УДК 019.963.2:004.62

А. М. Федотов, О. Л. Жижимов, А. А. Князева, О. С. Колобов, Н. А. Мазов, И. Ю. Турчановский, О. А. Федотова

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

Институт вычислительных технологий СО РАН, Новосибирск
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

Отдел проблем информатизации ТНЦ СО РАН
пр. Академический, 10/4, Томск, 634028, Россия

Институт нефтегазовой геологии и геофизики им. А. А. Трофимука СО РАН
пр. Акад. Коптюга, 3, Новосибирск, 630090, Россия

Институт сильноточной электроники СО РАН
пр. Академический, 2/3, Томск, 634055, Россия

Государственная публичная научно-техническая библиотека СО РАН
ул. Восход, 15 Новосибирск, 630200, Россия
E-mail: fedotov@nsc.ru; okolobov@hcei.tsc.ru

ПРОБЛЕМЫ АВТОРИТЕТНОГО КОНТРОЛЯ ДЛЯ РАСПРЕДЕЛЕННЫХ ЭЛЕКТРОННЫХ БИБЛИОТЕК И БИБЛИОГРАФИЧЕСКИХ БАЗ ДАННЫХ *

В статье обсуждаются проблемы авторитетного контроля и идентификации информационных ресурсов. Представлен алгоритм автоматического авторитетного контроля, позволяющий связывать библиографические и авторитетные записи. Анализ факторов, влияющих на соответствие записей и расчет параметров алгоритма, производился с помощью статистического эксперимента на библиографических и авторитетных базах данных. Описанный подход можно применять для решения задачи слияния дублетных записей (как библиографических, так и авторитетных).

Ключевые слова: авторитетный контроль, поиск информации, электронные библиотеки, библиографические базы данных, распределенные информационные ресурсы, авторитетные файлы, задача дискриминации.

Введение

Проблема поиска информации – одна из вечных проблем человеческого сообщества. Чтобы решить проблему доступа к информации, человечество создало библиотеки – универсальную систему хранения «информации и знаний», их систематизации и каталогизации [1].

В настоящее время научно-исследовательский процесс неотделим от использования различных электронных ресурсов доступных в сети Интернет. Значительную часть своего времени научные сотрудники проводят за компьютерами в поиске и анализе информации. Все большую роль в этом процессе начинает играть использование электронных библиотек или электронных каталогов обычных библиотек.

Электронная библиотека – структурированная каталогизированная коллекция разнородных электронных документов (в отличие от печатных изданий, микрофильмов и других носителей),

* Работа выполнена при частичной поддержке РФФИ (проекты № 08-07-00229, 09-07-00277, 10-07-00302), президентской программы «Ведущие научные школы РФ» (грант № НШ-931.2008.9) и интеграционных проектов СО РАН.

снабженная средствами навигации и поиска. Электронные библиотеки – явление достаточно новое, но тем не менее популярное и достойное быть темой обучения студентов [2]. Однако электронные библиотеки сегодня следует рассматривать как множество слабосвязанных сущностей, объединяемых, на первый взгляд, только общим названием.

Под термином «электронная библиотека» могут фигурировать совершенно различные объекты, такие как архивы цифрового контента и наборы программного обеспечения для управления этим контентом. Электронной библиотекой может называться система сетевых сервисов, предоставляющих доступ к цифровому контенту, объединенных единой системой управления этим доступом. Кроме того, некоторые организации, которые берут на себя ответственность не только за исполнение функций управления цифровым контентом и предоставления к нему доступа всем заинтересованным лицам. Такое определение электронной библиотеки полностью соответствует определению традиционной библиотеки как организации в системе, например, министерства культуры [3]. Именно это определение электронной библиотеки наиболее импонирует руководству наших ведущих традиционных библиотек (РГБ, РНБ, ГПНТБ, ГПНТБ СО РАН и др.), поскольку обеспечивает преемственность методов работы в эпоху перехода к цифровому способу обработки и хранения информации. Однако при этом зачастую забывается тот факт, что работа с цифровым контентом для обеспечения наибольшей эффективности от его использования требует совершенно нового подхода к процессу обработки информации. В частности, для работы необходимы персонал, обладающий новыми знаниями и умениями; правила, регламентирующие технологические процессы обработки цифрового контента и организации доступа к нему. При этом технологические процессы, несомненно, должны регламентироваться не региональными, не федеральными, а международными правилами для обеспечения глобальной интероперабельности не только интерфейсов доступа, но и схем и форматов представления цифрового контента (ресурсов).

Задача электронной библиотеки – не только обеспечить многосторонний поиск в каталоге, но и предоставить пользователю найденный ресурс (публикацию, фотографию, описание научного факта и др.), а также дополнительные сведения о нем, например, об авторах, редакторах, библиографии, организации и т. п. Важным фактором электронных библиотек является определение метаданных для описания ресурсов и выделение ключевых видов субъектов и объектов.

