119270, Москва, Лужнецкая наб., д. 6,
стр.1, офис 214, ООО «ЭР СИ О»
Тел./факс: (495) 287-98-87
E-mail: info@rco.ru
http://www.rco.ru

Руководство администратора

RCO News Clustering Engine

Москва, 2020

В содержание данного документа могут быть внесены изменения без предварительного уведомления. Названия организаций, имена и даты, используемые в качестве примеров, являются вымышленными, если не оговорено обратное.

Содержание

Общие сведения

Назначение

Программное обеспечение (ПО) RCO News Clustering Engine (далее RCO RNE, Агрегатор новостной ленты) предназначено для связывания сообщений, описывающих одни и те же события, в кластеры – сюжеты.

Агрегатор новостной ленты использует алгоритмы разбора текста, взаимного взвешивания документов, кластеризации документов. При построении кластеров Агрегатор на каждой итерации рассматривает временной интервал, называемый окном кластеризации. Итерации повторяются со сдвигом окна кластеризации на заданный временной отрезок, называемый шагом кластеризации.

Агрегатор новостной ленты в штатном режиме предназначен для работы с СУБД (дополнительно предусмотрен режим отладки, в котором обращения к СУБД заменяются на работу с файловой системой). Входной информацией агрегатора являются документы новостной ленты, хранящиеся в базе данных новостей заданного формата. Агрегатор новостной ленты сохраняет результат кластеризации документов в базе данных новостей в специально разработанных таблицах. Результатом кластеризации является набор кластеров. Каждый кластер имеет набор документов, собственно образующих кластер, и набор терминов, характеризующих кластер. Указанные документы и термины имеют свой вес в кластере.

Помимо штатного режима, в котором происходит обработка новых документов, Агрегатор новостной ленты имеет также ретроспективный режим, в котором за явно указанный интервал проводится кластеризация документов с заданными окном и шагом кластеризации. В базе данных сохраняются результаты всех итераций кластеризации.

Внутренняя структура агрегатора допускает достаточно легкую замену одного или нескольких алгоритмов агрегатора на этапе компиляции и сборки ПО.

Требования к квалификации персонала, необходимом для поддержки ПО

Для обслуживания RCO NCE требуется следующий персонал:

- Системный администратор RCO NCE:
  - Установка и запуск сервиса RCO NCE;
  - Контроль работоспособности сервиса RCO NCE.
- Администратор базы данных:
  - Контроль работоспособности БД.

Системный администратор RCO NCE должен иметь следующие навыки:

- администрирования информационных систем;
- технического обслуживания средств вычислительной техники, на которых устанавливается ПО RCO NCE;
- работы с операционными системами Linux/Windows.

Администратор базы данных RCO NCE должен иметь следующие навыки:

- администрирования информационных систем;
- работы с операционными системами Linux/Windows;

- работы с базой данных (PostgreSQL).

Системные требования

Требования сервиса зависят от дневного потока новых документов, размера окна кластеризации и числа порождаемых сервисом NCE параллельных процессов разбора документов RCO Fact Extractor.

Типичные требования

Для нагрузки до 10тыс документов в час, обработки в 2 RCO FX, окне 4 дня:

- Четырёх ядерный процессор 3.5GHz с SMT;
- Оперативная память 8GB (+500MB на каждый дополнительный обработчик RCO FX);
- Дисковое пространство 1GB;
- Операционная система Windows/Linux x64 бит.

Минимальные требования:

- Двухъядерный процессор 3.0GHz;
- Оперативная память 4GB;
- Дисковое пространство 1GB;
- Операционная система Windows/Linux x64 бит.

Подготовка к работе

Состав дистрибутива

AgregatorHDD.exe - исполняемый файл программы.

RCO_NCE_Dev_Guide.pdf – документация.

AgregatorDB.cmd и AgregatorHDD.cmd - примеры вызова программы в режиме работы с базой данных и в режиме работы с файловой системой.

service_DB_install.cmd, service_DB_uninstall.cmd, service_HDD_install.cmd,

service_HDD_uninstall.cmd - установка и удаление программы в качестве сервиса

cb.ini и RCOAggregator.ini - файлы с параметрами программы - описаны в документации

ServiceLauncher.exe - программа для установки и удаления программы в качестве сервиса

FXModuleWrapper.exe и ModuleList.xml - программа и файл настройки модуля извлечения терминов из текста новостных сообщений из папки **TermExtractorApp**

word_...txt, world_...txt - результаты промежуточных этапов работы программы - используются для отладки и быстрого продолжения итерации

clusters_added_...txt - файлы с результатом работы программы в виде таблицы с символом табуляции в качестве разделителя колонок и следующими названиями колонок: cluster_id – номер кластера, doc_id – идентификатор документа, входящего в кластер, weight – вес документа в кластере

AgregatorHDD_...log - файлы журнала запуска данной программы

WrapperLog_...log - файлы журнала запуска подпрограммы извлечения терминов из текста новостных сообщений

zlibwapi.dll - свободная кроссплатформенная библиотека для сжатия данных zlib, используется для работы со сжатыми входными данным

Каталог **by_date** - входные данные - новостные сообщения в специальном формате, описанном в разделе «Запуск программы без использования БД» документации RCO_NCE_Dev_Guide.pdf, сжатые архиватором zlib (для режима работы с файловой системы)

Каталог **TermExtractorApp** - библиотека извлечения терминов из текста новостных сообщений и её лингвистическая конфигурация.

