Методы анализа текстов в технологиях «Big Data»

© А.А. Хорошилов
ИПИ РАН
khoroshilov@mail.ru

© О.Н. Пошатаев
ЦИТиС
Москва
pon-52@inevm.ru

Аннотация

В статье представлены методы и алгоритмы базовых процедур системы анализа текстов и показана возможность их применения в технологиях «big data». Использование этих технологий для решения задач смысловой обработки текстовой информации требует значительной доработки как процедур реализующих парадигму решений MapReduce, так и процедур, реализующих технологический процесс семантической обработки и анализа текстовой информации.

1 Введение

Возникшая в последние годы диспропорция между огромными объемами информации и технологическими возможностями их обработки привела к возникновению проблемы «big data». Особенно остро эта проблема стоит при решении задач 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. Обработка текстов выполняется последовательно комплексом процедур семантико-синтаксического и концептуального анализа. Используемый технологический процесс часто не позволяют обеспечить требуемые в настоящее время скорости обработки текстов.

Между тем существенное повышение скоростей обработки текстов может лежать в плоскости уже апробированных решений проблемы «big data». В качестве базового принципа обработки “больших данных” предлагается массово-параллельная обработка, масштабируемая без деградации на множество узлов обработки. Такая обработка выполняется в рамках таких технологий, как NoSQL, MapReduce, Hadoop и др.

Авторами настоящего доклада совместно с рядом сотрудников ИПИ РАН и ЦИТиС разработано и введено в промышленную эксплуатацию ряд систем обработки и анализа текстовой информации [1,2,3,4]. В этих системах также стоит проблема существенного увеличения их производительности. Ниже рассмотрим методы и алгоритмы базовых процедур анализа текстов, реализованные в этих системах и возможности их применения в технологиях «big data».

2 Формализация смысловой структуры текстов

В системах семантической обработки текстовой информации основной задачей является формализация 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 [3,5,6].

При анализе текстов необходимо также учитывать, что в них одни и те же объекты и процессы могут описываться с различной степенью общности и с помощью различных языковых средств. Поэтому при решении задач автоматической смысловой обработки текстовой информации необходимо учитывать такие явления как синонимия, гипонимия (родо-видовые отношения), разнообразие средств выражения межфразовых связей.

Основной структурной единицей текста традиционно считается предложение. Некоторые лингвисты склонны даже рассматривать его в качестве основной единицы смысла [7,8]. Предложения выступают в тексте не изолированно друг от друга, а в тесной смысловой связи. В основе этой связи лежат мыслительные образы тех конкретных или абстрактных объектов (ситуаций, явлений), которые человек имеет в виду, когда он порождает текст. Образы этих объектов имеют определенную структуру. Кроме того, они дополнительно структурируются человеком при их описании на естественном языке. Соответственно этому структурируется и текст.

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

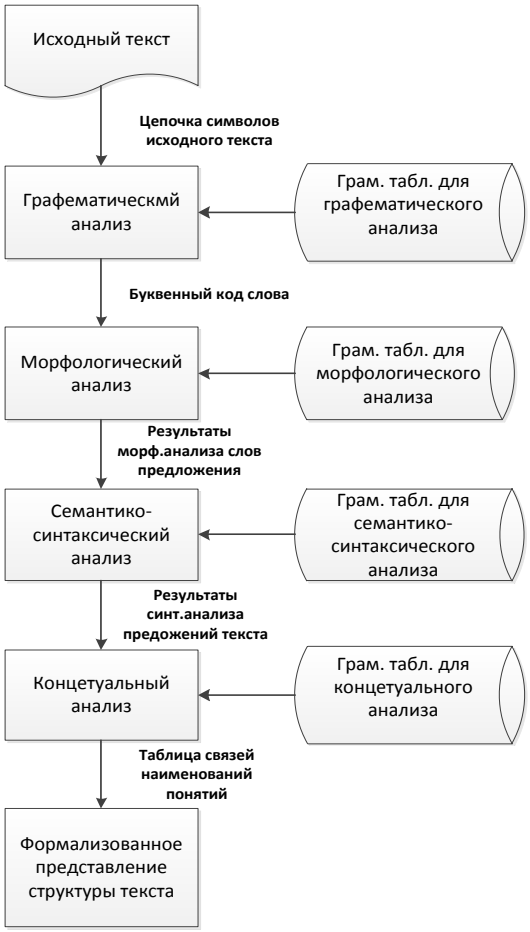
При прочтении текста у читателя, как и у автора текста, возникнет определенный мыслительный образ. Мыслительные образы автора текста и его читателя обычно не тождественны, но в основе своей они должны быть сходными. Целью передачи информации с помощью текста является не столько исчерпывающее описание мыслительных образов его автора, сколько инициация процесса порождения соответствующих мыслительных образов у читателей. Письменный текст, как и звуковая речь, развертывается последовательно во времени, т.е. имеет линейную структуру, тогда как мыслительные образы “многомерны”. При их словесном описании может быть принят различный порядок линейной развертки, но цель описания должна быть в основном одна и та же – воссоздание в сознании читателей мыслительных образов, подобных мыслительным образам автора текста. Такое воссоздание осуществляется постепенно, путем восприятия предложения за предложением и “монтажа” возникающих при этом частичных образов в целостный мыслительный образ, соответствующий содержанию текста. При этом в каждом предложении элемент его актуального членения “тема” выполняет роль “стыковочного узла”, служащего для подключения нового частичного мыслительного образа, обозначаемого этим предложением, к ранее построенному мыслительному образу.

