УДК 001.81:004.91]-057.875

Профессиональный поиск научной литературы для исследовательских работ магистрантов

Н. Н. Дацун
Пермский государственный национальный исследовательский университет,
nndatsun@inbox.ru, OrcID: 0000-0001-8560-7036, SPIN-код: 3896-4544

Аннотация
В статье рассматривается проблема повышения эффективности исследовательской работы ИКТ-магистрантов при профессиональном поиске научной литературы с помощью методологии систематического картографического исследования. Выявлены ограничения цифровизации процесса по этой методологии. Сформулирована постановка задачи цифровизации этого процесса. Представлена цифровизация процесса в системе BibReader для решения проблемы эффективности исследования. Применение рассмотренной методологии является перспективным направлением при поиске научной литературы для студентов других образовательных программ.

Введение

Научная работа магистранта по образовательным программам магистратуры в области информационных и компьютерных технологий (ИКТ) должна начинаться с анализа состояния исследований предметной области, используемых подходов, методов и алгоритмов, применяемых архитектур, инструментальных и программных средств, фреймворков, библиотек и т. д. Результат такого «разведочного анализа» позволяет не только обнаружить пробелы, разрывы, ограничения в существующих методологиях и выявить системы-аналоги, но и уточнить формулировку требований к работе, а также детализировать постановку задачи.

Магистрантам необходимо усовершенствовать свои навыки профессионального поиска и анализа научных публикаций в процессе подготовки первого (аналитического) раздела своей выпускной квалификационной работы (ВКР).

В домене программной инженерии рекомендована к использованию методология систематического обзора литературы (Systematic Literature Review, SLR) [1] и ее «легковесный» вариант для обучающихся – систематическое картографическое исследование (Systematic Mapping Study, SMS).

В эпоху цифровизации актуальные результаты научных исследований доступны в профессиональных источниках: цифровых (электронных) библиотеках, реферативных базах данных, сайтах научных издательств и т.п. Таким образом, этап оцифровки этих результатов («digitization») уже выполнен. В данной статье фокус исследования сосредоточен на цифровизации (digitalization, дигитализации), т.е. использовании оцифрованной информации и цифровых технологий для внесения изменений в процессе поиска научной литературы.

Анализ динамики проведения SMS

Результаты SMS представлены двумя видами публикаций. Первый из них – это диссертации, защищенные в университетах [2]. Второй – это научные публикации в журналах [3] или материалах научных мероприятий [4]. Для изучения публикационной активности в области SMS был выполнен поиск документов в ACM Digital Library (ACM DL) [5], IEEE Xplore Digital Library (IEEE Xplore DL) [6], ScienceDirect [7] и SpringerLink [8]. По поисковому запросу (НАЗВАНИЕ ДОКУМЕНТА = "Systematic Mapping Study" AND 2007<= ГОД <= 2024) было найдено 772 документа, из них 749 уникальных. Распределение публикаций о проведении SMS по годам показывает тенденцию роста интереса к рассматриваемой нами методологии проведения профессионального поиска литературы (рис. 1).

Рисунок 1 – Публикационная активность по теме SMS

Цель и задачи статьи

Объектом исследования этой статьи является профессиональный поиск литературы для проведения научного исследования в области ИКТ по методологии систематического картографического исследования. Предмет исследования – выявление проблем цифровизации этапов SMS, сравнение сервисов/систем автоматизации выполнения этих этапов.

Цель работы – проанализировать существующие методы и средства цифровизации профессионального поиска научной литературы при подготовке магистрантами аналитического раздела выпускной квалификационной работы.

Задачи статьи – проанализировать цифровые средства поддержки выполнения этапов методологии SMS, выявить ограничения цифровизации процесса SMS на основе опыта выполнения SMS студентами-магистрантами, предложить решение снятия этих ограничений.

Методология SMS

Методология систематического картографического исследования предусматривает выполнение пяти последовательных этапов (рис. 2). Результаты текущего этапа SMS являются исходными данными для следующего.

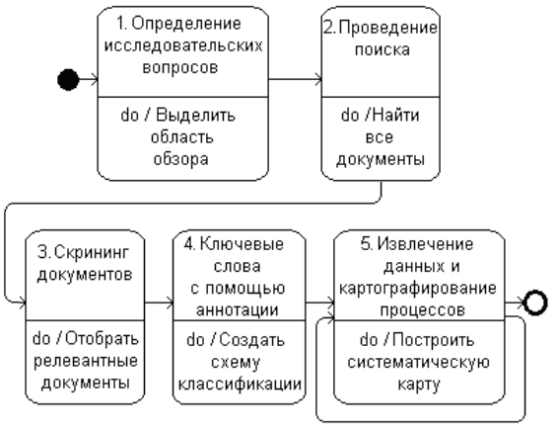

Рисунок 2 – Модель SMS в виде диаграммы состояний языка UML [9, с. 3]

Далее рассматриваются этапы SMS в контексте их сравнения с SLR, выделяются проблемы, с которыми могут столкнуться студенты при их выполнении, а также анализируются цифровые технологии для реализации соответствующих этапов исследования.

Этап 1: определение исследовательских вопросов

В отчете [1] выделены основные различия между SLR и SMS для каждого из этапов методологии. В SMS на первом этапе часто ставится множество исследовательских вопросов, потому что в его основе лежат более широкие исследовательские задачи.

Магистрантам при работе над SMS рекомендуется сформулировать не менее трех исследовательских вопросов (ИВ) в соответствии с разделами ВКР. «ИВ1: Как выглядит ландшафт исследований по теме ВКР?» ориентирован на определение текущего состояния в этом домене. Ответ на «ИВ2: Какие подходы и методы используются для решения задач…?» помогает в формализации и оформлении научной составляющей ВКР, а ответ на «ИВ3: Какие инструменты / системы / платформы / фреймворки / библиотеки применяют для решения задач…? – проектной составляющей.

