Технологии потоковой обработки новостей

для персонализированной поисковой выдачи новостного контента*

Т.В. ИГНАТОВА, В.А. ИВИЧЕВ, ООО «Медиалогия», Москва. E-mail: tignatova@mlg.ru, vivichev@mlg.ru

В статье описывается реализованный подход к персонализированной выдаче новостей, который учитывает содержание и другие характеристики новостного сообщения, заметность и актуальность информационных новостных поводов, личные предпочтения пользователей, а также представляет результат персонализированной выдачи в структурированном, динамическом и удобном для пользователя виде. *Ключевые слова:* информационная система, персонализированная поисковая выдача, кластеризация, лингвистическая обработка, профиль пользовательских предпочтений.

Новостные сервисы становятся все более распространенными (Google News, Yahoo! News и др.), благодаря тому, что Интернет обеспечивает быстрый доступ к новостям, статьям из различных источников информации по всему миру. С гигантским ростом количества новостных статей ключевым вопросом качества становится предоставление пользователям наиболее интересной информации основе объективных характеристик новостного контента и персональных предпочтений пользователя. В описываемом в данной статье подходе этот вопрос решается путем автоматической персонализированной выдачи пользователю новостного контента.

Это непростая задача, по крайней мере, по трем причинам. Первая причина – необходимость обрабатывать непрерывно растущие огромные массивы неструктурированной информации в режиме реального времени. Вторая причина заключается в том, что новостные статьи публикуются непрерывно по различным поводам, и пользователю интересны значимые и актуальные конкретно для него информационные поводы. Третья причина – в том, что значимость и актуальность новостных сообщений быстро меняются (в отличие, скажем, от научных или искусствоведческих статей).

* Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации. УДК 004.031.42.

Компания Google предлагает новостную персонализацию на основе коллоборативной фильтрации[1].

В предлагаемой системе реализуется автоматическая персонализированная выдача, базирующаяся на трех главных принципах: отбор новостного контента для пользователя на основе характеристик новостных сообщений; кластерный анализ потока новостных сообщений, служащий для выявления наиболее заметных и свежих информационных поводов; представление новостей в порядке, формируемом на основе анализа статистических данных поведения пользователя в системе.

Формирование и применение основных характеристик новостных сообщений

Использование основных характеристик новостных сообщений для персонализированного их представления широко применяется в мире[2]. Но в разрабатываемой системе представлен, пожалуй, самый большой набор таких характеристик. Их формирование начинается с процесса сбора новостных данных, включающего следующие этапы.

1. Загрузка сообщений – получение материалов в электронном виде, копий выпусков печатных СМИ и др. выполняется для нескольких форматов представления новостных сообщений – файлы формата pdf, представляемые печатными изданиями, файлы других форматов и сообщения открытых источников, размещенные в Интернете. Из источников Интернета загрузка осуществляется по расписанию путем сканирования и загрузки в систему найденных новых (или измененных) сообщений.

2. Разбор, синтаксический анализ и предварительная обработка загруженных сообщений производят их преобразование в XML-файлы, разметку текста на заголовок и тело сообщения, очистку текстов загруженных сообщений от HTML-тегов, удаление сообщений-дубликатов (в том числе и плагиата).

3. Лингвистическая обработка загруженных и предварительно обработанных сообщений заключается в извлечении знаний из

1 *Das A.S., Datar M., Garg A., and Rajaram S.* Google news personalization: scalable online collaborative filtering // In WWW. – 2007. – P. 271–280.

2 *См. например: Lei Li, Dingding Wang, Tao Li, Knox D., Balaji Padmanabhan*. SCENE: a scalable two-stage personalized news recommendation system. – SIGIR. – 2011. – P. 125–134.

неструктурированных текстов, когда происходят синтаксический и морфологический анализ текста, выделение в текстах информационных объектов, их жанровая и тематическая классификация, выделение прямой и косвенной речи, ранжирование важности упоминания информационных объектов и определение характера упоминания, а также расчет заметности сообщений для выявленных информационных объектов.

