Автоматическое наполнение информационных систем библиографическими сведениями о научных публикациях♣

© Ю.А Загорулько, О.О. Дяченко

Институт систем информатики имени А.П. Ершова СО РАН, г. Новосибирск
zagor@iis.nsk.su, dyachenko.oleg@gmail.com

Аннотация

В докладе описывается подход к автоматизации наполнения информационных систем библиографическими сведениями о научных публикациях. В рамках этого подхода разработан метод автоматического построения формальных описаний научных статей, а также метод автоматического добавления таких описаний в контент портала научных знаний.

1. Введение

В связи с бурным ростом числа публикаций по различным отраслям знаний становится все труднее найти нужную печатную работу. Для облегчения этой задачи создаются электронные архивы научных статей [8, 9], информация о них представляется в базах данных (БД) цитирования [7, 12, 16] и порталах научных знаний [11, 17]. Главным требованием к таким информационным системам является обеспечение пользователя гибкими и удобными средствами поиска, навигации и доступа к представленным в них статьям. Причем эти системы должны предоставлять пользователю не только метаданные статьи (название, авторов, аннотацию, ключевые слова, библиографическую ссылку и т.п.), но и информацию о ее связях с другими публикациями.

Для того, чтобы информационная система оставалась актуальной, необходимо постоянно пополнять ее сведениями о новых публикациях. Сбор таких сведений является очень трудоемким процессом – для каждой научной статьи необходимо получить ее формальное описание, включающее ее основные атрибуты и список содержащихся в ней библиографических ссылок, а затем добавить это описание в конвент информационной системы, обеспечив его согласованность и связанность с ранее введенными данными. Вручную выполнить такую работу достаточно сложно, поэтому предпринимаются попытки ее автоматизации.

Рассмотрим наиболее значимые подходы к созданию баз данных цитирования.

Согласно предложенной Дэвидом Сонгом модели универсальной базы данных цитирования [18] все публикации должны быть описаны в стандартном формате. Ссылки на цитаты (библиографические ссылки) должны даваться в XML-документе, построенном согласно определенной структуре. Дэвидом Сонгом была представлена такая XML-структура, в которой были определены правила описания всех публикуемых статей.

Разработанная исследовательским институтом корпорации NEC автономная система индексирования цитат ResearchIndex [3] автоматически индексирует публикации по информатике, обнаруженные в Web.

Отдельный класс составляют так называемые менеджеры ссылок (Reference managers) [1, 2, 4], позволяющие пользователям, как правило, авторам публикаций или исследователям, создавать и использовать свои локальные базы данных цитирования. Большинство из предлагаемых программ этого класса не обеспечивают автоматического получения информации о статье из внешних файлов. Исключение составляет лишь Mendeley Desktop [4], работающий с PDF-файлами, однако он обеспечивает низкое качество извлечения информации о статье. Помимо этого ни одна из рассмотренных программ не работает с неразмеченными документами.

Каждый из рассмотренных подходов имеет свои достоинства и недостатки. В одних подходах вся «работа» возложена на автора или издательство, в других наоборот – ручной труд практически исключен, но при этом ухудшается качество поиска, так как некоторые атрибуты статьи (например, автор или заголовок) не всегда правильно извлекаются из текста.

Таким образом, задача разработки системы, которая бы позволяла автоматически извлекать библиографические сведения из неразмеченных текстов научных публикаций и заносила бы их в

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

информационные системы, является актуальной. Рассмотрению одного из подходов к решению этой задачи и посвящен данный доклад.

2. Построение формального описания научных статей

Рассмотрим метод автоматического построения формальных описаний научных статей. Прежде всего, отметим, что формальное описание статьи состоит из двух блоков. Первый из них, включает основные характеристики статьи, извлекаемые из ее вводной части: «название», «авторы», «аннотация», «ключевые слова». Второй блок содержит описание содержащихся в статье библиографических ссылок (цитируемых в статье публикаций). Каждое из этих описаний включает такие атрибуты, как «название», «авторы», «год издания», «название журнала», «том», «выпуск», «первая страница», «последняя страница», «URL» и другие.

