресурс] URL: https://www.textapp.ru/ (дата обращения 07.11.2022).

[21] Соченков И. В. Реляционно-ситуационные структуры данных, методы и алгоритмы решения поисково-аналитических задач: дис. ... канд. физ.-мат. наук: 05.13.17 / Соченков Илья Владимирович. Москва; 2014. 148.

[22] «Znanium безвозмездно»: цивилизованный подход к ресурсам и технологиям открытого доступа [Электронный ресурс] URL: http://www.unkniga.ru/biblioteki/vuzbiblio/12817-znanium-bezvozmezdno-tsivilizovanniy-podhod-k-resursam-itehnologiyam.html (дата обращения 06.11.2022).

[23] Барышев Р.А., Цветочкина И.А., Касянчук Е.Н., Бабина О.И. Модернизация процесса обслуживания пользователей университетских библиотек. Научные и технические библиотеки. 2022; 3: 39-60. https://doi.org/10.33186/1027-3689-2022-3-39-60

[24] Свирюкова В.Г. Организация и методика справочно-библиографи-чес-кого обслуживания: конспект лекций /отв. ред. Е. Б. Артемьева; Гос. публич. науч.-техн. б-ка Сиб. отд-ния Рос. акад. наук. – 2-е изд., испр. и доп. ГПНТБ СО РАН. Новосибирск; 2007.

64.

[25] Михайлова Е.В. Мобильные технологии в современной библиотеке: выбираем лучшее. Библиотеки вузов Урала: проблемы и опыт работы: По материалам научно-практической конференции: Научно-практический сборник, Екатеринбург, 30 сентября – 01 октября 2014 года. Ответственный редактор Г. Ю. Кудряшова; научный редактор Г. С. Щербинина. Уральский федеральный университет имени первого Президента России Б.Н. Ельцина. Екатеринбург; 2014. 75-80.

[26] ИРБИС: J-ИРБИС: Развитие J-ИРБИС 2.0 [электронный ресурс] URL: http://irbis.elnit.org/read.php?43,90737 (дата обращения 21.11.2022).

REFERENCES

[1] Yumasheva Yu.Yu. Historical Science, Archives, Libraries, Museums and Artificial Intelligence: What Does Tomorrow Hold? Dokument. Arkhiv. Istoriya. Sovremennost. 2021; 21, 247-279 (in Russian).

[2] Ivanov V.K., Meskin P.I. Realizatsiya geneticheskogo algoritma dlya effektivnogo dokumentalnogo tematicheskogo poiska. Programmnyye produkty i sistemy. [Implementation of a genetic algorithm for efficient documentary thematic search. Software products and systems]. 2014; 4(108): 118-126. (in Russian). https://doi.org/10.15827/0236-235X.108.118-126

[3] Kostin V.V. K voprosu sozdaniya sistemy podderzhki raboty s nauchnymi publikatsiyami Vestnik NGU. Seriya: Informatsionnyye tekhnologii. [On the issue of creating a support system for working with scientific publications. Bulletin of the Novosibirsk State University. Series: Information Technology]. 2014; 12(4): 32-37. (in Russian).

[4] Arentova T.N. Ot dostupa k informatsii – k prostranstvu znaniy i tsifrovoy kulture: novyye realii dlya bibliotek. Tsifrovoye prostranstvo biblioteki kak sreda intellektualnogo razvitiya lichnosti : materialy mezhregionalnoy nauch.-prakt. konf. 22–23 maya 2019 goda. [From access to information to the space of knowledge and digital culture: new realities for libraries. Digital space of the library as an environment for the intellectual development of the individual: materials of the interregional scientific and practical. conf. May 22–23, 2019] M-vo kultury i arkhivnogo dela Sakhalin. obl., Sakhalin. obl. univers. nauch. b-ka; red.-sost. D. A. Uskova. Yuzhno-Sakhalinsk; 2019. 9-17. (in Russian).

