Аналитика рынка на основе чековых данных
Решение задачи сопоставления текстовых наименований товаров

Цифровизация 2019
28 октября
МГУ имени М.В.Ломоносова

Артем Меликджанян
Директор по развитию бизнеса
Такском

Даниил Каневский
Директор по аналитике
GoodsForecast

О компаниях

С 2000 года разрабатывает и сопровождает системы электронного документооборота (отчетность через Интернет, электронные счета-фактуры и т.п.)

Оператор фискальных данных ФНС России

(более 500 000 касс на обслуживании)

С 2004 года на рынке прогнозирования спроса, оптимизации процессов и аналитики для производственных и торговых предприятий

Компания обладает уникальными математическими алгоритмами и компетенциями по постановке и решению реальных бизнес-задач

GoodsForecast входит в группу компаний Forecsys и является членом Консорциума в области технологий хранения и анализа больших данных

Большие данные

Более 17 млрд чеков

20 млн. чеков ежедневно

20% российского рынка

Зачем?

Возможность получать аналитику по своему сегменту розничного рынка Российской Федерации в режиме реального времени - неоценимый инструмент для развития любого бизнеса.

Производителям

- Оценка эффективности рекламы и промо-акций
- Анализ потребительской корзины (совместно с какими товарами приобретают вашу продукцию?)
- Случаи остановки продаж (Out-of-shelf)
- Анализ и кластеризация торговых точек
- Анализ продаж конкурентов

Ритейлерам

- Сравнительный анализа показателей компании со средними показателями на рынке
- Случаи остановки продаж (Out-of-shelf)
- Эффективность работы кассиров
- Эффективность промо
- Анализ продаж конкурентов

Основная проблема

В чеках передаются текстовые наименования товаров. Для одного и того же товара наименования в разных торговых точках могут различаться.

Как сопоставлять разные наименования и понимать, что это один и тот же товар?

Информация о чеке

- Касса
- Торговая точка
- Тип оплаты (нал/безнал)
- Время операции
- Сумма чека

Информация о «теле» чека

- Товар (наименование)
- Количество
- Сумма

Постановка задачи сопоставления наименований

В «широком» смысле:

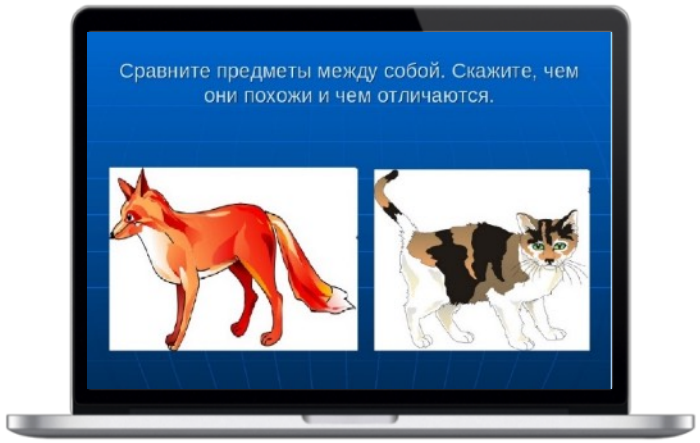
- Формирование на основании получаемых данных единого справочника и классификатора товаров

В «узком» смысле:

- Сопоставление наименований из чеков с заданным списком товаров

Подход к решению задачи

Традиционный подход: «Сравнение строк»



Оценка меры близости двух строк.

