Применение фасетной классификации для систематизации газетных публикаций в электронной коллекции казанской периодической печати 19-го – начала 20-го веков

А.Г. Абросимов, В.Ю. Кузьмина, И.И. Салосина
Научная библиотека им. Н.И. Лобачевского Казанского государственного университета

Аннотация

Описываются лингвистические средства электронной библиотеки Казанского государственного университета, применяемые при формировании электронной коллекции казанской периодической печати 19-го – начала 20-го веков.

Введение

В Научной библиотеке Казанского государственного университета (НБ КГУ) при поддержке Российского гуманитарного научного фонда (проект № 04-01-12032в) формируется электронная коллекция казанской периодической печати 19-го – начала 20-го веков. К настоящему моменту завершено создание первого раздела коллекции – оцифрован полный комплект газеты «Казанские известия» (818 номеров), проведено редактирование изображений страниц газет, определены формат и размеры изображений. Низкое качество бумаги и печати периодических изданий приводит к тому, что затраты на распознавание текста газеты сопоставимы с затратами на его ручной набор. Поэтому наиболее реальным решением является организация электронной коллекции в виде структурированной совокупности электронных графических образов страниц периодических изданий и метаданных (подробнее см. [1], [2]). На основе разработанного профиля метаданных строятся основные технологические процессы: навигация по электронной коллекции, организация хранения, разнообразное представление электронных документов, поиск электронных документов по элементам библиографического описания и другим признакам газетного материала и т. д. Характерными свойствами создаваемой электронной коллекции являются доступ в среде WEB и профиль метаданных для описания коллекции и отдельных изданий.

Выбор лингвистических средств

Традиционно в информационной системе значительное место занимает лингвистическое обеспечение (ЛО) – комплекс информационно-поисковых языков (ИПЯ) и лингвистических процессоров, предназначенных для обработки, представления и поиска электронных документов на семантическом уровне [3]. ЛО создаваемой электронной библиотеки включает следующие виды ИПЯ: метаданные; в том числе библиографические (подробнее см. [2]); классификационный и вербальный языки. Эти ИПЯ на разных уровнях отображают информацию, представленную в электронных документах (ЭД), что должно обеспечивать полноту и точность информационного поиска.

Обычно вербальные языки, среди которых широко распространен язык ключевых слов, являются центральным лингвистическим средством для выражения смыслового содержания текста. Достоинства координатного индексирования общеизвестны, но свойства предметной области коллекции и способ представления электронных документов, в нашем случае, делают этот процесс проблематичным и неоправданно трудозатратным. Среди основных причин отказа от применения этого подхода можно выделить следующие: во-первых, невозможность применения методов автоматической обработки текста для полнотекстового индексирования и основанного на них полнотекстового поиска; во-вторых, неконтролируемое и неуправляемое применение ключевых слов неизбежно приводит к значительным потерям в характеристиках полноты и точности поиска, для устранения этих недостатков как минимум необходимо решение проблемы дискрипторизации ключевых слов, а для организации действительно содержательного поиска предстоит сложный и трудоемкий, часто занимающий годы, процесс создания тезауруса. Для нашей коллекции мы сочли наиболее приемлемым вербальным ИПЯ – аннотацию. Она сжато характеризует тематическое содержание публикации, что предоставляет пользователю возможность определить релевантность результатов поиска и решить, следует ли обращаться к полному тексту.

В качестве основного инструмента для систематизации и поиска газетных публикаций нами был выбран подход, основанный на многоаспектной классификации текстов. Классификационные языки обладают рядом преимуществ перед другими типами поисковых языков, прежде всего, наглядностью, простотой для пользователя и независимостью от естественного языка. Фасетная система классификации позволяет выбирать признаки классификации (фасеты) независимо друг от друга. Каждый фасет содержит совокупность однородных значений данного классификационного признака. Фасетная система классификации позволяет многоаспектно (всесторонне) охарактеризовать специфический газетный материал.

Содержательная классификация (систематизация) документов широко используется при составлении библиографических пособий, в том числе во всевозможных указателях к содержанию периодических изданий. Так, еще в 1880 году был опубликован «Полный систематический указатель статей местно-областного содержания, напечатанных в "Казанских известиях", издававшихся при имп. Казанском университете с 1811 по 1821 гг.» (Казань, 1880), составленный П. Пономаревым. В нем краеведческий материал "Казанских известий" представлен в девяти разделах: космография, естественная история, топография, этнография, археология и история, промышленность, медико-топографические статьи, общественная жизнь, топо- и этнография.

