Проект электронной библиотеки методик и результатов исследований текстовых коллекций для системы «Источник»♣

© Н.В. Каргинова, И.В. Кравцов, Н.Д. Москин, А.Г. Варфоломеев

Петрозаводский государственный университет
avarf@psu.karelia.ru

Аннотация

Данная статья посвящена применению формата записи произвольных правил RuleML для сохранения методик и результатов исследований в цифровых библиотеках исторических документов. Предлагается формализация процесса исследования текстовых коллекций: сначала в текстах выделяются логические или семантические фрагменты, затем из них образуются структуры текстов в виде векторов или графов, эти структуры разбиваются на типы, и делаются выводы о зависимости типов структур от других характеристик текстов. Эти выводы записываются в виде правил, пригодных для машинной обработки, и сохраняются в библиотеке.

Рассмотренный подход используется авторами на практике в разработке информационной системы «Источник», предназначенной для организации работы сетевых сообществ исследователей текстовых исторических источников.

1 Введение

С развитием компьютерных технологий все чаще исследования по истории и лингвистике опираются на большие коллекции текстовых документов. Такие коллекции, представленные в Интернете, составляют основу для формирования сетевых сообществ исследователей, разделяющих между собой форматы представления данных (TEI [20], MEP [15], CEI [13]), а иногда и сами тексты для совместного изучения и редактирования («Манускрипт» [8], Monasterium [16], TextGrid [21]). Естественным выглядит следующий шаг – предоставить в распоряжение сообщества не только данные и инструменты, но и описания проведенных ранее исследований. Казалось бы, для этого достаточно научных публикаций в их традиционной форме. Однако традиционные публикации не являются машиночитаемыми, с их помощью невозможно строить базу знаний сообщества. Хотелось бы дополнить такие публикации некими формализованными описаниями проведенных исследований, на основе которых компьютер не только мог бы производить поиск подходящих методик и результатов, но и выдвигать собственные гипотезы.

Технологии для достижения поставленной цели существуют. В рамках современного направления Semantic Web разрабатываются стандарты представления бизнес-правил [19], в виде которых могут быть записаны и научные выводы, гипотезы, формулы или алгоритмы. Примерами таких стандартов являются язык PMML [18], служащий для записи регрессионных и других предиктивных моделей анализа данных, и форматы группы MKM [22] для обмена математическими результатами. Однако в историко-филологических исследованиях текстовых источников формализация методик и результатов в виде бизнес-правил также может быть осуществлена. К сожалению, пока нельзя привести ни одного примера электронных коллекций текстов, предоставляющих подобную информацию. На наш взгляд, причиной является то, что пока подавляющее большинство текстовых коллекций ориентированы на предоставление данных для локальной работы индивидуальных исследователей, и не предусматривают обмен формализованными методиками и результатами в рамках сетевых сообществ.

Задача разработки формата представления методик и результатов исследований текстовых коллекций была поставлена в [11] в рамках проекта информационной системы «Источник» [7]. Разработка этой системы, ориентированной на поддержку сетевого научного сообщества исследователей текстов, отражена в публикациях [4, 9, 11]. В качестве особенностей данной системы следует указать ее многофункциональность, а также последовательное применение технологий XML на разных этапах работы с текстом.

В данной статье мы предлагаем проект библиотеки методик и результатов исследований текстовых коллекций для системы «Источник» с возможностью публикации результатов в машиночитаемой форме с помощью языка RuleML [23].

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

2 Информационная система «Источник»

2.1 История развития системы

Первоначально информационная система «Источник» возникла как сетевой сервис для подготовки печатных изданий средневековых исторических документов, прежде всего комплекса «Moscowitica-Ruthenica» [5]. Для реализации была выбрана технология XML. Однако текст, один раз размеченный, мог быть повторно использован, причем не только для публикации, но и для анализа его структуры и содержания. Так возникла идея разрабатывать многофункциональный сетевой сервис, включающий в себя различные инструменты анализа текста [11]. Одним их важных способов анализа является визуализация текстовой информации, для этого в системе «Источник» было решено широко применять SVG-графику [4].

