Система автоматической обработки русскоязычных текстов

Повышенное внимание к проблеме автоматической обработки текстов на естественных языках и появление новых методик анализа текстов — одна из главных тенденций ИТ-индустрии. Однако на сегодняшний день нет единого подхода к решению задачи обобщения и визуализации больших объемов текстовых данных.

Ключевые слова: аналитика Больших Данных, анализ неструктурированных данных, неструктурированные данные
Keywords: Big Data analytics, analysis of unstructured data, text mining, unstructured data

Михаил Дубов, Борис Миркин, Артем Шаль

Современные программные средства анализа, обобщения и визуализации коллекций текстов варьируются от простых построителей облаков тегов типа Wordle до более интеллектуальных средств — например, SAS Text Miner, IBM Watson или IBM Content Analytics. Имеются специальные сайты, такие как newsanalytics.net, публикующие обзоры подобных программных продуктов и решаемых ими задач. Обычно речь идет о программах, позволяющих находить в Сети и собирать тексты по одной тематике, проводить их осмысленную каталогизацию, выявлять события, связанные с одним и тем же лицом, компанией или местом, и т. п. Среди программ, выполняющих более глубокую обработку текстов, можно упомянуть Grasshopper [1], формирующую краткий документ из коллекции текстов, в который отбираются «наиболее значимые» фразы из коллекции. Ряд приложений построены на технологии контент-анализа, заключающегося в определении для заранее заданного множества категорий, наиболее подходящих или нет для коллекции [2]. Вместе с тем подавляющее большинство программных средств способно обрабатывать тексты лишь на одном-двух естественных языках, прежде всего на английском, а систем, работающих с русским языком, чрезвычайно мало (OntosMiner, ABBYY Compreno, RCO for Oracle, «Гитика»), и решают они задачи ограниченного класса.

В основе системы LM Monitor (Latent Meaning Monitor) лежит идея использования графа референций, что в некотором смысле близко к контент-анализу, однако если в последнем речь идет об изучении распределения категорий, то в LM Monitor — об их связанных парах.

LM Monitor представляет собой систему для автоматического сбора и анализа веб-корпуса новостных статей из русскоязычных источников с последующим построением графа связей между ключевыми словосочетаниями, задаваемыми пользователем или выделяемыми из текстов автоматически. Получая на вход список ключевых словосочетаний, система осуществляет анализ степеней их релевантности каждому из текстов собранного корпуса, используя «нечеткий» метод [3, 4]. Полученные в результате оценки релевантности словосочетаний в текстах используются для определения взаимосвязи словосочетаний. Например, если 60% текстов, где степень релевантности ключевого словосочетания A достаточно велика, содержат также ключевое словосочетание B, то можно считать, что A «отсылает» к B. Вычислив все подобные связи между ключевыми словосочетаниями, можно визуализировать их в виде направленного графа «референций», то есть отсылок, между ключевыми словосочетаниями. Этот граф, а также результаты анализа его структуры и являются выходным результатом работы системы (рис. 1). В таком графе вершины аннотированы ключевыми словосочетаниями (с указанием

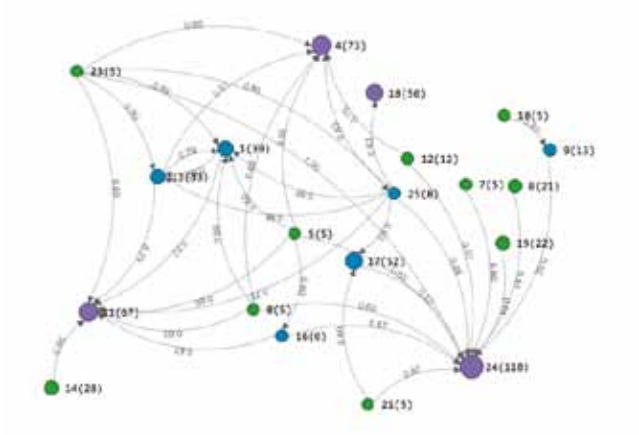

Рис. 1. Пример графа связей между ключевыми словосочетаниями

их поддержки, то есть числа релевантных текстов). Направленное ребро из вершины A в B обозначает тот факт, что ключевое словосочетание, соответствующее вершине A, отсылает к ключевому словосочетанию, соответствующему вершине B.

