Извлечение библиографии из текстов регулярными выражениями

Колмогорцев С.В., Академия ФСО России
warrior63rus@mail.ru
Сараев П.В., Академия ФСО России
mayder_95@mail.ru

Аннотация

В данной работе на основе рассмотрения существующих международных стандартов библиографического описания и методов выделения библиографических списков, предложен алгоритм распознавания библиографических записей по ГОСТ 7.1 в русскоязычных текстах.

1 Введение

Всепроникающая информатизация общества привела к тому, что наблюдается взрывной рост числа ежегодно публикуемых книг и статей, в том числе электронных. Атрибутом любого издания является наличие библиографической информации, которая представляет собой по определённым правилам организованную информацию о документах, содействующую реализации соответствий между документами и их потребителями [Коршунов, 1985]. Разновидностями библиографической информации являются библиографические записи, совокупность которых составляет библиографический список. В свою очередь один и более списков задают библиографическое описание. При этом библиографическое описание можно рассматривать как структурированный объект, состоящий из последовательности сущностей. Причем ряд из них представлен именованными сущностями (named entity), которые являются наименованиями реальных объектов, например, персон, географических локаций, организаций, и т.д.

Проблематика автоматического извлечения именованных сущностей из текстов на естественных языках получила широкое распространение за последние 20 лет, что подтверждается динамикой встречаемости словосочетания «named entity recognition» в изданиях, рис. 1. Полученные за это время теоретические и практические результаты могут быть применены к разработке алгоритмов автоматического извлечения библиографической информации из текстов. Такие алгоритмы могут использоваться в качестве элементов перспективных систем машинного перевода для реализации этапа предпереводческого анализа текста [Гращенко и др., 2011]; в информационно-поисковых системах; в электронных библиотеках (для внедрения метаинформации в оцифрованные архивы и вычисления различных наукометрических показателей); а также для систем квалиметрии научных и учебных работ [Гращенко, Романишин, 2015].

Всё вышесказанное позволяет судить о важности выбранной тематики. В связи с этим в данной статье предпринята попытка систематизировать имеющиеся сведения в данной области, описать имеющиеся методы, применяемые к решению данной проблемы, а также определить направления для дальнейшей работы.

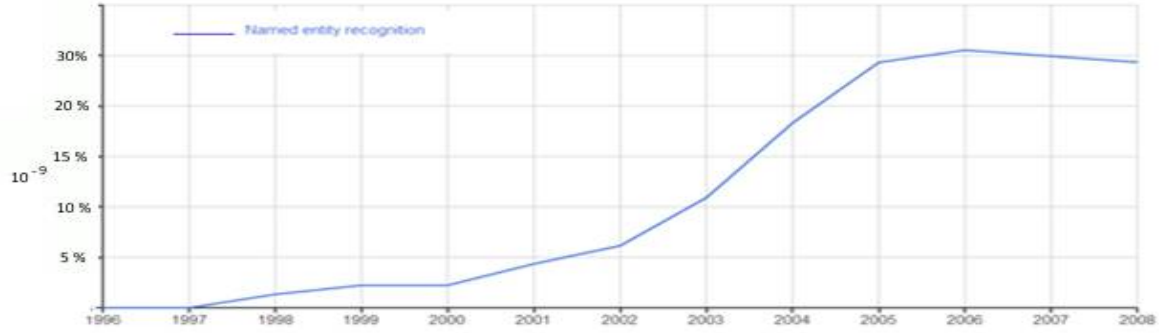

Рис. 1 Динамика встречаемости словосочетания «named entity recognition» (рисунок разработан авторами на основе онлайн-сервиса Google Ngram Viewer)

2 Описание предметной области

Согласно ГОСТ 7.1–2003, библиографическое описание представляет собой сведения о документе, приведенные по определенным правилам, устанавливающим наполнение и порядок следования областей и элементов, и предназначенные для идентификации и общей характеристики документа. Библиографическое описание является основной частью библиографической записи и составным элементом библиографического списка. Элементы библиографического описания, представлены на рис. 2.

