ОБЪЕДИНЕНИЕ БИБЛИОТЕЧНЫХ КАТАЛОГОВ С РАЗЛИЧНЫМИ БАЗАМИ ДАННЫХ АВТОРИТЕТНЫХ ЗАПИСЕЙ

COMBINING LIBRARY CATALOGS WITH VARIOUS DATABASES OF AUTHORITY RECORDS

Григянец Ромуальд Брониславович *– заведующий лабораторией информационного обеспечения научных исследований, кандидат технических наук, доцент, государственное научное учреждение «Объединенный институт проблем информатики Национальной академии наук Беларуси» (Республика Беларусь), griganec@bas-net.by*

Griganets Romuald B. *– Head of the Laboratory information support of scientific research, candidate of technical sciences, assistant professor, State Scientific Institution "United Institute of Informatics Problems of the National Academy of Sciences of Belarus" (Belarus), e-mail: griganec@bas-net.by*

Степанцова Елена Вячеславовна *– главный конструктор проекта лаборатории информационного обеспечения научных исследований, государственное научное учреждение «Объединенный институт проблем информатики Национальной академии наук Беларуси (Республика Беларусь)», stelena@basnet.by*

Stepantsova Alena V. *– chief architect of the project of the laboratory information support of scientific research, State Scientific Institution "United Institute of Informatics Problems of the National Academy of Sciences of Belarus" (Belarus), e-mail: stelena@basnet.by*

Рабушко Кристина Анатольевна *– ведущий инженер программист лаборатории информационного обеспечения научных исследований, государственное научное учреждение «Объединенный институт проблем информатики Национальной академии наук Беларуси (Республика Беларусь)», mitskevich@basnet.by*

Rabushka Krystsina A. *– leading software engineer of the laboratory information support of scientific research, State Scientific Institution "United Institute of Informatics Problems of the National Academy of Sciences of Belarus", e-mail: mitskevich@basnet.by*

Аннотация. *Рассматривается задача объединения электронных библиотечных каталогов с различными базами данных авторитетных записей (АЗ). Представлен алгоритм связывания библиографических и авторитетных записей в автоматическом режиме, что позволит сократить число растущих прототипов АЗ и уменьшить трудозатраты и время библиотекарей на их обработку.*

Abstract. *The problem of combining electronic library catalogs with various databases of authorities records (AR) is considered. An algorithm for linking bibliographic and authorities records in an automatic mode is presented, which*

allows you to reduce the number of growing AR prototypes, reduce the labor costs and time of libraries for their processing.

Ключевые слова: *база данных, электронный каталог, авторитетная запись, библиографическая запись.*

Keywords: *date base, electronic catalog, authorities record, bibliographic record.*

В современном мире научно-исследовательский процесс тесно связан с использованием ресурсов, доступных в Интернете. Обилие информации, которую необходимо изучить и обработать для принятия решений в различных областях, вынуждает конечного пользователя проводить большую часть времени за поиском и анализом данных. Все большее значение в этом процессе начинают приобретать электронные каталоги библиотек (ЭК) и электронные библиотеки, содержащие большие массивы библиографической информации, задачей которых является осуществление эффективного поиска в каталоге и предоставление пользователю не только найденного документа, но и дополнительных сведений о нем (например, авторы, организации, тематика, географические и временные привязки).

Полнота и качество информации, предоставляемой конечному пользователю, неразрывно связано с авторитетными записями, благодаря нормализованным заголовкам которых обеспечивается идентификация различных объектов, используемых в библиографических записях (БЗ). Большинство крупных библиотек в своих каталогах поддерживают базы авторитетных записей (АЗ) и применяют технологии авторитетного контроля. Это позволяет упростить работу каталогизаторов и повысить качество библиографических записей, решая проблему идентификации персон и ключевых слов.