Детальность формулировки исследовательских вопросов может быть пересмотрена по результатам выполнения этапа проведения поиска.

Этап 2: проведение поиска

На этом этапе определяется стратегия поиска и выполняется поиск научных публикаций в соответствии с ней.

Стратегия поиска включает:

– определение временной глубины поиска,

– выбор первоначального списка источников публикаций,

– формулировку поисковых запросов.

По сравнению в SLR поисковые запросы для SMS менее сфокусированные, потому что цель SMS состоит в обеспечении широкого охвата области исследования.

Для принятия решения о временной глубине поиска в проводимом SMS необходимо предварительно найти ранние SMS/SLR по близкой тематике и год их опубликования. Точки 1–5 на кривой популярности темы исследования соответствуют некоторым ранним SMS (рис. 3).

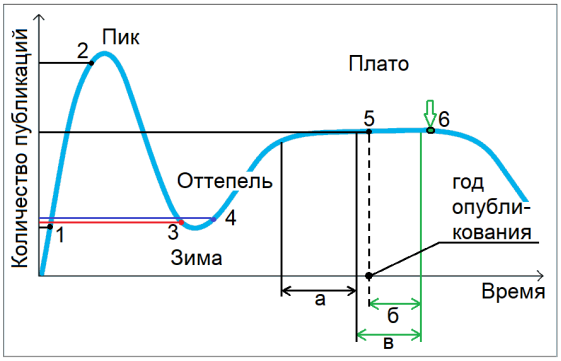

Рисунок 3 – Кривая популярности темы исследования

Часть из них может явно содержать информацию о временном диапазоне

исследованных публикаций в своем названии [10, 11]. В остальных случаях необходимо определить этот интервал из аннотации или текста соответствующей публикации. Это интервал «а» на оси времени для SMS с номером 5 (рис. 2). Следует учитывать, что временной промежуток между правой границей интервала «а» и годом опубликования SMS с номером 5 может составлять 1–3 года. Он включает время на индексирование публикаций, включенных в SMS, и время работы исследователей над SMS. Поэтому в стратегии поиска для проводимого SMS с номером 6 временная глубина поиска может соответствовать интервалам «б» или «в» (рис.2).

В отчете [1] обсуждаются два варианта списков источников публикаций (табл. 1).

Таблица 1 – Названия источников для SLR/SMS в области программной инженерии

Вариант 1	Вариант 2
IEEE Xplore DL	Inspec
ACM DL	EI Compendex
Google scholar	ScienceDirect *
Citeseer library	Web of Science
Inspec	IEEE Xplore DL
ScienceDirect	ACM DL
EI Compendex	
SpringerLink	
Scopus *	

В современных условиях доступ к источникам, помеченным символом «*» отсутствует, они далее в нашей статье не обсуждаются. Поэтому магистрантам рекомендуется основной список источников:

– электронные/цифровые библиотеки НЭБ [12], ACM DL [5], IEEE Xplore DL [6],
– платформа ScienceDirect [7],
– сайт издательства Springer (SpringerLink) [8].

Дополнительно могут быть использованы Google scholar [13], а с учетом мультидисциплинарности тематики ВКР магистрантов – Wiley Online Library [14] и Medline [15].

Поиск рекомендуется проводить по названиям публикаций. Поэтому базовый поисковый запрос по тематике распознавания диаграмм классов за последние пять лет может быть сформулирован так:

((YEAR >= 2020 AND YEAR<=2024) OR (ГОД >= 2020 И ГОД<=2024)) AND ((TITLE = "class diagram recognition") OR (НАЗВАНИЕ="распознавание диаграмм классов"))

Однако выполнение поиска с помощью базового запроса не обязательно будет удовлетворять его цели, и результат может быть:

– пустым,
– неполным,
– избыточным,
– неверным.

Если выполнение запроса по названиям публикаций не дало результата, то поиск необходимо повторить для ключевых слов (при отсутствии таковых в публикации часть источников извлекает их на основе своих внутренних словарей [6]) и/или для аннотаций. В этом случае в базовом запросе необходимо изменить наименование метаданного «TITLE» на «KEYWORDS», «ABSTRACT» или «ALL».

Если при выполнении поискового запроса результатов мало (не более трех), то рекомендуется добавить к значениям метаданного варианты синонимов. Например,

((TITLE = "class diagram recognition") OR (НАЗВАНИЕ="распознавание диаграмм классов")) OR ((TITLE = "extracting model class diagram from images") OR (НАЗВАНИЕ="извлечение модели диаграммы классов из изображения"))

Также следует учитывать, что результаты запросов с ключевыми словами в единственном и множественном числе у некоторых источников отличаются [5, 6]. Так, одинаковые результаты будут для двух запросов

(TITLE = "MOOC")
(TITLE = "MOOCs"),

но будут отличаться для запросов

(TITLE = "massive open online course")
(TITLE = "massive open online courses")

Поэтому запрос с комбинацией ключевых слов в единственном и множественном числе дает более полный результат. Для SMS с номером 1 (рис. 2) результатов поиска будет немного, т.к. исследования по этой теме только начинаются. Поэтому для более широкого охвата рекомендуется применить метод снежного кома (snow-ball), изучая ссылки в списках литературы публикаций, найденных с помощью базового поискового запроса.

Если при выполнении поискового запроса получены избыточные данные, то их удаление производится на этапе 3 применением критериев исключения.