Автоматическое построение формального описания научной статьи выполняется с использованием эвристических правил и иерархической системы шаблонов и включает следующие этапы:

(1) с помощью эвристических правил и с опорой на маркеры (характерные слова или словосочетания) выделяются основные разделы статьи (заголовок, список авторов, аннотация, ключевые слова, список литературы),

(2) на основе анализа выделенных разделов определяются основные характеристики статьи,

(3) на основе иерархической системы шаблонов и регулярных выражений выполняется синтаксический разбор библиографического списка и формируется его формальное представление,

(4) все полученные данные о статье заносятся в базу данных цитирования (библиографических ссылок).

Под синтаксическим разбором элемента библиографического списка или цитаты в п.3. понимается определение входящих в нее полей и нахождение их значений. Согласно нашему подходу, синтаксический разбор осуществляется путем сопоставления цитаты шаблонам, описанным следующим образом:

<шаблон> ::= {<блок-поле>|<символьный блок>}+

Блок-поле в записи шаблона представляет собой имя поля, заключенное в угловые скобки. Определение в цитате значения некоторого блок-поля происходит при помощи низкоуровневых (частичных) шаблонов, описанных на языке регулярных выражений (PCRE), путем нахождения им соответствий в цитате. В случае, если цитата подходит под шаблон, то указанным полям ставится в соответствие их текстовые значения. Каждый из шаблонов имеет свой вес, используемый как величина его авторитетности. Символьные блоки – это просто набор символов, как правило, присутствующий в шаблоне для описания характерных для библиографической ссылки элементов.

Символьные блоки располагаются в шаблоне между блоками-полями и позволяют существенно улучшить результаты работы отдельного шаблона.

Применение каждого шаблона происходит по следующему алгоритму:

1) Если в шаблон входят символьные блоки, проверяется их наличие и правильная последовательность в ссылке, иначе – переход к следующему шаблону;

2) Каждая из частей цитаты, заключенная между символьными блоками, проверяется на соответствие блокам-полям, стоящими в шаблоне между ними, в случае несоответствия – переход к следующему шаблону;

3) Если всем полям цитаты установлены соответствия, то цитата считается разобранной.

Достоинством предложенного метода построения формальных описаний научных статей является возможность его декларативной настройки на коллекцию документов, подлежащих обработке. Это необходимо в связи с тем, что правила оформления статей, особенно списка цитируемой литературы, у разных изданий различаются. Такая настройка обеспечивается путем модификации набора высокоуровневых (полных) шаблонов, обеспечивающих обработку статей, и задания им весовых коэффициентов, определяющих их авторитетность при разборе библиографических ссылок. (Более подробно метод автоматического построения формального описания научной статьи описан в [14].)

К настоящему времени разработан конструкторский интерфейс, позволяющий редактировать полные, иерархические шаблоны, и частичные, низкоуровневые шаблоны на языке PCRE [5]. С использованием конструкторского интерфейса был сформирован набор шаблонов, покрывающий наиболее часто используемые форматы описания библиографических ссылок. Разработан модуль, реализующий указанный метод автоматической обработки текста статьи, средствами СУБД MySQL реализована база данных цитирования, а также пользовательский интерфейс, позволяющий просматривать и редактировать полученные формальные описания статей.

При построении формального описания статьи порождаются два основных вида объектов – авторы и публикации. Рассмотрим их структуру подробнее.

Каждый объект-автор имеет набор полей, которые могут иметь либо одно, либо множество значений. Схема представления такого объекта имеет вид:

Author = <Id_Author, Last_name, First_name, Second_name, Email, Work_phone, Cordinates, Initials, Syn_name, Url, Place_work, Place_work_short, Place_live>.

Первый элемент такого объекта (*Id_Author*), задает уникальный идентификатор персоне – автору публикации, смысл других полей понятен из их

названия. Заметим, что автор может иметь несколько мест работы и рабочих телефонов.

Объект-публикация представляется следующей структурой:

Publication = < Id_Publication, Title, Year, Journal, Volume, Issue, Chapter, Number, Start_page, End_page, Keywords, Alternate_title, Abstract, Url, Date, Publisher, Conference_location, Series_title, City, ISBN, Number_of_pages, Conference_name, Edition, Language, AUTHORS, REFERENCES >.

Объекты такого типа также имеют уникальный идентификатор (*Id_Publication*) и множество атрибутов, представляющих метаданные статьи. Особенностью объекта типа *Publication* является наличие двух наборов ссылок – *AUTHORS* и *REFERENCES*. Первый задает ссылки на авторов данной публикации, второй – на публикации, на которые ссылается публикация, описываемая данным объектом.

