IBM OmniFind Enterprise Edition:

Стратегическая платформа для поиска и анализа текста в масштабах предприятия.

Потребность предприятия в поиске и анализе текста.

Во многих организациях поиск документов и их извлечение из электронных архивов, файловых хранилищ и приложений представляют проблему. Неспособность найти нужную информацию в нужное время может негативно влиять на производительность сотрудников, качество обслуживания клиентов и скорость ведения бизнеса в целом. Ожидания пользователей сегодня – иметь возможность находить информацию, которая накоплена в организации в реальном времени, используя парадигму поиска в виде запросов на естественном языке, широко используемую для поиска информации в Интернет. Однако корпоративная информация отличается от Web – контента и алгоритмы поиска, используемые для Web-запросов, не дают адекватного результата для поиска в масштабах крупной организации.

К тому же, значительная часть данных организаций хранится в неструктурированном виде или в «свободной» форме. Эти данные включают в себя документы, электронные сообщения, Web-контент, комментарии, поля сообщений или отчеты.

Постоянный рост объема неструктурированных данных и их ценность для бизнес-приложений вызывает потребность для расширенного анализа текста для того, чтобы эта информация была снабжена более детальными и семантическими метаданными. Руководители современных организаций осознают ценность неструктурированной информации и ее полезность для пользователей и приложений.

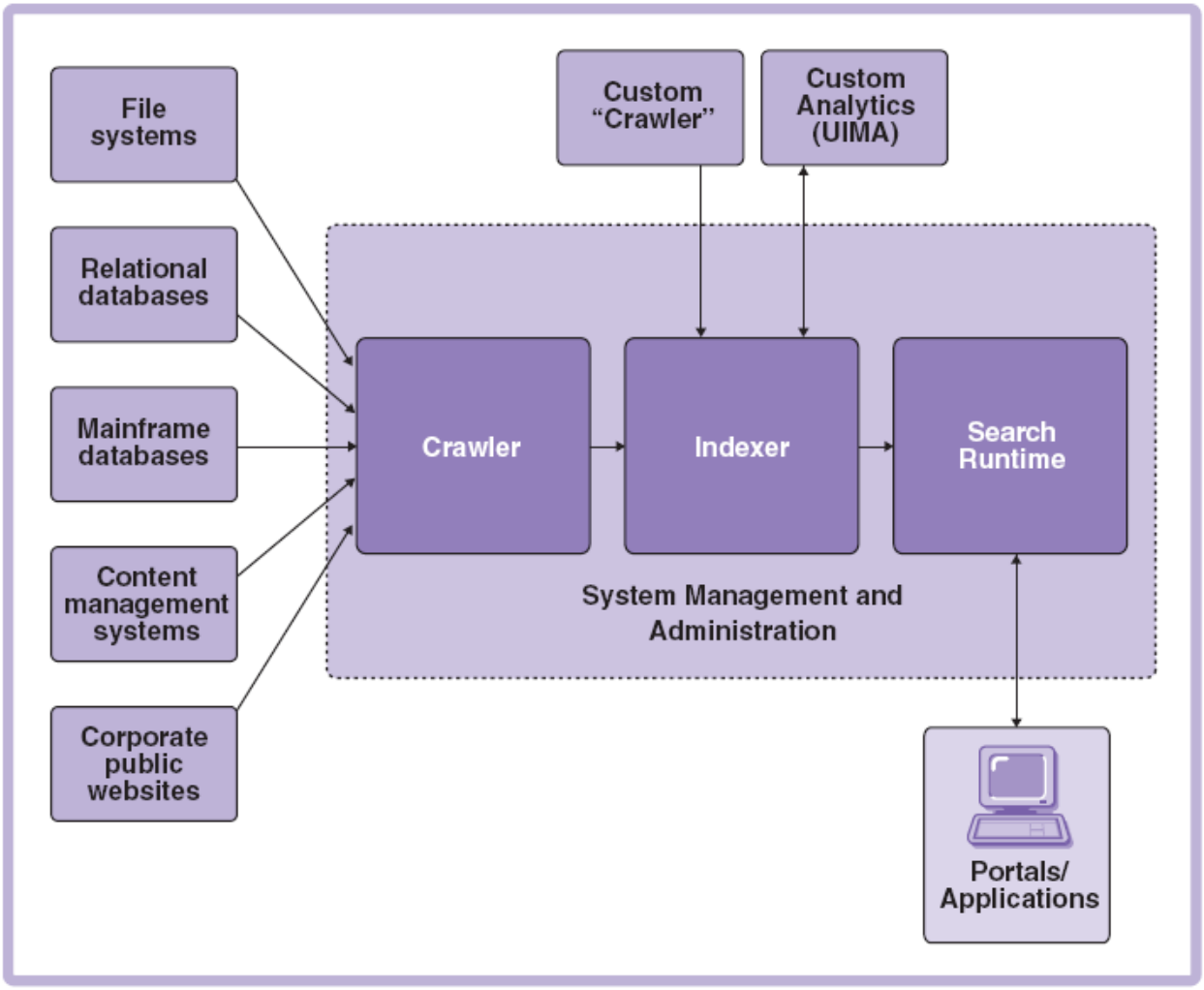
Решение IBM OmniFind Enterprise Edition

Решение OmniFind Enterprise Edition – ключевое предложение в рамках портфеля решений IBM для работы с неструктурированной информацией (контентом) от IBM. Продукт IBM OmniFind предоставляет высококачественный, масштабируемый, защищенный поиск в свободной форме, позволяющий находить наиболее релевантную корпоративную информацию сотрудникам, поставщикам, партнерам и клиентам. Простым введением ключевых слов или фразы, пользователи могут легко найти информацию в интранете, корпоративных Web-сайтах, базах данных, файловых системах и хранилищах контента, получая качественные результаты. У пользователей также имеется возможность пойти дальше стандартного полнотекстового поиска, используя параметрические и семантические запросы, что значительно повышает релевантность результатов поиска.

