КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

УДК 004.048; 004.912

М. Н. Мокроусов, кандидат технических наук
Н. Н. Чиркова, магистрант
ИжГТУ имени М. Т. Калашникова

ИЗВЛЕЧЕНИЕ ДАННЫХ ИЗ КОММЕРЧЕСКИХ ВЕБ-ФОРУМОВ*

В статье описываются существующие подходы поиска данных в тексте и предлагается способ извлечения данных с коммерческих веб-форумов на основе регулярных выражений, словарей и анализе соседствующих атрибутов. Приводятся структура и примеры хранения регулярных выражений и правил поиска атрибутов, описание эксперимента в разработанной программной системе поиска и результаты эффективности извлечения данных.

Ключевые слова: автоматическая обработка текста, извлечение данных, регулярные выражения, информационный поиск.

Введение

Во многих областях деятельности людям часто приходится выполнять рутинные действия: расчеты, сопоставление фактов, анализ больших объемов информации. Информатизация общества привела к тому, что люди сталкиваются с подобными задачами и дома. Интернет содержит большое количество сведений по требуемой теме. Существует множество тематических форумов, содержащих большой объем полезной информации по определенным вопросам. Сбор, фильтрация и анализ этой информации вручную очень трудоемок, особенно в тех случаях, когда форумы насчитывают сотни и тысячи сообщений. В настоящее время процесс может быть автоматизирован с применением поисковых систем и каталогов. К сожалению, подобные системы слишком универсальны и не предназначены для решения конкретных задач.

Классификация текстов

Существуют разные классификации текстов. По одной из таких классификаций [1] тексты делятся на структурированные, неструктурированные и частично структурированные.

Полностью структурированные тексты организованы определенным образом. Анализ таких текстов очень легок и не составляет труда, так как место расположения данных и их формат заранее известны. Примером такого текста может служить текст телефонного справочника:

Фамилия: Иванов, телефон: 435052, улица: Ленина, номер дома: 123.

Записи имеют вполне определенный и одинаковый формат. Узнать все 4 параметра записи не составляет труда: фамилия, телефон, улица и номер дома расположены справа от соответствующих обозначений, сами данные также отформатированы одинаково.

Неструктурированным является любой текст, не обладающий заранее определенной структурой, о форматировании текста также ничего неизвестно. Такой текст может содержать большое количество избыточных данных или не содержать части требуемых данных. В неструктурированном тексте существует большое количество неопределенностей, нередки случаи, когда интерпретация и трактовка данных неоднозначны. Данная особенность делает такие тексты самыми трудоемкими для обработки компьютером. Неотъемлемым этапом обработки неструктурированных текстов всегда выступает анализ грамматических взаимозависимостей между отдельными фрагментами текста[1]. Примером неструктурированного текста может служить текст художественного произведения или текст настоящей статьи.

Частично структурированные тексты обладают определенным форматом, однако нельзя сказать, что этот формат заранее известен и одинаков в разных записях. Примером такого текста может являться объявление в газете:

Продам автомоб. Ford Focus 2006 г. пробег около 95 тыс км 1.6 AT, передний привод, седан, лев. руль. Не бит не крашен

В текстах данной группы нередки сокращения и ошибки. Грамматические связи между отдельными фрагментами текста могут отсутствовать, поэтому анализ грамматических зависимостей в данном случае почти бесполезен. Рассматривать частично структурированные тексты как полностью структурированные нельзя. Тем не менее определенная структура в них присутствует, значения помечаются своеобразными лингвистическими тэгами: *пробег* 95 *тыс км, площадь 20 кв.м*. и т. д.

Подходы к извлечению и формализации данных из текстов

В подходах, основанных на правилах, используется набор логических правил, которые система использует для поиска и анализа информации. Правила должны быть составлены заранее исходя из предметной области [2]. Многие коммерческие системы используют данный подход, несмотря на то, что для эффективной работы системы необходимо составить

[1] Здесь и далее под фрагментом текста следует понимать: букву, буквосочетание, слово, словосочетание, простое предложение, различные обороты.

