Использование средств лингвистической обработки текстов в системе мониторинга информационных ресурсов по пользовательским предпочтениям

Н.Л. Сомс, А.В. Добров, А.Е. Доброва

ООО «Гелайн», Университет ИТМО
nsoms@aiire.org, adobrov@aiire.org, adobrova@aiire.org

Аннотация

Современные системы мониторинга СМИ и социальных медиаресурсов в большинстве случаев не позволяют конечному пользователю осуществлять непосредственную индивидуальную настройку процедур фильтрации данных, которая давала бы ему возможность отслеживать публикации по созданным этим пользователям темам с учетом показателей охвата интернет-аудитории публикациями и темами. В данной статье описывается подход к решению этой задачи, основанный на методах лингвистической обработки текстов, примененный коллективом разработчиков ООО «Гелайн» при создании программного комплекса «СМиРТЕО».

Введение

На сегодняшний день сеть Интернет стала если не основным, то одним из основных средств получения новостной, познавательной и развлекательной информации конечным пользователем. К положительным качествам этой глобальной сети, безусловно, следует отнести объем генерируемого контента, который в равной степени можно считать и ее существенным недостатком. В потоке информации пользователю становится все сложнее выделить те публикации, которые в действительности относятся к интересующим его темам, при этом в большинстве случаев конечный пользователь даже не пытается отслеживать все такие публикации, так как считает эту задачу невыполнимой.

Помимо выделения информации в информационном потоке, следующей задачей для пользователя может стать выделение доверенных или интересующих его источников информации — тех источников, данные которых не вызывают у пользователя сомнений в их достоверности или интересуют его по форме или содержанию. Действительно, прочно занявшие свои позиции в жизни пользователей интернет-среды блоги, социальные сети и интернет-СМИ поставляют такой поток информации, для работы с которым пользователю необходимо так или иначе преодолевать проблему оценки каждого ресурса по степени доверия к этому ресурсу, по степени соответствия тематики ресурса темам, интересным пользователю, и по многим другим показателям.

С данной задачей сталкиваются не только интернет-пользователи, но и различные аналитические службы, для которых важна не только возможность фильтрации информационного потока, но и возможность проведения анализа публикаций по различным показателям. Одной из метрик, позволяющих ранжировать интернет-публикации, является оценка охвата публикацией интернет-аудитории: количество просмотров, комментариев, «репостов», «лайков» и другие величины, количественно отражающие интерес аудитории к данной публикации.

Как для информационно-аналитических служб, так и для конечного пользователя важно выделять в информационном потоке публикаций те сообщения, которые относятся к теме, соответствующей каким-либо событиям окружающей действительности, либо отражающим интересы пользователя. Для работы аналитиков при этом не менее важно отслеживать вышеупомянутые показатели охвата аудитории, но уже не публикациями, а темами этих публикаций. Эта величина (охват темой аудитории) показывает значимость той или иной темы в каждый момент времени.

1. Постановка задачи

Для решения проблемы оценки и фильтрации данных используются различные подходы, которые можно разделить на 2 группы:

– основанные на ограничении источников информации (RSS-клиенты, клиенты отдельных новостных лент),

Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS-2014, Санкт-Петербург, 19 - 20 ноября 2014 г.

- основанные на фильтрации самой информации (новостные агрегаторы, системы мониторинга).

Основным недостатком подходов первой группы является сама идея ограничения источников информации и невозможность выделения интересующей пользователя информации по ее содержимому. Вторая группа лишена этого недостатка, однако характеризуется высоким уровнем информационного шума в связи с использованием ограниченных подходов к анализу текста при группировке сообщений по темам. К числу таких подходов можно отнести:

- подходы, основанные на кластеризации текстов (см. [6], [13], [14], [15]);
- подходы, основанные на классификации текстов (см. [3], [4], [5], [7], [8], [9], [10], [11], [12]).

Подходы, основанные на кластеризации текстов, предполагают использование различных методов кластерного анализа при выделении тем. При этом текст, как правило, считается вектором в n-мерном пространстве текстов, измерениями в котором являются абсолютные, относительные или взвешенные (например, при помощи метрики TF*IDF) частоты алфавитных цепочек (форм слов). Методы кластерного анализа позволяют весьма условно выделить в массиве текстов некие группы текстов, которые, с точки зрения системы, являются в некоторой степени сходными и потому объединяются в один кластер. Пользователь при этом не может напрямую управлять набором и составом кластеров; пользователю лишь иногда предоставляется возможность контролировать величины используемых пороговых значений и весовых коэффициентов.

Подходы, основанные на классификации (автоматической рубрикации) текстов, основываются на различных методах автоматической классификации. Эти методы предполагают, что набор классов (тем) задан заранее и не может изменяться в процессе работы системы мониторинга. Поэтому пользователь в общем случае не может создавать новые классы (темы): если метод классификации основан на алгоритмах машинного обучения, то создание новой темы требует переобучения классификатора; если же в основе классификатора лежит система правил (образов рубрик), то для создания новой темы пользователь должен создавать сложные правила (булевы формулы, правила продукций) не только для новой темы, но и для остальных имеющихся в системе тем, с целью недопущения ложных корреляций и проблем, связанных с неоднозначностью языковых единиц.