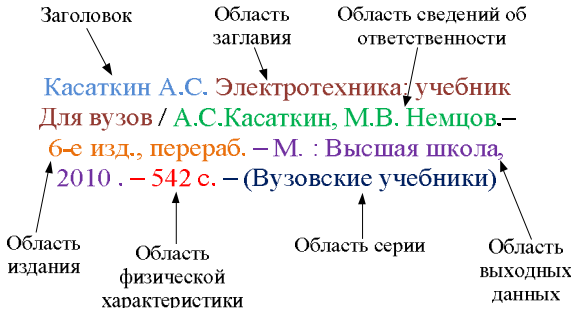

Рис. 2. Элементы библиографического описания (рисунок разработан авторами)

Ранее мы отмечали, что ряд этих элементов может рассматриваться как отдельная именованная сущность. Так, **заголовок** и **область сведений об ответственности** относятся к типу *PER (персоны)*, так как в них указываются авторы издания с указанием фамилии, имени и отчества. **Область заглавия**, содержит сведения о названии издания, относится к типу *TITLE (заголовок)*. **Область издания** содержит сведения об изменениях и (или) особенностях данного издания по отношению к предыдущему изданию того же документа где находится издательство, является типом *ORG (организации)*. **Область выходных данных** содержит сведения о месте и времени публикации, распространения и изготовления объекта описания, а также сведения об его издателе, распространителе, изготовителе, тип *LOCATION_OTHER (локация)*. Область **серии** – содержит сведения о многочастном документе, отдельным выпуском которого является объект описания, тип *MISC (разное)*. **Область физической характеристики** – содержит обозначение физической формы, в которой представлен объект описания, в сочетании с указанием объема и, при необходимости, размер документа, его иллюстраций и сопроводительного материала, являющегося частью объекта описания, тип *MISC (разное)*.

Предпочтительный способ и корректность извлечения библиографической информации из публикации зависит от ряда факторов. Одним из них является учет стиля оформления библиографической записи. В настоящее время в библиографоведении выделяют ряд основных международных стилей оформления публикаций, рис. 3. В данной статье за основу был взят стиль ГОСТ 7.1–2003.

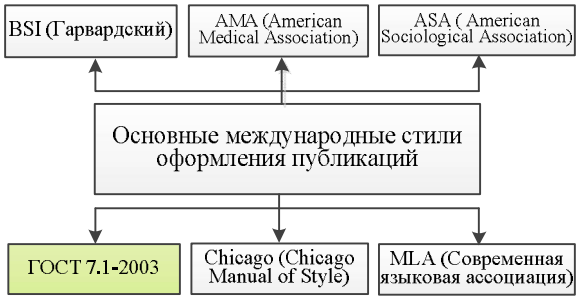

Рис. 3. Основные международные стили оформления публикация (рисунок разработан авторами)

Особенности решения задачи распознавания зависят от выбранного стиля библиографического описания, т.к. каждый вид оформления определяет как его структурные свойства (буквенно-цифровые индексы, детерминированные последовательности знаков пунктуации, заглавных букв и т.д.), так и статистические (распределение длин, вероятности появления отдельных символов и их последовательностей). Исходя из этого, к решению задачи извлечения библиографических описаний из текстов применимы как структурные, так и вероятностные методы распознавания, а также их комбинации. На практике широкое распространение получили методы, основанные на скрытых Марковских моделях (HMM), на основе классификации с помощью метода опорных векторов (SVM) и регулярных выражений [Васильев, 2015], рис 4. Далее рассмотрим особенности перечисленных методов и их

результативность применительно к текстам русскоязычных изданий.