Если при выполнении поискового запроса получены данные, не соответствующие предметной области, то некоторые источники [5, 6, 8] позволяют отфильтровать их до экспорта. Самым распространенной ситуацией, которая приводит к неверным результатам, является использование аббревиатур в качестве значений метаданных запроса. Например, в SMS по тематике обучения/образования/педагогики сокращение «SPOC» обозначает термин «Small Private Online Course» (небольшой закрытый онлайн-курс). Однако в других предметных областях смысл этого сокращения иной: «State and Parameters Observability Canonical», «Statistical Process Optimisation and Control», «Spatial Planar Optical Circuit», фреймворк

«Secure and Privacy-preserving Opportunistic Computing», лэптоп «Shuttle Portable On-Board Computer», «Space Operations Center» (SpOC) в Колумбии, «Scientific data Processing Operations Center», «Single Point of Contact» и т.д. Также аббревиатура может являться частью другого сокращения, например, «SPOC» как часть «iSPOC»/«i-SPOC» (Industry Single Point of Contact), «ISPOC» (IS Student Presentations Over the Cloud) и т.п. Удаление таких результатов производится на этапе 3.

Выполнение поиска в соответствии с выбранной стратегией производится на сайтах соответствующих источников. Результаты поиска представляются браузером в человеко-читаемом виде. Метаданные найденных публикаций в машиночитаемом виде можно получить с помощью сервисов экспорта. Источники позволяют выполнить экспорт в одном или нескольких форматах (табл. 2). Любой из этих форматов, кроме CSV, представляет собой набор пар: имя_тега и значение_тега.

Таблица 2 – Форматы экспорта данных из источников

Формат	Источник				
	НЭБ	ACM DL	IEEE Xplore DL	Science Direct	Springer Link
ACM Ref	-	+	-	-	-
BibTeX	-	+	+	+	-
CSV	-		+	-	+
EndNote	-	+	-	+	-
html	+	-	-	-	-
Plain Text	-	-	+	-	-
RefWorks	-	-	+	+	-
RIS	-	-	+	-	-

Далее для обозначения ограничений цифровизации процесса SMS используются идентификаторы, которые содержат номер этапа SMS и порядковый номер выявленного ограничения.

Ограничением О2.1 на этапе проведения поиска является выбор формата данных для экспорта. О2.2 заключается в том, что файлы одного формата с метаданными публикаций, полученные из различных источников, имеют различную структуру.

Таким образом, результатом этапа 2 после проведения поиска является корпус первичных документов (ПД) – файлы метаданных о всех найденных публикациях).

Этап 3: скрининг документов

На этом этапе выполняется удаление повторяющихся работ и отбор релевантных документов.

Дублирование информации о публикациях происходит в случае проведения совместных научных мероприятий (например, организаторы – ACM и IEEE, и каждый из них размещает материалы в своей цифровой библиотеке) или при индексировании опубликованных работ в библиотеках (НЭБ). Результатом удаления повторяющихся публикаций на этом этапе SMS является корпус уникальных документов (УД) – файл метаданных о неповторяющихся публикациях).

Для отбора релевантных документов формулируются критерии включения и исключения, которые затем применяются к корпусу уникальных публикаций.

Критерии включения позволяют оставить только те публикации, которые соответствуют исследовательским вопросам. Если SMS посвящено исследованию восприятия диаграмм UML студентами-программными инженерами российских вузов, то критерии включения можно сформулировать так:

– тематика публикаций связана с использованием диаграмм UML в профессиональном высшем образовании направления подготовки «Software engineering»/«Программная инженерия»,

– публикации подготовлены авторами из России.

Критерии исключения:

– документы представляют собой прелиминарии, оглавления, предисловия или аннотации книг, сопроводительные части изданий, постеры докладов на научных мероприятиях,

– публикации в научно-популярных изданиях (Magazines),

– публикации являются обзорами литературы,

– объем публикации менее пяти страниц текста.

Последний из критериев следует применять, чтобы исключить из рассмотрения тезисы докладов. Однако в случае использования источника Medline [15], который содержит в основном аннотации публикаций, этот критерий применять нет необходимости. Особые случаи расширения перечня критериев исключения были рассмотрены в предыдущем разделе.

Скрининг метаданных первичных документов в машиночитаемом виде можно выполнить с помощью библиографических менеджеров (БМ, «Reference Manager») Mendeley [16], Zotero [17] и систем автоматизации профессионального поиска литературы BibReader [18], SMS-Builder [19].

Ограничением О3.1 на этапе скрининга является объединение нескольких исходных файлов метаданных в один для удаления повторяющихся публикаций. Некоторые БМ работают только с одним файлом [17].

В системах [16, 17] дубликаты удаляются по одному, а не все сразу. Это существенно увеличивает время обработки корпуса первичных из десятков документов, что является ограничением О3.2 при получении корпуса уникальных публикаций. О3.3 на этапе скрининга – это отсутствие в экспортированных файлах значений метаданных, необходимых для автоматизации отбора релевантных документов при применении критериев исключения (табл. 3): диапазон/количество страниц [8].

