УДК 004.6

СЛАБОСТРУКТУРИРОВАННЫЕ ТЕКСТОВЫЕ ДОКУМЕНТЫ КАК ИСТОЧНИКИ ДАННЫХ

В.А. Губин

ХНУРЭ, г. Харьков, Украина, gubinv@rambler.ru

Исследованы возможности и предпосылки автоматизированной идентификации данных в электронных текстовых документах, размещенных в корпоративных сетях и в сети Internet. Выделен класс слабоструктурированных текстовых документов, представлены их основные признаки и свойства как источник данных. Приведены различные сценарии взаимного расположения в таких документах текстовых фрагментов, соответствующих атрибутам и значениям данных.

СЛАБОСТРУКТУРИРОВАННЫЙ ДОКУМЕНТ, ДАННЫЕ, АТРИБУТ ДАННЫХ, ЗНАЧЕНИЕ ДАННЫХ

Введение

Объем контента, содержащегося в сети Internet и в корпоративных сетях, растет с каждым годом в геометрической прогрессии. Большую его часть составляет мультимедиа: видео-, аудио-файлы, файлы с изображениями различных форматов. Но и объемы электронных текстовых документов в абсолютном измерении также увеличиваются все более возрастающими темпами. В связи с этим все более и более актуальной становится задача эффективного поиска и доступа к информации, содержащейся в этих документах.

Изначально предполагалось и преимущественно предполагается и в настоящий момент, что с документами, опубликованными в сети Internet, в первую очередь будет знакомиться человек. Этой цели подчинено и внутреннее представление документов и то, как они отображаются в окне просмотра того или иного Internet-браузера. Ориентированность на восприятие человеком той информации, которая содержится в подавляющем числе опубликованных в сети документов, мешает доступу к ней и обработке ее тем или иным программным обеспечением. В данной работе предпринимается попытка выделить класс документов, для которых эта проблема может быть преодолена.

1. Цели и задачи исследования

Целью данной работы является исследование возможностей и предпосылок автоматизированной идентификации данных в электронных текстовых документах, размещенных в корпоративных сетях и в сети Internet. Для этого из всего множества документов необходимо выделить класс текстовых документов, содержимое которых предназначено для непосредственного ознакомления человеком, но, в силу особенностей их внутренней структуры, имеющих предпосылки для автоматизированной идентификации содержащихся в них данных. Целью также является исследование основных признаков и свойств такого рода документов как источников данных.

2. Исследование особенностей слабоструктурированных текстовых документов

Главное предназначение электронных текстовых документов – быть источником информации. Причем в зависимости от формата текстовых документов информация представлена в виде данных, ориентированных, в первую очередь, на автоматизированную обработку, либо в виде, изначально предполагающем, что основным агентом, получающим к ним доступ, будет человек.

К первому типу документов можно отнести документы в формате xml, rdf и подобных им. Эти форматы ориентированы на описание данных. Их содержимое аналогично содержимому таблиц базы данных. Поэтому при автоматизированной обработке таких документов не возникает проблем с идентификацией и доступом к данным, содержащимся в них. Причем сами эти документы, как правило, также получены как итог работы того или иного программного обеспечения.

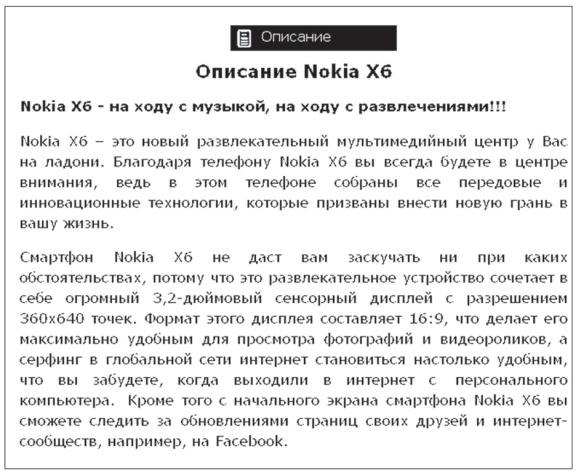
Ко второму типу документов можно отнести документы в форматах html, doc, rtf и подобных им. В этом случае изначально предполагается, что с содержимым таких документов будет знакомиться человек. Разумеется, речь не идет о непосредственном доступе человека к содержимому. Человек знакомится с экранным представлением таких документов в окне просмотра браузера Iternet Explorer, например, либо в редакторе текстов Microsoft Word, либо в любой другой, аналогичной, среде.

