НАЛИЧИЕ ИНФОРМАЦИИ ДЛЯ СВЯЗЫВАНИЯ НА ПРИМЕРЕ БАЗЫ ДАННЫХ "MEDART"

А.А. Князева, О.С. Колобов, И.Ю. Турчановский

Институт вычислительных технологий СО РАН
e-mail: aknjazeva@ict.nsc.ru
Институт сильноточной электроники СО РАН
e-mail: okolobov@hcei.tsc.ru
Институт вычислительных технологий СО РАН
e-mail: tur@hcei.tsc.ru

В работе рассматривается библиографическая информация, на основе которой можно производить автоматическое связывание библиографических и авторитетных документов. Приводятся результаты анализа базы библиографических документов "MedArt" с точки зрения наличия такой информации.

Введение

Задача автоматического авторитетного контроля, или автоматического связывания библиографических документов с соответствующими авторитетными документами в настоящее время не решена. Для сопоставления авторитетных и библиографических документов и уставновления между ними связей без использования ручного труда авторами ранее был предложен алгоритм автоматического связывания. Более подробно о предлагаемом подходе можно прочитать в работах [1,2]. В данной работе приводится анализ наличия информации, которую можно использовать в процессе связывания, на примере базы данных “MedArt” и делаются выводы о принципиальной возможности применения подхода на практике. Цель проведенного анализа — выявление наиболее часто встречающейся информации для связывания и определение полей библиографического документа, на наличие которых можно рассчитывать в первую очередь.

1. Основная информация

В качестве библиографических документов в данной работе выступали библиографические записи в формате RUSMARC [3]. В качестве примера приводится фрагмент такой записи в текстовом формате (рисунок 1).

```
001 П15/А437-114799
701 1$aПанов $b А. А. $g Андрей Алексеевич $c физиология $f 19700224
$3APanov_AndrA2004050763480700 $p кафедра нормальной физиологии ПФ СГМУ
701 1 $a Ковалев $b И. В. $p Сибирский медицинский университет (Томск)
701 1 $a Бородин $b Ю. Л.
71202 $a Сибирский медицинский университет $c Томск
606 1# $a МЫШЦА ГЛАДКАЯ $x действия лекарственных препаратов
$3 D009130Q000187 $ 2mesh $8 rus
606 1 $a ФЕНИЛЭФРИН $x терапевтическое применение $3 D010656Q000627
$2 mesh $8 rus
```

Рисунок 1 - Фрагмент библиографической записи

Стандартами библиографического описания публикации предусматривается возможность упоминания персон, имеющих отношение к описываемой публикации, а также некоторой дополнительной информации, которая касается этих персон. В данной работе не

делается различий между авторами, редакторами, иллюстраторами и другими персонами, упоминаемыми в библиографическом описании. Для простоты все эти персоны считаются авторами публикации. С точки зрения формата RUSMARC это означает, что следующие поля рассматриваются единообразно:

700 — Имя лица — Первичная ответственность;

701 — Имя лица — Альтернативная ответственность;

702 — Имя лица — Вторичная ответственность.

Всего было рассмотрено 41746 упоминаний авторов. В таблице 1 приводятся поля библиографического документа, задействованные в работе и относящиеся непосредственно к автору. В рамках данной работы информация, указанная в этих полях носит название основной.

Таблица 1. Основная информация — перечень полей

Код	**Название поля**	**Информация**
$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 в полной форме	Имя и отчество
$p	Наименование/адрес организации	Место работы автора
$3	Номер авторитетной /нормативной записи	Авторитетный код

Поля $a и $b являются обязательными и присутствуют во всех рассмотренных упоминаниях. Поле $3 указывает на установленную связь с авторитетным документом. Следует отметить, что в настоящий момент поле $3 присутствует всего в 20% упоминаний, что указывает на актуальность проблемы установления связей.

Основная информация, относящаяся непосредственно к автору, является достаточно надежной с точки зрения связывания. Однако на практике она встречается достаточно редко. Так, два и более поля основной информации (исключая обязательные поля с фамилией и инициалами) присутствуют только в 21% упоминаний. Таким образом, не имеет смысла разрабатывать процедуру связывания, которую можно применить лишь к пятой части упоминаний авторов.

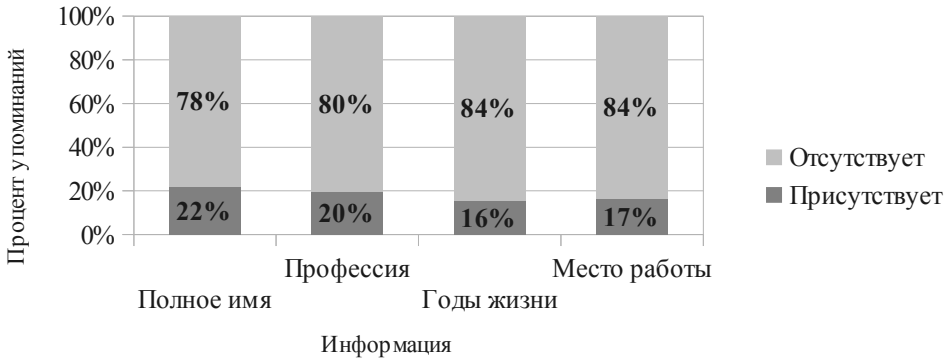

Рисунок 2 - Наличие основной информации

Для расширения области применения нашего подхода было принято решение задействовать косвенную информацию для связывания. Прежде всего, такая информация содержится в тех полях библиографического документа, которые не относятся непосредственно к автору, но характеризуют публикацию.

