УДК 004.415.23 DOI 10.25559/SITITO.2017.3.437

Отраднов К.К., Алёшкин А.С., Жуков Д.О.

Московский технологический университет (МИРЭА), г. Москва, Россия

АРХИТЕКТУРА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ КОМПЛЕКСОВ ОБРАБОТКИ НОВОСТНЫХ ТЕКСТОВ ДЛЯ РЕШЕНИЯ ЗАДАЧ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА

Аннотация

В статье предложена архитектура программного обеспечения обработки коллекций новостных текстовых сообщений, а также соответствующий состав и структура информационной системы. Система представляет собой несколько ступеней получения и обработки информации, функционирующих на базе гибридного вычислительного кластера. Каждый этап получения, обработки и сохранения информации в обобщенной информационной системе представлен микросервисом как отдельной программной единицей. При этом подчёркнута возможность использования различных стеков технологий для каждого микросервиса так, что правильно подобранные специализированные решения повышают эффективность и качество результата, а недостатки классической микросервисной архитектуры нивелируются внутренней неоднородностью микросервисов, выраженной в виде гибкой модуляризации. Суть предлагаемого подхода заключается в применении принципа конвейерного параллелизма на базе микросервисной архитектуры с динамическими границами сервисов.

Ключевые слова

Архитектура программного обеспечения; микросервисная архитектура; микросервисы; модуляризация; конвейерный параллелизм.

Otradnov K.K., Alyoshkin A.S., Zhukov D.O.

Moscow Technological University (MIREA), Moscow, Russia

SOFTWARE ARCHITECTURE OF HIGH-PRODUCTION COMPLEXES FOR TEXT OF THE NEWS PROCESSING FOR SOLVING THE PROBLEMS OF DATA MINING

Abstract

The article proposes the architecture of the software for processing collections of news text messages, as well as the corresponding composition and structure of the information system. The system is a few steps of obtaining and processing information, functioning on the basis of a hybrid computing cluster. Each stage of receiving, processing and storing information in a generalized information system is represented by a microservice as a separate program unit. At the same time, the possibility of using different technology stacks for each service is emphasized so that properly selected specialized solutions increase the efficiency and quality of the result, and the shortcomings of the classical micro-service architecture are offset by the internal heterogeneity of the micro-services expressed in the form of flexible modularization. The essence of the proposed approach is the application of the principle of pipeline concurrency based on a microservice architecture with dynamic service boundaries.

Keywords

Software architecture; microservice architecture; microservices; modularization; pipeline concurrency.

Введение

Построение программного обеспечения для анализа неструктурированных или слабоструктурированных естественно-языковых текстовых данных является актуальной задачей с точки зрения создания инструментария поиска скрытых закономерностей в больших объемах

информации, реализации прогностических моделей различных процессов (в том числе социальных и экономических) и извлечения знаний. При анализе больших объемов неструктурированных или слабо структурированных текстовых данных и поиске в них скрытых закономерностей возникает проблема отбора признаков существенной для анализа информации. С одной стороны, необходимо, чтобы выполнение поиска происходило за приемлемое время, а с другой стороны необходимо, чтобы существенные данные не были потеряны. Отбор большого количества разнообразных признаков приводит к увеличению вычислительной сложности модели, и, как следствие, к необходимости использования значительных вычислительных ресурсов и затрат времени. Однако отбрасывание признаков, кажущихся несущественными или проявляющихся на уровне шума, может привести к потери значимой для поиска скрытых закономерностей информации. Таким образом, априори является неизвестным, какая часть имеющихся данных может понадобиться, а какая – нет. Определение связи шумовых признаков с целевой переменной решаемой задаче всегда является проблемой.

Следует отметить, что достаточно часто используемый для ускорения вычислений на практике метод, когда новые признаки формируются на основе старых, в данном случае является ненадежным, поскольку скрытые закономерности могут быть не линейным образом связаны с имеющимися данными и не являться их простой суперпозицией. Вместе с тем можно отметить, что формирование новых признаков на основе старых позволяет уменьшить их число для использования в моделях информационного поиска и приводит к ускорению выполнения модели.

