На правах рукописи

Князева Анна Анатольевна

АВТОМАТИЧЕСКОЕ СВЯЗЫВАНИЕ ЗАПИСЕЙ БИБЛИОГРАФИЧЕСКИХ БАЗ ДАННЫХ НА ОСНОВЕ УНИФИЦИРОВАННЫХ ПОИСКОВЫХ ПРИЗНАКОВ

05.25.05 – «информационные системы и процессы»

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Новосибирск - 2014

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте вычислительных технологий Сибирского отделения Российской академии наук, г. Новосибирск

Научный руководитель	член-корреспондент РАН, доктор физико-математических наук **Федотов Анатолий Михайлович**
Официальные оппоненты	**Бобров Леонид Куприянович**, доктор технических наук, доцент, профессор Новосибирского государственного университета экономики и управления «НИНХ», г. Новосибирск
	Тихомиров Илья Александрович, кандидат технических наук, старший научный сотрудник Института системного анализа РАН, г. Москва
Ведущая организация	Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук (ГПНТБ СО РАН), г. Новосибирск

Защита состоится «16» сентября 2014 г. в 16.30 на заседании диссертационного совета ДМ 003.046.01 на базе Федерального государственного бюджетного учреждения науки Института вычислительных технологий Сибирского отделения Российской академии наук по адресу: 630090, г. Новосибирск, пр. Академика Лаврентьева, 6 (dsovet@ict.nsc.ru).

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Института вычислительных технологий Сибирского отделения Российской академии наук http://www.ict.nsc.ru/sitepage.php?PageID=17

Автореферат разослан 2014г.

Ученый секретарь
диссертационного совета ДМ 003.046.01
к. ф.-м. н., доцент  Лебедев А.С.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Вопросы идентификации объектов реального мира в библиографических и других данных становятся все более актуальными в связи с постоянным ростом объемов информации, а также развитием наукометрических исследований. В качестве объектов, которые необходимо идентифицировать, могут выступать персоны, организации, географические объекты и т.п. Необходимость идентификации объектов реального мира в библиографических данных рассматривалась бельгийским социологом Полем Отле[1] еще в конце XIX века. Такую идентификацию можно осуществить с помощью установления связи со специальной авторитетной/нормативной[2] записью, однозначно указывающей на данный объект. В качестве такой записи может выступать любой структурированный документ, содержащий информацию об объекте и удовлетворяющий требованиям, разработанным международными организациями[3].

В настоящее время различные системы учета публикаций, например, Scopus, Web of Science, SCIENCE INDEX (на базе РИНЦ) используют различные идентификационные коды авторов[4]. При этом один автор может упоминаться в различных базах с разными кодами. Таким образом, является актуальной задача связывания этих кодов между собой, а также их привязки к базам авторитетных записей имен лиц, которые используются в библиотеках. Развитие данного подхода способно улучшить качество наукометрических показателей за счет учета публикаций автора, учтенных в различных базах данных.

Также большое значение имеет задача идентификации объектов реального мира в библиографических записях, которая является задачей авторитетного контроля электронного каталога. В настоящее время установление связей с авторитетными записями производится каталогизаторами вручную и однократно (только на этапе создания библиографической записи). Как следствие, при объединении ресурсов нескольких библиотек

[1] Отле П. Библиотека, библиография, документация

[2] Далее в рамках данной работы используется термин *авторитетная запись*

[3] Функциональные требования к авторитетным данным : заключительный отчет рабочей группы Международной федерации библиотечных ассоциаций и учреждений (IFLA)

[4] ORCID, ResearcherID, SPIN-код соответственно

возникают задачи выявления дубликатов записей и восстановления утерянных или отсутствующих связей между авторитетными и библиографическими записями. Решению этих задач в автоматическом режиме (без участия человека) и посвящена данная работа.

В настоящее время существуют различные АБИС[5], такие как «Руслан», «ИРБИС», «БУКИ», «Нева» и др. Большинство из них позволяет осуществлять авторитетный контроль электронного библиотечного каталога. При этом под *авторитетным контролем* понимается процесс поддержания единообразия форм авторитетных заголовков, определяющих одно и тоже лицо, организацию, предмет и так далее в библиографическом файле, контроль за адекватностью присвоения предметных рубрик и индексов библиотечно-библиографических классификаций документам, а также контроль за последовательным соблюдением принципов, методик, инструкций и правил по представлению поисковых признаков.