В существующих реализациях электронных библиотек, как правило, плохо решается проблема идентификации документов и субъектов, поскольку метаданные рассматриваются только для целей извлечения документов. Это зачастую приводит к тому, что, например, встретив упоминание персоны в одном месте, невозможно точно установить соответствие с ее упоминанием в другом месте.

Решение этой проблемы может быть достигнуто следующим способом. При формировании метаданных того или иного ресурса (при его каталогизации) необходимо использовать авторитетные базы данных (авторитетные файлы), с помощью которых устанавливать конкретные ссылки на персоны.

В настоящее время большинство крупных библиотечных каталогов формируется с применением технологии авторитетного контроля записей, что позволяет упростить работу каталогизаторов и повысить качество библиографических записей [4].

Применение технологии авторитетного контроля записей само по себе решает проблему идентификации персон. Однако иногда связи между авторитетными и библиографическими записями отсутствуют или становятся некорректными (в условиях объединения нескольких каталогов, в каждом из которых используются свои коды авторитетных записей) и далеко не всегда все персоны, имеющие отношение к ресурсу, описываются.

Другой проблемой, связанной с идентификаций персоны, является в целом неудовлетворительное качество авторитетных записей (на этом мы остановимся ниже более подробно), что не позволяет качественно решать ряд задач, связанных с обеспечением научной деятельности.

Однако основной целью данной работы является создание отсутствующих связей между ресурсами и персонами, а также восстановлением утраченных.

Хранение информации

Основной канал пополнения фондов электронных библиотек – перевод в цифровую форму (сканирование) традиционной печатной продукции. Получаемые в результате этого процесса цифровые объекты не содержат никакой новой информации по сравнению со своими материальными оригиналами, а функциональность созданной на их основе электронной библиотеки содержит некоторую ущербность, поскольку при работе с цифровыми объектами человечество уже выработало определенный набор стереотипов, отсутствие которых вызывает дискомфорт [3]. Одним из элементов этого набора является требование наличия взаимных ссылок между цифровыми объектами, проявляющееся, например, в виде гиперсвязей в пользовательских графических интерфейсах просмотра информации. Реализация взаимных ссылок в цифровых документах не представляет большой сложности, однако при этом проявляются специфические моменты. Во-первых, электронный объект с реализованными связями уже не совсем соответствует своему печатному оригиналу. Это уже другой объект. И этот факт должны учитывать все юридические нормы. Во-вторых, внедренные в объект связи должны быть гарантированно актуальными. Никого, например, не интересуют гиперссылки, ссылающиеся на несуществующие документы. Так появляется отличное от традиционных библиотек требование обеспечения ссылочной целостности данных. Это очень жесткое требование, которое тяжело обеспечить даже в хорошо формализованных системах управления базами данных. Результат – новый цифровой объект как самосогласованное хранилище цифрового контента, или база данных цифровых объектов.

С другой стороны, в электронной библиотеке объекты хранения могут содержать информацию, которая не имеет к объектам хранения традиционных библиотек вообще никакого отношения. Речь может идти

- об электронных копиях элементов хранения традиционных архивов;
- об изображениях элементов хранения традиционных музеев;
- о видео-, аудиоинформации, полученной разными способами, например, видеозапись доклада, сделанного на конференции;
- о научных или других фактах;
- и т. д. и т. п.

Заметим, что существование перечисленных объектов регламентируется нормами, о которых традиционные библиотеки не имеют, как правило, никакого понятия. Последний тезис будет проиллюстрирован в разделе о каталогизации.

Каталогизация и метаданные

В традиционных библиотеках каталогизация реализует основную парадигму упорядочивания информации и обеспечения ее поиска по заранее определенным критериям. При каталогизации порождается новый вторичный информационный ресурс как массив стандартизованных описаний основных информационных объектов – элементов учета и хранения в традиционной библиотеке. Создание вторичного информационного ресурса регламентируется некоторыми правилами каталогизации, которые фиксируются в специальных нормативных документах. Первоначально конечным результатом каталогизации первичного объекта было создание каталожной карточки (в электронном каталоге библиографической записи в базе данных), в которой прописывались основные свойства первичного объекта в соответствии с общими правилами. Появление технологий машинного учета и баз данных привело к переводу вторичных информационных массивов традиционных библиотек в так называемые электронные каталоги, которые упростили доступ к вторичным ресурсам, но сохранили некоторую неполноту связей как следствие существующих правил каталогизации. В настоящее время многие библиотеки вынуждены сочетать электронные технологии и бумажные каталоги, что приводит к ограничениям в каталогизации цифровых объектов, характеристики которых отличаются от традиционных библиотечных носителей информации.