Существуют задачи, в которых прогнозы нужно строить очень быстро, например, выдача рекомендаций товаров на сайте интернет-магазина. В этом случае необходимо, чтобы модель была очень быстрой, и один из подходов к ускорению модели – это отбор признаков, которых достаточно, чтобы прогнозы были хорошими. В данном случае осуществляться обработка значительного числа текстовых документов, однако сами по себе документы имеют небольшой размер по числу слов и являются однотипными (описание товаров с их характеристиками), что существенно облегчает отбор признаков для модели информационного поиска.

При работе с большими объемами неструктурированных или слабо структурированных естественно-языковых текстовых данных возникает необходимость осуществлять операции информационного поиска (в том числе поиска скрытых закономерностей), извлечения знаний и классификации, что требует создания соответствующих моделей и инструментов для автоматизации этих процессов. Следует отметить, что в данном случае нет физически измеряемых величин, с которыми можно проводить математические операции и строить формальные модели. [1 – 6].

Статистическая семантическая гипотеза позволяет формализовывать представление текстовых документов в виде векторной модели. Главная идея векторной модели семантики (vector space model, VSM) заключается в возможности представления каждого текстового документа в качестве точки в многомерном пространстве (вектора в векторном пространстве, размерности R^M). Близко лежащие друг к другу точки соответствуют семантически схожим документам.

Перед векторизацией текстовых документов необходимо проведение ряда предварительных процедур их лингвистической обработки, таких как разметка текста (линеаризация), нормализация текста, комментирование.

Линеаризация необходима для распознавания и извлечения слов – терминов и составных терминов (n – грамм) из исходного текста. При разметке учитывается пунктуация и переносы слов, и кроме того игнорируются стоп – слова (малоинформативные, но часто встречающиеся в текстах слова, например, предлоги, союзы, причастия, междометия и т.д.).

При нормализации проводится удаление из исходного текста грамматической информации (падежи, числа, глагольные виды и времена, залоги причастий, род и так далее). Слова в разных регистрах, а также их аббревиатуры приводятся к одной форме. Нормализация требует морфологического анализа, распознающего части речи с учетом контекста и многочисленных правил согласования (без него нормализация будет давать значительное количество ошибочных результатов). В ряде случаев нормализация может быть заменена стеммингом или лемматизацией. В процессе стемминга проводится нахождение основы слова, которая необязательно совпадает с его морфологическим корнем. Стемминг важен при небольшом наборе текстов и флективных языков (например, русского). Для реализации стемминга может быть использован алгоритм Портера. Основная проблема, возникающая при

использовании стеммера – это обработка слов, которые при образовании разных грамматических форм меняют не только окончание, но и основу слова. Лемматизацию можно рассматривать как упрощенную процедуру нормализации при которой происходит упрощенный анализ без учета контекста. При стемминге и лемматизации глубина преобразования текста в меньше, чем при нормализации.

Комментирование в частности необходимо потому что, одно и то же слово в тексте может быть употреблено в различных значениях. Комментирования заключаются в добавлении в документ метаданных с информацией о части речи, о наличии у слова других значений и синтаксический разбор предложения. Синтаксическое комментирование включает сегментацию и разметку частей речи. Это используется для устранения неоднозначности в сокращениях и поиск ассоциаций среди ключевых слов. Комментирование полезно для измерения семантической схожести слов и понятий (для моделей, основанных на матрице слово-контекст).

После разметки, нормализации и комментирования текстов каждый документ коллекции представляется набором терминов, которые в него входят, и их частотами вхождения, что позволяет создать векторную модель документа.

Подготовка текстов и их векторизация являются технически и вычислительно трудоемкими этапами информационного анализа. И их основная задача – формализация представления документов при сохранении их смыслового содержания.

В рамках данных исследований создан и развивается инструментарий реализации технологий прогнозирования новостных событий.

Архитектура программного обеспечения обработки новостных текстов

Разработанное программное обеспечение представляет собой микросервисную информационную систему, построенную на основе модульной модели предметной области, распределённой по сервисам. Разработка ведётся с применением предметно-ориентированного подхода к проектированию распределённых систем.