IBM OmniFind Enterprise Edition является также и первой коммерчески доступной платформой для обработки текстовой информации в архитектуре Unstructured Information Management Architecture (UIMA). UIMA позволяет бесшовно интегрировать компоненты текстового анализа, извлекать знания и идентифицировать высокоуровневые объекты, такие как люди, места, организации, продукты, и другие «объекты», которые скрыты в неструктурированных данных. Эти значения могут быть использованы для создания расширенных индексов и семантического поиска или перенаправлены в традиционную витрину данных или хранилище данных для использования в BI-приложениях (Business Intelligence) или аналитических приложениях.

Гибкая архитектура для предоставления высоко релевантных результатов

Тремя основными компонентами IBM OmniFind Enterprise Edition являются: сборщики (crawlers), сервер индексации и поисковая подсистема (search runtime). Сборщики извлекают контент из различных источников; сервер индексации разбирает предложение и анализирует документы, а затем строит индексы; поисковая система обрабатывает поисковые запросы, находя наиболее релевантные документы по индексу, и менее чем через секунду возвращает результаты поиска.



Сборщики позволяют обращаться за информацией к различным источникам данных. Так же, информацию можно передать "вручную", минуя сборщиков, на следующий шаг процесса, который включает разбиение документа на лексемы и выполнение соответствующего лингвистического анализа. Далее, если это определено, содержимое может быть ассоциировано с категориями на основании таксономических правил. На шаге индексации происходит статическое ранжирование и удаление дублирующих записей.

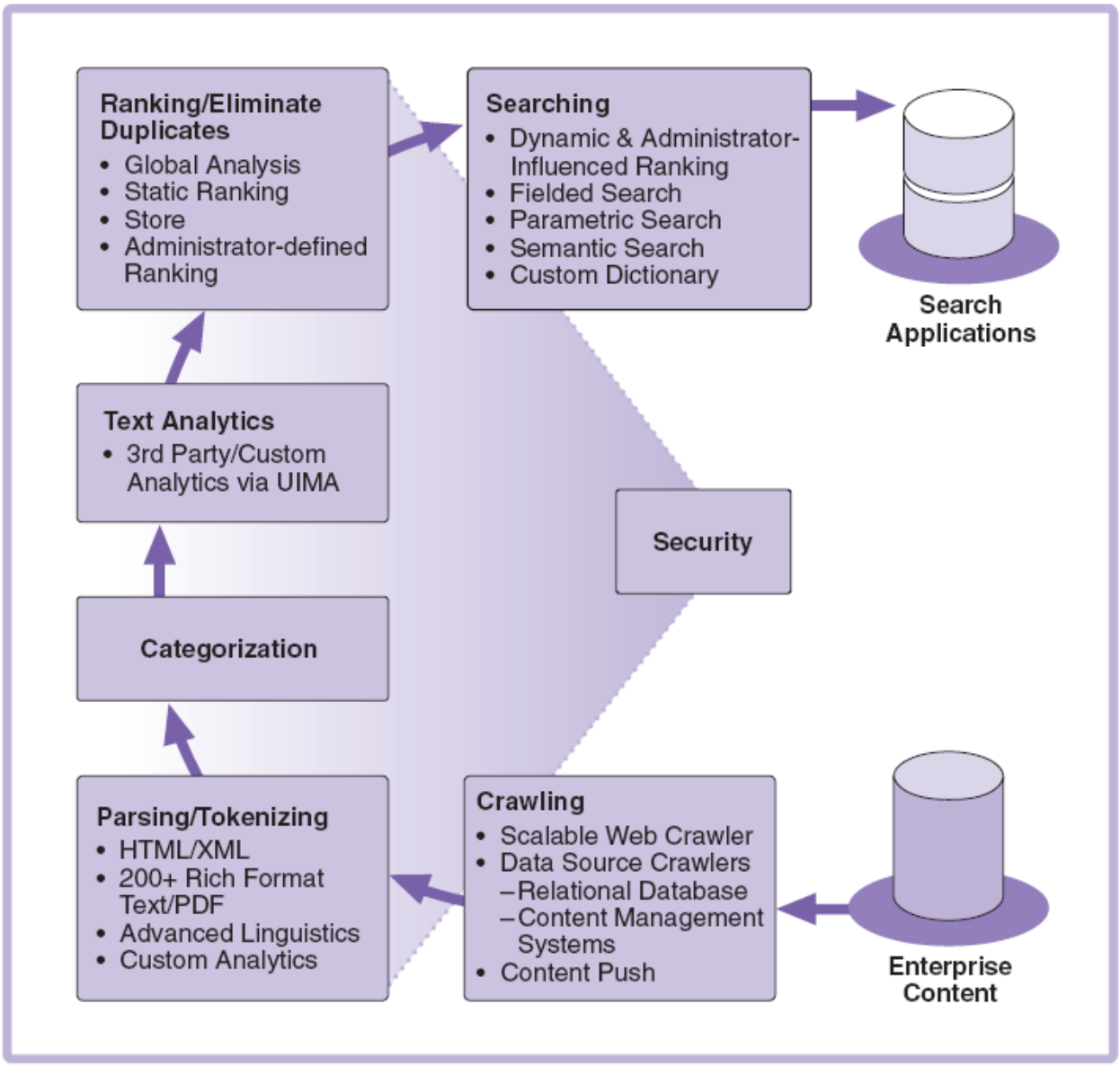
Наконец, во время поиска, запросы могут быть скорректированы по орфографии, или списку терминов, которые возникают в сравниваемой коллекции, по словарям синонимов или специфичного отраслевого словаря, и результаты будут динамично ранжироваться, основываясь на терминах, содержащихся в запросе.

