ОБРАБОТКА, ХРАНЕНИЕ И ПРЕДСТАВЛЕНИЕ ДАННЫХ В ОТКРЫТЫХ ФОРМАТАХ ДОКУМЕНТОВ

Аслами К. З., студентка 4 курса НИУ ИТМО, kamilla.1201@gmail.com

Аннотация

Рассматривается организация обработки и хранения данных в открытых форматах документов. Описывается модель представления полученной текстовой информации в памяти. Анализ проводился путем изучения исходного кода проекта LibreOffice с целью выделить основные модули, участвующие в процессе чтения и обработки файла. В результате исследования была определена стратегия использования рассмотренной реализации для разработки модуля предварительной обработки документов в системе семантического анализа.

Введение

В настоящее время в связи с увеличением массивов неструктурированной текстовой информации возрастает потребность в автоматизированном интеллектуальном анализе. Обработка текста может быть произведена методом извлечения семантической информации – смыслового содержания, функций и связей слов. Автоматизированные системы семантического анализа позволяют обрабатывать большие массивы данных в сравнительно короткие сроки практически без участия человека.

Во время создания подобной системы была поставлена задача организовать предварительное извлечение текстовой информации из документов различных форматов. При реализации модуля извлечения текста необходимо учесть, что текстовые типы данных ,как правило, имеют собственные детали реализации, разметку и способы хранения текста. Особенностью разрабатываемой системы является анализ не только семантики всего текста, но и расположение каждого параграфа в документе, соответственно, модуль предварительной обработки должен не только извлекать текстовую информацию, но также сохранять его структуру. Эта структура в дальнейшем может быть передана анализатору.

Возможность обработки текстовых документов различных форматов

реализована в большинстве современных текстовых процессоров, в задачи которых также входит сохранение исходного форматирования содержимого обрабатываемого файла.

Целью данной работы является исследование одной из подобных реализаций, используемой в текстовом процессоре LibreOffice Writer.

Офисный пакет LibreOffice

LibreOffice - свободно распространяемый пакет офисных программ с открытым исходным кодом, обладающий широкой поддержкой операционных и аппаратных систем. Лицензия GNU Lesser General Public License v3 позволяет использовать программное обеспечение как для персональных, так и для коммерческих целей, копировать и распространять его, а также модифицировать для создания производных программ. Текстовый процессор LibreOffice Writer позволяет обрабатывать наиболее распространенные форматы документов, такие как RTF, TXT, XML, HTML, DOC и DOCX, а по умолчанию используется расширение ODT, являющееся частью международного стандарта Open Document Format.

Обработка документов

Работа с различными текстовыми форматами реализована при помощи так называемых фильтров. По ссылке на документ текстовый процессор определяет расширение файла и выбирает соответствующий этому формату фильтр. Формат определяется одним из следующих способов: если это возможно, то информация о файле считывается путем анализа полного имени и MIME-спецификации[1], в противном случае используются метаданные, хранящиеся внутри документа. Для каждого типа данных есть собственный фильтр, который содержит набор инструкций для корректной обработки файла. Также некоторые форматы (например, DOCX и ODT) представляют из себя архив, соответственно перед чтением файла происходит восстановление сжатых данных.

Хранение и представление данных

Полученный текст, стили и форматирование сохраняются в специальной структуре SwDoc, представляющей модель документа в памяти. Эта модель универсальна и не зависит от исходного формата

[1] MIME (англ. Multipurpose Internet Mail Extensions — многоцелевые расширения интернет-почты) — спецификация механизмов передачи разного рода информации внутри текстовых данных.

данных. Вся текстовая информация управляется контейнером SwNodes, являющимся частью SwDoc. Он представляет из себя массив указателей на так называемые узлы или ноды (Node) - логические объединения текста, например, параграфы или таблицы. В каждом узле может быть любое количество вложенных объектов. Массив SwNodes, вне зависимости от действительного содержимого документа, содержит пять основных разделов. Первый из них всегда пустой и не используется, следующий раздел содержит сноски и примечания. Далее следует секция верхнего и нижнего колонтитулов. Четвертый раздел, как правило, скрыт для пользователя и содержит удаленный текст для осуществления операций повтора или отмены совершенного действия. В последней, пятой секции сохраняется основное содержимое документа. На изображении 1 представлен пример структуры SwNodes документа, включающего в себя верхний колонтитул с двумя параграфами, условно обозначенными как X и Y, а также основное содержимое документа - параграф A, секции S, в которой содержится абзац B, и еще двух параграфов C и D вне секции.