Таблица 3 – Теги, используемые источниками публикаций в файлах экспорта

Тег		Источник				
id тега	Название	НЭБ	ACM DL	IEEE Xplore DL	ScienceDirect	SpringerLink
T1	Content Type	+	+	+	+	+
T2	author	+	+	+	+	+
T3	title	+	+	+	+	+
T4	publication_title	+	+	+	+	+
T5	number	+	+	+	+	+
T6	pages	+	+	+	+	-
T7	year	+	+	+	+	+
T8	DOI	-	+	+	+	+
T9	URL	+	+	-	+	+
T10	volume	-	-	+	+	+
T11	ISBN	-	+	-	-	-
T12	ISSN	-	-	+	+	-
T13	month	-	-	+	-	-
T14	note	-	-	-	+	-
T15	address	-	+	-	-	-
T16	publisher	-	+	-	-	-
T17	series	-	+	-	-	+
T18	location	-	+	-	-	-
T19	abstract	-	+	+	+	-
T20	keywords	-	+	+	+	-
T21	numpages	-	+	-	-	-
T22	articleNumber	-	+	-	-	-

Ограничением О3.4 на этом этапе является отсутствие значений метаданных, необходимых для формирования библиографического описания (БО) публикации в соответствии:

– с требованиями издательства, для которого готовится статья с результатами SMS (краткое, расширенное или полное библиографическое описание),

– с видом ресурса (печатный или электронный),

– с типом ресурса (монографическое издание, статья или раздел из монографического или сериального издания, сайт в сети Интернет и т.д.),

– со стилем форматирования (ГОСТ, APA, IEEE, Harvard, Vancouver и т.п.) (табл. 3).

Значение тега T1 идентифицирует тип ресурса. значения T2–T8 используются в БО любой публикации, значение T9 присутствует в БО электронных ресурсов. Большинство источников поставляют в файлах экспорта значения указанных выше тегов для формирования БО.

Теги T11–T13 для монографических и сериальных изданий в некоторых стилях являются факультативными, отсутствие значений этих тегов не влияет на корректность БО. Но не все источники предоставляют значения T10, хотя для сериальных изданий это обязательный элемент БО. Кроме этого, в современных электронных сериальных изданиях диапазон страниц не указывается (T6 отсутствует), в БО его должен заменить элемент «номер статьи» (T22). Соответствующий тег предусмотрен только у одного источника [6].

В БО книжных изданий обязательным элементом является количество страниц. Но T21 экспортируется только из одного источника [5], но не для книг.

В библиографическом описании научного мероприятия требуется указание его названия (T14), места проведения (T15), сведения об издательской функции (T16), факультативно – серии (T17), локации издательства (T18). Однако только источник [5] предоставляет значения тегов T15–T18, при этом у него отсутствует T14.

Также обязательным в БО научного мероприятия является дата(-ы) его проведения, но эту информацию источники не предоставляют.

Таким образом, при выполнении этапа скрининга в корпусе уникальных документов авторам SMS требуется выполнить ручную работу по внесению в файлы метаданных недостающей информации как перед отбором релевантных публикаций, так и после формирования корпуса релевантных – для последующего формирования корректных библиографических описаний.

Ограничение О3.5 заключается в создании списка библиографических описаний корпуса релевантных документов. Для ее решения применимы два вида сервисов: узкоспециализированные (генераторы БО для стилей ГОСТ [20], APA [21], IEEE [22], и др.) и универсальные (БМ [16, 17, 23, 24]). Создание БО проводится для отдельной публикации в интерактивном режиме заданием значений всех элементов БО с дальнейшей автоматической их компоновкой либо получением элементов БО публикации для компоновки из Интернет или имеющегося БО. Библиографические менеджеры могут импортировать файлы метаданных различных форматов и выполнять их пакетную обработку. В контексте выполнения SMS из всего многообразия форматов отобраны те, которые являются форматами экспорта метаданных из источников, рекомендуемых магистрантам (табл. 4)

Таблица 4 – Форматы файлов импорта библиографических менеджеров, совместимые с источниками SMS

Название БМ	Форматы импорта
Citavi	BibTeX, CSV, RIS
JabRef	RIS
Mendeley	BibTeX, RIS
Zotero	BibTeX

Таким образом, результатами этапа 3 после скрининга является корпус релевантных документов (РД) – файл метаданных релевантных публикаций, а также список их библиографических описаний.

Этап 4: ключевые слова с помощью аннотаций

На этом этапе на основе текстов аннотаций как наиболее репрезентативного элемента метаданных выполняется первоначальная категоризация релевантных публикаций. Для этого можно использовать термины (ключевые слова), встречающиеся в аннотациях.

Сервисы построения облаков тегов [25, 26] позволяют загрузить текст, задать количество выводимых слов облака, строят частотный словарь слов текста и визуализируют его. Дополнительно можно управлять фильтрацией спецсимволов [26]. При визуализации результата настраиваются параметры фона, выводимого текста и размер слов с учетом его семантики (частота, ранг и пр.). Эти возможности настройки визуализации облаков тегов могут быть использованы на следующем этапе SMS при картографировании. Изображение построенного облака доступно для импорта в графическом файле. Семантическое ядро текстов аннотаций можно получить с помощью сервисов SEO-анализа [27, 28].

Ограничение О4.1 на этом этапе состоит в том, в корпусе релевантных документов тексты аннотаций – это значения тега T19 (табл. 3) публикаций. Хотя они доступны в машиночитаемом виде, но их извлечение не автоматизировано.

Ограничением О4.2 является отсутствие возможности импорта частотного словаря/семантического ядра: они доступны в человеко-читаемом виде в браузере.

О4.3 заключается в ограничении размера исходного текста сервисов SEO-анализа: у Advego [27] оно составляет 100000 символов, у бесплатной версии Text.ru [28] – 15000. Для корпусов релевантных документов из нескольких десятков публикаций использование этих сервисов будет не всегда возможным.

Этап 5: извлечение данных и картографирование процесса

Картографирование в SMS выполняют в виде «систематических карт» (таблиц или графиков/диаграмм). Поэтому предварительно необходимо подготовить соответствующую информацию, собрав количественные данные из метаданных корпуса релевантных публикаций.

Процесс извлечения данных для SMS шире, чем для SLR: в [1] его называют этапом классификации (категоризации), чтобы ответить на общие исследовательские вопросы и определить статьи для последующего рассмотрения.

