***ШОРИН Олег Николаевич*** -
*заместитель генерального директора*
*по информатизации ФГБУ*
*«Российская национальная библиотека»*
*Адрес: 191069, г. Санкт-Петербург,*
*ул. Садовая, 18*
*e-mail: shorin@nlr.ru*

***УДК 004.4***

# ИНТЕГРАЦИЯ БИБЛИОГРАФИЧЕСКИХ ЗАПИСЕЙ

***Введение***

В Министерстве культуры Российской Федерации предпринимаются попытки, направленные на реализацию нового этапа развития Национальной электронной библиотеки (НЭБ). Основной целью этого этапа является обеспечение свободного, равного и всеобщего доступа граждан нашей страны к документной информации историко-культурного, научного и образовательного назначения через сеть Интернет, предоставляемой на основе единой общенациональной системы создания и эффективного использования цифровых библиотечно-информационных ресурсов и сервисов.

Основной инфраструктурной особенностью функционирования НЭБ является одновременное наличие центрального портала и распределенного характера процессов создания, каталогизации, подключения, использования и хранения ресурсов среди участников НЭБ. Центральный портал обеспечивает навигацию и поиск по полным текстам и библиографическим записям распределенного фонда, доступ к ресурсам с учетом ограничений, накладываемых авторскими правами, функционирование единого электронного читательского билета, сбор статистики доступа к документам и составления отчетности, координацию процессов оцифровки различными библиотеками.

Основным отличием НЭБ от проектов подобного рода, например, сводного каталога электронных ресурсов (СКЭР), является наличие единого интерфейса для работы с документами из различных библиотек. Предыдущие проекты предоставляли только ссылки на электронные ресурсы, для получения доступа к которым необходимо было регистрироваться в каждой конкретной библиотеке, устанавливать и настраивать специализированное программное обеспечение, применяемое в данной библиотеке.

***Задача интеграции библиографических записей и выявления дублетных записей***

Выполняя свои уставные функции, библиотеки создают библиографические записи на экземпляры, хранящиеся в их фондах. Общее количество записей, хранящихся только в двух крупнейших библиотеках страны - РГБ и РНБ, составляет несколько десятков миллионов. Поиск и получение новых и обновленных записей в таком огромном массиве постоянно меняющейся информации представляет из себя отдельную задачу - сбора библиографических записей. Если выбранный способ будет сильно загружать центральный сервер, то при увеличении количества библиотек, участвующих в проекте, работоспособность всей системы в целом не может быть гарантирована.

Априорно известно, что при получении библиографических записей из разных библиотек часто будет возникать ситуация, когда на один и тот же экземпляр будет существовать несколько записей. Эти

записи будут отличаться как по формату, так и по полноте заполнения, поскольку в различных учреждениях процессы каталогизации отличаются друг от друга. Например, состав дополнительных элементов, точек доступа может быть разным, системы классификации и предметизации, использование аббревиатур также могут разниться от одной организации к другой. К тому же, запись может содержать ошибки, опечатки, вызванные банальной человеческой оплошностью и невнимательностью.

Вопросы интеграции библиографических записей из различных источников с последующим их объединением и обогащением давно находятся в фокусе внимания ученых. Однако большинство наиболее распространенных алгоритмов использует метод сравнения всех записей со всеми, из-за этого сложность алгоритма составляет $O(n^2)$. Возникает необходимость создания алгоритма, позволяющего существенным образом сузить множество библиографических записей - потенциальных кандидатов на дублетность. Большинство подобных алгоритмов основано на разбиении всего множества данных на кластеры, внутри которых содержатся подобные друг другу записи.

### *Сбор библиографических записей*

Поскольку структура Национальной электронной библиотеки представляет из себя распределенную конфигурацию с выделенным центральным сервером [1], на котором агрегируются записи из различных источников, то наиболее подходящим протоколом для сбора информации является узкоспециализированный протокол, изначально предназначенный для решения именно этой задачи - OAI-PMH. Именно он и используется для агрегации библиографических записей из АБИС различных библиотек.