Следует еще раз обратить внимание на то, что, с одной стороны, в электронных библиотеках имеют право существовать цифровые объекты, не имеющие аналогов в традиционных библиотеках и, как следствие, не попадающие под действующие правила каталогизации. С другой стороны, развитие пользовательских интерфейсов для доступа к информации требует возможности расширения списка атрибутивной информации, подлежащей вводу при

каталогизации первичных объектов [3]. Например, уже сегодня прослеживается потребность привязки контента к географическим координатам, которая полностью игнорируется действующими правилами и сложившейся практикой каталогизации. Наконец, существует необходимость описания не только информационного контента первичного объекта (заметим, что в основном именно информационный контент первичного объекта описывается в традиционных библиотеках), но и общего контекста существования первичного объекта с фиксацией всех событий в процессе его существования и непосредственной связи с персонами, имеющими к нему отношение.

Подчеркнем, что при попытке описания контекста перестают работать все действующие библиотечные правила каталогизации.

В качестве иллюстрации можно привести попытку создать разумное описание цифрового объекта, который является изображением глиняной таблички, найденной в точке с координатами (x_1, y_1) в момент времени t_1, помещенной в хранилище с координатами (x_2, y_2) в момент времени t_2, сфотографированной в момент времени t_3 на выездной выставке в точке с координатами (x_3, y_3). При этом на глиняной табличке описано событие, имеющее место быть в момент времени t_0 в точке (x_0, y_0). Обязательным требованием к структурированному описанию первичного объекта должно быть требование возможности поиска по всем временным и пространственным характеристикам как контекста так и контента.

Следует заметить, что невозможность создания нужного вторичного ресурса в рамках существующих правил требует переосмысления правил создания метаданных для информационных ресурсов.

Нельзя не отметить еще один момент. Переход к хранению и учету цифровых объектов делает неэффективной существующую парадигму каталогизации и создания массивов вторичных ресурсов. Действительно, необходимость во вторичных ресурсах объяснялась многие века разнородностью носителей первичной информации, необходимостью систематизации первичных ресурсов и организации поиска информации, хотя бы методом перебора каталожных карточек в алфавитном или систематическом каталогах. Переход к электронным каталогам существенно расширил возможности поиска в массивах вторичной информации. Однако современные технологии позволяют, во-первых, внедрять метаданные в первичные объекты, и, во-вторых, организовывать поиск по первичному ресурсу. При этом метаданные образуют с первичным объектом единое целое, а функциональность сервисов доступа к массивам информации не страдает.

Как уже отмечалось, в традиционной библиотеке возможности поиска ограничивались поиском по алфавитному или систематическому каталогам для вторичных информационных ресурсов с прямой ссылкой (указания шифра хранения) на соответствующий первичный ресурс. Использование электронных каталогов расширило поисковые возможности, но сохранило основным типом поиска поиск по предопределенным поисковым атрибутам. Это атрибутивный поиск, именно этот тип поиска является основным в традиционных библиотеках, в том числе и в библиотеках цифровых объектов. Фактически при этом поиск производится только по массивам вторичной информации, оставляя открытым вопрос соответствия последней первичным информационным ресурсам. Заметим, что внедрение атрибутивной информации в первичный цифровой объект, как упоминалось выше, могло бы вместе с отказом от привычной процедуры каталогизации существенно упростить технологии атрибутивного поиска.

Другой возможный тип поиска – поиск по заданным шаблонам – имеет смысл только в массивах первичных цифровых ресурсов. Наконец, поиск с привлечением онтологии является поиском более интеллектуальным, для его реализации требуется дополнительная информация о предметной области, включающая определения терминов, сущностей и связей. Следует отметить, что представление этой дополнительной информации должно соответствовать глобальным договоренностям – международным стандартам, иначе поиск с привлечением онтологии всегда будет ограничен текущей системой, а интероперабельность не будет реализована [5].

Использование публикаций (информационных ресурсов) в научно-исследовательском процессе выдвигает необходимость быстрого ознакомления с содержимым публикации, и аннотация здесь может оказаться недостаточной. В связи с этим должны быть разработаны средства полуавтоматического выделения оглавления и выделения фактов (научных результатов в соответствии с онтологией, понятиями) с обеспечением ссылок на соответствующие разделы документа, а также средства работы с библиографическими ссылками. Достаточно

важными отношениями при составлении метаописания является связь с предметной областью и ее понятиями (концептами).

Правильный авторитетный контроль информационных ресурсов должен давать конкретное указание на персоны с учетом их отношений к данному ресурсу [1]: автор, редактор, персонаж, владелец, рецензент и т. д. Это позволит корректно решать задачу идентификации объектов.

Отметим, что помимо общепринятых описательных метаданных основные сущности электронной библиотеки должны быть снабжены именованными отношениями, из которых можно выделить следующие.

