ISP RAS
Institute for System Programming
of the Russian Academy of Sciences

Автоматическое обогащение неформальной онтологии на основе текстов определенной предметной области

Никита Астраханцев

28 марта 2013

План

- Введение
- Обогащение онтологии
  - Разбиение на этапы
  - Для каждого этапа:
    - Краткий обзор существующих работ
    - Реализованные методы
    - Результаты тестирования
- Заключение

Текстерра

- Система для обработки текстов
- Основана на знаниях, извлекаемых из Википедии
- Приложения:
  - Разрешение лексической многозначности
  - Поиск ключевых слов
  - Семантический поиск (BlogNoon)

Семантический поиск по блогам

BlogNoon machine learning [Explore]

(•) Search posts ( ) Search blogs

Your query: machine learning;

Sorted by **relevance** | date

Rand Fishkin Interviewed by Eric Enge

from Strategic SEO Planning, Technical SEO, Link Building, and PPC - 23.05.2011

new for Google. They are using the aggregated opinions of their quality raters, in combination with **machine learning** algorithms, to filter and reorder the results for a better user experience. That's a mouthful, but essentially what it means is that Google has this huge cadre of human workers who

Machine learning ; Clickthrough rate ; User-generated content ; Facebook ; Grey hat ; Algorithm ; Social graph ;

What Is Machine Learning Good For?

from e-Literate - 06.05.2012

interested in this robot grader technology? And why? [Emphasis added.] This is a classic case of a market gone awry. **Machine learning** is sold as an "efficiency" tool, because there is money in squeezing cost out of education. In and of itself, there's nothing wrong with wanting education to be cost

Machine learning ; Discourse community ; Conversation ; Message ; Data ; System ; Information ;

Your Next Query relevance | bundles

Artificial intelligence
- Machine learning
- Support vector machine
- Fuzzy logic

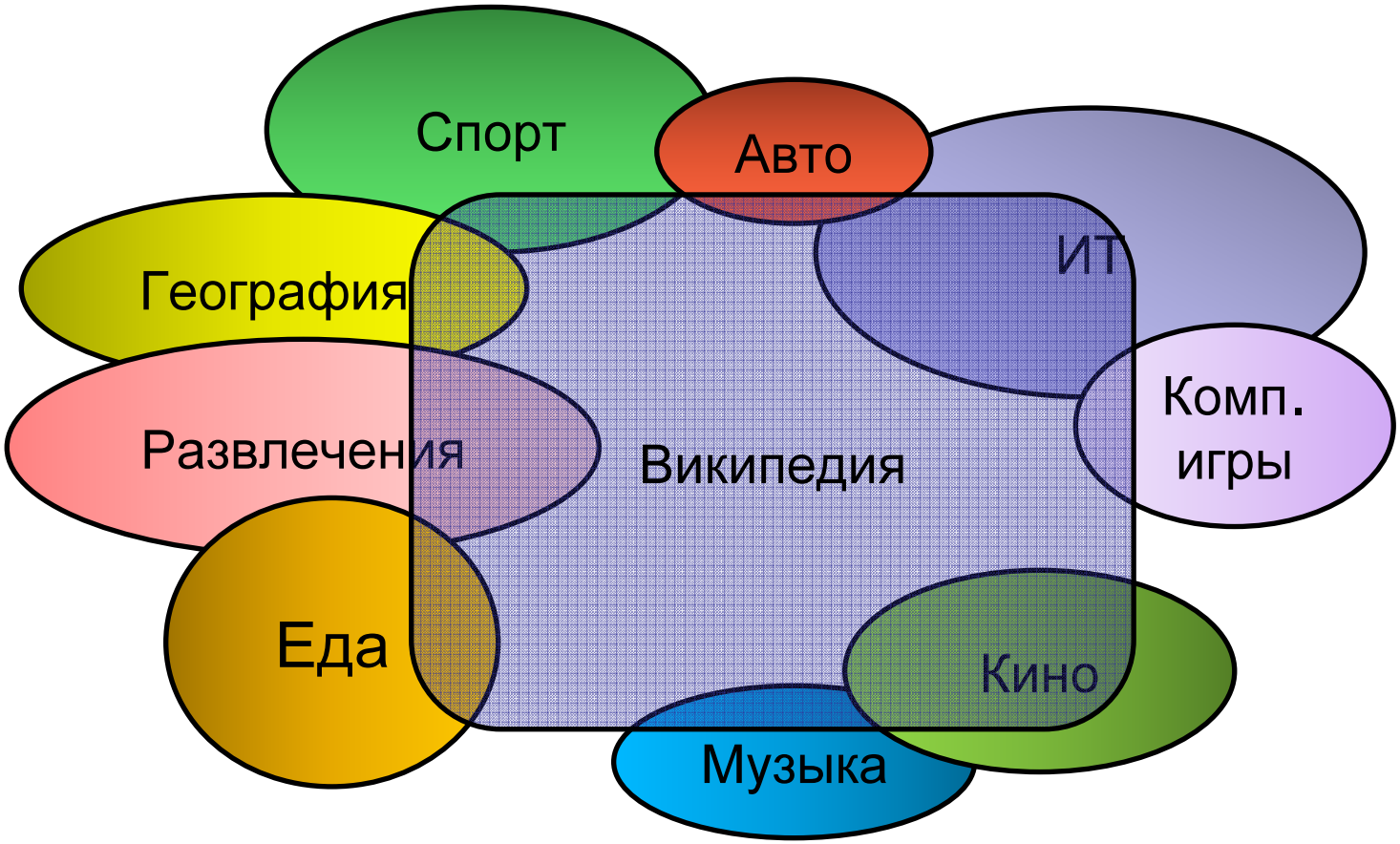
Social networking services
- User-generated content
- Facebook
- Social graph

Stanford University
- Stanford University
- Daphne Koller
- Andrew Ng

Algorithms
- Algorithm
- Markov chain Monte Carlo

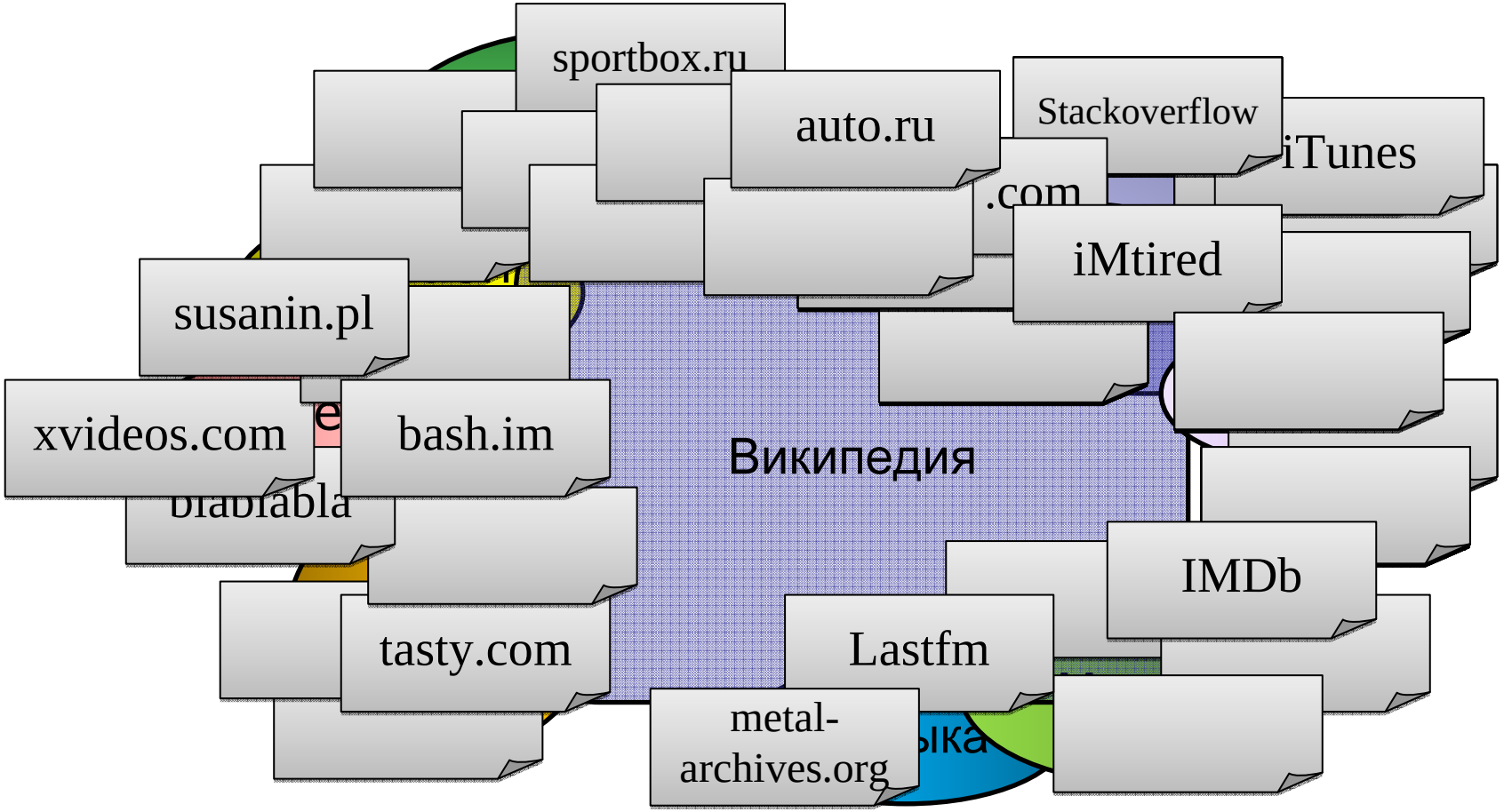
Проблема неполноты

- Википедия покрывает специфичные предметные области не полностью



Проблема неполноты

- Тексты покрывают специфичные области



Попытка построения онтологии

- Построение онтологии на основе текстов определенной предметной области

Попытка построения онтологии

- Построение онтологии на основе текстов определенной предметной области

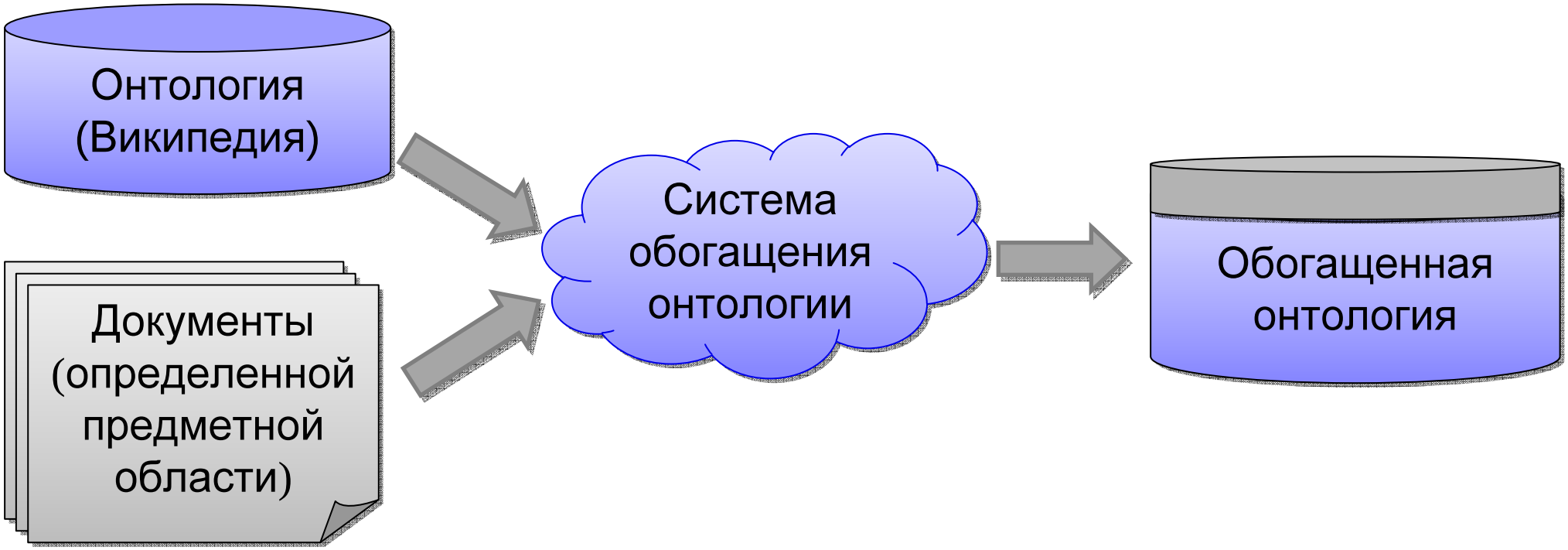
- Точность получившегося решения оказалась недостаточной для практического применения

