Социальные медиа

Прикладные задачи анализа. Проблематика.

Макаренко А.В.

avm.science@mail.ru

Научно-исследовательская группа
«Конструктивная Кибернетика»
www.rdcn.ru

Институт проблем Управления
Российская академия наук
Москва, Россия

09 декабря 2015 г.

❑ План мини-курса «Социальные медиа»

Мини-курс «Социальные медиа»:

Лекция x.1. Большие Данные. Цели, задачи и парадигма анализа.

Лекция x.2. Дескриптивный и структурный анализ. Подходы и Методы.

Лекция x.3. Прикладные задачи анализа. Проблематика.

❑ План III лекции

1. Прикладные модели и области применения результатов анализа.

2. Проблемы построения объясняющих моделей.

3. Проблемы построения предсказывающих моделей. Горизонты прогноза.

4. Организация вычислительного процесса по анализу социальных медиа.

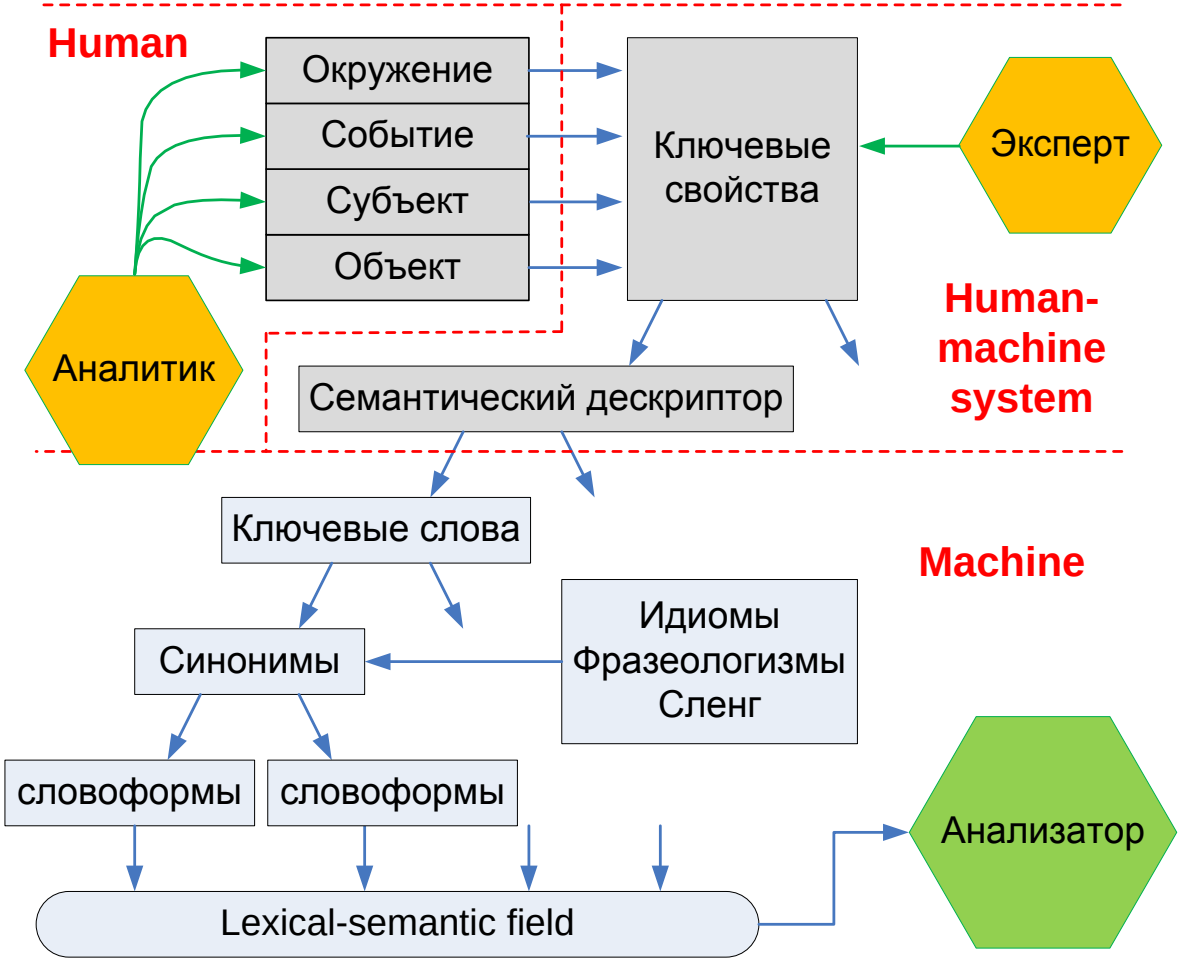
5. Организация бизнес процесса по анализу социальных медиа.

6. Возможности языка R в прикладном анализе социальных медиа.

7. Задания для самостоятельной проработки.

