Megaputer

Your Knowledge Partner™

АВТОМАТИЗИРОВАННАЯ КЛАССИФИКАЦИЯ И МАРШРУТИЗАЦИЯ ВХОДЯЩИХ ОБРАЩЕНИЙ

PolyAnalyst 6.5

ПОТОК ВХОДЯЩИХ ОБРАЩЕНИЙ

ПРОБЛЕМА

- Крупный российский дистрибьютор ПО и вычислительной электроники столкнулся с проблемой валового роста числа входящих обращений.
- Подавляющее число поступающих документов составляют запросы о резервировании или заказе продукции от компаний-партнеров. Число таких организаций превышает 6000 тысяч, а объем входящих писем составляет порядка нескольких тысяч в сутки.
- Ручная обработка, включающая процедуру определения сути обращения и отправки его в соответствующее структурное подразделение, крайне трудоемка и приводит к существенному временному лагу.

ЗАДАЧА

Реализовать решение по классификации (категоризации) входящих e-mail сообщений с запросами на оборудование, поставляемое компанией, по сочетаниям категорий

- Вендор (бренд производителя оборудования: Cisco, IBM, Asus и т.д.);
- Вид оборудования (категория запрашиваемого оборудования: ноутбуки, мониторы и т.п.).

СТРУКТУРА ОБРАЩЕНИЙ

MESSAGE

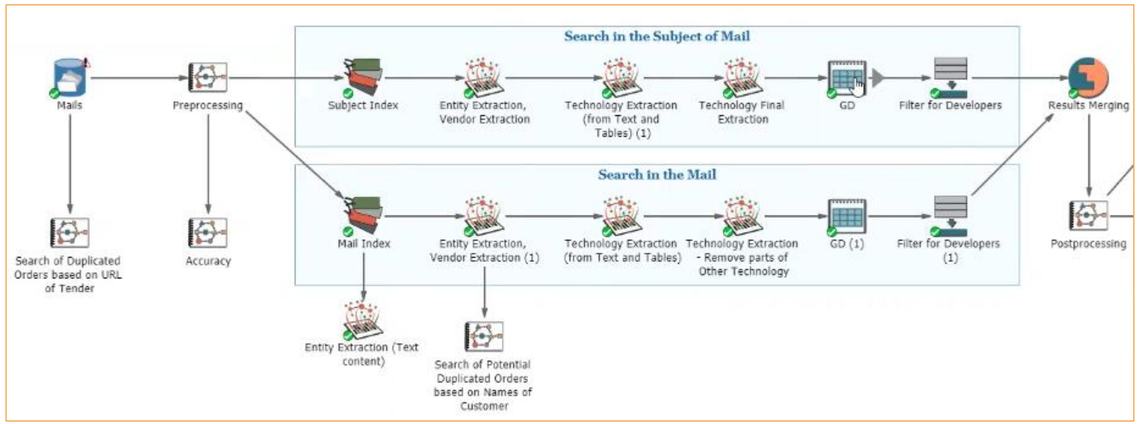
Одно письмо может содержать запрос сразу на несколько связок категорий, например:

- IBM+Ноутбуки;
- LG+Мониторы;
- Оперативная память (без четко обозначенного вендора);
- APC (без четко обозначенного вида оборудования).

АВТОМАТИЗИРОВАННОЕ РЕШЕНИЕ

Команда аналитиков Мегапьютер разработала на базе системы интеллектуального анализа данных и текстов PolyAnalyst специализированное решение, извлекающее из текста электронного письма связку категорий «производитель» + «тип продукции».

Аналитический сценарий проекта



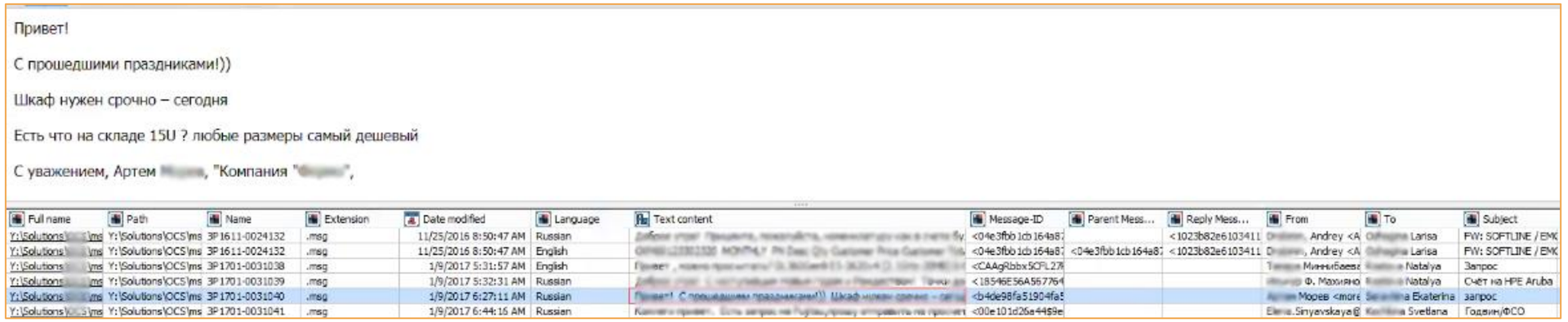
ЭТАПЫ РАБОТЫ СИСТЕМЫ:

