



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

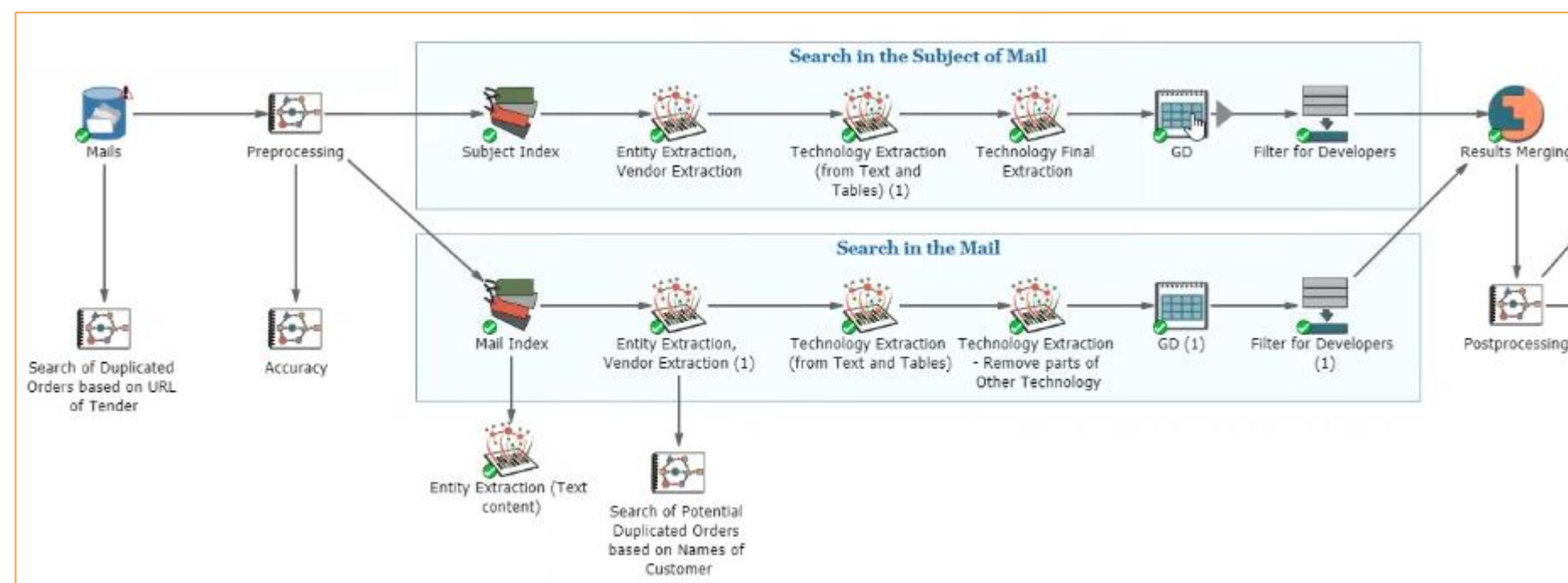
Одно письмо может содержать запрос сразу на несколько связок категорий, например:

- 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

The screenshot shows an email client interface. The top part displays the content of an email in Russian: "Привет! С прошедшими праздниками!)) Шкаф нужен срочно – сегодня Есть что на складе 15U ? любые размеры самый дешевый С уважением, Артем [redacted], 'Компания ' [redacted]'". Below the email content is a table listing the loaded messages.