Описанная модель восприятия текста позволяет объяснить тот факт, что связи между предложениями выражаются в большинстве случаев с помощью лексических повторов: в “стыковочных узлах” предложений повторяются наименования понятий предшествующего текста либо буквально, либо в виде синонимических и эллиптических конструкций, либо в виде родовых наименований понятий и местоимений. Для связи с предыдущим текстом применяются также средства, основанные на указании координат его фрагментов (слов и выражений типа “на основании вышеизложенного”, “рассмотренный нами ранее”, “описанный в главе…”, “в приведенном выражении” и т.п.

Исходя из вышесказанного, 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. Ниже рассмотрим базовый набор процедур семантико-синтаксического и концептуального анализа, а также декларативные средства, обеспечивающие их функционирование, методы, входные и выходные данные.

3 Базовые процедуры анализа текста

Базовые процедуры анализа текста рассмотрим на примере системы семантического анализа текстов, ориентированной на построение формализованной смысловой структуры текста. На Рис.1 приведена общая схема системы семантического анализа текстов.



На Рис.1. Общая схема системы семантического анализа текстов.

Ниже более подробно рассмотрим основные процедуры семантической обработки и анализа текстов.

3.1 Графематический анализ текста

Графематический анализ предназначен для предварительного анализа текста по представляющей его последовательности символов. Методы и алгоритмы, реализующие этот анализ описаны в ряде работ [1,2,3]. В результате этого анализа определяется язык текста, устанавливаются местоположения слов, предложений, абзацев, фамильно-именной группы, дат и электронных адресов. Для автоматического определения указанной информации о формальной структуре текста в соответствующих методах графематического анализа используются следующий набор грамматических таблиц и словарей:

- Словарь для установления языка текста;
- Таблица признаков для выделения слов и разделителей в тексте;

- Таблица признаков для выделения дат в цифровых форматах;
- Таблица признаков для выделения фамильно-именной группы;
- Таблица признаков для выделения электронных адресов;
- Таблица признаков для разделения текста на предложения;
- Таблица признаков для разделения текста на абзацы;
- Таблица признаков для выделения примечаний.

Исходными данными является последовательность символов исходного текста.

Выходными данными является информация о местоположении слов, разделителей, дат, установленных фамильно-именных групп, электронных адресов, предложений, заголовков и примечаний.

В таблице 1 приведен класс CGraphemAnalys, реализующий основные методы графематического анализа.

Таблица 1.