2. Дополнительная информация

В качестве дополнительной информации выступает, например, содержимое поля 712 «Коллективный автор». В этом поле форматом RUSMARC предусматривается возможность указания организации, в которой работает один или несколько упоминаемых авторов. Информация косвенная, поскольку не предусмотрено явное указание того, к какой именно из указанных организаций относится автор. Тем не менее, такая информация также может быть полезна. Наряду с наименованием организации, как правило, указывается и ее географическое местоположение. Информация о коллективном авторе присутствует для примерно 30% упоминаний авторов, поэтому необходимо дальнейшее расширение информации.

Следующий этап вовлечения дополнительной информации в процедуру связывания основан на сопоставлении соавторов, то есть всех персон, которые также указаны в библиографическом документе. При этом хранение информации о соавторах в авторитетном документе не предусмотрено форматом RUSMARC AUTHORITIES, поэтому необходимо воспользоваться технологией расширенных авторитетных документов. Суть технологии очень проста — рассматривать наряду с самим авторитетным документом еще и все библиографические, уже связанные с ним. Такой подход позволяет существенно расширить объем привлекаемой информации.

Относительно информации о соавторах следует иметь в виду, что ее можно разбить на два типа:

- более точная информация — указан код соответствующего авторитетного документа;
- менее точная — код отсутствует.

С точки зрения процедуры связывания эти типы различаются степенью надежности, поэтому будем рассматривать их по-отдельности (рисунок 3).

Еще один тип информации, которую можно использовать, содержится в полях 606 «Наименование темы как предмет». В рассматриваемой базе данных эти поля используются для указания предметных рубрик в соответствии с тезаурусом MeSH (Medical Subject Headings) [3], разработанным Национальной медицинской библиотекой Соединенных Штатов (NLM). В базе данных "MedArt" используется русскоязычная версия тезауруса, разрабатываемая в Центральной научной медицинской библиотеке [4].

Информация о предметных рубриках содержится в 70% упоминаний авторов. Причем, как правило, указывается не одна рубрика, а сразу несколько, что повышает ценность данной информации.

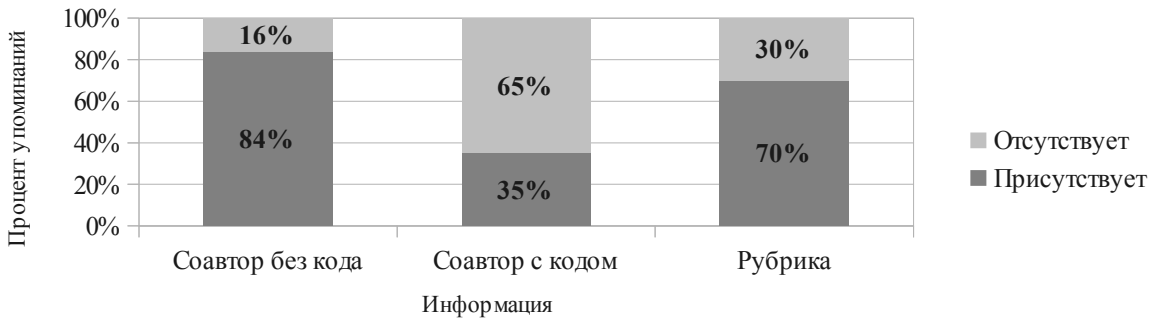

Рисунок 3 - Наличие дополнительной информации

Чтобы оценить насколько удалось расширить область применения подхода следует рассмотреть всю информацию вместе (рисунок 4). Основная либо дополнительная информация присутствует в 98% упоминаний авторов. В основном, такой результат обеспечивается наличием предметных рубрик и указания соавторов без кода авторитетных документов. Поэтому следует делать ставку именно на такую информацию.

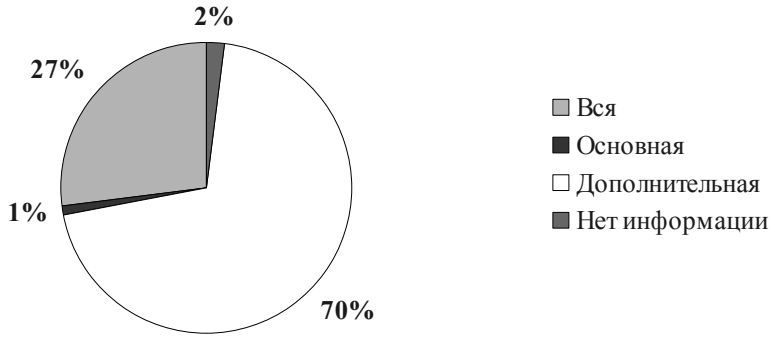

Рисунок 4 - Основная и дополнительная информация

Для того, чтобы повысить качество связывания было принято решение использовать те упоминания авторов, в которых присутствует как минимум два поля из списка полей для связывания.