На рис. 1 изображен фрагмент электронного текстового документа в формате html, содержащий информацию о мобильном телефоне.

Совершенно очевидно, что представленный выше документ содержит некоторые сведения. Они относятся к тем или иным характеристикам мобильного телефона. Просматривая этот документ, человек может легко извлечь информацию, например, о параметрах дисплея, об объеме встроенной памяти, о разрешении фотокамеры телефона и т.д.

Но если ставится задача автоматизированного извлечения данных из такого рода текстов, то, в общем случае, есть риск столкнуться с сущест-

венными трудностями, связанными с необходимостью понимания смысла естественно-языковых источников.


Описание

Описание Nokia X6

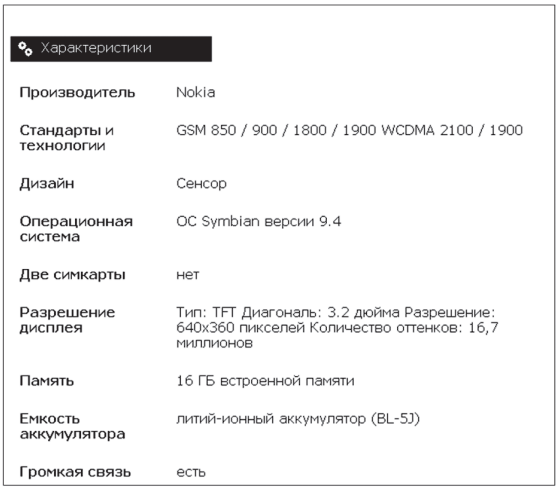
Nokia X6 - на ходу с музыкой, на ходу с развлечениями!!!

Nokia X6 – это новый развлекательный мультимедийный центр у Вас на ладони. Благодаря телефону Nokia X6 вы всегда будете в центре внимания, ведь в этом телефоне собраны все передовые и инновационные технологии, которые призваны внести новую грань в вашу жизнь.

Смартфон Nokia X6 не даст вам заскучать ни при каких обстоятельствах, потому что это развлекательное устройство сочетает в себе огромный 3,2-дюймовый сенсорный дисплей с разрешением 360x640 точек. Формат этого дисплея составляет 16:9, что делает его максимально удобным для просмотра фотографий и видеороликов, а серфинг в глобальной сети интернет становиться настолько удобным, что вы забудете, когда выходили в интернет с персонального компьютера. Кроме того с начального экрана смартфона Nokia X6 вы сможете следить за обновлениями страниц своих друзей и интернет-сообществ, например, на Facebook.

Рис. 1. Фрагмент html-документа с описанием мобильного телефона

Если нажать на этой же странице на кнопку «Характеристики», то получится несколько иное текстовое представление тех же сведений о мобильном телефоне (рис. 2).


Характеристики

Производитель	Nokia
Стандарты и технологии	GSM 850 / 900 / 1800 / 1900 WCDMA 2100 / 1900
Дизайн	Сенсор
Операционная система	ОС Symbian версии 9.4
Две симкарты	нет
Разрешение дисплея	Тип: TFT Диагональ: 3.2 дюйма Разрешение: 640x360 пикселей Количество оттенков: 16,7 миллионов
Память	16 ГБ встроенной памяти
Емкость аккумулятора	литий-ионный аккумулятор (BL-5J)
Громкая связь	есть

Рис. 2. Фрагмент html-документа с характеристиками мобильного телефона

Это представление уже более напоминает то, что имеет место в таблицах базы данных или в xml-документах. Здесь имеется больше предпосылок для автоматизированного извлечения данных. Это объясняется тем, что, в отличие от документа, представленного на рис. 1, данный документ содержит обособленные форматированием текстовые фрагменты. Причем некоторые из них объективно относятся к значениям данных («Nokia», «GSM 850 / 900 / 1800 / 1900 WCDMA 2100 / 1900» «Сенсор»). Другие же обособленные текстовые фрагменты («Производитель», «Стандарты и технологии», «Дизайн», «Операционная система») напоминают имена полей в таблице базы данных и, по сути, являются атрибутами данных, помогающими правильно интерпретировать содержимое тех обособленных текстовых фрагментов, которые выше были отнесены к значениям данных.