* Работа выполнена при частичной поддержке Госзадания МОиН РФ Проект № 625 и гранта РФФИ № 15-07-08077.

такое количество правил, которое наиболее полно описывает предметную область. Зачастую правила привязаны к определенному домену, и при смене домена требуется обновлять базу правил. Тем не менее этот подход является наиболее точным при наличии хорошей базы знаний и правил и подходит для текстов любого из трех типов.

Еще одним подходом к извлечению данных является применение специальных масок для поиска. Чаще всего для этого используются регулярные выражения [3]. Подход к анализу текстов на основе шаблонов может использоваться при условии, что в анализируемом тексте присутствуют фрагменты, обладающие определенной структурой.

Третий подход основан на методах машинного обучения [4], которые делятся на две группы: обучение с учителем и обучение без учителя. Суть методов первой группы состоит в том, чтобы обучить машинный классификатор на коллекции заранее размеченных текстов, а затем использовать полученную модель для анализа новых документов. Для обучения без учителя заранее размеченная коллекция документов не требуется. С научной точки зрения это наиболее интересная группа методов, но в настоящее время данные методы показывают менее точные результаты [5], чем методы, основанные на правилах и масках.

Предлагаемый подход к извлечению данных из тематических форумов

Процесс извлечения данных из объявлений делится на два этапа:

1) извлечение самих объявлений;

2) анализ текстов объявлений и извлечение атрибутов объявления.

Первый этап осуществляется на основе анализа структуры форума, а также текстовых блоков на отдельной странице темы обсуждения. Анализ может быть проведен вручную предварительно или программной системой в необходимый момент времени. В данной работе данный этап не рассматривается.

Рассмотрим второй этап.

Тексты объявлений чаще всего представляют собой частично структурированные тексты. Для извлечения атрибутов из таких текстов предлагается использовать правила трех типов:

1) правила, основанные на регулярных выражениях;

2) правила, основанные на словарях;

3) правила, основанные на анализе связей соседствующих атрибутов.

Рассмотрим работу правил на примере объявлений рынка жилья.

Каждое правило помечает фрагмент текста лингвистическим тэгом. Например, текст *сдается квартира по Ленина, 126 5000р. услуги посредника: 2000* будет размечен следующим образом:

<сдача>сдается</сдача> <объект>квартира</объект> по <адрес>Ленина, 126</адрес> <стоимость>5000р.</стоимость> <с услугами посредника>услуги посредника</с услугами посредника>: <стоимость услуг посредника>2000</стоимость услуг посредника>

Регулярные выражения применяются для информации, которая является структурированной или частично структурированной и может быть описана с применением шаблона. Примерами здесь выступают: количество комнат, номер телефона, площадь, стоимость, этаж и т. п.

Правила второго типа используют словари ключевых слов, имен собственных, сложных аббревиатур. Этот тип правил применяется для атрибутов, которые представляют собой слова естественного языка, например: *сдается*, *сниму, куплю, услуги посредника* и т. п., а также названия улиц, учреждений, строений и т. д. В случае использования правил на словарях необходимо решить задачу поиска в словаре слов, написанных не в нормальной форме (им. падеж, ед. число, инфинитив). Данная задача может быть решена с использованием применения библиотек морфологического анализа текста.

Правила третьего типа анализируют атрибуты, выделенные с использованием правил первого и второго типа, выделяют дополнительные атрибуты или изменяют существующие. Например, атрибут *стоимость* изменяется на атрибут *стоимость услуг посредника* при наличии слева от него атрибута *с услугами посредника*. Еще одним примером может служить идентификация числа справа от атрибута *улица* как номера дома и объединение номера дома и улицы в атрибут *адрес*.

При этом правила должны быть организованы определенным образом и применяться в заданном порядке. Требуется обеспечить возможность удобного представления и анализа информации о коммерческих объявлениях на рынке товаров и услуг, размещаемых на текстовых тематических форумах. У таких объявлений всегда есть автор, который предоставляет или желает получить товар или услугу. Автор обычно сообщает контакты для связи с ним. Объявления содержат в себе ряд атрибутов, описывающих тот или иной товар. Применяя некоторое правило к тексту объявления, мы можем извлечь этот атрибут. Любая категория товара может подразделяться на ряд подкатегорий, образуя, таким образом, иерархическую структуру.