Сам процесс установления связи между авторитетными и библиографическими записями в рамках существующих АБИС выполняется каталогизатором. С одной стороны, эксперт может устанавливать достаточно надежные связи между записями, за счет привлечения дополнительной информации, не содержащейся в самих записях. С другой стороны, такой подход предполагает большой объем ручного труда, сложность ретроспективного анализа и множество «упущенных» связей между записями.

Задача создания автоматического авторитетного контроля была впервые поставлена в рамках данной работы. Для решения данной задачи было решено использовать принципы и методики *связывания записей* (record linkage) в применении к библиографическим данным. В настоящее время существует множество работ в области связывания записей или выявления дубликатов зарубежных[6] и отечественных[7] авторов. Однако, среди них нет работ, в которых рассматривается задача связывания записей разной

[5]АБИС – Автоматизированная библиотечно-информационная система

[6]William E. Winkler, Mikhail Y. Bilenko, Jeremy A. Hylton, Mauricio A. Hernández и Salvatore J. Stolfo, Peter Christen и Tim Churches, Pawel Jurczyk и др.

[7]Серебряков В.А., Антопольский А.Б., Каленкова А.А., Атаева О.М., Шиолашвили Л.Н., Чудин А., Зелепухина В.А., Пинжин А.Е., Тарасов С. и др.

структуры в форматах семейства MARC[8], в которых, на сегодняшний день, представлена практически вся библиографическая информация.

Существуют также различные системы связывания записей, такие как MARLIN, TAILOR, Febrl и др. Данные системы нацелены на работу по связыванию адресов, информации о пациентах или библиографических ссылок одной строкой. Применить данные системы к решению поставленной задачи не представляется возможным, поскольку они не поддерживают работу с записями в MARC-форматах.

Задача выявления и слияния нескольких авторитетных записей для одного автора решалась в рамках проекта *VIAF*[9] Международной федерации библиотечных ассоциаций и учреждений (IFLA). Целью проекта является обеспечение возможности автоматического сопоставления и связывания авторитетных записей из различных национальных источников. Подход, применяемый в проекте VIAF, не может быть применен к решению нашей задачи, поскольку он основан на экспертной оценке значимости признаков, участвующих в сопоставлении.

Цель диссертационной работы. Разработать технологию автоматического авторитетного контроля, позволяющую устанавливать связи между библиографическими записями, относящимися к одному объекту реального мира.

Задачи. Реализация данной цели предполагает решение следующих задач:

1. Сформулировать и проанализировать основные требования к процедуре связывания, исходя из особенностей библиографических данных;
2. Разработать модель связывания библиографических записей;
3. Разработать технологию связывания авторитетных и библиографических записей, относящихся к одному и тому же автору;
4. Сформулировать рекомендации по наполнению библиографических баз данных для повышения качества связывания.

[8]Machine-Readable Cataloging(англ.) – формат машиночитаемой каталогизационной записи
[9]The Virtual International Authority File (англ.) – Виртуальный авторитетный файл

На защиту выносятся:

- Аналитическая и концептуальная модели связывания библиографических записей, основанные на методах машинного обучения;
- Технология идентификации библиографических данных, позволяющая связывать авторитетные и библиографические записи в формате RUSMARC, относящиеся к одному автору;
- Ранжированный набор признаков и весовые коэффициенты, полученные на основе реальных данных;
- Разработанный программный комплекс «ААК-персоны», позволяющий проводить обучение на основе библиографических данных и устанавливать связи между библиографическими и авторитетными записями в формате RUSMARC без участия эксперта.

Научная новизна. На основе общих принципов связывания записей впервые сформулированы требования к системе *автоматического авторитетного контроля (ААК)*, позволяющей идентифицировать объекты реального мира в библиографических записях без участия эксперта. Предложен набор моделей связывания библиографических записей в условиях неполноты данных и взаимозависимости признаков. Модель предусматривает возможность использования информации об уже установленных связях. Реализован алгоритм обучения системы на основе набора пар записей с отметками о принадлежности к одному из двух классов: пар записей с упоминанием одного объекта и пар записей с упоминанием разных объектов реального мира. Предложена процедура отбора наиболее значимых признаков для связывания. На основе сформулированных требований и в соответствии с предложенными моделями разработана технология автоматического авторитетного контроля персон в библиографических записях в формате RUSMARC.

