УДК 004.9


А.Н. Швецов, А.В. Колосов
Вологодский государственный университет

АВТОМАТИЧЕСКАЯ ОБРАБОТКА ИНФОРМАЦИИ НА ОСНОВЕ ФОРМАЛЬНОЙ СЕМАНТИКИ

Статья посвящена проблемам анализа текстов. Ввиду увеличения количества информации возникает необходимость разработки инструментов с целью фильтрации, формирования смыслового портрета, а также навигации по базам данных и получения необходимой текстовой информации.

Информационный поиск, обработка естественно-языковых текстов, разработка системы семантического анализа.

Объемы информации, циркулирующие в информационно-вычислительных системах, с каждым годом возрастают все быстрее. В 2011 году мировой объем данных составил более 1,8 зеттабайт (1,8 трлн Гб). По данным исследования International Data Corporation (IDC) установлено, что объемы данных удваиваются каждые два года. Так, к 2020 году объемы информации, которой необходимо управлять, вырастут в 50 раз. При этом доля полезной информации составит всего 35 %. Количество текстовой информации также будет пропорционально увеличиваться. Вместе с этим, учитывая сложную структурированность естественно-языковых текстов, анализ текстов представляет собой актуальную проблему [1]. Разработка эффективных подходов к обработке текстов с целью фильтрации, формирования смыслового портрета, навигации по базе текстов и т.д. является одним из наиболее актуальных направлений современных информационных технологий.

Возможность управления и анализа больших объемов текстовой информации, получение из хранилища необходимых данных предоставляют методы и инструменты текстовой аналитики, широко используемые в сфере информационного поиска.

Под информационным поиском (information retrieval) чаще всего понимают поиск в некоторой коллекции неструктурированных данных (чаще всего текстовых), которые удовлетворяют информационным потребностям лица, проводящего поиск. Под неструктурированными данными понимают информацию, которая не имеет строгой организации, подразумевающей быструю автоматическую обработку.

В большинстве традиционных поисковых систем используются следующие виды поиска:

1. Двоичный поиск. Поисковая машина определяет наличие/отсутствие слов запроса в целевом тексте. В качестве запроса используется список слов. Результаты поиска не сортируются или сортируются на основе внешних по отношению к поиску данных (например, по дате создания документов).

2. Частотный поиск. Учитывается частота встречаемости ключевых слов в исходных документах. Результаты сортируются на основе частоты встречаемости. Этот вид поиска отличается от двоичного дополнительной метрикой.

3. Поиск по рубрикатору. Рубрикатор может создаваться как вручную, так и автоматически. Когда речь идет об автоматическом создании рубрикаторов, то говорят о кластеризации – объединении документов на основе сходства или поиск документов определенной тематики.

4. Поиск по вопросу на естественном языке. Пользователь вводит вопрос на естественном языке. А система пытается ответить на поставленный вопрос.

5. Поиск с учетом ссылочных характеристик. Большинство методов являются расширениями двоичного поиска. В общем смысле все подобные методы реализованы на основе поиска по ключевым словам. Следовательно, результат поиска зависит от присутствия или отсутствия ключевых слов в исходном тексте. Однако основная проблема естественного языка заключается в том, что он обладает двусмысленностью. В этом случае слова могут быть связаны с другими словами внутри предложения, вне его, и, возможно, с общим окружением.

Инструменты текстовой аналитики позволяют собирать, систематизировать и анализировать текстовые данные в автоматическом режиме при помощи лингвистических правил, статических методов и алгоритмов семантического разбора естественно-языковых сообщений.

Семантика занимается анализом отношений между знаками и обозначаемыми объектами, между словами и соответствующими им понятиями, а также изучает отношения между значениями простых знаков и значениями сложных знаков, составленных из простых. Например, отношения между значением слов и значением предложений, построенных из этих слов.

В сфере информационного поиска семантический анализ используется для решения задачи понимания машиной естественных текстов.

Целью семантического анализа является переход от структуры синтаксических связей к ее смысловой интерпретации [2]. На выходе формируется множество семантических структур, построенных в соответствии с принятой формальной нотацией (семантической моделью).