Предложенный подход к визуализации достаточно прост, компактен и информативен. При этом тематическая направленность информации полностью контролируется пользователем, задающим список ключевых словосочетаний, хранящийся в системе. Например, если пользователя интересует, как в газете «Известия» отображаются вопросы развития индустрии ИТ в России, то ему следует, используя LM Monitor, отобрать публикации за определенный период и сформировать список словосочетаний, отображающий понятия или действия, характерные для ИТ. Такой список может включать словосочетания «широкополосный доступ» и «электронное правительство», а в ответ LM Monitor выдаст граф референций между ключевыми словами с визуализацией элементов первичного анализа, как представлено на рис. 1, где зеленым помечены «входные», а сиреневым — «выходные» вершины; в скобках же проставлен уровень поддержки — количество статей, релевантных данному словосочетанию. В какой-то степени эта визуализация напоминает облака тегов — графы референций позволяют определить связи между ключевыми словосочетаниями в текстах и их направление, а также дают пользователю возможность контролировать список ключевых словосочетаний, участвующих в визуализации.

Граф (рис. 1) позволяет определить наиболее часто встречающиеся в анализируемых текстах ключевые словосочетания (такие как «электронное правительство» и «информационные технологии», соответствующие самым крупным вершинам графа 24 и 11), а также структуру связей между ними. Так, большое количество входящих ребер вершины 24 свидетельствует о том, что многие словосочетания встречаются в текстах именно в связи с «электронным правительством». Или же, например, ребро из вершины 21 («Телемедицина») в вершину 17 («Программист») можно интерпретировать как индикатор потребности первой в программных инженерах.

Эффективность системы LM Monitor непосредственно зависит от ее способности обеспечить: сбор больших объемов неструктурированных данных и их первичную обработку; интеграцию внешних инструментов для морфологического анализа текстов; использование нереляционных баз данных для хранения текстов и информации об отдельных словах; учет возможных языковых явлений (синонимии, морфологической омонимии и т. д.); проведение сопоставления ключевых словосочетаний с текстами для эффективного определения уровня релевантности; эргономичную визуализацию результатов. Поддержка данных функций определяет требования к архитектуре системы LM Monitor.

Система (рис. 2) представляет собой веб-сервер, реализованный на языке программирования Python с использованием фреймфорков Spring Python и CherryPy. У сервера есть планировщик, пополняющий хранилище текстов новыми данными, и аналитический модуль, работающий с собранным корпусом текстов и поддерживающий веб-интерфейс.

Планировщик дает пользователю возможность задавать регулярность, с которой система должна осуществлять выгрузку статей из сетевых источников. При выгрузке статей автоматически выполняются их индексация и размещение в хранилище текстов на сервере.

Аналитический модуль вычисляет, насколько определенные пользователем ключевые словосочетания релевантны текстам из собранного системой корпуса, после чего формирует граф референций между словосочетаниями на основе этого анализа. Данные процедуры реализованы с использованием специально разработанного метода аннотированного суффиксного дерева для «нечеткого» поиска строк. Алгоритмы анализа ключевых словосочетаний и построения графов референций, основанные на индексации корпуса с помощью аннотированных суффиксных деревьев, реализованы в отдельной открытой библиотеке EAST (Enhanced Annotated Suffix Trees) [5].