Другим важным направлением развития системы «Источник» было превращение ее в систему поддержки работы сетевого сообщества исследователей произвольных текстов с возможностью накопления размеченных текстов и отчетов о проведенных исследованиях в виде своеобразной «базы знаний» сообщества [9,11]. Но описания методов и результатов исследований, хоть и были структурированы, не имели вид правил «если-то» и не могли быть использованы для автоматического порождения гипотез и получения новых знаний. Настоящая статья как раз направлена на решение проблемы записи методик и результатов в виде правил вывода.

2.2 Анализ текстовых коллекций

Работая в информационной системе «Источник», исследователь размечает тексты коллекции с помощью одной или нескольких XML-разметок. Затем, изучая размеченные тексты, исследователь может делать выводы об обнаруженных закономерностях.

Например, в ходе формулярного анализа средневековых грамот [6] исследователь вручную выделяет в текстах устойчивые смысловые фрагменты, такие как «приветствие», «описание ситуации», «ссылка на предыдущий документ» и сравнивает последовательности таких фрагментов для разных текстов, делая выводы о наличии норм и классифицируя документы по их «формулярам». В контент-анализе массовых источников [1] исследователь тоже выделяет фрагменты текста – смысловые единицы на основании присутствия в них словосочетаний-«индикаторов», указывающих на некие «категории анализа», затем подсчитывает количество или объем смысловых единиц, классифицируя документы по степени отражения в них тех или иных категорий. При структурном анализе фольклорных песен в текстах выделяются устойчивые фрагменты (мотивы), упоминания объектов (например, основных персонажей, животных, предметов обихода и т.д.), и глаголы (или отглагольные формы), указывающие на отношения между объектами [3].

Для организации работы исследователей с методиками и результатами информационная система, на наш взгляд, должна обладать следующими функциями:

- предоставлять пользователю интерфейс для разметки текстов, записи правил разметки и правил вывода (методики исследования), а также самих выводов (результатов исследования);
- предоставлять сетевому сообществу возможность накапливания не только коллекций размеченных текстов, но также библиотеки правил и выводов как основы базы знаний сообщества;
- обеспечивать возможность поиска, просмотра и сравнения методик и результатов других исследователей;
- предлагать пользователю ознакомиться с гипотезами, выведенными системой автоматически на основе имеющихся данных и правил;
- предоставлять возможность публикации методик и результатов исследований в машиночитаемой форме для того, чтобы они также становились доступными для произвольных семантических сервисов в сети.

Возможны различные варианты автоматической генерации гипотез в зависимости от полученных результатов исследователя и уже имеющихся результатов других исследователей. Например, если исследователь получает какой-либо вывод на основании определенного правила, то система может попробовать применить это правило к другим похожим коллекциям и предложить сравнить результаты. Если есть несколько исследований с похожими целями или методиками, проведенных на одной коллекции, система может самостоятельно сравнить полученные результаты и сделать выводы об их сходстве или различии. Если проводится повторное исследование коллекции или текста, система может сообщить все полученные ранее результаты, чтобы исследователь мог их подтвердить или опровергнуть.

3 Формализация процесса исследования

Для успешного решения задачи формализованной записи методик и результатов исследований необходимо предложить некоторую обобщенную схему, формализующую сам процесс исследования текстовых коллекций. На наш взгляд, процесс исследования можно представить в виде следующей последовательности этапов.

Исследователь разбивает текст на логические или семантические фрагменты. При этом он может выбрать разметку из уже имеющихся в системе вариантов или предложить свою. В последнем случае исследователь должен описать принципы, положенные в основу разметки, для того, чтобы ее в дальнейшем могли использовать другие исследователи.

Для представления разбиения текста в системе «Источник» используется язык XML. Для хранения синтаксических правил разметки используются форматы DTD или XSD. На данном этапе также могут фиксироваться некоторые исследовательские

правила, с помощью которых производится разбиение, чтобы в дальнейшем можно было частично автоматизировать этот процесс. Например, если в тексте уже присутствует какая-то разметка, в частности, на абзацы, то наличие в абзаце слова-индикатора может служить основанием для преобразования данного абзаца в логический фрагмент определенного типа. Такое преобразование может быть записано с помощью языка XSLT.