Причины неудачи

- Понимание текстов, в том числе предметно-специфичных, требует общих знаний
- Общие знания содержатся неявно в самой языковой модели (К. Бименн)
- Необходима внешняя онтология

Постановка задачи

- Обогащение онтологии на основе текстов определенной предметной области



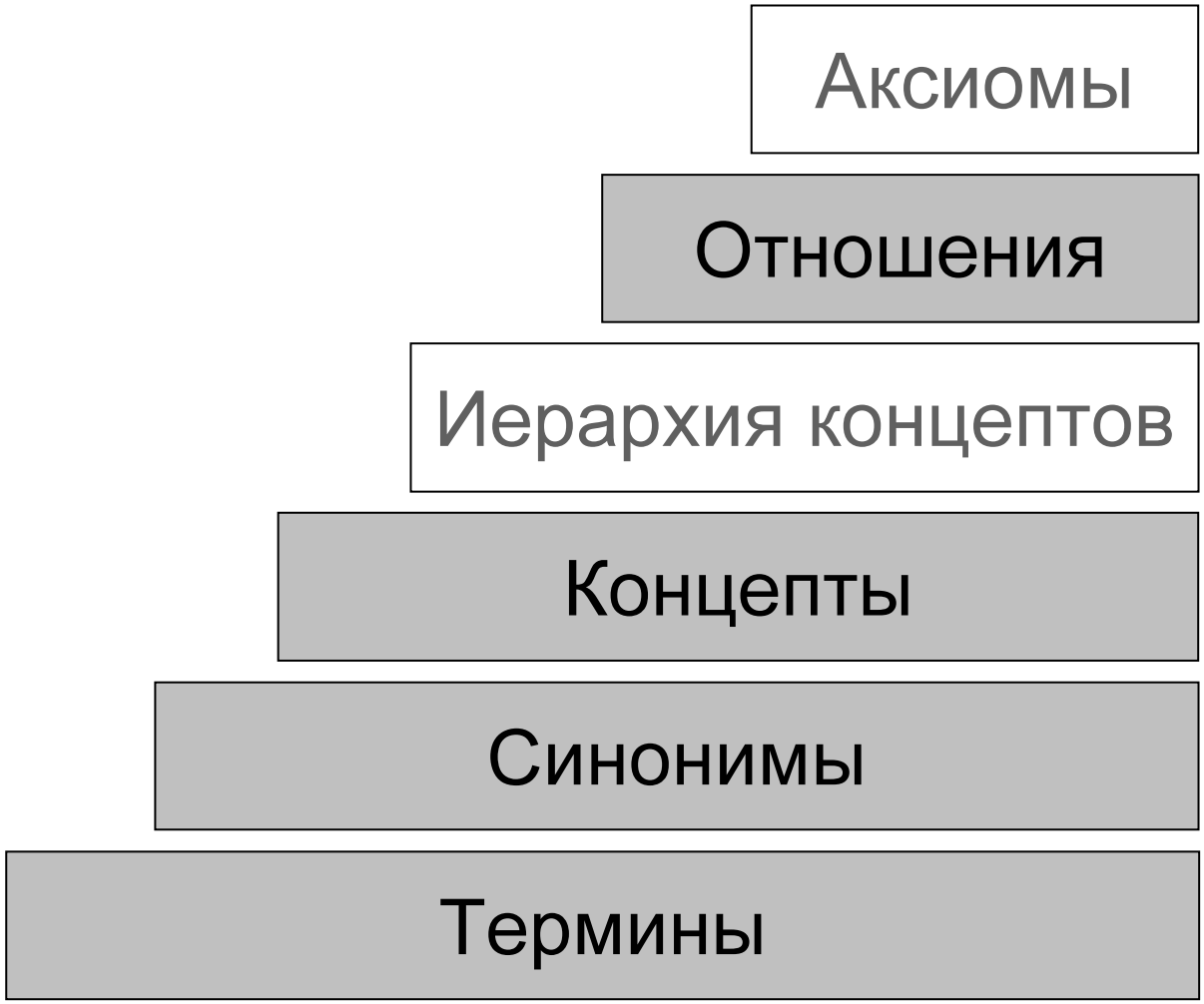
Общее определение

- Онтология – концептуализация предметной области
- Концептуализация — результат выведения понятий из наблюдений, результат формулирования утверждений общего характера
- Концепт – понятие, сущность реального мира
- Термин – текстовое представление концепта

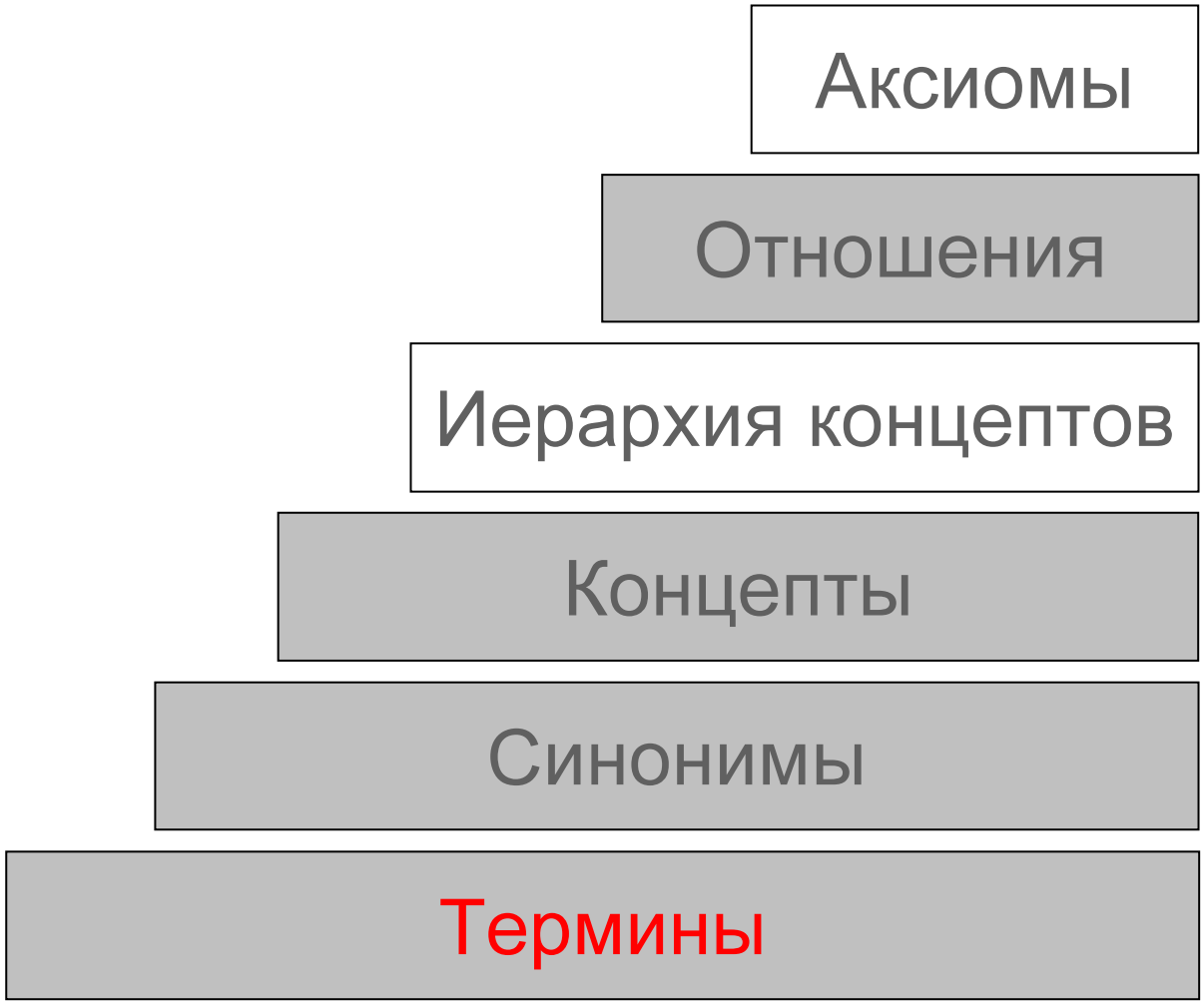
«Слоеный пирог» построения онтологии

- L. Drumond, R. Girardi

Построение онтологии



Построение онтологии



Извлечение терминологии

— «выбор такого множества представлений концептов, которое бы лучшим образом описывало набор текстов определенной предметной области с точки зрения экспертов»

(M. Pazienza, M. Pennacchiotti, F. Zanzotto, "Terminology extraction: An analysis of linguistic and statistical approaches,")

Примеры терминов

- “The United States of America is a large country in North America, often referred to as the "USA", the "US", the "United States", "America", or simply "the States". It has the third largest population among all countries. (…)”

(wikitravel.org/en/United_States_of_America)

Термин

- базовое представление концепта определенной предметной области
- Характеристики термина:
  - Терминоподобность, или предметная специфичность (termhood)
  - Связность (unithood)

Постановка задачи

В заданной коллекции документов определенной предметной области найти термины, отсутствующие в текущей онтологии

Существующие работы

- Лингвистические методы
- Статистические методы
  - Связность
  - Терминоподобность

Лингвистические методы

- Шаблоны частей речи
  - [сущ. сущ.], [прил. сущ.]
- Определенные корни и суффиксы
  - «-ция», «-ость»
- Фильтрация по стоп-словам
  - «этот», «такой», «лучший»

Статистические методы: связность

- T-критерий (T-test)
- Хи-квадрат (chi-square)
- Критерий отношения правдоподобия (log likelihood)