- Входит в состав (ссылка: объект) – данный ресурс является физически или логически частью указанного ресурса.
- Включает (ссылка: объект) – данный ресурс физически или логически включает указанный ресурс.
- Работал (ссылка: субъект; атрибут: время).
- Преподавал (ссылка: субъект; атрибут: время).
- Изображен (ссылка: объект; атрибут: время).
- Ученик (ссылка: субъект).
- Автор (ссылка: объект; атрибут: время).
- Персонаж; (ссылка: объект; атрибут: время).

Авторитетные файлы

Возвращаясь к проблемам идентификации документов относительно персон в них упомянутых, мы приходим к необходимости создания полнофункциональных баз данных описания персон (авторитетных файлов).

Следует отметить, что создание авторитетного файла (базы данных по персонам) – трудоемкий и достаточно дорогой процесс, требующий, как правило, привлечения дополнительной информации из множества источников [6] и пока не до конца поддающийся формализации.

Как видно из табл. 1, на которой показан фрагмент авторитетной записи, значения атрибутов достаточно произвольные и требуют стандартизации.

Таблица 1

Фрагмент авторитетной записи

Номер поля	И1	И2	Значение поля
001			AIvanovVladV2004042963480700
200		1	$aИванов $b В. В. $c биохимия $f 19530130 $g Владимир Владимирович $y Томск
830			$a Образование: в 1975 г. окончил Томский университет, биолого-почвенный факультет, аспирантуру в Томском медицинском институте. $a Ученая степень: в 1975 г. защитил кандидатскую диссертацию. Кандидат биологических наук. $a Трудовая деятельность: работал на кафедре физики сначала инженером, затем ассистентом в Томском медицинском институте. Преподавал курс физики студентам медико-биологического факультета, был руководителем студенческого конструкторского бюро. С 1981 г. работает на кафедре биологической химии в должности ассистента, а с 1984 г. доцент кафедры биохимии и молекулярной биологии Сибирского медицинского университета (Томск). $a Направления научной деятельности: специалист в области биохимии и экспериментального диабета. Изучение влияния активных форм кислорода на секрецию инсулина островками Лангерганса и периферические эффекты гормона. Исследовалось влияние перикисного окисления на рецепцию и деградацию инсулина адипоцитами, метаболизм глюкозы в жировых клетках. $a Научные труды: автор 65 научных публикаций. $a Награды и звания: доцент.

[1] Заметим, что в рекомендациях Dublin Core имеется более 20 квалификаторов персон.

За рубежом наиболее близкие задачи, по-видимому, решаются в рамках проекта VIAF (виртуальный авторитетный файл) Международной федерации библиотечных ассоциаций и учреждений (ИФЛА). Виртуальный авторитетный файл (VIAF) – совместный проект OCLC, Библиотеки конгресса, Немецкой национальной библиотеки и национальной библиотеки Франции для разработки виртуальной комбинации имен авторов из каждого института в единый авторитетный сервис [2].

Целью проекта VIAF является обеспечение возможности автоматического сопоставления и связывания авторитетных записей из различных национальных авторитетных файлов. Изначально VIAF связывал авторитетные файлы имен Библиотеки Конгресса и Немецкой Библиотеки в единый виртуальный авторитетный файл имен. В настоящий момент в проекте участвует 15 организаций и обрабатывается 18 различных типов авторитетных записей. Система проектировалась с учетом возможности обучения для любого числа национальных авторитетных файлов. Алгоритмы сопоставления и верификации имен достаточно сложны и включают данные как из авторитетных, так и библиографических записей с целью создания расширенных авторитетных записей, повышающих точность сопоставления [7].

Наряду с задачей сопоставления и слияния дублирующихся авторитетных файлов возникает задача автоматического связывания библиографических записей с соответствующими авторитетными в условиях объединения нескольких электронных каталогов. Работы по решению последней задачи нам не известны. Этой проблеме и посвящена данная статья.

Авторитетный контроль

Оставив в стороне формирование правильной электронной библиотеки, отвечающей всем задачам идентификации документов и формирования авторитетных записей, рассмотрим вопрос связанный с работой с распределенными библиографическими каталогами.

Предполагается, что записи, с которыми имеем дело, достаточно полны и содержат правильную информацию. Однако и в этом случае при объединении двух и более электронных каталогов неизбежно приходится сталкиваться со следующей проблемой: записи на материалы одного и того же автора содержат разные контрольные номера авторитетной записи (или совсем не содержат их). Такая ситуация возникает вследствие того, что для каждого каталога применяется свой набор авторитетных файлов. В отечественной практике нет примеров (точнее нам они неизвестны), которые бы показали, как можно разрешать данную проблему без участия человека (или с минимальным участием человека). С другой стороны, задача объединения библиотечных каталогов встречается на практике достаточно часто.