Full name	Path	Name	Extension	Date modified	Language	Text content	Message-ID	Parent Mess...	Reply Mess...	From	To	Subject
Y:\Solutions\...ms	Y:\Solutions\OCS\ms	ЗР 1611-0024132	.msg	11/25/2016 8:50:47 AM	Russian	Доброе утро! Присоединяюсь к поздравлениям с прошедшими праздниками!)) Шкаф нужен срочно – сегодня Есть что на складе 15U ? любые размеры самый дешевый С уважением, Артем [redacted], 'Компания ' [redacted]'	<04e3fbb1cb164a87...>	<04e3fbb1cb164a87...>	<1023b82e6103411...>	Druskin, Andrey <A...>	Druskin, Andrey <A...>	FW: SOFTLINE / EMC
Y:\Solutions\...ms	Y:\Solutions\OCS\ms	ЗР 1611-0024132	.msg	11/25/2016 8:50:47 AM	English	DRUSKIN, ANDREY <A...> FW: SOFTLINE / EMC	<04e3fbb1cb164a87...>	<04e3fbb1cb164a87...>	<1023b82e6103411...>	Druskin, Andrey <A...>	Druskin, Andrey <A...>	FW: SOFTLINE / EMC
Y:\Solutions\...ms	Y:\Solutions\OCS\ms	ЗР 1701-0031038	.msg	1/9/2017 5:31:57 AM	English	Привет! Шкаф нужен срочно – сегодня Есть что на складе 15U ? любые размеры самый дешевый С уважением, Артем [redacted], 'Компания ' [redacted]'	<CAAgrbbx5CFL27...>			Тимофей Мухоморов <...>	Druskin, Andrey <A...>	Запрос
Y:\Solutions\...ms	Y:\Solutions\OCS\ms	ЗР 1701-0031039	.msg	1/9/2017 5:32:31 AM	Russian	Доброе утро! Присоединяюсь к поздравлениям с прошедшими праздниками!)) Шкаф нужен срочно – сегодня Есть что на складе 15U ? любые размеры самый дешевый С уважением, Артем [redacted], 'Компания ' [redacted]'	<18546E56A56776...>			Druskin, Andrey <A...>	Druskin, Andrey <A...>	Счет на HPE Aruba
Y:\Solutions\...ms	Y:\Solutions\OCS\ms	ЗР 1701-0031040	.msg	1/9/2017 6:27:11 AM	Russian	Привет! С прошедшими праздниками!)) Шкаф нужен срочно – сегодня Есть что на складе 15U ? любые размеры самый дешевый С уважением, Артем [redacted], 'Компания ' [redacted]'	<b4de98fa51904fa5...>			Artem Morev <more...>	Druskin, Andrey <A...>	запрос
Y:\Solutions\...ms	Y:\Solutions\OCS\ms	ЗР 1701-0031041	.msg	1/9/2017 6:44:15 AM	Russian	Коллеги привет! Есть запрос на покупку сервера на складе	<00e101d26a4469e...>			Elina.Sinyevskaya@...>	Korotkova Svetlana	Годовиц/FCO

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 разделение вызвано различиями в структуре этих двух компонент:

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

Тема **Срочно необходимы ACER Nitro 5**
Кому Мне ★

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 XPDЛ) три типа сущностей (объектов) в данных:

Вендор

Правила на языке XPDЛ 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

The screenshot shows a search interface with a search bar containing 'Dallmeier' and a 'Count: 1' indicator. Below the search bar is a list of company names, with 'Dallmeier' selected. To the right, there are input fields for 'Name', 'Alternative names', 'Short names', 'Abbreviations', 'Type', 'Location', 'Industry', 'Stock Symbol', and 'Stock Exchange'. The 'Name' field is filled with 'Dallmeier'. At the bottom, there are 'Delete variant' and 'Add variant' buttons.

Технология (продукт)

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

The screenshot shows a search result for 'CIVS-IPC-6000P'. A tooltip displays the following information: Name: IP-камера Cisco CIVS-IPC-6000P, Category: Системы видеонаблюдения/Камеры, Developer: CISCO, Subtype: IP-камера. Below the tooltip is a table with columns: Name, Category, Developer, Subtype.

Name	Category	Developer	Subtype
IP-камера Cisco CIVS-IPC-6000P	Системы видеонаблюдения/Камеры	CISCO	IP-камера
IP-телефон Avaya 1603 (700)	Телефония/Телефоны	Avaya	IP-телефон
IP-телефон Avaya B169 (700)	Телефония/Телефоны	Avaya	IP-телефон
LCD панель Samsung UN8500	Телевизоры и медиа/Телевизоры	Samsung	LCD панель
Блок Panasonic KX-N5500RU	Телефония/Опции и аксессуары	Panasonic	Блок
Блок питания Cisco CP-PWR-	Сетевое оборудование/Блоки	CISCO	Блок питания
Видеоконференция Polycom	Оборудование для конференц	Polycom	Видеоконференция
Видеоконференция Polycom	Оборудование для конференц	Polycom	Видеоконференция
Видеоконференция Cisco	Оборудование для конференц	CISCO	Видеоконференция
Видеорегистратор Hikvision	Системы видеонаблюдения/Видеорегистраторы	Hikvision	Видеорегистратор
Видеотерминал Polycom HDX	Оборудование для конференц	Polycom	Видеотерминал
Видеотерминал Polycom RealView	Оборудование для конференц	Polycom	Видеотерминал
Жесткий диск HP 2Tb SAS (J7100)	Серверное оборудование/Жесткие диски	HP Inc.	Жесткий диск
Жесткий диск Intel 53710 E7	Серверное оборудование/Жесткие диски	Intel	Жесткий диск
Жесткий диск LENOVO 00WK	Серверное оборудование/Жесткие диски	Lenovo	Жесткий диск

Части технологии

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

Запрос цен на [жесткие диски для серверов СИБИНТЕК](#)

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 Аficio MP301SP/SPE (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 (Mr sw2200 - Картридж, Сена MP с6003 - 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 информацией.

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 технологии, 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 отличие 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