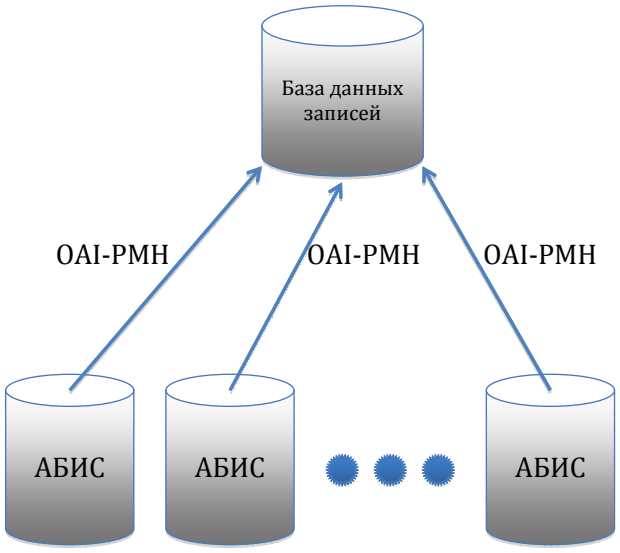


***Рис. 1.*** *Сбор библиографических записей*

Стоит отметить, что объединенный мировой каталог WorldCat на базе протокола OAI-PMH создал Java-сервлет - OAICat, который держатели метаданных могут адаптировать и установить на своем оборудовании для автоматического сбора обновленных записей центральным сервером WorldCat. Европейская цифровая библиотека Europeana также создала сервис REPOX [7], который основан на использовании протокола OAI-PMH в качестве базового протокола для сбора ресурсов с тысяч различных серверов, расположенных по всей Европе. Нельзя не сказать, что обновленная версия Сводного каталога библиотек России, известная под названием СКБР2, агрегирует записи с помощью протокола OAI-PMH. Использование протокола OAI-PMH для аккумуляции библиографических записей из различных источников такими крупными мировыми проектами еще раз убеждает в правильности выбора протокола доступа.

Для осуществления сбора библиографических записей из библиотек - участников НЭБ - были созданы модули интеграции для автоматизированных библиотечных интегрированных систем, используемых в библиотеках: Aleph, Ирбис, MarcSQL и OPAC-Global. Схема сбора библиографических записей приведена на **рис. 1.**

Центральный сервер, на котором располагается хранилище данных, раз в сутки посылает запрос автоматизированным системам библиотек. Формат запроса выглядит следующим образом: http://aleph.nlr.ru/OAI?verb=ListRecords&metadataPrefix=marc21&from=2015-03-12&until=2015-03-12 . Параметрами from и until ограничивается временной период - одни сутки. В качестве ответа выдается список библиографических записей в формате MARCXML, измененных за указанную в запросе дату. Полученные от библиотек библиографические записи сохраняются в базе данных на центральном сервере для последующей обработки.

### *Выявление дублетных библиографических записей и их слияние*

Как было сказано ранее, основной задачей алгоритма выявления дублетных библиографических записей является сужение множества записей, являющихся претендентами на дублетность. Для достижения этой цели в разработанной системе используется подход разбиения всего множества записей на кластеры. Для этого применяется функция хэширования, основными свойствами которой являются:

- незначительное изменение значения функции при незначительном изменении аргумента;
- и, наоборот, значительное изменение значения функции при значительном изменении аргумента.

Одной из наиболее часто используемых функций, обладающих такими свойствами, является функ-

ция SimHash [2]. Ее реализация является достаточно простой, к тому же на практике показано [6], что существует взаимосвязь между побитовым совпадением двух хэшей, посчитанных для двух множеств с использованием функции SimHash, и мерой Жаккара [4] для этих множеств. Именно это свойство функции SimHash и используется для разбиения библиографических записей на кластеры.

Поля «Автор» и «Название» библиографических записей сначала проходят через процесс нормализации, в результате которого все символы приводятся к нижнему регистру, а множества пробельных символов заменяются на один пробел. Затем поля «Автор» и «Название» всех записей разбиваются на биграммы - двухбуквенные сочетания. Например, для строки «Достоевский» будет сгенерировано следующее множество биграмм: «До», «ос», «ст», «то», «ое», «ев», «вс», «ск», «ки», «ий». Для множеств биграмм, полученных из полей «Автор» и «Название», вычисляются два 32-битных хэша с использованием функции SimHash.

В базе данных Национальной электронной библиотеки собрали 21 313 009 библиографических записей. Для указанного количества записей используется следующий алгоритм разбиения на кластеры: каждый из хэшей полей «Автор» и «Название» разбивается на 4 части, и в качестве претендентов на дублетность выбираются записи, в хэшах которых совпадает 2 из 4 частей для поля «Автор» и 2 из 4 частей для поля «Название». То есть для объединенного 64-битного хэша, имеющего вид

$hash_1$-$hash_2$-$hash_3$-$hash_4$-$hash_5$-$hash_6$-$hash_7$-$hash_8$,

в один кластер попадут все записи, в которых совпадают части

$hash_i$ - $hash_j$ - $hash_p$ - $hash_q$, *где* $1 \le i < j \le 4$ *и* $5 \le p < q \le 8$.