❑ Прикладные модели и области применения результатов анализа

Семантический дескриптор – спецификатор «смысла»



R. Jackendoff, Semantic Structures, MIT Press, Cambridge, MA, 1990.

J. Euzenat, Ontology Matching, Springer-Verlag, Berlin Heidelberg, 2007.

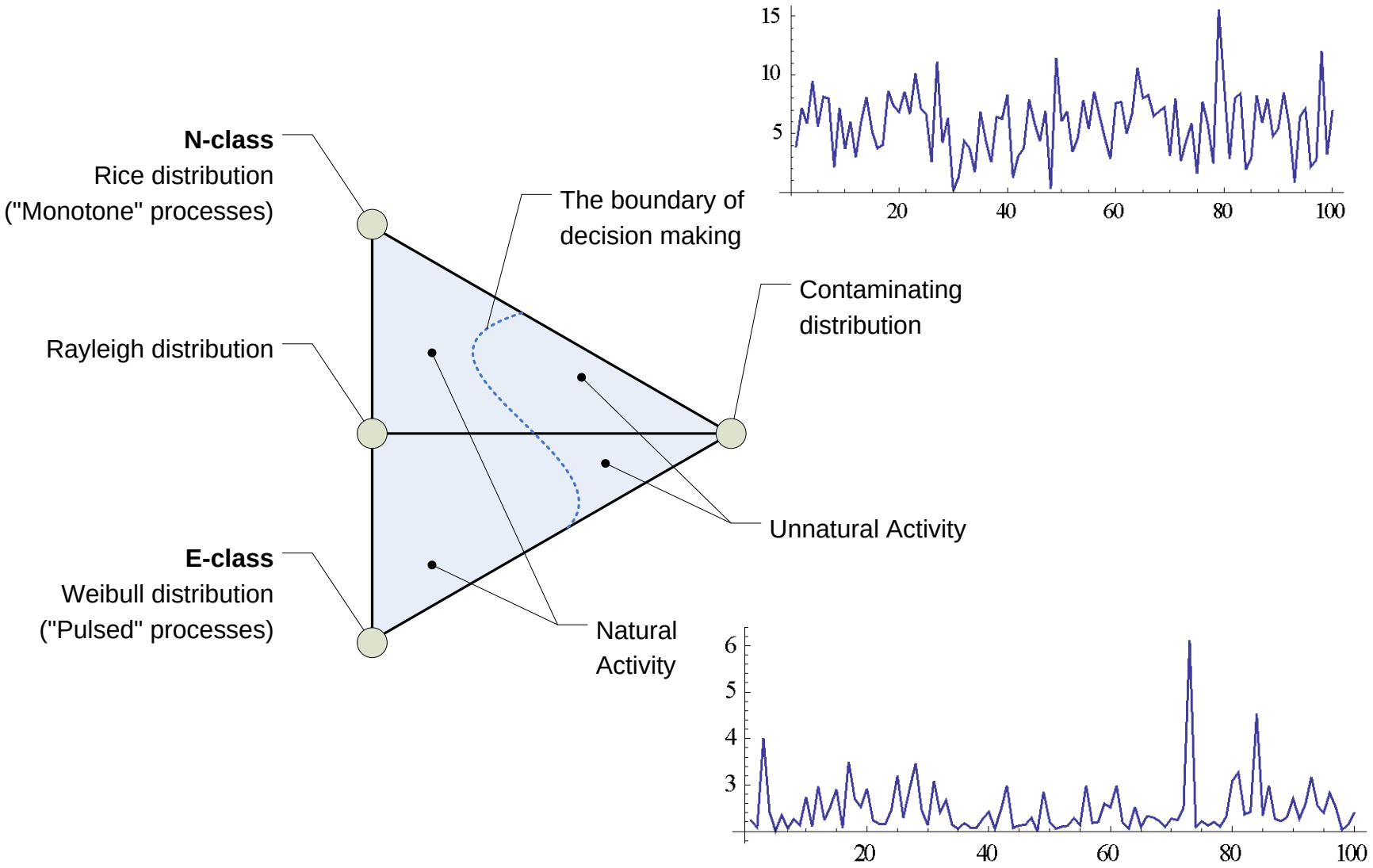
Образцы идиом:

“Дать дуба” <умереть>

“ходить вокруг да около” <намекать>

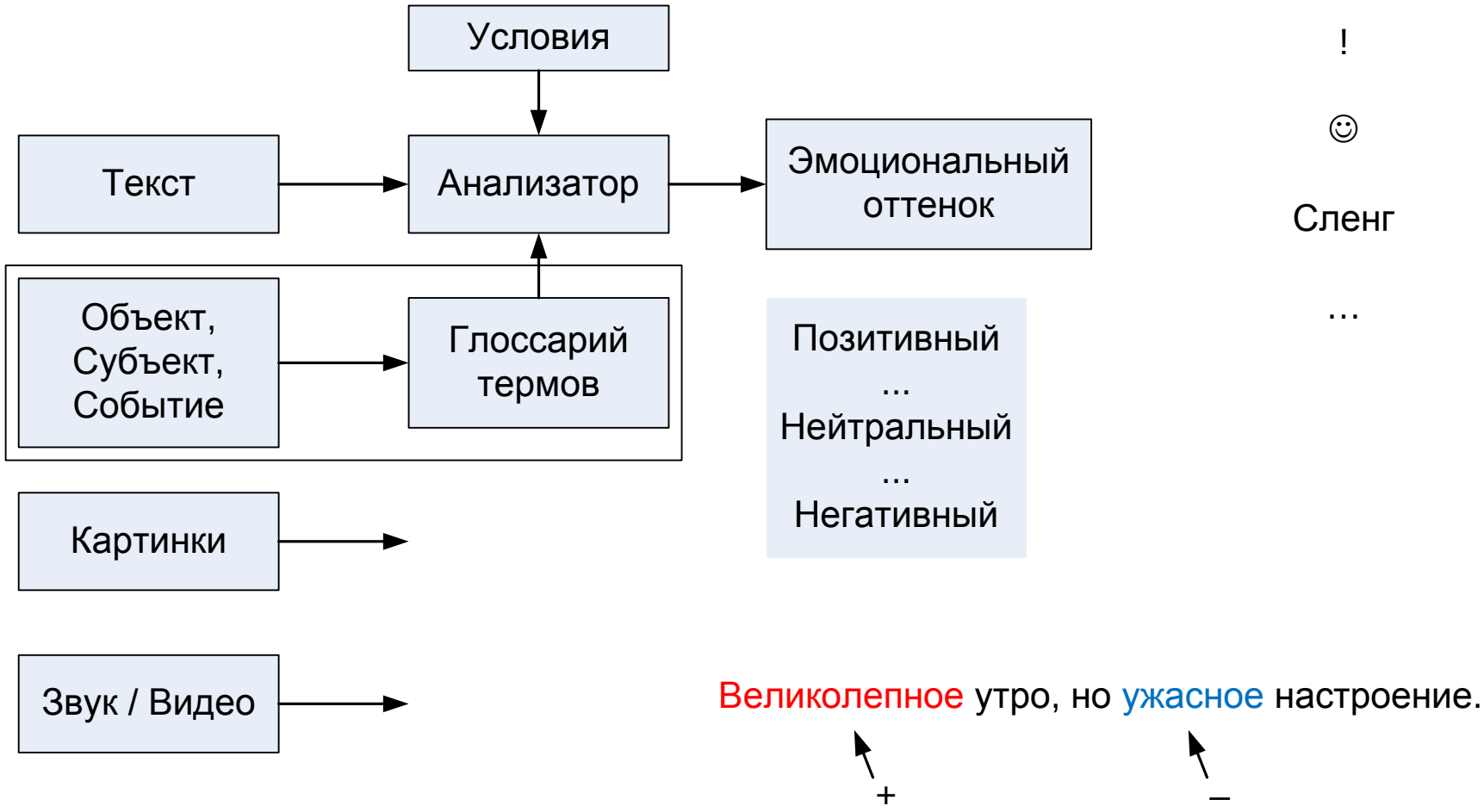
❑ Прикладные модели и области применения результатов анализа

Естественные / Неестественные информационные процессы



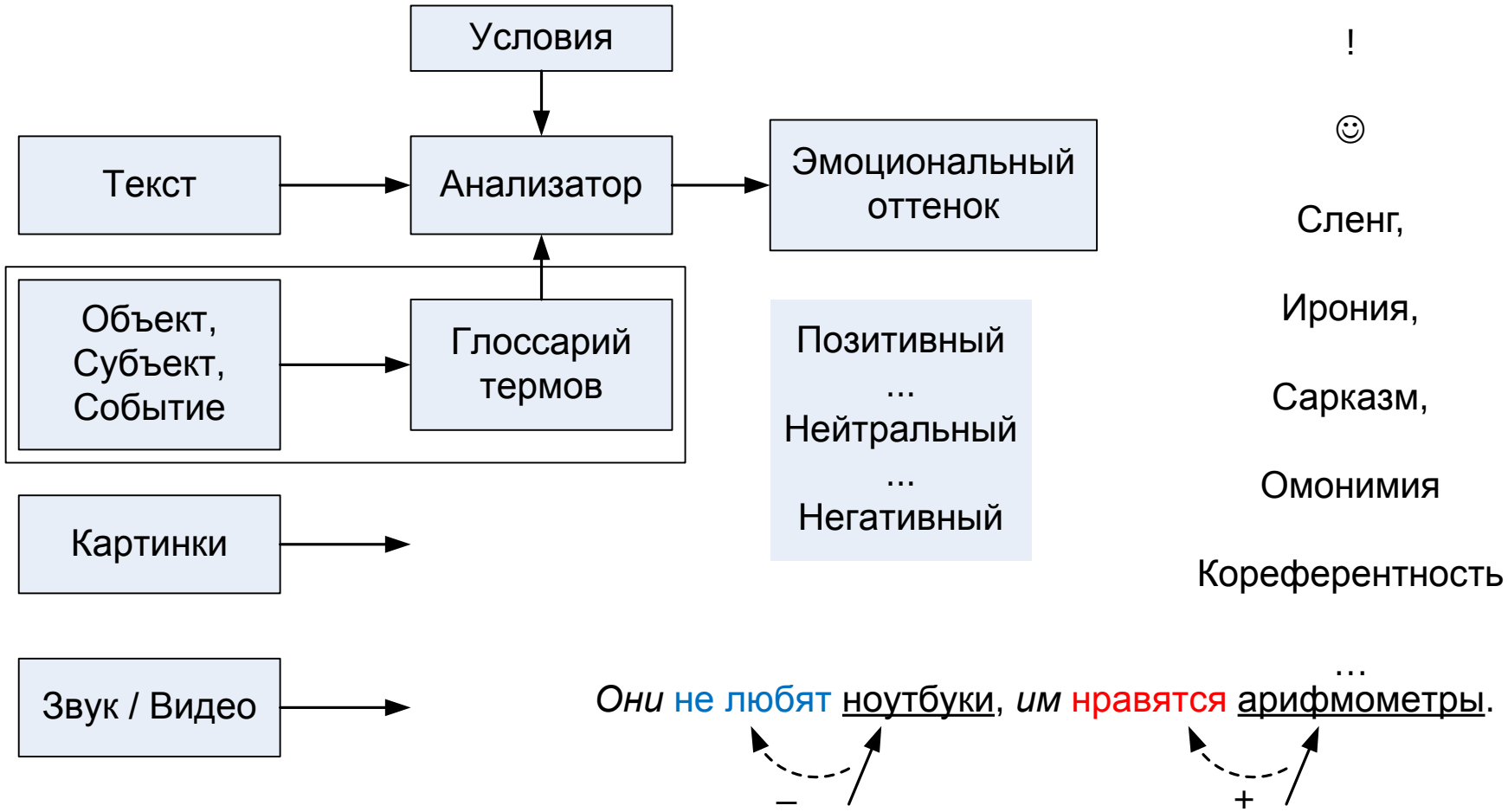
❑ Прикладные модели и области применения результатов анализа