Сайты основного списка источников [5, 6, 7, 8] формируют статистику по результатам поиска в виде списков пар:

показатель (количество_публикаций) .

У источников [6, 7, 8] эти списки выполняют роль фильтров, которые применяются к результатам поиска. Некоторые показатели могут быть использованы для извлечения данных для ответа на ИВ1 (табл. 5).

Ограничение О5.1 на этом этапе состоит в том, что статистики доступны только в человеко-читаемом виде.

Таблица 5 – Распределение публикаций по классификационным признакам на сайтах источников

Название признака	Источник			
	ACM DL	IEEE Xplore DL	ScienceDirect	SpringerLink
Год опубликования	+	+	+	-
Тип публикации (книга/журнал/материалы научного мероприятия)	+	+	+	+
Домен деятельности (отрасль знаний)	-	+	+	+
Каналы публикаций (журналы)	+	+	+	-
Каналы публикаций (научные мероприятия)	+	+	+	-
География (аффилиация) авторов	+	+	-	-
Язык	-	-	+	+

Результаты SMS и их оформление

Согласно отчету [1] для подведения итогов исследования результаты должны быть представлены таким образом:

- неколичественные обзоры в табличной форме (табл. 1–5);
- количественные результаты – в виде таблиц и графиков (рис. 1).

Обсуждение результатов SMS как часть раздела ВКР или публикации

Методологией систематического картографического исследования в его отчетности предусмотрен раздел «Обсуждение результатов», который магистрантам можно оформить как часть аналитического раздела ВКР или часть публикации.

Этап анализа результатов SMS (в сравнении с SLR) [1] состоит в обобщении данных для получения ответов на исследовательские вопросы. Визуальное представление распределения исследований по типам классификации повышает эффективность отчетности SMS.

Таким образом, этот раздел ВКР или публикации должен содержать выводы, соответствующие выводам раздела "Результаты". В них могут быть представлены:

– детализация сильных и слабых сторон данных, отобранных в SMS;
– сравнение или связь с другими SMS,
– обсуждение достоверности (внутренней и внешней) доказательств для принятия решения читателем надежности и важности этих доказательств,
– применимость, обобщаемость, распространение результатов,
– обсуждение преимуществ, побочных эффектов и рисков,
– обсуждение различия в эффектах и их причины.

В сравнении с SLR [1] распространение результатов SMS более ограниченное: в контексте нашей статьи оно ограничивается академическими публикациями с целью повлиять на будущее направление первичных исследований.

Для студентов процесс создания этого раздела является одновременно существенно важным и непростым, но способствует формированию необходимых профессиональных компетенций исследователя и критического мышления.

Устранение ограничений цифровизации процесса SMS

Сформулируем постановку задачи цифровизации процесса SMS.

Исходные данные

F: метаданные {множество файлов },

$F = \{f_i^k\}$, где $k -$ номер файла из источника i

Ограничения

$F =< F_{\text{нэб}}, F_{acm}, F_{ieee}, F_{scdirect}, F_{sprlink} >$

$F \neq \emptyset$

$\forall\, i \in [1;5], k \in [1; |F_i|]\ length(f_i^k) \neq 0$

где $i -$ номер источника,

$k -$ номер файла из источника i

формат $f_i^k \in \{.html, .bib, .csv\}$

Результаты

KS: метаданные {файл корпуса ПД, формат $.bib$}

KU: метаданные {файл корпуса УД, формат $.bib$}

KR: метаданные {файл корпуса РД, формат $.bib$}

RF: БО {файл библиографических описаний, $.doc$},

где

BO: < стиль, тип документа >,

стиль = {ГОСТ, $APA, IEEE, harvard$ },

тип документа = { книга, журнал, мероприятие}.

FDF: словарь {файл частотного словаря аннотаций}

где словарь = < термин, число вхождений >

$SF =< S_{\text{нм}}, S_{\text{геогр}}, S_{\text{журн}}, S_{\text{год}}, S_{\text{ист}}, S_{\text{т док}}, S_{\text{авт}} >$

{ файл статистики},

где

$S_{\text{нм}} =<$ мероприятие, число публикаций >,

$S_{\text{геогр}} =<$ страна, число публикаций >,

$S_{\text{журн}} =<$ журнал, число публикаций >,

$S_{\text{год}} =<$ год, число публикаций >,

$S_{\text{ист}} =<$ источник, число ПД, число УД, число РД >

$S_{\text{т док}} =<$ тип документа, число публикаций >

$S_{\text{год}} =<$ год, число публикаций >.

В разделе «Связь» (рис. 4) использованы Ti – id тегов.

Связь
$KS = \{ ks_l \}$,
$KU = \{ ku_j \}$, где:
$ku_j = ks_l$, $\exists l \in [1; |KS|](ks_l \notin KU)$
$KR = \{ kr_j \}$, где:
$kr_j = ku_l$, где:
$\exists l \in [1; |KU|](ku_l.\text{T21} \geq 5)$
$RF = \cup_{i=1}^{kr} BO_i$, где $kr = |KR|$
$\forall i \in [1; N_k] \exists k \in [1; |KR|], KR[k].\text{T1} = \text{"conference"}$:
$S_{\text{НМ}}[i].\text{мероприятие} = KR[k].T4$,
$S_{\text{НМ}}[i].\text{число публикаций} \mathrel{+}= 1$
где:
Nk = число уникальных названий научных мероприятий.
Остальные таблицы статистик формируются аналогично.

Рисунок 4 – Раздел «Связь» постановки задачи

Для устранения ограничений цифровизации процесса SMS была разработана система BibReader [18, 29, 30] (рис. 5). С ее помощью выполнение этапов 3 и 4 автоматизировано.