Класс CGraphemAnalysis, реализующий методы графематического анализа текста

CGraphemAnalysis
МЕТОДЫ
• Определение языка текста; • Разделение входного текста на слова, разделители и т.д. • Выделение дат в цифровых форматах; • Выделение электронных адресов; • Выделение предложений; • Выделение абзацев, заголовков, примечаний;
ДАННЫЕ
• Исходный текст (Plain Text); • Длина текста (в символах); • Язык текста; • Массив адресов слов; • Массив адресов предложений; • Массив адресов абзацев;

3.2 Морфологический анализ слов текста

Морфологический анализ слов естественных языков предназначен для определения структуры слов и назначения им грамматических признаков, необходимых для выполнения последующих процедур автоматической обработки текстовой информации (например, морфологического синтеза слов, синтаксического анализа и синтеза текстов и их концептуального анализа). Методы и алгоритмы, реализующие этот анализ описаны в работах [1,2,3]. В результате этого анализа определяется структура слова и набор грамматической информации, определяющей это слово вне контекста. Для автоматического определения указанной информации о слове в соответствующих методах морфологического анализа используются следующий набор грамматических таблиц и словарей:

- Словарь слов-исключений (коротких и служебных слов) с назначенной ГИ;
- Словарь конечных буквосочетаний слов русского языка;
- Таблица флективных классов слов;
- Таблица супплетивных форм слов;
- Таблица установления наличия чередований в основах слов;
- Таблица подстановок для реализации чередований в основах слов;
- Таблица наборов грамматической информации слов;
- Таблица окончаний русских слов.

Таблица 2.

Класс CMorphoAnalysis, реализующий методы морфологического анализа текста

CMorphoAnalysis
МЕТОДЫ
• Поиск в словаре слов-исключений; • Поиск в словаре конечных буквосочетаний слов; • Поиск в таблице супплетивных форм слов; • Установления наличия чередований в основах слов; • Определение флективного класса слов; • Назначение слову грамматической информации;
ДАННЫЕ
• Буквенный код слова; • Длина слова; • Длина окончания; • Лексикограмматический класс слова; • Флективный класс слова; • Наборы грамматических признаков (род, число, падеж, лицо);

Исходными данными является буквенный код отдельного слова.

Выходными данными является информация о структуре слова и набор грамматической информации, определяющей это слово вне контекста.

3.3 Семантико-синтаксический анализ предложений текста

Семантико-синтаксический анализ текстов проводится с целью формализованного представления их структуры – выделения в них смысловых единиц и установления связей между ними. При этом структура текстов может интерпретироваться по-разному и описываться на различных формализованных языках. При описании синтаксической структуры текстов удобно опереться на какую-либо ее формализованную модель, например, на модель дерева зависимостей. Согласно этой модели каждое предложение представляется в виде дерева, в узлах которого находятся слова. Слова соединяются друг с другом

стрелками, выражающими отношения непосредственной доминации, и направленными от подчиняющего (определяемого) слова к подчиненному (определяющему). Степень дифференциации этих отношений может быть разная. Причем, чем больше степень дифференциации, тем сложнее процесс описания текстов. Методы и алгоритмы, реализующие этот анализ описаны в работах [1,2,3]. В результате этого анализа определяется синтаксическая структура предложения: производится членение на простые предложения, определяются главные и второстепенные члены предложения и устанавливаются смысловые связи между ними, строится дерево зависимости предложения и для каждого слова определяется однозначная грамматическая информация, соответствующая контексту. Для автоматического определения указанной информации о структуре предложения в соответствующих методах семантико-синтаксического анализа используются следующий набор грамматических таблиц и словарей:

- Таблица описаний правил синтаксического анализа
- Таблица описаний правил установления смысловых связей между словами предложения
- Таблица описаний правил определения однозначной грамматической информации слов с учетом контекста

Таблица 3.

Класс CSyntAnalysis, реализующий методы семантико-синтаксического анализа текста

CSyntAnalysis
МЕТОДЫ • Членение на простые предложения; • Определение главных членов предложения; • Определение второстепенных членов предложения; • Установление однородных второстепенных членов предложения; • Построение дерева зависимостей; • Разрешение многозначности грамматической информации слов;
ДАННЫЕ • Адрес предложения; • Длина предложения; • Массив адресов простых предложений; • Массив адресов словосочетаний; • Массив адресов слов;

Исходными данными являются слова предложения и результаты их обработки процедурой морфологического анализа.

Выходными данными является информация о семантико-синтаксической структуре предложения.