На основе структурированного представления статей могут быть установлены ассоциативные связи между публикациями и их авторами, а также между самими публикациями. Это значительно облегчает поиск нужных статей, хранящихся в информационной системе, и делает возможной навигацию по ним. Кроме того, такое представление статей создает хорошие предпосылки для автоматизации процесса занесения их описаний в контент информационной системы.

3. Автоматическое добавление формальных описаний научных статей в контент портала научных знаний

Задача автоматического добавления описаний статей в контент портала знаний является довольно сложной в связи с тем, что необходимо обеспечить не только корректность, но и согласованность и связанность вводимых данных с ранее введенными данными. Сделать это непросто, потому что статья в портале знаний представляется не одним, а целым набором связанных объектов (практически каждый элемент описания есть объект, будь то автор публикации, место его работы и др., представляются отдельными объектами), и каждый такой объект нужно проверить на корректность и существование в контенте портала.

Поясним сказанное на примере портала по компьютерной лингвистике [13, 17]. Каждая публикация в контенте такого портала представляется связанным набором объектов следующих классов: «Публикация», «Персона», «Организация» и др. В связи с этим при занесении формального описания публикации, представленной в БД цитирования, выполняются отображения ее элементов в объекты портала, при этом только часть полей формального описания статьи отображается в атрибуты объекта класса «Публикация», остальные – в атрибуты объектов других классов. Например: структура *Author* отображаются в объекты класса «Персона», структура *Publication* в объект класса «Публикация», а их поля *Place_work* (Место работы) и *Publisher* (Издательство) – в объекты класса «Организация». Кроме того, все полученные объекты связываются между собой различными отношениями. Например, объект класса «Публикация» связывается с объектами класса «Персона» и «Организация», соответственно, отношениями «Автор публикации» и «Издан в», а объект класса «Персона» с объектами класса «Организация» отношением «Работает в».

В соответствии с предложенным методом формальные описания статей последовательно извлекаются из базы данных цитирования и вносятся в контент портала в соответствии со схемой, описанной на Рис.1.

1. Осуществляется поиск в контенте портала знаний статьи с таким же названием (*Title*).

Поскольку в названии статьи могут содержаться различного рода ошибки (опечатки и пр. ошибки, совершенные автором-составителем статьи, а также ошибки, появившиеся из-за погрешностей в работе модуля генерации формального описания статей), считается, что название найдено, если оно совпадает с названием публикации, присутствующей в контенте портала, с некоторым предопределенным уровнем точности [10].

2. Если название не найдено, то «тело публикации», т.е. значения всех полей, кроме авторов и связей между публикациями, добавляется в контент портала; при этом запоминается идентификатор (*ID*) добавленной статьи и осуществляется переход к п.6.
3. Если название найдено, то запоминается идентификатор (*ID*) найденной статьи, после чего выполняются дополнительные проверки, описанные в п.4.
4. Сравниваются множества авторов и названий статей, на которые ссылается добавляемая статья и статья, запомненная под идентификатором *ID*.

a. Если множество авторов одной из статей – вносимой или имеющей идентификатор *ID* – полностью содержится в множестве авторов другой статьи, и списки библиографических ссылок у них содержат общие элементы, либо хотя бы один из них пуст, то считается, что объект найден, и осуществляется переход к п.5.

b. В противном случае считается, что статья не найдена и выполняются действия из п.2.

5. Выполняется объединение «тел статей» по правилу:

a. Если какое-то поле в статье с идентификатором *ID* отсутствует, то оно добавляется в контент.

6. Если соответствующие поля совпадают с некоторым предопределенным уровнем точности (может варьироваться для различных полей), то в зависимости от уровня привилегий автоматического добавления – либо статья с идентификатором *ID* остается неизмененной, либо записывается новый .