Во время индексирования, текст извлекается и анализируется, используя передовые лингвистические технологии. Процессы в сервере индексации анализируют структуру ссылок для интранет контента, перемещают дублирующийся контент и производит другие

действия с коллекциями документов для улучшения качества результата. С помощью UIMA может быть проведен расширенный анализ документа для генерирования дополнительных метаданных, которые при добавлении в индекс позволяют осуществлять поиск по объектам или фактам, скрытым внутри неструктурированной информации. Опция инсталляции двух поисковых подсистем предоставляет собой резервирование, для гарантии того, что поиск всегда будет доступен.

В дополнении к этим возможностям, архитектура решения – открытая и расширяемая для более легкой поддержки широкого разнообразия индустриальных или специфичных для предприятия приложений поиска. Используя интерфейс Data Listener API, сборщики (crawlers) настраиваются на нестандартные пользовательские приложения для отражения их контента в индексе. Это позволяет ИТ департаментам добавлять их собственные, разработанные приложения в список источников, по которым осуществляется поиск. Эти API также могут использоваться для «подталкивания» контента в сервер индексации по расписанию, которое определяется бизнес-процессами, такими как публикация зависимых от времени материалов, как пресс-релизы или информация о новых продуктах на внешних сайтах.

Для доступа к возможностям посика из внешних приложений определен Java API для легкого встраивания в существующие системы предприятия. С использованием UIMA, IBM OmniFind Enterprise Edition также предоставляет основу для подключения к расширенной аналитики, что позволяет осуществлять домен – специфичный поиск и т.д.



Высокая релевантность результатов посика в зависимости от специфики языка

Решение IBM OmniFind Enterprise Edition использует продвинутый алгоритм ранжирования (ranking), разработанный IBM Research и расширенный лингвистический анализ, основанный на IBM LanguageWare, для предоставления результатов поиска очень высокого качества. Эти механизмы не только оценивают и ранжируют документы в течение предварительной обработки (до того, как индекс построен), они также включают запатентованные особенности ранжировать результаты, основываясь на типе запроса, который пользователь ввел. Другими словами, когда пользователь вводит строку поиска, система может динамично распознать тип запроса и упорядочивать ранжирующие факторы соответственно. Ранжирующие факторы различны, например, по запросы «ibm» от конкретного пользователя, система предложит вариант «как поменять пароль к системе». Эта технология поиска сейчас применяется на корпоративном портале IBM.

Поддержка пользовательских словарей синонимов

Использование синонимических словарей (иногда они называются «тезаурус») улучшают результаты поиска и помогают пользователям расширять запрашиваемые термины с

помощью списка синонимов. Так как качество результата зависит от качества запроса, очень важно сохранять качественный уровень каждого запроса высоким. Синонимы предоставляют способ улучшить качество запроса с помощью сходных ключевых слов, которые могут предоставляться пользователям немедленно, и быть предопределенными и, возможно, подразумеваемыми. Синонимы полезны для названий продуктов, для отраслей промышленности и даже для специфических для компании терминов. Синонимы также чрезвычайно полезны для поиска акронимов (слова, образованные по первым звукам или слогам словосочетания, которые они заменяют) или их расширенных форм в запросах.
Например, в среде центра обработки звонков, этот расширенный поиск может быть очень ценен. Представитель по работе с заказчиками может искать информацию по термину «машина», однако это потребует довольно длительного времени, а с помощью синонимов, можно значительно сократить это время, сразу определив тип информации, который ему необходим в разделе «машина». Решение IBM OmniFind Enterpsise Edition может автоматически расширять поиск с помощью списка синонимов, которые были определены для, например, машин. Поиск по термину «машина» может означать поиск «фургона», «грузовика», «кабриолета», «универсала» и т.д. Пользователь в строке запроса пишет только один термин, хотя, подразумевать он может целую группу слов, описывающих машину. С помощью синонимичного ряда, IBM OmniFind Enterprrise Edition помогает понять и предугадать запрос пользователя.

Влияние на ранжирование

Алгоритм ранжирования может быть настроен администратором для подчеркивания важности определенного содержимого:

- Создание пользовательского словаря стоп-слов. Стоп-слова – часто используемые слова, не включаемые в поисковые индексы, как например, определенные и неопределенный артикли («a» и «the» в английском языке). Вы можете расширить стандартный список стоп-слов IBM OmniFind Enterprise Edition.
- Создание пользовательского словаря слов-акселераторов для увеличения или уменьшения “веса” перечисленных в нем терминов. Решение IBM OmniFind Enterprise Edition изначально устанавливает “веса” терминов на основании относительной частоты этих терминов в коллекции[1]. Например, предлоги встречаются в тексте очень часто, однако их значимость при определении наиболее релевантных документов крайне низка. Возможность влиять на “веса” позволяет управлять значимостью терминов в рамках коллекции.
- Назначение “весов” полям метаданных. Данная возможность позволяет влиять на важность слов в зависимости от того, в каких атрибутах документа они могут быть найдены.
- Повышение релевантности документов на основании соответствия URI документа заданному шаблону. URI-шаблон задает тип документов и информацию в иерархической форме, определяющую положение документов в хранилище – чем глубже маршрут, тем уже область документов, соответствующих ему. Эта возможность позволяет повысить важность отдельных документов или их наборов (например, таблиц DB2 или объектов Lotus Notes).