В нашей работе мы будем исследовать возможности создания технологии автоматического авторитетного контроля (далее просто ААК). Мы также надеемся, что, имея четкие представления о том, как можно связывать библиографические и авторитетные записи, можно будет применить разработанный подход и к задаче выявления дублетных записей (как авторитетных, так и библиографических). Впервые на решение этой проблемы обратил внимание еще Дж. Солтон в своей ставшей уже классической монографии [8].

Суть задачи автоматического авторитетного контроля заключается в сопоставлении и связывании авторитетных записей (АЗ) и библиографических записей (БЗ) на основе имеющейся в них информации. Для решения этой задачи необходимо разработать алгоритм, позволяющий на основе содержимого определенных полей БЗ и АЗ делать предположение о соответствии либо несоответствии этих записей. Если составить пару из авторитетной и библиографической записей, такую что фамилия и инициалы автора в этих записях совпадают, то все множество таких пар можно разбить на два класса: соответствующие и несоответствующие пары. Рассматривать записи в паре представляется оптимальным еще и потому, что сама по себе, например, дата рождения автора дает не так много информации, как факт совпадения этой информации в АЗ и в БЗ. Такие факты совпадения (или несовпадения) будем учитывать с помощью «факторных» (независимых) переменных. А соответствие (или несоответствие) данной БЗ, рассматриваемой АЗ в целом с помощью «результирующей» (зависимой) переменной.

[2] См. подробнее: http://viaf.org/.

Итак, в качестве отдельного наблюдения выступает пара АЗ-БЗ, характеризующаяся набором факторов (они отражают степень совпадения информации в записях) и значением результирующей переменной (она указывает на принадлежность к одному из двух классов).

Рассматриваемую задачу можно свести к задаче дискриминации [9] наблюдений на два класса при наличии обучающей выборки. Для этого достаточно представить наблюдения точками в многомерном пространстве, определив в этом пространстве понятие расстояния и обучающую выборку.

Под обучающей выборкой понимается набор наблюдений, для которых известны значения факторных переменных и результирующей, т. е. уже известна принадлежность к тому или иному классу.

В качестве расстояния предлагается использовать расстояние Махаланобиса [10], характеризующееся тем, что оно учитывает корреляции между переменными и является инвариантным к масштабу. При этом наряду с расстояниями между двумя отдельными наблюдениями можно рассчитать также расстояние между центроидами двух классов (в качестве центроида будем рассматривать среднее по классу) и расстояние от отдельного наблюдения до центроида класса.

Квадрат расстояния Махаланобиса между двумя точками X и Y, определенными в p-мерном пространстве, можно записать в виде

$$D^2(X,Y) = (X-Y)C^{-1}(X-Y)^T, \quad (1)$$

где X и Y – векторы координат размерности p; C^{-1} – матрица, обратная ковариационной матрице.

Заменив один или оба вектора в формуле (1) на вектор координат центроида первого класса μ_1 или второго μ_2, получим расстояние от точки до класса или расстояние между классами. На практике оценить расстояние Махаланобиса можно, подставив соответствующие оценки средних значений и матрицы ковариации. В качестве оценки матрицы ковариации в формуле (1) будем использовать внутригрупповую матрицу ковариации W, элементы которой находятся по формуле

$$W_{ij} = \frac{1}{n.-2}\sum_{k=1}^{g}\sum_{m=1}^{n_k}\left(X_{ikm}-X_{ik.}\right)\left(X_{jkm}-X_{jk.}\right), \quad (2)$$

где g – число классов; n_k – число наблюдений в k-*м* классе; $n.$ – общее число наблюдений по всем классам; X_{ikm} – величина переменной g для m-го наблюдения в k-*м* классе; $X_{ik.}$ – средняя величина переменной i в k-*м* классе.

Воспользовавшись расстоянием Махаланобиса, можно произвести отбор наиболее информативных признаков (т. е. признаков, позволяющих наиболее четко разделить классы), а также спрогнозировать принадлежность к классу для новых наблюдений (вычислив расстояния до обоих классов и выбрав наиболее близкий класс).

Алгоритм

В общих чертах алгоритм автоматического авторитетного контроля выглядит следующим образом.

1. В момент загрузки библиографической записи в базу данных анализируется поле, в котором указан автор источника и находятся все записи из авторитетной базы данных, с такими же ФИО.
2. Для каждой пары «авторитетная запись – библиографическая запись»:
   а) рассчитываются значения факторных переменных и расстояния до классов (соответствующих и несоответствующих пар);
   б) на основе того, к какому классу пара ближе, принимается решение о ее соответствии либо несоответствии.
3. В случае, если соответствующей признана ровно одна пара, в библиографической записи указывается код авторитетной записи, если больше одной – запись отправляется на до-

полнительный контроль с привлечением специалистов, в противном случае никаких отметок не делается и запись попадает в базу.

Описанный алгоритм можно применять и для записей, которые уже хранятся в базе данных, но не получили контрольных номеров соответствующих авторитетных записей.