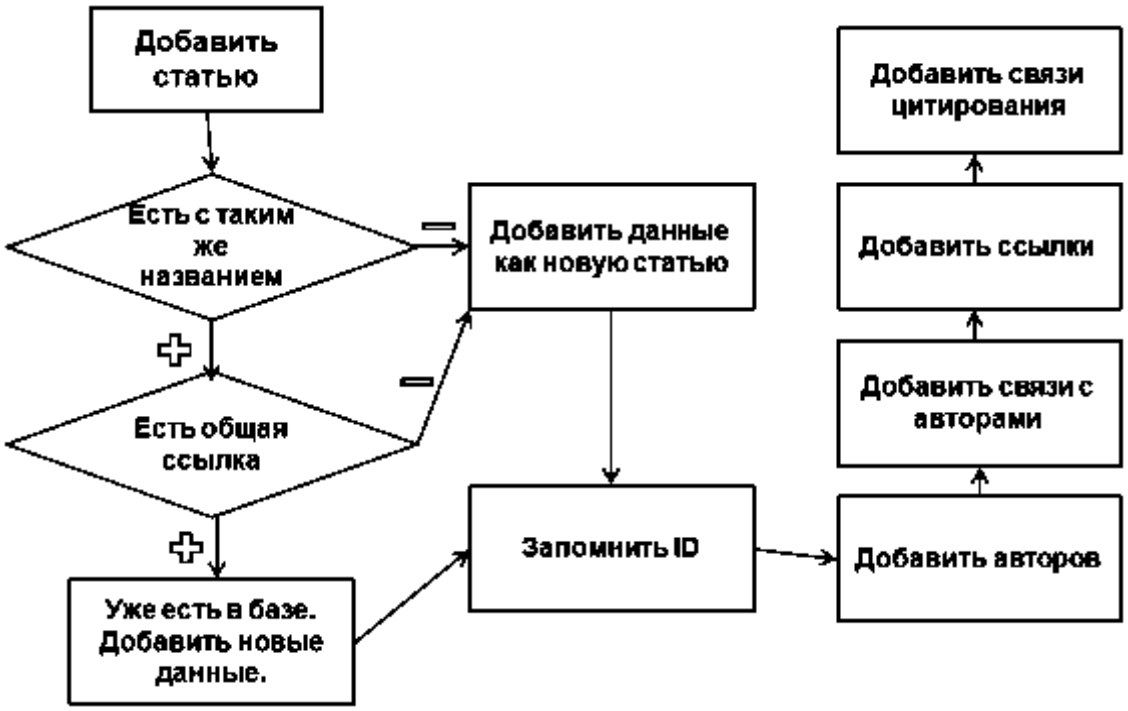

Рис.1. Схема добавления статьи в информационную систему

вариант, либо *ID* остается неизмененной, но при этом логируются данные о незаписанном поле в *ID*

7. Сравниваются авторы новой статьи и статьи с идентификатором *ID* по алгоритму, описанному ниже.

a. Для каждой пары авторов, признанных алгоритмом совпадающими, данные объединяются аналогично схеме из п. 5.

b. Все оставшиеся авторы просто добавляются в контент, а соответствующие им объекты связываются с объектом публикации.

8. Библиографические ссылки, содержащиеся в новой статье, добавляются в контент по описанному выше сценарию.

Алгоритм сравнения авторов статьи состоит в следующем (см. Рис.2):

1. Авторы сравниваются по именам (ФИО) и по месту работы/жительства.

a. Если ФИО авторов различны, либо одинаковы, но места работы/жительства у авторов различны, то считается, что авторы не совпадают.

b. В противном случае, авторы совпадают.

Предложенный метод автоматического добавления формальных описаний статей обладает рядом важных свойств:

1. Обеспечивается недублируемость данных,

2. Метод различает объекты с похожими характеристиками путем сравнения объектов не только по ключевым характеристикам, но и по второстепенным уникальным свойствам.

Таким образом, с помощью этого метода возможно эффективное импортирование данных из БД цитирования в информационные системы, в частности, порталы знаний.

4. Практические результаты

С целью исследования эффективности работы модуля генерации формальных описаний и метода пополнения контента портала знаний был проведен ряд тестов.

4.1 Тестирование и настройка модуля генерации формальных описаний

Процесс обработки корпуса текстов данным модулем включает два этапа:

1) Настройка набора шаблонов на корпус текстов;

2) Использование модуля для автоматического получения формальных описаний статей.

Настройка шаблонов выполнялась на небольшой выборке, включающей 50 докладов конференции “Диалог-2008”. По результатам экспериментов был доработан исходный набор шаблонов.