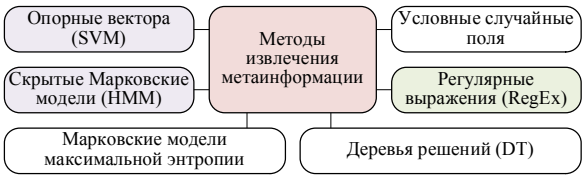

Рис. 4. Методы извлечения метаданных из текстов [Астахова, 2015]

3 Обзор методов извлечения библиографического описания

3.1 Метод, основанный на Марковских моделях

В работе [McCallum, 2005] для извлечения библиографического описания используется метод машинного обучения, основанный на скрытых Марковских моделях (HMM). Для его работы необходимо иметь промежуточное представление статьи, которое содержит символы перевода строк. Метод HMM двухэтапный. Вначале производится настройка модели на основе обучающей выборки, в которой размечены все элементы метаданных. При добавлении новых стилей оформления списка требуется увеличение обучающей выборки и повторное обучение. Далее на этапе распознавания уже настроенная модель используется для извлечения библиографического списка.

Недостатком данного статистического метода распознавания применительно к рассматриваемой задаче является то, что математический аппарат HMM основывается на предположении о линейной независимости компонентов вектора признаков, чего трудно добиться в практическом плане при обработке текстов.

3.2 Метод опорных векторов

В работе [Giles, 2003] для извлечения библиографического описания рассматривается статистический метод, основанный на классификации с помощью метода опорных векторов (SVM). Последовательность этапов и состав входных данных такие же, как в ранее рассмотренном методе. Особенностями текстовой классификации является как высокая размерность пространства признаков - десятки, иногда сотни тысяч, так и большое количество классов (рубрик) – от нескольких десятков до сотен. Проблема использования SVM на реальных коллекциях текстов связана как с указанными особенностями задачи текстовой классификации, так и с тем, что методы решения задачи квадратичной оптимизации известны, но вычислительно сложны.

В результате время обучения оказывается неприемлемо длительным, что является основным недостатком данного метода. Повышение скорости обучения на основе SVM возможно за счет использования многопроцессорных вычислительных систем и комплексов [Пескишева, 2013].

Опубликованные результаты экспериментов показывают превосходство данного метода над методом HMM, табл. 2.

3.3 Метод, основанный на регулярных выражениях

Для извлечения библиографического описания могут применяться структурные методы распознавания, среди которых выделим использование регулярных выражений (Regular Expressions - RegEx) для каждого элемента библиографического описания (примеры приведены в таблицах приложения).

Подстрока, соответствующая образцу поиска, заданному регулярным выражением, ищется по всему тексту. Образец содержит в себе слова, при нахождении которых в строке можно сделать вывод либо о принадлежности всей строки или ее части к соответствующему классу, либо о принадлежности соседнего фрагмента к какому-либо классу. Например, при нахождении словосочетаний, представленных в табл. 1 можно сделать вывод, что далее в тексте следует библиографический список.

Табл. 1. Возможные обозначения раздела с библиографическим списком

Название раздела
Библиографический список
Библиография
Источники литературы
Литература
Литературный перечень
Литературный список
Перечень библиографии
Перечень использованной литературы
Перечень использованных источников
Перечень использованных источников и литературы
Перечень использованной литературы и источников
Перечень используемой литературы
Перечень используемой литературы и источников
Перечень используемых источников
Перечень используемых источников и литературы

Перечень источников
Перечень источников и литературы
Перечень книг
Перечень литературы
Перечень цитируемой литературы
Список библиографии
Список использованной литературы
Список использованных источников
Список использованных источников и литературы
Список используемой литературы
Список используемых источников
Список используемых источников и литературы
Список источников
Список литературы

Табл. 2. Точность извлечения русскоязычной библиографии различными методами (в %)

Класс	Извлечено			Не извлечено		
	SVM	**HMM**	**RE**	**SVM**	**HMM**	**RE**
Ссылка	83,2	91,6	n/a	0	0	n/a
Заглавие	9,6	21,4	14	0	22,2	0
Авторы	14,9	37,5	18	0	9,8	0
Дата	61,3	90,8	n/a	3,8	4,6	n/a
URL	54,5	64,3	n/a	0	14,3	n/a
Страницы	25	66,7	n/a	25	16,7	n/a

Недостатком описанного метода является его высокая чувствительность к ошибкам, которые могут случайно или систематически возникать как в ходе оцифровки текста, так и при его формировании. Именно человек является основным фактором недетерминированности при размещении библиографической информации в текстах. В качестве примера стоит отметить склонность некоторых авторов к смешению стилей библиографического описания и опусканию необходимых элементов (год издания, инициалы авторов, номера страниц).