Для хранения и анализа информации о коммерческих объявлениях предлагается использовать структуру данных, указанную на рис. 1.

Описанная выше структура данных может быть представлена в виде онтологии. В данной работе была использована система PraDict [6]. Созданная в PraDict структура показана на рис. 2, *а*.

Сущности *словарь* и *маска* соответствуют правилам первого и второго типа и содержат в себе справочные сведения. На рис. 2, *б* показан фрагмент словаря *улица,* на рис. 2, *в* показан пример правила-маски для сущности *телефон*.

Описание атрибутов сущностей включает ссылки на эти правила. Пример описания атрибутов сущностей приведен на рис. 2, *г*.

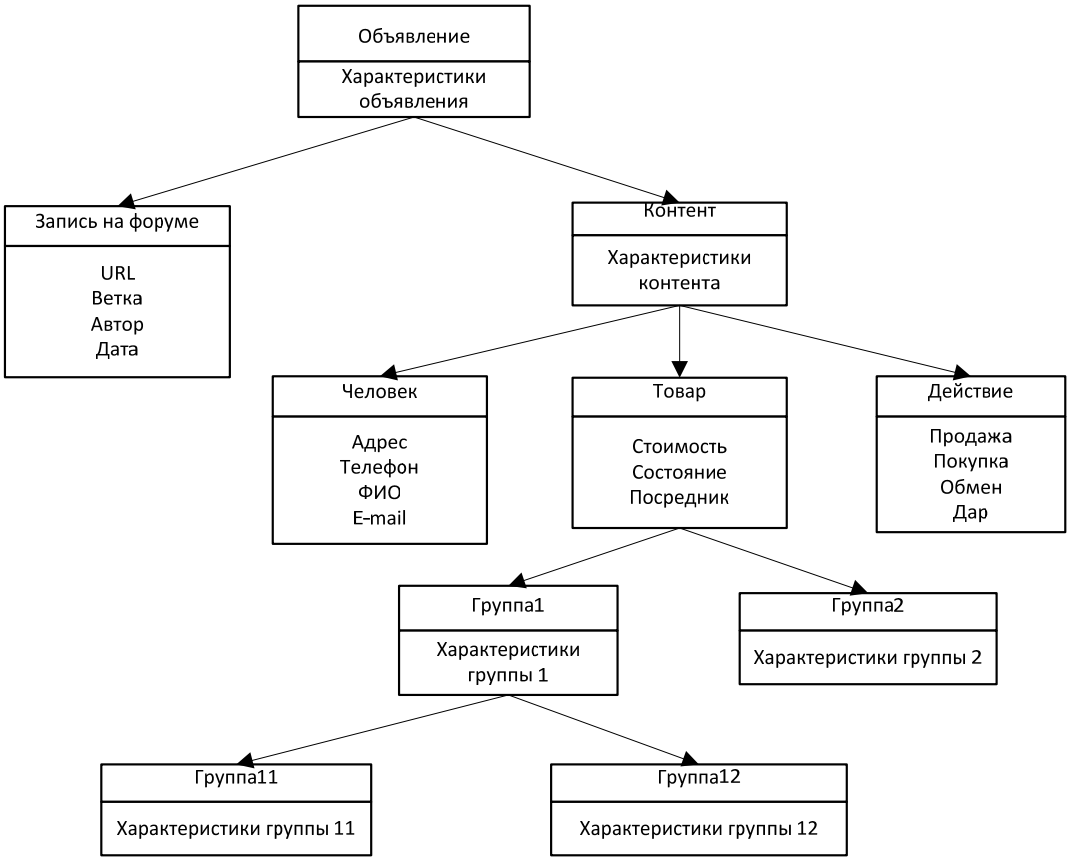
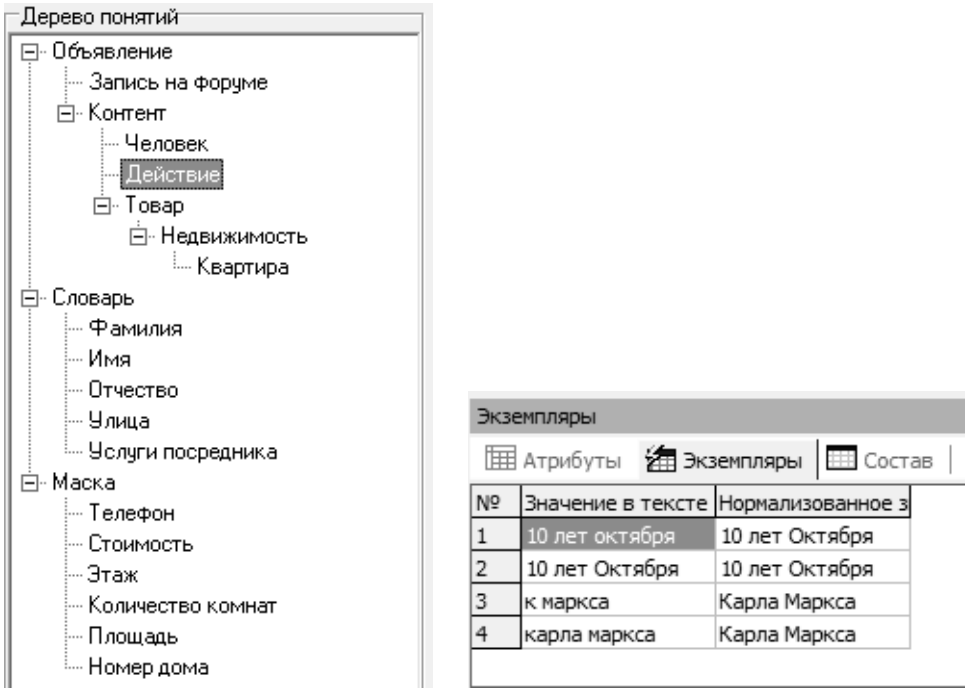
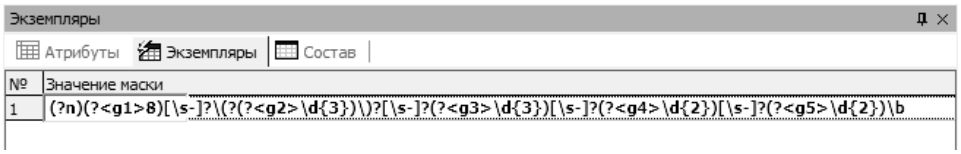