Описание эксперимента

Исходные данные. Эксперимент проводился на системе, включающей:

1) библиографическую базу данных, более 300 000 записей в формате RUSMARC;

2) авторитетный файл (АФА), около 10 000 записей в формате RUSMARC/Authorities.

Примеры авторитетной и библиографической записей приводятся в табл. 1 и 2. На основе АФА был составлен список фамилий с инициалами, соответствующих сразу двум и более авторитетным записям.

Затем полученный список из 42 элементов случайным образом был поделен поровну. Первая часть использовалась для обучающей выборки, вторая – для тестовой. Кроме наличия однофамильцев, к авторитетным записям предъявлялось требование полноты: наличие информации о дате рождения, географических и профессиональных дополнений, аннотации (наличие полей 001, 200 ($a, $b, $c, $f, $y), 830$a).

При этом в рамках эксперимента намеренно игнорировалась информация о расшифровке инициалов (200$g) для увеличения области совпадения авторитетных и библиографических записей и, следовательно, объема обучающей выборки. Разумеется, рабочий алгоритм не будет игнорировать эту информацию, что позволит повысить его точность.

В свою очередь, библиографическая запись также удовлетворяет условиям:

1) содержит указание на авторитетную запись (наличие поля 701$3), что позволяет ответить на вопрос о принадлежности записи;

2) требование полноты к БЗ не предъявляется так строго, однако очевидно, что при отсутствии сразу всех подполей $c, $f и $p вряд ли удастся правильно соотнести записи, поэтому в эксперименте рассматривались только те библиографические записи, для которых определены значения как минимум двух переменных.

Таблица 2

Фрагмент библиографической записи

Номер поля	И1	И2	Значение поля
001			61/Н340-682478
700		1	$a Шилов $b Б. В. $g Борис Владимирович $c цитолог $f 19710323 $3 AShilov_BoriB2003100663480700
701		1	$a Иванов $b В. В. $g Владимир Владимирович $c биохимик $f 19530130 $3 AIvanovVladV2004042963480700 $ркафедра биохимии и молекулярной биологии СГМУ
701		1	$a Казанский $b В. Е.
712	0	2	$a Сибирский медицинский университет $c Томск

Факторы. Рассмотрим подробнее переменные, по которым осуществляется связывание записей. Результирующая переменная out отвечает за принадлежность БЗ данному автору (другими словами за принадлежность наблюдения к одной из двух групп). Остальные семь переменных в нашем эксперименте являются факторными и указывают на степень соответствия информации о годах жизни автора, профессиональной деятельности, географическом положении и месте работы (табл. 3).

Таблица 3

Переменные

Переменная	Значение	Код	Комментарий	АЗ	БЗ
out (соответствие)	не соответствует	1	точное совпадение	001	701$3
	соответствует	2			
birth (дата рождения)	не совпадает	1	совпадение с точностью до года	200$f	701$f
	не указана дата	2			
	совпадает	3			
death (дата смерти)	не совпадает	1	совпадение с точностью до года	200$f	701$f
	не указана дата	2			
	совпадает	3			
addition (профессиональное дополнение)	нет совпадений	1	совпадение усеченных форм	200$c	701$c
	не указано	2			
	одно совпадение	3			
	два совпадения	4			
	…	…			
place1 (географическое)	не совпадает	1	совпадение усеченных форм	200$y	712$c
	не указано место	2			
	совпадает	3			
place2 (географическое)	нет совпадений	1	вхождение усеченных форм	200$y	712$a
	не указано место	2			
	хотя бы одно совпадение	3			
work1 (место работы)	не совпадает	1	совпадение усеченных форм	830$a	701$p
	не указано место	2			
	совпадает	3			
work2 (место работы коллектива)	нет совпадений	1	вхождение усеченных форм	830$a	712$a
	не указано место	2			
	хотя бы одно совпадение	3			

Вычисляя значения перечисленных переменных для записей, приведенных в примере, а затем проделав аналогичное сравнение для всех пар из обучающей выборки, получим исходные данные эксперимента, фрагмент которых приведен в табл. 4.

Таблица 4

Фрагмент исходных данных

Результирующая переменная out	Факторные переменные						
	addition	birth	death	place1	place2	work1	work2
2	3	3	2	3	1	1	3
2	1	3	2	2	2	2	2
1	1	1	2	3	2	2	2

На основе табл. 4 и будут строиться дальнейшие статистические расчеты. В первой строке даны значения переменных для приведенной пары записей. Заметим, что переменные place2 и workl для этой пары равны 1, т. е. выявлено несоответствие. Введение переменной place2 было основано на том факте, что в библиографических записях при заполнении поля 712$a иногда в скобках указывают место расположения организации, кроме того, в названиях некоторых организаций содержится указание на географическое положение (например, можно найти подстроку «Томск» в названии «Томский государственный университет»). Переменная workl отвечает за указание места работы автора в аннотации. В данном случае 701$p='кафедра биохимии и молекулярной биологии СГМУ', тогда как 830$a содержит подстроку 'кафедры биохимии и молекулярной биологии Сибирского медицинского университета'. Сопоставить эти строки можно, если использовать словари. Если словарь сокращений не привлекать, то получим значение workl = l, хотя очевидно, что налицо совпадение информации. В то же время пара записей относится к классу соответствующих (переменная out принимает значение 2).