Методы исследований. В работе применялись методы классификации, непараметрической статистики, нечеткого сопоставления строк и принципы связывания записей.

Достоверность результатов подтверждается проведенным экспериментальным исследованием по связыванию записей из библиографической

и авторитетной баз данных Некоммерческого партнерства по содействию медицинским библиотекам «МедАрт»[10], а также использованием результатов, что подтверждено соответствующими документами.

Практическая значимость. Результаты диссертационной работы могут использоваться для решения задач автоматического связывания библиографических записей. В частности, предлагаемая технология позволяет организовать ААК библиографических данных с учетом особенностей конкретной коллекции и информации об уже установленных связях. В работе представлены рекомендации по наполнению библиографических баз данных, позволяющие повысить качество связывания записей. Предлагаемый подход является достаточно общим и может быть перенесен на задачу выявления дубликатов среди записей, как библиографических, так и авторитетных.

Представление работы. По теме диссертации были сделаны сообщения и доклады на научно-практических конференциях: DICR (Российская конференция с международным участием «Распределённые информационные и вычислительные ресурсы», г. Новосибирск, 2010, 2012 гг.), Современные проблемы математики, информатики и биоинформатики (Международная конференция «Современные проблемы математики, информатики и биоинформатики», посвященная 100-летию со дня рождения члена-корреспондента АН СССР Алексея Андреевича Ляпунова, г. Новосибирск, 2011), ”RCDL” (Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», г. Переславль-Залесский, 2012, г. Ярославль, 2013), ”MTE” (Всероссийская конференция молодых ученых «Материаловедение, технологии и экология в третьем тысячелетии», г. Томск, 2012), «Корпоративные информационно-библиотечные системы: технологии и инновации» (XI международная конференция и выставка, г. Санкт-Петербург, 2013). Работа выполнялась при финансовой поддержке Министерства образования и науки Российской Федерации (грант №07.514.11.4130[11]).

[10]Объем коллекций около 300 тысяч и 10 тысяч записей соответственно

[11]Разработка принципов и программных средств виртуальной интеграции распределённых источников данных на основе международных стандартов для создания масштабных информационных инфраструктур (шифр «2012-1.4-07-514-0022-004»).

Реализация и внедрение результатов работы. Разработанные в диссертации методы и алгоритмы внедрены и использованы при выполнении Государственных контрактов в ИВТ СО РАН, а также в рабочем процессе Некоммерческого партнерства по содействию медицинским библиотекам «МедАрт», Научно-медицинской библиотеки Сибирского государственного медицинского университета и Ленинградской областной универсальной научной библиотеки, что подтверждено актами о внедрении, прилагаемыми к диссертационной работе.

Личный вклад автора. Работы по теме диссертации выполнены в Томском филиале Института вычислительных технологий (ИВТ) СО РАН автором совместно с ведущим инженером Института сильноточной электроники (ИСЭ) СО РАН Колобовым О.С. Все результаты, включенные в диссертацию, получены автором лично или в неделимом соавторстве. Автором были предложены модели и технология ААК, а также проведена статистическая обработка массивов данных, полученных в ходе эксперимента, проведенного совместно с Колобовым О.С.

Публикации. По теме диссертации опубликовано 10 печатных работ (объемом 9,8/7,9 печатных листов), в том числе 3 статьи [1–3] в изданиях, рекомендованных ВАК для представления результатов кандидатских диссертаций (в скобках в числителе указан общий объем публикаций, в знаменателе – объем, принадлежащий лично автору). Основные результаты диссертации содержатся в работах [1–9] список которых приведен в конце автореферата.

Структура и объем диссертации. Диссертация состоит из введения, 3-х глав, заключения и 7-ми приложений. Объем диссертации составляет 147 страниц, включая основное содержание, список литературы и приложения. Список литературы содержит 140 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность проблемы *автоматического связывания библиографических записей*. Формулируются цель и задачи работы. Приведены требования к системе ААК: возможность выявления значимых признаков и присвоение им соответствующих весов без привле-

чения экспертов; учет косвенной информации в процессе сопоставления записей и возможность работы с неполными данными.

