# HP Autonomy IDOL: анализ совсем неструктурированных данных

***В конце августа 2012 г. компания HP в составе организационной структуры в России ввела новый бизнес-юнит – "Information Management". В обязанности нового департамента вошло продвижение решений HP Autonomy с новой платформой IDOL 10, доступной с конца января 2012 г., которая способна работать как с неструктурированной информацией (аудио-, видео-, социальные ресурсы, электронная почта, веб-контент и др.), так и со структурированными данными (журналы транзакций, показатели счетчиков и др.), позволяя автоматически извлекать из огромного объема данных ценные сведения, увязывая смысловой контент данных, признаки поведения и др. с отдельными социальными группами.***

**Евгений Олейник** – руководитель департамента Autonomy IM в Центральной и Восточной Европе, HP.

## *Введение*

*Человеческое сообщество стремительно развивается и вместе с этим резко возрастают объемы информации, которые оно генерирует. К ним можно отнести: аудио- и видеоданные, контент социальных сетей и веб-контент, электронную почту и др. До недавнего времени это был практически только информационный шум, полноценных технических средств для его анализа не существовало. Этому препятствовали многие проблемы, среди которых можно выделить две основные. Первая – необходимость автоматизации перевода в текстовый формат аудио- и видеозаписей с последующим их аннотированием и каталогизацией. Вторая – необходимость обработки колоссальных объемов данных.*

*По мере развития ИТ как аппаратных, так и программных технологий эти задачи стали решаемыми и более того – почти в реальном времени или близком к нему. Практически полностью решена задача распознавания речи (для более сотни языков) – и не только с точки зрения перевода ее в текстовый вид, но и с точки зрения понимания ее смыслового содержания и даже интонационных оттенков.*

*Стала доступна для анализа видеоинформация, например, с точки зрения выделения из всего видеоряда отдельных объектов по определенным признакам: к примеру, распознавание лиц, а также мимики лица, по которой появилась возможность(машинная) автоматически определять эмоциональное и психологическое состояние человека. Теперь можно, например, выбрать какой-либо объект и отследить его перемещение в пространстве по результатам видеонаблюдения со множества камер, сделав обобщенную вырезку из всех файлов.*

*В результате стало доступно аннотирование аудио- и видеоинформации с последующей каталогизацией и индексацией данных – и все это без какого-либо участия человека.*

## Autonomy в мире и в России

Решения Autonomy стали продвигаться в России совсем недавно – после ее приобретения HP в 2011 г. за $12 млрд. Сама Autonomy была создана в 1996 г. профессором Кембриджского университета и до покупки была одной из самых успешных компаний в мире. В настоящее время компания является лидером по обработке и пониманию неструктурированных данных с общим числом заказчиков более 65 тыс., среди которых – большинство топовых компаний. Более 400 вендоров (Adobe, Citrix, Dassault Systemes, EMC, IBM Global Services, Novell, Sybase, TIBCO и др.) интегрировали отдельные компоненты решений Autonomy в свои разработки (в основном, это поиск). Есть компании, которые полностью стандартизировались на Autonomy. Это значит, что разные департаменты – маркетинг, юридический, ИТ и т.д. – используют для поиска информации только движки от Autonomy. Например, BBC после закачивания видеоинформации осуществляет ее дальнейший анализ и поиск по ней только с помощью решений Autonomy.

Ведущие аналитические мировые агентства признают лидерство Autonomy в своих исследованиях: Digital Asset Management Wave report, Forrrester, 2012, Q2; Magic Quadrant for Information Archiving, Gartner, 2011; Magic Quadrant for Web Content Management, Gartner, 2010 и 2011.

В конце августа 2012 г. компания HP в составе своей организационной структуры ввела новый бизнес-юнит – "Information Management". В обязанности нового департамента вошло продвижение решений HP Autonomy с новой платформой IDOL 10, доступной с конца января 2012 г.

## Решения Autonomy

Платформа IDOL 10 – это первое ее обновление в составе компании HP. Она способна работать со всей "человеческой" информацией (аудио, видео, социальные ресурсы, электронная почта, веб-контент и др.) и структурированными машинными данными (журналы транзакций, показатели счетчиков и др.), позволяя извлекать смысловую информацию из большого объема неструктурированных данных, включая видео- и аудиозаписи без предварительной обработки. Платформа IDOL 10 включает ПО от компании Autonomy для автоматической обработки неструктурированных данных, а также высокопроизводительный модуль анализа структурированных данных от компании Vertica, входящей в состав HP.