После того, как текст разбит на логические фрагменты (блоки), его структуру необходимо представить в виде некоторого объекта, с которым будет в дальнейшем работать система. В самом простом варианте это может быть список (вектор), в котором каждый элемент будет являться идентификатором соответствующего блока в разметке. Можно представить структуру теста в виде следующей таблицы, где номер – это порядковый номер блока в структуре текста, тип блока – значение из заданного набора типов блоков выбранной разметки, характеристики блока могут использоваться для хранения дополнительной информации о том фрагменте текста, который относится к данному блоку. В качестве характеристик могут выступать объем фрагмента текста, вычисленный в словах или символах, или процентное соотношение объема блока к общему объему текста.

№	Тип блока	Характеристика блока
1	A	20
2	B	25
…	…	…

Приведенная таблица адекватно передает линейную структуру текста, используемую, например, в формулярном анализе или в контент-анализе. В случае использования более сложной иерархической или графовой структуры, как, например, в анализе фольклорных песен, таблица представления результатов тоже должна быть усложнена.

После представления структуры текста в выбранном стандартном виде можно проводить анализ при помощи правил вывода.

Правила в системе делятся на заданные изначально и формируемые в ходе работы исследователей. Заданные изначально правила обеспечивают анализ исследований, получение новых выводов и применение уже существующих закономерностей к текущему исследованию. Также изначально может быть задан набор фактов, устанавливающих, например, что какая-то структура является общей для какого-либо набора текстов или что две структуры похожи друг на друга:

структура 1 – общая для структур (структура 2, …, структура n)

структура 1 похожа на структуру 2 на n %

Некоторые факты могут быть представлены в виде функций (алгоритмов). Например, функция *compare_structure* с двумя аргументами – структурами возвращает их сходство в процентах.

compare_structure(структура 1, структура 2) = n %

К заданным изначально правилам также относятся шаблоны, с помощью которых могут записываться и новые правила, полученные в ходе исследования. Ниже приведены примеры шаблонов.

Шаблон 1

ЕСЛИ *текст 1 — структура 1* *текст 2 — структура 2* *….* *текст n — структура n* *(текст 1, …, текст n) – тип 1* *структура k – общая для структур (структура 1, …, структура n)*
ТО *структура k – общая структура для типа 1*

В соответствии с данным шаблоном, если для каждого текста из набора была получена определенная структура, и известно, что все тексты относятся к одному типу, то можно сделать предположение об общей структуре для текстов данного типа.

Шаблон 2

ЕСЛИ *текст 1 — структура 1,* *текст 2 — структура 2,* *….* *текст n — структура n,* *(текст 2, …, текст n) имеют тип 1,* *(структура 1, структура 2…, структура n) похожи на m%,* *m больше, чем пороговое значение*
ТО *текст 1 имеет тип 1*

В соответствии с этим шаблоном, если имеется несколько текстов достаточно близкой структуры, и про часть из них известно, что они относятся к определенному типу, можно предположить, что оставшиеся тексты также относятся к этому типу. Следует отметить, что шаблоны могут отражать вовсе не только дедуктивные правила вывода, но, как в указанных примерах, «правдоподобные» рассуждения, основанные на индукции и аналогии [2].

Правила, формирующиеся в ходе работы исследователей, представляют собой формализованную запись выводов исследователей. Они могут формироваться как конкретизация имеющихся в системе шаблонов. В таком случае на место свойств, типов и структур подставляются конкретные значения. Эти правила могут быть сохранены в библиотеке системы и использованы в дальнейших исследованиях.

Для каждого правила должны задаваться область видимости и коэффициент доверия. По умолчанию исследователь видит только свои правила, но при желании может ознакомиться с правилами других исследователей. Если в ходе исследования какое-то

уже сохраненное в системе правило подтверждается или опровергается, это может вызвать изменение его коэффициента доверия. Таким образом, в дальнейшем исследователи могут выбирать наиболее подтвержденные правила.