Взаимодействие пользователя с системой происходит через веб-клиент (View), включающий в себя средства для планирования выгрузки текстов, управления источниками текстов и просмотра содержимого базы данных. Администратор

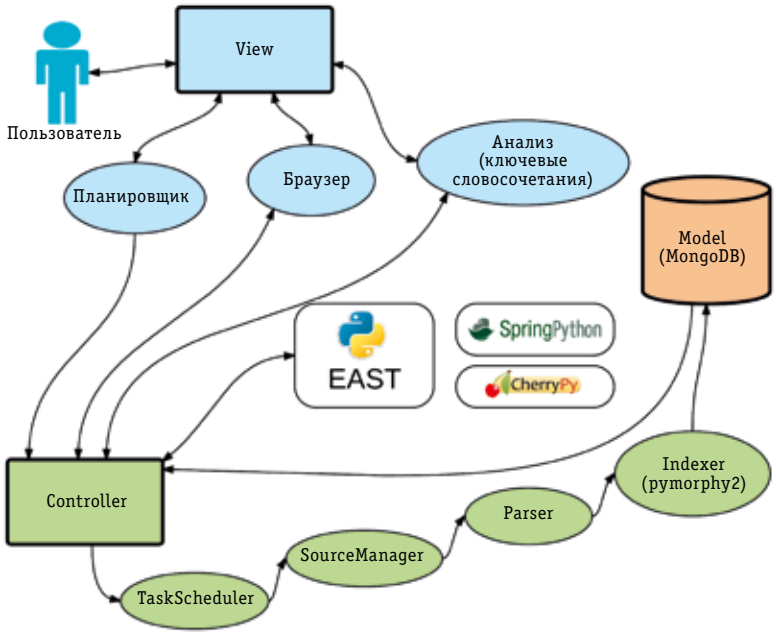

Рис. 2. Архитектура системы LM Monitor

имеет доступ к детальной информации о базе текстов (статистика по количеству статей, уникальных употреблений слов и т. д.) и может выполнять визуализацию текстов корпуса посредством графа заданных словосочетаний (рис. 3).

Собственно тексты сохраняются средствами нереляционной документоориентированной СУБД MongoDB, которая позволяет хранить тексты и их разметку в удобном для последующих запросов виде и дает возможность задействовать встроенную в нее функцию полнотекстового поиска по базе текстов, что как нельзя лучше подходит под задачи, решаемые LM Monitor.

Особенностью системы является то, что тексты хранятся не только в исходном виде, но и вместе с морфологической разметкой (снабжением отдельных слов информацией об их части речи и грамматических характеристиках). Сам процесс выгрузки и предобработки текстов из заданных источников производится в подсистеме сбора и анализа текстов Controller. В состав этой подсистемы входят: планировщик задач TaskScheduler — выгрузка текстов и сохранение их в хранилище; модуль SourceManager — поиск статей на сайтах соответствующих издательств, выгрузка неструктурированных текстов статей и их разбор; парсеры, используемые компонентом SourceManager для синтаксического анализа html-разметки соответствующих веб-страниц или разбора RSS-лент. Работа парсеров основана на информации о специфике разметки страниц той или иной газеты — зная особенности разметки страниц в конкретном издании, можно из их неструктурированного содержимого получить не только текст статьи, но и такие семантически значимые элементы, как заголовок, подзаголовок, информация об авторе и аннотация. Подобная структура статей учитывается при их сохранении в базе и в алгоритме вычисления оценок релевантности.

Визуализация состоит в рендеринге графов словосочетаний на стороне клиента (генерации страницы в браузере) с помощью JavaScript-библиотеки d3. Основная функция модуля визуализации — сделать работу пользователя с графом удобной и быстрой. Корни (узлы, имеющие только исходящие дуги), листья (узлы, имеющие только входящие дуги) и остальные узлы помечаются цветом, а изолированные вершины не отображаются. Размер узла на экране зависит от уровня его поддержки — количества статей, релевантных вершине.

Стрелкам дуг в графе приписан вес — численный параметр доверия, показывающий степень достоверности соответствующей связи. При попадании курсора на изображение узла показывается соответствующее словосочетание и подсвечиваются инцидентные ребра.