Статистические методы: терминоподобность

- Частота
- Domain Consensus
- Domain Relevance
- Weirdness
- CValue

Выбранный метод

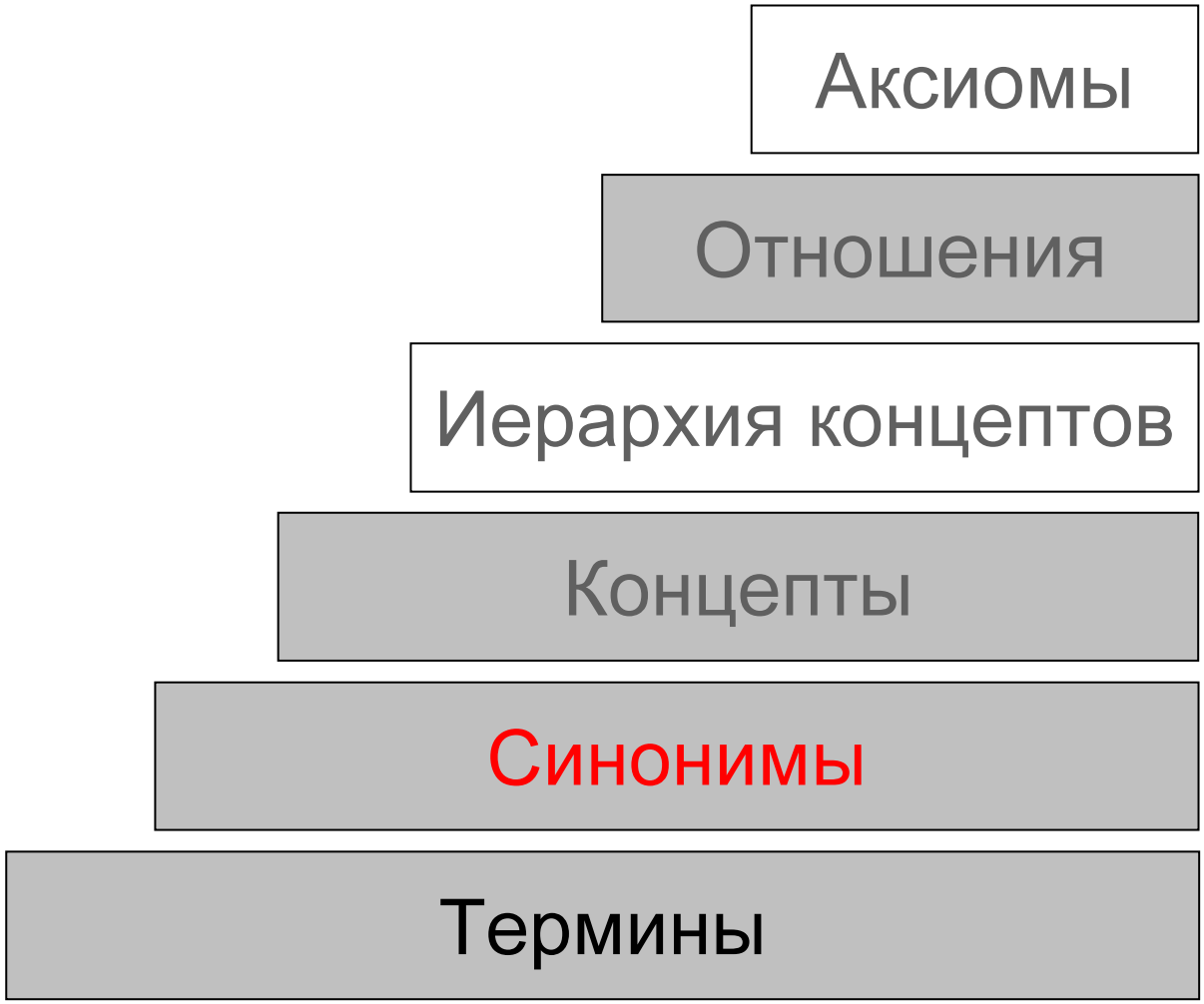
- Кандидаты
  - Именные фразы – OpenNLP Chunker
- Признаки
  - Все вышеперечисленные
- Машинное обучение
  - Метод максимальной энтропии

Результаты

- GENIA – медицинские тексты
  ~2000 статей, 400000 слов, 35000 терминов

Метрика	**Значение**
Средняя точность	0.78
Точность	0.75
Полнота	0.69
Полнота (глобальная)	0.002
F-мера	0.72
Accuracy	0.83

Построение онтологии



Синонимы

- “The United States of America is a large country in North America, often referred to as the "USA", the "US", the "United States", "America", or simply "the States". It has the third largest population among all countries. (…)”

Постановка задачи

- Среди найденных терминов (кандидатов в термины) найти пары, соответствующие одному и тому же концепту

Виды синонимов

1) орфографические:
   - *amino acid — amino-acid*;
2) морфологические:
   - *cellular gene — cell gene — cell genes*;
3) лексические:
   - *carcinoma — cancer*;

Виды синонимов

4) структурные
   a) предложные:
      - *clones of human — human clones*;
   b) согласованные:
      - *RNA polymerases II and III — RNA polymerases II и RNA polymerases III*;
5) акронимы:
   - *DNA — deoxyribonucleic acid*.

Выбранные методы

1) Орфографические
2) Морфологические

} стемминг

3) Лексические
4) Структурные
   a) предложные
   b) согласованные

} Инверсия + словарь

5) Акронимы

} Регулярное выражение + специальные признаки

Обработка предложных синонимов

- Инверсия:

$$(JJ\,|\,NN)_1 \quad NN_1\,\mathrm{IN}\,(JJ\,|\,NN)_2 \quad NN_2$$

$$\downarrow$$

$$(JJ\,|\,NN)_2\,NN_2\,(JJ\,|\,NN)_1\,NN_1$$

"Governor of Louisiana" - "Louisiana Governor”

- Словарь прилагательных для стран:
  - "Queen of **Britain**" - "**British** Queen"

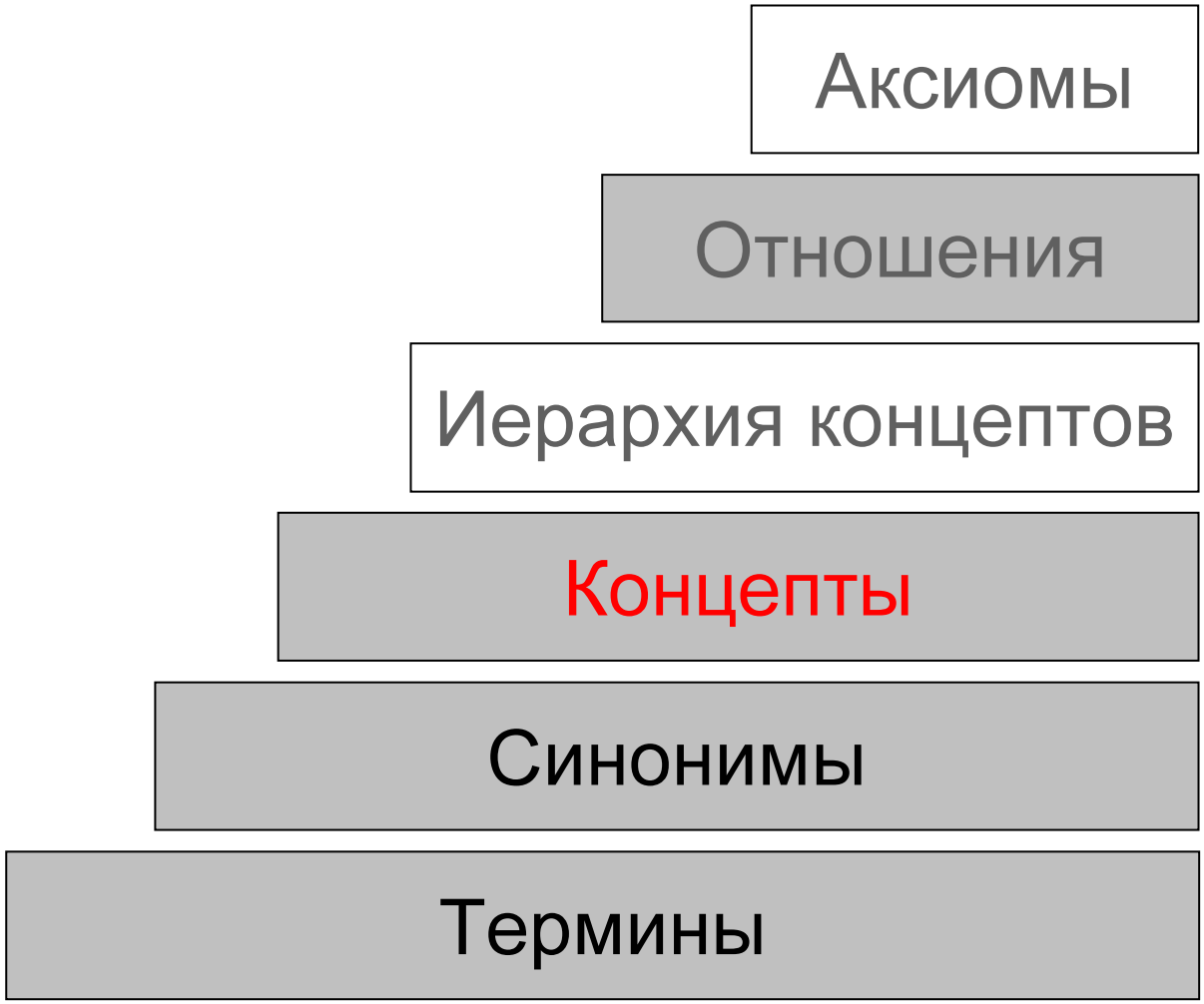
Обработка акронимов

- Регулярное выражение:
  - [A-Z]{2,5}
- Признак, в одном ли документе
- Расстояние между вхождениями терминов в исходном документе

Результаты

- Неразмеченный корпус из 1246 текстов (настольные игры)
- 117 пар предложных синонимов
  - 89% точность
- 33 пары акронимов
  - 100% точность
  - 73% полнота