Результаты лингвистической обработки потока новостей позволяют указать в профиле пользователя для отбора новостных сообщений для каждого информационного раздела пользователя следующие *характеристики отбора*: конкретные источники сообщений (печатные издания, интернет-СМИ, блоги, форумы, сети); автор сообщения; информационные объекты; тематические рубрики; жанр (например, новость, репортаж, аналитическая статья, интервью, беседа, опрос, ток-шоу и т.д.); наличие прямой и косвенной речи; степень важности упоминания (главная, второстепенная или иная роль); характер упоминания (позитивный, негативный, нейтральный).

Кластерный анализ

Данный анализ предназначен для выявления таких характеристик новостного потока, как значимость и актуальность информационных поводов. Под информационным поводом (кластером, событием) будем понимать группы сообщений, объединенные на основе смысловой и временной близости. Существуют различные методы кластеризации[3]. Перечислим наиболее распространенные.

– *Алгоритм К-средних*. Этот метод прост в реализации и обеспечивает быструю кластеризацию. Его недостатки – чувствительность к выбросам, медленная работа на больших базах данных, необходимость задания количества кластеров, невозможность применения алгоритма на данных с пересекающимися кластерами.
– *Fuzzy C–means* позволяет определить вероятность вхождения объекта в кластер. Его недостатки – вычислительная сложность,

необходимость задания количества кластеров; неопределенность с объектами, удаленными от центров всех кластеров.

– *CURE*. Данный метод выполняет кластеризацию на высоком уровне даже при наличии выбросов, выделяет кластеры сложной формы и различных размеров, обладает линейно зависимыми характеристиками. Недостатки – необходимость задания пороговых значений и количества кластеров.

[3] *Нейский И.М.* Классификация и сравнение методов кластеризации.

– *Гравитационный метод* выполняет автоматическое определение количества кластеров, обеспечивает устойчивую работу с любым количеством входных данных, которыми являются пороги расстояний между объектами и кластерами. Он применен в нашей перспективной информационной системе.

Для определения смысловой близости текстов сообщений ключевым является понятие *терм-вектора*. Он представляет собой совокупность всех слов текста в начальной форме (терминов) с указанием веса каждого слова (обычно используют до 200 слов, упорядоченных по убыванию веса). Вес термина определяется частотой его появления в сообщении, в общем индексе системы, положением в сообщении и т.д. Например, термин в названии текста используется с большим весом, чем слова из тела текста. Для определения веса термина используют статистическую меру TF-IDF[4], в соответствии с которой он пропорционален количеству употребления этого слова в тексте и обратно пропорционален частоте его употребления в других документах текста.

В результате построения терм-вектора текст каждого сообщения можно представить в виде вектора в многомерном пространстве, в котором термины определяют оси координат, веса терминов – координаты по этим осям, а смысловую близость текстов сообщений можно установить по мере сходства их терм-векторов.

Задача кластерного анализа – это определение групп текстов таким образом, чтобы расстояние между текстами в группе было минимальным. Для оценки близости текстов, в дополнение к смысловой близости, используется также параметр временной близости.

Для группировки сообщений по информационным поводам в разрабатываемой системе применяется метод потоковой гравитационной кластеризации. Его использование позволяет сделать процесс выделения групп информационных подмножеств более гибким, избежать необходимости разбиения уже сформированных кластеров. Использование временного показателя сообщения в качестве одного из параметров кластеризации позволяет получать наиболее актуальные кластеры сообщений.

[4] От англ. TF – term frequency, IDF – inverse document frequency.

Заголовок кластера выбирается из множества заголовков входящих в него сообщений. Выбор делается по величине рейтинга заголовка кластера, который рассчитывается на основании критериев количества цитирований заголовка, его длины и содержимого.

Ранжирование кластеров при потоковой обработке осуществляется по следующим критериям: размер кластера, общая скорость прироста сообщений, количество новых за последнее время (свежесть), суммарный вес источников сообщений кластера.

Наиболее информационно значимое (главное) сообщение кластера удовлетворяет следующим критериям (в порядке убывания):

1) связано с максимальным количеством сообщений в кластере;
2) имеет наибольшую смысловую близость сообщения к терм-вектору кластера;
3) имеет самое актуальное время публикации;
4) опубликовано в самом влиятельном источнике.

Для расчета индекса влиятельности источника используется показатель его цитируемости в других источниках. При этом учитывается влиятельность тех СМИ, которые ссылаются на данное издание. Чем чаще источник информации цитируют другие влиятельные издания, тем выше его индекс влиятельности.