Результаты поиска в НЭБ должны быть сохранены как html-страница, из SpringerLink экспортированы в формате CSV, из остальных источников – экспортированы в формате BibTeX (так выполняется снятие ограничения О2.1). Полученные после поиска файлы загружаются в систему в произвольном порядке или в пакетном режиме. Они автоматически объединяются в единый файл корпуса ПД (снятие О3.1), с метаданными которых далее можно выполнять следующие этапы SMS.

Метаданные публикаций проходят идентификацию – принадлежность источнику, унификацию – приведение тегов-синонимов к единому обозначению и единому регистру представления, очистку – удаление избыточных символов у значений тегов и т.п. (снятие О2.2).

При удалении повторяющихся публикаций сохраняются метаданные документа, полученного из источника с наибольшим количеством тегов.

Метаданные корпусов первичных, уникальных и релевантных документов можно выбирать, в том числе с помощью фильтров (снятие ограничения О4.1), просматривать, редактировать и сохранять. Это позволяет выполнить необходимую обработку перед автоматическим отбором релевантных документов, а также внести недостающие данные в публикации после завершения скрининга (возможность снятия О3.3 и О3.4).

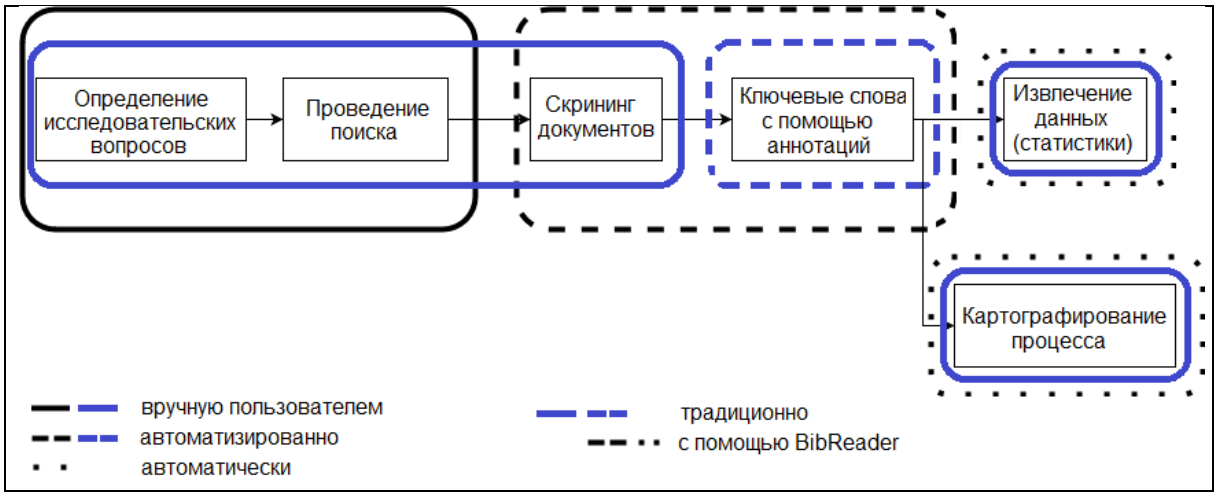

Рисунок 5 – Цифровизация процесса SMS с помощью системы BibReader

При необходимости исследователь может сохранить файл метаданных на любом шаге, получая последовательно корпуса первичных, уникальных и релевантных документов.

С помощью BibReader формируется список библиографических описаний для выбранного корпуса документов, оформленных по ГОСТ или стилям APA, IEEE, Harvard. Этот результат сохраняется в файл текстового процессора.

При выполнении этапа 4 предусмотрены возможности извлечения данных из тегов аннотации документов, построения частотного словаря, удаления стоп-слов в частотном словаре, задания количества слов в формируемом облаке, настройки визуализации облака, сохранения частотного словаря в файл табличного процессора (снятие О4.2) и облака – в графический файл. При эксплуатации BibReader

ограничений на размер текста аннотаций корпуса РП не выявлено (снятие О4.3).

С помощью BibReader извлечение данных этапа 5 в виде статистики для ответа на первый исследовательский вопрос выполняется автоматически. Формируются распределения публикаций для выбранного корпуса по источникам, годам опубликования, типу документа, каналам публикаций (журналы и научные мероприятия), странам аффилиации авторов публикации, количественному составу авторских коллективов. При этом существует возможность выбрать один из вариантов картографирования извлеченных данных в виде диаграммы. Данные распределений могут быть сохранены (снятие О5.1) в файл табличного процессора для дальнейшей обработки, диаграмма – в графический файл.

Таким образом, интеграция стека технологий, выполняющих цифровизацию процесса SMS, в единую систему позволяет устранить частично или полностью все ограничения, выявленные в разделах этой статьи. Это приводит к значительному сокращению времени выполнения проведения SMS, получению данных и созданию отчетности для обсуждения результатов SMS. Также формируются некоторые систематически карты, которые могут быть использованы для представления SMS в тексте ВКР или в виде научной публикации [31, 32]. В учебном процессе BibReader используется с 2019-2020 учебного года.

Применимость методологии систематического картографического исследования

Определим место SMS в научных исследованиях.

Большинство дисциплин учебного плана образовательной программы магистратуры в области ИКТ предусматривают рассмотрение тем/вопросов, связанных с выполнением ими научных исследований по теме ВКР. В случае продолжения работ по теме ВКР бакалавра студентам проще выполнять SMS, т.к. с предметной областью они уже знакомы. При обучении «стороннего» магистранта (выпускника другого направления подготовки или другого университета) выполнять SMS студентам сложнее, но удается и с «низкого старта». Результаты SMS используются магистрантами в аналитическом разделе ВКР.