Программный комплекс «Система мониторинга и рейтингования тем с точки зрения их ежедневного охвата аудитории» (http://asm.aiire.org/smirteo), разработанный коллективом ООО «Гелайн» и описываемый в данной статье, позволяет частично преодолеть недостатки обоих подходов путем их комбинирования и применения методов компьютерной лингвистики к задаче тематической атрибуции публикаций таким образом, чтобы пользователь системы имел возможность самостоятельно создавать интересующие его темы, пользуясь средствами естественного языка. Назначением разработанного комплекса является мониторинг СМИ, блогов и социальных сетей с возможностью оценки охвата аудитории по имеющимся у публикации показателям охвата — количествам комментариев, «лайков», «репостов» и т.п., а также выделении публикаций, относящихся к пользовательским темам и совокупной оценке показателей охвата тем. Данный проект разрабатывался с использованием уникальных лингвистических технологий, разработанных коллективом.

В ходе выполнения данной работы, коллективом было сформулировано и решено несколько основных проблем, связанных с поставленной задачей:

- мониторинг публикаций СМИ, блогосферы и социальных сетей;
- мониторинг информации о публикациях — показателях охвата, ключевых словах, «метках», указанных в источнике для публикации;
- выделение заголовков, текстов, ключевых слов на страницах публикаций;
- определение соответствия текста публикации пользовательским интересам — темам;
- вычисление величины показателей охвата публикации или темы.

2. Задачи мониторинга

С развитием сети Интернет как сферы функционирования медиадискурса все большую актуальность приобретает такая особенность медиатекста, как характер его авторства. Так, для новостных сообщений актуальна информация о средстве массовой информации, опубликовавшем данное сообщение, вне зависимости от того конкретного информационного ресурса, на котором оно опубликовано — официального сайта данного СМИ, группы в социальной сети, блога или иного ресурса. Той же особенностью характеризуются и публикации блоггеров, а также иных лиц, обладающих различными аккаунтами в социальных сетях: для таких публикаций соотнесение их с авторами в большинстве прикладных задач оказывается более актуальным, чем информация об интернет-ресурсе. Здесь и далее лица (персоны и организации), являющиеся авторами публикаций, будут именоваться *инфосубъектами*.

Вместе с тем, с точки зрения особенностей процедуры мониторинга различные интернет-ресурсы и их разделы представляют собой принципиально разные объекты анализа. Как будет показано ниже, различные интернет-ресурсы

обладают собственной спецификой во всех аспектах, значимых для мониторинга: различаются и способы организации данных (структура представления публикаций), и состав предоставляемых ресурсом реквизитов публикаций, в том числе — показателей охвата аудитории. Поэтому для каждого интернет-ресурса (а в некоторых случаях — и раздела интернет-ресурса) требуется либо отдельный алгоритм загрузки данных, либо особый набор настроек, регулирующих эту загрузку. В различных интернет-ресурсах различаются даже режимы доступа к данным и протоколы их загрузки (некоторые ресурсы, например, сайт «vk.com», не предназначены для непосредственного доступа приложений к их содержимому по протоколу HTTP, однако предоставляют API, позволяющий загружать данные в формате JSON, в то время как большинство новостных сайтов не предоставляют открытого API и доступны только по протоколу HTTP). Поэтому при построении систем мониторинга информация об интернет-ресурсах, предоставляющих публикации, является определяющей. Здесь и далее интернет-сайты и их разделы, обладающие инвариантными относительно публикаций структурой представления данных и режимом доступа к ним, будут называться *инфоресурсами*. Объект (в терминах объектно-ориентированной парадигмы), выполняющий мониторинг инфоресурса, далее будет называться *монитором*.

Соответствие между инфосубъектами и инфоресурсами является инъективным, но не биективным: один инфоресурс соответствует одному инфосубъекту, но один инфосубъект может публиковать тексты на разных инфоресурсах.

Процедура мониторинга инфоресурса предполагает постоянное получение данных о последних по дате создания или редактирования публикациях, размещенных в этом инфоресурсе. Многие инфоресурсы предоставляют такую информацию в формате RSS, однако частота обновления RSS в большинстве случаев оказывается недостаточно высокой (например, требуется загрузка новых публикаций в систему мониторинга в течение одной секунды после их опубликования, а RSS обновляется раз в 15 минут). Кроме того, даже те инфоресурсы, которые предоставляют RSS-ленты, часто предоставляют отдельные ленты для разных разделов или рубрик, однако не предоставляют информации о полном перечне всех новых публикаций. Поэтому, помимо загрузки списка публикаций при помощи RSS, может возникнуть необходимость реализации альтернативных методов загрузки этой информации монитором. В частности, данная задача может быть решена при помощи создания парсера главной страницы веб-сайта, если она содержит список новых публикаций, или иным путем, если такая информация впрямую не предоставляется (вплоть до полного перебора всех размещенных на инфоресурсе публикаций и сличения их со списком уже загруженных).