Таким образом, представленный на рис. 2 документ объективно содержит данные, представленные парами «атрибут-значение». Но с формальной точки зрения содержимое документа – набор равнозначных фрагментов текста. Следовательно, если рассмотреть фрагмент внутреннего представления соответствующего html-документа (рис. 3), то можно увидеть, что нет формальных признаков, указывающих на то, чем является тот или иной обособленный фрагмент текстового документа: значением данных либо атрибутом данных. Из этого следует, что данные в таких документах не имеют строгого описания. Для того чтобы эти данные получили строгое описание, необходимо идентифицировать каждый такой обособленный фрагмент текстового документа либо как атрибут данного, либо как значение данного.

```
 width="25%"><b>Производитель</b></td><td>Nokia
 </td></tr><tr class="in_tab_row1"><td
 width="25%"><b>Стандарты и
 технологии</b></td><td>GSM 850 / 900 / 1800 /
 1900 WCDMA 2100 / 1900</td></tr><tr
 class="in_tab_row2"><td
 width="25%"><b>Дизайн</b></td><td>Сенсор</td><
 /tr><tr class="in_tab_row1"><td
 width="25%"><b>Операционная
 система</b></td><td>ОС Symbian версии
 9.4</td></tr><tr class="in_tab_row2"><td
 width="25%"><b>Две
 симкарты</b></td><td>нет</td></tr><tr
 class="in_tab_row1"><td
 width="25%"><b>Разрешение
 дисплея</b></td><td>Тип: TFT
Диагональ: 3.2 дюйма
Разрешение: 640x360 пикселей
Количество оттенков: 16,7
 миллионов</td></tr><tr
 class="in_tab_row2"><td
 width="25%"><b>Память</b></td><td>16 ГБ
 встроенной памяти </td></tr><tr
```

Рис. 3. Фрагмент внутренней разметки документа с характеристиками мобильного телефона

Таким образом, из всего множества текстовых документов можно выделить класс текстовых документов, называемых слабоструктурированными. Можно выделить следующие их признаки и свойства как источников данных:

1. Документ содержит внутреннюю разметку.
2. Содержимое документа разбито внутренним форматированием на обособленные текстовые фрагменты.
3. Каждый фрагмент объективно представляет собой либо значение данных, либо атрибут данных.
4. Во внутренней разметке документов нет формальных признаков, указывающих на то, что есть значение данных, а что есть атрибут данных.

Для такого рода документов будет наиболее эффективным их представление в виде совокупности объектов [1] с последующей возможной идентификацией текстовых фрагментов как атрибутов и значений данных.

Примеры типов слабоструктурированных текстовых документов:

1. Анкеты.

2. Страховые формы.
3. Налоговые декларации.
4. Счета.
5. Транспортные накладные.
6. Контракты.
7. Технические параметры изделия.
8. Прайс-листы.
9. Типовые договоры.
10. Карточки учета.
11. Результаты спортивных матчей.

Разумеется, этот перечень далеко не исчерпывает все возможные типы слабоструктурированных текстовых документов. Причем каждый тип может иметь свои подтипы.

Если бы данные в текстовых документах всегда были представлены в таком виде: таблица, состоящая из двух столбцов, в первой ячейке каждой строки таблицы – атрибут данных, а во второй – значение данных, то проблем с автоматизированной классификацией того, что есть что, не было бы. Но сложность автоматизированной классификации обособленных фрагментов текстового документа заключается и в том, что топологически эти фрагменты могут располагаться относительно друг друга по-разному, а не только так, как показано на рис. 2. Пример фрагмента реального документа, подтверждающего данную мысль, представлен на рис. 4.