1. Загрузка писем в PolyAnalyst
2. Предобработка данных
3. Анализ и извлечение информации
   3.1 Из Темы письма
   3.2 Из Текста письма
4. Постобработка
5. Выгрузка

ЗАГРУЗКА ПИСЕМ В POLYANALYST

Используя встроенный коннектор, система автоматически подключается к корпоративному хранилищу входящей корреспонденции и с периодичностью раз в минуту загружает в себя новые письма.

Загруженные письма на сервере PolyAnalyst формируют таблицу данных



ПРЕДОБРАБОТКА ДАННЫХ

На данном этапе данные подготавливаются к извлечению необходимой информации:

1. Определяется язык
2. Исправляются орфографические ошибки;
3. Замена нижних подчеркиваний на пробелы;
4. Прочие мелкие исправления данных.

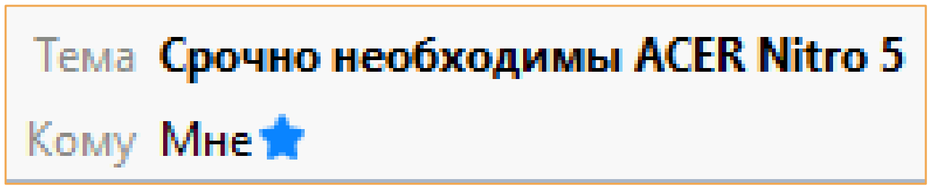
АНАЛИЗ И ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ

Данный этап разделен на два параллельных процесса:

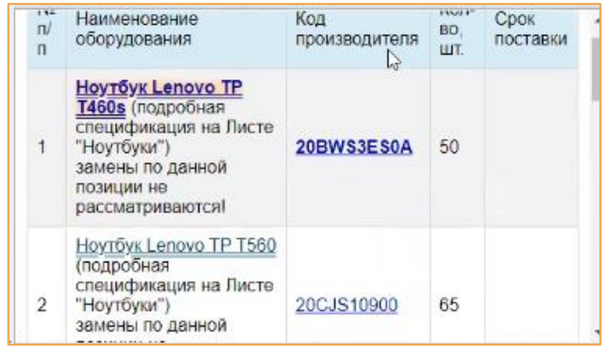
1. Работа с темой письма;
2. Работа с текстом письма.

Такое разделения вызвано различиями в структуре этих двух компонент:

Тема письма имеет простую структуру и чаще всего просто включает информацию о том, что необходимо автору письма:



Текст письма может иметь сложную структуру- включать подпись с перечислением вендоров, таблицы и т.д.



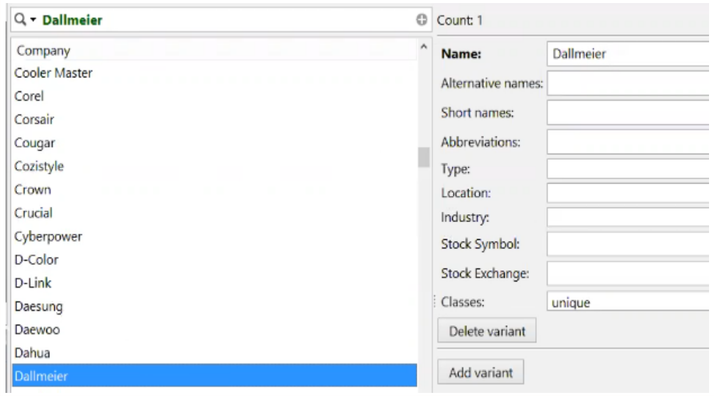
№ п/п	Наименование оборудования	Код производителя	Кол-во, шт.	Срок поставки
1	Ноутбук Lenovo TP T460s (подробная спецификация на Листе "Ноутбуки") замены по данной позиции не рассматриваются!	20BWS3ES0A	50	
2	Ноутбук Lenovo TP T560 (подробная спецификация на Листе "Ноутбуки") замены по данной	20CJS10900	65	