Анализ средних значений. Факторные переменные, используемые в работе, измеряются в интервальной шкале. Это позволяет вычислять такие статистические характеристики, как среднее и ковариация, однако не подчиняются нормальному распределению, что исключает применение параметрических критериев. Для проверки гипотезы значимости различия в средних по группам использовался критерий Хи-квадрат. Переменные, для которых принималась гипотеза об отсутствии различий (при уровне значимости 0,01), исключались из работы (табл. 5). Как видно из табл. 5, переменную place2, уровень значимости для которой больше 0,01, можно исключить из рассмотрения.

Таблица 5

Анализ различий в средних значениях

Переменная	$\overline{x}_1$	$\overline{x}_1$	Min	Max	χ^2	*p*-value	Разница в средних
addition	1,122	1,924	1	3	128	$< 2{,}2*10^{-16}$	значима
birth	1,076	2,912	1	3	386	$< 2{,}2*10^{-16}$	значима
death	2	2,063	2	3	14	$1{,}6*10^{-4}$	мало значима
placel	1,855	2,151	1	3	44	$3{,}3*10^{-10}$	значима
place2	1,786	1,874	1	3	9	0,01216	не значима
workl	1,794	1,937	1	2	15	$1{,}2*10^{-10}$	мало значима
work2	1,786	1,899	1	3	13	0,0014	мало значима

Отбор факторных переменных. Воспользовавшись расстоянием Махаланобиса можем произвести отбор наиболее информативных признаков. Для этого на каждом шаге отбираем по одной переменной, дающей наибольшее расстояние между центроидами классов в сочетании с уже выбранными (табл. 6). Так, на первом шаге будет выбрана переменная birth (отмечена *), а на втором – переменные birth и work2, и т. д. Чем раньше включаются переменные, тем больше информации в них содержится. Очевидно, что переменная workl, включенная последней, наименее информативна.

Таблица 6

Процесс отбора факторных переменных

Шаг отбора	Переменная	$D^2(g1\backslash g2)$
шаг 1	addition	0,64
	birth *	3,37
	death	0,004
	placel	0,087
	workl	0,021
	work2	0,013
шаг 2	addition	52,84
	death	44,98
	placel	45,15
	workl	45,26
	work2 *	54,63
шаг 3	addition *	61,66
	death	54,65
	placel	54,99
	workl	54,69
шаг 4	death *	62,49
	placel	61,99
	workl	61,73
шаг 5	placel *	62,91
	workl	62,55

В результате процедуры отбора был получен список факторных переменных в порядке их значимости для дискриминации. Пользуясь этой информацией, можно исключить наименее информативные переменные и сократить время работы алгоритма.

Проверка качества дискриминации. С помощью расстояния Махаланобиса можно прогнозировать принадлежность наблюдения к одной из групп. Для этого достаточно рассчитать расстояния до центроидов обоих классов, подставив в (1) координаты этого наблюдения, координаты центроида класса и матрицу внутригрупповой ковариации, рассчитанную по формуле (2). После этого следует выбрать в качестве прогноза тот класс, расстояние до которого наименьшее.

Проведя расчеты для тестовой выборки (табл. 7), наблюдения которой не использовались при вычислении параметров алгоритма, получим так называемую матрицу классификации, на основе которой можно рассчитать долю правильно классифицированных объектов и оценить точность прогноза.

Таблица 7

Пример классификации

Факторные переменные						Расстояния до классов		Прогноз класса	Значение переменной out
addition	birth	death	placel	work1	work2	до первого $D^2(X, X^1)$	до второго $D^2(X, X^2)$		
3	3	2	3	1	3	124,5	27,8	2	2
1	3	2	2	2	2	65,6	2,6	2	2
1	1	2	3	2	2	7,8	76,9	1	1

В табл. 8 приведены результаты прогнозирования принадлежности пар записей к классу соответствующих и несоответствующих для разных наборов факторных переменных.