Таким образом, используя вышеописанный алгоритм для поиска близких хэшей для полей «Автор» и «Название», все исходное множество библиографических записей разбивается на кластеры, которые содержат претендентов на дублетность.

Внутри каждого кластера для всех пар записей вычисляется мера Жаккара для множеств биграмм, полученных из полей «Автор» и «Название». Мера Жаккара для двух множеств A и B представляет собой отношение количества совпадающих элементов этих множеств к суммарному количеству элементов в этих множествах:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Если мера Жаккара превышает пороговое значение, то между этими библиографическими записями устанавливается соответствие. Пороговое значение было получено экспериментальным путем, и оно составляет:

$T$=0,7.

Для записей, между которыми установлено соответствие, запускается процесс создания обогащенной записи. Обогащение осуществляется по свойствам формата MODS. Для библиографической записи в формате MODS существуют простые и составные свойства. Для составных свойств происходит объединение наборов данных из различных записей. Для простых свойств возможны две ситуации. В случае, когда свойство присутствует в одной записи и отсутствует в других, берется свойство из той записи, в которой оно присутствует. В случае присутствия простого свойства в нескольких записях, используется подход, описанный Jeremy A. Hylton [3].

В соответствии с этим подходом для каждого возможного написания свойства подсчитывается количество записей, в которых присутствует такое написание. В обогащенную запись попадает свойство, присутствующее в наибольшем количестве записей. Иногда различные написания присутствуют в одинаковом количестве записей. В таком случае в обогащенную запись попадает свойство, имеющее наибольшую длину, поскольку более длинное свойство содержит в себе больше информации.

### *Результаты работы алгоритма выявления дублетных библиографических записей*

Алгоритм выявления дублетных библиографических записей и создания обогащенных на их основе реализован на языке Java с использованием фреймворка Spring MVC, объектно-реляционного отображения Hibernate и библиотеки XMLBeans. Выбор Spring MVC основан на том, что он позволяет создавать легко поддерживаемый исходный код за счет деления его на модули, соответствующие разным функциональностям. Подобное деление полностью согласуется с концепцией Model-View-Controller (MVC) [5].

Hibernate обеспечивает отображение классов Java на таблицы реляционных баз данных, таким образом освобождая разработчика от лишней работы. Этот подход не ограничивает разработчика исходно выбранной СУБД. Для обработки данных в формате XML используется библиотека XMLBeans. Она позволяет на основе схемы XML получать библиотеку или набор классов для представления данных, описанных в этой схеме.

Процесс выявления дублетов и слияния записей исполняется на центральном сервере системы. Данные берутся из хранилища, в которое попадают записи, собранные из различных библиотек с использованием протокола OAI-PMH.

В своей работе алгоритм использует временную базу данных, состоящую из 3 таблиц:

1. Таблица для хранения уникальных библиографических записей - BibRecords.

2. Таблица для хранения обогащенных библиографических записей - EnrichedBibRecords.

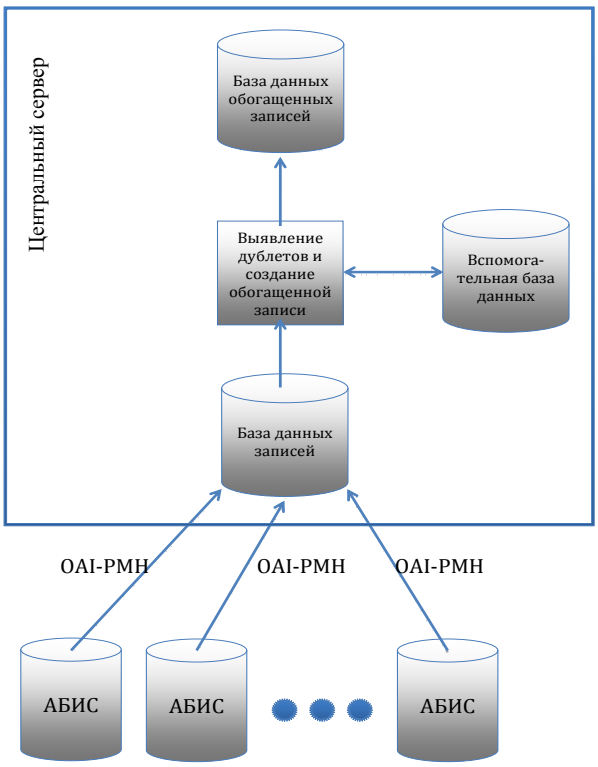


***Рис. 2.*** *Схема хранения данных на центральном сервере*

3. Таблица для хранения хэшей - StringHash.

