Вопросы автоматизации обработки библиографических ссылок для справочно-библиографической системы по инженерной геометрии

Some issues of automation of information extracting from the bibliographic references for a bibliographic system on engineering geometry

Васнев И.П.
Студент РТУ МИРЭА
e-mail: ivasnev2002@gmail.com

Vasnev I.P.
Student, MIREA – Russian Technological University
e-mail: ivasnev2002@gmail.com

Аннотация
Рассматривается задача автоматической обработки библиографической ссылки на публикации в области инженерной геометрии. В ходе обработки из текста ссылки выделяются существенные элементы данных – фамилии и инициалы авторов, название публикации, название сборника или журнала, город и год издания, издательство, номера и число страниц. Автоматическая обработка не только позволяет создавать перекрестные ссылки между публикациями, но и, в случае успешного выделения элементов данных из текста ссылки, генерировать (дополнять) сведения о публикациях, изданиях, авторах и др. Для автоматизации обработки текстов ссылок в работе предлагается использовать регулярные выражения. Для этого подробно описывается формат элементов данных. Приводятся результаты работы программного модуля, выполняющего обработку текста ссылок. Делаются выводы по работе.
Ключевые слова: инженерная геометрия, библиографическая система, регулярные выражения.

Abstract
Automatic processing of bibliographic references to publications is considered into a field of engineering geometry. During the course of processing the essential data elements are extracted from the text of links. For example, the surnames and initials of the authors, the title of the publication, the title of the collection or journal, the city and year of publication, publisher, numbers and number of pages. Automatic processing not only allows you to create cross-references between publications, but also, in case of successful selection of separating data elements from the text links, generate (supplement) information about publications, authors, etc. To process link texts in the work, it is proposed to use regular terms. For the deal, the format of the data items is described in detail. The results of the work of the software module, that performs the processing of the text of links are presented. Conclusions are drawn on the work.
Keywords: engineering geometry, reference bibliographic system, regular expression.

1. В работах [1–2] была поставлена задача создания справочно-библиографической системы по инженерной геометрии (СБСИГ), намечены основные этапы ее создания. В текущем состоянии СБСИГ может стать серьезным помощником на этапе сбора информации в научных исследованиях, например, в [3–8]. В [1] следующими этапами работы над СБСИГ назва-

ны – внесение сведений о диссертациях и авторефератах и обработка библиографических ссылок. В настоящей работе анализируются результаты, полученные в ходе разработки программного модуля для автоматического анализа текста библиографических ссылок.

2. Задача анализа библиографической ссылки состоит в следующем. Ссылка представляет собой текстовую строку определенного формата. Как правило, это информация о публикации в сборнике (журнале), рукописи (автореферат или диссертация), монографии или учебнике, многотомном издании, реже – авторском свидетельстве или патенте. Приведем примеры ссылок из библиографических списков некоторых публикаций:

Глазунов Е. А., Четверухин Н. Ф., Аксонометрия, Госизд. теоретико-технической литературы, М., 1953.

Четверухин Н. Ф., Введение в высшую геометрию. Учпедгиз. М., 1934.

Русскевич Н. Л., Новые методы вычерчивания наглядных изображений в аксонометрических и центральных проекциях, М.—Л., 1953.

Jouffret E. Traitéelèmèntaire dc gèomètrie a quatre dimensions. — Paris, 1903. — 212 p.

Семушин А. Д. Методика обучения решению задач на построение по стереометрии, АПН РСФСР, М., 1959.

Джапаридзе И. С. Геометрические преобразования пространства и их применение. Сб. «Методы начертательной геометрии и ее приложение» (под ред. Н. Ф. Четверухина), М., 1955.

Авторское свидетельство № 134127, от 1/X 1960 г. «Бюллетень изобретений», № 23, 1960.

Гаспар Монж. Приложение анализа к геометрии. М., Объединенное научно-техническое издательство. 1936.

Котов И. И. Графические способы задания и построения технических форм поверхностей. Диссертация, М., 1959.

Как видно из приведенных примеров, несмотря на значительное сходство в структуре различных ссылок, которое легко позволяет человеку распознать содержание ссылки, наблюдается значительная вариативность, которая делает задачу распознавания довольно сложной.

В настоящей работе мы будем ограничиваться ссылками на статьи, монографии, учебники, доклады и прочие печатные публикации. Диссертации, авторефераты, отчеты и пр. имеют сходную, но несколько иную структуру (указывается шифр специальности, ученая степень, научная область и др.), поэтому в настоящей работе они не рассматриваются.

Примерная структура ссылки выглядит следующим образом:

<Авто-
ры><Название><Издание><Город><Издательство><Год><Номера страниц><Число страниц>

Не все ссылки обязательно содержат полный набор элементов: например, <Издание>есть у статей и докладов, но не у монографий или учебников; издательство может отсутствовать, а <Номера страниц> и <Число страниц> – взаимоисключающие элементы.