4 Опытные разработки

Для опытной работы был выбран метод, основанный на регулярных выражениях, как наиболее простой в реализации. Для этого на основании литературного обзора, анализа положений ГОСТ 7.1-2003 и эмпирического изучения разнообразных образцов изданий был подготовлен справочник регулярных выражений, описывающих элементов библиографической записи (см. приложение). Далее был предложен алгоритм извлечения библиографической информации и разработан реализующий его проблемно-ориентированный программный продукт, рис. 5. В результате последовавшей опытной работы был сделан ряд выводов, направленных на совершенствование имеющихся решений.

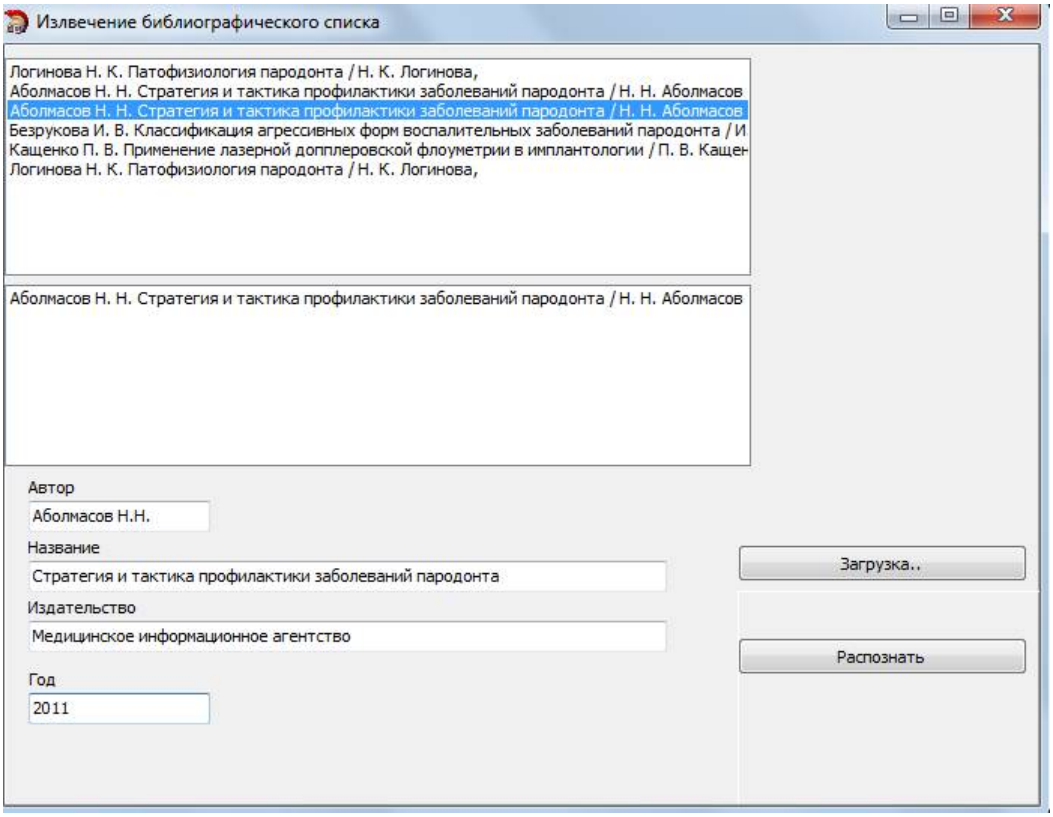

Рис. 5. Экранная форма рабочего окна программы

Предлагаемый порядок извлечения библиографической информации описывается блок-схемой, приведенной на рис. 6. В его основе лежит необходимость первичной сортировки текстов, так как существуют отличия между структурированными текстами (статья, реферат, диссертация и т.д.) и неструктурированными.