ТЕМА ПИСЬМА

Оба этапа имеют похожую структуру. Система последовательно ищет с применением ruled-based подхода (поиск на основе выявления паттернов с помощью языка XPDL) три типа сущностей (объектов) в данных:

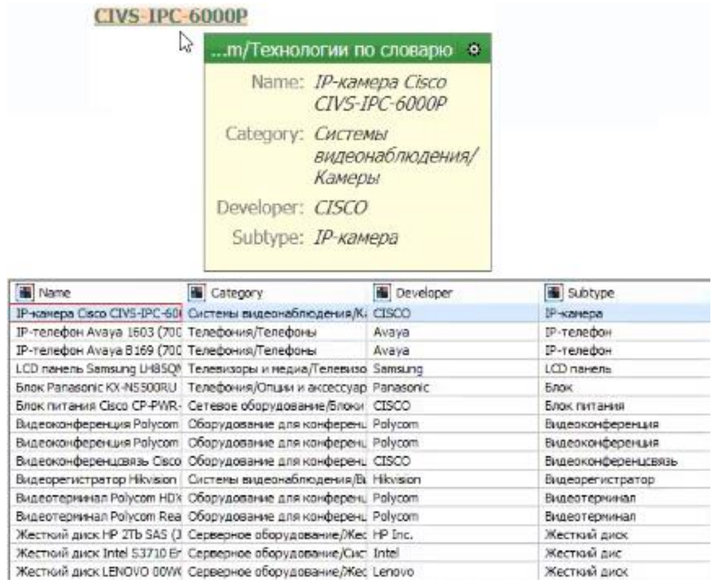
Вендор

Правила на языке XPDL ищут в тексте лингвистические паттерны с упоминанием Вендора. Для этого в PolyAnalyst на основе полученного от заказчика списка Вендоров (таблица Excel) был разработан словарь Вендоров, с необходимым количеством и типом полей



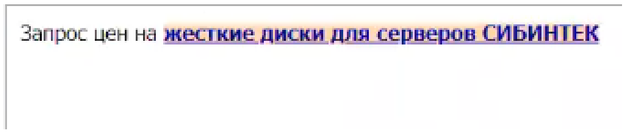
Технология (продукт)

Технологии извлекаются аналогично с применением XPDL и специально собранного сотрудниками Мегапьютер словаря технологий



Части технологии

Правила ищут некоторые сопроводительные части технологии: извлекается технология, и идущие перед ней составные части



ТЕКСТ ПИСЬМА

Здесь выполняются все те же этапы извлечения информации, что и на этапе анализа Темы письма, но добавляются специализированные правила по извлечению данных из таблиц и не извлечению из подписи.

Таблицы

Ищется вендор и продукт, когда они разнесены в разные колонки. Применяется специальная функция для поиска среди таблиц и сведение найденных сущностей воедино.

841636	Ricoh	Картридж голубой тип MP CW2200 для MPCW2200/2201SP (100мл,440стр A1)
841635	Ricoh	Картридж черный тип MP CW2200 для MPCW2200/2201SP (200мл,834стр A1)
D1272110	Ricoh	Блок фотобарабана для Aficio MP301SP/SPF (45000стр, входит в стартовый комплект)

Подпись

Разработаны правила, определяющие в тексте письма подпись, в составе которой находится употребление названий Вендоров или продукции, что не является полезной информацией в рамках данного проекта.

С уважением,
%ФИО%
Ассистент менеджера проектов
ООО «%%%»
тел. %%%
моб. %%%

Dell EMC Platinum Partner, HP Inc. Gold Partner, HP Enterprise Business partner, Cisco Premier Partner, Intel Platinum Partner, Lenovo Silver Partner, IBM Registered Partner, NetApp Registered Reseller, VmWare Partner, APC Select Partner, Microsoft Silver Partner, Oracle Silver Partner, Zebra Business Partner, Juniper Partner

ПОСТОБРАБОТКА ДАННЫХ

Извлеченная из письма Технология (продукт) имеет не только свое наименование, но определяется принадлежностью к подтипу (Mp cw2200 - Картридж, Cena MP c6003 - тонер и т.д.). Задачей этапа постобработки является отнесение Технологии исходя из подтипа к более общей **категории** технологии.

Система автоматически определяет категорию извлеченной технологии и дополняет собранные данные этой информации.

