ктн Янин Дмитрий Михайлович, 27 ЦНИИ МО РФ
(8(985)311-27-77)

Автоматический анализ текстов на основе использования технологий *«Big Data»*

Возникшая в последние годы диспропорция между огромными объемами информации и технологическими возможностями их обработки привела к возникновению так называемой проблемы «больших данных» («Big Data»). Особенно остро эта проблема стоит при решении задач обработки неструктурированной текстовой информации в процессе выполнения семантического анализа текстов и связано это, прежде всего, с тем, что при таком анализе тексты подвергаются сложной и многоступенчатой обработке с использованием словарей больших объемов. При этом обработка каждого текста выполняется последовательно процедурами семантико-синтаксического и концептуального анализа. Используемый технологический процесс часто не позволяет обеспечить требуемые в настоящее время скорости обработки текстов.

Между тем существенное повышение скоростей обработки текстов может лежать в плоскости уже апробированных решений, применяемых для параллельных вычислении Здесь в качестве базового принципа вычислений используется массово-параллельная обработка, масштабируемая без деградации на множество узлов обработки. Такая обработка выполняется в рамках таких технологии, как NoSQL, Map Reduce, Hadoop, R и др.

В системах семантической обработки текстовой информации основной задачей является построение формализованного представления смысловой структуры текстов -выделения в них смысловых единиц и установления связей между ними. Центральной процедурой при решении этой задачи является процедура семантико-синтаксического концептуального (понятийного) анализа текстов. Важнейшим средством автоматической смысловой обработки текстовой информации являются мощные словари наименований понятий, представленные преимущественно фразеологическими словосочетаниями

При решении задачи формализации смыслового содержания текстов необходимо методами семантико-синтаксического и концептуального анализа обработать текст, разделить его на предложения, выделить из него единицы смысла (наименования понятий) - слова и словосочетания, выражающие понятия Основными параметрами, по которым может оцениваться функционирование систем семантического анализа текстов, являются качество анализа текстов и скорость их обработки. Качество анализа текстов определяется, прежде всего, использованием адекватной модели представления их смыслового содержания, эффективными методами и алгоритмами анализа текстов, наличием декларативных средств, обеспечивающих высокое покрытие анализируемых текстов.

Скорость обработки текстов зависит от эффективности методов и алгоритмов семантической обработки, числа проходов по тексту при его

обработке и от объемов грамматических таблиц и словарей, используемых при обработке текста. Для оценки быстродействия текущей версии системы анализа текста необходимо подсчитать скорости обработки текстов на различных этапах его обработки. Поскольку обработка текста выполняется последовательно процедурами графематического, морфологического, семантико-синтаксического и концептуального анализа, то необходимо просуммировать время обработки на каждом из этих этапов. Повышение качества обработки и анализа текстов и связанное с этим неизбежное значительное усложнение алгоритмов обработки текстов, а также увеличение объемов их декларативных средств неизбежно приведет к существенному снижению скоростей обработки текста. Так, при сравнении скоростей обработки текстов на одном и том же программном комплексе со стандартными и модифицированными словарями и грамматическими таблицами, суммарный объём которых соответственно равен 1.8 млн. и 3.6 млн. словарных статей, скорость обработки текстов на этом комплексе с модифицированными словаря снизилась на 34%.

Между тем постоянное возрастание потоков текстовой информации во всех сферах человеческой деятельности требует существенного повышения производительности систем обработки и анализа текстовой информации, поэтому возникает необходимость поиска новых технологических решений этой проблемы и одним го таких решений могут быть технологии "Big Data"(больших данных)

Наибольший интерес представляет технология Hadoop (проект фонда Apache Software Foundation) представляющая собой свободно распространяемый набор утилит, библиотек и программный каркас для разработки и выполнения распределённых программ, работающих на кластерах из любого числа узлов. Технология Hadoop разработана в рамках вычислительной парадигмы Map Reduce, согласно которой приложение разделяется на большое количество одинаковых элементарных заданий, выполнимых на узлах кластера и естественным образом сводимых в конечный результат. Ядром этой технологии является распределённая файловая система (HDFS).

При автоматическом анализе текстов можно воспользоваться алгоритмом Map Reduce, представляющим собой модель для распределенных вычислений. В рамках этой модели происходит распределение входных данных на рабочие узлы (individual nodes) распределенной файловой системы для предварительной обработки map-шаг) и затем свертка (объединение) уже предварительно обработанных данных (reduce-шаг). При реализации этой модели необходимо технологический процесс обработки текста разделить на элементарные семантические процедуры, выполняемые над промежуточными результатами обработки текста. При этом текст и результаты его обработки на различных этапах также могут быть разделены на различные фрагменты.

На каждом узле должен выполняться определенный этап обработки конкретного фрагмента текста. Входными данными для них должны быть результаты работы рабочих узлов, выполняющих предыдущие технологические операции, и. соответственно, выходные данные работы

конкретного рабочего узла должны быть входными данными для рабочих узлов, выполняющих следующие технологические операции.

Так, например, на узлы, реализующие морфологический анализ слов можно подавать отдельные слова текста, которые будут объединяться в результаты обработки предложения, а эти результаты также будут поданы на узлы, реализующие семантико-синтаксический анализ предложений

Таким образом, для обеспечения возможности использования технологий «Big Data» для решения различных задач автоматической обработки текстовой информации необходимо определить состав элементарных семантических процедур, структуру их входных и выходных данных и декларативные средства, а также разработать идентификаторы, регламентирующие последовательность выполнения элементарных операций.

Вышеописанная технологическая схема требует некоторых уточнении:

1. При семантической обработке текста необходимо соблюдать строгую последовательность при разделении текста на фрагменты, их технологическую обработку и последующую сборку как частных, так и конечных результатов обработки текста. Поэтому каждый фрагмент текста, каждый частный и конечный результат обработки должен сопровождаться информацией об их местоположении в исходном тексте или в результатах его обработки, а также информацией об выполненном этапе их обработки.
2. Текст недопустимо произвольным образом делить на фрагменты. Это может привести к разрушению его смысловой структуры. Необходимо разработать процедуры, выполняющие такое деление текста на основе его упрощенного формального анализа. Необходимо также разработать процедуры корректного разделения и объединения частных или конечных результатов анализа текста.
3. Для обеспечения функционирования предлагаемой технологии в распределённой файловой системе HDFS необходимо, чтобы все исходные данные и выходные данные были представлены в виде файловой структуры. Это требование можно обеспечить путем преобразования всех данных в XML-структуру.
4. Все файлы, содержащие информацию об входных и выходных данных конкретного текста, должны сопровождаться идентификатором, в котором содержится вся необходимая информация для его обработки в распределённой файловой системе HDFS.
5. Современные вычислительные системы с массово-параллельной архитектурой имеют, как правило, неоднородную структуру. При формировании параллельных ветвей вычислительных процессов (mapping) возникает задача эффективного выделения свободных вычислительных модулей для решения конкретной задачи с целью минимизации времени ее выполнения. Процедуры загрузки вычислительных ресурсов системы должны учитывать алгоритмические особенности планируемых вычислительных процессов, технические характеристики используемых процессоров и свойства коммуникационной среды.