1) Извлечение списка литературы из структурированных текстов происходит следующим образом:

1.1) Выделение в тексте структурных частей (сегментация) с отнесением их к одному из трех классов: 1 - Библиографический список, 2 - Основная часть, 3 - Вспомогательные элементы (такие как введение, заключение, аннотации, приложения и т.п.).

Данная процедура облегчает вычисления по нахождению необходимой информации, так как каждый из классов имеет свои особенности по обработке. Например, большая часть библиографической информации в публикации приходится именно на библиографический список. Подобная информация так же может встречаться во «Введении», «Заключении», «Аннотации». Чаще всего библиографическая информация находится в начале и конце публикации.

1.2) Для сегментов текста первого класса применяются регулярные выражения. Выделенная информация сохраняется.

1.3) Во втором классе сегментов библиографические данные могут не содержаться в явном виде. Зачастую библиографические описания и затекстовые ссылки заключают в круглые или квадратные скобки. Поэтому следующим этапом алгоритма является поиск этих меток в тексте. Вспомогательным условием поиска скобок могут служить кавычки, обозначающие прямую речь, после которой предположительно может находиться ссылка на литературу, откуда бралось высказывание, например:

В конце 30-х – начале 40-х годов В.И. Вернадский сам писал по поводу этой работы: «Многое теперь пришлось бы в ней изменить, но основа мне представляется правильной» (Вернадский В.И. Размышления натуралиста. М., 1977. Кн. 2: Научная мысль как планетное явление. С. 39).

При нахождении таких конструкций в тексте, они целиком записываются в массив для дальнейшей обработки и выделения именованных сущностей.

1.4) В третьем классе сегментов наиболее частыми являются затекстовые ссылки, заключенные в квадратные скобки. Содержимое между скобок может быть различным, для его отыскания используются регулярные выражения, перечисленные в таблице 1 приложения.

1.5) Результаты по всем трем классам сводятся в один файл (базу данных).

2) Для работы с неструктурированными текстами необходимо следующее:

2.1) Сегментировать текст на приблизительно равные части относительно небольшого объема с учетом имеющейся разметки абзацев или предложений. Такая сегментация необходима для снижения затрат памяти при возможном применении рекурсивных функций.

2.2) Каждая из частей проверяется на наличие в ней внутритекстовых или затекстовых ссылок. В случае если такая проверка ничего не дала, то переходим на следующий шаг.

2.3) Проверка наличия имен собственных. Наличие фамилий, имен, отчеств может указывать на наличие библиографических данных, ведь составитель текста мог просто не указать скобки, что приведет к неправильной работе алгоритма. На такой случай необходимо:

2.3.1) Удалить слова, с которых начинаются предложения.

2.3.2) Выделить имена собственные, после них проверить структуру внутритекстовой ссылки.

2.3.3) В случае если имя собственное стояло в начале предложения, необходимо проверить, встречается ли оно в других частях текста, что даст повторы и укажет на то, что это действительно имя собственное.