В первой главе приводится краткий обзор развития авторитетного контроля библиотечного каталога и используемая терминология. В рамках работы используется определение авторитетной записи, предложенное разработчиками российского коммуникативного формата[12]. *Авторитетная запись (АЗ)* – это машиночитаемая запись, исходным элементом которой является принятая точка доступа для имени лица, наименования организации, произведения, торговой марки, издателя/типографа, предметной рубрики, географического названия, формы, жанра или физических характеристик, в форме, установленной каталогизирующим агентством. В дополнение к принятой точке доступа запись может содержать: информационные примечания; запись всех вариантных и связанных точек доступа, от которых созданы ссылки; принятые точки доступа на другом языке или в другой графике; информацию о классификационных индексах и истории объекта записи; примечания об источнике информации и т.п.; информацию о библиографическом учреждении, ответственном за запись; постоянный идентификатор записи и международные стандартные идентификаторы объектов, описанных в записи.

Таким образом, основной элемент *авторитетной записи* - это авторитетный заголовок, то есть унифицированная формулировка (имени индивидуального или коллективного автора, предметной рубрики или индекса библиотечно-библиографической классификации).

Также, в первой главе кратко описываются этапы автоматизации работы библиотек и место авторитетного контроля в АБИС. Далее анализируются подходы и технологии применяемые для решения задачи связывания записей в различных областях. Рассмотрены возможные варианты решения типичных задач, возникающих в процессе связывания записей: *нормализации*, *составления пар*, *сравнения* на уровне полей и *вынесения решения* о соответствии записей друг другу.

[12]Российский коммуникативный формат представления библиографических и авторитетных/нормативных записей в машиночитаемой форме (http://www.rusmarc.ru)

В работе приводится анализ существующих систем с точки зрения следующих критериев: 1) отказ от эмпирических правил для принятия решения о соответствии записей; 2) отсутствие требования независимости признаков; 3) связывание записей разной структуры; 4) работа с записями в форматах семейства MARC; 5) возможность работы с неполными данными; 6) учет информации об уже установленных связях в массиве данных. Как показал проведенный анализ, в настоящее время не существует системы, которая отвечала бы всем приведенным требованиям.

В заключении к первой главе приводится вывод о необходимости разработки модели связывания записей и технологии автоматического авторитетного контроля, которые учитывают особенности предметной области и удовлетворяют перечисленным критериям.

Во второй главе предлагается набор моделей связывания записей: аналитическая и концептуальная модели, а также процедурные модели для каждого из функциональных блоков системы. Приводится описание технологии автоматического авторитетного контроля, которая позволяет связывать библиографические записи с авторитетными записями имен лиц.

Аналитическая модель связывания записей была разработана в терминах «Функциональных требований к библиографическим записям»[13]. В соответствии с данными требованиями, публикация, персона и организация рассматриваются как некоторые объекты, а информация об этих объектах записывается в значениях атрибутов, из которых и состоит запись. Соответствие записей друг другу означает, что в них описывается один и тот же объект реального мира, а сопоставлять записи возможно лишь по набору значений атрибутов.

Итак, пусть даны две коллекции записей A и B. Допустим $\alpha(g)$ – запись из коллекции A, описывающая некоторый объект g; $\beta(f)$ – запись из коллекции B, описывающая объект f. Объекты принадлежат некоторому общему множеству (например, всех авторов мира) $g \in G, f \in G$. Множество пар записей, описывающих один и тот же объект реального мира g будем обозначать как $M(g)$:

$$M(g) = \langle \alpha(g), \beta(f) \rangle; g = f; \alpha(g) \in A; \beta(f) \in B. \quad (1)$$

[13] Functional requirements for bibliographic records, FRBR - исследование в рамках IFLA

Если объединить множества, построенные для каждого объекта из G, получим общее множество соответствующих пар записей M. Дополнение множества $M(g)$, которое будем обозначать как $U(g)$, представляет пары записей, описывающие различные объекты:

$$U(g) = \langle \alpha(g), \beta(f) \rangle; g \neq f; \alpha(g) \in A; \beta(f) \in B. \tag{2}$$

Аналогично, объединив $U(g)$ для всех возможных объектов, получим множество несоответствующих пар записей U.