Поскольку появление новых публикаций необходимо отслеживать непрерывно, а загрузку конкретной публикации необходимо производить только по факту ее появления, в разработанной системе мониторинга эти процессы разведены и выполняются параллельно, в разных потоках. Соответственно, структура монитора должна обеспечивать независимость этих процессов, а программный интерфейс — предоставлять различные методы для них. Вместе с тем, в некоторых случаях это может вызвать ряд проблем, так как часть реквизитов публикации может быть доступна только в списке публикаций, в то время как остальные реквизиты загружаются одновременно с ее текстом. Более того, загрузка нескольких новых публикаций осуществляется параллельно (в противном случае исключается возможность гарантии мгновенной загрузки публикации). Поэтому в мониторе должно быть предусмотрено хранение в разделяемой потоками области оперативной памяти реквизитов, загруженных из списка публикаций, причем доступ к этой области должен регулироваться одноместными семафорами (мьютексами) или иными средствами блокирования ресурсов, гарантирующими безопасность работы монитора в многопоточном режиме.

Помимо загрузки новых публикаций, система мониторинга постоянно отслеживает наличие изменений в уже загруженных публикациях. Для этого для каждого монитора выделяется отдельный поток, выполняющий повторную загрузку уже загруженных публикаций, и производящий их обновление в используемой базе данных. В отличие от процедуры загрузки новых публикаций, процедура обновления уже загруженных предполагает значительно большие объемы данных и регулярное повторение запросов на получение каждой публикации. Такой режим работы системы мониторинга противоречит требованиям многих инфоресурсов, в частности, платформы «Живой Журнал», рассчитанных на однопоточный режим загрузки данных, с существенными ограничениями на повторные запросы. Для решения этой проблемы в созданной системе мониторинга был разработан специализированный модуль, реализующий функциональность настраиваемого кеширующего прокси-сервера.

Работа с различными инфоресурсами требует разных подходов к обработке предоставляемых данных. Для построения монитора для каждого инфоресурса необходимо экспериментальное исследование этого инфоресурса, позволяющее выявить особенности его организации и режимов доступа к нему, а также выбрать оптимальный алгоритм работы с этим ресурсом. В рамках данного исследования были выявлены следующие виды мониторов, соответствующие различным

алгоритмам работы с инфоресурсами в зависимости от их особенностей:

- мониторы ресурсов, предоставляющих программный API;
- мониторы RSS-лент, предоставляющих содержимое в виде RSS;
- мониторы лент, предоставляющих заголовки публикаций и ссылки на содержимое в виде html-страниц;
- мониторы, требующие разбора html-содержимого.

Кроме того, в рамках данного исследования были выделены следующие виды получаемых данных, каждый из которых требует особой, но вместе с тем единообразной для всех инфоресурсов обработки:

- списки публикаций инфоресурса: список новых публикаций, список всех публикаций (при наличии), список публикаций по заданной рубрике (при наличии) и др.;
- заголовок и текст конкретной публикации;
- инфосубъект, являющийся автором публикации;
- показатели охвата при их наличии;
- ключевые слова, используемые на инфоресурсе для отождествления публикаций сходной тематики;
- дата и время подготовки, публикации, последнего редактирования и др.

Создание мониторов, предоставляющих API доступа к публикациям, требует исключительно следования требованиям, описанным в спецификациях к программным интерфейсам. Вместе с тем, в некоторых случаях показатели охвата публикаций не могут быть получены посредством предоставляемых API, и для получения показателей охвата публикаций необходимо использовать сторонние службы. Так, API ресурса «vk.com» не позволяет получить информацию об охвате публикаций, размещенных на данном ресурсе, в иных инфоресурсах — например, количество цитирований данной публикации в социальной сети «Twitter».

Мониторы инфоресурсов, предоставляющих RSS-ленту, используют стандартные средства доступа к таким лентам и, получив список публикаций, создают потоки для загрузки каждой публикации из ленты.

Наиболее сложным для программной реализации видом мониторов является группа мониторов, осуществляющих разбор html-содержимого страниц для получения списка публикаций, заголовка и текста конкретной публикации, ее показателей охвата и ключевых слов при их наличии. Такие мониторы требуют реализации для каждого инфоресурса особого набора настроек, выявляемых вручную путем проведения серии экспериментов с различными публикациями, размещенными на этом инфоресурсе. Так, помимо ряда служебных данных, в мониторе информационного ресурса *Aljazeera* определены:

- путь к RSS (*http://www.aljazeera.com/Services/Rss/?PostingId=2007731105943979989*);
- путь к полному тексту статьи на языке XPATH (*//td[@id='ctl00_cphBody_tdTextContent' and @class='DetailedSummary']*);
- путь к ключевым словам публикации на языке XPATH (*//td[@class='DivTitles']*);
- наименование файла с таблицей соответствий между ключевыми словами инфоресурса *Aljazeera* и концептами онтологии AIIRE (*aljazeera_tags.txt*).

Отдельного внимания требует обработка ключевых слов, используемых на инфоресурсе и указанных у публикаций. Для обеспечения возможности отождествления рубрик, используемых различными инфоресурсами, а также создания механизма быстрой фильтрации публикаций, для каждого монитора должен быть указан файл связей ключевых слов инфоресурса и концептов универсальной компьютерной онтологии (в данном исследовании использовалась онтология AIIRE). Принцип построения отношений в онтологии AIIRE позволяет делать сложные выборки публикаций с учетом их классификаций и таксономий концептов онтологии — например, по концепту *Центральный регион* выделить публикации, помеченные ключевыми словами *Москва*, *Тула*, *Московская область* и т.п.