Информационная система функционирует на базе гибридного вычислительного кластера со следующими характеристиками:

- 7 TFlops расчетная вычислительная мощность (с учетом GPU);
- 9 вычислительных узлов;
- 1 управляющий узел;
- 80 вычислительных ядер CPU;
- 4 NVidia Tesla K20x;
- 10752 вычислительных ядра CUDA;
- скоростная шина обмена данными Infiniband 40Gb/s;
- lan 1000Mb/s.

Для построения системы параллельной обработки данных учитывались принципы конвейерного параллелизма. Система представляет собой несколько ступеней получения и обработки информации. Каждый этап получения, обработки и сохранения информации в обобщенной информационной системе представлен отдельной программной единицей – микросервисом.

На данный момент развёрнуто и функционирует 9 микросервисов в составе информационной системы (рисунок 1).

Микросервис 1: управление информационными ресурсами. Каждый ресурс имеет свою собственную структуру и соответствующие правила выделения содержательной части доступной информации, особенности доступа, авторизации и пр. На основе этого для каждого ресурса строится шаблон обработки. Данные о ресурсах и их доступность регулярно проверяются, при изменениях шаблон ресурса переводится в режим редактирования, высылаются уведомления. На данный момент в качестве ресурсов рассматриваются новостные порталы, справочные и аналитические сайты, прочие ресурсы с информацией нужного типа, содержащие html и pdf файлы. Посредством расширения перечня информационных ресурсов осуществляется постепенное увеличение объёмов информации, хранящейся в информационной системе.

Микросервис 2: сбор документов. Осуществляются мониторинг информационных ресурсов и загрузка файлов. В первую очередь, в зависимости от выбранного источника информации применяются правила и процедуры для получения нужных файлов и соответствующей метаинформации: источник, привязка к шкале времени, уникальный идентификатор (ссылка, URL) и пр. Затем загружаются нужные данные и осуществляется первичная обработка информации, куда входит удаление ненужных данных (например, пользовательских скриптов, визуальных стилей оформления и т. д.), приведение данных к нужному формату (например, формирование корректной xml-структуры документов, правка ссылок на изображения и т. д.), выделение метаданных. В результате получаем набор метаданных и файлов, доступных для дальнейшей обработки.

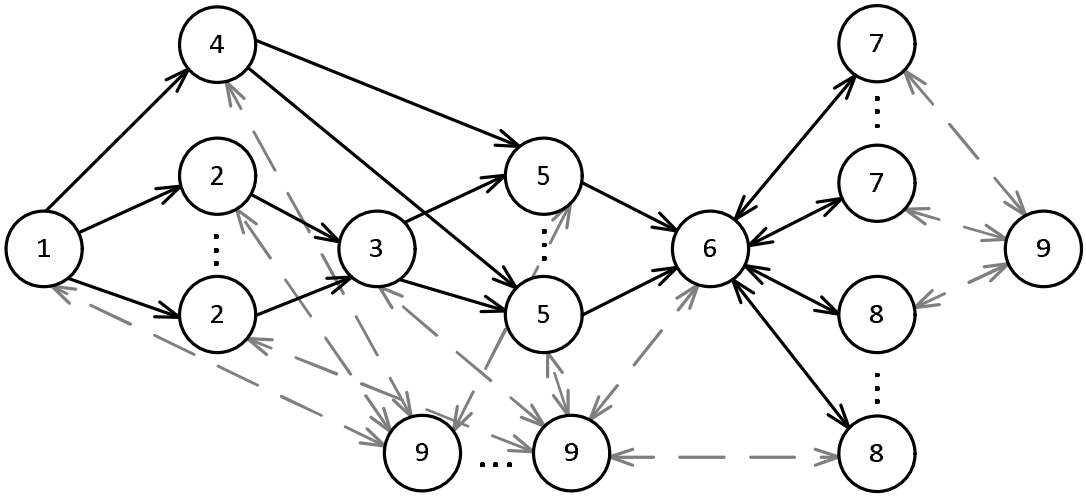

Рисунок 1 – Архитектура информационной системы обработки новостных текстов

Микросервис 3: ведение хранилища документов. Файловое хранилище со всем сопутствующим функционалом. Файлы из хранилища используются неоднократно при проведении экспериментов с целью увеличить вычислительную эффективность и качество обработки текстов.