В результате работы алгоритма обогащенные записи попадают в отдельную базу данных, также располагающуюся на центральном сервере. Общая схема хранения исходных, вспомогательных и выходных данных на центральном сервере приведена на **рис. 2.**

Алгоритм выявления дублетных библиографических записей выполняется на сервере, который, по данным программы синтетических тестов AIDA64, обладает следующими характеристиками:

1. Процессор - двухпроцессорный Intel(R) Xeon(R) CPU E5645 @ 2.40 GHz, обладающий 12 реальными и 24 виртуальными ядрами.

2. Жесткий диск - RAID0 Intel Multi-Flex SCSI Disk Device. Средняя скорость чтения данных - 150 MB/s.

3. Оперативная память - 98 238 MB DDR3 1066 MHz.

На сервере установлена операционная система Microsoft Windows Server 2012 Standard, а в качестве СУБД используется Microsoft SQL Server 2012 - 11.0.2100.60 (X64).

Алгоритм выявления дублетных библиографических записей обработал 21 313 009 записей. Для обработки такого объема данных на указанном сервере потребовалось 90 часов. Столь небольшого времени работы алгоритма удалось добиться за счет того, что вся хранимая в базе данных информация, занимающая всего 60 гигабайт, была полностью подгружена в оперативную память. Это позволило избавиться от относительно медленных операций чтения/записи из/в СУБД.

В результате работы алгоритма было выявлено, что 1 239 293 библиографические записи являются дублетами какой-нибудь другой записи. Уникальных записей, т.е. записей, для которых не существует ни одного дублета, оказалось 19 066 057.

***Заключение***

В процессе создания системы интеграции библиографических записей необходимо было решить ряд проблем, связанных со сбором и обработкой больших массивов. Изначальное решение о создании модульной системы позволило разбить основную задачу на ряд мелких подзадач, для которых можно применить эффективные алгоритмы, использовать высокопроизводительные протоколы взаимодействия и узкоспециализированное программное обеспечение. Основываясь на теоретическом и практическом опыте ведущих зарубежных и отечественных специалистов, были приняты решения, которые позволили увеличить масштабируемость, гибкость, отказоустойчивость и интероперабельность всей системы в целом.

В частности, для сбора библиографических записей из библиотек используется узкоспециализированный протокол OAI-PMH, который был разработан непосредственно для решения задачи аккумуляции данных из разнородных источников информации.

Для всех записей, собранных из различных источников, применяется вероятностный алгоритм, который позволяет эффективно выявить дублетные записи и создать обогащенную запись, основываясь только на информации из основных полей записи. Использование данного алгоритма позволяет существенным образом уменьшить количество попарных операций сравнения.

***Список литературы:***

*1. Шорин О. Н. Национальная электронная библиотека: технологии реализации / О.Н. Шорин // Национальная библиотека.- 2015. - № 3(06). - С. 60-65.*

*2. Charikar M. S. Similarity estimation techniques from rounding algorithms / Moses S. Charikar // Proceedings of the forty-fourth annual ACM symposium on Theory of computing - STOC'02, Montreal, QC, Canada, May 19 - 21, 2002. - New York: ACM, 2002. - P. 380-388.*

*3. Hylton J. Identifying and merging related bibliographic records: [thesis/diss.] / Jeremy A. Hylton; Massachusetts Inst. of Technology, Dept. of Electric. Engineering a. Computer Science. - Cambridge, MA, USA: Massachusetts Inst. of Technology, 1996. - 99 p.: ill.*

*4. Jaccard P. Distribution de la flore alpine: dans le Bassin des Dranses et dans quelques régions voisines / Paul Jaccard // Bull. de la Soc. Vaudoise des sciences naturelles. - 1901. - Vol. XXXVII, № 140. - P. 241-272.*

*5. Reenskaug T. The Model-View-Controller (MVC): its past and present [Electronic resource] / Trugve Reenskaug // University of Oslo, Department of Informatics: [webserver for all the personal homepages]. - Electronic data. - [Oslo, Norway], 2003. - 16 p. - Mode of access: http://home.ifi.uio.no/trygver/2003/javazone-jaoo/MVC_pattern.pdf, free. - Title of screen.*