Файл связей ключевых слов и концептов онтологии для вышеприведенного монитора инфоресурса *Aljazeera,* aljazeera_tags.txt, содержит в себе таблицу рубрик инфоресурса, которые привязаны к соответствующим концептам онтологии AIIRE. Примеры таких соответствий приведены в Таблице 1.

Таблица 1. Соответствие между ключевыми словами инфоресурса «Aljazeera» и идентификаторами концептов онтологии AIIRE

Ключевые слова	Идентификаторы
Africa	268127
Americas	224438
Asia-Pacific	189725
Central & South Asia	189725
Europe	602891
Middle East	1394383
Sport	1390995

Вышеприведенный инфоресурс *Aljazeera* характеризуется относительным постоянством состава используемых ключевых слов. Вместе с тем, некоторые ресурсы, например, *Кавказ Узел*, постоянно пополняют состав и увеличивают количество используемых ключевых слов, что означает, что

список таких слов существенно больше с самого начала и постоянно пополняется в ходе работ по обновлению и усовершенствованию монитора. Например, состав ключевых слов монитора *Кавказ Узел* включает в себя не только набор относительно стандартных тем (примеры приведены в табл. 2), но и постоянно пополняющийся набор географических объектов (см. табл. 3).

Таблица 2: Ключевые слова инфоресурса «Кавказ-Узел», соответствующие стандартным темам

Ключевые слова	Идентификаторы
Права человека	1204815
Политика	602894
Общество	2047418
Конфликты	802176
Экономика	1918520
Культура	845387
Преступность	2047495
Природа и экология	2047414
Происшествия	1220505
Туризм	2047466

Таблица 3: Ключевые слова инфоресурса "Кавказ-Узел", соответствующие географическим объектам

Ключевые слова	Идентификаторы
Кабардино-Балкария	709541
Карачаево-Черкесия	732863
Дагестан	533264

Постоянные изменения в составе ключевых слов инфоресурса влекут за собой не только необходимость в регулярном отслеживании таких изменений, но и часто приводят к необходимости редактирования онтологии (добавления в нее новых концептов или уточнения атрибутов уже имеющихся концептов), с тем, чтобы структура онтологии была актуальной с точки зрения изменяющихся особенностей инфоресурсов и разработанных мониторов.

Следует также отметить некоторые сложности при выделении XPATH заголовков, текста публикаций и других элементов, выявленные в процессе данного исследования. Низкое качество верстки html-страниц часто приводит к необходимости конструирования сложных алгоритмов выделения исходного текста, иногда основанных не на форматировании страниц, а на их текстовом содержимом. Так, текст публикации может не отличаться по форматированию от следующего непосредственно за ним списка наиболее обсуждаемых тем, не имеющих к тексту никакого отношения. Если монитор не производит разграничения текста и этого списка, то при дальнейшей лингвистической обработке публикация может быть отнесена к некорректной теме. Кроме того, частые смены дизайна веб-сайтов или просто изменение шаблона дизайна требует повторного выделения путей XPATH. Аналогично, выделение пути к ключевым словам, часто вообще неотделимых друг от друга или от тела текста сообщения по их форматированию, может представлять определенные сложности. Те же проблемы возникают и при идентификации показателей охвата в тех случаях, когда их значения входят в html-содержимое публикации, а не загружаются асинхронным методом.

3. Задачи лингвистической обработки

Помимо задач, связанных с обработкой ключевых слов и их отнесения к концептам онтологии AIIRE, отнесение публикации к заданным пользователями темам (пользовательским интересам) требует лингвистической обработки. В рамках поставленной задачи пользовательские темы задаются набором ключевых слов, при этом публикация считается принадлежащей теме в том и только том случае, если все ключевые слова присутствуют в тексте публикации. Если какое-то слово из набора ключевых слов отсутствует в публикации, то, по постановке задачи, она не должна быть отнесена к этой теме.

Например, тема «Антитабачный закон» описана пользователем системы с помощью набора ключевых слов «табак антитабачный закон».

3.1.Морфологический анализ

Очевидно, в неизменном виде (в словарной форме) данные ключевые слова могут встретиться далеко не во всех публикациях. Для снятия этого ограничения заголовки и тексты всех публикаций подвергаются морфологическому анализу с помощью универсального лингвистического процессора AIIRE, выполняющего морфологический анализ на основе морфологического словаря. Морфологический словарь AIIRE состоит из словарных статей следующего вида:

```
<entry hw="санкция">
  <inflection template="сущ ru f ina 7a">
    <variable name="основа"
    value="са́нкци"/>
  </inflection>
  <attributes>
    <attr name="lemma" value="санкция"/>
    <attr name="anim" value="0"/>
    <attr name="isName" value="0"/>
    <attr name="pos" value="noun"/>
    <attr name="gender" value="f"/>
  </attributes>
</entry>
```

В теге *entry* заключено всё тело словарной статьи, атрибутом элемента *entry* служит элемент *hw(head word)* и его значение, т.е. заголовочное слово статьи. В данном случае это слово *“санкция”*. Внутри элемента *entry* существует два вложенных элемента - *inflection* и *attributes*. Любое изменяемое

по формам слово может быть записано таким образом и представлено в виде словарной статьи.