Платформа IDOL 10 предлагает:

- *единый уровень обработки информации* вне зависимости от ее формы и местонахождения;
- *ПО Autonomy* для автоматической обработки неструктурированных данных и высокопроизводительный *модуль Vertica* для обработки машинных данных в режиме реального времени — в одном решении;
- *уникальные технологии сопоставления*, основанные на модуле аналитики и статистических алгоритмах и способные распознавать идеи, концепции и контекст в режиме реального времени;
- *пять новых наборов решений*: HP Big Data Solutions, HP Social Media Solutions, HP Risk Management Solutions, HP Cloud Solutions и HP Mobility Solutions;
- *технологию Manage-in-place*, которая индексирует все формы данных, позволяя использовать информацию без изменения ее местоположения. Это исключает необходимость создания копий данных, снижает расходы на сис-

темы хранения и устраняет потребность в переносе данных;

– *интерфейс NoSQL,* который обеспечивает единый уровень обработки для анализа структурированных и неструктурированных данных;
– *функции повышения гибкости* обеспечивают динамическое расширение и сжатие кластеров при любом сценарии развертывания (облачном, виртуальном или физическом). Это позволяет оперативно выделять дополнительную емкость по мере необходимости.

Autonomy создает решения в трех областях Power, Protect и Promote:

– *решения Power* – все, что связано с обработкой неструктурированных данных: поиск, аналитика, управление бизнес-процессами и др. Применения:
  - Pan-Enterprise Search;
  - Probabilistic Analysis;
  - Business Process Management;
  - Collaboration and Expertise Networks;
  - Knowledge Management & eLearning;
  - Personalization Operations;
– *Protect* – все, что касается защиты организации и сокращение расходов за счет понижения рисков. Применения:
  - eDiscovery;
  - Compliance;
  - Content Management
  - Legal Market;
  - Records Management;
  - Content Archiving;
  - Backup & Recovery;
– *Promote* – решения, которые позволяют увеличить эффективность и прибыльность компаний. Применения:
  - Web Content Management;
  - Revenue Optimization;
  - Advanced Analytics;
  - Social Media;
  - eCommerce;
  - Mobile;
  - Online Advertising Management;
  - Segmentation & Targeting;
  - Multichannel Optimization;
  - Contact Center;
  - Rich Media Management.

Все решения строятся на базе одной платформы – IDOL (Intelligent Data Operating Layer).

***Как строятся решения на основе IDOL***

Основой всех решений является платформа IDOL с подключенными к ней т.н. коннекторами, которые позволяют понимать тот или иной формат данных. Есть коннекторы для твиттера, блогов, для видеообработки потоков данных конкретных ТВ каналов, т.е. под конкретный источник данных пишется свой коннектор. На текущий момент разработано более 400 коннекторов, которые понимают более 1000 файловых форматов. На основе конкретных коннекторов и платформы IDOL строятся конкретные решения: "анализ по предпочтениям", "увеличение воронки продаж" и др.

В IDOL более 500 разных функций и более 70 различных приложений, которые и позволяют создавать специализированные решения для различных бизнес-задач. В настоящее время Autonomy – одна из немногих компаний в мире, предоставляющая решения на основе понимания неструктурированных данных, в которых поиск является самым начальным базовым приложением, на базе которого строятся уже другие более высокоуровневые приложения Autonomy.

***Особенности поисковых систем, концептуальный поиск***

В прошлом и, в основном, в настоящем поиск основывается на ключевых словах. Например, если вводится какое-либо слово в поисковой системе Google, то в качестве результата выдаются страницы, где упоминается это слово или его варианты. Autonomy с помощью своего движка IDOL может это сделать. Но, если говорить, например, о безопасном поиске в корпоративной среде обычные движки, в основном, не смогут дифференцировать и показывать информацию в зависимости от профиля работника, осуществляющего поиск, и способа доступа к корпоративной инфраструктуре. Юридический поиск также имеет свою специфику и в большинстве случаев стандартные средства поиска не могут удовлетворить все его требования.