Построение онтологии

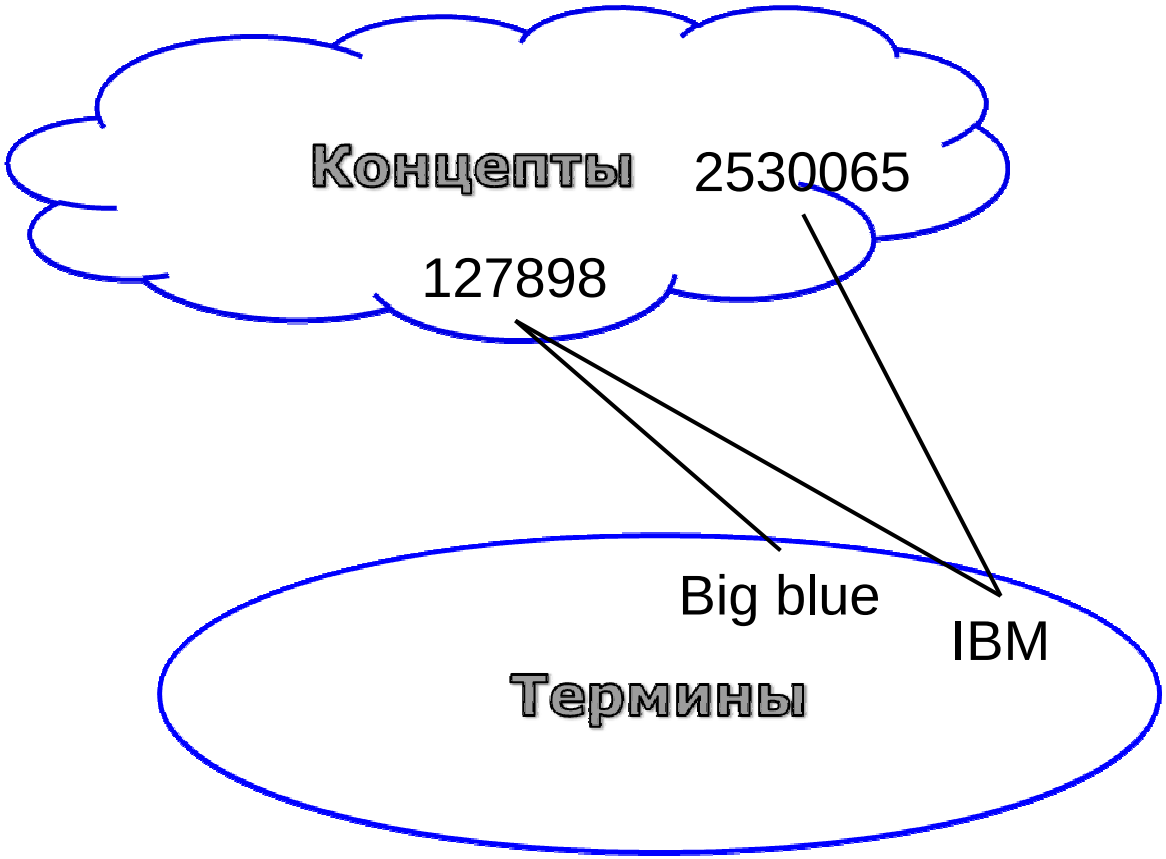


Постановка задачи

- Для найденных терминов образовать концепты

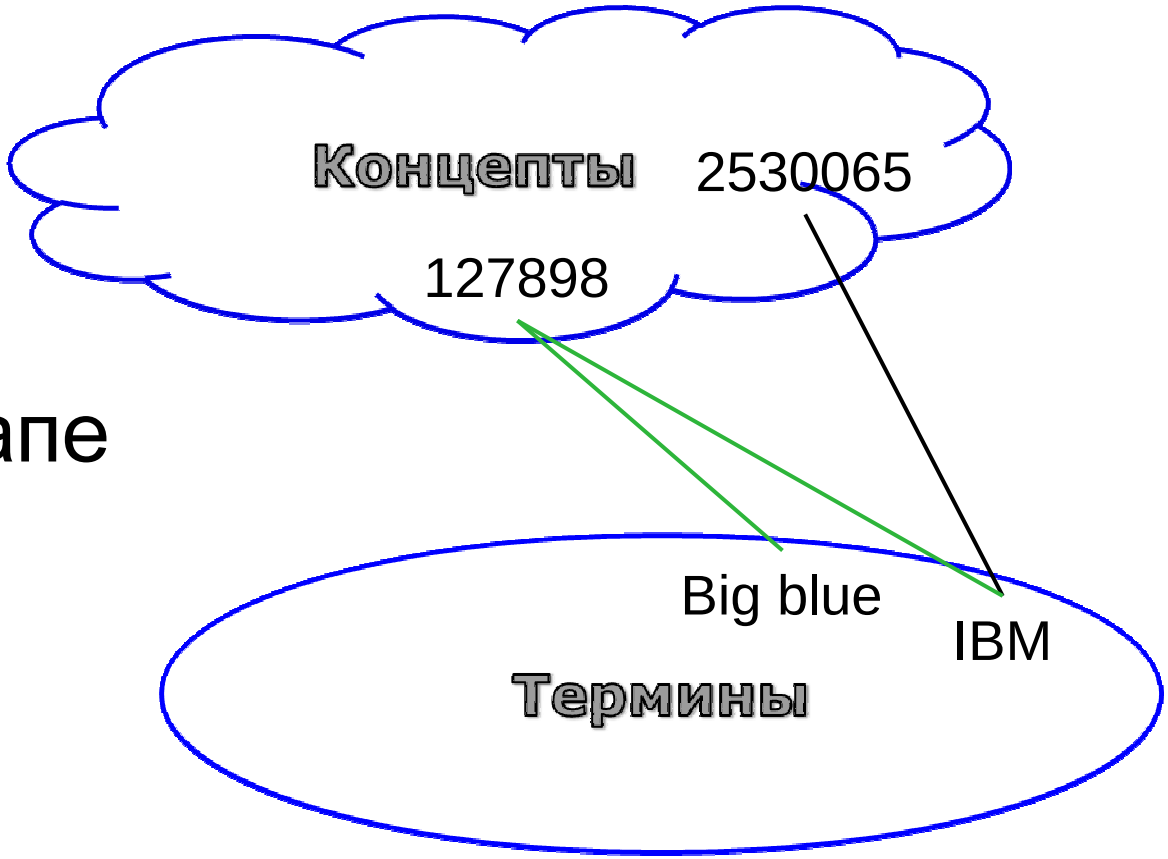
Связь между концептами и терминами

- Термин – концепт: многие ко многим



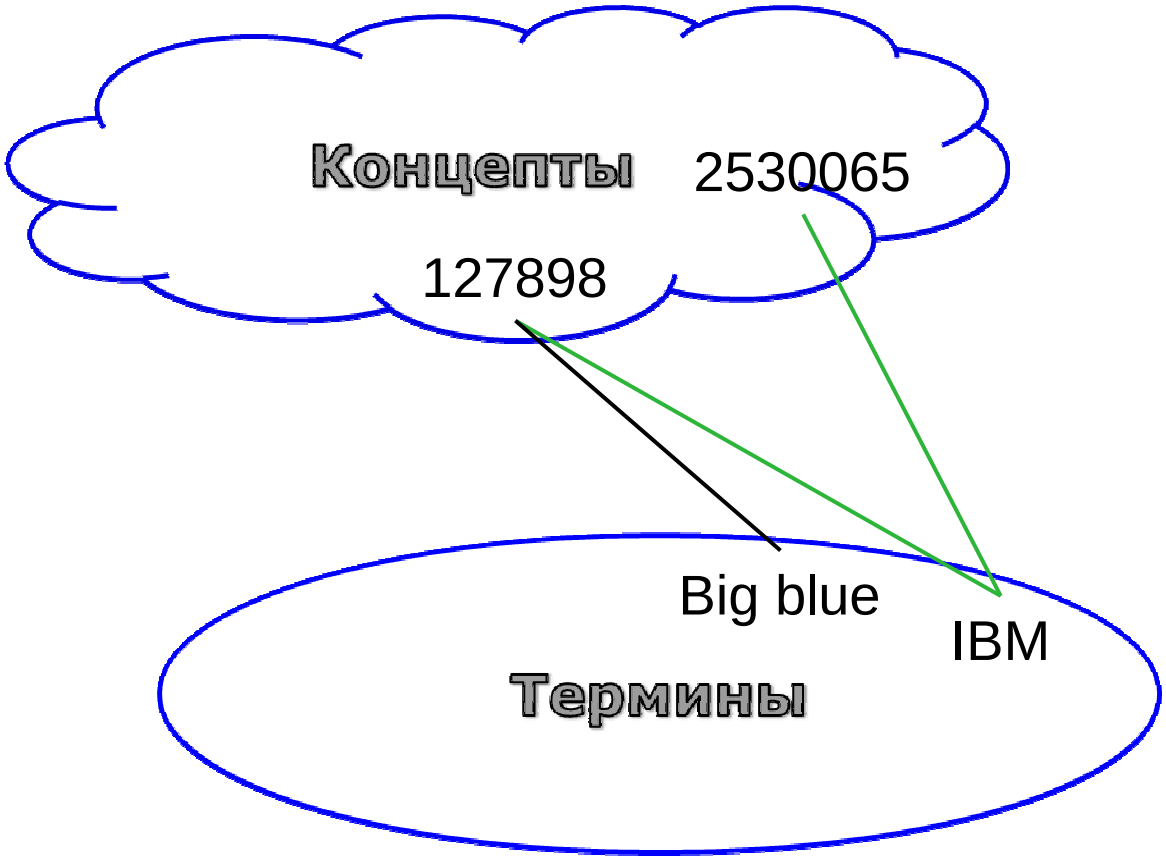
Связь между концептами и терминами

- Термин – концепт: многие ко многим
- Синонимия
  - Решена на предыдущем этапе



Связь между концептами и терминами

- Термин – концепт: многие ко многим
- Синонимия
- Омонимия



Образование концептов

- Для каждого термина:
  - новый термин – новый концепт
  - существующий термин – определить, новый ли концепт

Задача определения новых значений существующих терминов

- Для заданной коллекции документов найти такие термины:
  - существуют в исходной онтологии
  - концепт этого термина характерен для предметной области
  - концепт этого термина отсутствует в исходной онтологии

Определение новых значений слов

"…Using **ML** to solve this problem seems to be breaking a fly upon the wheel…"

- *Millilitre?*
- *Machine learning?*
- *Новое значение: Monolayer?*

Существующие работы

- Отбрасывание концептов по порогу уверенности алгоритма разрешения лексической многозначности
- Определение выбросов (Outlier detection)

Определение выбросов

- Все термины — точки в n-мерном пространстве признаков:
  - контекстные слова
  - части речи
  - именованные сущности

- Если термин достаточно сильно удален от тренировочного множества => он использовался в тексте в ранее неизвестном значении

Выбранный метод

- Бинарная классификация терминов, присутствующих в базе знаний:
  - Термин с существующим значением
  - Термин с новым значением
- Машинное обучение (с учителем)
  - Метод максимальной энтропии

Признаки

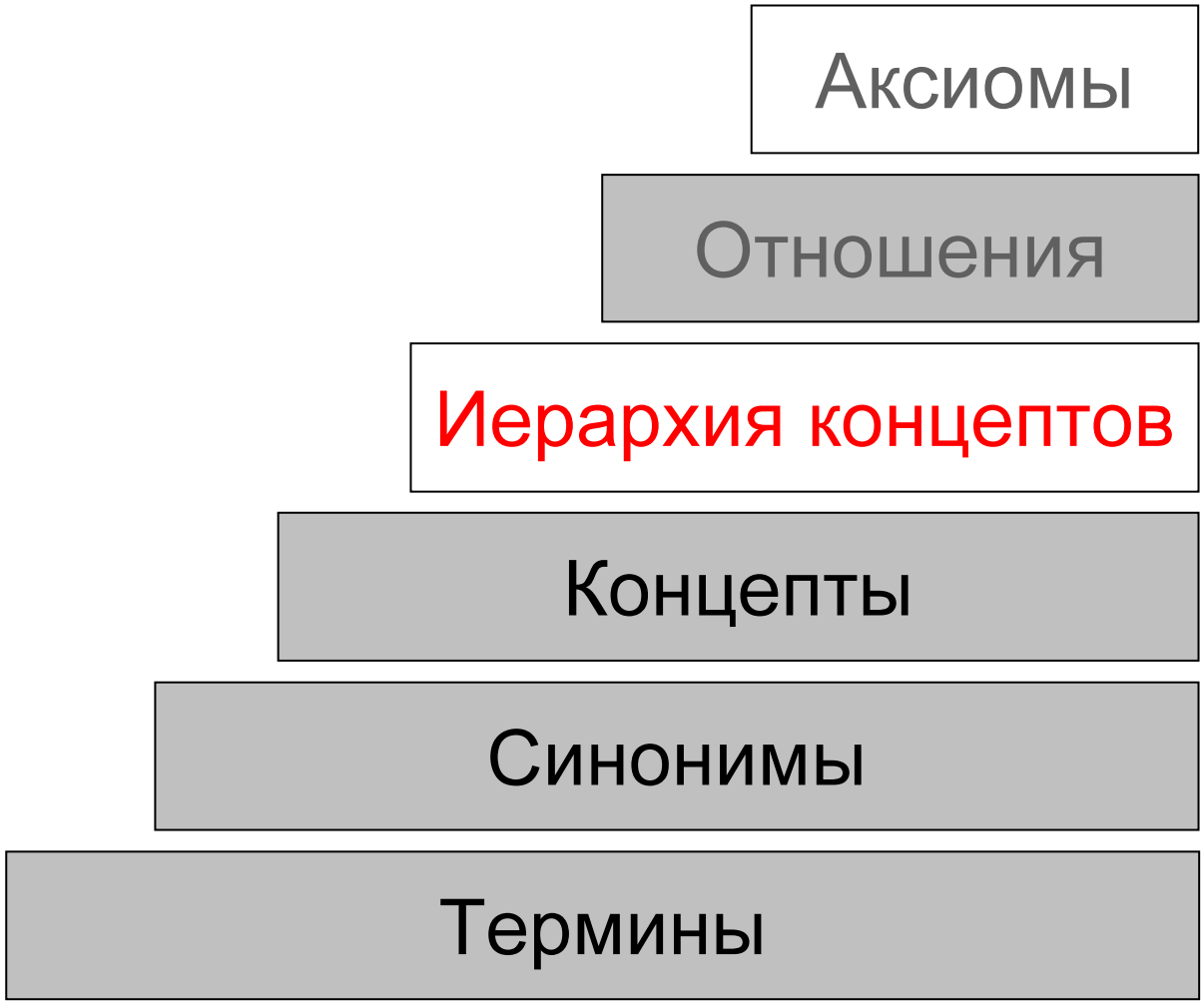
1) Признаки терминов, измеряющие специфичность
2) Максимальная семантическая близость к ключевым концептам предметной области
3) Количество концептов на число вхождений термина
4) Количество выбранных концептов
5) Общее количество концептов в базе знаний

Методология тестирования

- 1246 текстов про настольные игры
- Вручную найдены:
  - 77 примеров с явно новым значением
  - 100 примеров с пограничными случаями

Метрика	Явные	Пограничные
Точность	0.733	0.584
Полнота	0.793	0.604

Построение онтологии



Постановка задачи

- Найти все пары концептов, которые находятся в отношении «является» (IS–A или SubClassOf)
- «Яблоко» — «Фрукт»

Иерархия концептов

- Лингвистические методы
  - Шаблоны вида:

    "NP_1, NP_2 and other NP_3"
- Статистические методы
  - Кластеризация
  - Классификация