***

Система LM Monitor готова к применению и может использоваться, например, для анализа периодики. Собранный корпус русскоязычных статей и публикаций — RuNeWC (Russian Newspaper Web Corpus) на данный момент содержит 4 тыс. статей по теме «Экономика», а всего в корпусе содержится более 100 тыс. уникальных словоупотреблений. Наибольший интерес система представляет как средство выявления отношений (зачастую скрытых) между ключевыми понятиями той или иной предметной области. Предложенный в системе подход использует относительно простую математическую модель вычисления частот встречаемости и оценок релевантности ключевых словосочетаний, однако он позволяет получить гораздо больше информации о корпусе текстов и о соответствующей предметной области, чем при использовании облаков тегов. ■

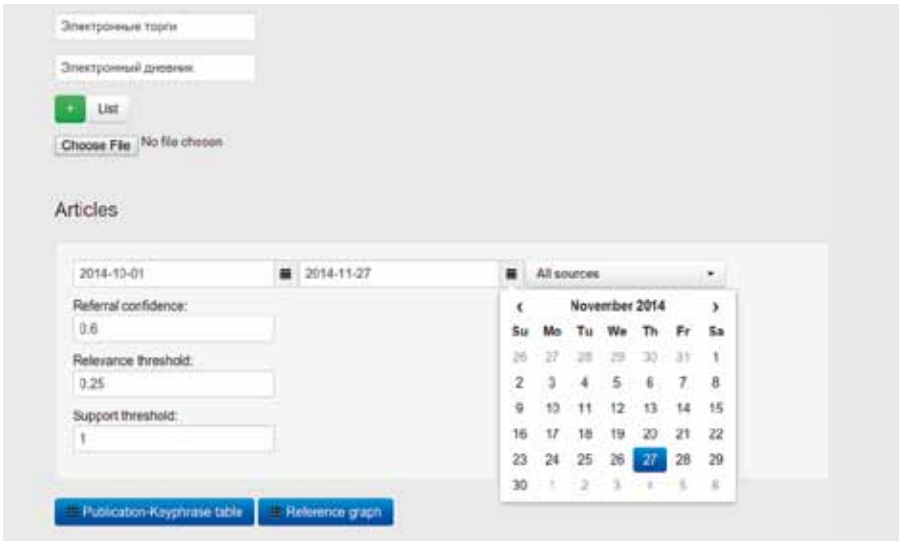

Рис. 3. Панель управления системы LM Monitor

Системы автоматической обработки текстов

Многообразие систем автоматической обработки неструктурированных текстов сегодня вызывает необходимость их систематизации и классификации с целью упрощения выбора решения, наиболее адекватного для конкретной задачи.

Дмитрий Ильвовский, Екатерина Черняк

«Открытые системы», № 01, 2014

ЛИТЕРАТУРА

1. Zhu, X., Goldberg, A. B., Van Gael, J., & Andrzejewski, D. Improving Diversity in Ranking using Absorbing Random Walks. In HLT-NAACL. April, 2007. P. 97–104.

2. Krippendorff, K. Content analysis: An introduction to its methodology: Sage Publications, Inc. 2004.

3. Миркин Б. Г., Черняк Е. Л., Чугунова О. Н. Метод аннотированного суффиксного дерева для оценки степени вхождения строк в текстовые документы // Бизнес-информатика. — 2012. — Т. 3, № 21. — С. 31–41.

4. Дмитрий Ильвовский, Екатерина Черняк. Системы автоматической обработки текстов // Открытые системы. СУБД. — 2014. — № 1. — С. 51–53. URL: http://www.osp.ru/os/2014/01/13039687 (дата обращения: 15.12.2014).

5. EAST — Text analysis library based on the annotated suffix tree method. URL: https://pypi.python.org/pypi/EAST (дата обращения: 15.12.2014).

Михаил Дубов (msdubov@gmail.com) — студент магистратуры факультета компьютерных наук, Борис Миркин (bmirkin@hse.ru) — профессор департамента анализа данных и искусственного интеллекта факультета компьютерных наук, НИУ ВШЭ (Москва). Артем Шаль (artiom.shal@gmail.com) — эксперт-разработчик, компания «Полимедиа» (Москва). *Работа выполняется в рамках программы «Научный фонд НИУ ВШЭ» в 2013–14 гг., грант № 13-05-0047.*