МЕТОДЫ ОБРАБОТКИ СТРУКТУРИРОВАННЫХ ДОКУМЕНТОВ

Ляхвацкая О.Н.
Национальный технический университет
"Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: *belsquirrel213@mail.ru*

Большинство исследователей под обработкой естественно-языкового текста традиционно понимают обработку текста, которая представляет собой набор предложений без выраженной структуры. В настоящее время становится актуальной задача обработки документов, обладающих высокой степенью формальности и, как следствие, внутренней иерархической структурой [1].

Сложность задачи анализа иерархически структурированных текстов обусловлена следующими их свойствами:

1) как правило, разметка заголовков и маркеров (с помощью стилей, тегов и т.д.) в документе присутствует лишь частично или отсутствует;

2) заголовки с различных уровней иерархии могут не отличаться по виду;

3) название и ссылка в тексте могут иметь одинаковый вид;

4) большое количество конфигураций непрерывных текстовых фрагментов: предложение может состоять из нескольких таких фрагментов, один фрагмент может включать несколько предложений, группа предложений может быть вложена в предложение в виде комментариев [1].

Документы можно различать по стилю форматирования:

1. Структурированные документы, в которых расположение и размеры полей фиксированы.

2. Частично структурированные документы, для которых известен перечень реквизитов, но не регламентировано их расположение и количество.

3. Гибко-структурированные документы или «гибкие документы» – документы, в которых состав и порядок следования их частей по горизонтали и вертикали одинаков, но части могут отличаться по размерам или масштабу.

4. Свободно структурированные документы, у них нет обязательных реквизитов, и форматирование не ограничено [2].

Исходя из видов структурированных документов, рассматриваются такие методы обработки (Data Mining):

1. Алгебраические методы. Исходные данные в них представляются в виде алгебраических структур.

2. Статистические методы. Они используют аппарат теории вероятностей и математической статистики.

3. Методы мягких вычислений. В них используются нечеткое представление данных (нейросети, генетический алгоритм и т.д.) [3].

В данной работе подробно рассматриваются такие виды структурированных документов, как патенты. Патенты – это документ, который свидетельствует о праве изобретателя на его изобретение.

Преимущества системы патентования:

- поощряет изобретателей изобретать;
- обнародование помогает другим исследователям;
- после окончания срока - вседозволенности.

Недостатки:

- опасность монополии.

В работе проанализированы поля данных патентов, рассмотрено из чего состоит патент. Проведено сравнительный анализ патентов Национального Украинского Патентного Бюро с Американским Патентным Бюро и найдено ряд отличий. Для применения и создания программы решено использовать методы Data Mining.

Применение методов DM оправдано при наличии достаточно большого количества данных, которые имеются в структурированных документах. А так же методы DM позволяют доступно и понятно провести парсирование документов, в данном случае патентов.

Список литературы

1. *Лахути Д.Г.* Автоматический анализ естественно-языковых текстов. // М.: ВИНИТИ, 2003.

2. *Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.* Методы и модели анализа данных: OLAP и Data Mining.

3. *Шереметьева С.О.* Теоретические и методологические проблемы инженерной лингвистики. // М.: ВИНИТИ, 1998.