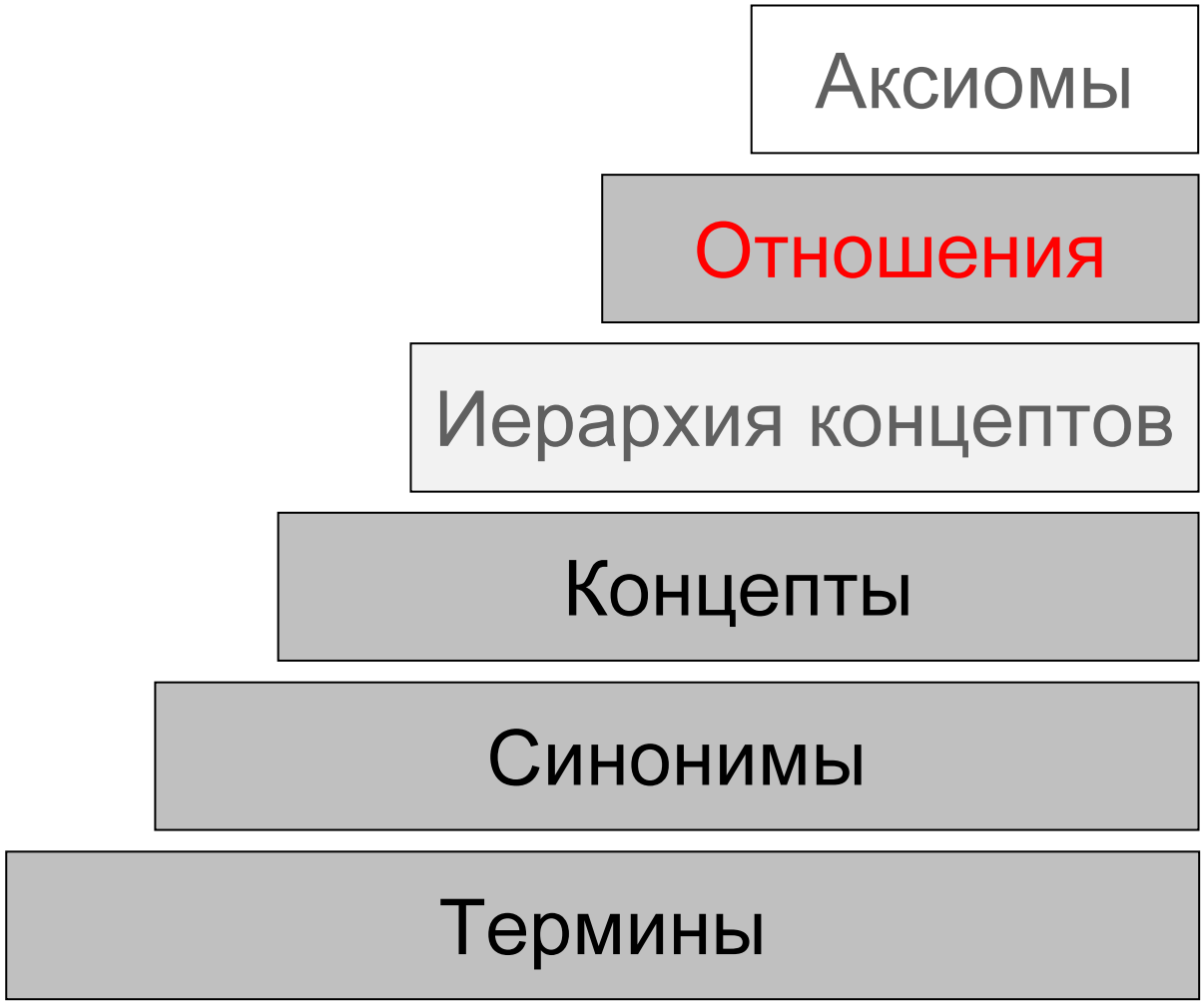
Методы, основанные на кластеризации

- Построение новой таксономии
- Объединение кластеров
  - Одиночное связывание
  - Полное связывание
  - Среднее связывание

Методы, основанные на классификации

- Обогащение существующей таксономии
- Обход дерева и прикрепление к максимально близкой вершине
  - Нисходящий обход
  - Восходящий обход

Построение онтологии

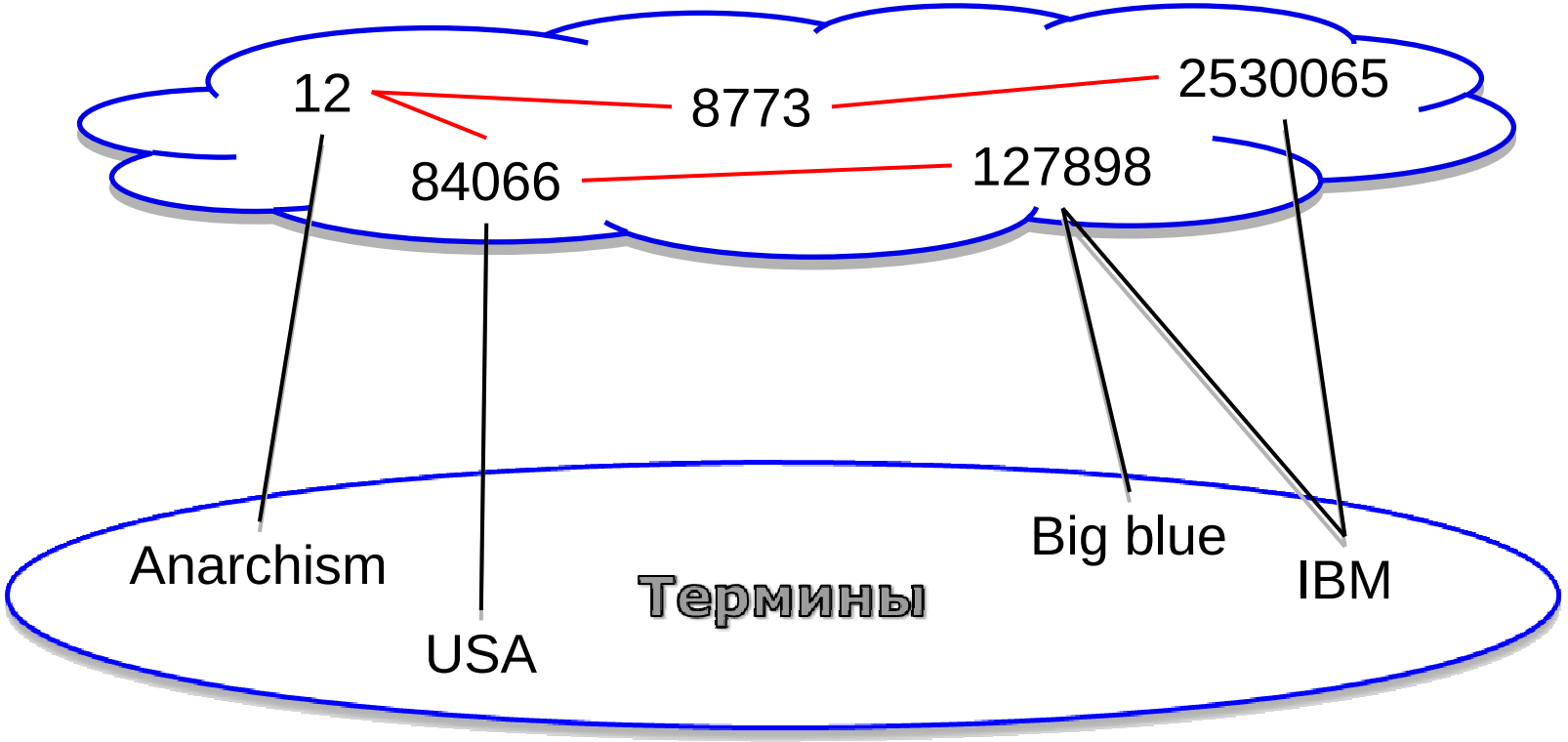


Типы отношений

- Специфичные отношения:
  - "является" (IS–A или SubClassOf)
  - "часть–общее" (PART–OF)
- Семантическая близость
  - высокая: «алый» и «красный»
  - средняя: «табак» и «сигарета»
  - низкая: «велосипед» и «философия»

Онтология Текстерры

- Два концепта связаны ⇔ есть гиперссылка между статьями



Семантическая близость Текстерры

- Вещественная функция от двух концептов, учитывающая связи этих концептов
  - мера Дайса: $$sim(A,B) = 2\frac{|n(A) \cap n(B)|}{|n(A)| + |n(B)|}$$
  - нулевая при отсутствии общих соседей

Постановка задачи

- Для данного (нового) концепта найти связанные концепты, максимально похожие на те, что были бы проставлены в Википедии, если бы в ней существовал данный концепт
- Похожесть в смысле относительной семантической близости

Дистрибутивная семантика

«Слова, встречающиеся в похожих контекстах, имеют близкие значения»

Harris, Z. (1954). "Distributional structure"

Распознавание семантически близких концептов

- Тип контекста
  - текст или лингвистические единицы
- Размер контекста
- Взвешивание частот слов
  - энтропия, взаимная информация
- Уменьшение размерности
  - SVD-разложение, LSI, PLSI
- *Семантическая близость*

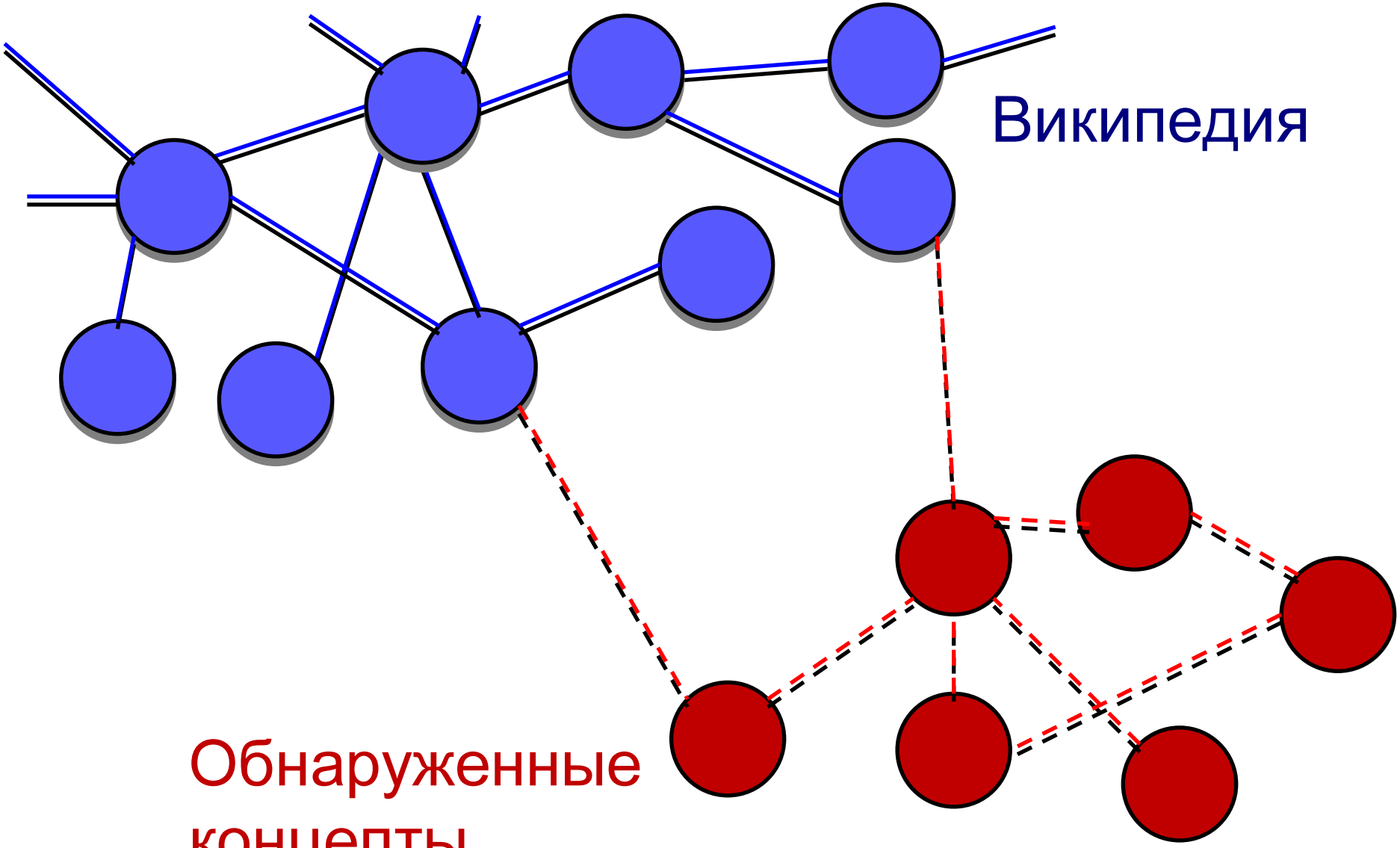
Реализованный метод

- *Вход:* концепт, набор документов
- *Выход:* связанные концепты
- *Алгоритм:*
  - Для каждого термина:
    - Для каждого вхождения:
      - Взять все концепты из **определенного окна** (в мультимножество)
  - **Взвесить** полученные концепты
  - Отбросить по **пороговому** значению