В данной статье рассмотрено использование методологии SMS в домене исследований ИКТ. Но систематическое картографическое исследование – это методология не только «в» и «для» ИКТ. Анализ 749 SMS, опубликованных в 2009–2024 гг., по измерению «домен исследования» показал следующее.

Во-первых, методология SLR/SMS до применения в ИКТ была и остается базовой в медицинских исследованиях. Среди указанных SMS представлены исследования в области телемедицины и здравоохранения.

Во-вторых, большинство исследований посвящено методам и технологиям ИКТ, актуальным в каждый момент их развития: качество программного обеспечения (ПО), требования к ПО, геймификация, облачные технологии, архитектуры информационных и программных систем, пользовательские интерфейсы, тестирование ПО, блокчейн, онтологический подход, DSL и Model-driven подходы, машинное, глубокое и федеративное обучение и т.д.

В-третьих, цифровизация охватывает другие предметные области, результаты обсуждаются в SMS из разнообразных доменов:

– Индустрия 4.0, моделирование в строительстве, авионика, градостроительство, сохранение геонаследия,
– сельское и лесное хозяйство,
– цифровая экономика, фондовые рынки,
– обучение, в том числе инженерное и STEM.

Выявленная междисциплинарность методологии SMS показывает перспективы ее применения в современных условиях при подготовке магистрантов различных направлений подготовки, в первую очередь, инженерных.

Выводы

В работе проанализированы особенности выполнения этапов методологии систематического картографического исследования в контексте цифровизации этого процесса. Выявлены ограничения, связанные с разрывом между представлением данных, подлежащих последовательной обработке и возможностями сервисов, автоматизирующих эту обработку. Основная причина ограничений – это представление данных в человеко-читаемом виде и необходимость ручной обработки данных, представленных уже в машиночитаемом виде.

Сформулирована постановка задачи цифровизации процесса SMS. Предложено решение задачи с помощью системы BibReader, которая реализует конвейер поэтапной обработки исходных метаданных публикаций, формируя необходимые файлы отчетности SMS.

BibReader используется в учебном процессе для профессионального поиска научной литературы студентами образовательной программы магистратуры в области ИКТ при подготовке SMS и аналитической части ВКР.

Анализ публикаций-SMS за 2007-2024 гг. показал не только тренд роста интереса к методологии систематического картографического исследования, но ее перспективность для профессионального поиска литературы в предметных областях, отличных от ИКТ.

Литература

1. Kitchenham, B., Charters, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001.– Keele University, Durham University Joint Report. – 2007. – [Электронный ресурс]. – Режим доступа: https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf (дата обращения: 26.05.2024).

2. Latifaj, M. A Systematic Mapping Study on Quality of Service in Industrial Cloud Computing. Master's Thesis in Software Engineering. Malardalen University, Sweden. 2020. [Электронный ресурс]. – Режим доступа: https://www.diva-portal.org/smash/get/diva2:1438758/FULLTEXT01.pdf (дата обращения: 26.05.2024).

3. Truger, F. Warm-Starting and Quantum Computing: A Systematic Mapping Study / F. Truger, J. Barzen, M. Bechtold [et al.] // ACM Computing Surveys. – 2024. – Vol. 56, Iss. 9. – Article No. 229. – Pp. 1–31. [Электронный ресурс]. – Режим доступа: https://dl.acm.org/doi/10.1145/3652510 (дата обращения: 26.05.2024).

4. Alahiane, A. The Use of Deep Learning, Image Processing, and High-Performance Computing: A Systematic Mapping Study / A. Alahiane, K. El Asnaoui, S. Chadli [et al.] // AI2SD' 2023. Lecture Notes in Networks and Systems. – Vol. 931. – Cham: Springer, 2024. – Pp. 223–235. [Электронный ресурс]. – Режим доступа: http://link.springer.com/chapter/10.1007/978-3-031-54288-6_21 (дата обращения: 26.05.2024).

5. ACM Digital Library: сайт. – URL: https://dl.acm.org/ (дата обращения: 07.04.2024). – Текст: электронный.

6. IEEE Xplore Digital Library: сайт. – URL: https://ieeexplore.ieee.org/ (дата обращения: 07.04.2024). – Текст: электронный.

7. ScienceDirect: сайт. – URL: https://www.sciencedirect.com/ (дата обращения: 07.04.2024). – Текст: электронный.

8. SpringerLink: сайт. – URL: https://link.springer.com/ (дата обращения: 07.04.2024). – Текст: электронный.

9. Дацун, Н. Н. Совместное оценивание деятельности обучающихся в массовых открытых онлайн курсах: систематический обзор литературы // Мир науки. – 2015. – № 3. – С. 1–24.

10. Wolny, S. Thirteen Years of SysML: A Systematic Mapping Study / S. Wolny, A. Mazak, C. Carpella [et al.] // Software and Systems Modeling. – 2020. – Vol. 19. – Pp. 111–169. [Электронный ресурс]. – Режим доступа: https://doi.org/10.1007/s10270-019-00735-y (дата обращения: 26.05.2024).

11. Shaikh, A. More Than Two Decades of Research on Verification of UML Class Models: A Systematic Literature Review / A. Shaikh, A Hafeez, A. A. Wagan [et al.] // IEEE Access. – 2021. – Vol. 9. – Pp. 142461–142474. [Электронный ресурс]. – Режим доступа: https://ieeexplore.ieee.org/document/9579419

12. eLIBRARY.RU: сайт. – URL: https://elibrary.ru (дата обращения: 26.05.2024). – Текст: электронный.