Эмоциональная оценка содержания информационных блоков



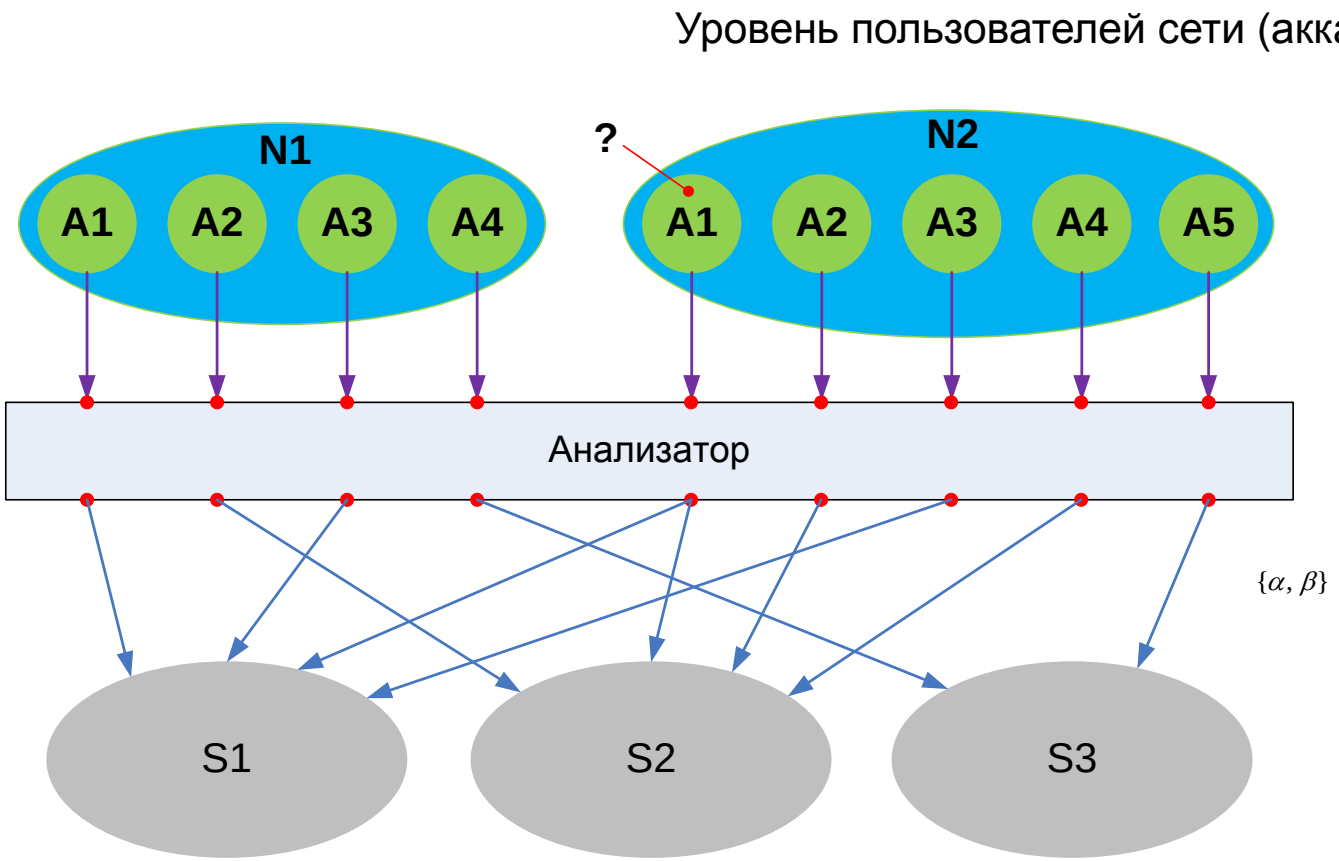
❑ Прикладные модели и области применения результатов анализа