Примером применения многоаспектного анализа печатных материалов при раскрытии содержания газеты является комплекс указателей к газете 18-го века «Санктпетербургские ведомости» [5]. Комплекс указателей (тематико-видовой, по странам и регионам, алфавитный, топографический указатель по Петербургу и его окрестностям, указатель имен, видов хозяйственно-культурной деятельности, учреждений) дает возможность подойти к искомой информации с разных сторон или провести поиск по совокупности признаков, исключая тем самым потери информации.

Особенности предметной области – газетного материала

Газеты 19-го века, в особенности его первой половины, значительно отличались от современных газет. «Казанские известия» (1811 – 1820 гг.) – первая провинциальная газета – имеет много общего с другими периодическими изданиями этого периода. В ней помещалось то же, с некоторым выбором и сокращением, что и в столичных газетах.

Вскоре после основания (с 19-го номера) она стала изданием Казанского университета. Главный контингент корреспондентов сложился из преподавателей гимназий и училищ всего востока Европейской России и Сибири, входивших в Казанский учебный округ. Содержание газеты представляет собой многоплановый синтетический материал, включающий в себя информацию, самую разнообразную по жанру, происхождению и содержанию. Наряду с официальными сообщениями и документами, законодательными актами, объявлениями, некрологами, письмами, городской хроникой в ней публикуется обширный научно-образовательный и литературно-художественный материал. Богатство и разнообразие информации делают ее многоплановым историческим источником. Особенность структуры газеты заключается в том, что в ее номерах нередко нет однозначно определенных рубрик, которые можно было бы использовать для тематического раскрытия содержания газеты, хотя бы на уровне ее разделов. Другой особенностью газеты этого времени является то, что значительная часть публикаций не озаглавлена и не подписана авторами, что делает невозможным полное библиографирование по заголовкам и фамилиям. Газета состоит из более или менее кратких сообщений, объявлений и других видов информации, многообразие которых затрудняет организацию поиска нужных сведений. Поэтому в качестве индексируемой единицы содержания выбрана часть текста, ограниченная определенной темой (сообщение о международных отношениях, сообщение о научных открытиях, всевозможные объявления, научные статьи и т. д.).

Газеты более позднего времени, например, «Казанский телеграф» (1893 – 1907 гг.), ближе к современным изданиям. В них более четко выражена структура газеты, существуют рубрики, присутствующие в каждом номере, некоторые рубрики привязаны к определенным страницам номера, но выбор индексируемой единицы содержания остается прежним, так как и в них достаточно много коротких, не озаглавленных заметок, объявлений и т. п. С другой стороны, присутствие таких рубрик, как реклама, позволяет в некоторых случаях описывать рубрику целиком, не детализируя отдельные объявления (отношение к рекламе заслуживает отдельного рассмотрения, так как с точки зрения некоторых исследователей газетного материала отдельные рекламные объявления должны быть обязательно выделены, например, театральные афиши, связанные с культурной жизнью города).

Аспекты анализа содержания газеты, используемые в процессе индексирования

При разработке системы классификации мы опирались на существующие методы, применяемые при составлении библиографических пособий, учитывали особенности конкретных периодических изданий и информационные запросы читателей.

В процессе аналитико-синтетической переработки газетного текста «Казанских известий» описывалась информация каждой единицы содержания в соответствии с разработанной системой классификации. Систематизация содержания газеты осуществлялась по следующим аспектам: виду информации, сфере общественной жизни, персонам (именам, встречающиеся в газете), учреждениям, географическим названиям мест, датам событий, приведенным в тексте. Внутри фасетов значения признаков либо просто перечисляются, либо образуют иерархическую структуру, если существует соподчиненность выделенных признаков.