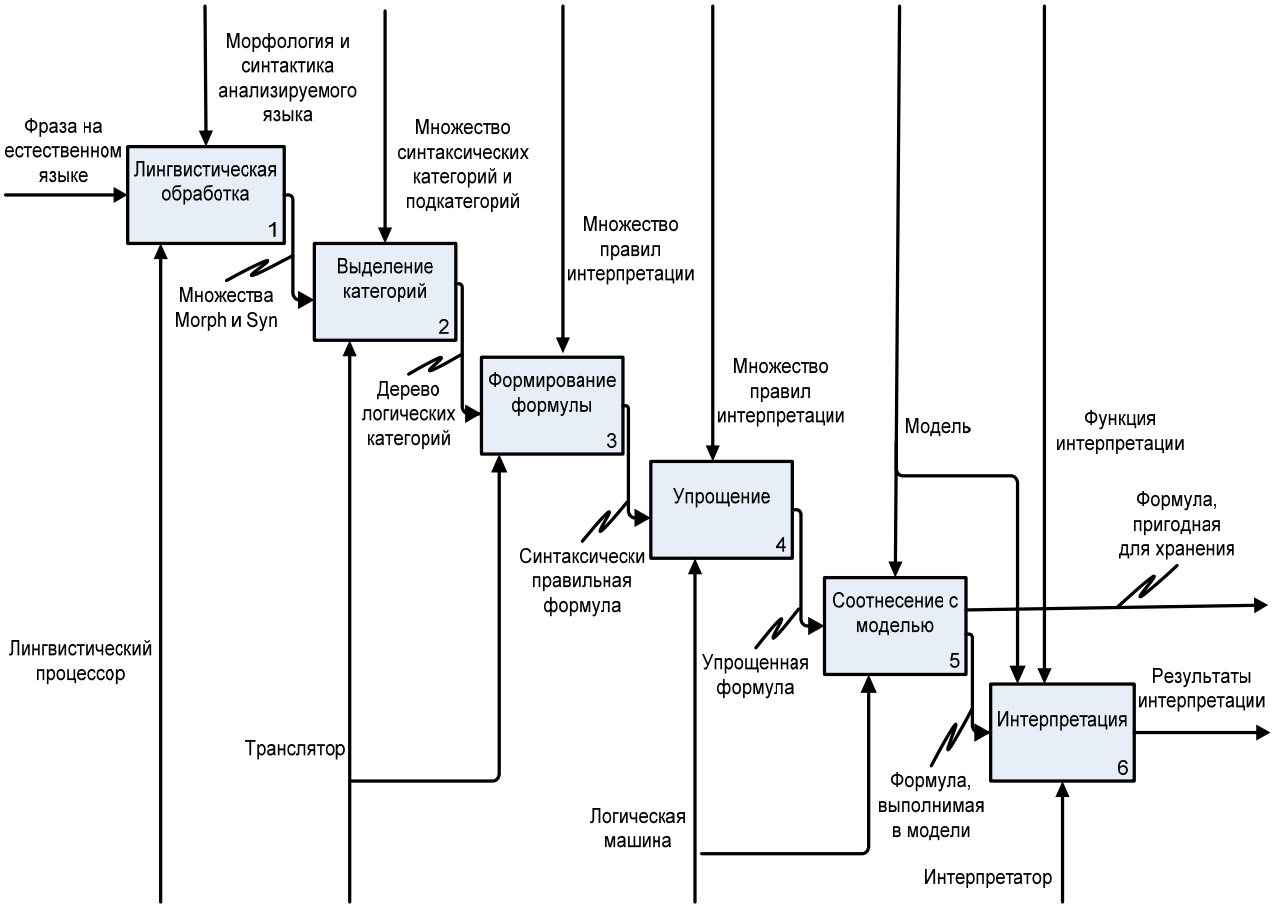

Рис. 1. Метод автоматической семантической обработки информации на основе формальной семантики

В работе [3] была рассмотрена проблема системы семантического поиска информации на предприятии. В связи с этим актуальным становится разработка системы обработки естественно-языковых текстов для повышения качества результатов поисковых запросов.

Для разработки используется метод автоматической семантической обработки информации на основе формальной семантики. В общем виде алгоритм будет выглядеть следующим образом (рис. 1).

В ходе выполнения система должна пройти все шесть этапов. Рассмотрим каждый этап выполнения.

1. Морфологическая и синтаксическая обработка фразы (лингвистическая обработка). Формируется два множества значений. В первое множество выделяются слова и их характеристики. Во второе заносятся синтаксические категории, выделяемые данным анализатором. Другими словами второе множество определяет отношения подчиненности древовидной структуры синтаксического построения групп. Второе множество при этом связано с характеристиками из первого с помощью функции соотнесения.

Можно выделить следующие основные характеристики: начальная форма, часть речи. К данным характеристикам по необходимости могут добавляться дополнительные.