Ценность решений Autonomy существенно возрастает при концептуальном поиске т.е когда поиск документов осуществляется не на встречаемость какого-либо слова, а на тему, которая описывается этим словом, и само это слово в этих документах может вообще не встречаться. В данном случае IDOL пытается понять смысл документа и его соответствие смысловому содержанию поискового слова и уже на основании этого делает селекцию документа.

Еще больше возрастает сложность концептуального поиска, когда он осуществляется в аудио- и видеофайлах. В настоящее время поиск по этим файлам, в основном, осуществляется только в тех случаях, где для них есть метаданные. Autonomy способен осуществлять концептуальный поиск по аудио- и видеофайлам без предварительной подготовки для них метаданных. Он способен распознавать речь (включая русский язык) на более чем 120 языках и увязывать каждую фразу с видеокартинкой.

Движок Autonomy основывается на двух математических моделях. Первая гласит, что чем больше у нас источников информации, тем выше вероятность того, что можно понять в полной мере то, что происходит. Вторая говорит о том, что необязательно понимать каждую часть информации (например, в документе), чтобы понять суть.

Например, есть запись телефонного разговора, где отдельные слова не понятны. Autonomy подставляет их в текст разговора, основываясь на совместном использовании трех алгоритмов: лингвистическом, звуковом и вероятностном. Первый – это понимание того, какое слово должно быть в данный момент в тексте по смыслу, второй – это словарь слов по звучанию, из которого выбирается слово, имеющее наименьшее отклонение от того, которое нужно распознать. И третье – какие слова упоминаются чаще всего вместе.

***Возможности Autonomy по анализу видеоинформации***

Autonomy может отслеживать на основе своих правил (например, того, что должно происходить, а что нет) и математических моделей может выявлять:

– связи между различными объектами (например, между отдельными гражданами/гражданином) по видеозаписи/видеозаписям;
– неправомерные действия граждан (например, многочисленные попытки открыть дверь или снять деньги с банкомата) и автоматически их фиксировать;
– несовпадение/соответствие лица человека, осуществляющего какие-либо действия с магнитной картой, с фотографией человека (хозяина карты), записанной на самой карте (или хранящейся в БД). В этом случае автоматически могут генерироваться сообщения службе безопасности, а этот человек автоматически попадает под дальнейшее более пристальное наблюдение. Более того, Autonomy может выявить поведение человека и его поступки, предшествующие этому моменту, по уже сделанным видеозаписям. При этом в каждом случае Autonomy сама выявляет характерные признаки человека (одежда, манера поведения, лицо, мимика лица, прическа и т.д.), по которым она его будет отслеживать, характер и число этих признаков каждый раз меняется.

Небольшая иллюстрация результативности технологий Autonomy. В небольшом городе Хулл (Англия) после введения системы видеоаналитики на основе Autonomy в течение трех лет удалось сделать задержание более 6,5 тыс. человек по причинам правонарушений, т.е. система по видеоматериалам в онлайне автоматически выявляла какие-либо незаконные действия граждан на основе своих правил, отбирала эти сюжеты из общего видеоряда и передавала в полицейское управление. После их анализа принимались соответствующие действия: или видеоданные игнорировались или на их базе проводились действия по задержанию/пресечению нарушения.

***Семейство решений Virage Security & Surveillance (VSS)***

Семейство Virage Security & Surveillance включает 6 решений:

– Digital/Network Video Recording (DVR/NVR);
– Automatic Number Plate Recognition (ANPR);
– Intelligent Scene Analysis System (iSAS), Electronic Point of Sale Monitoring (EPOS);
– Container Surveillance and Management System (CSM);
– fingerprint and audio analysis and 3-D face recognition,

которые позволяют на их базе строить разнообразные приложения для разных применений.