[5] Noskov M.V., Shershneva V.A., Baryshev R.A., Manushkina M.M. Informatizatsiya obrazovaniya v vuze: Aktualnyye voprosy razvitiya elektronnykh bibliotek. [Informatization of

education at the university: Topical issues of the development of electronic libraries]. Vestnik TGPU. 2016; 1(166): 151-155. (in Russian).

[6] Kutuyeva B.K. Informatizatsiya vuzovskikh bibliotek Kyrgyzstana. [Informatization of university libraries in Kyrgyzstan]. M. Ryskulbekov atyndagy Kyrgyz ekonomikalyk universitetinin kabarlary. 2018; 3(45): 78-80. (in Russian).

[7] Korshunov O.P. Bibliografovedeniye. Obshchiy kurs. Osnovy teorii bibliografii: uchebnik. [Bibliography. General course. Fundamentals of the theory of bibliography: textbook] Moskva. MGU kultury i iskusstv. Moskva; 2000. 149. (in Russian).

[8] Neshcheret M.Yu. Tsifrovizatsiya protsessov obsluzhivaniya v bibliotekakh - eto uzhe realnost. [Digitalization of service processes in libraries is already a reality] Bibliosfera. 2019; 2: 19-25. (in Russian).

[9] Isayeva S.K. Avtomatizatsiya i informatizatsiya biblioteki im. D.R. Novikova. Epokha na knizhnykh stranitsakh: Sbornik statey (k 180-letiyu Belorusskoy gosudarstvennoy selskokhozyaystvennoy akademii i biblioteki im. D.R. Novikova). [Automation and informatization of the library. D.R. Novikov. Epoch on book pages: Collection of articles (on the occasion of the 180th anniversary of the Belarusian State Agricultural Academy and Library named after D.R. Novikov)] Pod redaktsiyey A.I. Malko. Belorusskaya gosudarstvennaya selskokhozyaystvennaya akademiya. Gorki; 2020. 16-20 (in Russian).

[10] Neshcheret M.Yu. Tsifrovaya bibliografiya: biblioteki v poiskakh innovatsionnykh instrumentov bibliograficheskoy deyatelnosti. [Digital Bibliography: Libraries in Search of Innovative Bibliographic Tools] Nauchnyye i tekhnicheskiye biblioteki. 2021; 7: 33-50. (in Russian). https://doi.org/10.33186/1027-3689-2021-7-33-50

[11] Timoshenko I.V. Artificial Intelligence in Library Technologies. Is it time yet? The Rumyantsev readings — 2019: Proceedings International Scientific and Practical Conference (April 23–24, 2019) Part 3. Pashkov Dom. Moscow; 2019. 153-157. (in Russian).

[12] Dementyev V.Ye., Kireyev S.Kh. Vybor algoritmov mashinnogo obucheniya dlya klassifikatsii tekstovykh dokumentov. [Selection of machine learning algorithms for classifying text documents] Tekhnika sredstv svyazi. 2022; 2(158): 22-52. (in Russian).

[13] Yeprev A.S. Avtomaticheskaya klassifikatsiya tekstovykh dokumentov. [Automatic classification of text documents] MSiM. 2010; 1(21): 65-81. (in Russian).

[14] Mochalova A.V. Algorithm for semantic text analysis by means of basic semantic templates with deletion. Scientific and Technical Journal of Information Technologies,

Mechanics and Optics. 2014; 5(93): 126-132. (in Russian).

[15] Yewno Discover: Annual Commitment 2020-2022 Available: https://subscriptionsmanager.jisc.ac.uk/catalogue/2408 (Accessed 15.11.2022).

[16] Redkina N.S. Development vectors for research libraries: the review of the key reports at the IFLA World Library and Information Congress 2019. Bibliosphere. 2020; 2: 71-81. (in Russian). https://doi.org/10.20913/1815-3186-2020-2-71-81

[17] Yewno - Silverchair Available: https://www.silverchair.com/the-silverchair-platform/silverchair-universe/yewno/ (Accessed 15.11.2022).