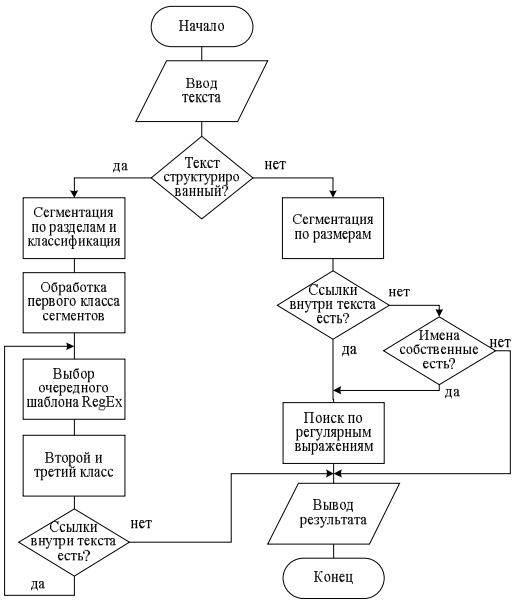

Рис. 6. Алгоритм работы выделения библиографической информации из текста

5 Выводы

Практическое использование описанного алгоритма позволило расширить и систематизировать состав проблем, стоящих перед разработчиками перспективных систем извлечения библиографической информации из русскоязычных текстов. Во-первых, для более-менее корректного извлечения всей библиографической информации необходимы десятки-сотни регулярных выражений. Это вызвано тем, что многие авторы при составлении библиографических описаний своей публикации не всегда следуют определенным стилям. К тому же, возможны многочисленные виды систематических и случайных ошибок, возникающих при наборе, конвертации или оцифровке текстов. Поэтому применение метода регулярных выражений для извлечения библиографической информации на практике оказывается недостаточно. Поэтому применение структурных методов распознавания в чистом виде не продуктивно. Целесообразно сочетать аппарат регулярных выражений со статистическими методами для локализации возможного размещения библиографической информации, оценки стилей оформления и т.д.

В связи с этим представляется важным учет информации о длинах библиографических описаний для сегментации текста. Так, из электронной научной библиотеки Российского индекса научного цитирования[1] были отобраны 128 библиографических описаний научных статей различных тематик и построена гистограмма распределения их длин в символах, рис. 7.

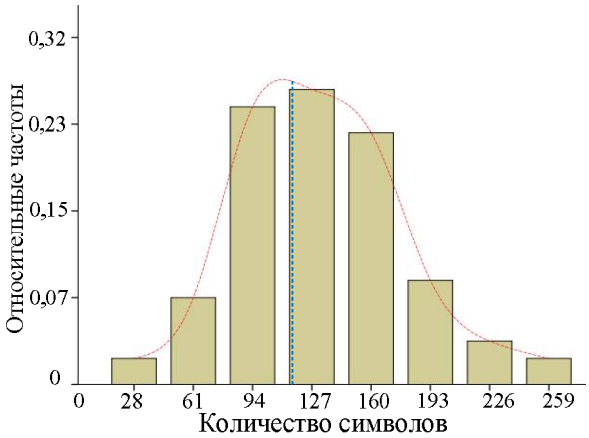

Рис. 7. Гистограмма и огибающая распределения длин библиографических описаний (рисунок разработан авторами)

Установлено, что средняя длина одной записи составила 116 символов, а распределение близко к нормальному. Предполагается, что при увеличении выборки огибающая распределения длин библиографических записей будет приближаться логнормальным законом. Данную информацию использовать для оконной обработки текстовых сегментов, где при нахождении с помощью регулярных выражений в ограниченной области нескольких именованных сущностей будет приниматься решение об отнесении текущего участка текста к библиографическому описанию.

Обзор тематики статьи и полученный авторами опыт показали целесообразность комплексного подхода к решению задачи извлечения библиографических описаний. Таким образом, возможно нивелирование недостатков отдельных методов и получения синергетического эффекта. Следующим шагом в направлении извлечения библиографической информации станет доработка предложенного алгоритма за счет добавления процедур статистической обработки.

[1] www.elibrary.ru

Список литературы

Коршунов О. П. *Библиографическая информация как научное понятие* // Советская библиография. – 1985. – № 3. – С. 31–42.

Гращенко Л.А., Клышинский Э.С., Тумковский С.Р., Усманов З.Д. *Концептуальная модель системы таджикско-русского машинного перевода* // Доклады Академии наук Республики Таджикистан. – 2011. – №4. Том 6. – С. 279-286.

Гращенко Л.А., Романишин Г.В. *Опыт автоматизированного анализа повторов в научных текстах* // Новые информационные технологии в автоматизированных системах. – 2015. – №18. – С. 582-590.