Платформа Virage Command and Control (VCC) предлагает службам безопасности полностью законченное решение, значительно упрощая многие операции и поддерживая аналитику в реальном времени. VCC обеспечивает обширный диапазон автоматических функциональных возможностей, которые включают:

- *автоматическое соединение со связанным контентом:* по заданной части информации (текст, видео или аудио) VCC будет автоматически идентифицировать другую релевантную информацию и взаимоотношения между различными частями данных. Результат выдается в форме гиперссылок;
- *динамическая кластеризация:* VCC способна обеспечить полный обзор базы знаний. Анализируя концептуальное содержание различных частей информации, VCC идентифицирует кластеры связанной информации, автоматически сортируя их в соответствующие группы для немедленного доступа к ним. Этим способом VCC позволяет выявить скрытые взаимоотношения и идентифицировать потенциальные угрозы безопасности, которые без этого могли бы быть пропущены;
- *профилирование и своевременное информирование:* за счет автоматического профилирования информации VCC (на основе заданного профиля клиента) VCC формирует высоко целенаправленную информацию в реальном масштабе времени, выдавая ее на мобильное устройство через SMS, по электронной почте или по Internet.

*Digital/Network Video Recording*

VCC поддерживает полный спектр цифровых и сетевых видеорекордеров. Одна система может расширяться до 65 000 входных потоков. Virage использует различные техники сжатия видео и позволяет:

- осуществлять любые варианты фиксации происходящего: запланированная запись; запись, запускаемая событиями; запись после события; фотографирование и др.;
- динамически изменять разрешение видеозаписи в зависимости от событий и приоритезации между камерами;
- хранить видеоматериалы объемом до 1 Пбайт и др.

*Automatic License/Number Plate Recognition (ALPR/ANPR) – автоматическое распознавание лицензии/номера (автомобиля)*

ALPR/ANPR-система использует современные техники для распознавания символов. Ее интеграция с DVR-решением позволяет создавать приложения для: обслуживания различных автомобильных потоков, автомобильных стоянок и гаражей, бензозаправочных станций; обнаружения пропавших транспортных средств.

*Video Analytics: Intelligent Scene Analysis System (ISAS)*

Как показывают исследования, человеческие возможности крайне ограничены при осуществлении видеонаблюдения, которые могут значительно снижаться после даже 15-минутного внимательного просмотра видеокартинок. Наступление общей утомляемости и снижение концентрации внимания происходят после 45 минут наблюдения. Часть информации не фиксируется из-за особенностей мозговой деятельности человека. Как результат, до 85% информации на экране остается незамеченной (рис. 1). Автоматизация видеонаблюдения помогает во многом решить эту проблему.


**Рис. 1.** До 85% информации, связанной с видеонаблюдением только человеком, остается незамеченной.

Интеллектуальная система анализа сцен – расширенное решение, которое помогает пользователям любой системы наблюдения идентифицировать важную активность типа потенциальной угрозы, незаконного действия или ситуации, где требуется помощь. Фактически, система может быть обучена детектировать все, что требуется пользователю.

ISAS интегрирует с любой новой или существующей инсталляцией CCTV и улучшает эффективность в широком диапазоне ситуаций, включая:

- контроль доступа к определенным зонам;
- мониторинг трафика с целью идентификации транспортного средства в аварийных зонах, перехват правонарушителя или контроль трафика на автостраде;
- мониторинг железнодорожных или станций метрополитена с целью идентификации пассажиров;
- мониторинг в аэропортах с целью идентификации подозрительного поведения или багажа.

ISAS хорошо подходит для многих инфраструктур и легко интегрируется с системами от третьих фирм.

*Electronic Point of Sale (EPOS) Monitoring*

Вместе с DVR решение мониторинга электронной точки продаж представляет мощный пакет, позволяя обнаруживать и фиксировать мошенническую деятельность в точке продажи по различным действиям. Отдельные интерфейсы EPOS позволяют подключаться к кассовым аппаратам (300 типов), банкоматам, паркоматам и др.

*Container Surveillance and Management (CSM)*

CSM представляет одну из лидирующих в мире систем наблюдения и управления контейнерными перевозками. За счет высококачественного DVR управления, CSM позволяет пользователям вести точный учет того, когда и где контейнеры проходят через транспортные хабы типа морских портов, внутренних контейнерных терминалов и железнодорожных контейнерных терминалов.

CSM автоматически фиксирует информацию типа:

- номера транспортного средства;
- контейнерные ISO номера;
- тип контейнера;
- месторасположение контейнера;
- состояние контейнера.