Многоязыковая поддержка

[1] Коллекция (Collection) – экземпляр системы IBM OmniFind, содержащий набор данных (документов) из разных источников, текстовый анализатор, сервер индексации и поисковый механизм. В рамках решения может существовать несколько коллекций, администрация которых происходит независимо друг от друга.

В сегодняшних организациях, высококачественное решение по поиску должно бесшовно поддерживать множество языков. Это может быть достигнуто только за счет создания механизма расширенного поиска и технологии интеграции лингвистических особенностей разных языков. Такая лингвистическая технология требует не только механизм, предоставляющий высококачественный многоязычный поиск, но также основу, на которой более продвинутые поисковые возможности могут бесконечно достраиваться.
Несколько примеров сложностей, которые могут возникнуть при работе с многоязычным поиском, и их решение в IBM OmniFind Enterprise Edition:

- Нормализация слова (обычно называемая «лемманизацией») важна при работе с «гибкими» языками, но она критично необходима для работы с высоко-гибкими языками, такими как русский. Лемманизация – это приведение слова к его словарной форме. Этот алгоритм широко используется в IBM OmniFind Enterprise Edition для обеспечения эффективного поиска всех словоформ. В общем случае, лемманизация не может быть достоверно заменена простым морфологическим поиском, который механически убирает окончание слова для создания его корня, что отличается от механизмы нахождения леммы. Эти «корни» не всегда являются словами с семантическим значением, или даже хуже, могут иметь абсолютно противоположный смысл в отличие от его словоформы. Например, слова «организовывать», «организация» - похожи, но механически отделенные окончания от них останется корень «орган», который в семантическом смысле не будет иметь ничего общего с первыми двумя словами. Иными словами, простое механическое убирание окончания, не подскажет вам взаимоотношения между словом и его словоформой.
- Сегментация текста необходима для точного понимания смысла запроса в таких языках, как китайский и японских.
- Сегментация слова необходима для разбивки слова на семантические единицы в таких комплексных языках, как германский.
- Написание слова в нескольких вариантах (варианты нормы) встречаются довольно часто во многих языках. Например, в немецком языке – есть значительная разница между reform и pre-reform, английский язык отличается от американского во многих случаях написания слов, китайский традиционный и упрощенный - также существенно разнятся.
- Префиксы и суффиксы могут маскировать семантический компонент слова в таких языках как арабский.
- Нормализация знаков - приведение их к стандартному виду - также важная языковая характеристика, которая может повлиять на качество запросов. Например, в европейских языках существуют буквы, употребляющиеся с определенными значками (например, в немецком – умляуты (две точки над буквами), во французском - буквы с аксанами (или их называют акцентами) - наклонными палочками). Суть нормализации знаков – сведение всех "форм" знаков к одному виду: например, u с умляутом можно свести к обычному u и производить поиск без учета наличия дополнительных знаков и т.д.

IBM OmniFind Enterprise Edition предоставляет решение по многоязыковому поиску посредством интеграции с лингвистической технологией IBM LanguageWare.
IBM LanguageWare – сложная лингвистическая технология, является результатом многолетних исследований с привлечением специалистов из многих стран мира. В противоположность многим другим лингвистическим технологиям, данная технология описывает в точности, как нужно анализировать язык и позже индексировать. Технология LanguageWare – открытое, гибкое и изменяемое средство, которое благодаря этим качествам может быть изменено в соответствии с потребностями конкретного

предприятия. Есть возможность дополнить стандартные лексико-семантические ресурсы LanguageWare пользовательской терминологией, правилами и отношениями. В результате получается многоязыковое поисковое приложение, индексирующее содержимое, учитывая как особенности языков, так и пользовательские словари, используемые в компаниях.

Быстрое нахождение ключевых ресурсов

Для того чтобы пользователю было проще просмотреть результаты с различных сайтов, IBM OmniFind Enterprise Edition поддерживает возможность запрета появления дополнительных результатов поиска с одного и того же сайта в списке приоритетных результатов, позволяя большому количеству результатов из различных источников появляться в браузере пользователя. Например, предоставление результатов поиска может быть конфигурировано таким образом, чтобы в списке результатов появлялось не больше двух высоко-релевантных результатов с каждого сайта.
Каждая организация имеет набор ключевых сайтов, быстрый и надежный доступ к которым необходим для пользователей, как, например, медицинские формы, информация отдела кадров, процедуры командировок и т.д. Не имеет значения, как назван документ, или насколько плохие у него метаданные, у администратора есть возможность настроить предопределенный результат по некоторым запросам, который будет представлен вместе с другими релевантными результатами поиска. Эти быстрые ссылки (QuickLinks) могут значительно сократить время поиска для сотрудников, а также повысить удовлетворенность клиентов от поиска на внешних сайтах предприятия. Например, такие запросы как «пенсионный возраст», «пенсия», «пенсионный план» могут быть распознаны быстрыми ссылками как поиск сотрудником информации, находящейся в отделе кадров и предоставить ему ссылку на начальную страницу отдела кадров.

Релевантные, динамические резюме документов

Названия документов могут не предоставлять достаточно информации о его содержимом. По этим причинам, пользователям необходима возможность понять суть документа без дополнительных, требующих времени, усилий, таких как прохождение по каждой из предложенных ссылок. IBM OmniFind Enterprise Edition использует слова в запросе поиска, на основе которых формируется резюме документа, основываясь на фразах, которые лучше всего представляют суть документа, который пользователь ищет. Динамическое резюме делается более наглядным путем подчеркивания искомых терминов. Выделенное резюме может быть отображено в списке результатов, так чтобы пользователь мог легко оценить контент и релевантность документа. В противоположность другим решениям поиска, которые предоставляют статическое резюме, извлекая специфичное, заранее сохраненное резюме документа, IBM OmniFind Enterprise Edition создает интеллектуальное и динамичное резюме из содержимого документа.