Авторитетный контроль данных представляет собой процесс поддержания единых форм авторитетных/нормативных данных, контроль за адекватностью присвоения поисковых реквизитов, последовательным соблюдением принципов, методик, инструкций и правил по представлению поисковых признаков документа [1]. Применение технологии авторитетного контроля позволяет однозначно идентифицировать АЗ, однако на практике приходится сталкиваться с определенными проблемами:

- связи между БЗ и АЗ могут отсутствовать частично либо полностью (в БЗ описываются не все лица, а только несколько основных);
- связи между БЗ и АЗ со временем становятся некорректными (например, при объединении библиотечных каталогов, когда в каждом ЭК используются свои идентификаторы АЗ), либо теряют свою актуальность;
- качество и полнота АЗ является неудовлетворительным (не описываются либо не полностью описываются вспомогательные идентифицирующие признаки);

- информация, содержащаяся в описании АЗ, имеет довольно произвольный вид и требует определенной стандартизации.

Создание и поддержка качественных баз данных АЗ, таким образом, приобретает все большее значение, их ценность возрастает, несмотря на то, что это является довольно трудоемким процессом, зачастую требующим анализа дополнительной информации из множества источников [2].

Решение подобного рода задач было предложено в рамках проекта VIAF (Virtual International Authority File, виртуальный международный авторитетный файл) Международной федерации библиотечных ассоциаций и учреждений (ИФЛА). Проект был инициирован в 2000 г. Немецкой национальной библиотекой, Библиотекой конгресса США и Национальной библиотекой Франции. VIAF представляет собой систему классификации на базе Онлайнового компьютерного библиотечного центра, признанную на международном уровне. Основной задачей проекта VIAF является сопоставление и автоматическое связывание АЗ из различных национальных авторитетных файлов. Механизмы сравнения и слияния дубликатов АЗ достаточно сложны и включают анализ данных как авторитетных, так и библиографических записей из различных каталогов, создавая расширенные (или объединенные) АЗ.

Параллельно с задачей слияния баз авторитетных данных возникает задача автоматического связывания БЗ с АЗ при объединении нескольких электронных каталогов, что встречается довольно часто. Даже в случае, если сливаемые записи достаточно полны и содержат корректную информацию, зачастую приходится сталкиваться с тем, что для каждого каталога применяется свой набор авторитетных файлов, либо АЗ не используются вовсе. В итоге библиографические записи на материалы одного и того же автора содержат разные идентификаторы авторитетной записи, или не содержат их вообще.

На постсоветском пространстве А.М. Федотовым, О.Л. Жижимовым и другими авторами был предложен алгоритм автоматического связывания БЗ и АЗ при объединении ЭК библиотек [3]. Данный алгоритм был адаптирован авторами статьи для формата BelMarc с учетом технологий загрузки в систему корпоративной каталогизации (СКК) библиотек Беларуси.

Задача автоматического авторитетного контроля состоит в сопоставлении и связывании авторитетных и библиографических записей на основании имеющейся в них информации. Рассматривать записи в паре представляется целесообразным, поскольку при сравнении можно учесть большее количество факторов.

В общих чертах алгоритм автоматического авторитетного контроля записей можно описать следующим образом:

1. При добавлении новой БЗ в ЭК анализируются поля, содержащие ссылки на АЗ;
2. В базе данных выполняется поиск соответствующих АЗ для составления пар «АЗ-БЗ»;
3. Для каждой пары рассчитывается оценка соответствия, в зависимости от которой принимается решение о соответствии/несоответствии самой пары;
4. В том случае, если для добавляемой БЗ находится только одна АЗ, то в БЗ подставляется ее идентификатор, если больше одной – запись отправляется на доработку с привлечением специалистов. Если соответствий не найдено, то БЗ попадает в базу данных без каких-либо отметок и формируется прототип АЗ.

При сопоставлении записей применяются методы линейного дискриминантного и факторного анализа по коррелирующим переменным.

Если составить пару из библиографической и авторитетной записей, в которых фамилия и инициалы автора совпадают, то все множество таких пар можно условно разбить на два класса: соответствующие и несоответствующие пары. Сопоставление записей происходит путем анализа данных определенных полей. Для оценки соответствия полей используются такие критерии, как «не совпадает», «не определено», «совпадает» и «частично совпадает», которым присваиваются числовые характеристики (что позволяет вычислять такие статистические характеристики как среднее и ковариация). Сравниваться могут как слова целиком, так и их усечения (словоформы), в зависимости от рассматриваемого поля. Например, на рисунке 1 представлена часть полей, используемых для анализа имен лиц.