*Biometrics and Audio Recognition*

Технологии распознавания речи позволяют подразделениям/организациям, отвечающим за безопасность, осуществлять поиск в файлах в видео-, радио- и телефонных систем. Технология распознавания речи в VCC существенно отличается от традиционных, в которых в основе лежит фонетический подход, и используется только акустическая информация. Virage достигает бОльшего уровня понимания через моделирование языка. Акустически-фонетический подход не дифференцирует, например, между “can I” and “can eye”. В этом случае технологии Virage используют интеллектуальное вероятностное моделирование языка, чтобы понять контекст того, о чем говорится.

Функциональные возможности Virage для распознавания речи включают:

- *независимость от спикера*: система обучена на большом количестве данных, затрагивающих много различных переменных типа различных акцентов, мужская или женская речь, тональность речи и др. Решение работает " из коробки " без необходимости какого-либо обучения, хотя настройка для определенных акцентов или спикеров может быть сделана;
- *обширные словари* без ограничения на размер словаря;
- *возможность идентификации спикера:* система может быть обучена для распознавания отдельных спикеров;
- *определение слова и распознавание фразы:* система может осуществлять поиск по стандартным ключевым словам, а также концептуальным методом. При концептуальном способе информация оценивается на релевантность искомой фразе/слову;
- *пониженные затраты использования ЦПУ и ОП* за счет патентованной технологии Autonomy;
- *поддержку аудио как высокого качества, так и телефонных разговоров.*

Технологии распознавания отпечатков пальцев и объемного изображения лица дополняют решение по распознаванию речи. VSS имеет те же техники распознавания отпечатков, что и полицейские эксперты, и позволяет использовать БД с информацией на миллионы людей. Методы распознавания по лицу часто используются, например, где подозреваемые должны быть идентифицированы на расстоянии. Однако традиционные двухмерные изображения всегда имели фундаментальные ограничения, связанные с положением лица, его выражением, освещенностью и др. VSS использует для распознавания трехмерные техники, которые во многом снимают эти ограничения.

## Примеры использования

*1) Autonomy Legal Hold – сбор информации для судебных разбирательств*

Autonomy предлагает законченный портфель решений – eDiscovery – для электронной обработки документов в рамках базовой модели EDRM (Electronic Discovery Reference Model) – поиска, выемки и предоставления электронной информации.

В рамках этого семейства для судебных разбирательств может использоваться решение Legal Hold.

В США большое значение придается своевременному предоставлению информации во время судебных разбирательств. Если этого не происходит, то во многих подобных случаях суды проигрываются. При наличии большого числа источников информации (например, в крупной организации) ее сбор крайне осложнен. Решение Autonomy Legal Hold (рис. 2) в значительной степени упрощает эту задачу. Оно собирает относящуюся к делу информацию из корпоративных архивов, а также из собственного и сохраняет это в отдельном кейсе, гарантируя ее целостность.


**Рис. 2.** Этапность сбора и предоставления информации при судебных разбирательствах с помощью решения Autonomy Legal Hold.

При этом сама информация остается в первоначальных источниках данных. Она лишь защищается от стирания. Параллельно выборка документов может быть помещена во внутренний архив Autonomy для бОльшей целостности и гарантированной защиты от стирания.

Собранная информация используется как на ранней стадии рассмотрения дела, так и на последующих стадиях судебного процесса.

Это решение может быть также использовано и для внутрикорпоративных расследований.

*2) Autonomy Investigator – поиск информации, связанной с фактами нарушения комплайнса (мошенничество)*

На этом примере проиллюстрируем факты мошенничества в компании Enron (6-летней давности). Все данные для этого примера были скачаны с публичных сайтов.

Например, нас интересуют все документы в организации, которые связаны с нарушением этики в связи с мошенничеством. При этом мы не знаем что искать, поэтому в поисковой строке (Query Text) вводим самую общую фразу: “ethical issues with fraud” (рис. 3). После этого система нам выдает (Retrieval Results) все документы, связанные с этой темой (при этом в

**Рис. 3.** Окно приложения на базе движка IDOL для расследования фактов нарушения комплайнса.

них может вообще не присутствовать целиком фраза для поиска или ее отдельные слова), а под поисковой фразой (Query Expansion List) предлагаются другие предложения – по “идеям”, которые могут относиться к этой теме. Среди них мы видим слово “raptor” (хищник). Оно вызывает подозрение, потому что логически к этой теме не имеет никакого отношения. Чтобы внести ясность в этот вопрос, можно посмотреть файлы по ссылке и установить, о чем говорят люди, когда употребляют слово “raptor”.