Таблица 8

Проверка на тестовой выборке

Набор переменных	Классифицированы			
	ошибочно		правильно	
birth	13	2,1 %	611	97,9 %
birth, work2	2	0,32 %	622	99,68 %
birth, work2, addition	2	0,32 %	622	99,68 %
birth, work2, addition, ..., workl	2	0,32 %	622	99,68 %

Как видно из табл. 8, процент ошибок при обработке тестовой выборки достаточно мал. Можно заметить, что применение оценки всего по двум переменным birth и work2 дает тот же результат, что и применение большего числа переменных. Кроме того, при ближайшем рассмотрении в ошибочно классифицированных записях оказались неправильно указанные коды авторитетных файлов, т. е. на самом деле эти записи были классифицированы верно.

Заключение

В работе представлен алгоритм автоматического авторитетного контроля, позволяющий сделать вывод о связи библиографической и авторитетной записей без участия человека. В результате эксперимента, проведенного на библиографических и авторитетных базах данных, были получены достаточно обнадеживающие результаты, позволяющие утверждать, что данная техника имеет право на существование. Был проведен статистический анализ факторов, указывающих на соответствие записей, их ранжирование по степени информативности и проверка работы алгоритма на тестовой выборке. Рассматриваемый алгоритм также может быть полезен для проверки уже связанных записей и выявления ошибок. Для адекватной работы необходимо периодическое уточнение его параметров с ростом базы данных. Устойчивость алгоритма к ошибкам дискриминации можно повысить за счет разработки методики привлечения неиспользованной информации, например, информации о соавторах и тематике работы автора.

Основным недостатком предложенного алгоритма является предположение о том, что вновь поступающие в систему записи аналогичны уже имеющимся. Так, например, если в авторитетной базе данных нет записей с одинаковыми датами рождения, то алгоритм может оказаться несостоятельным в случае, когда необходимо дискриминировать несоответствующую пару АЗ-БЗ с совпадающей датой рождения. Это происходит из-за того, что параметры алгоритма зависят от характера обучающей выборки.

Список литературы

1. *Федотов А. М., Барахнин В. Б.* Проблемы поиска информации: история и технологии // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2009. Т. 7, вып. 2. С. 3–17.

2. *Земское А. И., Шрайберг Я. Л.* Электронные библиотеки: Учеб. пособие для студентов ун-тов и вузов культуры и искусств и др. учеб. заведений,. 3-е изд., испр. и доп. М.: ГПНТБ России, 2004. 130 с.

3. *Жижимов О. Л., Мазов Н. А., Федотов А. М.* Некоторые заметки об эволюции цифровых репозитариев традиционных библиотек к полнофункциональным электронным библиотекам // Вестн. Владивостокского гос. ун-та экономики и сервиса. Территория новых возможностей. №3 (7). 2010. С. 55–63.

4. *Мешечак Н. А., Шамардина Л. А., Карауш А. С.* Опыт создания и использования авторитетных записей на Томских ученых-медиков в научно-медицинской библиотеке Сибирского медицинского университета / Современные пользователи автоматизированных ин-

формационно-библиотечных систем: проблемы обслуживания, изучения и обучения: материалы VI и VII науч.-практ. конф. СПб.: РБА, 2006. С. 158–161.

5. *Шокин Ю. Ж., Федотов А. М., Барахнин В. Б.* Проблемы поиска информации. Новосибирск: Наука, 2010. 198 с.

6. *Ковалева А. М.* Авторитетный файл Имя лица // Библиотечное краеведение в информационном пространстве региона Барнаул, 2008. С. 172–178.

7. *Rick B., Hengel-Dittrich Ch., O'Neill E. T., Tillett B.* VIAF (Virtual International Authority File): Linking the Deutsche Nationalbibliothek and Library of Congress Name Authority Files // International Cataloging and Bibliographic Control. 2007. Vol. 36(1). P. 12–19.

8. *Солтон Дж.* Динамические библиотечно-информационные системы: Пер. с англ. М.: Мир, 1979.

9. Факторный, дискриминантный и кластерный анализ: Пер. с англ. / Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка М.: Финансы и статистика, 1989. 215 с.

10. *Афифи А., Эйзенс С.* Статистический анализ. Подход с использованием ЭВМ: Пер. с англ. М.: Мир, 1982. 488 с.

Материал поступил в редколлегию 15.02.2011

**A. M. Fedotov, O. L. Zhizhimov, A. A. Knyazeva, O. S. Kolobov,
N. A. Mazov, I. Yu. Turchanovsky, O. A. Fedotova**

PROBLEMS OF AUTHORITY CONTROL FOR DISTRIBUTED DIGITAL LIBRARIES AND BIBLIOGRAPHIC DATABASES

Problems of authority control and identification of information resources are discussed. The algorithm of automatic authority control to link bibliographic and authority records is presented. The analysis of the factors affecting correspondence of records and calculation of algorithm parameters were carried out using a statistical experiment on bibliographic and authority databases. The described approach can be applied to solve the problem merging duplicate records (both bibliographic and authority).

Keywords: authority control, information retrieval, digital library, bibliographic databases, distributed information resources, authority files, discrimination problem.