Автоматический авторитетный контроль

Князева Анна Анатольевна, младший научный сотрудник, Институт вычислительных технологий Сибирского отделения Российской академии наук, aknyazeva22@gmail.com

Колобов Олег Сергеевич, директор, ООО «Комлог», okolobov@gmail.com

В работе дано описание технологии автоматического авторитетного контроля для одной или нескольких коллекций библиографических записей. В основе технологии лежит модель связывания документов, которые представляют один объект реального мира. В качестве такого объекта рассматривается автор, а отношение воссоздается для принятой формы – имя лица. В работе описывается применение представленной технологии на примере коллекций аналитических и авторитетных/нормативных записей в одном из проектов некоммерческого партнерства «МедАрт». Приводится ранжированный список признаков, на основе которых принимается решение, а также техническое решение для интерпретации этой технологии в виде отдельного сервиса сети.

Введение

В отечественной литературе под авторитетным контролем подразумевают процесс поддержки единообразия форм представления заголовков, определяющих одно и тоже лицо, организацию, предмет и т.д. в библиографической записи, а также контроль над процессом отображения содержания документов [Yurchenko2003]. Заголовок представляется в виде отдельной авторитетной записи, которая однозначно определяет заголовок. Однократное введение авторитетной записи в систему позволяет любую ее модификацию автоматически соотносить с другими авторитетными записями, связанными с ней, а также с соответствующими библиографическими записями [Yurchenko2003]. Это позволяет упростить работу каталогизаторов и других специалистов, а также повысить качество записей за счет их унификации.

В общем случае применительно к авторитетным записям можно рассматривать четыре функции, которые представляют различные аспекты авторитетного контроля:

1. функция *авторитетный контроль* (обеспечивает единообразие заголовков);
2. функция *поиск* (обеспечивает связь с различными вариантами форм заголовков);
3. функция *информационная* (обеспечивает предоставление для заголовка примечаний и комментариев каталогизатора с указанием используемых источников);
4. функция *техническое обслуживание* (обеспечивает поддержку выявления ошибок и корректировки библиографических записей в ручном или автоматическом режиме).

Как правило, авторитетный контроль осуществляется следующим образом: при внесении в базу новой библиографической записи каталогизатор находит авторитетную запись, с которой необходимо установить связь и вносит в определенное поле библиографической записи соответствующий номеру авторитетной записи. Далее будем называть его авторитетным *кодом*. Поле с авторитетным кодом выступает в качестве точки доступа, что позволяет легко находить все библиографические записи, связанные с определенной авторитетной записью. Проще говоря, становится возможным найти все произведения конкретного автора, не включая в результат поиска произведения его однофамильцев. Кроме уточнения поиска авторитетные файлы позволяют унифицировать и ускорить ввод документов - нужные элементы описания берутся из авторитетной записи, а не вводятся каждый раз заново [Yurchenko2003].

Целью нашей работы является разработка механизма автоматического авторитетного контроля в аспекте *технического обслуживания* библиографических записей. Этот механизм требует участия человека в процессе связывания только в тех случаях, когда в библиографической записи недостаточно информации или возникают коллизии при связывании. Под коллизией понимается ситуация, когда для одной библиографической записи находятся более одной авторитетной записи, и неясно какой из них следует отдать предпочтение.

В рамках данной работы решалась задача автоматического авторитетного контроля в рамках работ по развитию научно-образовательного портала НП «МедАрт»[MEDARTCO]. Разработка технологии автоматического сопоставления библиографической записи с авторитетными записями и вынесения решения о соответствии, либо несоответствии для каждой из них. Под авторитетными данными будем понимать авторитетный файл имен авторов, хотя общий подход, изложенный в работе, может быть применен и к другим видам авторитетных документов (таких, как наименования организаций или предметные рубрики).

Идея автоматического авторитетного контроля

Анализ соответствия пары записей авторитетной и библиографической (далее АЗ-БЗ) – достаточно сложная задача. Соответствие записей в целом необязательно означает совпадение информации в отдельных полях[Bilenko2002]. Причинами такого несоответствия могут быть опечатки, транспозиции символов, измененный порядок слов, использование сокращений и аббревиатур, разночтения в зарубежных транскрипциях, неполнота данных и т.п.[Rubtsov2009ngu]. Поэтому после непосредственного сравнения значений в отдельных полях записей, принимать решение о соответствии АЗ-БЗ.