Микросервис 4: ведение словарей. Для векторного представления текстов необходим словарь терминов (модель «bag of terms»), который может создаваться в процессе обработки документов или адаптироваться готовый. Следует отметить, что использование для анализа заложенного в текст смысла, словаря, состоящего только из единичных слов, является недостаточным. Для повышения точности анализа текстов (более глубокого сохранения смысла) необходимо добавлять в словарь словосочетания (би-, три -, три-, тетра и т.д. граммы), более точно отображающие смысл описываемого объекта. Однако увеличение размера словаря приводит к увеличению времени обработки текстов, информационного анализа и поиска. На данной стадии необходимо сформировать словари, по которым будет транслироваться информация из полученных статей в векторное представление информации. Данная стадия может быть выполнена как отдельная и независимая компонента системы, а также может опираться на собранную текстовую информацию (статьи) для создания словаря в процессе работы («на лету»).

Микросервис 5: векторизация. Собранные ранее тексты преобразуются в векторную форму, проводится нормализация текста: удаляются незначащие для поставленных задач частей речи (например, предлогов), приводятся слова в тексте к базовому словарю (использование базовых словоформ), а также организуется векторное представление текста (в соответствии с выбранным словарем).

Микросервис 6: ведение хранилища данных. Полученная ранее информация попадает в общее хранилище данных информационной системы. Следует отметить, что обычно микросервисный подход требует соответствующего разделения данных, но в данном случае необходимо учесть, что функциональное разделение базы данных сильно усложняет поддержание целостности данных, а также реализацию многих запросов.

Микросервис 7: кластеризация данных. Кластеризацию документов можно определить, как процесс обнаружения в их коллекции групп, имеющих примерно одинаковое смысловое значение (однородных классов в произвольной предметной области). Обратим внимание на две проблемы. Во-первых, заранее неизвестны кластеры, которые можно отнести к истинным. Во-вторых, неизвестно число истинных кластеров, на которые предстоит разделить коллекцию. Большинство используемых способов кластеризации базируются на определении расстояний между векторами и между кластерами. Вектора относят к определённым кластерам таким образом, чтобы расстояние между векторами, принадлежащими одному кластеру, было минимальным. Построение моделей кластеризации текстовых документов основывается на методологии отбора существенных признаков для решения поставленной задачи информационного поиска. Не существует модели кластеризации, которая одновременно имела бы очень небольшое время выполнения, обладала бы абсолютной точностью и обрабатывала бы большие объемы неструктурированных или слабоструктурированных данных. Кроме того, следует отметить, что набор признаков необходимых для анализа скрытых закономерностей в больших объемах данных априори является неизвестным, поэтому при построении модели кластеризации и информационного поиска необходимо основываться на минимальном числе

допущений, при максимальном сохранении разнообразных признаков, используемых для кластеризации. Однако следует учесть, что время выполнения модели при этом должны оставаться приемлемым, а реализация вычислений не требовать значительных аппаратных ресурсов.

Микросервис 8: анализ спектров информационных процессов. После формирования спектров информационных процессов проводится их анализ на основе: теории почти- периодических функций, теории вейвлетов, теории самоподобия Хёрста (частный случай фрактальности), теории фрактальных точечных случайных процессов, теории фрактального броуновского изменения спектров (частоты) с течением времени.

Микросервис 9: графический интерфейс пользователя. Каждый микросервис может реализовать графический интерфейс посредством передачи соответствующих стандартизированных структур данных.