[18] Schreur PhE. Yewno: Transforming Data into Information, Transforming Information into Knowledge. Paper presented at: IFLA WLIC 2019 - Athens, Greece - Libraries: dialogue for change in Session 114 - Knowledge Management with Information Technology and Big Data.

[19] Elektronno-bibliotechnaya sistema Znanium Ispolzovaniye modulya Discovery Znanium: rukovodstvo chitatelya. [Znanium Digital Library System Using the Znanium Discovery Module: A Reader's Guide]. 2020; 25. Available: https://znanium.com/help/reader-discovery (Accessed 06.11.2022) (in Russian).

[20] TextAppliance Available: https://www.textapp.ru/ (Accessed 07.11.2022) (in Russian).

[21] Sochenkov I.V. Relyatsionno-situatsionnyye struktury dannykh, metody i algoritmy resheniya poiskovo-analiticheskikh zadach: dis. ... kand. fiz. -mat. nauk: 05.13.17 [Relational-situational data structures, methods and algorithms for solving search-analytical problems: dis. ... cand. Phys.-Math. Sciences: 05.13.17] Sochenkov Ilya Vladimirovich. Moskva; 2014. 148. (in Russian).

[22] "Znanium for free": a civilized approach to open access resources and technologies. Available: http://www.unkniga.ru/biblioteki/vuzbiblio/12817-znanium-bezvozmezdno-tsivilizovanniy-podhod-k-resursam-itehnologiyam.html (Accessed 06.11.2022) (in Russian).

[23] Baryshev R.A., Tsvetochkina I.A., Kasyanchuk Ye.N., Babina O.I. Modernization of user services at academic libraries. Scientific and technical libraries. 2022; 3: 39-60. (in Russian). https://doi.org/10.33186/1027-3689-2022-3-39-60

[24] Sviryukova V.G. Organizatsiya i metodika spravochno-bibliografi-ches-kogo obsluzhivaniya: konspekt lektsiy. [Organization and methodology of reference and bibliographic services: lecture notes] otv. red. Ye. B. Artemyeva; Gos. publich. nauch. -tekhn. b-ka Sib. otd-niya Ros. akad. nauk. – 2-ye izd., ispr. i dop. GPNTB SO RAN. Novosibirsk;

2007. 64. (in Russian).

[25] Mikhaylova Ye.V. Mobilnyye tekhnologii v sovremennoy biblioteke: vybirayem luchsheye. Biblioteki vuzov Urala: problemy i opyt raboty: Po materialam nauchno-prakticheskoy konferentsii: Nauchno-prakticheskiy sbornik, Yekaterinburg, 30 sentyabrya – 01 oktyabrya 2014 goda. Otvetstvennyy redaktor G. Yu. Kudryashova; nauchnyy redaktor G. S. Shcherbinina. [Mobile technologies in the modern library: choosing the best. Libraries of universities of the Urals: problems and work experience: Based on the materials of the scientific and practical conference: Scientific and practical collection, Yekaterinburg, September 30 - October 01, 2014 / Editor-in-chief G. Yu. Kudryashova; scientific editor G. S. Shcherbinina]. Uralskiy federalnyy universitet imeni pervogo Prezidenta Rossii B.N. Yeltsina. Yekaterinburg; 2014. 75-80 (in Russian).

[26] IRBIS: J-IRBIS: Razvitiye J-IRBIS 2.0 [IRBIS: J-IRBIS: Development of J-IRBIS 2.0] Available: http://irbis.elnit.org/read.php?43,90737 (Accessed 21.11.2022) (in Russian).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Юрченко Павел Вячеславович, кафедра безопасности информационных систем, Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича, Санкт-Петербург, Россия
e-mail: pactmon9@mail.ru

Pavel Yurchenko, department of information systems security, The Bonch-Bruevich Saint-Petersburg State University of Telecommunication, Saint-Petersburg, Russia

e-mail: pactmon9@mail.ru

Статья поступила в редакцию 25.01.2023; одобрена после рецензирования 13.02.2023; принята к публикации 14.02.2023.

The article was submitted 25.01.2023; approved after reviewing 13.02.2023; accepted for publication 14.02.2023.