Элемент *inflection* представляет собой основную с точки зрения морфологии часть статьи, а именно, он содержит тип склонения слова, помещённого в элемент *hw*. Тип склонения представлен концепцией склонения российского лингвиста А.А. Зализняка. Концепция состоит из алгоритмического описания склонения имён в письменной форме русского языка. Этот словарь, составленный А. А. Зализняком [1], стал основой практически для всех программ морфологического анализа, используемым в системах машинного перевода и информационного поиска. Автор словаря указал точные модели словоизменения для 100 тыс. слов русского языка.

Данная система акцентно-словоизменительных типов используется также в русскоязычном Викисловаре для описания морфологии русских имён и глагола (следует, однако, отметить, что глагольное словоизменение в словаре [1] описано значительно менее подробно, чем именное, поэтому в проекте AIIRE, помимо данного словаря, в качестве основного источника информации о глагольной морфологии служат материалы докторской диссертации С.А. Кузнецова [2]).

В атрибуте *template* указано наименование шаблона словоизменения. В приведенном выше примере используется шаблон, название которого содержит в себе указание на часть речи (в данном случае — имя существительное), анализируемый язык (русский), род (женский), одушевлённость/неодушевлённость и тип склонения. Наименования типов склонения А.А. Зализняка включают в себя цифры, буквы и другие обозначения при них. В данном случае шаблон, помимо части речи, языка и одушевлённости, даёт информацию о том, что основа существительного оканчивается на *-и* (*7*), имеет ударение всегда на основу (*a*), и в ней не происходит чередований. При наличии чередования, оно отражается в наименовании шаблона знаком *.

Элемент *variable* позволяет определять значения переменных, используемых при порождении формы слова. В приведенном примере такой переменной является «основа», к которой в указанном в *inflection* шаблоне указано присоединение конкретных флексий (см. ниже). Значение переменной (т. е. сама основа) представлено в атрибуте *value*.

```
<variable name="основа" value="са́нкци»/>
```

Во многих случаях, когда в основе происходит чередование (часто — несколько чередований), две или три переменных основы для порождения корректных форм, например:

```
<entry hw="мышонок">
  <inflection template="сущ ru m a 3°a">
    <variable name="основа2"
    value="мыше́нк"/>
    <variable name="основа"
    value="мыше́нок"/>
    <variable name="основа1"
    value="мыша́т"/>
  </inflection>
  <attributes>
    <attr name="lemma" value="мышонок"/>
    <attr name="anim" value="1"/>
    <attr name="isName" value="0"/>
    <attr name="pos" value="noun"/>
    <attr name="gender" value="m"/>
  </attributes>
</entry>
```

В данном примере представлены особый тип чередования *о/-* и неравносложность основ, обозначенные знаком °, что потребовало указать вторую основу. Третья основа понадобилась для корректного образования форм множественного числа, требующих несистемного чередования *онок/ат*.

Элемент *attributes* содержит в себе значения атрибутов слова. Атрибут «лемма» позволяет в конечном итоге отождествить различные формы одного и того же слова и в то же время различить омонимы. Атрибут *«anim»* может принимать значения *«0»/«1»* (неодушевленный/одушевлённый) соответственно. Атрибут *«isName»* работает таким же образом, как и *«anim»,* принимая те же значения и обозначает, является ли имя существительное именем собственным. Атрибуты *«pos»* и *«gender»* принимают значения заданной части речи и рода, соответственно.

3.2. Формообразовательные (словоизменительные) шаблоны

Структура шаблона словоизменения представляет собой набор соответствий между типами основ и присоединяющийся к ним окончаний и/или иных аффиксов. Например, для представленной выше статьи «санкция» определен шаблон «*сущ ru f ina 7a*», представленный ниже.

```
<template name="сущ ru f ina 7a">
  <flection context="основа" content="ях">
    <attr name="case" value="loc"/>
    <attr name="num" value="pl"/>
  </flection>
  <flection context="основа" content="й">
    <attr name="case" value="gen"/>
    <attr name="num" value="pl"/>
  </flection>
  <flection context="основа" content="и">
    <attr name="case" value="gen"/>
    <attr name="num" value="sg"/>
  </flection>
  <flection context="основа" content="и">
    <attr name="case" value="loc"/>
    <attr name="num" value="sg"/>
  </flection>
  <flection context="основа" content="и">
    <attr name="case" value="dat"/>
    <attr name="num" value="sg"/>
  </flection>
  <flection context="основа" content="и">
    <attr name="case" value="acc"/>
    <attr name="num" value="pl"/>
  </flection>
  <flection context="основа" content="и">
    <attr name="case" value="nom"/>
```

```
    <attr name="num" value="pl"/>
  </flection>
  <flection context="основа" content="ею">
    <attr name="case" value="ins"/>
    <attr name="num" value="sg"/>
  </flection>
  <flection context="основа" content="ей">
    <attr name="case" value="ins"/>
    <attr name="num" value="sg"/>
  </flection>
  <flection context="основа" content="я">
    <attr name="case" value="nom"/>
    <attr name="num" value="sg"/>
  </flection>
  <flection context="основа" content="ю">
    <attr name="case" value="acc"/>
    <attr name="num" value="sg"/>
  </flection>
  <flection context="основа"
content="ями">
    <attr name="case" value="ins"/>
    <attr name="num" value="pl"/>
  </flection>
  <flection context="основа" content="ям">
    <attr name="case" value="dat"/>
    <attr name="num" value="pl"/>
  </flection>
</template>
```