Коррекция запросов, основываясь на ваших данных.

Использование IBM LanguageWare в IBM OmniFind Enterprise Edition расширяет процессы индексирования и запросов. Представьте интерфейс запросов, который был бы достаточно умен, чтобы распознать, когда запрос неоптимален и предложить другую, интеллектуальную альтернативу. IBM разрабатывает интерфейс запросов таким образом, чтобы привнести эти инициативы в решение путем интеграции лингвистических инструментов с интерфейсом запросов. Эта интеграциz помогает легче преодолеть

ошибки, которые могут возникнуть при неправильном написании слов и предложит интеллектуальную альтернативу пользователю.
IBM OmniFind Enterprise Edition может генерировать альтернативы запросы при неправильном вводе слов в запросе. Решение IBM OmniFind Enterprise Edition содержит знание об общем написании слова или об ошибке при наборе слова для всех поддерживаемых языков и может предоставить значимые замечания при неправильном написании. IBM OmniFind Enterprise Edition также использует словарь LanguageWare, который настроен таким образом, чтобы предлагаемый на замену термин был специализирован и использовался в вашей пользовательской группе.

Гибкая поддержка категорий

Таксономия – это способ категоризировать контент для того, чтобы пользователи легче находили релевантную информацию. Решение IBM OmniFind Enterprise Edition включает классификатор, основанный на правилах, который предоставляет простой, гибкий способ настраивать правила, используемые в создании категорий в таксономии. Пример правила «Найти любой документ, содержащий термин «автомобиль», но не содержащий слово «дилер» и поместить его в категорию «автомобильная индустрия». Классификаторы, основанные на правилах, быстры, легко настраиваемы и работают лучше всего с небольшим количеством связанных категорий.

Поддержка запросов для новых и опытных пользователей

Каждый человек сейчас знает, как найти информацию, задавая вопрос в свободной форме. Запросы в свободной форме – просты и интуитивно понятны. При поиске, задаваемом в свободной форме, пользователи просто могут вписать слово, фразу или предложение, и механизм поиска найдет документы, релевантные запросу. Эта концепция поиска широко используется в Интернете.
Однако опытные пользователи и специализированные приложения нуждаются в более богатом и более эффективном языке запросов. Сложные запросы с логическими операторами позволят вам идентифицировать специфичные атрибуты документов в вашем запросе, такие как язык документа, тип или источник. Дополнительно, IBM OmniFind Enterprise Edition распознает структуру документа и может использовать эту структуру для специально сконструированного поиска по секциям документа или для специализированных метаданных, связанных с этим документом или для ограничения границ поиска в рамках сайтов. Поисковый запрос должен включать имя поля. Иначе, эта возможность должна быть сделана внутри приложения для поиска. Несколько примеров:

- Для баз данных Lotus Notes, поиск по полям.
- Для XML документов, поиск внутри определенных элементов (<AUTHOR>, <TITLE>, <SUBJECT>). Для HTML документов, поиск внутри определенных признаков (элементы <META>, <TITLE>, URL), например, найти все документы с названием «OmniFind» на сайте ibm.com
- Для новостных групп, поиск в индивидуальных группах
- Для MS Office или IBM Smartsuite документов, поиск по свойствам документа, таким как автор, название или описание документа.

Дополнительно, IBM OmniFind Enterprise Edition может производить сравнительные или оценочные запросы в цифровых и информационных полях и метаданных (параметрический поиск). Например, с файловой системой UNIX, вы можете искать документы определенного размера или созданные после определенной даты. Для баз данных, вы можете специализировать поиск, основанный на полях в базах данных, как документы, в которых цена за единицу больше чем рыночная стоимость.

Семантический поиск - отличительная особенность IBM OmniFind Enterprise Edition

Хотя поиск по ключевым словам может быть мощным средством в поиске документов, основанных на метаданных и ключевых словах, как таковых, он не предоставляет возможность поиска по высокоуровневым объектам, таким как имена людей, телефонные номера, по частям документа или условиям его создания. Приложения семантического поиска расширяют запросы пользователей от простых, по ключевым словам, до нахождения таких высокоуровневых объектов как люди, места, организации, продукты, и другие объекты, которые могут возникнуть в тексте. Кроме того, эти приложения позволят пользователям специфицировать взаимоотношения между этими концептами. Механизм такой работы базируется на возможности поиска улучшать метаданные, которые аналитика текста предоставляет в базе информации. Семантический поиск – это возможность понять и двигаться через расширенные метаданные, предоставляющие пользователям более релевантные результаты, и возможность, например, найти объекты вне зависимости от того, как они описаны в тексте (например, поиск документа, в котором упомянута некая персона, но имя ее неизвестно, или документ, содержащий дату, но день этот неизвестен).

Высокая доступность и маштабируемость

В дополнение к высококачественным результатам поиска, IBM OmniFind Enterprise Edition разработан для предоставления непревзойденной маштабируемости и производительности. Единичный сервер индексации может в настоящий момент маштабироваться до 20 миллионов документов, а набор серверов индексации в настоящий момент может обслуживать даже более крупные предприятия.
IBM OmniFind Enterprise Edition был испытан на прочность и масштабируемость в собственном интранете компании IBM w3.ibm.com с количеством пользователей более чем 300 тысяч.