13. Google Академия: сайт. – URL: https://scholar.google.com/ (дата обращения: 26.05.2024). – Текст: электронный.

14. Wiley Online Library: сайт. – URL: https://onlinelibrary.wiley.com/ (дата обращения: 26.05.2024). – Текст: электронный.

15. Medline: сайт. – URL: https://www.ebsco.com/products/research-databases/medline (дата обращения: 26.05.2024). – Текст: электронный.

16. Mendeley: сайт. – URL; https://www.mendeley.com/?interaction_required=true (дата обращения: 26.05.2024). – Текст: электронный.

17. Zotero: сайт. – URL; https://www.zotero.org/ (дата обращения: 26.05.2024). – Текст: электронный.

18. Субботин, Е. А. Система автоматизации скрининга публикации для систематического обзора литературы / Е. А. Субботин, Н. Н. Дацун // Математика и междисциплинарные исследования – 2019. Материалы Всероссийской научно-практической конференции молодых ученых с международным участием. – Пермь: ПГНИУ, 2019. – С. 363–367.

19. Candela-Uribe, C.A. SMS-Builder: An Adaptive Software Tool for Building Systematic Mapping Studies / C.A. Candela-Uribe, L.E. Sepúlveda-Rodríguez, J.C. Chavarro-Porras [et. al.] // SoftwareX. – 2021. – Vol. 17. – Article No. 100935. – Pp. 1–10. [Электронный ресурс]. – Режим доступа: https://www.sciencedirect.com/science/article/pii/S2352711021001710/pdfft?md5=c8b896312ab85de39d17e3eb64027c85&pid=1-s2.0-S2352711021001710-main.pdf (дата обращения: 26.05.2024).

20. Список литературы и сноски онлайн: сайт. – URL: https://open-resource.ru/spisok-literatury (дата обращения: 26.05.2024). – Текст: электронный.

21. Citation Machine. APA Citation Generator: сайт. – URL: https://www.citationmachine.net/apacitationmachine (дата обращения: 26.05.2024). – Текст: электронный.

22. Cite This For Me. Free IEEE Citation Generator: сайт. – URL: https://www.citethisforme.com/citation-generator/ieee (дата обращения: 26.05.2024). – Текст: электронный.

23. Citavi: сайт. – URL: https://www.citavi.com/en (дата обращения: 26.05.2024). – Текст: электронный.

24. JabRef: сайт. – URL: http://www.jabref.org/ (дата обращения: 26.05.2024). – Текст: электронный.

25. Word Cloud: сайт. – URL: https://www.jasondavies.com/wordcloud/ (дата обращения: 26.05.2024). – Текст: электронный.

26. Word it out: сайт. – URL: https://worditout.com/word-cloud/create (дата обращения: 26.05.2024). – Текст: электронный.

27. Advego: сайт. – URL; https://advego.com/ (дата обращения: 26.05.2024). – Текст: электронный.

28. Text.ru: сайт. – URL; https://text.ru/seo (дата обращения: 26.05.2024). – Текст: электронный.

29. Шукшина, М. И. Совершенствование реализации этапов систематического картографирования литературы в системе BibReader / М. И. Шукшина, Н. Н. Дацун // Математика и междисциплинарные исследования – 2020. Материалы Всероссийской научно-практической конференции молодых ученых с международным участием. – Пермь: ПГИУ, 2020. – С. 78–82.

30. Скоробогатова, М. М. Адаптация системы BibReader к источникам публикаций и стилям форматирования библиографических описаний для проведения систематического картографирования литературы // Электронные системы и технологии. Сборник материалов 58-й научной конференции аспирантов, магистрантов и студентов БГУИР. – Минск: БГУИР, 2022. – С. 281-284.

31. Колесников, А. С. Методы и средства распознавания UML-диаграмм: систематическое картографирование литературы / А. С. Колесников, Н. Н. Дацун // Инновационные технологии: теория, инструменты, практика, 2022. – Т. 1. – С. 59–66.

32. Беляков, К. В. Систематическое картографирование литературы: использование транспайлеров / К. В. Беляков, Н. Н. Дацун // Вестник компьютерных и информационных технологий, 2023. – Т. 20, № 7 (229). – С. 3–10.

Дацун Н. Н. Профессиональный поиск научной литературы для исследовательских работ магистрантов. *В статье рассматривается проблема повышения эффективности исследовательской работы ИКТ-магистрантов при профессиональном поиске научной литературы с помощью методологии систематического картографического исследования. Выявлены ограничения цифровизации процесса по этой методологии. Сформулирована постановка задачи цифровизации этого процесса. Представлена цифровизация процесса в системе BibReader для решения проблемы эффективности исследования. Применение рассмотренной методологии является перспективным направлением при поиске научной литературы для студентов других образовательных программ.*

Ключевые слова: *систематическое картографическое исследование, магистрант, цифровизация, ограничение, критерии включения и исключения, библиографическое описание.*

Datsun N.N. Professional Searching the Scientific Literature for Research Papers of master's students. *The article deals with the problem of increasing the effectiveness of research work of ICT master's students in the professional search for scientific literature using the methodology of Systematic Mapping Study. Digitalization constraints of the process according to this methodology are revealed. The problem statement of this process digitalization is formulated. The digitalization of the process in the BibReader system is presented to solve the problem of research effectiveness. The application of the considered methodology is a promising direction in the search for scientific literature for students of other educational programs.*

Keywords: *Systematic Mapping Study, Master's degree student, digitalization, constraint, inclusion and exclusion criteria, bibliographic description.*

Статья поступила в редакцию 20.05.2024
Рекомендована к публикации профессором Мальчевой Р. В.