Процедура обработки ссылки должна решать две задачи:

1. Распознавание данных в строке. Если обнаруживается, что такая публикация есть в базе данных, тогда необходимо создать перекрестную ссылку на нее.
2. Порождение данных на основе информации в строке. Если обнаруживается, к примеру, что статьи нет в базе данных, но есть сборник (внесенный в базу данных частично), в котором она напечатана, информацию сборника требуется дополнить. Основные процедуры порождения данных включают следующее:
    - публикация есть, но информация о ней не полна (например, нет номеров страниц) – информация дополняется;

- публикации нет, но авторы есть в базе данных, и есть издание, частью которого она является – создается элемент данных для публикации с перекрестными ссылками на авторов и издание;
- публикации нет, есть издание, но нет одного или нескольких авторов – создаются элементы данных для авторов и публикации с соответствующими перекрестными ссылками;
- нет публикации и издания, возможно, одного или нескольких авторов – создаются все соответствующие элементы данных с перекрестными ссылками.

Таким образом, простого распознавания по тексту ссылки имеющейся в базе данных публикации, которое сравнительно просто реализовать на основе словаря, не достаточно. Требуется тщательный разбор текста ссылки, точное распознавание каждого элемента, который затем может быть использован для порождения данных.

3. Основной способ обработки текста ссылки – использование регулярных выражений, которые распознают вышеперечисленные элементы информационной структуры. Перечислим их особенности:

1. <Авторы>:

Формат: Одно или два слова, разделенные дефисом, пробелы, один или два инициала, оканчивающиеся точкой. Несколько авторских элементов разделяются союзом ‘и’ или запятой. Примеры:

- СЕМЕНЦЕВ-ОГИЕВСКИЙ М. А.
- ГЛАЗУНОВ Е. А.
- МОНЖ Г.

2. <Название>:

Формат не важен, выделяется по остаточному принципу между элементом <Авторы>и <Издание> (если есть) или <Город>.

3. <Издание>:

Формат: В данный элемент попадает любая информация, стоящая после длинного тире ‘—’, если таковая имеется. Также проверяется наличие в строке слов (‘Труды’, ‘Тр.’, ‘Записки’, ‘Сб.’ и т.п.), если таковые находятся – выделяется информация до элемента<Город>.

4. <Город>:

Формат: Дефис или длинное тире, за которым следует название города из таблицы данных city, в том числе краткие названия – ‘М.’, ‘Л.’, ‘М.—Л.’. Примеры:

- - Л.
- —М.—Л.
- —Омск

5. <Издательство>:

Формат: Название организации из таблицы данных org, включая краткий вариант, или последовательность символов после ‘:’, следующего за элементом <Город>.

6. <Год>:

Формат: Четыре десятичные цифры. Диапазон лет может быть задан добавлением дефиса или длинного тире и еще двух или четырех десятичных цифр. Примеры:

- 1952
- 1951—1954
- 1951—54

7. <Число страниц>:

Формат: Дефис или длинное тире, за которым следует число и литера, означающая страницы (‘p’, ‘c’, ‘s’ и т.п.), с точкой. Примеры:

- — 212 p.
- —54 S.
- -420 c.
- -420 C.

8. Номера страниц:

Формат: Дефис или длинное тире, за которым следует литера, означающая страницы (‘pp’, ‘c’, ‘C’, ‘s’, ‘S’ и т.п.), с точкой, число, дефис или длинное тире и снова число. Примеры:

- c. 229—242.
- S. 109—148.

4. Результаты экспериментов:

Перечисленные выше особенности элементов структуры библиографической ссылки могут быть представлены в виде регулярных выражений. Был создан программный модуль на языке Python, содержащий процедуру обработки списка библиографических ссылок. Список разделяется на строки. Каждая строка проверяется последовательно созданными регулярными выражениями до момента, когда хотя бы одно подтвердит корректность ее структуры. Если ни одно выражение не подтверждает корректность структуры, строка отвергается.

Приведем примеры работы программного модуля:

Корректное распознавание:

1. Входные данные:

ГЛАЗУНОВ Е. А., ЧЕТВЕРУХИН Н. Ф., Аксонометрия, Госизд. теоретико-технической литературы, М., 1953.

Результат работы:

Авторы ('ГЛАЗУНОВ Е. А.', 'ЧЕТВЕРУХИН Н. Ф.'), Название (Аксонометрия), Издательство (Госизд. теоретико-технической литературы), Издание(), Город издания (Москва), Число страниц (), Номера страниц (), Год издания (1953).

2. Входные данные:

Четверухин Н. Ф., Введение в высшую геометрию. Учпедгиз. М., 1934.

Результат работы:

Авторы ('Четверухин Н. Ф.'), Название (Введение в высшую геометрию.), Издательство (Государственное учебно-педагогическое издательство Министерства Просвещения РСФСР), Издание(), Город издания (Москва), Число страниц (), Номера страниц(), Год издания (1934).