Поскольку структура записей α и β может различаться, необходимо выработать правила для сопоставления пары записей. Зададим правила сравнения записей c_j, $j = \overline{1, K}$. Каждое правило состоит из функции сравнения и перечисления полей авторитетной и библиографической записей, которые необходимо сравнить. Функция сравнения реализует один из методов сравнения текстовых строк, это может быть строгое или нечеткое сравнение. Поля записей можно выбирать по одному из каждой записи или группой, если интересующая информация записана сразу в нескольких полях. Результат сравнения предполагает следующие варианты: несовпадение информации, пропуск данных, полное или частичное совпадение информации. Результат применения каждого из правил выражается числом, и таким образом осуществляется переход от качественных характеристик (значений текстовых полей) к количественным значениям.

После того, как было проведено сравнение информации из пары записей в соответствии с принятыми правилами, результат этого сравнения можно представить как точку в пространстве признаков размерности K, то есть $\gamma = (X_1, \ldots, X_K)^T$. Здесь X_j – результат применения правила сравнения c_j, который является оценкой соответствия определенной информации в записях. Эта оценка выражается в заранее заданных градациях, которые могут быть различными для разных правил. Для решения задачи связывания записей необходимо построить решающую функцию

$$D(\gamma) = \begin{cases} 1, & \langle \alpha(g), \beta(f) \rangle \in M, \\ 0, & \langle \alpha(g), \beta(f) \rangle \in U, \end{cases} \tag{3}$$

служащую оценкой истинного статуса соответствия объектов

$$s(g,f) = \begin{cases} 1, & g = f, \\ 0, & g \neq f, \end{cases} \tag{4}$$

на основе имеющегося набора прецедентов. Прецеденты – это пары $\langle \alpha(g), \beta(f) \rangle$ с известным статусом $s(g,f)$, из которых составляется обучающая выборка. Представим обучающую выборку как два непересекающихся множества точек в пространстве признаков. Первое множество Γ^M объединяет те пары записей, которые описывают один объект, второе множество Γ^U включает пары, описывающие различные объекты.

Задача отнесения новой пары записей к одному из классов M и U может быть сведена к задаче классификации на основе вычисления некоторого расстояния до множеств Γ^M и Γ^U. В рамках данной работы было использовано расстояние Махаланобиса, которое учитывает возможную взаимозависимость признаков и инвариантно к масштабу.

Квадрат расстояния Махалонобиса до центроида класса M:

$$Dist^2(\gamma, \mu^M) = (\gamma - \mu^M) W^{-1} (\gamma - \mu^M)^T, \tag{5}$$

где γ - вектор значений признаков $\gamma = (X_1, \ldots, X_K)^T$;
μ^M - центроид класса M;
W^{-1} - матрица, обратная внутригрупповой матрице ковариации.

Расстояние до центроида класса U рассчитывается аналогично. В качестве центроида выступает вектор арифметических средних признаков:

$$\mu_i^M = \frac{1}{n^M} \sum_{k=1}^{n^M} X_{ik}^M, \tag{6}$$

где μ_i^M - i-я компонента вектора μ^M, X_{ik}^M - значение i-й компоненты вектора $\gamma_k \in \Gamma^M$, $k = \overline{1, n^M}$.

Элементы матрицы W^{-1}, которые вычисляются на основе обучающей выборки, можно рассматривать в качестве весовых коэффициентов, отражающих степень важности того или иного признака (или отдельного правила сопоставления записей). Также, можно ранжировать признаки по их вкладу в расстояние Махаланобиса между центродами двух классов, что позволяет выделить наиболее значимые для связывания правила.

В качестве критерия для построения решающей функции можно предложить минимизацию числа ошибок классификации пар из тестовой выборки по набору правил сравнения $\{c_j\}$, $j = \overline{1, K}$:

$$\min_{\{c_j\}} \sum_{i=1}^{N} I\{D(\gamma_i) \neq s(g, f)\}, \tag{7}$$

где I – индикаторная функция, γ_i – вектор значений признаков для i-й пары записей из тестовой выборки, $i = \overline{1, N}$.

Очевидно, различные наборы правил будут давать разное число ошибок. После проведения проверки качества классификации на основе тестовой выборки следует определить набор, который позволяет добиться наименьшего числа ошибок.