*6. Shrivastava A. In defense of MinHash over SimHash / Anshumali Shrivastava, Ping Li // The JMLR: Workshop and Conference Proceedings. Vol. 33: Proceedings of the seventeenth international conference on artificial intelligence and statistics (AISTATS), Reykjavik, Iceland, Apr. 2, 2014 / [ed. by: Samuel Kaski, Jukka Corander]. - Cambridge, Mass.: [MIT Press], 2014. - P. 886-894.*

*7. D5.3.1 - Europeana OAI-PMH infrastructure - documentation and final prototype Appendix - REPOX User Manual [Vienna, Austria, 11 Oct. 2010] [Electronic resource] / Gilberto Pedrosa, Petz Georg, Cesare Concordia, Nicola Aloia; Europ. Union, Austr. Nat. Libr. // Europeana.eu. Connect: [offic. website] / Europeana Foundation, Europ. Union. - Electronic data. - Den Haag, Netherlands [2008-2010]. - 33 p. - Mode of access: http://www.europeanaconnect.eu/documents/01a_Europeana_OAI_PMH_APPENDIX_User%20Manual.pdf, free. - Title of screen.*

## НАША ИНФОРМАЦИЯ

**ФГБУ «Российское энергетическое агентство» Минэнерго России (РЭА) и Центр возобновляемых источников энергии и энергосбережения Греции (CRES) выступили организаторами Российско-греческого форума по возобновляемой энергетике и энергоэффективности. Он завершился 18 сентября 2016 г. в г. Салоники на полях 81-й Международной торговой выставки. Мероприятие было проведено в рамках реализации Программы «перекрестных» годов России и Греции.**

В работе Форума приняли участие представители исполнительной власти России и Греции, руководители крупнейших компаний, а также ведущих российских и греческих общественных и отраслевых организаций, экспертов. Открывая работу Форума, министр энергетики Российской Федерации А.В. Новак отметил, что в сфере энергоэффективности и развития возобновляемых источников энергии у наших стран имеется большой потенциал для наращивания сотрудничества. Сегодня и общество, и бизнес во всем мире проявляют к этим вопросам повышенный интерес. Энергоэффективность рассматривается как ключевой фактор обеспечения конкурентоспособности не только отдельных компаний, но и национальных экономик в целом. В ней заложен уникальный потенциал, помогающий обеспечить энергетическую безопасность, экономический рост и улучшение качества жизни граждан.

А. Новак также сообщил, что объем возобновляемых источников энергии с учетом ГЭС в России к 2024 году достигнет порядка 20% в общем объеме. Российским правительством поставлена цель по расширению возобновляемых источников энергии в энергобалансе. Без учета выработки гидрогенерации - до 2,5%, а с учетом производства гидроэнергии за счет гидроэлектростанций эта доля будет около 20% к 2024 году.

Основная задача Форума - осветить потенциал развития двустороннего российско-греческого сотрудничества в области энергоэффективности и возобновляемой энергетики и стать благоприятной площадкой для обмена опытом в области перспективных технологий и разработок, презентации инвестиционного потенциала обеих сторон в указанных областях, а также обсуждения предложений по созданию совместных предприятий с участием российского и греческого капитала.

Участники Форума обсудили в ходе пленарных заседаний и панельных дискуссий перспективы развития возобновляемых источников энергии (ВИЭ), повышения энергоэффективности, вопросы правового регулирования в области ВИЭ и энергосбережения в Греции, практические аспекты государственной политики и законодательства в области ВИЭ и энергосбережения в Российской Федерации, обменялись опытом перспективных технологий и разработок ВИЭ.

Российские и греческие участники мероприятия также провели «круглый стол» (биржу контактов) по определению направлений взаимодействия в области ВИЭ и энергоэффективности.

Как отметил генеральный директор РЭА А.В. Тихонов, в проведении совместной системной работы заложен большой потенциал. Таким образом создаются предпосылки для повышения конкурентоспособности наших стран в системе глобальных экономических отношений, формируются условия для дальнейшего гармоничного развития других отраслей, связанных с энергетикой.

Проведение Форума в рамках деловой программы 81-й Международной торговой выставки в г. Салоники позволило придать дополнительный импульс развитию российско-греческого взаимодействия в области энергетики, способствовало установлению новых деловых контактов между организациями и компаниями обеих стран.

**Сайт**: ***http://www.rosenergo.gov.ru/cur_news/2016-09-12/262/***