На втором этапе были использованы 200 статей конференции «Диалог» за 2005, 2006, 2007, 2009 и 2010 годы. Так как все статьи были взяты с сайта конференции, где они были представлены в формате html, а модуль генерации формальных описаний работает только с текстовым форматом (txt), то перед обработкой статей потребовалось удаление из их текстов тегов разметки.

В результате проведения экспериментов были получены следующие результаты:

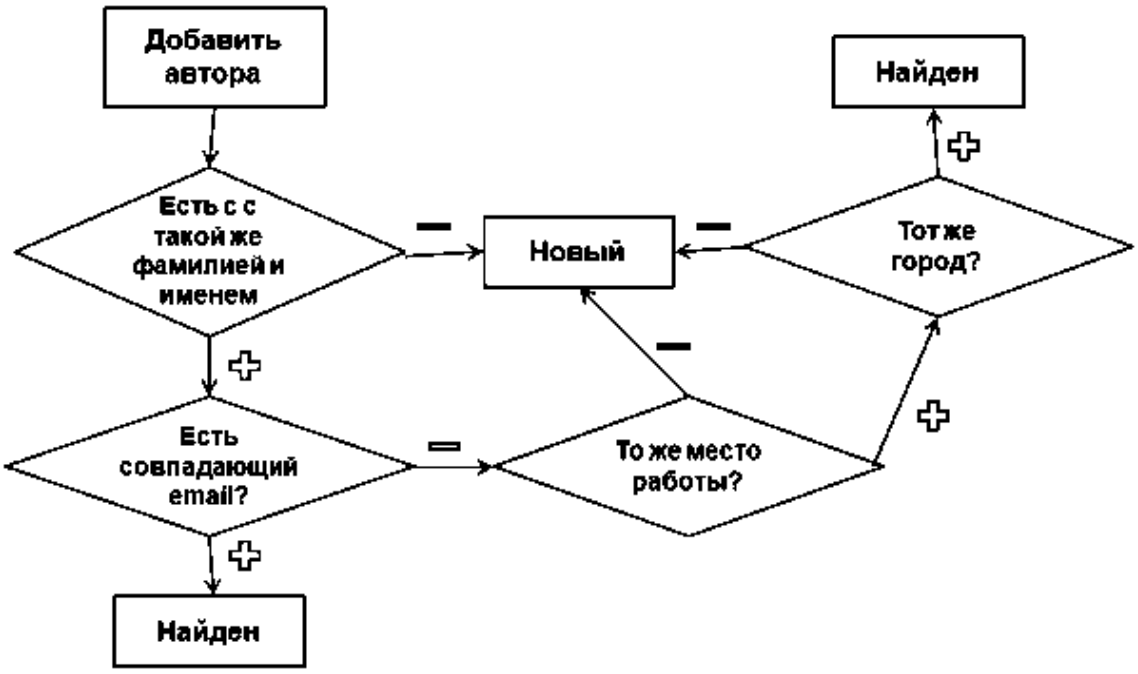

Рис.2. Схема алгоритма сравнения авторов статьи

1) Количество выделенных из текстов названий статей (включая библиографические ссылки) составило 2194; после исключения ошибочно определенных названий эта цифра уменьшилась примерно до 2000;

2) Количество извлеченных из статей различных авторов – 1378 (с учетом повторений – около 2000); заметим, что при обработке авторов было допущено существенно меньше ошибок – 29;

3) Процент правильно обработанных вводных частей статей, содержащих основную информацию о них, составил около 90%;

4) Процент правильно выделенных (подошедших) под шаблоны цитат составил около 85%; следует заметить, что это, отчасти, было вызвано ошибками при составлении цитат самими авторами статей, например, не проставлением знаков препинания, а также огрехами удаления разметки.

Часть статей была исключена из тестовой выборки, поскольку предварительное удаление разметки сторонними инструментами привело к повреждению внутренней структуры их текста, что не позволило выполнить их автоматическую обработку удовлетворительно.

4.2 Тестирование метода пополнения контента портала

Тестирование метода проводилось на БД портала по компьютерной лингвистике.

Поскольку метод пополнения контента портала использует сравнение строк и понятие “допустимой точности”, необходимо было определить, каким образом их можно было сравнивать, учитывая, что строки

“Опыт теории лингвистических моделей "Смысл Ы Текст". М.:”

и

“Опыт теории лингвистических моделей Смысл <-> Текст”

должны определяться как совпадающие.