Васильев А., Козлов Д., Самусев С., Шамина О. *Извлечение метаинформации и библиографических ссылок из текстов русскоязычных научных статей* // Труды конференции RCDL2007, Переславль. – 2007. – Том 1. – С. 175-181.

Астахова Д.И. *Извлечение именованных сущностей с использованием Википедии* // Московский государственный университет имени М.В. Ломоносова. – 2015. - С. 40.

Пескишева Т.А. *Параллельная реализация алгоритма обучения системы текстовой классификации* // Вятский государственный гуманитарный университет. – 2013. - С. 597-607.

Чуприна К.В. *Исследование и разработка методов извлечения именованных сущностей из текстов с использованием структуры категорий Википедии*: дипломная работа. – М.: МГУ им. М.В. Ломоносова, 2014. - 46 с.

Герасимов, А.М. Елизаров, Е.К. Липачев, Ш.М. Хайдаров. А.Н. *Методы автоматизированного извлечения метаданных научных публикаций для библиографических и реферативных баз цитирования* // Казанский (Приволжский) федеральный университет, 2016. - С. 41-473.

Зацман И.М. *Метод извлечения библиографической информации из полнотекстовых описаний изобретений* / И. М. Зацман, В.А. Хаванских, С.К. Шубников // Информатика и ее применение. – 2013. - Т. 7, Вып.4. - С.52-65.

Milosavljević B. Danijela D. Surla *Retrieval of bibliographic records using Apache Lucene*. – 2010.

McCallum A. Information *Extraction Distilling Structured Data from Unstructured Text* // ACM Queue. - 2005. - Vol. 3. – P. 4.

Giles L. et al. *Automatic Document Metadata Extraction using Support Vector Machines* // JCDL. - 2003. - P. 12.

Приложения

Табл. 1. Примеры регулярных выражений для внутритекстовых ссылок

Пример	Регулярное выражение
[59].	\[[0– 9]+\]
[10, с. 81]	\[[0– 9]+\,( )*[с|С]+\.( )*[0– 9]+\]
[Пахомов, Петрова]; [Сергеев, Латышев, 2001; Сергеев, Крохин, 2000]	\[(( )?[А– Я]{1}[а– я]+,?( )?)*([0– 9]+)?(;)?)+\]
Пахомов В.И., Петрова Г.П. Логистика. М.: Проспект, 2006. 232 с.	(([А– Я]{1}[а– я]+,?( )?)* ([А– Я]\.)+,?)+ ?[А– Яа– я]+\.
[Нестационарная аэродинамика баллистического полета]	\[([А– Яа– я]+( )?)+\]
[Бахтин, 2003, с. 18] [Целищев, ч. 1, с. 17] [Гордлевский, т. 2, с. 142; Алькаева, Бабаев, с. 33– 34]	\[[А– Яа– я]+,( )?([0– 9]+|[а– я]\.( )?[0– 9]),( )?[с|С]?\.( )?[0– 9]+\]
[Философия культуры ... , с. 176]	\[([А– Яа– я]+)+(\.)+( )?,( )?[с|С]?\.( )?[0– 9]+\]

Табл. 2. Примеры регулярных выражений для элементов библиографического описания

Элемент библиографического описания	Регулярные выражение
Заголовок	[А– Я]{1}[а– я]{1,20}( )?([А– Я]{1}\.)*
Область заглавия	( )?[А– Я]{1}([а– я:,]+)\/
Область сведений об ответственности	(([А– Я]\.( )?){1,2}[А– Я]{1}[а– я]{1,20}(,)?)+\.
Область издания	( )?[0– 9]+(–)?[е|Е]( )?[А– Яа– я\.\, \–]+
Область физической характеристики	[0– 9]+( )?[с|С]\.
Область серии	\([А– Яа– я]+\)

Область выходных данных	:( )?([А– Яа– я,]+)+[0– 9]+