Внутренняя структура микросервисов также неоднородна. Каждый микросервис состоит из стандартизированных модулей. Мы разделяем понятия микросервиса и модуля. Применение модуля ограничивается конструкциями используемого языка программирования в каждом микросервисе, а также внутрипроцессными взаимодействиями: взаимодействие между модулями осуществляется в рамках одного процесса. Большинство микросервисов состоит из следующего набора модулей:

- модуль A: модуль предметной области, реализует функционал согласно соответствующему фрагменту предметной области;
- модуль B: GUI-сопряжение, служит для формирования данных графического интерфейса пользователя;
- модуль C: локальная база данных. Каждый микросервис имеет свою локальную базу данных (используется SQLite), тем самым реализуется принцип децентрализованного управления данными;
- модуль D: самодиагностика. Микросервисные архитектуры отличаются на порядок возрастающим количеством и сложностью взаимодействий, при этом многокомпонентная распределённая среда затрудняет обнаружение ошибок и сбоев. Данных модуль призван нивелировать эти недостатки архитектуры;
- модуль E: коммуникации, обеспечивается транспорт и эффективный обмен сообщениями между микросервисами;
- модуль F: безопасность, осуществляется управление доступом ко внутренней информации, шифрование и реализация предустановленных политик безопасности;
- модуль G: документация, организация документации как для пользователей, так и разработчиков;
- модуль H: командный интерфейс пользователя.

Заключение

Разработанное программное обеспечение на базе представленной архитектуры активно развивается и совершенствуется. Это оказалось возможным не в последнюю очередь благодаря тому, что микросервисы позволяют вести параллельную разработку и апробацию решений. Отсутствие свойственных монолитным системам ограничений позволяет выбирать лучшие и наиболее удобные решения, а также экспериментировать с новыми технологиями. Существует возможность использования различных стеков технологий для каждого микросервиса так, что правильно подобные специализированные решения повышают эффективность и качество результата. Микросервисы и с точки зрения аппаратных ресурсов имеют преимущества, так как естественным образом подходят для процессов параллельной конвейерной обработки данных.

Недостатками представленной архитектуры является необходимость дополнительного управления, вопросы безопасности и прочие моменты, связанные с необходимостью построения инфраструктуры и микросервисной среды, что для монолитного приложения не столь критично.

Особенностями приведённого решения является ориентация на потоковую обработку данных, а также внутренняя неоднородность микросервисов, которая делает границы микросервисов динамическими, но в то же время жёстко очерченными в каждый момент времени. Это позволяет пользоваться преимуществами как микросервисных, так и монолитных архитектур в зависимости от текущих приоритетов.

На данный момент выборочная статистика работы системы следующая:

Проработано ресурсов – 5 шт.

Количество загруженных новостных текстов – 1 078 779 шт.

Объем занятого дискового пространства (включая изображения) – 62,1 ГБ (66 770 733 975 байт).

Объем занятого дискового пространства (исключая изображения) – 9,03 ГБ (9 700 293 510 байт).

Количество выделенных N-грамм – 87 144 027 шт.

Разработка информационной системы продолжается.

Благодарности

Работа выполнена за счет финансирования Министерством образования и науки Российской Федерации конкурсной части государственных заданий высшим учебным заведениям и научным организациям по выполнению инициативных научных проектов, номер проекта 28.2635.2017/ПЧ, наименование «Разработка моделей стохастической самоорганизации слабоструктурированной информации и реализации памяти при прогнозировании новостных событий на основе массивов естественно-языковых текстов».

Acknowledgments

The work is executed due to financing by the Ministry of Education and Science of the Russian Federation as the competitive part of government tasks of higher education and scientific organizations for the implementation the initiative scientific projects; the number of the project – 28.2635.2017 / IF named «Development of models of stochastic self-semistructured information and implementation of memory in predicting the news events array-based natural language texts».