4 Хранение правил

Рассмотренные выше правила могут быть представлены в системе в каком-либо внутреннем формате для ускорения процессов поиска, сравнения и вывода. Но при публикации результатов необходимо, чтобы правила были понятны не только данной системе, но и другим системам, ориентированным на работу с текстами. Для этого необходимо представить тексты, разметки и правила с помощью стандартных форматов, позволяющих сохранить семантику. Также, так как система «Источник» является web-ориентированной, чтобы обеспечить работу исследователей из разных мест, необходимо использовать стандарты, позволяющие представить информацию в Интернете.

Для записи правил в библиотеке методик и результатов системы «Источник» предлагается использовать активно разрабатываемый в настоящее время стандарт RuleML.

RuleML (Rule Markup Language) [23] представляет собой язык разметки для описания правил. С помощью данного языка можно публиковать и обмениваться правилами, созданными в рамках разных систем и задач. Он включает в себя такие языки, как PMML (для предиктивных моделей), MathML (для записи функциональных зависимостей), XSLT (для правил преобразований из XML в XML). Но ядром RuleML является XML-вариант языка Datalog. Он позволяет записывать факты и правила в форме продукций.

Пример описания шаблона 2 на языке RuleML.

```
<Implies>
    <head>
        <Atom>
            <Var>text 2</Var>
            <Rel>type of text</Rel>
            <Var>type 1</Var>
        </Atom>
    </head>
    <body>
        <Atom>
            <Var>text 1</Var>
            <Rel>type of text</Rel>
            <Var>type 1</Var>
        </Atom>
        <Atom>
            <Var>text 1</Var>
            <Rel>structure of text</Rel>
            <Var>structure 1</Var>
        </Atom>
        <Atom>
            <Var>text 2</Var>
            <Rel>structure of text</Rel>
            <Var>structure 2</Var>
        </Atom>
        <Atom>
            <Var>structure 1</Var>
            <Rel>is -like</Rel>
            <Var>structure 2</Var>
            <Ind>m %</Ind>
        </Atom>
    </body>
</Implies>
```

5 Примеры записи результатов анализа фольклорных текстов

Теперь рассмотрим, как можно применить правила, записанные на языке RuleML, для описания закономерностей, полученных в результате анализа фольклорных песен на основе их теоретико-графовых моделей [3]. Одним из способов анализа таких моделей является сравнение их агрегирующих графов с небольшим числом вершин и ребер. Эти графы позволяют выделить структуру основных отношений в тексте, отбросив несущественные объекты и связи. Поскольку в общем виде задача агрегации трудно решаема, данные графы были построены при помощи метода, предложенного в работе [10]. В нем накладывается следующее ограничение: разбиение объектов осуществляется на непересекающиеся группы, объединение которых дает исходное множество объектов. Число групп должно быть установлено заранее – в нашем случае оно, как правило, определялось по числу мотивов песни.

На рис. 1 приведен пример агрегирующего графа для бесёдной песни «Широкая борода» (текст песни и ее теоретико-графовая модель приводятся в [3]). Здесь группа объектов №2 – это *сад, плечи, головушка, отец-мать, род-племя, дом, блюдечко, серебряный поднос, высок терем,* т.е. объекты, связанные с девушкой.

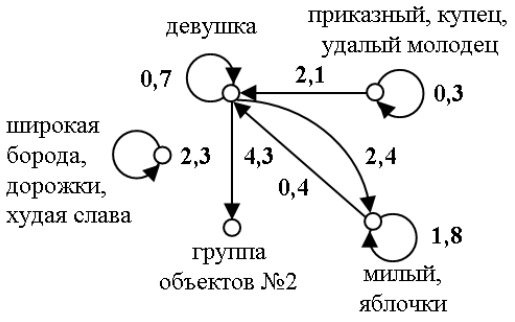


Рис. 1

На основе коллекции бесёдных песен была составлена выборка из 50 текстов. Если не учитывать изолированные вершины и петли, у большинства песен агрегирующий граф имеет вид дерева (см. рис. 2). При этом вершина с максимальной степенью (с номером 1), как правило, соответствует главному действующему лицу – парню или девушке.