3.4 Концептуальный анализ текстов

Концептуальный анализ текстов предназначен для определения смысловой структуры текстов, выявления их понятийного (концептуального) состава текстов и установления связей между наименованиями понятий. Эту задачу невозможно решить только путем анализа синтаксической структуры текстов без привлечения семантических признаков. Сложность этой задачи связана, прежде всего, с вариативностью форм представления наименований понятий в текстах. Авторами было разработано несколько вариантов решения этой задачи [3,5]. Наиболее эффективным решением являлся метод концептуального анализа текстов с контролем по тезаурусу (эталонному словарю), включающему более 1.8 млн. понятий и свыше 400 тыс. связей между ними. В результате этого анализа в тексте выявляются наименования понятий, устанавливается концептуальная структура текста и строится таблица связей между наименования понятий. Для автоматического определения указанной информации о структуре текста в соответствующих методах концептуального анализа используется следующий набор грамматических таблиц и словарей:

- Эталонный словарь наименований понятий;
- Словарь смысловых связей между наименованиями понятий;
- Словарь смысловых связей слов.

Таблица 4.

Класс CConceptAnalysis, реализующий методы концептуального анализа текста

CConceptAnalysis
МЕТОДЫ • Выявление наименований понятий в тексте; • Установления парадигматических отношений между понятиями; • Разрешение анафорических ссылок; • Установление синтагматических отношений между понятиями; • Приведение понятий к их каноническим формам; • Построение таблицы связей меду понятиями;
ДАННЫЕ • Массив адресов наименований понятий в тексте; • Массив длин наименований понятий в тексте; • Массив адресов наименований понятий-отношений в тексте;

Исходными данными является информация о семантико-синтаксической структуре предложения.

Выходными данными является таблица связей между наименованиями понятий

3.5 Формализованное представление смысловой структуры текстов

Таблица связей наименований понятий представляет собой машинное представление смысловой структуры текста. Его визуальное представление можно сформировать в виде семантической карты текста, представляющего

собой ориентированный граф, в узлах которого находятся объекты, события или темы документов, а дугами являются смысловые отношения между ними. Связи могут быть либо типизированными (определен семантический тип связи), либо логическими (установлен факт их наличия).

Семантическую карту можно построить с помощью пакета утилит Graphviz, разработанного специалистами лаборатории AT&T [9]. В качестве исходных данных для этой утилиты используется описание графа на специальном языке dot, а на выходе формируется граф в виде графического, векторного или текстового файла. При этом также возможен более сложный выход, например с использованием координатной сетки, которую потом можно использовать для обозначения областей при показе на странице гипертекста.

4 Перспективы развития систем семантического анализа текстов

Основными параметрами, по которым может оцениваться функционирование систем семантического анализа текстов, являются качество анализа текстов и скорость его обработки. Качество анализа текстов определяется, прежде всего, использованием адекватной модели представления их смысловой структуры, эффективностью методов и алгоритмов анализа текстов, составом декларативных средств, обеспечивающих высокое покрытие анализируемых текстов.

Скорость обработки текстов зависит от быстродействия применяемых методов и алгоритмов семантической обработки, числа проходов по тексту при его обработке и от объемов грамматических таблиц и словарей, используемых при обработке текста.

4.1 Пути повышения качества анализа текстов

Из предыдущих рассуждений следует, что ориентация на фразеологические словосочетания как на основную форму представления наименований понятий в естественных языках позволяет более точно учитывать семантико-синтаксическую структуру текстов и построить более эффективную систему смысловой обработки текстовой информации. Построение такой системы неизбежно связано с выявлением понятийного состава русского языка, которое по нашим представлениям содержит несколько сот миллионов наименований понятий. Косвенным подтверждением этой оценки являются данные Международного терминологического центра ИНФОТЕРМ[3,10]. Согласно этим данным, количество различных терминов в языках достигает 50 миллионов, а количество наименований товаров - 100 миллионов. Но многообразие устойчивых фразеологических единиц в естественных языках далеко не исчерпывается только этими двумя типами лексических единиц. В связных текстах нетерминологические фразеологические единицы встречаются чаще, чем терминологические. Следовательно, есть основания предположить, что их больше, чем терминов и наименований товаров вместе взятых. Поэтому любая система анализа текстов в перспективе должна включать в свой состав систему мощных политематических словарей наименований понятий, содержащих нескольких миллионов (или десятков миллионов) словарных статей состоящих преимущественно из фразеологических словосочетаний. В словарях должны содержаться также сведения об отношениях синонимии и о родо-видовых отношениях между понятиями.