С точки зрения точности связывания было решено рассматривать два типа требований к упоминаниям авторов: наличие хотя бы одного поля и наличие минимум двух полей (рисунок 5).

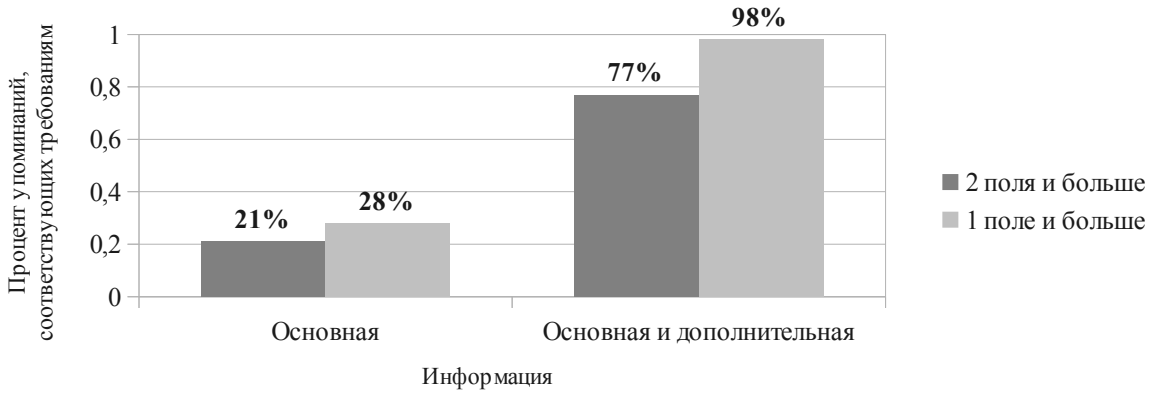

Рисунок 5 - Охват базы данных для разных требований

Таким образом, привлечение дополнительной информации позволило существенно увеличить объем упоминаний авторов, которые можно обрабатывать с помощью процедуры связывания с авторитетными документами.

Заключение

Как показал проведенный анализ базы данных "MedArt", процедура автоматического авторитетного контроля может применяться к рассмотренным данным только в том случае, если она будет использовать дополнительную информацию, в частности упоминания соавторов и предметных рубрик. Расширение информации позволило увеличить долю упоминаний с которыми может работать процедура связывания с 21% до 77% для более жесткого требования и с 28% до 98% для более мягкого.

В ходе дальнейшей работы мы планируем как можно более полно использовать всю рассмотренную информацию, а также по возможности вовлечь новую. Внедрение автоматического авторитетного контроля в распределенный электронный каталог медицинских библиотек НП «МедАрт» позволит улучшить возможности поиска и предоставления информации о медицинских публикациях, а также уменьшить трудоемкость поддержания электронного каталога в актуальном состоянии.

ЛИТЕРАТУРА

[1] Князева А.А., Турчановский И.Ю., Колобов О.С. Автоматический авторитетный контроль для распределенных библиографических баз данных [Электронный ресурс] // XIII Российская конференция с участием иностранных ученых «Распределенные информационные и вычислительные ресурсы»(DICR'2010): материалы конф. — Электрон. дан. — Новосибирск: ИВТ СО РАН, 2010. — 1 электрон. опт. диск (CD-ROM).— № гос. регистрации 0321100051.— http://conf.nsc.ru/dicr2010/ru/reportview/29244

[2] Князева А.А., Турчановский И.Ю., Колобов О.С. Автоматическое связывание документов //Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XIV Всероссийской научной конференции RCDL'2012. Переславль-Залесский, Россия, 15-18 октября 2012 г. — г. Переславль-Залесский: изд.-во «Университет города Переславля», 2012. — С. 360-369.

[3] Российский формат машиночитаемой каталогизации (RUSMARC): [Электрон. ресурс] / Мин-во культуры РФ, Российская библиотечная ассоциация. — Электрон. текст. дан. — СПб., 2004. — Режим доступа: http://www.rba.ru:8100/rusmarc/index.html

[3] Medical Subject Headings (MeSH) / U.S. National Library of Medicine [Электронный ресурс] – 8600 Rockville Pike, Bethesda, MD 20894, 2003-2012. – Режим доступа: http://www.nlm.nih.gov/mesh/, свободный. – Загл. с экрана.

[4] Колобов О. С. Представление тезауруса MeSH в формате RUSMARC посредством протокола Z39.50 / О. С. Колобов, Н. А. Мешечак, А. С. Карауш // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: Доклады и тез. докладов. — М.: ГПНТБ России, 2004. — 1 CD-ROM.