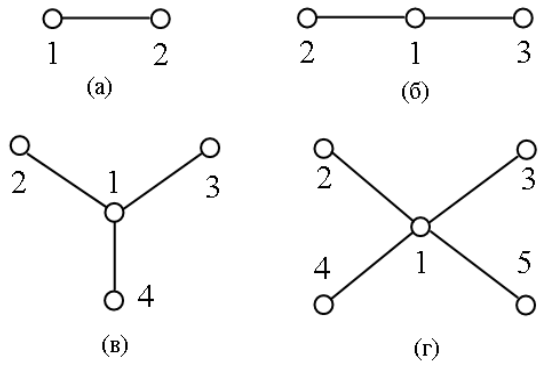


Рис. 2

Как показали эксперименты, данные структуры можно связать с основными характеристиками песен, например: *«если песня имеет тип «утушная песня», то агрегирующий граф с большой вероятностью имеет структуру 1»*. На языке RuleML данная закономерность может быть записана следующим образом:

```
<Implies>
    <head>
        <Atom>
            <Rel>имеет структуру</Rel>
            <Var>агрегирующий граф</Var>
            <Var>1</Var>
            <Var>с большой вероятностью</Var>
        </Atom>
    </head>
    <body>
        <Atom>
            <Rel>имеет тип</Rel>
            <Var>песня</Var>
            <Var>утушная песня</Var>
        </Atom>
    </body>
</Implies>
```

Также можно выделить закономерности с обратным порядком, например: *«если агрегирующий граф имеет структуру 1, то эта песня с большой вероятностью имеет частый темп исполнения»*. На языке RuleML данная закономерность может быть записана следующим образом:

```
<Implies>
    <head>
        <Atom>
            <Rel>имеет темп</Rel>
            <Var>песня</Var>
            <Var>частый</Var>
            <Var>с большой вероятностью</Var>
        </Atom>
    </head>
    <body>
        <Atom>
            <Rel>имеет структуру</Rel>
            <Var>агрегирующий граф</Var>
            <Var>1</Var>
        </Atom>
    </body>
</Implies>
```

Описав таким образом полученные закономерности, исследователь получает возможности для автоматизированного анализа правил и получения новых знаний

6 Реализация логического вывода

Для успешной работы системы необходимо обеспечить логический вывод на основе имеющихся фактов и правил. Для этого нужна машина логического вывода.

Возможны два варианта реализации. Первый заключается в использовании в системе какого-либо внутреннего формата для хранения фактов и правил. В таком случае в качестве такого формата можно, например, взять форматы представления знаний таких сред как CLIPS или SWI-Prolog. Эти среды предназначены для систем, основанных на знаниях, и являются свободно-распространяемыми продуктами. Программы, написанные в обеих этих средах, могут быть интегрированы с программами на других языках, в том числе с Java, что позволит создать web-интерфейс для системы. Сама программа, осуществляющая логический вывод может выполняться на сервере, а передача входных и выходных данных может осуществляться через web-интерфейс.

Стандарты XML, RDF и RuleML будут играть роль форматов для представления результатов и методики исследования другим сообществам исследователей и другим системам. Для преобразования из внутренних форматов во внешние необходимо будет использовать уже имеющиеся программы (например, SWI-Prolog содержит библиотеку для разбора данных в формате RDF) или реализовать собственную.

Второй вариант реализации заключается в использовании RuleML и в качестве внутреннего формата представления знаний. В таком случае, необходимо использовать машины логического вывода, работающие с форматом RuleML. Примерами таких машин является Bossam [12], которая позволяет строить приложения в рамках концепции Semantic Web, а также OO jDREW [17] – объектно-ориентированная дедуктивная машина вывода для RuleML. OO jDREW представляет собой библиотеку, написанную на языке Java. Таким образом можно написать набор приложений на Java, которые будут реализовывать функции системы, и осуществлять логический вывод, используя функции библиотеки OO jDREW.