Данный шаблон содержит все флексии, которые могут присоединяться к основе и, тем самым, образовывать формы слова *санкция*. Для каждой флексии указывается контекст (имя переменной, содержащей основу, к которой присоединяется флексия), определенный в словарной статье (см. выше), орфографическая запись флексии и те грамматические атрибуты, которые обозначаются этой флексией. В данном случае указываются атрибуты «*case*» (падеж) и «*num*» (число).

4. Показатели охвата аудитории

В отличие от иных видов медиатекстов, интернет-публикации характеризуются особым набором атрибутов (реквизитов), имеющих особую значимость. К числу таких реквизитов относится не только заголовок, ключевые слова и различные даты (создания, публикации, редактирования), но и реквизиты, специфичные для сети Интернет, свидетельствующие о степени охвата данной публикацией интернет-аудитории: количество просмотров, внутренних и внешних комментариев, «репостов», «лайков» и др. Такие реквизиты далее будут именоваться *показателями охвата*. Они специфичны для каждого инфоресурса и в общем случае имеют различную интерпретацию: например, количество просмотров новостной статьи имеет природу, отличную от количества просмотров публикации того или иного пользователя социальной сети (в последнем случае просмотры обусловлены непосредственными отношениями между пользователями социальной сети).

4.1. Внутренние и внешние реквизиты публикаций

Поэтому, прежде всего, важно разграничивать *внутренние* и *внешние* показатели охвата. К числу внутренних показателей охвата относятся данные об охвате аудитории в пределах того инфоресурса, в котором размещена публикация. Как правило, внутренние показатели охвата отображаются в том же html-файле, в котором представлен текст публикации, однако в некоторых случаях даже внутренние показатели охвата требуют отдельной загрузки, так как на странице публикации загружаются асинхронным методом. В отличие от внутренних, внешние показатели охвата свидетельствуют об охвате публикацией аудитории инфоресурсов, отличных от инфоресурса этой публикации. Например, количество «ретвитов» новостной статьи свидетельствует об охвате этой статьей аудитории платформы «Twitter». Такие показатели охвата по определению не могут быть загружены из того же инфоресурса, что и публикация, и требуют отдельных загрузчиков.

4.2. Виды показателей охвата аудитории, особенности различных ресурсов

В созданной системе мониторинга была разработана общая классификация показателей охвата аудитории, актуальных для различных инфоресурсов. Помимо деления показателей охвата на внутренние и внешние, данная классификация позволяет разграничивать просмотры и реакции; к числу реакций относятся комментарии, оценки (лайки) и цитирования («репосты», «ретвиты», «ссылки» и др.). Реакции во внешних системах классифицируются по системам («Количество лайков в VK», «Количество репостов в Surfingbird» и т. д.)

Данная классификация содержит абстрактные виды показателей охвата и позволяет обеспечивать наследование некоторых характеристик (например, коэффициента значимости показателя охвата) более частным видом от более общего, однако интерпретации в системе мониторинга подвергаются конкретные показатели охвата, а не их виды, и свойства вида учитываются лишь в тех случаях, когда они не переопределены в конкретном показателе (под конкретным показателем охвата понимается пара «инфоресурс, вид показателя охвата»).

В частности, значимость количества комментариев для оценки охвата аудитории в блогосфере значительно ниже, чем значимость, казалось бы, того же количества комментариев в СМИ: комментирование новости говорит об активной реакции постороннего читателя на описываемое в ней событие, в то время как комментирование сообщения блоггера может быть обусловлено отношением подписчика этого блоггера к нему или даже к другим подписчикам, ранее прокомментировавшим данное сообщение. Таким образом, показатели охвата индивидуальны для каждого инфоресурса.