Фасет **«Вид информации»** содержит перечень видов опубликованных в газете материалов и может быть представленным в виде следующей иерархической структуры:

Официальные материалы	– Дипломатические и др. документы и официальные письма;
	– О чинопроизводстве, прохождении службы, наградах;
	– Законодательные материалы;
	– О родившихся и умерших в губерниях, входивших в Казанский учебный округ
Сообщения	– О Международных отношениях;
	– О вооруженных конфликтах;
	– О внутреннем положении и общественно-политической жизни;
	– Дворцовая и светская хроника;
	– О стихийных бедствиях, эпидемиях и явлениях природы;
	– Об отъезжающих из Казани и приезжающих в Казань
Объявления	– О купле/продаже;
	– О найме/предложении рабочей силы;
	– Об аренде помещений и др. собственности;
	– О зрелищах и развлечениях;
	– О подписке на печатные издания (книги и периодические издания)
Финансово-экономическая информация	– Вексельный и денежный курсы;
	– О ценах, установленных полицией;
	– О сплаве грузов;
	– О таможенных сборах;
	– О ценах
Статьи и другие материалы	– Статьи естественного направления (химия, физика, естественная история);
	– Статьи гуманитарного направления (история, археология, этнография, словесность, экономика);
	– Ученые известия;
	– Художественные произведения (стихи, курьезы и чудеса, оды, повести, остроумные изречения, анекдоты, главы из книги и др.);

	– Публицистика; – Речи; – Описание увеселений (городские праздники); – О разном; – Описание церемоний (Общественных торжеств, городских событий)
Разное	– Статистические данные; – Некрологи; – Метеонаблюдения; – От редакции (опровержения, исправление опечаток и т. д.); – Полезные советы

Фасет **«Сферы общественной жизни»** содержит:

– Торговля (ярмарки, таможня);

– Промышленность;

– Сельское хозяйство, промысловая деятельность;

– Строительство. Благоустройство городов;

– Культура (Литература, театры, музыка, живопись, архитектура, скульптура, общественные развлечения);

– Коммуникации (перевозки, транспорт, почта);

– Просвещение;

– Наука и техника (новые изобретения, новые технологии);

– Здравоохранение;

– Благотворительность;

– Религии;

– Быт и нравы. Уголовная хроника.

Фасеты **«Персона»**, **«Учреждения»** и **«Географические названия мест»** в настоящий момент имеют линейную структуру и представляют собой перечни имен лиц, названий учреждений, географических названий мест, встречающихся в публикациях. В перспективе они должны быть сгруппированы внутри фасета в комплексы на основе общих признаков. Персоны по сословно-профессиональному признаку. В фасете

«Учреждения» будут выделены государственные, учебные и научные учреждения. **«Географические названия мест»** будут сгруппированы по принадлежности к странам. Внутри Российской империи будут выделены названия, связанные с Казанским учебным округом.

Совместное использование перечисленных выше аспектов рассмотрения газетных материалов позволяет получить самые разнообразные тематические комбинации.

Поскольку заранее невозможно определить все необходимые исследователю аспекты рассмотрения материалов периодического издания и необходимо учитывать неизбежный субъективизм систематизации, предусмотрена группировка текстов и по такому формальному признаку, как газетная рубрика. Несмотря на непостоянство названий рубрик в разных номерах газеты и неадекватное отражение ими тематики текстов, размещенных в разделе, такая систематизация наиболее точно и полно отражает структуру номеров газеты и может предоставить ряд дополнительных возможностей, в том числе и при поиске.

Дополнительно в процессе аналитико-синтетической обработки газетного текста формируются авторитетные файлы авторов, жанров периодики, список периодических изданий, из которых перепечатан новостной материал.