В качестве метода вычисления расстояния между строками были предприняты попытки использовать расстояние Левенштейна [10], определяемое как минимальное количество операций вставки, удаления или замены одного символа на другой, необходимых для превращения одной строки в другую. Однако, к сожалению, для сравнения названий статей этот показатель не подошел, поскольку разница в написании даже одного слова, например, из-за орфографической ошибки, замены предлога знаком пунктуации или сокращения, приводили к резкому возрастанию расстояния между строками, что делало использование порогового значения расстояния неудовлетворительным, каким бы оно не было.

По этой причине было решено использовать несколько модифицированное расстояние, выраженное в количестве процентов, которое составляет расстояние Левенштейна от минимальной длины двух сравниваемых строк:

$$\rho(s_1,s_2) = \begin{cases} 100, & \text{если } |s_1| = 0 \text{ или } |s_2| = 0 \\ 100 * \dfrac{leven(s_1,s_2)}{\min\{|s_1|,|s_2|\}}, & \text{иначе} \end{cases}$$

где $leven(s_1,s_2)$ – функция расстояния Левенштейна, $|s_i|$ – длина строки.

Сравнение строк производилось следующим образом:

1) Выполняется приведение строк к нижнему регистру;

2) Из строк удаляются все небуквенные символы;

3) Вычисляется расстояние между строками;

4) В случае, если расстояние больше заданного порогового расстояния, то строки считаются различными, иначе – совпадающими.

Для определения величины порогового значения были проведены эксперименты по импортированию результатов работы модуля генерации формальных описаний в локальную БД портала по компьютерной лингвистике с заданием различных пороговых значений

Из примерно 2000 выделенных названий статей, только 71 уже присутствовало в БД.

В результате эксперимента были получены данные, представленные в таблице 1.

Таблица 1. Зависимость числа отождествленных статей от порогового значения.

Порог	Число отожд.	Число ошиб.
90	1468	1397
50	114	43
30	75	4
25	72	1
20	71	0
10	64	7
5	63	8

Таким образом, исходя из приведенных в Таблице 1 данных эксперимента, пороговую величину целесообразно выбирать в интервале от 10 до 20, что соответствует 1-2 опечаткам на каждые 10 букв. Большее пороговое значение будет приводить к большему числу ошибочно отождествленных статей.

Среди извлеченных из статей 1378 авторов, только 105 оказались уже представленными в БД. Это указывает на практическую целесообразность автоматического пополнения контента портала и информационных систем в целом.

4.3 Сравнение с существующими системами

Следует заметить, что выполненная работа была нацелена не на создание универсальных индексов цитирования, а на получение данных о публикациях для порталов знаний и других информационных систем, интегрирующих знания и информационные ресурсы определенной тематики. В настоящий момент, имеющиеся в портале данные не используются для оценки эффективности научной деятельности, но их наличие создает предпосылки к построению сетей цитирования и соавторства, которые позволят выявлять наиболее значимые (цитируемые) публикации, скрытые научные сообщества и т.п.

К сожалению, такие реферативно-библиографические базы данных научного цитирования, как отечественный РИНЦ, а также зарубежные Web of Science и SCOPUS не предоставляют в открытой форме алгоритмов сравнения вновь добавляемых статей со статьями, уже представленными в их базах данных, поскольку частично эта работа выполняется вручную (в случае с SCOPUS и РИНЦ).

Web of Science описывает свой метод сравнения названий статей [6] на примере сравнения названий двух журналов:

International Journal of Manufacturing and Production Systems и

International Journal of Manufacturing and Production Services

При этом сообщается, что различать подобные названия статей достаточно просто, приведя их к сокращенной (аббревиатурной) форме, используя БД сокращений. Например, для первого названия такой формой будет: *INT J MANUF PROD SYS*.

Так как подход, использованный в Web of Science к различению названий, основывается на англоязычной базе данных сокращений, которой она располагает, это не позволяет нам сравнивать подход Web of Science с нашим подходом. В то же время, модификация нашего подхода с использованием идей, предложенных в подходе Web of Science, является одним из возможных направлений его развития.