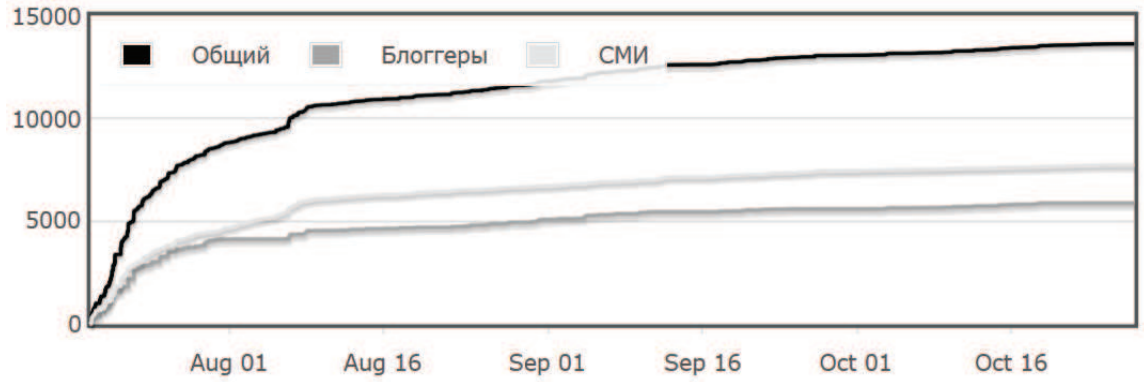

Рис. 1: График изменения суммарного показателя охвата публикаций по теме «Boeing» (2014 г.)

4.3. Усреднение показателей охвата аудитории, расстановка весовых коэффициентов

Вместе с тем, как отмечалось выше, разные инфоресурсы могут соответствовать одному инфосубъекту. Более того, для решения задачи отслеживания охвата аудитории не конкретными публикациями, а их темами, необходимо неким образом усреднять конкретные показатели охвата публикаций и вводить весовые коэффициенты, обеспечивающие возможность сравнения разных тем по признаку охвата аудитории. В частности, количество комментариев чаще всего существенно меньше, чем количество просмотров, поэтому при общей оценке охвата аудитории вес количества комментариев должен быть большим, чем вес количества просмотров. Весовые коэффициенты показателей охвата аудитории определяются эмпирическим путем, исходя из следующих требований:

- средние значения произведений w_x*x и w_y*y, где w_x — вес показателя x, w_y — вес показателя y, должны быть равны (расчет средних производится на всем массиве загруженных публикаций);
- дисперсии вышеупомянутых произведений (на том же массиве публикаций) должны быть равными;
- весовые коэффициенты должны измеряться в интервале от нуля до единицы.

Соблюдение вышеперечисленных требований позволяет производить усреднение показателей охвата публикации. В данном исследовании был выбран метод усреднения при помощи расчета среднего гармонического: в отличие от среднего арифметического, этот метод обеспечивает приоритет публикаций, обладающих высокими значениями всех показателей охвата, над теми, у которых по тем или иным причинам чрезвычайно высоко значение только одного показателя, в то время как остальные показатели близки к нулю. Поскольку при расчете среднего гармонического требуется деление единицы на значение усредняемой величины, возникла проблема нулевых значений: например, при нулевом количестве комментариев и даже ненулевом количестве просмотров расчет среднего гармонического оказывается невозможным из-за деления на нуль. Поэтому к каждому показателю прибавляется поправочная константа (единица), которая далее вычитается из полученного значения среднего гармонического. Таким образом, для отдельно взятой публикации усредненный показатель охвата оценивается по формуле:

$$\overline{cov} = \frac{n}{\sum_{i=1}^{n} \frac{1}{w_{cov_i} * cov_i + 1}} - 1$$

, где cov — показатель охвата, w — вес показателя охвата.

Следует отметить, что величина поправочной константы до сих пор остается предметом изучения; авторам данной статьи на сегодняшний день не удалось установить, является ли выбранное значение (1) оптимальным. Проблема состоит в том, что на данный момент единственным способом верификации данной величины является экспертная оценка качества результатов усреднения, однако такая оценка выявила лишь следующее:

- с ростом поправочной константы результаты усреднения увеличиваются в абсолютном выражении;
- зависимость роста усредненных значений от роста самих значений уменьшается с увеличением поправочной константы;
- с уменьшением поправочной константы повышается величина ошибки при операциях с дробными величинами.

Значение поправочной константы должно выбираться в соответствии с вышеуказанными принципами.

4.4. Расчет динамики охвата аудитории темой

Общий охват аудитории темой оценивается как сумма усредненных показателей охвата публикаций, относящихся к этой теме. С появлением новых публикаций данная величина увеличивается. В созданной системе мониторинга

предусмотрен интерфейс построения графиков изменения показателей охвата тем во времени, при этом на графике также отдельно отображаются изменения суммарного показателя темы в социальных инфоресурсах (блогах и социальных сетях) и в СМИ (рис. 1).

Кроме того, в системе предусмотрена возможность отслеживания первой производной этих функций, показывающей динамику изменения скорости роста суммарных показателей охвата во времени.

Заключение

Хотя система «СМиРТЕО» представляет собой опубликованный программный комплекс, доступный для пользователей и выполняющий вышеописанные функции, описанное в данной статье исследование нельзя назвать завершенным. На сегодняшний день показана эффективность использования средств лингвистического анализа текстов при тематической атрибуции текстов, предполагающей динамическое изменение набора и состава тем пользователями системы, однако далеко не весь арсенал средств лингвистического анализа задействован в вышеуказанном программном комплексе. На следующих этапах исследования планируется использовать средства синтаксического и семантического анализа текстов, ранее применявшимися коллективом разработчиков комплекса в системах интеллектуального поиска и автоматической рубрикации, для повышения качества отнесения текстов к пользовательским темам. Кроме того, отдельного исследования требует вопрос о способах расчета оптимальных значений весовых коэффициентов различных показателей охвата, а также поправочной константы в формуле их усреднения.