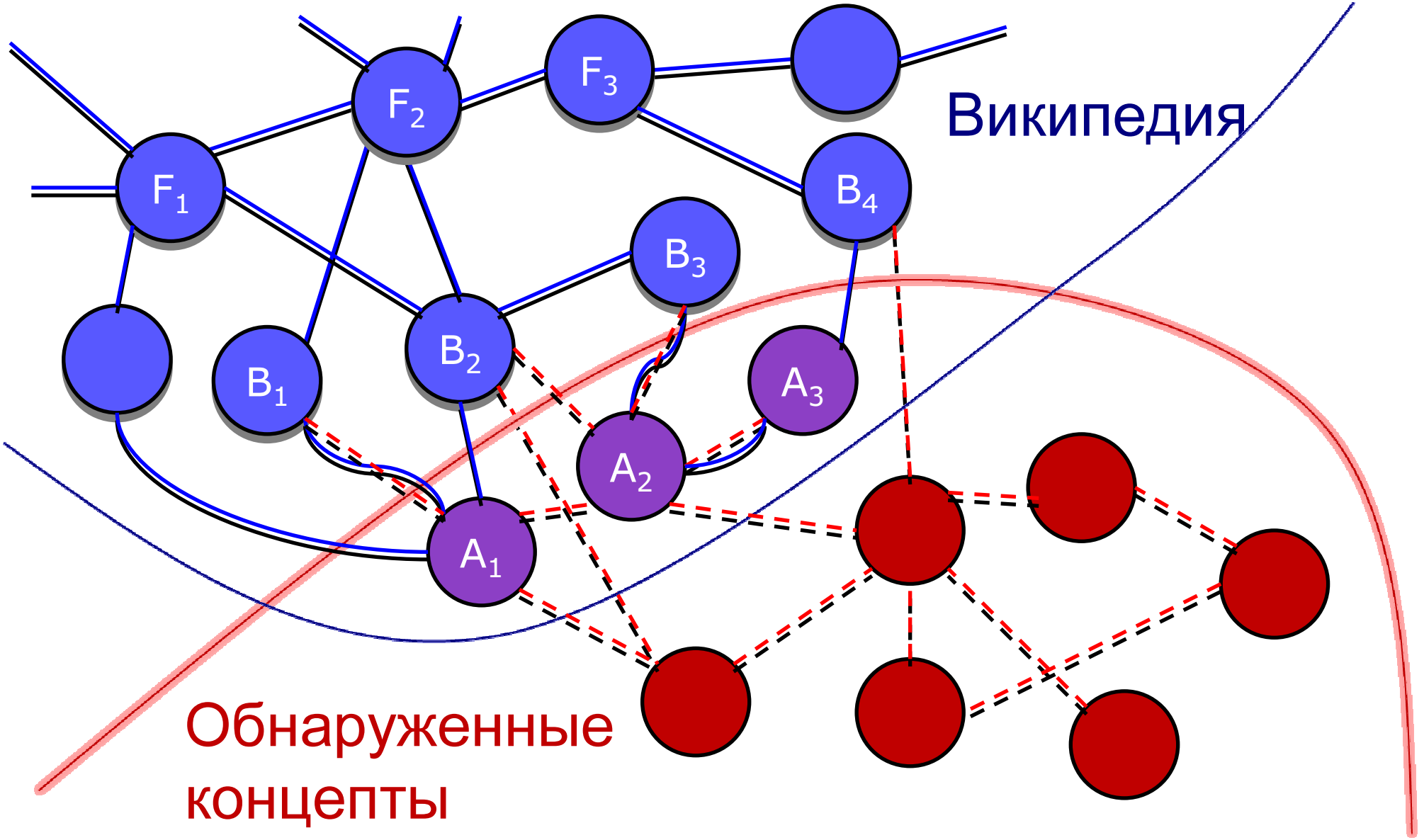
Параметры метода

- Размер окна – число концептов
- Способ взвешивания концептов
  - Частота
  - Т-критерий
- Пороговое значение

Методология тестирования



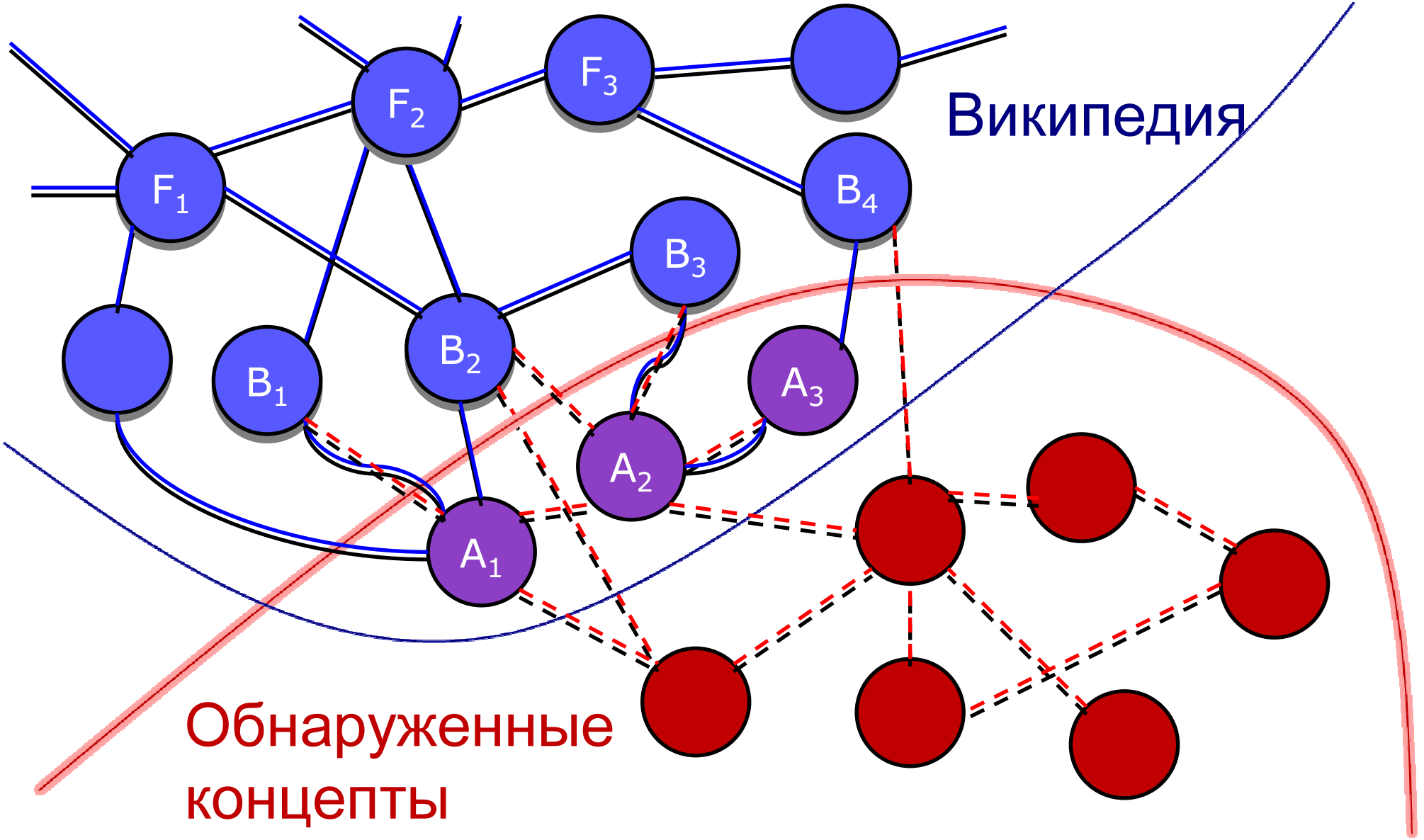
Методология тестирования



Меры качества

- Mean Average Precision
  - Средняя точность между списками, отсортированными по семантической близости (над разными онтологиями) к определенному опорному концепту
  - Усреднение по каждому опорному коцепту

Методология тестирования



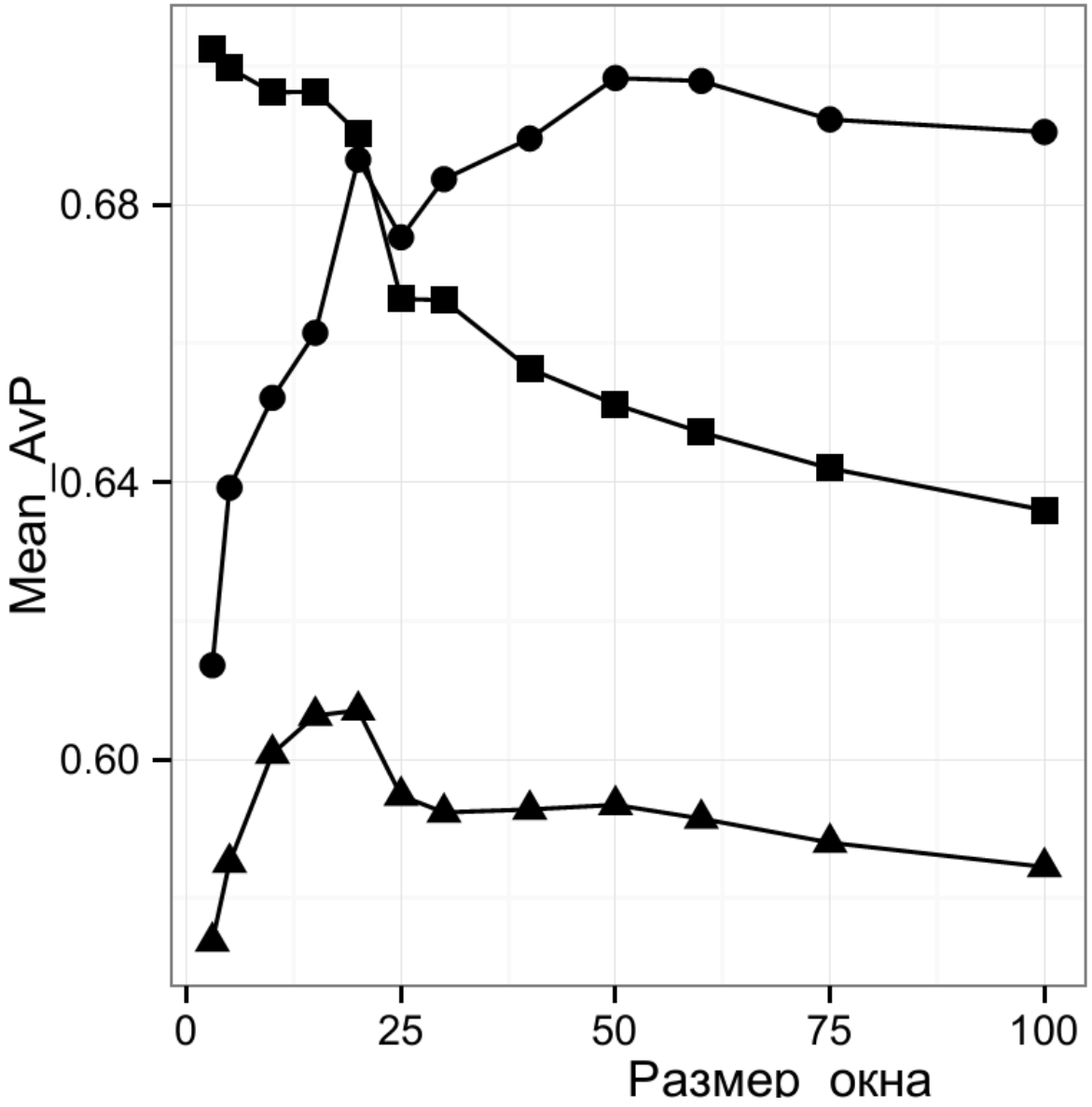
Меры качества

- Mean Average Precision
  1. Среди концептов ***A*** – присутствующих и в Википедии, и в «обогащенной» онтологии
  2. Между ***A*** и ***B*** – соседями ***A*** (одновременно в двух онтологиях)
  3. Между ***A*** и ***F*** – соседями ***B*** (одновременно в двух онтологиях)