Также существует система DR-DEVICE [14], позволяющая осуществлять рассуждения в условиях неполной и противоречивой информации. Это актуально в условиях Semantic Web, например, конфликты могут возникнуть при объединении онтологий. Система поддерживает использование фактов, правил и приоритетов правил.

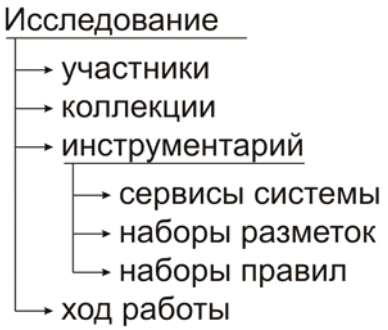
DR-DEVICE позволяет загружать RDF документы из Интернета и использовать их в качестве входной информации – фактов – для программы. Прави-

ла могут быть записаны либо на языке, подобном языку CLIPS, либо с использованием синтаксиса RuleML. Полученные результаты можно выгружать в Интернет в виде RDF документов.

7 Проект библиотеки

В рамках системы «Источник» создается библиотека методик и результатов исследований коллекций текстов. Цель такой библиотеки – организация хранилища аналитических публикаций, которые в своем составе содержат не только исходные источниковые материалы и описание полученного результата в виде научной статьи, но и сам инструментарий исследования с промежуточными выкладками.

Рассмотрим структуру информации в библиотеке об одном «типовом» исследовании.



Каждое исследование является фактом связывания набора составляющих его объектов и имеет собственный уникальный идентификатор и также является отдельным объектом системы.

В рамках одного исследования определяется состав участников, работающих над ним. Как правило, это руководитель исследования, который определяет состав остальных объектов исследования, и исполнители, которые получают возможность работать с определенными в исследовании объектами.

Исследование проводится на основе фиксированной коллекции документальных источников. Как правило, это полная обособленная коллекция документов.

Исследование заключается в применении к источниковому материалу специализированного инструментария: программных сервисов и алгоритмических модулей, фиксированных наборов структурных и семантических разметок, определенных на данных разметках наборов правил получения результатов.

Инструментарий применяется последовательно, согласно заранее описанному ходу работы. Например, сначала производится физическая разметка текстов (выделение словоформ), затем – структурно-семантическая разметка, после этого производится обработка полученной структуры с помощью заранее определенных правил (функций, преобразований) с целью получения новой разметки или новых правил, содержащих выявленные закономерности информации.

Результат всего исследования либо отдельной его стадии записывается в виде XML-документа, состоящего из трех частей: результата, правил, посылок. В качестве посылок (аргументов правил) выступают первично размеченные тексты, результатом является новая разметка или новые правила (закономерности).

В случае отсутствия заранее разработанного инструментария для анализа проводится «экспериментальное» исследование, результатом которого являются новые, выработанные исследователем, схемы разметки текстов или правила получения результатов.

Накопленные результаты могут быть рассмотрены как исходный материал для выдвижения и проверки новых гипотез и проведения новых исследований.

Кроме рассмотренного разреза «исследования», навигацию по библиотеке можно будет осуществлять на базе остальных составляющих ее объектов. Например, при просмотре с точки зрения участников можно будет видеть, в каких исследованиях они участвовали. Для коллекций текстов можно будет просмотреть примененные к ним разметки и правила, и, наоборот, для той или иной разметки получить примеры её использования.

Литература

[1] Бородкин Л.И. Контент-анализ и проблемы изучения исторических источников // Математика в изучении средневековых повествовательных источников. – М., 1986. – С. 8–30.

[2] Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В. Достоверный и правдоподобный вывод в интеллектуальных системах. – М., 2004.

[3] Варфоломеев А.Г., Кравцов И.В., Москин Н.Д. Проект специализированного Интернет-ресурса для представления и анализа фольклорных песен // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Пятой Всероссийской научной конференции RCDL'2003 (Санкт-Петербург, 29–31 октября 2003 г.). – СПб., 2003. – С. 339–343.