В качестве применения предложенной модели была рассмотрена задача идентификации персон, упоминаемых в электронном каталоге библиотеки, или задача *автоматического авторитетного контроля* имен лиц. Таким образом, в качестве коллекции B выступает база библиографических записей, содержащая описания публикаций, а в качестве коллекции A – база авторитетных записей имен авторов.

Концептуальная модель связывания записей и процедурные модели функциональных блоков системы связывания приводятся в тексте работы. На основе предложенного набора моделей была разработана технология автоматического авторитетного контроля имен лиц. В рамках описания данной технологии указываются задействованные поля записей и правила для их сравнения.

В третьей главе рассматриваются программные средства, позволяющие выполнить исследование проблемы автоматического авторитетного контроля имен авторов. В частности, рассматривается *программный комплекс «AAK-персоны»*, предназначенный для связывания библиографических записей в формате RUSMARC с авторитетными записями в формате RUSMARC/Authorities. Описана архитектура программного комплекса и основные этапы его работы.

Программные средства, входящие в состав комплекса: 1) базы библиографических данных, доступные по протоколу Z39.50[14], 2) консольный клиент *aak* для обращения к базам и вычисления значений признаков, 3) модуль статистического анализа *stat*.

В работе используются базы данных в формате RUSMARC, доступные через Z39.50 интерфейс и поддерживающие стандартный набор атрибутов.

Консольный клиент *aak* является центральный модулем комплекса, построенным на основе XML-ориентированных технологий (XSLT[15] и XPath[16]). Его задача – обращаться к базам библиографических данных по протоколу Z39.50 и подготавливать данные для следующего модуля.

Модуль статистического анализа *stat* представляет собой набор программ для выполнения в среде статистических вычислений R[17]. Он позволяет решать следующие задачи: 1) принятие решения о соответствии двух записей на основе набора значений признаков, вычисленных с помощью модуля aak, 2) обучение (вычисление параметров решающей функции) и 3) тестирование качества связывания.

Приводится описание экспериментального исследования, целью которого была проверка работоспособности предложенной технологии для конкретных данных. Был проведен ряд экспериментов. Результаты каждого эксперимента оценивались по охвату библиографических записей и количеству ошибок связывания. Под охватом понимается процент библиографических записей в базе данных, соответствующих требованиям полноты, т.е. содержащих достаточное количество информации для связывания. В качестве ошибок связывания рассматривались неверное отрицание связи (ошибка I рода) и неверно установленная связь (ошибка II рода).

Все проведенные эксперименты можно разделить на три группы по используемому набору правил сравнения (или признаков). В каждой группе экспериментов проводилось многократное разбиение выборки на обу-

[14] The ANSI/NISO Z39.50 Protocol: Information Retrieval in the Information Infrastructure

[15] eXtensible Stylesheet Language Transformations (англ.)

[16] XML Path Language (англ.)

[17] Free software environment for statistical computing and graphics (англ.), http://www.r-project.org/

чающую и тестовую части. Затем вычислялось среднее количество ошибок связывания для данного набора правил. Результаты экспериментов приводятся в таблице 1. При проведении *первой группы экспериментов* использовался *минимальный набор признаков*. Такой набор использует только саму авторитетную запись α, без подключения информации из связанных с ней записей. *Во второй группе экспериментов* использовался *стандартный набор признаков*. Этот набор был расширен за счет привлечения информации о соавторах и предметных рубриках, содержащейся в записях β, связанных с данной записью α. Такое расширение позволило существенно увеличить охват базы данных до (77%). Однако, при этом возросло количество ошибок I и II рода за счет использования менее полных записей.

Таблица 1: Результаты экспериментов: Ошибки I и II рода

Группа экспериментов	Ошибки I рода, %	Ошибки II рода, %	Сумма ошибок, %	Охват записей, %	Ошибки скорректированные
I	0,1	0,51	0,61	21	37,4
II	1,055	0,761	1,816	77	1,816
III	0,676	0,461	1,137	77	1,137

В *третьей группе экспериментов* использовался *расширенный набор признаков*. В этот набор были добавлены правила сравнения, в которых информация из минимального набора была заново оценена на основе записей β, связанных с записью α. Поскольку при этом количество задействованных в процессе сопоставления полей не изменилось, процент охвата остался равным 77%, при улучшении качества связывания.