3. Входные данные:

Jouffret E. Traitéelèmèntaire dc gèomètrie a quatre dimensions. - Paris, 1903. — 212 p.

Результат работы:

Авторы('Jouffret E.'), Название(Traitéelèmèntaire dc gèomètrie a quatre dimensions.), Издательство(), Издание(), Город издания(Paris), Число страниц(212), Номера страниц (), Год издания (1903).

Некорректное распознавание:

1. Входные данные:

Громов М. Я. Начертательная геометрия, ч. II, М., Изд. ВЗПИ. 1954.

Результат работы:

Авторы ('Громов М. Я.'), Название (Начертательная геометрия, ч.), Издательство (Изд. ВЗПИ.), Издание(II), Город издания (Москва), Число страниц (), Номера страниц (), Год издания (1954).

2. Входные данные:

РУССКЕВИЧ Н. Л., Новые методы вычерчивания наглядных изображений в аксонометрических и центральных проекциях, М.—Л., 1953.

Результат работы:

Авторы ('РУССКЕВИЧ Н. Л.'), Название (Новые методы вычерчивания наглядных изображений в аксонометрических и центральных), Издательство (), Издание (проекциях,), Город издания (Москва, Ленинград), Число страниц (), Номера страниц (), Год издания (1953).

3. Входные данные:

Семушин А. Д. Методика обучения решению задач на построение по стереометрии, АПН РСФСР, М., 1959.

Результат работы:

Авторы ('Семушин А. Д.'), Название (Методика обучения решению задач на построение по стереометрии, АПН) , Издательство (), Издание(РСФСР), Город издания (Москва), Число страниц (), Номера страниц (), Год издания (1959)

Как показали эксперименты, программный модуль корректно распознает только 40–50% библиографических ссылок. Основные ошибки связаны со сложной формализацией ключевого элемента структуры ссылок – сведениями об издательстве. Так, названия организаций могут быть записаны необычным способом, организация может отсутствовать в базе данных. Поскольку элемент <Издательство>является опорным при выделении структуры в разработанном способе, ошибка его распознавания приводит к другим ошибкам.

5. Выводы по работе. Проведен подробный анализ проблемы распознавания структуры библиографической ссылки. Установлено, что процедура обработки текста ссылки должна решать две задачи – распознавания и порождения данных. Показано, что задача распознавания полной структуры библиографической ссылки является трудно формализуемой. Разработан программный модуль, который позволяет распознать примерно 40–50% библиографических ссылок на монографии, учебники, статьи и доклады в журналах и сборниках.

Дальнейшие исследования позволят улучшить показатели распознавания:

- Накапливать и использовать специальный словарь аббревиатур и сокращений, встречающихся в ссылках.
- До запуска основного алгоритма потребовать ручной расстановки маркеров начала ключевых элементов библиографической ссылки.

Литература

1. *Бойков А.А., Варфоломеева А.А., Идрисова Ф.С., Пентюрина В.Р.* О создании библиографической базы публикаций по инженерной геометрии // Надежность и долговечность машин и механизмов. Сборник материалов IX Всероссийской научно-практической конференции. – Иваново, 2018. – С. 404-407.
2. *Бойков А.А.* О текущем состоянии справочно-библиографической системы по инженерной геометрии // Журнал технических исследований. – 2020. – Т.6. – №2. – С. 29–34.
3. *Антонова И.В., Соломонова Е.В., Кадыкова Н.С.* Математическое описание частного случая квазивращения фокуса эллипса вокруг эллиптической оси // Геометрия и графика. – 2021. –Т.9. – №1. – С. 39–45. – DOI: 10.12737/2308-4898-2021-9-1-39-45
4. *Бойков А.А.* Компьютерная проверка решений задач начертательной геометрии для инженерно-графического образования // Геометрия и графика. – 2020. – Т. 8, №2. – С. 66–81.–DOI: 10.12737/2308-4898-2020-66-81
5. *Назарова О.Н.* Адаптация дисциплины «прикладная геометрия» к программам бакалавриата эксплуатационных направлений авиационного вуза// Геометрия и графика. – 2020. – Т. 8. – №1. – С. 57–64.–DOI: 10.12737/2308-4898-2020-57-64
6. *Назарова О.Н.* Анализ некоторых задач курса теоретической механики, решаемых методами начертательной геометрии // Геометрия и графика. – 2019. – Т. 7. – №4. –С. 76–83. –DOI: 10.12737/2308-4898-2020-76-83
7. *Бойков А.А*. О построении моделей объектов пространства четырех и более измерений в учебном процессе // Геометрия и графика. – 2018. – Т. 6. – № 4. – С. 54–71. – DOI: 10.12737/article_5c21f96dce5de8.36096061
8. *Абдурахманов Ш.А.* Применение механизмов, отмечающих центры тяжестей симплексов в их 2-мерных проекциях как аксонографов многомерных пространств // Геометрия и графика. – 2020. –Т.8. – № 4. –С. 13–23. –DOI: 10.12737/2308-4898-2021-8-4-13-23