Переменная	АЗ	БЗ	Значение	Код	Комментарий
идентификатор	001	700$3 701$3 702$3	не соответствует	1	точное совпадение
			соответствует	2	
начальный элемент ввода (фамилия)	200$a	700$a 701$a 702$a	не совпадает	1	точное совпадение
			совпадает	2	
часть имени от нач.элемента	200$b	700$b 701$b 702$b	не совпадает	1	точное совпадение
			совпадает	2	
идентифицирующий признак (кроме дат)	200$c	700$c 701$c 702$c	нет совпадений	1	совпадение усеченных форм
			не указано	2	
			одно совпадение	3	
			два совпадения	4	
			...	n	
дополнительный идентифицирующий признак	340$a		одно совпадение	1	совпадение усеченных форм
			...	n	
римские цифры	200$d	700$d 701$d 702$d	не совпадает	1	точное совпадение
			совпадает	2	

Рисунок 1 – Фрагмент анализируемых полей имен лиц

Каждую такую пару БЗ-АЗ можно представить некоторой точкой в n-мерном пространстве, где n – это количество анализируемых полей. Имея обучающую выборку и определив расстояние между точками, можно рассчитать средние значения по классу (центроиды). Под обучающей выборкой понимается набор пар БЗ-АЗ, для которых известна

принадлежность к тому или иному классу. В качестве расстояния используется расстояние Махаланобиса [4], учитывающее корреляции между переменными и не зависимое от масштаба.

Далее для каждой рассматриваемой пары БЗ-АЗ по набору факторов, отражающих степень совпадения данных, вычисляется результирующая переменная, определяющая принадлежность пары к одному из двух классов. Для этого необходимо определить расстояние от точки до центроидов обоих классов и выбрать класс, который ближе.

Используя расстояние Махаланобиса можно также на обучающей выборке произвести отбор наиболее информативных признаков и отсечь слабо информативные, что позволит увеличить скорость обработки данных.

На обучающей выборке по данному алгоритму рассчитываются статистические параметры алгоритма, после чего алгоритм можно применять на тестовой выборке. Результат применения алгоритма зависит от качества загружаемых БЗ, в частности, от полноты описаний авторов (подразумевается, что существующий массив АЗ обладает достаточной полнотой описания). Чем больше информации об авторах присутствует в описании БЗ (кроме обязательных фамилии и инициалов), тем выше процент автоматического связывания библиографической и авторитетной записи.

Можно сделать вывод, что при слиянии электронных каталогов, использующих АЗ (даже со своими собственными, отличающимися идентификаторами), использование данного алгоритма позволит существенно снизить процент формирования прототипов АЗ и, тем самым, сократит время и трудозатраты библиотекарей на поддержку и сопровождение базы АЗ. Повысить точность прогнозов соответствия, можно дополнив данный алгоритм сравнением тематики (поля 606, 610) рассматриваемой БЗ и документов, связанных с рассматриваемым автором (АЗ из пары), а также анализом списка соавторов.

Следует отметить, что данный метод может быть применен не только при объединении различных каталогов, но и для выявления дубликатов БЗ, АЗ и их прототипов. Что в свою очередь позволит улучшить качественное наполнение электронного каталога. При этом следует учитывать, что с ростом количества разнородных записей в ЭК необходимо время от времени выполнять уточнение статистических параметров алгоритма.

СПИСОК ЛИТЕРАТУРЫ

1. Масхулия, Т. Л. Национальные авторитетные/нормативные файлы предметных рубрик и заголовков, содержащих наименование организаций : принципы и подходы к формированию и поддержке / Т. Л. Масхулия, Ю. Г. Селиванова // Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества : материалы X юбил. Междунар. конф. «Крым 2003», Судак, 7–15 июня 2003 г. – М. : ГПНТБ России, 2003. – Т. 1. – С. 209–210.

2. Ковалёва А. М. Авторитетный файл «Имя лица» / А. М. Ковалёва // Библиотечное краеведение в информационном пространстве региона : материалы межрегион, науч.-практ. конф., Барнаул, 26-27 февр. 2008 г. –Барнаул : РИО АКУНБ, 2008. – С. 172–178.

3. Федотов А. М. Проблемы авторитетного контроля для распределенных 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.

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