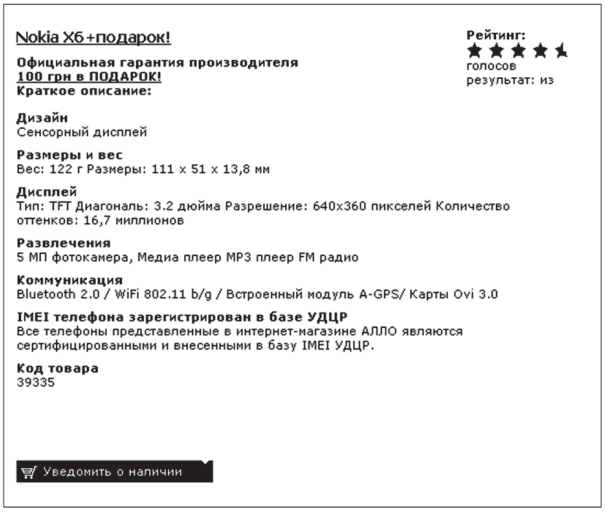


Рис. 4. Альтернативное представление характеристик мобильного телефона

Можно выделить основные варианты взаимного расположения значения и атрибута данных в слабоструктурированных текстовых документах:

1. Атрибут данных располагается в первом столбце таблицы, а значения данных – в ячейках соответствующих строк.

2. Атрибут располагается в первой строке таблицы, а значения данных — в ячейках соответствующего столбца.

3. Атрибут и значение данных располагаются в одном и том же абзаце.

4. Атрибут и значение данных располагаются в смежных абзацах.

5. Атрибут является элементом списка, а значения данных являются элементами вложенного списка.

6. Атрибут присутствует в документе, но не имеет явного соответствующего значения данных.

7. Значение данных присутствует в документе, но не имеет явного соответствующего атрибута.

8. Возможны комбинации вышеприведенных сценариев и, возможно, другие сценарии.

Возможный подход к определению того, чем является тот или иной обособленный фрагмент текстового документа, определяется тем, что в совокупности документов одного и того же типа фрагменты, соответствующие атрибуту данных, будут оставаться без изменений, а фрагменты, соответствующие значению данных, будут изменяться от документа к документу. Таким образом, при попытке извлечения данных из слабоструктурированных текстовых документов необходимо оценить только степень устойчивости к изменению каждого фрагмента текстового документа без необходимости проникновения в семантику текста.

Выводы

Исследования, представленные в данной работе, показывают, что из всего спектра электронных текстовых документов можно выделить подкласс документов, имеющих потенциал для автоматизированной идентификации данных. Приведены примеры и выделены особенности слабоструктурированных текстовых документов как источников данных.

Намечен возможный подход к определению того, чем является тот или иной обособленный фрагмент слабоструктурированного текстового документа: атрибутом данных или значением данных.

Научной новизной работы является следующее: из множества текстовых документов выделен класс слабоструктурированных текстовых документов и представлены их основные признаки и свойства как источник данных.

Список литературы: 1. *Губин, В.А.* Объектное представление электронных текстовых документов [Текст] / В.А. Губин, А.Н. Гвоздинский // Радиоэлектроника и информатика. –2007. – № 1 (36). – С. 61-63.

Поступила в редколлегию 8.09.2010.

УДК 004.6

Слабоструктуровані текстові документи як джерела даних / В.О. Губін // Біоніка інтелекту: наук.-техн. журнал. – 2010. – № 3 (74). – С. 109–111.

У статті виділено клас слабоструктурованих текстових документів. Представлені їх основні ознаки і властивості як джерел даних.

Іл. 4. Бібліогр.: 1 найм.

UDK 004.6

Semi-Structured Text Documents as Data Sources / V.A. Gubin // Bionics of Intelligence: Sci. Mag. – 2010. – № 3 (74). – P. 109–111.

In article the class of semi-structured text documents is isolated. Their main features and properties as data sources are represented.

Fig. 5. Ref.: 1 items.