Subtype	Категория	Упоминание вендора
Расходные материалы	Периферия	OKI
Расходные материалы	Периферия	Panasonic
Расходные материалы	Периферия	Samsung
Расходные материалы	Периферия	Sharp
Расходные материалы	Периферия	Xerox
ИБП	Инженерная инфраструктура	APC
Электрика	Инженерная инфраструктура	APC
Печать и сканирование	Периферия	Samsung
Расходные материалы	Периферия	HP Inc.
Расходные материалы	Периферия	Xerox
		Panasonic
		Zebra Technologies
Серверы неспециализированные	Серверы	HP Inc.
		Hewlett Packard Enterprise

ПОСТОБРАБОТКА ДАННЫХ

Таким образом у нас получаются две таблицы (по результатам анализа темы и текста), содержащих извлеченную информацию. Далее оба набора объединяются в одну общую базу данных, где представлены: номер письма, название вендора, название технологи, подтип технологии, категория технологии.

РЕЗУЛЬТАТ АНАЛИЗА ТЕМЫ

Subtype	Категория	Упоминание вендора
Расходные материалы	Периферия	OKI
Расходные материалы	Периферия	Panasonic
Расходные материалы	Периферия	Samsung
Расходные материалы	Периферия	Sharp
Расходные материалы	Периферия	Xerox
ИБП	Инженерная инфраструктура	APC
Электрика	Инженерная инфраструктура	APC
Печать и сканирование	Периферия	Samsung
Расходные материалы	Периферия	HP Inc.
Расходные материалы	Периферия	Xerox
		Panasonic
		Zebra Technologies
Серверы неспециализированные	Серверы	HP Inc.
		Hewlett Packard Enterprise

РЕЗУЛЬТАТ АНАЛИЗА ТЕКСТА

Subtype	Категория	Упоминание вендора
Расходные материалы	Периферия	OKI
Расходные материалы	Периферия	Panasonic
Расходные материалы	Периферия	Samsung
Расходные материалы	Периферия	Sharp
Расходные материалы	Периферия	Xerox
ИБП	Инженерная инфраструктура	APC
Электрика	Инженерная инфраструктура	APC
Печать и сканирование	Периферия	Samsung
Расходные материалы	Периферия	HP Inc.
Расходные материалы	Периферия	Xerox
		Panasonic
		Zebra Technologies
Серверы неспециализированные	Серверы	HP Inc.
		Hewlett Packard Enterprise

Name	Subtype	Категория	Упоминание вендора
0031052	Расходные материалы	Периферия	OKI
0031052	Расходные материалы	Периферия	Panasonic
0031052	Расходные материалы	Периферия	Samsung
0031052	Расходные материалы	Периферия	Sharp
0031052	Расходные материалы	Периферия	Xerox
0031053	ИБП	Инженерная инфраструктура	APC
0031053	Электрика	Инженерная инфраструктура	APC
0031054	Печать и сканирование	Периферия	Samsung
0031054	Расходные материалы	Периферия	HP Inc.
0031054	Расходные материалы	Периферия	Xerox
0031054			Panasonic
0031054			Zebra Technologies

ВЫГРУЗКА ДАННЫХ

Итоговая таблица с определенной периодичностью импортируется из PolyAnalyst в информационную систему заказчика и на основании представленных данных производится пересылка писем соответствующим корреспондентам.

ПРЕИМУЩЕСТВА СИСТЕМЫ

Разработанное решение за счет продуктивности интеллектуальных инструментов текстового анализа PolyAnalyst продемонстрировало высокую (свыше 90%) точность классификации писем. Интегрированная аналитическая схема позволила освободить сотрудников от рутинных операций и автоматизировать процессы классификации и маршрутизации.

• В отличии от ручного труда, система практически в режиме реального времени проводит анализ документации, что существенно сократило временной промежуток между получением письма и отправкой его в ответственное подразделение.

• В силу рутинности данных операций обработка документов человеком неизбежно приводит к возникновению ошибок и неточностей, что почти не характерно для автоматизированного процесса.

• Система адаптивна к росту количества запросов и исключает необходимость привлечения дополнительного персонала.

• Система имеет гибкую структуру, ясную логику и интуитивный интерфейс. При необходимости модернизации она может быть изменена или дополнена (новые сущности, словари, правила) сотрудниками Заказчика самостоятельно без обращения к вендору (Мегапьютер).

МЕГАПЬЮТЕР ИНТЕЛЛИДЖЕНС

PolyAnalyst 6.5

МЕГАПЬЮТЕР ИНТЕЛЛИДЖЕНС

Извлекаем и структурируем факты из текстовых документов

Оцифровываем и роботизируем бизнес-процессы

Строим модели на основе аналитики и Искусственного Интеллекта

107 разработчиков, 16 лингвистов и аналитиков, 9 кандидатов наук

Предоставляем кластерную платформу для анализа Больших Данных

Поддерживаем четверть компаний из списка Fortune 100 и еще более 100 клиентов

Член Ассоциации Разработчиков Программных Продуктов «Отечественный софт»

Платформа PolyAnalyst включена в реестр Российского ПО. Свидетельство №4414