Отношение к Объекту, Субъекту, Событию



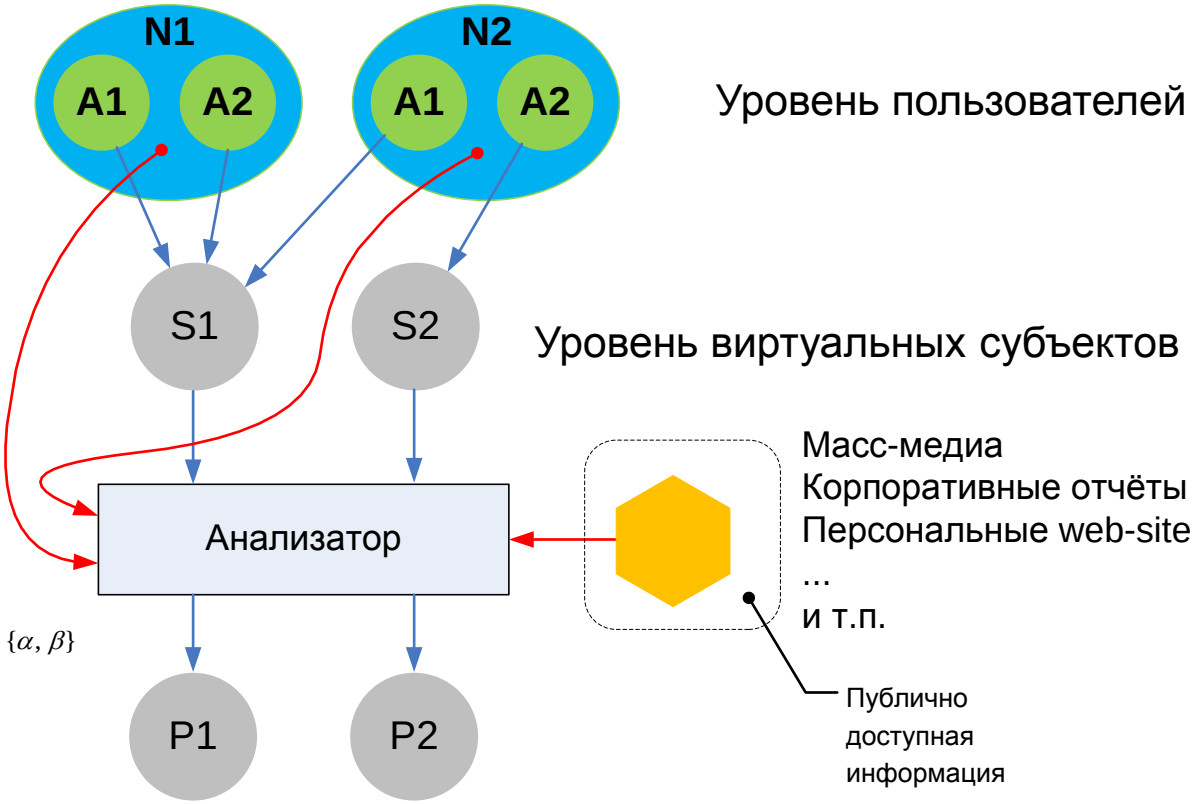
❑ Прикладные модели и области применения результатов анализа

Идентификация пользователей



❑ Прикладные модели и области применения результатов анализа

Де-анонимизация пользователей



Уровень пользователей сети (аккаунтов)