Открытая архитектура для расширяемости и интероперабельности

Архитектура IBM OmniFind Enterprise Edition предоставляет открытую и расширяемую инфраструктуру для интеграции со сторонними приложениями и подключения пользовательских расширений. Это предоставляет богатый набор средств и интерфейсов для упрощения и улучшения интеграции с широким разнообразием индустриальных или специфичных для компаний приложений поиска:

- Пример поискового приложения, разработанного с помощью стандартного API, переведенного на 21 язык и позволяющего легко настраивать внешний вид. Быстро устанавливается и может быть использовано как для демонстрации и тестирования, так и в качестве основы для построения приложения для заказчика. Включены функции простого и расширенного поиска, навигации по дереву категорий, анотации к найденным документам, поиск в результатах поиска и многое другое;

- Пример портлета для легкого встраивания в приложение WebSphere Portal;

- Интеграция с универсальным центром поиска (Universal Search Center) в WebSphere Portal, позволяющая пользователям центра поиска включать коллекции IBM OmniFind Enterprise Edition для поиска, и позволяющая администраторам добавлять или удалять коллекции из центра поиска;

- Интеграция с инструментами для поиска на компьютере, включая Google Desktop Search для предприятий и X1 Desktop Search, позволяющая пользователям использовать знакомый интерфейс для работы с корпоративным поиском;

- Мощная платформа для анализа текста основанная на Unstructured Information Management Architecture (UIMA), позволяющая разработчикам подключать дополнительные, специфичные для предметной области текстовые анализаторы;

- Удобный Java API, упрощающий интеграцию с существующими приложениями.

Java API построен на базе стратегического интерфейса IBM для поисковых решений, использующих индексацию (Search и Indexing API, или "SIAPI"), позволяющий разработчикам создавать приложения, которые могут работать с любыми решениями, предоставляющими сервис поиска (например, WebSphere Portal и IBM OmniFind Enterprise Edition). API также включает возможности администрирования коллекций и добавления документов в коллекции. Дополнительно, для приложений с постоянно-обновляющейся информацией, API позволяет передавать данные поисковой подсистеме вместо того, чтобы ждать пока система заново соберет информацию из источников данных - таким образом, вы можете быть уверены, что последняя информация будет доступна для поиска. Это также позволит вам создавать пользовательских сборщиков.

Простое, гибкое администрирование

Консоль администрирования IBM OmniFind Enterprise Edition предоставляет единую точку входа для инсталляции, конфигурирования, мониторинга и других административных задач. Легкий в использовании, основанный на web, интерфейс предоставляет администратору средства для поддержания непрерывной работы системы.

- Вам необходимо только специфицировать, что искать и когда начать индексирование.
- Искомый контент может быть динамично расширен без постройки новых индексов поиска.
- Администратор может следить за активностью поиска с помощью резюме статусов, которые включают производительность, сферу деятельности, возникающие ошибки, время отклика, поисковый рейтинг и оценку запросов.

Система может посылать напоминания на любой электронный адрес для помощи администраторам. Например, система может определять, если оставшееся пространство диска для одного поискового механизма меньше, чем 100 мегабайт, или время отклика больше обозначенного значения, или скорость сбора информации сборщика меньше указанного значения, система пошлет напоминания о существующей проблеме.
Внутри типичного предприятия, множественные коллекции будут созданы для различных приложений. Отдельная коллекция может быть создана для финансистов (с их собственным набором пользователей, ищущих внутри этой коллекции), и другая – для отдела кадров. Коллекции могут содержать документы из множества источников данных. Каждое приложение и ему соответствующая коллекция может требовать экспертизы разных людей в организации для помощи в настройках и конфигурировании коллекций.
Решение IBM OmniFind Enterprise Edition поддерживает эти разнообразные требования, чтобы позволить администраторам устанавливать роли для одной и более специфичных

коллекций. Поэтому индивидуальные коллекции могут иметь своих собственных администраторов и операторов, гарантируя, что люди, которые знают и понимают контент будут иметь права администрировать их.

Надежный набор элементов безопасности

Технология поиска внутри предприятия должна предоставлять строгую систему безопасности для гарантии того, что пользователи в результате поиска не увидят документы, доступ к которым они не имеют. Решение IBM OmniFind Enterprise Edition реализует безопасность за счет механизмов аутентификации и авторизации на двух уровнях доступа.

Аутентификация – это процесс, с помощью которого система верефицирует пользователей по принципу «кто они сказали они такие». Решение IBM OmniFind Enterprise Edition было разработано для работы с уже существующими компонентами аутентификации и не требует отдельного процесса регистрации для конечных пользователей. Когда IBM OmniFind Enterprise Edition требует идентификации зарегистрированного пользователя, решение взаимодействует с изначальной средой (например, WebSphere Portal) или приложением, к которому пользователь уже имеет доступ. Этот подход позволяет IBM OmniFind Enterprise Edition бесшовно интегрироваться с уже существующими политиками аутентификации на предприятии без необходимости отдельно регистрироваться пользователям.

Авторизация – это механизм, с помощью которого система дает или аннулирует права доступа к данным, или производит некоторые действия. В решении IBM OmniFind Enterprise Edition разработано несколько уровней контроля доступа, которые могут использоваться как совместно, так и раздельно, для предоставления возрастающего уровня авторизации. Доступ к коллекциям может быть ограничен только приложением поиска, которому может быть дан доступ администратором.

Например, многие копии поискового портлета IBM OmniFind Enterprise Edition могут быть созданы в сервере WebSphere Portal и размещены на различных страницах портала. Каждая копия портлета поиска может быть конфигурирована таким образом, что только определенные коллекции доступны для поиска через данную копию поискового портлета. Контроль доступа портала (Portal Access Control) может использоваться для контроля, кто имеет доступ и к какому поисковому портлету. Следовательно, одна копия поискового портлета может быть настроена для обслуживания финансового департамента и сконфигурирована в соответствии с тебованиями данного департамента, в терминах такой коллекции, которая доступна для поиска.