Вместе с тем, представленное исследование является одной из первых попыток изучения вопроса о способах создания системы, которая могла бы позволить пользователям получить полный контроль над выделяемыми системой темами, вне зависимости от их актуальности с точки зрения статистических данных. В созданной системе полностью реализован механизм такого контроля — от создания темы до отслеживания публикаций по этой теме и ее показателей охвата. Дальнейшие исследования могут быть направлены на повышение качества работы этой системы.

Литература

[1] Зализняк А.А. Грамматический словарь русского языка: Словоизменение. М.: «АСТ-ПРЕСС», 2008. — 794 с.

[2] Кузнецов С.А. Глагольное словоизменение и формообразование в современном русском языке: дис. … докт. фил. наук: 10.02.01. СПб., 2000. — 314 с.

[3] Агеев М.С. Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов: дис. ... канд. физ.-мат. наук: 05.13.11 / Московский гос. унив. — М., 2005. — 136 с.

[4] Агеев М.С. УИС РОССИЯ в РОМИП'2007: поиск и классификация / М.С. Агеев, Б.В. Добров, П.В. Красильников, Н.В. Лукашевич, А.М. Павлов, А.В. Сидоров, С.В. Штернов // Российский семинар по оценке методов информационного поиска. Труды РОМИП 2007-2008: Семинар в рамках Всероссийской науч. конф. RCDL'2007. 18 окт. 2007 г., Переславль-Залесский. — СПб.: НУ ЦСИ, 2008. — С. 199-220.

[5] Агеев М.С., Добров Б.В., Луцкашевич Н.В. Автоматическая рубрикация текстов: методы и проблемы // Ученые записки Казанского Государственного Университета. Серия Физико-математические науки. Т. 150, кн. 4. — Казань: Казанский государственный университет, 2008. — С. 25-40

[6] Антонов А.В. Галактика-Zoom на РОМИП'2009 / А.В. Антонов, С.Г. Баглей, В.С. Мешков, В.А. Стоян // Российский семинар по оценке методов информационного поиска. Труды РОМИП 2009. — СПб.: НУ ЦСИ, 2009. — 198 с.

[7] Белов А.А., Волович М.М. Автоматическое распознавание тематики сверхкоротких текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая - 3 июня 2007 г.) / Под ред. Л.Л. Иомдина, Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. — М.: Изд-во РГГУ, 2007. — С. 35-37.

[8] Агеев М.С. Экспериментальные алгоритмы поиска/классификации и сравнение с "basic line" / М.С. Агеев, Б.В. Добров, Н.В. Лукашевич, А.В. Сидоров // Российский семинар по оценке методов информационного поиска (РОМИП 2004): Семинар в рамках Всероссийской науч. конф. RCDL'2004. 1 окт. 2004 г. (Пущино, 2004). — СПб.: Изд-во НИИ химии СПбГУ, 2004. — С. 62-89.

[9] Агеев М.С. УИС РОССИЯ в РОМИП 2008: поиск и классификация нормативных документов / М.С. Агеев, Б.В. Добров, Н.В. Лукашевич, С.В. Штернов // Российский семинар по оценке методов информационного поиска. Труды РОМИП 2007-2008: Семинар в рамках Всероссийской науч. конф. RCDL'2008. 9 окт. 2008 г., Дубна. — СПб.: НУ ЦСИ, 2008. — С. 44-58.

[10] Васильев В.Г. Комплексная технология автоматической классификации текстов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4-8 июня 2008 г.). Вып. 7 (14). — М.: РГГУ, 2008. — С. 83-91.

[11] Васильев, В.Г. Выделение фрагментов в текстах

при классификации // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15). — М.: РГГУ, 2009. — С. 83-90.

[12] Васильев, В.Г. Обучение классификаторов на основе выделения фрагментов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). — М.: Изд-во РГГУ, 2010. — С. 62-70.

[13] Кураленок И.Е., Некрестьянов И.С. Автоматическая классификация документов на основе латентно-семантического анализа // Труды первой всероссийской научно-методической конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — СПб., 1999. — С. 89-96.

[14] Лукашевич Н.В., Добров Б.В. Автоматическое аннотирование новостных кластеров на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15). — М.: РГГУ, 2009. — С. 299-305.

[15] Поддубный В.В., Шевелев О.Г., Бормашов Д.А. Сравнение качества подходов к кластеризации текстов на основе гипергеометрического критерия // Вестник Томского государственного университета. 2006. № 293. — С. 120-125.

Using Natural Language Processing Tools in the System of Information Resources Monitoring by User Preferences

N.L. Soms, A.V. Dobrov, A.E. Dobrova

Modern systems of monitoring of news and social media resources in most cases do not allow the end user to customize the immediate procedures of data filtering, which would be able to give him the opportunity to keep track of publications on topics created by him, in view of the indicators of publications' and topics' coverage of the Internet audience. This article describes an approach to this problem based on the methods of natural language processing, developed by the team of developers of «Geline» LLC, when creating «SMiRTEO» software system.