4.2 Оценка быстродействия систем анализа текстов

Для оценки быстродействия системы анализа текста необходимо подсчитать скорости обработки текстов на различных этапах их анализа. В качестве теста был взят текст по авиакосмической тематике объемом 13124 слова (103139 символа). В процессе обработки этого текста последовательно процедурами графематического анализа, морфологического анализа, семантико-синтаксического анализа и концептуального анализа подсчитывалось время работы каждого этапа и общее время его обработки в текущей версии системы. Далее были подсчитаны скорости функционирования экспериментальной перспективной системы анализа текстов с модифицированными словарями и грамматическими таблицами. При этом их объемы в перспективной версии системы значительно возросли. Так, например, суммарный объём тематических и политематических словарей увеличился с 1.8 млн. до 9.6 млн. словарных статей. Полученные результаты приведены в таблице 5.

Результаты сравнительной оценки быстродействия текущей версии системы и перспективной версии показывают, что повышение качества обработки и анализа текстов и связанное с этим неизбежное значительное увеличение объемов декларативных средств в традиционной реализации приводит к существенному снижению производительности этих систем.

Таблица 5.

Таблица замеров быстродействия выполнения процедур семантического анализа текстов и длительности их выполнения

Название процедуры семантического анализа текста	Скорость обработки (слов/ сек)	Длительность процесса (в %)	Скорость обработки (слов/ сек)	Длительность процесса (в %)
	Текущая версия		Перспективная версия	
Графематический анализ	52496	5%	27342	4%
Морфологический анализ	43746	6%	36455	3%
Семантико-синтаксический анализ	7954	33%	7291	15%
Концептуальный анализ	4687	56%	1402	78%
Система семантического анализа (Общее быстродействие)	**4780**	**100%**	**1093**	**100%**

4.3 Пути повышения быстродействия – технологии “big data”

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 из таких решений, как выше уже было сказано, могут быть технологии “big data”(больших данных).

Наибольший интерес представляет технология с (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 MapReduce, 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 (Hadoop Distributed File System) [12,13,14].

При автоматическом анализе текстов можно воспользоваться алгоритмом MapReduce, представлявшим 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 него должны быть результаты предыдущей обработки этого фрагмента, и, соответственно, выходные данные работы этого 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» при решении различных задач автоматической обработки текстовой информации необходимо определить состав элементарных семантических процедур, структуру их входных и выходных данных, а также разработать идентификаторы, регламентирующие последовательность выполнения элементарных операций. Кроме того необходимо создать дистрибутивы программного обеспечения, и обеспечить их автоматическую загрузку на рабочие узлы.

На рис. 2 представлена технологическая схема работы Hadoop при реализации процесса семантического анализа текста.

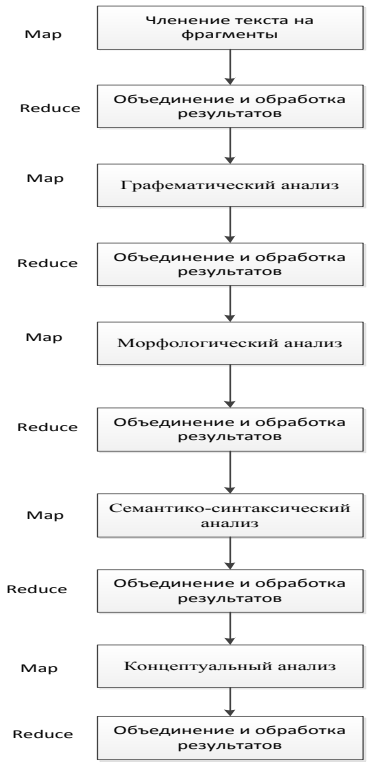


Рис 2. Технологическая схема работы Hadoop при реализации процесса семантического анализа текста.

Приведенные на схеме этапы обработки имеют различные характеристики вычислительных процессов – уровень параллелизма, вычислительную сложность, объем и интенсивность обмена данными и т.д., которые могут существенно зависеть от конкретного вида обрабатываемого текста. Выделение фиксированных сегментов кластера для каждого типа операций может привести к неравномерности загрузки вычислительных узлов системы, простою или чрезмерной загруженности части узлов, обрабатывающих различные этапы семантического анализа. Для исключения этой ситуации необходимо разработать процедуры балансировки загрузки вычислительных модулей системы за счет динамической реконфигурации размеров сегментов, выделяемых под каждый этап.