Тестовые данные

- 1246 статей про настольные игры
- 37 концептов (A) —
  - имеющие по одному термину
  - проверенные вручную

Baseline

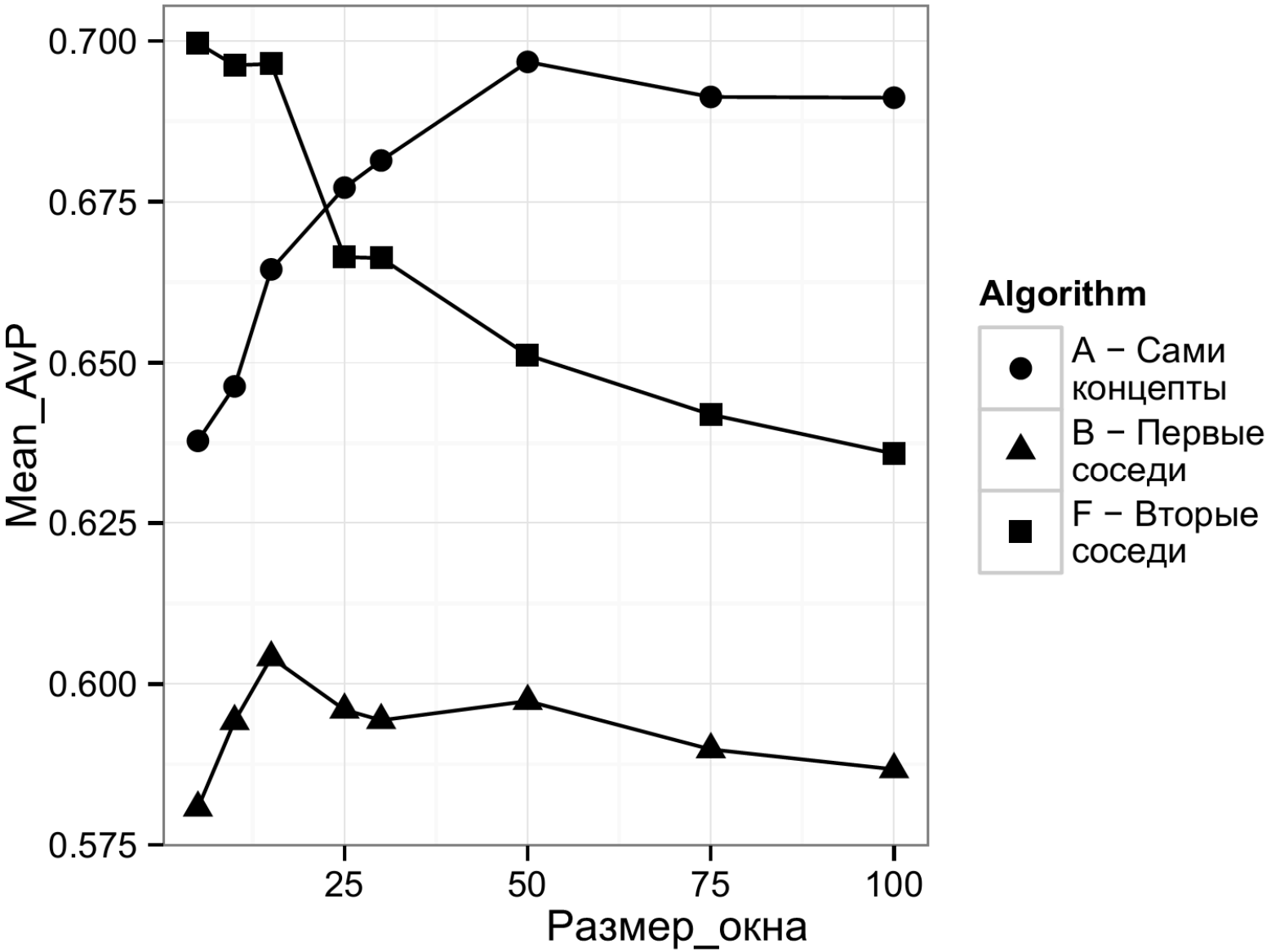


Отсутствие фильтрации кандидатов

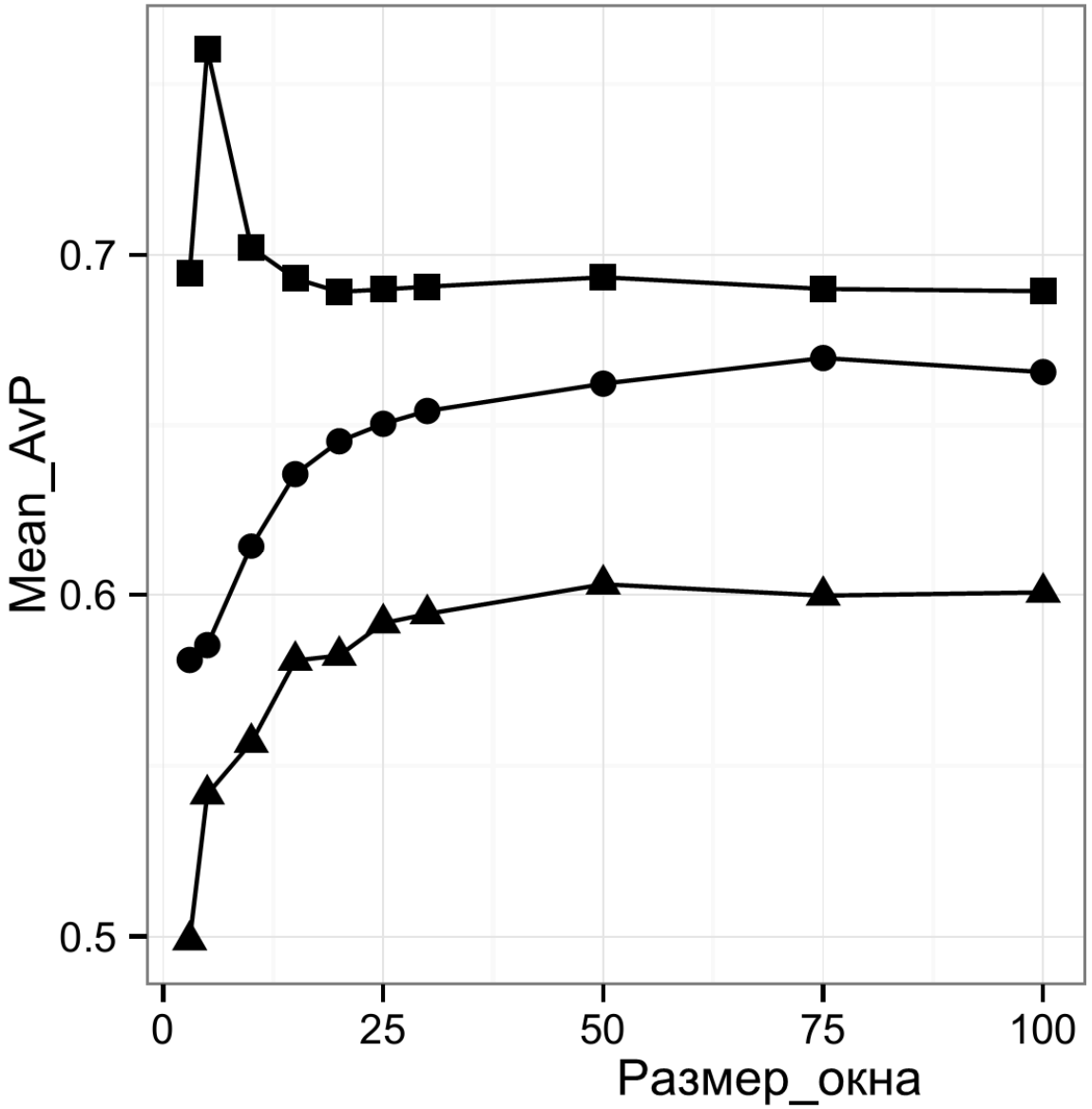
Algorithm

- A – Сами концепты
- B – Первые соседи
- F – Вторые соседи

T-критерий с порогом 1.0



Фильтрация по частоте

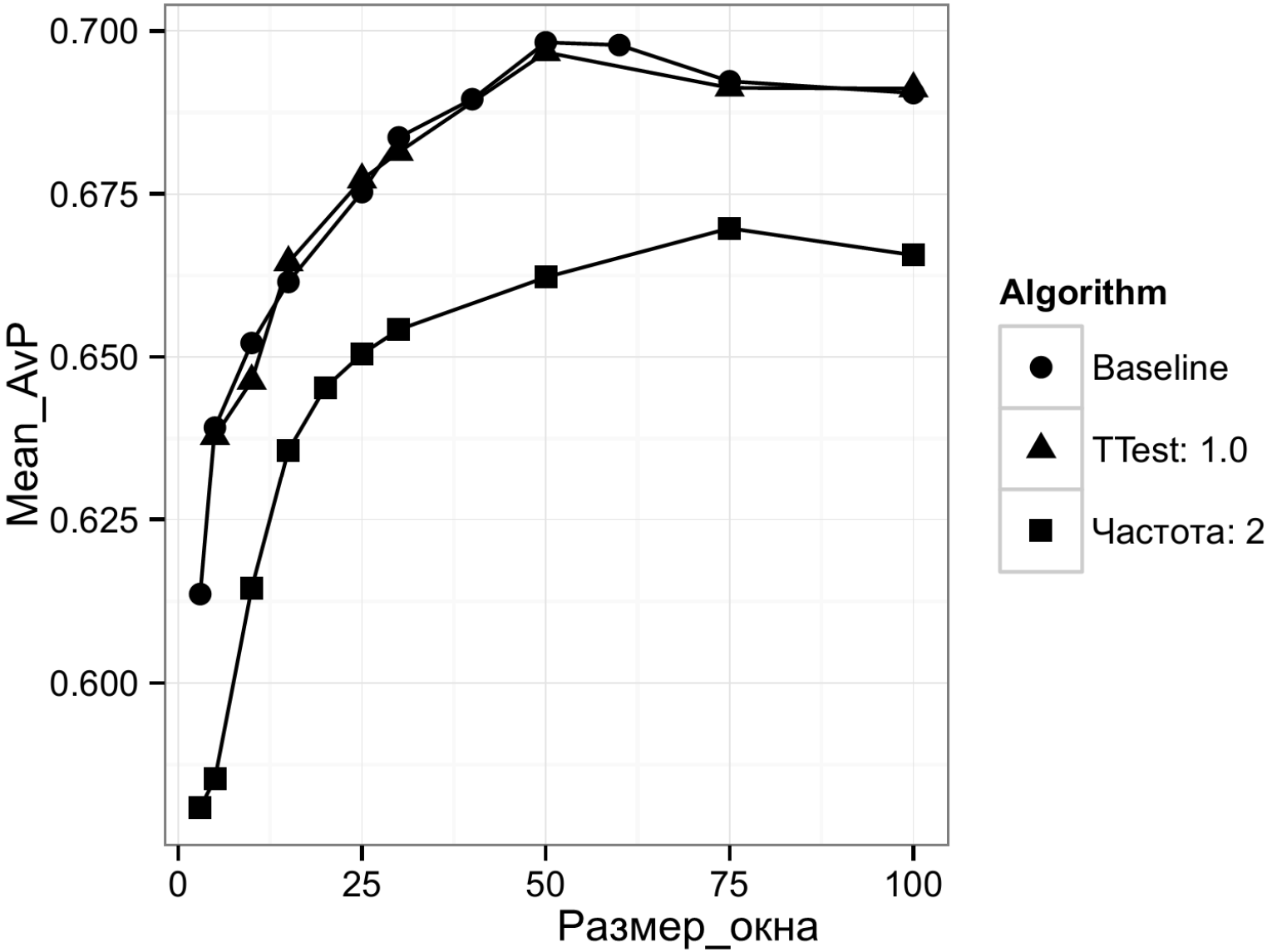


Порог частоты:
реже 2 раз

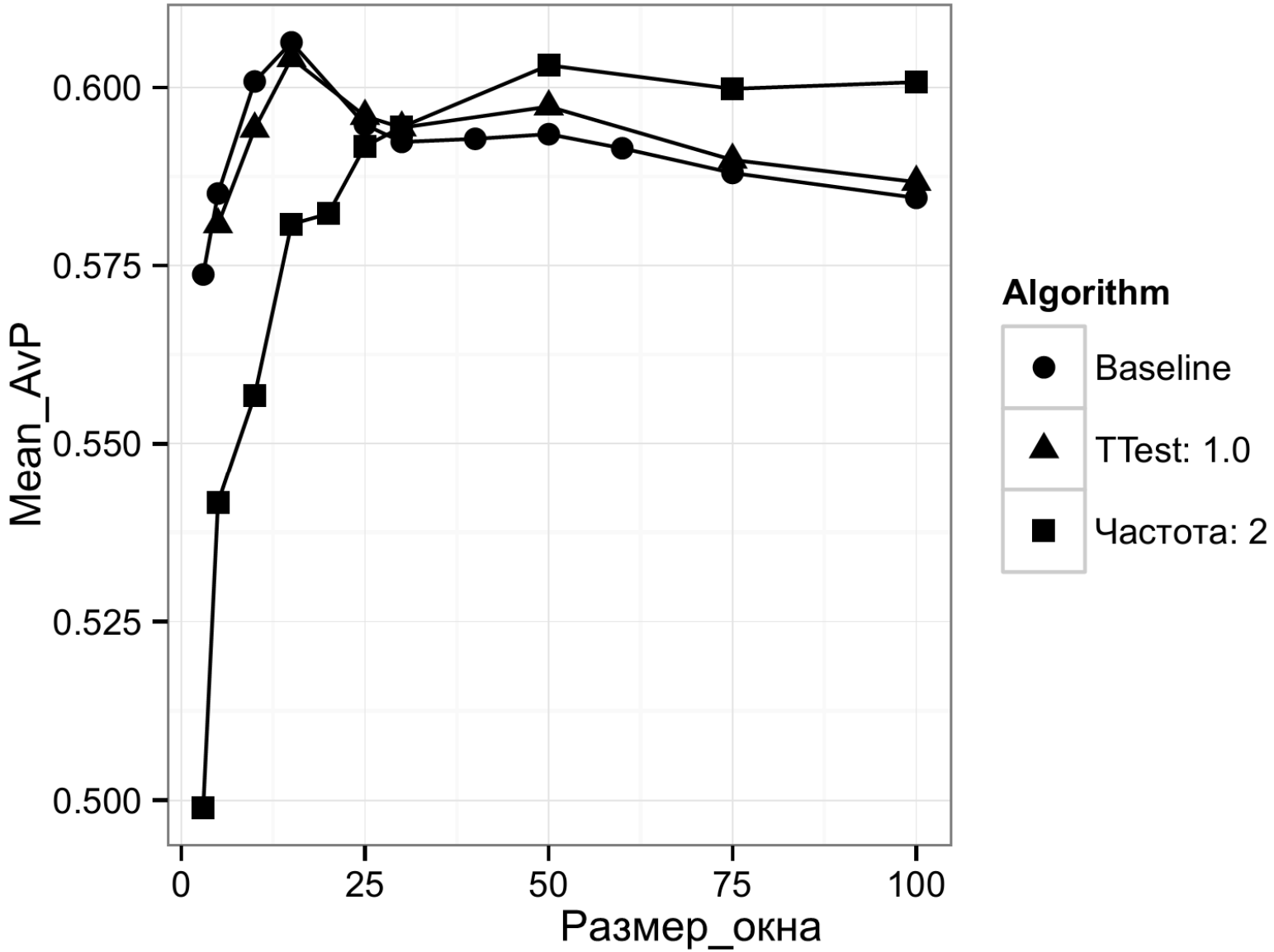
Algorithm

- A – Сами концепты