Предлагаемый подход основан на идее машинного обучения и опирается на так называемую *обучающую выборку*. Если есть некоторое множество уже связанных записей, то на их основе можно создать два набора данных: класс соответствующих пар записей, то есть записей с установленной связью, и класс несоответствующих пар, где связь не просто не установлена, но и должна отсутствовать. Анализируя затем пару АЗ-БЗ можно делать заключение о соответст-

вии записей на основе того, насколько данная пара близка к одному из указанных классов. Для определения близости задаются правила вычисления расстояний между парами записей.

Такой подход позволяет автоматически оценивать значимость признаков, поскольку, чем важнее признак, тем больший вес он получит в результате процедуры обучения. При этом учитывается взаимозависимость используемых признаков.

Анализ базы данных «MedArt»

В рамках данной работы использовались записи двух видов: авторитетные записи имен авторов и библиографические записи. Авторитетные записи полностью посвящаются персонам, являющимся авторами публикаций, поэтому, как правило, содержат достаточно подробную информацию об этих персонах. Основная проблема авторитетного контроля возникает из-за недостаточной полноты библиографических записей, в которых, зачастую, вся информация об авторе сводится к его фамилии и инициалам. Поэтому был проведен анализ наличия информации в библиографической базе данных.

В библиографическом описании одного источника может встречаться несколько разных персон (как авторов, так и редакторов, иллюстраторов и т.д.). Рассматривать, и связывать будем конкретное упоминание персоны. При этом для простоты все упоминаемые персоны считаются авторами источника, без уточнения конкретной роли. В контексте формата RUSMARC это означает, что поля 700, 701 и 702 рассматриваются одинаково. При анализе базы данных рассматривалась только часть упоминаний авторов. Эта часть была получена путем выборки из базы данных всех авторов, у которых есть совпадение по фамилии и инициалам, а также их соавторов. Всего было рассмотрено 41746 упоминаний авторов.

Непосредственно в поле с описанием автора, согласно правилам RUSMARC может присутствовать следующая информация (Таблица 1), которую будем называть *основной информацией.*

Таблица 1

Подполе	Наименование	Использование в работе
$a	Начальный элемент ввода	Фамилия
$b	Часть имени, кроме начального элемента ввода	Инициалы
$c	Дополнение к именам, кроме дат	Профессия
$f	Даты	Годы жизни автора
$g	Расширение инициалов личного имени	Имя и отчество в полной форме
$p	Наименование/адрес организации	Место работы автора
$3	Номер авторитетной /нормативной записи	Авторитетный код

В процессе связывания использовались данные о русскоязычных авторах с фамилией, именем и отчеством, поэтому подполе $d (Римские цифры) было исключено из рассмотрения. Также не используются подполя $o и $9. Подполе $4 (Код отношения) в настоящий момент не используется, хотя в дальнейшем его планируется подключать для идентификации персон, часто выступающих в роли редакторов, иллюстраторов и т.п. Подполя $a и $b являются обязательными и присутствуют во всех рассматриваемых записях, поэтому они не упоминаются далее.

Анализ основной информации в рассматриваемой базе данных привел к следующим результатам (Таблица 2).

Таблица 2

Признак	L	N0	P0 (%)	SP0 (%)	N1	P1 (%)	SP1 (%)	N	P (%)	SP (%)
Полная форма имени (ФИО)	No	32499	97.7	97.7	12	0.1	0.1	32511	77.9	77.9
	Yes	759	2.3	100.0	8476	99.9	100.0	9235	22.1	100.0
	All	33258	100.0		8488	100.0		41746	100.0	
Профессия	No	33224	99.9	99.9	291	3.4	3.4	33515	80.3	80.3
	Yes	34	0.1	100.0	8197	96.6	100.0	8231	19.7	100.0
	All	33258	100.0		8488	100.0		41746	100.0	
Годы жизни	No	33249	100.0	100.0	1950	23.0	23.0	35199	84.3	84.3
	Yes	9	0.0	100.0	6538	77.0	100.0	6547	15.7	100.0
	All	33258	100.0		8488	100.0		41746	100.0	
Место работы	No	30514	91.8	91.8	4334	51.1	51.1	34848	83.5	83.5
	Yes	2744	8.2	100.0	4154	48.9	100.0	6898	16.5	100.0
	All	33258	100.0		8488	100.0		41746	100.0	

В приведенной таблице в колонке «L» указаны возможные значения признака (наличие «Yes» либо отсутствие «No» информации), «N0» означает количество упоминаний, в которых не указан код соответствующего авторитетного документа, «N1» - количество упоминаний с указанным кодом и «N» - всех упоминаний. В следующих колонках указываются соответствующие проценты и кумулятивные проценты относительно всех упоминаний. Такое разбиение было сделано для иллюстрации еще одного важного замечания. Дело в том, что в процессе эксперимента для обучения алгоритма связывания и анализа его работы использовались только упоминания с указанным кодом. Это делалось для того, чтобы избежать необходимости вручную оценивать правильность связывания.