IBM OmniFind Enterprise Edition также контролирует доступ к документам для гарантии, что конечные пользователи могут видеть только строго разрешенные документы. В решение IBM OmniFind Enterprise Edition уровень доступа к документам достигается за счет разрешения администраторами объединить один или более защитных маркеров доступа с каждым документом во время работы сборщиков.

Следующие методы могут использоваться для спецификации защитных маркеров:

- Значение может быть извлечено из списков контроля доступа (ACL) репозиториев back-end системы, для которых эти списки доступа поддерживаются.
- Значение может быть задано администратором, использующим административную консоль
- Значение может быть извлечено из полей, созданных администратором в собранных документах.
- Значение может быть определено пользовательской Java-программой через API-интерфейс WSIIOFE под названием "плагины маркеров безопасности"

Во время поиска, приложение поиска гарантирует, что пользователь поиска в настоящий момент полностью аутентифицирован и получил корректные защитные маркеры. Когда пользователь подтверждает запрос, его защитные маркеры должны соотносится с маркерами сохраненного документа для гарантии, что пользователь не увидит документа, к которому у него нет доступа. Если исходный список доступа (ACL) используется как защитный маркер, поисковая подсистема проверяет проходили ли какие-либо изменения авторизации back-end системе после работы сборщиков.

Поисковый механизм в средах Lotus Notes/Domino

Решение IBM OmniFind Enterprise Edition – наиболее предпочтительная платформа для поиска в Lotus Notes/Domino. Она поставляется со специализированным сборщиком для Lotus Notes/Domino, который может быть сконфигурирован непосредственно внутри административного клиента IBM OmniFind Enterprise Edition. Также в комплекте поставки есть приложение для поиска, которое может открывать базы данных Lotus Notes и документы непосредственно из клиента Lotus Notes или в Web браузере. Сборщики IBM OmniFind Enterprise Edition могут быть конфигурированы также на поиск в представлениях Lotus Notes, которые созданы администратором системы.
IBM OmniFind Enterprise Edition использует исходные интерфейсы Domino (как NotesRPC или Domino IIOP); распознает структуру баз данных Lotus Notes от полей до приложений; поставляются готовые сборщики для QuickPlace и Domino.Doc (DDM), также как и возможность собирать документы из WorkPlace Web Content Management.
В последней версии IBM OmniFind Enterprise Edition, наиболее значимая функция безопасности – это поддержка безопасности коллекции из изначального репозитория Lotus Notes/Domino Server. Если опция проверки безопасности выбрана в течение обработки запроса, IBM OmniFind Enterprise Edition предложит Lotus Notes/Domino Server Access Control Lists (ACLs) в реальном времени определить, имеет ли пользователь доступ к документам Notes/Domino Server. Эта опция гарантирует защищенный поиск вне зависимости от того, насколько быстро происходят обновления в Access Control Lists.

Расширенные возможности IBM WebSphere Portal

Решение IBM OmniFind Enterprise Edition было разработано для работы в среде IBM WebSphere Portal. IBM WebSphere Portal поставляется с механизмом поиска для Portal, который может использоваться для индексирования Web, файловой системы, и документов Lotus Notes/Domino Server. Однако, оно разработано для поддержки коллекций документов от 800,000 и меньше, в то время как IBM OmniFind Enterprise Edition поддерживает миллионы документов и множество источников данных на предприятии.
Для построения индексов могут использоваться возможности поиска от WebSphere Portal, утилита, которая поддерживает миграцию этих конфигурационных настроек в IBM OmniFind Enterprise Edition, позволяя бесшовно мигрировать настройки в возможности предоставляемые с IBM OmniFind Enterprise Edition, в то же время позволяя поиску от WebSphere Portal использовать поисковый центр (Search Center) или портлеты поиска.

Образец поискового портлета WebSphere Information Integrator OmniFind Edition

Этот портлет похож по внешнему виду и функциям на портлет WebSphere Portal, но также предоставляет все расширенные возможности и настройки от IBM OmniFind Enterprise

Edition, такие как контроль отображения, категоризация, количество выводимых результатов поиска на страницу, как сортировать результаты поиска и т.д.