- B – Первые соседи
- F – Вторые соседи

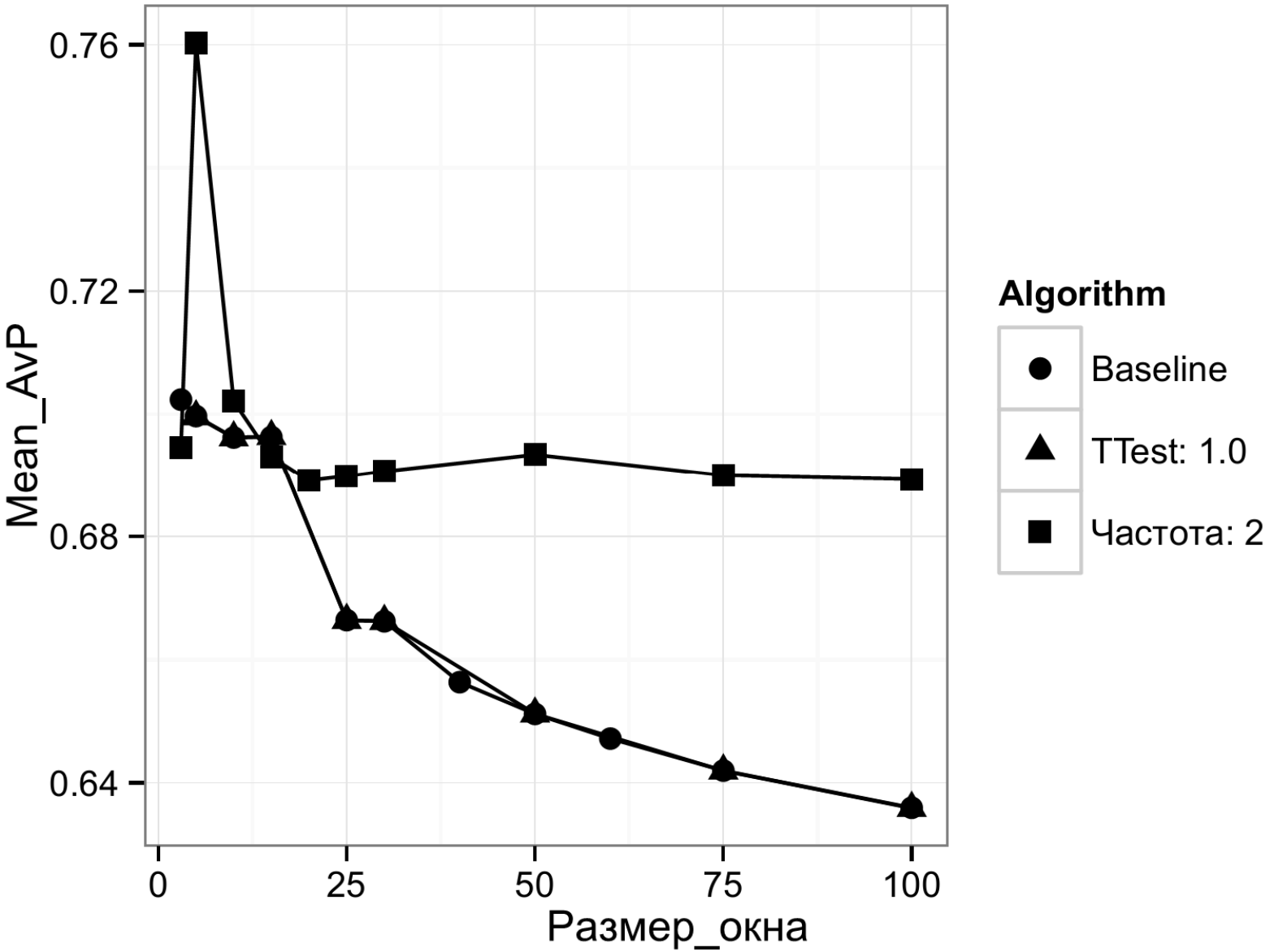
A – сами концепты



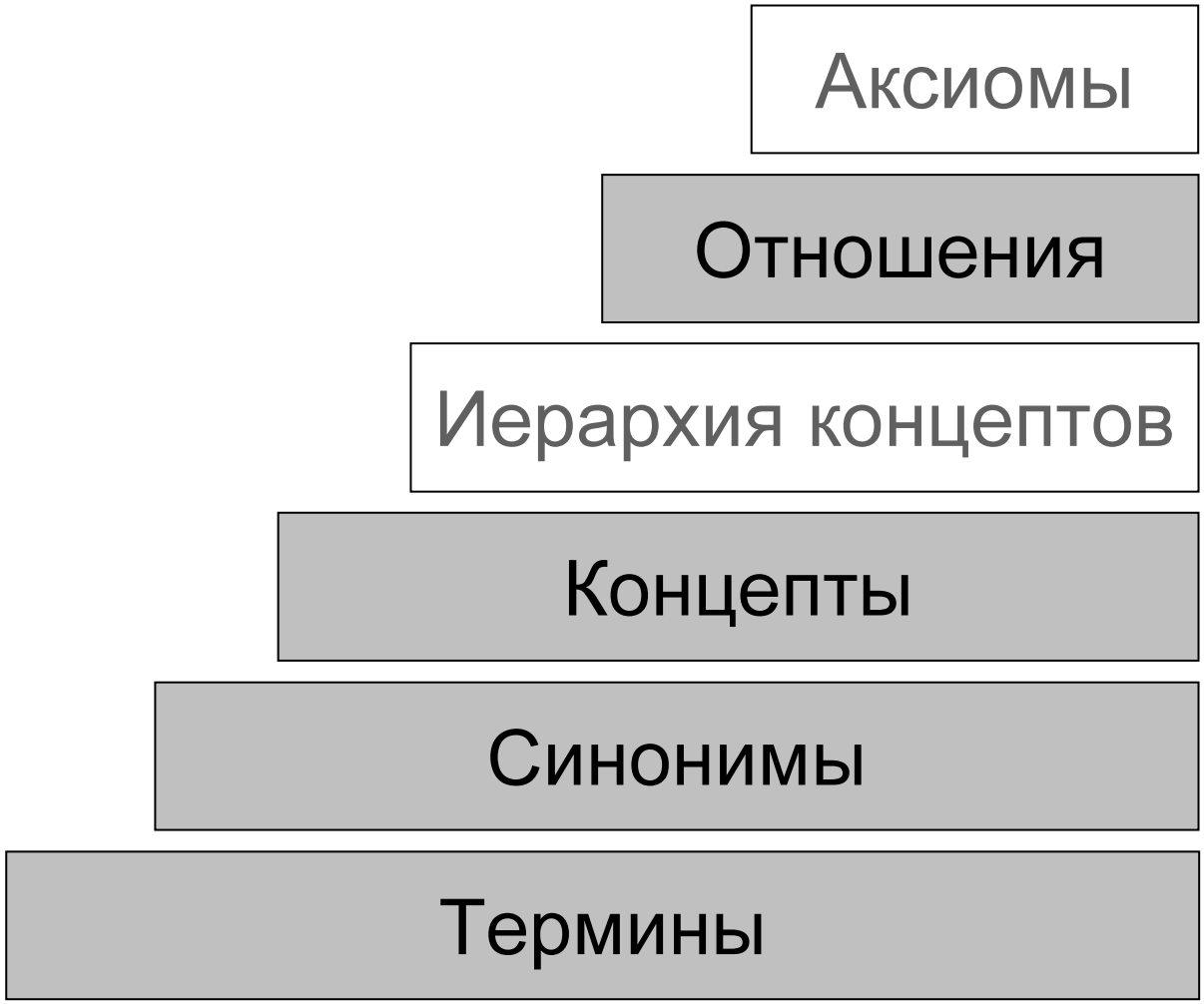
B – первые соседи



F – вторые соседи



Построение онтологии



Направления будущих исследований

- Тестирование на частично размеченном корпусе
- Тестирование в приложениях, опирающихся на онтологии
- Обогащение графа категорий

Заключение

Создан прототип, позволяющий для заданной коллекции документов определенной предметной области создавать соответствующую базу знаний