Проведенный эксперимент показал, что наличие двух или более подполей с основной информацией позволяет достаточно уверенно делать заключение о соответствии упоминания конкретному автору. Однако, в рассматриваемой базе данных упоминания авторов, которые содержат два и более подполя, составляют всего 21.18% всех упоминаний.

Дополнительная информация из библиографической записи может учитываться при связывании наряду с основной информацией. Эмпирическим путем для анализа была выбрана дополнительная информация, которая состоит из трех групп.

Группа 1. К дополнительной информации относится поле 712 (Коллективный автор), которое содержит наименование и географическое положение организации, имеющей отношение к публикации. При этом временные организации, такие как конференции, симпозиумы и т.п. не рассматриваются. Поскольку форматом RUSMARC не предусмотрена связь между коллективом и отдельным автором, нельзя с уверенностью говорить о том, что коллектив характеризует рассматриваемого автора, ведь он может относиться к одному из соавторов публикации. Поэтому будем считать, что информация о коллективе является косвенной.

Таблица 3

Признак	O	N	P (%)	SP (%)
Коллектив	0	29133	69.8	69.8
	1	10004	24.0	93.8
	2	2282	5.5	99.2
	3	286	0.7	99.9
	4	32	0.1	100.0
	5	9	0.0	100.0
	All	41746	100.0	
Коллектив без географической привязки	0	41535	99.5	99.5
	1	211	0.5	100.0
	All	41746	100.0	

В приведенной Таблице 3 в колонке «O» указанно количество упоминаний признака в отдельной записи. Как следует из Таблицы 3, подавляющее большинство упоминаний, а именно 69.8%, содержится в документах, в которых не указаны постоянные коллективные авторы. Это не даст существенного расширения доступной информации при включении данного фактора в процедуру связывания. Тем не менее, информация о коллективах не должна и просто отбрасываться в сторону из-за того, что она редко встречается. Дело в том, что согласно формату RUSMARC информация о местоположении организации не является обязательной, в отличие от ее наименования. Однако на практике оказалось, что такая информация как правило, присутствует в документе. Коллективами без географической привязки, количество которых всего 211 из 41746 можно пренебречь.

Группа 2. В качестве дополнительной информации выступает информация о соавторах. Рассматривая упоминание автора в контексте библиографической записи, можно учесть упоминания других персон, если они присутствуют в той же записи. При этом с помощью механизма *расширенных авторитетных запи-*

сей (расширение за счет включение в рассмотрение всех авторов связанных с данной авторитетной записью) можно учесть тот факт, что упоминания некоторых соавторов уже встречались в других библиографических записях, связанных с рассматриваемым автором. Такая информация позволяет существенно расширить категорию библиографических записей, для связывания. Следует отметить, что наличие соавторов высоко оценивается в рамках проекта VIAF [Bennett].

Группа 3.Последняя группа дополнительной информации содержит информацию о предметных рубриках.

Наличие информации о соавторах и предметных рубриках представлено в Таблице 4.

Таблица 4

Признак	**L**	**N**	**P (%)**	**SP (%)**
Рубрика	No	12537	30.1	30.1
	Yes	29173	69.9	100.0
	All	41746	100.0	
Соавтор без кода	No	6511	15.6	15.6
	Yes	35235	84.4	100.0
	All	41746	100.0	
Соавтор с кодом	No	26987	64.7	64.7
	Yes	14759	35.4	100.0
	All	41746	100.0	

При рассмотрении информации о соавторах проводилось деление на две группы в зависимости от наличия кода авторитетного документа. Дело в том, что соавторы с указанным кодом являются более надежной информацией, поскольку мы можем их более точно идентифицировать, чем соавторов, для которых такой код не указан. К сожалению, такая информация встречается реже.

Как видно из Таблицы 4 информация о рубриках присутствует в 69.88% записей (или для процента упоминаний). Особенная ценность предметных рубрик в том, что они подчиняются тезаурусу MeSH.

Еще один интересный вопрос заключается в том, насколько часто основная и дополнительная информация пересекаются между собой. В следующей таблице приводится пересечение этих двух типов информации в процентном соотношении.

Таблица 5

Основная (%) (без учета обязательных подполей)	**Дополнительная (%)**	
	Yes	**No**
Yes	26.877	0.755
No	70.033	2.336

На основании Таблицы 5 можно утверждать, что учет наличия любой имеющейся информации намного более обоснован. Так, вся рассматриваемая информация кроме фамилии и инициалов отсутствует всего в 2.34% упоминаний.