Уровень виртуальных субъектов

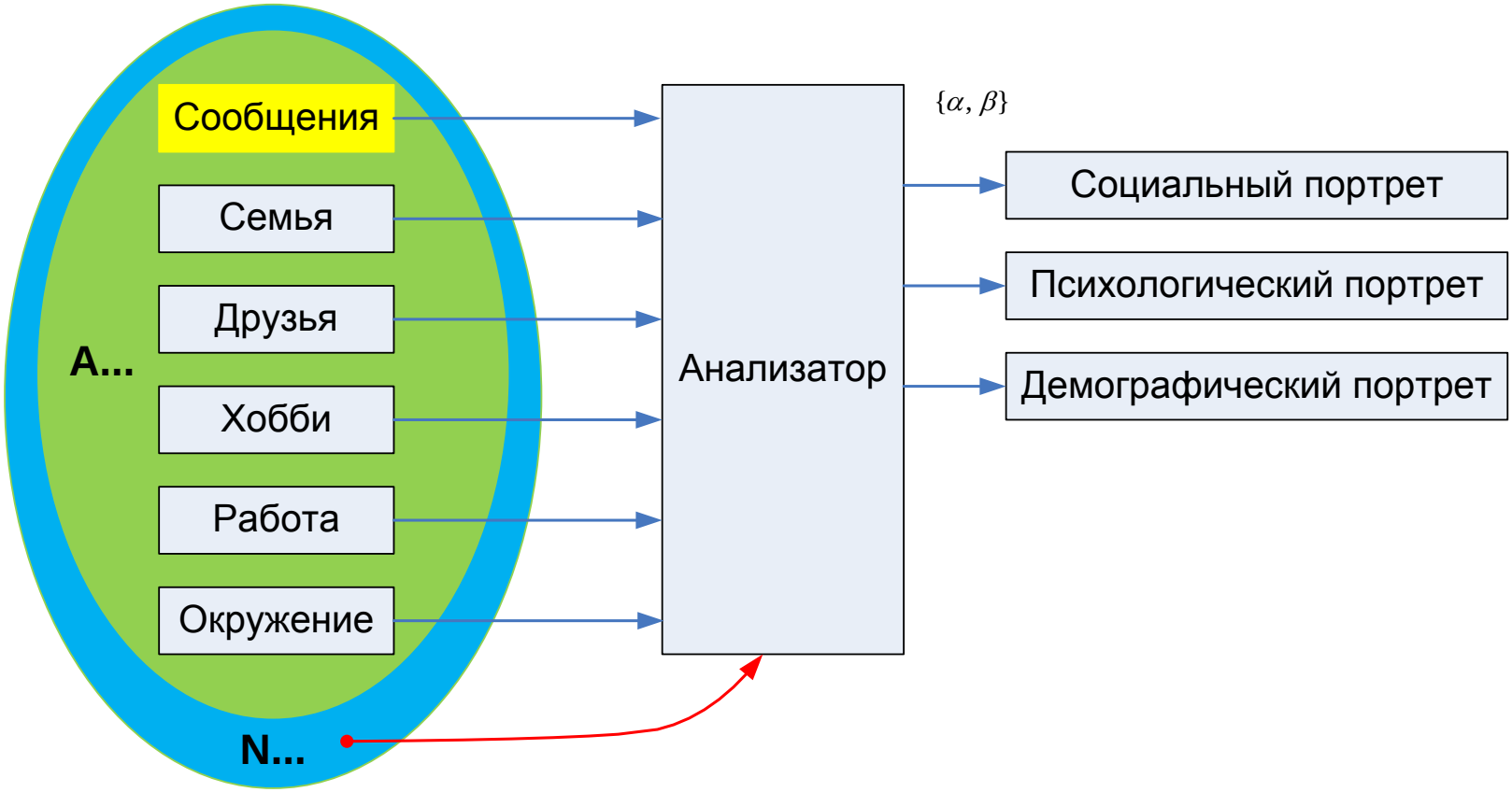
Уровень реальных субъектов

Читать:

E-Á. Horvát, M. Hanselmann, F.A. Hamprecht, K.A. Zweig, PLoS ONE, (2012), 7(4): e34740.

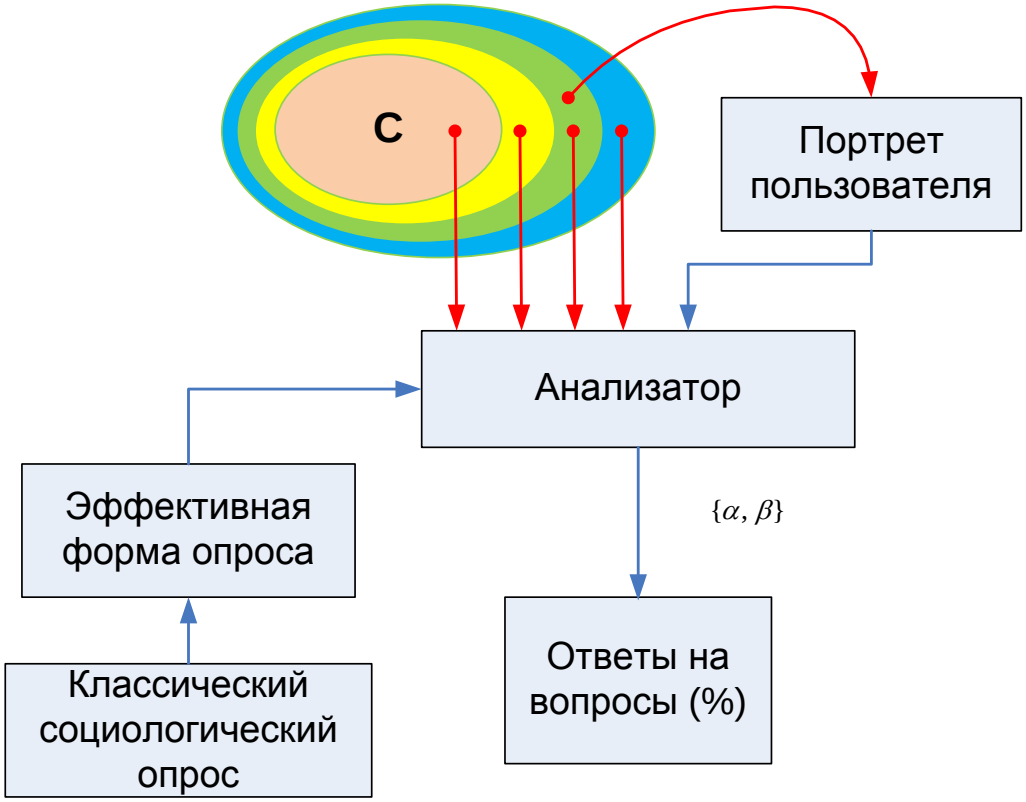
❑ Прикладные модели и области применения результатов анализа

Социально-психологический и демографический портрет пользователя



❑ Прикладные модели и области применения результатов анализа

Виртуальные социологические опросы в социальных медиа



❑ Прикладные модели и области применения результатов анализа

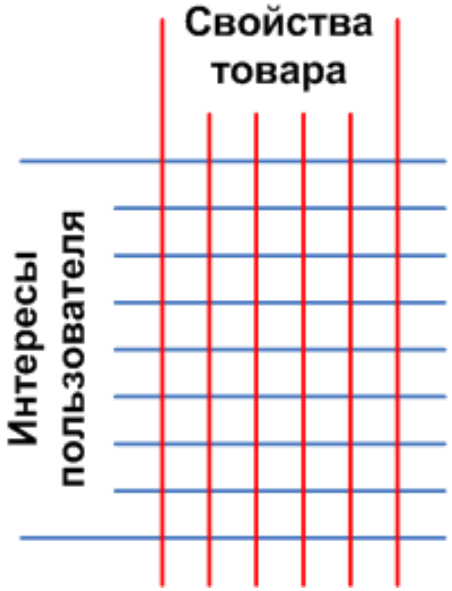
Рекомендательные системы

Рекомендательные системы (сервисы) – это компьютерные программы, которые пытаются предсказать, какие объекты будут интересны пользователю, имея определённую информацию о его статусе, предпочтениях, мнениях, социально-демографическом и психологическом портрете.