Для реализации множеств целесообразно использовать базу данных. Для первой таблицы можно выделить следующие поля: уникальный идентификатор (id), слово, часть речи, начальная форма, номер в предложении, номер предложения в тексте.

Для второй основными полями можно выделить уникальный идентификатор (id), наименование категории, множество подкатегории.

2. Выделение синтаксических категорий и подкатегорий. Рассмотрение альтернатив. Рассматриваем созданную древовидную структуру с целью большей формализации исходного документа. Происходит объединение категории в более крупные или разделение на подкатегории на основе созданной базы подчиненностей.

3. Перевод фразы в формулу логики на основе набора категорий и правил трансформации. На основе набора категорий и правил трансформации строится формула логики, являющаяся отображением фразы естественного языка на формальном языке формул.

Функция реализуется на основе таблицы категории, проходя по всем подкатегориям и соотнося синтаксическую группу с категориальным определением.

4. Упрощение полученной формулы. Полученная формула чаще всего будет избыточна. В целях упрощения дальнейшей обработки и хранения возможно ее упрощение. Поэтому упрощение связано с последовательным применением сокращением взаимно уничтожающих друг друга операторов получения интенсионала и экстенсионала выражения.

5. Соотнесение полученной формулы с выбранной моделью. Проверка на выполнимость полученной формулы в данной модели. Результатом этой операции является знание о том, что формула не противоречит модели. В этом случае ее можно интерпретировать или добавить в общее хранилище знаний.

6. Интерпретация полученной формулы в модели. На данном этапе возможно использование формулы в реальных системах, в которых может использоваться управление на естественном языке. Таким образом, алгоритм может перевести сообщения на язык, понятный машине: набор управляющих сигналов.

Правила, на которые система будет опираться, получая входные данные:

1. Документ содержит текст, который, в свою очередь, состоит из предложений.

2. Каждому предложению в тексте присваивается порядковый номер.

3. Предложения содержат множества слов, которые состоят из слов. Словам присваиваются порядковые номера.

4. Возможное число предложений определяется исходным текстом (в том числе – ноль).

5. Для продвижения по алгоритму каждое слово должно иметь как минимум одну характеристику – форму, в которой оно было употреблено в предложении. В зависимости от сложности и разнообразности слов может возникнуть ситуация невозможности определения дополнительных морфологических характеристик слова: начальная форма, часть речи, род, падеж и другие.

6. Предложения содержат множества слов.

7. Множества слов бывают двух типов – группы и клаузы. Клаузы могут содержать группы, но не наоборот.

Графическое представление алгоритма построения формул интенсиональной логики на основе дерева синтаксических категорий представлено на рисунке 2.