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.

При реализации приведенной технологической схемы необходимо также учесть ряд дополнительных требований:

1. В процессе семантической обработки текста необходимо соблюдать строгую последовательность при разделении текста на фрагменты, их технологической обработки и последующей сборки результатов обработки фрагментов текста и всего текста.

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 промежуточных 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 Заключение

В связи с лавинообразным ростом информации во всех областях деятельности современного общества решение проблемы создания высокопроизводительных систем обработки текстовой информации является чрезвычайно актуальной. Попытки ее решения путем разработки высокоскоростных методов обработки информации и создания многопоточных систем анализа текстовой информации не привели к кардинальному повышению их производительности. Сложность решения этой проблемы связана с применением сложной и многоступенчатой технологии обработки текстовой информации, базирующейся на использовании словарей больших объемов.

В докладе предпринята попытка показать возможность создания высокопроизводительных систем семантического анализа текстов на основе применения технологий “big data”. Но эти технологии, изначально создававшиеся для обеспечения выполнения высокоскоростных вычислительных задач, в контексте проблем смысловой обработки текстовой информации требуют значительной доработки как процедур, реализующих парадигму решений MapReduce, так и процедур, реализующих технологический процесс

семантической обработки и анализа текстовой информации.

В настоящее время коллектив разработчиков под руководством авторов настоящего доклада работает по этой проблеме и занимается определением состава элементарных семантических процедур, разработкой структуры их входных и выходных данных, а также разработкой идентификаторов, регламентирующих последовательность выполнения элементарных операций и обеспечивающих их корректную обработку в распределённой файловой системе HDFS.

Литература

[1] http://retrans.ru;

[2] http://metafraz.com;

[3] Белоногов Г. Г., Калинин Ю. П., Хорошилов А. А. Компьютерная лингвистика и перспективные информационные технологии. Теория и практика построения систем автоматической обработки текстовой информации. – Москва: Информационно-издательское агентство "Русский мир", 2004.

[4] Белоногов Г. Г., Гиляревский Р. С., Хорошилов Ал-др А., Хорошилов Ал-ей А. Развитие систем автоматической обработки текстовой информации.// Нейрокомпьютеры: разработка, применение. – 2010, № 8.

[5] Белоногов Г. Г., Гиляровский Р. С., Хорошилов А.А. Проблемы автоматической смысловой обработки текстовой информации.// Научно-техническая информация. сер. 2. Информационные процессы и системы/ Всероссийский институт научной и технической информации РАН.– 2012, № 11.

[6] Белоногов Г.Г., Хорошилов Ал-др А., Хорошилов Ал-сей А. Единицы языка и речи в системах автоматической обработки текстовой информации.// Сб.Научно-техническая информация. сер. 2.– М.: ВИНИТИ, 2005, № 11.

[7] Белоногов Г.Г., Быстров И.И., Козачук М.В. Новоселов А.П., Хорошилов А.А. Автоматический концептуальный анализ текстов. // Сб.Научно-техническая информация. сер. 2.– М.: ВИНИТИ, 2002, № 10

[8] Звегинцев В.А. Предложение и его отношение к языку и речи. - М.: Издательство Московского университета, 1976.

[9] Соссюр Фердинанд де. Курс общей лингвистики. – М.: Прогресс, 1977.

[10] Белоногов Г.Г., Гиляревский Р.С., Селетков С.Н., Хорошилов А.А. О путях повышения качества поиска текстовой информации в системе Интернет. Сб. Научно-техническая информация, Серия 2. – М.: ВИНИТИ, 2013, № 8.

[11] http://lib.custis.ru/Graphviz;

[12] http://www.compress.ru/article.aspx?id=23469&iid=1080;

[13] http://www.tadviser.ru/index.php;

[14] http://www.statsoft.ru/products/Enterprise/big-data.php.

Methods of text analysis in «big data» technologies

Alexandr A. Khoroshilov, Oleg. N. Poshataev

This paper presents the methods and algorithms of basic procedures for the system of text analysis and shows the possibility of their application in "Big Data" technologies. Use of these technologies for problem solving of semantic text information processing requires the essential updating both the procedures implementing the MapReduce solution paradigm, and the procedures implementing technological semantic processing and analysis of text information.