При выборе требуется сужение (формализация) области интересов.

Основные стратегии принятия решений:

- Фильтрация содержимого
- Коллаборативная фильтрация



Основные алгоритмические элементы:

- Семантический дескриптор
- Отношение к Объекту, Субъекту, Событию
- Социально-психологический и демографический портрет пользователя
- Виртуальные социологические опросы в социальных медиа

❑ Проблемы построения объясняющих моделей

Назначение математической модели

1. Объяснение – понимание этого явления.

2. Описание – формализация представления явления.

3. Предсказание – выход нового знания.

❑ Проблемы построения объясняющих моделей

Возможная структура модели

Читать:

I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2011.

S.H. Strogatz, Nonlinear Dynamics And Chaos: With Applications To Physics, Biology, Chemistry, And Engineering (Studies in Nonlinearity), Westview Press, 2001.

$$\mathrm{M}^t = \{A_1, A_2, A_1, A_3, \ldots\} \,|\, t \qquad \mathrm{R} = \left\{ \{\mathbf{s}_k\}_{k=1}^K, \{t_k\}_{k=1}^K \right\} \qquad \Gamma^t = \langle \mathrm{V}, \mathrm{E} \rangle \,|\, t$$

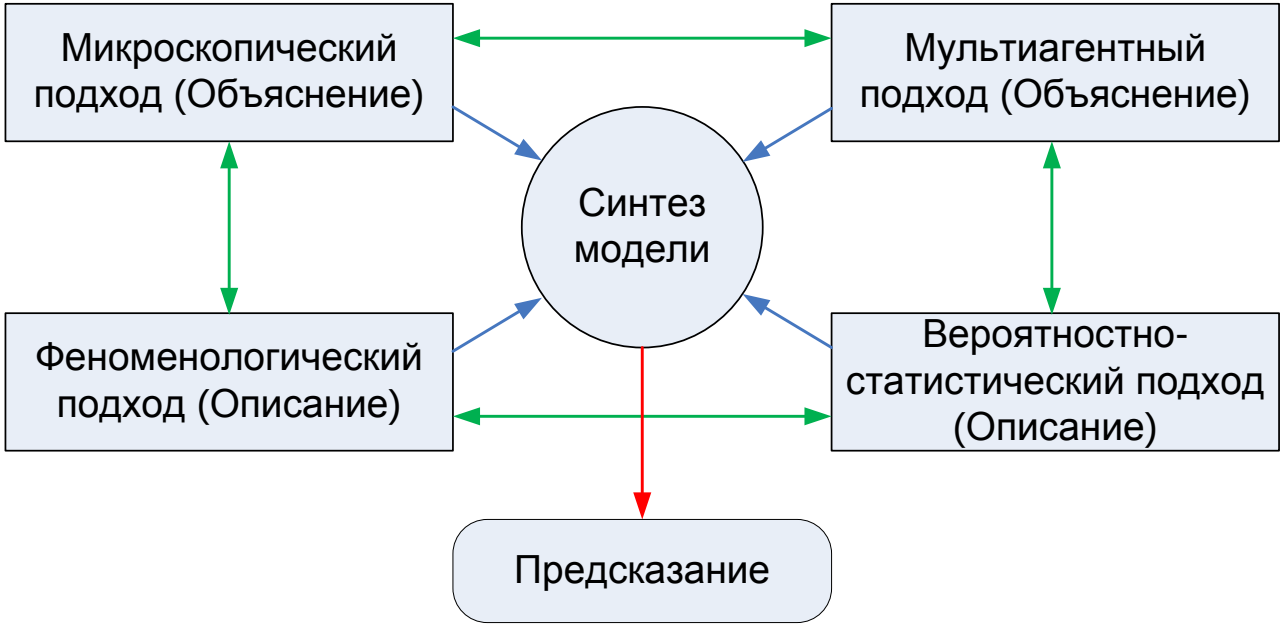
❑ Проблемы построения предсказывающих моделей. Горизонты прогноза.

Методология построения модели

1. Объяснение – понимание этого явления.