Моделью описания метаданных коллекции периодической печати была выбрана система RDF (Resource Description Framework) [2], поэтому **пример описания единицы содержания** приводится в виде XML-документа:

```
<?xml version="1.0" encoding="windows-1251"?>
<?xml-stylesheet type="text/xsl" href="ki.xsl"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dc="http://purl.org/dc/elements/1.1/"
     xmlns:lsl="http://lsl.ksu.ru/meta/">
<rdf:Description>
     <dc:Source>
          <lsl:Title>Казанские известия</lsl:Title>
          <lsl:Issue>6</lsl:Issue>
          <lsl:Date>19-01-1816</lsl:Date>
          <lsl:Page>1</lsl:Page>
     </dc:Source>
     <dc:Creator></dc:Creator>
     <dc:Title>Татарския училища</dc:Title>
     <dc:Subject>
          <lsl:Rubric></lsl:Rubric>
          <lsl:typeInform>Статьи и другие материалы</lsl:typeInform>
          <lsl:subType>Статьи гуманитарного направления
          (история, археология, этнография, словесность,
экономика)</lsl:subType>
          <lsl:field>Просвещение</lsl:field>
          <lsl:person></lsl:person>
          <lsl:institution></lsl:institution>
          <lsl:placename></lsl:placename>
          <lsl:event>  /  /    </lsl:event>
          <lsl:source></lsl:source>
     </dc:Subject>
     <dc:Description>Организация, система обучения, предметы
          преподавания и характеристика жизни в татарской
          школе</dc:Description>
     <dc:Language>
          <lsl:Language>rus</lsl:Language>
          <lsl:Grafic>old</lsl:Grafic>
```

```
      </dc:Language>
      <dc:Type>Image</dc:Type>
      <dc:Format>jpeg</dc:Format>
      <dc:Publisher>Научная библиотека КГУ</dc:Publisher>
      <dc:Contributor>Научная библиотека КГУ</dc:Contributor>
      <dc:Date>02/11/2005</dc:Date>
      <dc:Identifier>1816\1816-006-1.jpg</dc:Identifier>
</rdf:Description>
</rdf:RDF>
```

Заключение

Предлагаемый подход к организации лингвистических средств позволяет организовать удобный и эффективный информационный поиск, рассчитанный, в том числе, и на неподготовленного пользователя.

Как уже говорилось выше, предлагаемая система классификации была разработана для систематизации информационных материалов газеты «Казанские известия», которая обрабатывалась первой при создании коллекции казанской периодической печати 19-го – начала 20-го веков. В настоящее время коллектив Научной библиотеки КГУ продолжает дальнейшее формирование и развитие коллекции, используя разработанные подходы к систематизации материалов. Перечень основных категорий (фасетов) в незначительной степени изменяется для разных газет. Например, в газетах более позднего периода предполагается определять жанр публикации. В процессе анализа газетного материала вносятся уточнения и дополнения, но основной систематизации и в дальнейшем будет являться фасетный анализ содержания периодического издания. Визуализация классификационной схемы на странице поиска информационной системы должна быть обязательным элементом интерфейса пользователя. Она позволить пользователю осуществлять выбор и эффективно формировать поисковый запрос при любой последовательности выбора фасетов.

Литература

1. Абросимов А. Г., Кузьмина В. Ю. Электронная коллекция периодической печати конца XIX – начала XX вв.// Восьмая международная конференция «Libcom-2004». Сб. тезисов докл., 15 – 19 ноября 2004 г., Ершово. – 2004. – С. 5-7.
2. Абросимов А. Г. Метаданные описания коллекции периодической печати// Российский научный электронный журнал «Электронные библиотеки». – 2005. – Т. 8. – Вып. 2.
3. Антопольский А. Б. Лингвистическое обеспечение электронных библиотек / М.: ФГУП Науч.-техн. центр "Информрегистр", 2003. – 301 с.
4. Гендина Н. И. Состояние теории, практики и подготовки кадров в сфере лингвистического обеспечения информационно-библиотечной технологии: стимулы и препятствия. // Десятая Международная Конференция «Крым 2003». Сб. тезисов докл., 07-15 июня, 2003 г.
5. Газета "Санктпетербургские ведомости" XVIII века: Указатели к содержанию, 1756-1760/ Рос. АН, Б-ка; Сост. В.Ф. Воробьева и др.; Отв. сост. С.Н. Коротков. – СПб., 1994. – 588 с.

Об авторах

Абросимов Андрей Георгиевич - заместитель директора по информатизации Научной библиотеки им. Н.И. Лобачевского Казанского государственного университета
E-mail: aga@ksu.ru

Кузьмина Варвара Юрьевна - зав. отделом информатизации Научной библиотеки им. Н.И. Лобачевского Казанского государственного университета
E-mail: Varvara.Kuzmina@ksu.ru

Салосина Ирина Ивановна - инженер-программист Научной библиотеки им. Н.И. Лобачевского Казанского государственного университета
E-mail: Irina.Salosina@ksu.ru