Открыв список документов по слову “raptor”, мы можем воспользоваться опцией Link Map, которая позволяет посмотреть кто внутри данной выборки переписывался по данной теме с кем и когда. Это т.н. картина ссылок (рис. 4), в которой точки – это люди, линии – это коммуникации между людьми. Толстая линия означает, что коммуникаций было много, соответственно, тонкая – коммуникаций мало. Ярко-красная это – от кого исходило большинство сообщений. На этом графе также можно наблюдать, что Vince.J.Kaminski из компании Enron посылал письма VKaminski. При этом VKaminski обратно ничего не присылал. Т.е. можно сразу понять, что этот человек “сливает” информацию на свой личный адрес на публичном сервере

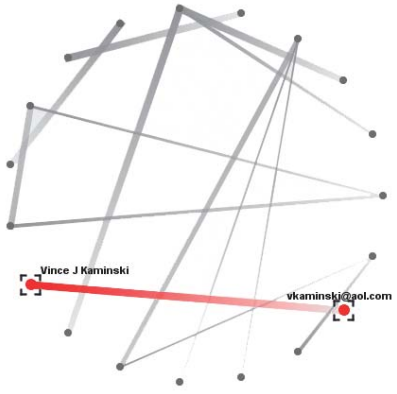


**Рис. 4.** Граф Link Map, иллюстрирующий интенсивность коммуникаций сообщества людей по выбранной теме.

(aol.com). Как выяснилось в дальнейшем, “raptor” – это было кодовое слово в выявленной группе мошенников.

Данное приложение можно использовать и в закрытых корпоративных сетях (для чего потребуется инсталляция данного приложения) для мониторинга выполнения комплайнса служащими на основе вышеприведенных запросов с целью предотвращения неправомерных попыток и/или при расследовании уже случившегося.

*3) Autonomy Voice Discovery – расследование аудиозвонков*

Данное приложение позволяет:

- выбрать из всего множества аудиофайлов только те, в которых встречается какое-либо слово, например, наука. При этом с помощью регулятора (от 0 до 100%) можно повышать “уверенность” системы в том, что данное слово определено верно. Это может существенно уменьшать объемы (в часах прослушивания) выбранной аудиоинформации. При этом при открытии файлов, отобранных системой, она автоматически будет “подсвечивать” (с помощью стрелок) место, где оно встречается (рис. 5);
- определить слова, которые совпадают по смыслу с написанным, но могут иметь одинаковое звучание с другим словом. Например, слово “know” одинаково произносится с “no”, но имеет совершенно отличный смысл;
- распознавать не только текст, но и эмоции;
- отобрать файлы по смысловому содержанию задаваемой фразы. При этом сама фраза (или ее отдельные слова) в выбранных аудиофайлах может (могут) не упоминаться вообще;
- сделать выборку людей, которые общаются на определенную тему;
- понимать смешение языков.

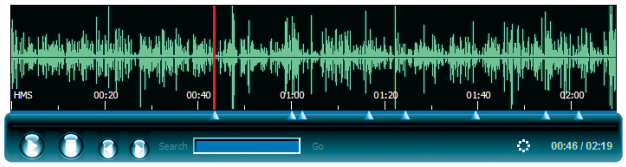


**Рис. 5.** Система автоматически выделяет (с помощью стрелок) то место, где встречается найденное слово .

*4) Фиксация результатов расследования*

После того, как были определены документы (с помощью Investigator и Voice Discovery) для их выемки и предоставления в качестве доказательств в ходе расследования, можно сделать их слепок (или заморозить) с помощью Legal Hold.

## *Заключение*

*Технологии, заложенные в семействе решений Autonomy, раскрывают очень большие возможности по анализу неструктурированных данных и извлечению из них гораздо больше ценной информации, чем это было доступно до недавнего времени.*

*Видео- и аудиоаналитика на порядок повышает эффективность служб безопасности при выявлении и ретроспективном анализе случаев возможных угроз, а также при расследовании чрезвычайных событий.*

*Autonomy для бизнеса – это возможность автоматизации управления большого числа процессов, происходящих в реальном времени, которые ранее были сопряжены со многими рисками и/или издержками.*

***Евгений Олейник,***
*компания HP*