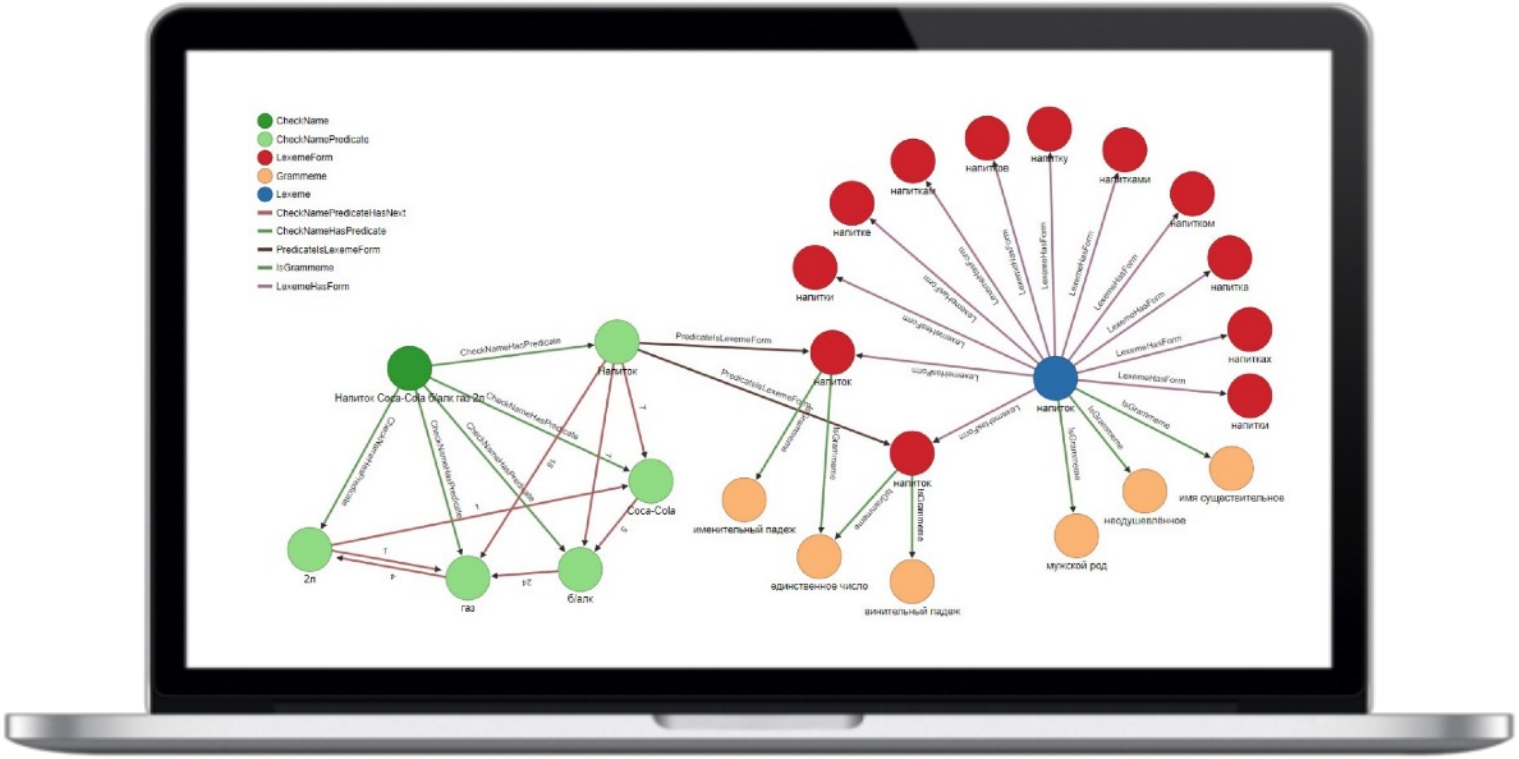
Основная гипотеза:
«схожие строки определяют схожие товары».

Разработанный GoodsForecast подход: Формирование «Базы знаний»

Формирование базы знаний на основании чековых данных, позволяющей определить уникальную специфику наименования различных товаров.

Основная гипотеза:
«при внесении наименования товара люди опираются на одни и те же ключевые свойства товара».