Относительно низкий процент ошибок при проведении экспериментов во многом обусловлен хорошим качеством записей из тестовой выборки. В нее включались записи β, в которых присутствовало указание на соответствующие записи α. На практике такие записи, как правило, содержат значительно больше информации об авторах, чем записи β без указаний на авторитетные записи.

Для оценки качества идентификации менее полных записей было проведено искусственное «ухудшение» качества данных путем стирания информации, используемой в минимальном наборе признаков. Таким об-

разом, идентификация проводилась лишь на основе данных о соавторах и предметных рубриках. Количество ошибок I и II рода в этом случае составило 13,63% и 3,29% соответственно для стандартного набора признаков, 20,02% и 1,69% – для расширенного. Следует отметить, что процент установления неверных связей между записями (ошибка II рода) по-прежнему достаточно низок, увеличивается лишь количество упущенных связей между записями. Итак, в самой «худшей» ситуации, когда система располагает минимальной информацией об авторе, процент ошибок идентификации составит приблизительно 17%.

В заключении к третьей главе приводятся выводы, основанные на результатах проведенных экспериментов. Основной вывод заключается в возможности использования предложенной технологии к решению поставленной задачи. При этом особое внимание следует обращать на то, какая информация лежит в основе сопоставления и насколько часто она в действительности присутствует в библиографических записях.

В заключении сформулированы основные результаты исследований по теме диссертации.

Приложение А Содержит примеры записей. **Приложение Б** Содержит паспорта используемых баз данных. **Приложение В** Содержит входные требования к библиографическим записям в формате RUSMARC и авторитетным записям в формате RUSMARC/Authorities, которые выступают в роли рекомендаций по созданию новых авторитетных и библиографических записей. **Приложение Г** Содержит листинг консольного клиента aak и модуля статистического анализа stat. **Приложение Д** Содержит ранжированные наборы признаков, использованные при проведении экспериментов. **Приложение Е** Содержит оценки матриц W^{-1}, полученные при проведении экспериментов. В **приложении Ж** приводятся акты внедрения результатов работы.

ОСНОВНЫЕ ВЫВОДЫ И РЕЗУЛЬТАТЫ РАБОТЫ

В диссертационной работе выполнен анализ перспективных подходов и технологий, применяемых для связывания записей. Поставлена задача автоматизации процесса авторитетного контроля, сформулированы основные требования, которые необходимо учитывать при ее решении.

Предложены аналитическая и концептуальная модели связывания библиографических записей, основанные на методах машинного обучения. Также предложена технология автоматического авторитетного контроля электронного каталога. Разработано программное обеспечение, позволяющее оценить качество связывания для конкретных баз библиографических и авторитетных записей, а также обучить систему автоматического авторитетного контроля на основе этих баз данных.

В процессе разработки технологии автоматического авторитетного контроля были определены входные требования к библиографическим и авторитетным записям, которые можно рассматривать как рекомендации по наполнению библиографических и авторитетных баз данных.

Основные научные выводы и практические результаты:

1. Предложены аналитическая и концептуальная модели связывания библиографических записей, позволяющие использовать информацию об уже установленных связях в массиве данных;
2. На основе предложенных моделей разработана технология автоматического авторитетного контроля, позволяющая связывать библиографические записи в формате RUSMARC и авторитетные записи в формате RUSMARC/Authorities;
3. Разработан программный комплекс «ААК-персоны», позволяющий идентифицировать персоны в библиографических записях;
4. Проведено экспериментальное исследование качества связывания записей для коллекций библиографических и авторитетных записей (около 300 тысяч и 10 тысяч записей соответственно). Проведенное исследование подтвердило достоверность и эффективность предложенных методик;
5. В результате применения технологии к реальным базам данных получены ранжированные списки признаков и матрицы весовых коэффициентов; сделан вывод о том, какой набор признаков дает меньшее число ошибок связывания;
6. Разработанные в диссертации методы и алгоритмы были использованы на практике, что подтверждено актами о внедрении, прилагаемыми к диссертационной работе.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК

1. Князева А. А. Ранжированный поиск в библиографических базах данных / А. А. Князева [и др.] // Вестн. НГУ. Сер. : Информ. технологии. – 2009. – Т. 7, вып. 4. – С. 81–96.