В системе **для определения предпочтений пользователя** применяется статистика прочитанных пользователем сообщений. На ее основе определяются действительно предпочитаемые пользователем тематические категории. Также возможно определять все другие действительно предпочитаемые пользователем характеристики сообщения.

Автоматическая персонализированная выдача новостных сообщений

Автоматическая персонализированная выдача новостного потока в системе основана на принципе *непрерывного проак-тивного представления* новостного контента, предназначенного для конкретного пользователя.

Главное окно системы содержит индивидуальный контент, сформированный для конкретного пользователя и состоящий из описанных ниже областей.

Область «Перечень информационных разделов пользователя» отображает наименования информационных разделов,

определенных в профиле пользователя, и предоставляет возможность выбора любого из разделов в качестве текущего. Переход между разделами автоматический (они сменяют друг друга по мере появления новых информационных сообщений) или в ручном режиме. Отбор кластеров для раздела осуществляется следующим образом: главное сообщение кластера соответствует характеристикам, указанным в профиле пользователя. Отобранный кластер, представленный в информационном разделе, содержит только сообщения, соответствующие характеристикам профиля пользователя.

Область ключевых индикаторов раздела содержит основные индикаторы текущего состояния – количество негативных и позитивных сообщений, ТВ-сюжетов, сообщений печатных и интернет-СМИ, блогосферы, количество цитат.

При выборе ключевого индикатора пользователь получает список публикаций, по которым определен выбранный индикатор.

Лента событий предназначена для последовательного (автоматического) отображения актуальных информационных поводов текущего информационного раздела. Когда события текущего раздела заканчиваются, происходит автоматический переход на следующий раздел, кластеры которого последовательно отображаются в ленте событий.

Каждый кластер отображается в отдельном элементе визуализации (плашке) ленты событий. Плашка содержит дату события, его наименование, количество публикаций по событию. Если событие новое (самое раннее сообщение кластера появилось не более трех часов назад, и событие еще не прочитано), в плашке события отображается соответствующий значок. Если событие имеет позитивный характер (количество позитивных сообщений в нем преобладает), то плашка события в ленте новостей окрашивается в зеленый цвет, при негативном – в красный.

Всем событиям, отображаемым в ленте событий по какому-либо разделу, предшествует плашка с общей информацией по

событиям раздела, содержащая его ключевые индикаторы.

Эта плашка, так же как и плашки с событиями, перемещается по ленте новостей, всегда оставаясь самой первой в информационном разделе.

Для просмотра списка сообщений события открывается окно, в котором первым отображается наиболее информационно важное сообщение, далее идет список остальных сообщений по данному событию в порядке, автоматически определенном личными предпочтениями пользователя. Эти сообщения могут фильтроваться по типу источника и сортироваться по заметности или времени сообщения. При отображении сообщения по теме, кроме его текста, отображается также список источников, перепечатавших данное сообщение.

Область картины дня предназначена для отображения главных событий на сегодня. В данном блоке последовательно выводятся заголовки главных событий дня, независимо от указанных в профиле пользователя характеристик информационных разделов. Каждый заголовок события сопровождается списком ключевых индикаторов и возможностью просмотра всех сообщений любого события картины дня.

Таким образом, **автоматическая персонализированная выдача новостей в разрабатываемой системе достигает главной цели: оперативное представление объективно наиболее важной и актуальной информации, определенной базовыми настройками пользователя в удобном структурированном виде. При этом пользователю предоставляется и наиболее важная информация дня, зачастую не связанная впрямую с его базовыми предпочтениями, но позволяющая быть своевременно осведомленным об основных событиях в мире.**

Главными условиями достижения целей предлагаемого подхода являются полнота, оперативность и непрерывность сбора и обработки потока новостной информации. В противном случае основные характеристики новостного контента будут определены некорректно (основные информационные поводы, заметность сообщения, влиятельность открытых источников и т.п.).

Исследования точности выявления основных характеристик новостного потока подтверждают высокую эффективность реализованных методов и алгоритмов, а опросы пользователей системы показывают их достаточную удовлетворенность.

Дальнейшие работы будут направлены на реализацию автоматически формируемых тематических порталов в различных отраслях: экономика, политика, наука, культура.