База Знаний



Информация графа обогащается за счёт:

- **Словаря брендов**
- **Словаря единиц измерения**
- **Словаря русского языка с определением частей речи**
- **Словаря синонимов-алиасов (gold = золото = голд и т.п.)**

Выполняется объединение вершин с разными написаниями брендов и единиц измерения, а также синонимы

Алгоритм сопоставления

Дано

Два набора строк:

1) **строки из чеков (полная выборка, либо частичная).**
2) **список искомых товаров**

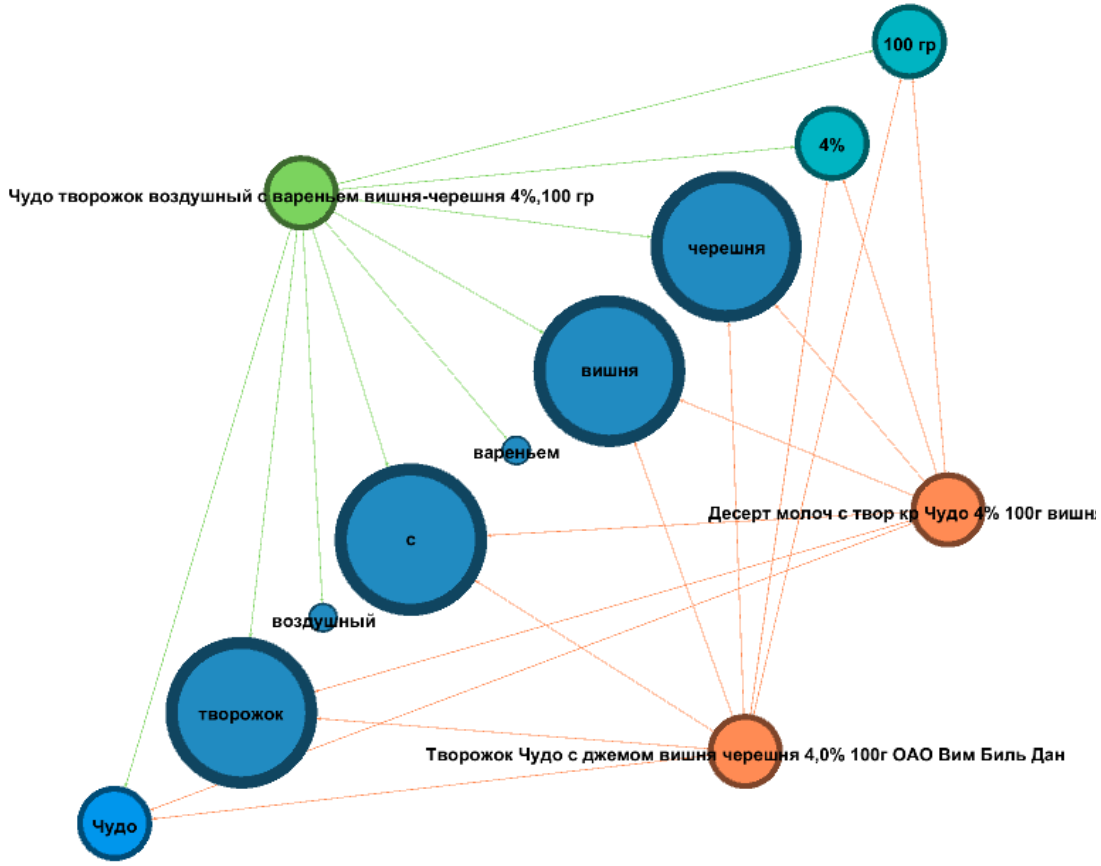
Требуется

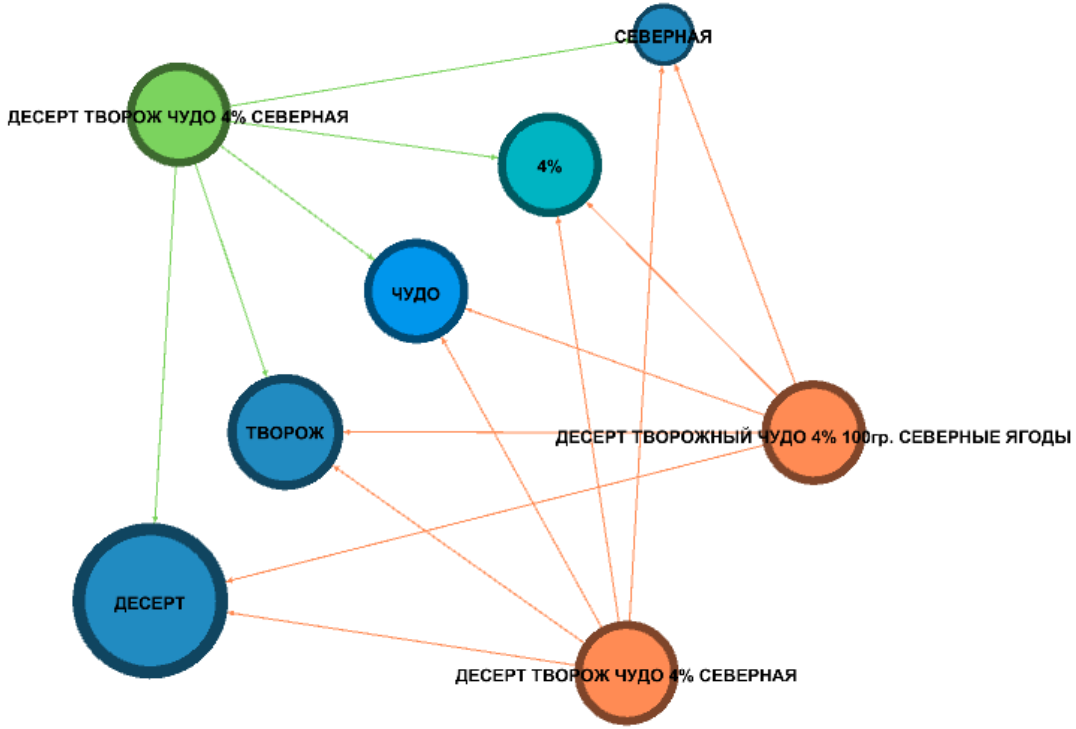
Найти в первом наборе все строки, которые соответствуют товарам из второго набора.

Решение

1) **Строки обоих наборов формируют базу знаний алгоритма. При загрузке происходит маркировка брендов, определение количественных показателей, а также дополнительное сопоставление слов (транслит, сокращения, синонимы и т.п.)**
2) **Каждая строка из чека разделяется на отдельные слова и словосочетания, которые сопоставляются вершинам графа базы знаний.**
3) **На основании полученных вершин определяется множество строк списка искомых товаров, которым может соответствовать анализируемая строка. Алгоритм учитывает частоту совместного расположения отдельных слов в базе знаний чеков, обнаруженные торговые марки и количественные показатели (вес, жирность, возраст и т.п.).**
4) **Если в списке строк кандидатов оказывается несколько строк, далее проводится сравнение уже на уровне найденных строк.**

Алгоритм сопоставления





Преимущества подхода

- **Быстрый этап настройки и тюнинга алгоритма матчинга (примерно неделя работы двух человек – разработчик и математик).**
- **Постоянное накопление экспертной информации (бренды, алиасы и т.п.).**
- **Единый алгоритм сопоставления, дополняемый по мере необходимости экспертной информацией.**
- **Высокая точность работы.**

Перспективы использования

- Проекты Такском
  - Высвобождение ресурсов на альтернативные задачи
- Проекты с участием нескольких ОФД
  - Сопоставление с разным качеством – стопор развития рынка
- Составление и поддержание единого классификатора товаров
  - Разный подход к классификации у разных заказчиков и исполнителей

Спасибо!

Артем Меликджанян
Директор по развитию бизнеса
Такском
artem@taxcom.ru

Даниил Каневский
Директор по аналитике
GoodsForecast
kanevskiy@forecsys.ru

Цифровизация 2019
28 октября
МГУ имени М.В.Ломоносова