Более подробно распределение количества имеющихся полей по видам информации представлено в Таблице 6, из которой видно, что наиболее часто встречаются упоминания, в которых присутствует 2-3 фактора из рассматриваемых признаков.

Таблица 6

Признак	O	N	P (%)	SP (%)
Основная информация	0	30211	72.4	72.4
	1	2692	6.5	78.8
	2	1940	4.7	83.5
	3	3273	7.8	91.3
	4	3630	8.7	100.0
	All	41746	100.0	
Дополнительная информация	0	1290	3.1	3.1
	1	10017	24.0	27.1
	2	14958	35.8	62.9
	3	10077	24.1	87.1
	4	5404	12.9	100.0
	All	41746	100.0	
Вся информация	0	975	2.3	2.3
	1	8507	20.4	22.7
	2	12959	31.0	53.8
	3	7109	17.0	70.8
	4	4116	9.9	80.6
	5	2527	6.0	86.7
	6	2225	5.3	92.0
	7	2134	5.1	97.1
	8	1194	2.9	100.0
	All	41746	100.0	

Среди всех 41746 упоминаний такие упоминания с указанием рубрики и соавтора без авторитетного кода составляют 59%.

Заключение по анализу базы данных «MedArt»

При автоматическом связывании естественно полагаться на наличие информации о предметных рубриках и соавторах без кода, поскольку именно такая информация встречается чаще всего. При этом необходимо добиваться приемлемого качества связывания для упоминаний, в которых не встречается дру-

гой информации. Добиться этого можно, например, с помощью уточнения процедуры сопоставления для предметных рубрик MeSH.

Элементы технологии автоматического авторитетного контроля библиографических записей

Технология автоматического авторитетного контроля основана на модели идентификации персон [KNYAZEVA]. Рассмотрим, кратко, основные ее этапы в применении к электронному каталогу. Все этапы представлены в виде отдельных блоков. При помещении в электронный каталог нового библиографического документа необходимо производить процедуру связывания с существующими авторитетными документами. В рассматриваемой системе используются записи двух типов: библиографические документы (БЗ) и авторитетные документы (АЗ).

Процедура связывания производится для каждой персоны. Пара образуется на основе связи «Имя лица», под этот тип подпадают следующие связи, определенные форматом RUSMARC:

1. 700 - Имя лица - Первичная ответственность;
2. 701 - Имя лица - Альтернативная ответственность;
3. 702 - Имя лица - Вторичная ответственность.

При связывании все эти поля являются равноправными, для каждого из них находятся авторитетные записи, которые могут быть связаны с данным полем. Под термином «автор» в данной работе понимается любой человек, имеющий отношение к документу и все поля седьмого блока рассматриваются одинаково. Условно будем обозначать то поле, по которому идет связывание, как поле номер 700. Все остальные лица, упомянутые в блоке «7--» – библиографической записи при этом считаются соавторами.

Функциональная схема процесса идентификации персон может быть представлена в виде отдельных блоков:

- ***Нормализация***. Нормализация документов на уровне полей обеспечивается правилами каталогизации, накладывающими ограничения на значения полей в библиотечных информационных системах. Нормализация на уровне записи осуществляется с помощью проверки на ее соответствие входным требованиям.
- ***Составление пар.*** При составлении пар значение составного ключа определяется подполями $a и $b поля 700 документа БЗ, поиск производится в подполях $a и $b поля 200 документа АЗ. Дополнение найденной АЗ до расширенного авторитетного документа производится следующим образом: по ключу, содержащему номер АЗ, в коллекции БЗ отбираются те документы, в 700, 701 или 702 поле которых в подполе $3 указано значение ключа. Полученное дополнение состоит из документов БЗ, уже загруженных в электронный каталог и связанных с рассматриваемым документом АЗ.
- ***Сравнение отдельных полей в паре документов.*** В зависимости от поля/пополя записи, в качестве сравнительных функций используются

следующие: 1) точное совпадение; 2) сравнение свободного текста с выделением основ слов; 3) сравнение дат. Формат RUSMARC предусматривает возможность повторения полей, когда несколько значений признака записываются в двух и более полях с одинаковым обозначением. Для сравнения записей по этому признаку необходимо агрегировать результаты сравнения по каждому из полей в одно значение, указывающее на степень соответствия по рассматриваемому признаку.