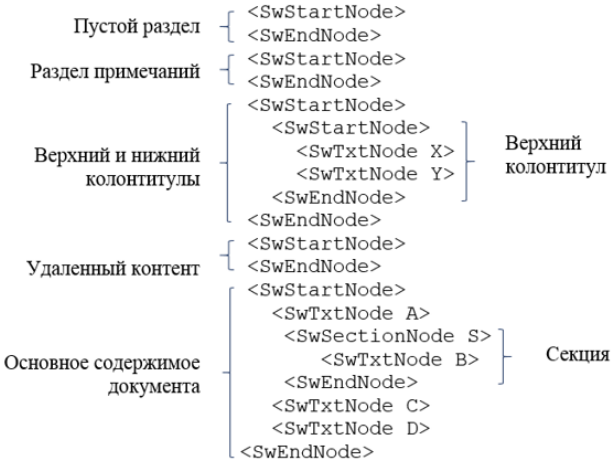

Рисунок 1: Пример модели документа. Структура SwNodes

Внешнее отображение документа определяется так называемыми фреймами (Frame) - набором атрибутов и стилей каждого фрагмента текстового файла. Каждый фрейм привязан к определенной части

документа и, в случае изменения содержимого, немедленно изменяет представление.

Более детальный анализ структуры SwDoc позволил определить обобщенную модель документа, используемую в текстовом процессоре. В таблице 1 представлен каждый компонент и соответствующие ему базовые классы. Основными объектами документа являются текстовые ноды (Node), как правило, в них содержится наибольшая часть данных. Текст таблиц хранится в структуре SwTableNode, управление которой осуществляют классы табличной модели. Текстовое поле - это набор атрибутов и стилей для определенного набора символов, оно не содержит текст, а лишь указывает на него. Менеджер текстовых полей управляет их жизненным циклом – созданием, изменением и удалением, а также связью с узлами. Закладки, сноски, заметки, оглавления и нумерованные списки аналогично таблицам имеют собственную логику работы с текстом. Атрибуты нод – это набор стилей и свойств для форматирования каждого узла в целом. Курсор – это, соответственно, указатель на текущий символ под курсором.

Компонент текстового документа	Класс модели документа
Текстовый параграф (нода)	SwNode
Таблица	SwTable, SwTableLine, SwTableRow
Текстовое поле	SwField
Менеджер текстовых полей	SwFieldType
Закладки	IMark
Текстовые сноски, заметки	SwFmtFtn
Оглавление	SwToxBase, SwForm
Нумерованный список	SwNumberTree
Атрибуты параграфов (нод)	SfxPoolItem
Курсор	SwXTextCursor

Таблица 1. Обобщенная модель документа.

Заключение

В результате проделанной работы была исследована организация чтения и обработки документов текстовым процессом LibreOffice Writer. Анализ исходного кода позволил определить модель представления документа в памяти. Благодаря результатам исследования была определена стратегия использования рассмотренной реализации для разработки модуля предварительной обработки документов в системе семантического анализа. Реализация модуля в виде отдельной бибилиотеки позволит использовать его не только как часть системы, но и в любом другом программном продукте.

Литература

1. OOoAuthors Team Staff. OpenOffice.org 3 Impress Guide/OOoAuthors, Friends of OpenDocument. - Lulu.com, 2010 - 291 с.

2. Writer application code source and explanation [Электронный ресурс]: http://opengrok.libreoffice.org/xref/core/sw/

3. Writer/Core And Layout explanation [Электронный ресурс]: https://wiki.openoffice.org/wiki/Writer/Core_And_Layout

4. LibreOffice Module sw (master) Documentation [Электронный ресурс]: https://docs.libreoffice.org/sw/html/index.html