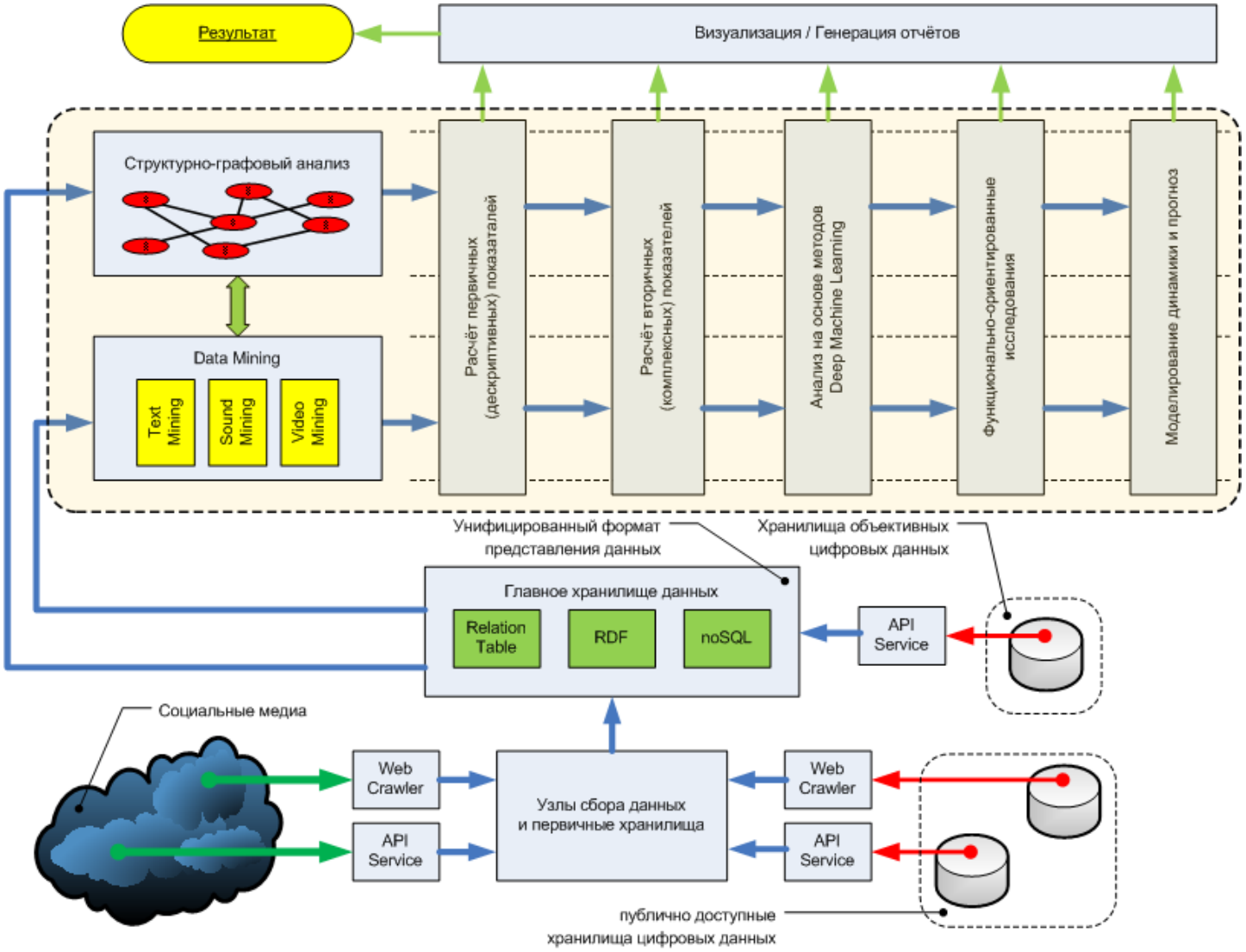
2. Описание – формализация представления явления.

3. Предсказание – выход нового знания.



❑ Организация вычислительного процесса по анализу социальных медиа

Архитектура системы



❑ Организация бизнес процесса по анализу социальных медиа

Быстрый старт

Для создания прототипа аналитической системы потребуется:

1. Источник данных.
2. Язык программирования.
3. Простейшее хранилище данных (Data Base).
4. Библиотека для анализа графов (Graph Analysis).
5. Библиотека для компьютерной лингвистики (NLP).
6. Библиотека алгоритмов машинного обучения (Machine Learning).
7. Некоторые знания и умения.
8. Понимание цели анализа.

❑ Возможности языка R в прикладном анализе социальных медиа

Пакет Caret

Пакет Caret [http://topepo.github.io/caret/index.html] является удобной обёрткой над широким набором алгоритмов машинного обучения.

Основные классы функций:

- Структурирование данных
- Нормализация и трансформация данных
- Классификаторы и регрессоры
- Тюнинг моделей
- Ранжирование переменных

Основной недостаток – сокрытие важных метапараметров, и, как следствие, снижение гибкости применяемых алгоритмов.

❑ Возможности языка R в прикладном анализе социальных медиа

Пакет igraph

Пакет *igraph* [http://igraph.org/r/] содержит широкий набор методов для работы с графами.

Основные классы функций:

- Генерация графов
- Импорт/Экспорт графов
- Оценивание метрических и топологических характеристик графов
- Визуализация (динамическая) графов

Основной недостаток – не очень интуитивный интерфейс.

❑ Возможности языка R в прикладном анализе социальных медиа

Пакеты NLP и tm

Пакеты

NLP [https://cran.r-project.org/web/packages/NLP/NLP.pdf]

tm [https://cran.r-project.org/web/packages/tm/tm.pdf]

являются одними из основных для работы с естественным языком (английский). Содержат широкий набор базовых методов компьютерной лингвистики.

Основной недостаток – узкая языковая специализация и слабая поддержка синтаксического и семантического анализа.

Методы компьютерной лингвистики «из коробки» – сильнейшая проблема!

❑ Задачи для самостоятельного решения

Теоретического плана (отчёт в электронном виде):

1. Разработать алгоритмическую архитектуру системы идентификации пользователей социальных медиа на уровне виртуальных субъектов.

Практического плана:

1. Изучить методы функционального программирования на языке R.
2. Изучить подходы к очистке и разведочному анализу данных. Продемонстрировать их реализацию в системе R.
3. Изучить пакет caret R. Написать программу для обучения модели, решающей задачу классификации методом Random Forest.