- ***Принятие решения***. Для сравнительного вектора, полученного в блоке 3 необходимо вычислить расстояния до центроидов классов соответствующих и несоответствующих пар.
- ***Обучение решающей функции***. Вычисляя значения перечисленных признаков для всех пар из обучающей выборки, получим таблицу исходных данных эксперимента. На основе данной таблицы сначала проводится предварительный анализ, а затем вычисляются параметры системы. В дальнейшем эти параметры могут использоваться при работе блока 4.
- ***Оценка качества связывания.*** Качество связывания оценивается по результатам тестов. При этом выделялись ошибки двух типов: неверное отрицание связи (ошибка I рода) и неверно установленная связь (ошибка II рода) [Gnedenko1988]. Количество ошибок зависит от того, как именно распределились пары документов по обучающей и тестовой выборкам. Поэтому рекомендуется проводить несколько экспериментов и вычислять средние проценты ошибок.

Заключение

Сервис автоматического авторитетного контроля может быть реализован как выделенный сервис в сети. Децентрализованная архитектура наиболее предпочтительная для создания такого сервиса. Так как в этом случае имеется возможность интегрировать ресурсы на логическом уровне, учитывать политику поставщиков данных в отношении ресурса и предоставить возможность поставщику данных заниматься вопросами хранения и предоставления доступа к информации.

На данном этапе мы определили подход для создания выделенного сервиса автоматического авторитетного контроля, который не исключает применение технологии Z39.50, но подразумевает применение XML-ориентированного протокола SRU. Это означает, что требования к поставщику данных проще, а возможностей для распространения информации о ресурсе и его документах больше. И более того, поставщик данных может воспользоваться нашей системой для организации работы со своим ресурсом, для выполнения функций поиска и метапоиска, включая выполнение автоматического авторитетного контроля. Для конечного пользователя предлагается автоматическое связывание документов одного автора, которые находятся у различных поставщиков данных.

Связывание документов одного автора из различных источников состоит из двух шагов. Первый шаг выполняется поставщиком данных на этапе индекси-

рования библиографических записей. На этом этапе вычисляется дополнительная связь загружаемого документа с авторитетными записями, и соответствующие номера авторитетных записей добавляется в метаданные индексируемого документа. Второй шаг, это создание ссылки на документы данного автора из других источников на основе сервиса. На этом этапе автоматически создается ссылка, переход по которой выполняет поиск документов во всех источниках по номеру авторитетной записи на автора, который был добавлен в метаданные на первом шаге.

Список литературы

1. [Yurchenko2003] Юрченко Я. Г. Авторитетный контроль как важнейший элемент интеграции [Электронный ресурс] /Я. Г. Юрченко // Фонды и каталоги Кузбасса. Опыт. Проблемы. Решения : науч.-практ. сб. – Кемерово : [Кемер. обл. науч. б-ка им. В. Д. Федорова], 2003. – Вып. 2. – URL:http://www.kemrsl.ru/documents/founds/vip2/vip2.5.htm, свободный. – Загл. с экрана (дата обращения: 04.06.2013).
2. [MEDARTCO] Мешечак Н. А. Опыт создания и использования авторитетных записей на томских ученых-медиков в научно-медицинской библиотеке Сибирского медицинского университета / Н. А. Мешечак, Л. А. Шамардина, А. С. Карауш // Современные пользователи автоматизированных информационно-библиотечных систем: проблемы обслуживания, изучения и обучения : материалы 6-й и 7-й науч.-практ. конф. – СПб. : РБА, 2006. – С. 158–161.
3. [Bilenko2002] Bilenko M. Learnable similarity functions and their application to record linkage and clustering [Electronic resource] : diss. . . . for the degree of DPh / Mikhail Yuryevich Bilenko ; Univ. of Texas. – Austin, 2006. – 136 p. – The electronic version of print. publ. – Access from ProQuest Dissertations and Theses. – Title from the screen.
4. [Rubtsov2009ngu] Рубцов Д. Н. Выявление дубликатов в разнородных библиографических источниках / Д. Н. Рубцов, В. Б. Барахнин // Вестн. НГУ. Сер. : Информ. технологии. – 2009. – Т. 7, вып. 3. – С. 86–93.
5. [Bennett] Bennett R. VIAF (Virtual international authority file): linking the Deutsche Nationalbibliothek and Library of Congress name authority files / R. Bennett [et al.] // Int. cataloging and bibliographic control. – 2007. – Vol. 36, № 1. – P. 12–19.
6. [KNYAZEVA] 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.
7. [Gnedenko1988] Гнеденко Б. В. Курс теории вероятностей : учеб. – 6-е изд., перераб. и доп. / Б. В. Гнеденко. – М.: Наука, 1988. – 448 с.