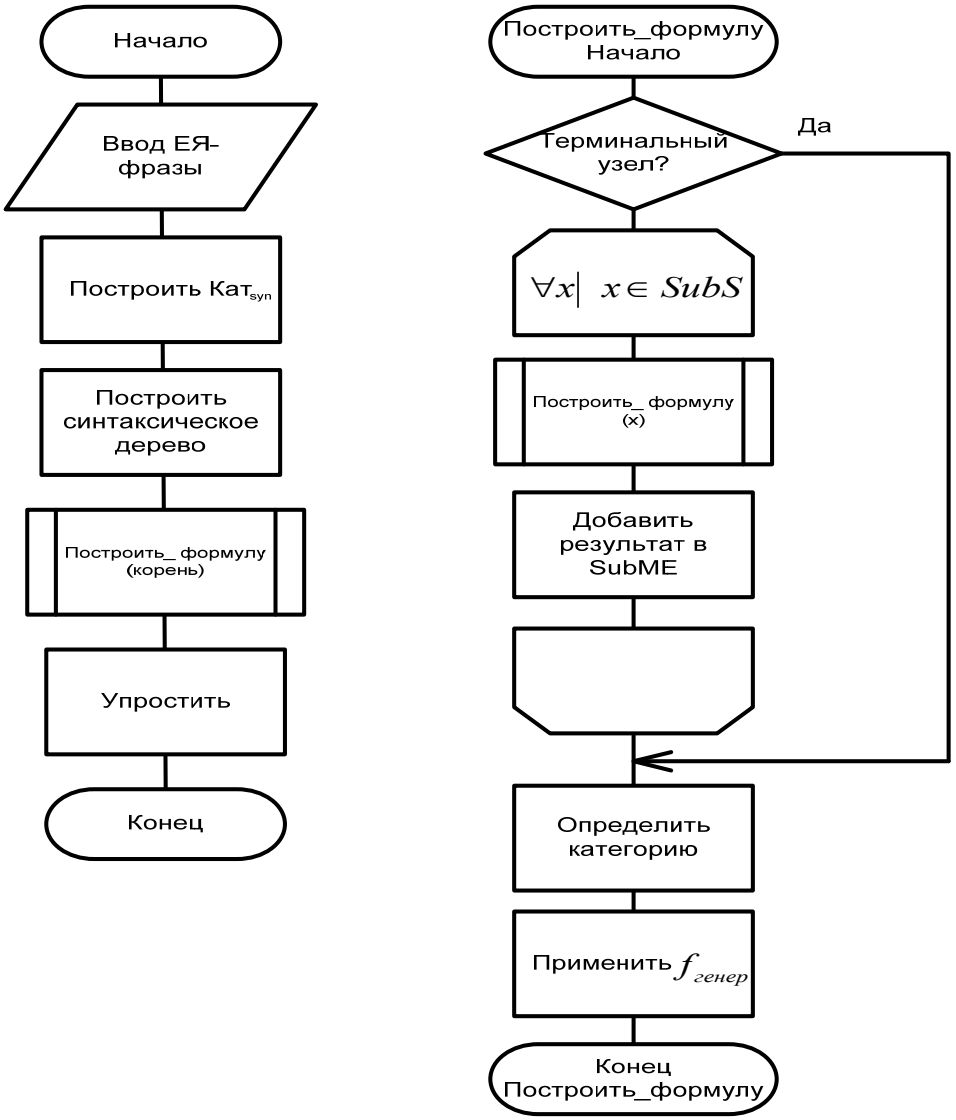

Рис. 2. Графическое представление алгоритма построения формул интенсиональной логики

Основную часть алгоритма занимает функция построения формулы, основанная на рекуррентном обходе дерева для построения подформул узлов. При правильно построенном дереве последовательность действий, описанная в данном алгоритме, всегда конечна.

Результатом данного алгоритма станет не только законченная формула интенсиональной логики, но и дерево категорий, породивших данную формулу. Обходя дерево, можно узнать, о каком действии говорится, кто его совершает (субъект), над чем (объект) и при каких обстоятельствах (место, время и т.д.). Это дерево может использоваться или перестраиваться для получения более адекватных формул на основе дополнительных правил.

Разработка алгоритма осуществляется средствами языка C++ с использованием следующих дополнительных инструментов:

– Microsoft Visual Studio – интегрированная среда разработки (IDE) программного обеспечения и других инструментальных средств;

– Библиотека Strutext – инструмент для обработки текстов на естественном языке на различных уровнях представления. Используется для разработки алгоритма первичной обработки исходного текста. При выделении множества категорий и подкатегории необходимо провести морфологический анализ всего текста. Для построения множеств, форм слов необходимо проверить каждое слово. В библиотеке реализована функция поиска по встроенному словарю с последующим выделением основы слова и его лексических атрибут;

– My Sql Community – система управления реляционными базами данных. Используется для реализации иерархической структуры при построении дерева множеств форм слов и их дополнительных характеристик;

– Библиотека Boost – набор библиотек для языка программирования C++. Используется как вспомогательное средство для расширенной работы с базами данных.

Анализ текстов на естественном языке позволяет не только выделить основные смысловые части текста, но и представить их в интуитивно понятном человеку графическом виде. Рассмотренный алгоритм планируется применять для анализа естественно-языковых текстов в корпоративных информационных системах предприятий и организаций.

Литература

1. Gantz, J. Extracting Value from Chaos / J. Gantz, E. Reinsel // IDC's Digital Universe Study, sponsored by EMC. – 2011. – 12 P.

2. Бах, Э. Неформальные лекции по формальной семантике : перевод с английского / под редакцией О. А. Митрофановой, О. В. Митрениной ; предисловие Б. Парти. – Москва : ЛИБРОКОМ, 2010. – 224 с.

3. Летовальцев, В. И. Мультиагентная система поиска информации на промышленном предприятии / В. И. Летовальцев, А. Н. Швецов // Программные продукты и системы. – 2012. – № 2. – С. 62 – 67.

A.N. Shvetsov, A.V. Kolosov

AUTOMATIC INFORMATION PROCESSING BASED ON FORMAL SEMANTICS

The article is devoted to the problems of text analysis. Considering the increase in the amount of information, it becomes necessary to develop tools for filtering, forming a semantic portrait, as well as navigating through databases and obtaining the necessary textual information.

Information search, processing of natural language texts, development of a semantic analysis system