Литература

1. Lesko, S.A., Zhukov, D.O. Trends, self-similarity, and forecasting of news events in the information domain, its structure and director. 2015 International Conference on Big Data Intelligence and Computing, Chengdu, China. -- 2015.
2. Zhukov, D.O., Lesko, S.A. Stochastic self-organization of poorly structured data and memory realization in an information domain when designing news events forecasting models. The 2nd IEEE International Conference on Big Data Intelligence and Computing, Auckland, New Zealand. -- 2016.
3. Sigov, A., Zhukov, D., Novikova, O. Modelling of memory realization processes and the implementation of information self-organization in forecasting the new's events using arrays of natural language texts. 1st International Scientific Conference Convergent Cognitive Information Technologies, Moscow, Russian Federation. -- 2016.
4. S. Sigov, D.A. Akimov, D.O. Zhukov, E.G. Andrianova, V. E. Sachkov, V.K. Raev. Psycholinguistic analysis of text messages in Russian based on their phono semantic statistical characteristics. Informatics and applications. 2017 volume 11 issue 3, pp. 77 -86.
5. D.O. Zhukov, A.M. Zamyshlyaev, O.A. Novikova. Model of Forecasting the Social News Events on the Basis of Stochastic Dynamics Methods. ITM Web of Conferences 10, 02009 (2017) 2017 Seminar on Systems Analysis, DOI: 10.1051/itmconf/20171002009.
6. Zhukov D.O., Novikova O.A., Otradnov K.K. Methods of analysis of news events in the information space based on the use of almost – periodic functions, wavelet transforms and Hurst's self-similarity. Proceeding The 7th International Conference on Information Communication and Management ICICM'17, August 28-30, 2017, Moscow, Russian Federation, ACM ISBN 978-1-4503-5279-6/17/08.

References

1. Lesko, S.A., Zhukov, D.O. Trends, self-similarity, and forecasting of news events in the information domain, its structure and director. 2015 International Conference on Big Data Intelligence and Computing, Chengdu, China. -- 2015.
2. Zhukov, D.O., Lesko, S.A. Stochastic self-organization of poorly structured data and memory realization in an information domain when designing news events forecasting models. The 2nd IEEE International Conference on Big Data Intelligence and Computing, Auckland, New Zealand. -- 2016.
3. Sigov, A., Zhukov, D., Novikova, O. Modelling of memory realization processes and the implementation of information self-organization in forecasting the new's events using arrays of natural language texts. 1st International Scientific Conference Convergent Cognitive Information Technologies, Moscow, Russian Federation. -- 2016.
4. A. S. Sigov, D.A. Akimov, D.O. Zhukov, E.G. Andrianova, V. E. Sachkov, V.K. Raev. Psycholinguistic analysis of text messages in Russian based on their phono semantic statistical characteristics. Informatics and applications. 2017 volume 11 issue 3, pp. 77 -86.
5. D.O. Zhukov, A.M. Zamyshlyaev, O.A. Novikova. Model of Forecasting the Social News Events on the Basis of Stochastic Dynamics Methods. ITM Web of Conferences 10, 02009 (2017) 2017 Seminar on Systems Analysis, DOI: 10.1051/itmconf/20171002009.
6. Zhukov D.O., Novikova O.A., Otradnov K.K. Methods of analysis of news events in the information space based on the use of almost – periodic functions, wavelet transforms and Hurst's self-similarity. Proceeding The 7th International Conference on Information Communication and Management ICICM'17, August 28-30, 2017, Moscow, Russian Federation, ACM ISBN 978-1-4503-5279-6/17/08

Поступила: 28.09.2016

Об авторах:

Отраднов Константин Константинович, соискатель, старший преподаватель кафедры автоматизированных систем управления Института комплексной безопасности и специального приборостроения, Московский технологический университет (МИРЭА), const.otradnov@yandex.ru

Алёшкин Антон Сергеевич, кандидат технических наук, доцент кафедры автоматизированных систем управления Института комплексной безопасности и специального приборостроения, Московский технологический университет (МИРЭА), antony@testor.ru

Жуков Дмитрий Олегович, доктор технических наук, профессор, заместитель директора по научной работе института комплексной безопасности и специального приборостроения, Московский технологический университет (МИРЭА), zhukovdm@yandex.ru

Note on the authors:

Otradnov Konstantin K., Applicant, Senior Lecturer, Moscow Technological University (MIREA), const.otradnov@yandex.ru

Alyoshkin Anton S., Candidate of Technical Science, Associate Professor of Department of automated management systems, Moscow Technological University (MIREA), antony@testor.ru

Zhukov Dmitry O., Doctor of Technical Science, Professor, Deputy Director on scientific work, Moscow Technological University (MIREA), zhukovdm@yandex.ru