2. Федотов А. М. Проблемы авторитетного контроля для распределённых электронных библиотек и библиографических баз / А. М. Федотов, О. Л. Жижимов, А. А. Князева [и др.] // Вестн. НГУ. Сер. : Информ. технологии. – 2011. – Т. 9, вып. 1. – С. 89–101.

3. Князева А. А. Принципы идентификации объектов в структурированных документах / А. А. Князева // Вестн. НГУ. Сер. : Информ. технологии. – 2013. – Т. 11, вып. 1. – С. 58–67.

Труды конференций

4. Князева А. А. Автоматический авторитетный контроль для распределённых библиографических баз данных [Электронный ресурс] / А. А. Князева, И. Ю. Турчановский, О. С. Колобов // XIII Рос. конф. с участием иностр. учен. «Распределённые информационные и вычислительные ресурсы» (DICR'2010), Новосибирск, 30 нояб.–4 дек. 2010 г. : материалы конф. – Новосибирск : ИВТ СО РАН, 1996–2013. – URL: http://conf.nsc.ru/dicr2010/ru/reportview/29244, свободный. – Загл. с тит. экрана (дата обращения: 04.06.2013).

5. Князева А. А. Автоматический авторитетный контроль [Электронный ресурс] : [доклад на конференции "Корпоративные информационно-библиотечные системы: технологии и инновации"(11; 2013; Санкт-Петербург)] / А. А. Князева, О. С. Колобов. – Электрон. текстовые дан. (1 файл : 229 Кб). – Санкт-Петербург, 2013. – Доклад опубликован на электрон.-опт. диске с материалами конференции (локальный шифр CD-670). – Свободный доступ из сети Интернет (чтение, печать, копирование). – Adobe Acrobat Reader 7.0. – <URL : http://dl.unilib.neva.ru/dl/2/3280.pdf>.

6. Князева А. А. Восстановление связей между библиографическими записями [Электронный ресурс] / А. А. Князева, О. С. Колобов // Междунар. конф. «Современные проблемы математики, информатики и биоинформатики», посвящ. 100-летию со дня рождения чл.-кор. АН СССР А. А. Ляпунова, Новосибирск, 11–14 окт. 2011 г. : материалы конф. – Новосибирск : ИВТ СО РАН, 1996–

2013. – URL: http://conf.nsc.ru/Lyap-100/reportview/74497, свободный. – Загл. с тит. экрана (дата обращения: 04.06.2013).

7. Князева А. А. Автоматическое связывание документов / А. А. Князева, И. Ю. Турчановский, О. С. Колобов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL'2012) : тр. XIV Всерос. науч. конф., Переславль-Залесский, 15–18 окт. 2012 г. – Переславль-Залесский : Изд-во «Университет города Переславля», 2012. – С. 360–369.

8. Князева А. А. Автоматическое связывание структурированных документов [Электронный ресурс] / А. А. Князева, И. Ю. Турчановский, О. С. Колобов // Материаловедение, технологии и экология в 3-м тысячелетии : сб. докл. V Всерос. конф. молод. учен. / Ин-т оптики атмосферы СО РАН. – Электрон. текст. дан. – Томск : ИОА СО РАН, 2012. – [С. 9–12]. – 1 электрон. опт. диск (CD-ROM). – №гос. регистрации 0321300235.

9. Князева А. А. Наличие информации для связывания на примере базы данных «MedArt» [Электронный ресурс] / А. А. Князева, О. С. Колобов, И. Ю. Турчановский // XIV Рос. конф. с междунар. участием «Распределённые информационные и вычислительные ресурсы» (DICR-2012), Новосибирск, 26–30 нояб. 2012 г. : материалы конф. – Новосибирск ИВТ СО РАН, 1996–2013. – URL: http://conf.nsc.ru/dicr2012/ru/reportview/139662, свободный. – Загл. с тит. экрана (дата обращения: 04.06.2013).

10. Князева А. А. Выявление дубликатов в библиографических базах данных / А. А. Князева, И. Ю.Турчановский, О. С. Колобов // Электронные библиотеки : перспективные методы и технологии, электронные коллекции (RCDL'2013) : тр. XV Всерос. науч. конф., Ярославль, 14–17 окт. 2013 г. – Ярославль : ЯрГУ, 2013. – С. 276–282.

Печ. л. 1.
Тираж 100 экз.