

ABBYY Retrieval & Morphology Engine

К.Анисимович, В.Балл, В. Селегей
компания ABBYY,

www.abbyy.com

Аннотация

В сообщении представляется новый программный продукт, относящийся к классу “engines”, то есть, предоставляющий пользователю некоторые технологии не в составе прикладных программных продуктов, а в “чистом” виде. Технологии полнотекстового поиска и морфологического анализа ABBYY Retrieval & Morphology Engine могут быть встроены в любые программные системы и комплексы как готовые решения.

ABBYY Retrieval & Morphology (ARM) Engine – это инструментальный разработчика программного обеспечения, предназначенный для интеграции технологий полнотекстового поиска и морфологического анализа компании ABBYY Software House в локальные и клиент-серверные приложения разработчика через 32-битный COM (Component ObjectModel) API (Application Programming Interface).

Retrieval API ARM Engine позволяет разработчикам встраивать в свои приложения функции быстрого и эффективного поиска нужных документов среди огромных массивов текстовой информации в разнообразных электронных хранилищах: корпоративных информационных системах, архивах на CD, WEB серверах и т.п. Для обеспечения этого результата в ARM Engine применяются эффективные поисковые алгоритмы и знания о морфологии 30 естественных языков.

Morphology API ARM Engine дает разработчику доступ к функциям морфологического анализа для 30 языков (с разным объемом интерфейса для разных групп языков).

Основная функциональность ARM Engine

Функции Retrieval API:

- Полнотекстовая индексация. Инструментарий позволяет создать эффективный полнотекстовый индекс многоязычных документов. Индекс хранит информацию о словах и их местоположении в информационном хранилище (файловых серверах, WEB-серверах или базах данных). При внесении каких либо изменений в информационный источник (удалении, копировании или редактировании) происходит обновление индекса в реальном времени. Скорость индексирования – до 15 MB/мин.
- Полнотекстовый поиск. ARM Engine позволяет осуществлять простой и расширенный поиск с использованием морфологии. При простом поиске запрос состоит из одного или нескольких слов, а также логических операторов И, ИЛИ, НЕ. Запрос для расширенного поиска представляет собой строку на специальном «языке запросов», что позволяет более точно задать параметры поиска.
- Нечеткий поиск. ARM Engine позволяет осуществлять нечеткий поиск, т.е. находить слова, которые отличаются от исходного несколькими символами. Такая возможность полезна в случае, если документы содержат слова с ошибками.

Функции Morphology API:

- Проверка орфографии (функции Check и Suggest);
- Лемматизация (определение словарной формы по произвольной форме);
- Определение словоизменительного типа несловарного слова;
- Генерация парадигмы по произвольной словоформе;
- Грамматическая атрибуция произвольной словоформы;
- Расширение словаря.

Практическое применение ARM Engine

ARM Engine – это инструментарий разработки приложений, технологии которого успешно используются как компанией ABBYY в ее прикладных разработках, так и другими компаниями. Для примера можно назвать следующие системы:

- *ABBYY Lingvo*, система англо-русско-английских электронных словарей, активно использующая ARM технологии для полнотекстового поиска в словарях и перевода слов и словосочетаний непосредственно из текста.
- *ABBYY FineReader*, многоязычная система оптического распознавания символов (OCR) Высокое качество распознавания, обеспечиваемое этой системой, получившей более 20 международных наград от ведущих компьютерных изданий, во многом определяется использованием ARM технологий для контекстного анализа и проверки орфографии.
- *ABBYY Internet Index Server*, приложение, созданное на основе ARM Engine, позволяет эффективно осуществлять многопоточный поиск документов в сети Internet.

- *Архив «Независимой» газеты.* Одна из наиболее авторитетных российских газет использует ARM Engine для ежегодно разрабатываемого архива материалов, выпускаемого на CD.
- *Компания Siemens Nixdorf* использует ARM технологии как составную часть электронного архива ARCIS - системы архивации, обработки и хранения информации.

Спецификация:

Поддерживаемые языки

ARM Engine использует актуальные словари и морфологические описания для 30 языков: Английский, Армянский (Восточный, Западный, Грабар), Болгарский, Венгерский, Голландский (Бельгийский, Стандартный), Греческий, Датский, Испанский, Итальянский, Латышский, Литовский, Немецкий, Немецкий (новая орфография), Норвежский (Букмол), Норвежский (Нюнорск), Польский, Португальский (Бразильский, Стандартный), Русский, Словацкий, Турецкий, Украинский, Финский, Французский, Хинди, Чешский, Шведский, Эстонский.

Возможен поиск и индексирование в многоязычных документах, содержащих любые комбинации вышеперечисленных языков.

Все языки, поддерживаемые ARM Engine делятся на три группы по степени поддержки функций:

1. Группа с полной поддержкой всех функций. В нее входят Английский, Армянский (Восточный, Западный, Грабар), Испанский, Итальянский, Немецкий, Немецкий (новая орфография), Русский, Украинский, Французский.
2. Группа языков с полной морфологической моделью, но со словарями, полученными полуавтоматически, в результате обработки больших корпусов текстов с использованием морфологической модели. Такие модели не обеспечивают 100% правильности грамматической атрибуции словоформ, лемматизации и получения парадигмы. Тем не менее, они позволяют производить полнотекстовый поиск, проверку орфографии, пополнение словаря (остальные индо-европейские языки).
3. Языки с неполной морфологической моделью и автоматически полученным словарем. К таковым сейчас относятся турецкие и финно-угорские языки (турецкий, венгерский, финский, эстонский). Эти языки обеспечивают только проверку орфографии и полнотекстовый поиск с использованием частичных парадигм.

Поддерживаемые форматы документов и кодировки

ARM Engine позволяет индексировать документы в форматах TXT и HTML в следующих кодировках:

- Русский – Windows-1251, KOI8-R, Cyrillic ISO (ISO-28595), OEM-866;
- Польский, Чешский – 1250;
- Болгарский, Украинский – Windows-1251;
- Греческий – Windows-1253;
- Турецкий – Windows-1254;
- Эстонский – Windows-1257;
- для остальных поддерживаемых языков – Windows-1252.

Среды разработки

ARM Engine успешно протестирован на совместимость со следующими языками программирования:

- Microsoft Visual C++ 4.x и выше;
- Microsoft VB 5.0 и выше;
- Delphi 3.0 и выше.

Системные требования:

- PC на основе процессора Intel от 486DX и выше;
- Windows 9x или Windows NT 4.0 для клиентской части, и Windows NT 4.0 для серверной части;
- MS Internet Explorer 3.x и выше для регистрации COM-серверов;
- 16 Мб RAM для Windows 9x, и 32 Мб для Windows NT 4.0;
- от 3 до 30 Мб свободного пространства на жестком диске для лингвистической базы данных (в зависимости от количества устанавливаемых языков);
- минимум 10 Мб свободного пространства для хранения временных файлов (в зависимости от размера индексной базы).

Испытательная версия

Для того, чтобы потенциальные пользователи ARM Engine могли протестировать этот инструмент в реальных условиях без каких бы то ни было функциональных ограничений, мы предлагаем бесплатную версию ABBYY Retrieval & Morphology Engine на ограниченный период времени. Испытательная версия работает 60 дней, после чего все ее функции становятся недоступными. Условия использования Испытательной версии ARM Engine регулируются Соглашением об использовании Испытательной версии. Для получения Испытательной версии, пожалуйста, обратитесь в компанию ABBYY по адресу engine@abbyy.ru.