Решения IBM OmniFind Enterprise Edition – платформа для аналитики текста

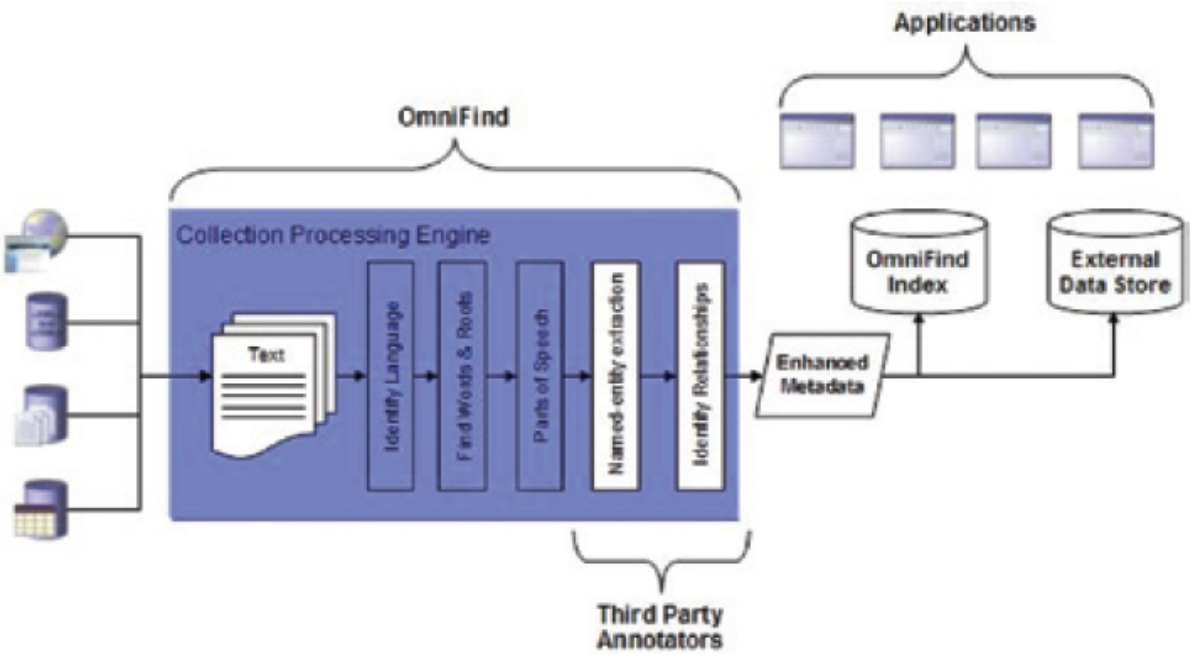
В дополнение, к предоставлению инфраструктуры поиска для предприятия, IBM OmniFind Enterprise Edition также и первая доступная, основанная на UIMA платформа для обработки текстовой информации.

Улучшение качества поиска результатов

UIMA – это открытая, стандартная основа для обработки неструктурированного контента для генерации более описательных и семантически – богатых метаданных. UIMA позволяет бесшовно интегрировать компоненты текстового анализа, которые анализируют документы и извлекают дополнительные знания для создания расширенных индексов для поиска.
Также на основе механизма UIMA можно идентифицировать высокоуровневые объекты, такие как люди, места, организации, названия продуктов, и другие «объекты», которые скрыты в неструктурированных данных.
Качество результатов значительно возрастает за счет поиска пользователями в более полных метаданных. Точность и релевантность результатов поиска можно значительно улучшить за счет усиления исследуемых зависимостей и фактов. Это позволит пользователям более легко находить специфические документы, релевантные зависимостям или фактам.

Платформа для текстовой аналитики

Решение IBM OmniFind Enterprise Edition предоставляет возможность собирать и синтаксически анализировать контент из различных источников, анализировать текст на основе механизма UIMA, индексировать расширенные метаданные для возможности более комплексного поиска, и посылать извлеченную информацию в базы данных, хранилища, и другие внешние базы данных знаний.
IBM OmniFind Enterprise Edition включает некоторые основные возможности текстовой аналитики, такие как расширенная лингвистика, извлечение корней слов и поддержка тезариуса для идентификации синонимов.
С другой стороны, компоненты текстовой аналитики могут быть вставлены и конфигурированы через стандартные средства администрирования для генерации дополнительных метаданных. В дополнение, IBM предоставляет заказчикам образцы экстракторов объектов, которые демонстрируют как улучшать индексирование контента для людей, мест, названий продуктов и других объектов, скрытых в неструктурированном контенте. Многие из этих образцов могут быть настроены для специальных объектов в компании. Наконец, IBM OmniFind Enterprise Edition позволяет пользователям улучшить поиск по ключевым словам или семантический поиск для нахождения наиболее релевантного контента.



Улучшенные метаданные, извлеченные из неструктурированного контента, могут также быть отосланы во внешние источники данных, такие как базы данных или хранилища, где стандартные средства отчета могут использоваться для улучшения анализа новых наборов знаний. Это позволяет компаниям более быстро находить несовместимости, идентифицировать посторонние значения и исследовать тенденции, которые были раньше скрыты в документах. Результатом более быстрого исследования информации будет являться возможность получать прямые выгоды для бизнеса, такие как сокращение затрат за счет быстрого определения и анализа проблемы, минимизации связанных с этим риском, и убыстрение времени для поиска на рынке новых предложений.

Максимизация ценности существующих внедрений текстовой аналитики

Большинство компаний, которые внедряли аналитику текста, делали это, как правило, для специфичных приложений или инициатив. Они инвестировали большое количество времени и денег для фиксирования и систематизирования методов и терминологий, связанных с их бизнесом, и извлекали знания из различных наборов неструктурированного контента. Это также обычно требует глубокой интеграционной работы. После всех этих усилий, компании обычно ограничивают использование наборов информации до наименьшего количества пользователей и исключительно для целей глубокой аналитики.
Однако, усиливая разработанную платформу при помощи интеграции с поиском внутри предприятия и бизнес-приложениями, для которых аналитика текста и была разработана, компании могут начать получать большие преимущества от этих инвестиций.

Примеры приложений

Большинство приложений для бизнес-аналитики создают отчеты и анализируют структурированные данные, но будут не в состоянии работать с неструктурированным контентом, таким как контракты, напоминания, корреспонденция заказчика, или данные в свободной форме – комментарии, картинки операций, рисками и т.д. Используемая аналитика текста может давать возможность извлекать факты из неструктурируемого

контента и соединять эти знания в отчет и анализ для более комплексной и аккуратной картины организации, включая их продукты, сервисы, заказчиков и поставщиков. Уникальная комбинация IBM OmniFind Enterprise Edition и механизма UIMA создает новый класс текстово-аналитических приложений. Некоторые из этих решений поставляются вместе с приложениями, разработанными бизнес-парнерами IBM.