[4] Варфоломеев А.Г., Кравцов И.В., Филатов В.О. SVG-визуализация в цифровых библиотеках рукописных документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Девятой Всероссийской научной конференции RCDL'2007 (Переславль-Залесский, Россия, 14–18 октября 2007 г.). – Переславль-Залесский, 2007. С. 230–235.

[5] Иванов А. С., Варфоломеев А.Г. Использование технологии XML для введения в научный оборот комплекса документов «Moscowitica-Ruthenica» // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Шестой Всероссийской научной конференции RCDL'2004 (Пущино, 29

сентября – 1 октября 2004 г.). – Пущино, 2004. – С. 285–289.
[6] Иванов А. С., Варфоломеев А.Г. Технология XML как инструмент компьютерного источниковедения (на примере формулярного анализа документов приказного делопроизводства) // Круг идей: Алгоритмы и технологии исторической информатики : Труды IX конференции Ассоциации «История и компьютер» / ред. Л.И. Бородкин, В.Н. Владимиров. – М. ; Барнаул, 2005. – С. 241–281.
[7] Источник. Информационная система для работы сообществ исследователей текстов. (http://istochnik.karelia.ru)
[8] Манускрипт. Древние славянские памятники. (http://manuscripts.ru)
[9] Кравцов И.В., Варфоломеев А.Г. Принципы организации информационного пространства сетевого сообщества исследователей рукописных текстов // Информационное общество. Интеллектуальная обработка информации. Информационные технологии : Материалы 7-й международной конференции НТИ-2007 (Москва, 24–26 октября 2007 г.). – М., 2007. – С. 383–386.
[10] Куперштох В. Л., Трофимов В. А. Алгоритм анализа структуры матрицы связи // Автоматика и телемеханика. – 1975, №11. – С. 170–180.
[11] Филатов В.О., Кравцов И.В., Варфоломеев А.Г. Информационная система для работы с полнотекстовыми базами данных исторических документов на основе технологии XML // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Восьмой Всероссийской научной конференции (RCDL'2006), Суздаль, 17–19 октября 2006 г. – Ярославль, 2006. – С. 337–344.
[12] Bossam Rule/OWL Reasoner (http://bossam.wordpress.com)
[13] Charters Encoding Initiative (CEI) (http://www.cei.lmu.de)
[14] DR-DEVICE. A Defeasible Logic Reasoner for the Semantic Web. (http://lpis.csd.auth.gr/systems/dr-device.html)
[15] Model Editions Partnership (MEP) (http://adh.sc.edu)
[16] Monasterium Project (http://monasterium.net)
[17] Object Oriented jDREW (http://www.jdrew.org/oojdrew)
[18] Predictive Model Markup Language (PMML) (http://www.dmg.org/pmml-v3-2.html)
[19] Rule Interchange Format Working Group (http://www.w3.org/2005/rules/wiki/RIF_Working_Group)
[20] Text Encoding Initiative (http://www.tei-c.org)
[21] TextGrid Project (http://www.textgrid.de)
[22] The MKM Interest Group (Mathematical Knowledge Management) (http://www.mkm-ig.org)
[23] The Rule Markup Initiative (http://www.ruleml.org)

Information system “Istochnik” e-library project for storing sets of rules and results in the area of texts collections analyzing

N. Karginova, I. Kravtsov, N. Moskin, A. Varfolomeyev

This paper represents one new way for special e-library organizing. The main idea is for simultaneously storing of linguistic text sources, analyzing methods, research services and final results. Also library is designed as instrument for remote collaborative work.

This project is the part of network information system “Istochnik” realization. There are many different XML-technologies using in project for text structure and semantic markup. Results of research are representing by the rules. Rules can be use for logical deduction and new formalized knowledge forming by machine. For rules we use RuleML technology.

♣ Статья написана в рамках проекта, поддержанного грантом Российского гуманитарного научного фонда (проект № 08-01-12136в).