5. Заключение

Рассмотрен подход к автоматическому наполнению информационных систем библиографическими сведениями о научных статьях. В рамках этого подхода разработана формальная структура представления статьи, разработан и реализован метод автоматического построения формальных описаний научных статей. Его достоинством является возможность декларативной настройки на коллекцию документов, подлежащих обработке. Средствами СУБД MySQL реализована база данных цитирования, а также пользовательский интерфейс, позволяющий просматривать и редактировать полученные формальные описания статей. Разработан и реализован метод автоматического добавления формальных описаний научных статей в информационные системы, интегрирующие знания и информационные ресурсы по определенной области знаний.

Для обеспечения возможности импорта сторонних публикаций в БД цитирования, а также использования формальных описаний статей, полученных предложенным методом, в других системах разработано представление этих описаний в формате XML и реализованы модули экспорта в такой формат и импорта из него.

Описанные в докладе методы были использованы для пополнения контента портала знаний, обеспечивающего содержательный доступ к систематизированным знаниям и информационным ресурсам по компьютерной лингвистике [13]. В частности, предложенные средства использовались для внесения в контент указанного портала

сведений о статьях, представленных на конференции «Диалог» в 2005-2010 гг. [15].

В дальнейшем предполагается расширить функциональность рассмотренного модуля генерации формальных описаний статей возможностью обработки документов, представленных не только в формате txt, но и других форматах (pdf, html и др.).

Литература

[1] BiblioScape 8, 2011. www.biblioscape.com/

[2] I, Librarian, 2011. www.bioinformatics.org/librarian/

[3] Lawrence, S., Giles, C.L. & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. In IEEE Computer. 32(6).

[4] Mendeley Desktop, 2011. www.mendeley.com/

[5] Perl Compatible Regular Expressions (PCRE). http://pcre.org

[6] Robertson J. Cited Title Unification. Thomson Reuters. http://thomsonreuters.com/products_services/science/free/essays/cited_title_unification/

[7] Scientific Literature Digital Library and Search Engine, 2011. http://citeseerx.ist.psu.edu

[8] The Internet Archive, non-profit Internet library, 2011. http://www.archive.org

[9] The Rhine Recearch Center, 2011. - http://www.rhine.org

[10] Wikipedia. Статья о Расстоянии Левенштейна. http://ru.wikipedia.org/wiki/Расстояние_Левенштейна

[11] Археологический портал знаний, 2011. - http://www.sati.archaeology.nsc.ru/classarch2/

[12] База данных цитирования по нанотехнологиям, 2011. http://thomson.collexis.com/nano/

[13] Боровикова О.И., Загорулько Ю.А., Загорулько Г.Б., Кононенко И.С., Соколова Е.Г. Разработка портала знаний по компьютерной лингвистике // Труды 11-ой национальной конференции по искусственному интеллекту с международным участием КИИ-2008. М.: ЛЕНАНД, 2008. Т.3. С.380-388.

[14] Дяченко О.О., Загорулько Ю.А.. Генерация формальных описаний научных статей для информационных систем // Труды 12-й национальной конференции по искусственному интеллекту с международным участием – КИИ-2010. – Москва: Физматлит, 2010.. -Т.1. -С.225-233.

[15] Материалы конференции «Диалог». - http://dialog-21.ru

[16] Научная электронная библиотека, российский информационный портал eLIBRARY.RU, 2011. http://elibrary.ru

[17] Портал по компьютерной лингвистике, 2011. - http://uniserv.iis.nsk.su/cl/

[18] Сонг Д. Новая модель базы данных цитирования на языке XML с использованием XQuery в качестве поискового языка // Сборник трудов десятой юбилейной международной конференции "Крым 2003. Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества". Украина, Крым, 2003.

Automatic Filling of Information Systems with Bibliographic Records of Scientific Publications

© Yu.A. Zagorulko, O.O. Dyachenko

The paper describes an approach to automation of the filling of information systems with bibliographic descriptions of scientific publications. In the framework of this approach, a method of generation of formal descriptions of scientific papers and a method of automatic addition of these descriptions in the content of a scientific knowledge portal were developed.

An advantage of the suggested method of generation of formal descriptions of scientific papers is a possibility of its declarative adjustment to collection of documents to be processed.

♣ Работа выполнена при финансовой поддержке РФФИ (проект № 09-07-00400).