Установка ПО

Запуск системы в режиме БД:

1. Убедиться, что на машине, где будет развернут модуль агрегации новостей, есть правильно настроенный клиент БД;

2. Скопировать полностью папку **\Aggregator** на жесткий диск;

3. В файле **\Aggregator\AgregatorDB.cmd** задать в командной строке параметры присоединения к БД, например: `RCOAggregator.exe connect=пользователь/пароль@БД days=4 interval=1/1/2017-31/12/2020`;

4. Зарегистрировать агрегатор новостей как сервис операционной системы, отредактировав и выполнив файл **service_DB_install.cmd**;

5. Запустить сервис **RCONewsAggr**.

Запуск системы в режиме без использования БД:

Доступен режим работы программы без подключения к базе данных, в котором входные данные считываются из файловой системы, туда же печатаются результаты работы, а промежуточные данные хранятся в оперативной памяти компьютера. Для запуска программы в этом режиме необходимо:

1. Разместить входные данные в каталоге, путь к которому задаётся параметром gz-dir в AgregatorHDD.cmd: данные входного потока новостей должны быть записаны по дням (один день – один файл) с именами вида docs_yyyymmdd, каждый файл заархивирован архиватором gzip в расширение gz, где yyyy – год, mm – месяц, dd – день числами. Структура файла – таблица с символом табуляции в качестве разделителя колонок и названий колонок в первой строке:

- DOCID – идентификатор новостного сообщения
- ISSUEDATE – дата публикации
- RECORDDATE – дата скачивания
- SOURCEID – идентификатор источника, опубликовавшего сообщение
- SOURCENAME – название источника
- TITLE – заголовок сообщение
- TEXT – текст сообщения

2. Задать в AgregatorHDD.cmd параметры запуска (описаны в разделе Запуск программы) и запустить его. Можно также установить программу в качестве сервиса, запустив service_HDD_install.cmd, и обращаться к этому сервису.

Результаты работы программы записываются в файл clusters_added_N.txt, где N – номер итерации, в виде таблицы с символом табуляции в качестве разделителя колонок: cluster_id – номер кластера, doc_id – идентификатор документа, входящего в кластер, weight – вес документа в кластере.

Проверка работоспособности

- Убедиться, что сервис с именем, заданным при установке в **service_*_install.cmd присутствует в системе и запущен, по умолчанию это сервис *RCO.AggrDB*.**
- Проверить список процессов, убедиться, что процесс **AgregatorHDD** и заданное в *RCOAggregator.ini* число процессов **FXModuleWrapper** работают.
- Просмотреть файл в корневом каталоге сервиса с именем *AgregatorHDD_YYYYMMDD.log* за текущую дату. В файле должны присутствовать записи об окончании кластеризации и сохранении результирующих кластеров не позже двух периодов кластеризации от текущего времени.

Дополнительные настройки

Запуск Агрегатора производится со следующими параметрами:

- days=*число_дней* – ширина окна кластеризации в днях. Изначальное значение – 5 дней. Для тестовых запусков был выбран период 7 дней;
- connect=*строка_соединения_с_БД* – строка соединения с БД новостей в формате имя_пользователя/пароль@строка_подключения_TNS. Обязательный параметр. Необходимый набор полномочий для пользователя описан в разделе «**Ошибка! Источник ссылки не найден.**». Одновременно можно иметь только один работающий агрегатор в одной базе данных;

- gz-dir – путь к директории с новостными сообщениями в режиме работы без базы данных. Формат описан в разделе Запуск программы без использования БД.
- every=*число_часов* – шаг кластеризации в часах. В штатном режиме данный параметр определяет, как часто агрегатор будет обрабатывать новые поступающие в БД документы. В ретроспективном режиме – на сколько сдвигать окно кластеризации при следующей итерации. Исходно – 1 час;
- interval=*dd/mm/yyyy-dd/mm/yyyy* – интервал кластеризации в днях и часах. Данный параметр определяет, в каком (штатном или ретроспективном) режиме запущен агрегатор. Если данный параметр задан (ретроспективный режим), агрегатор производит необходимое число итераций кластеризации по документам, хранящимся в базе и принадлежащих заданному интервалу. Если параметр не задан, агрегатор обрабатывает свежие новости (штатный режим);

Удаление ПО

Удалить агрегатор новостей из сервисов операционной системы, отредактировав и выполнив файл **service_DB_install.cmd**;

Удалить содержимое каталога *Agregator* с диска.

Регламентное обслуживание

Устранение неисправностей в ходе эксплуатации ПО

Просмотреть *./AgregatorHDD_YYYYMMDD.log* за последний период кластеризации на предмет сообщений об ошибках. Все строки с сообщениями об ошибках содержат "*Error*".

Если ошибка связана с соединением с БД, проверить соединение с машины где развёрнут сервис клиентом БД. Если ошибки связаны с запуском модуля разбора текстов, просмотреть файлы *WrapperLog_*.log* и *./TermExtractorApp/fx_log.txt*. Если ошибка не устранима на месте, файлы с ошибками переслать в техническую поддержку RCO.

Обновление ПО

Остановить сервис с именем, заданным при установке в **service_*_install.cmd,** по умолчанию это сервис ***RCO.AggrDB.***

Дождаться появления записи о завершении в текущем *./AgregatorHDD_YYYYMMDD.log***.**

Скопировать обновление в каталог сервиса.

Снова запустить сервис.

Совершенствование ПО

Производитель, ООО «ЭР СИ О», периодические выпускает новые версии ПО, содержащие новые функциональные возможности.

Для получения всех обновленных версий ПО, по мере их выхода, необходимо наличие действующего договора на техническую поддержку ПО с ООО «ЭР СИ О».