Рис. 1. Структура для хранения информации о коммерческих объявлениях

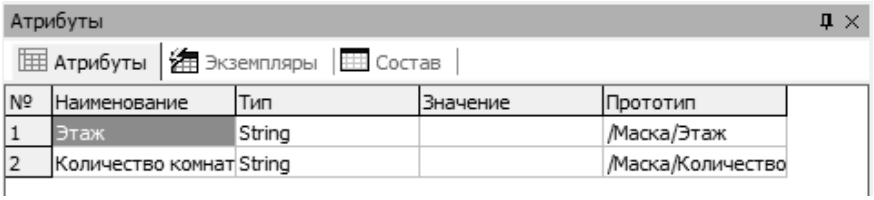


а *б*

Рис. 2. Структуры данных в PraDict: *а* – иерархия сущностей; *б* – словарь сущности *Улица* (см. также с. 72)



в



г

Рис. 2. Окончание: *в* – маска сущности *телефон*; *г* – атрибуты сущности *квартира*

Алгоритм извлечения информации с помощью такой структуры можно представить следующим образом:

1) получить ссылку на объявление;
2) получить текст объявления;
3) определить автора и дату;
4) сформировать по категориям товара верхнего уровня текущий список категорий;
5) применить правила для текущего списка категорий и выполнить извлечение данных;
6) отобрать все категории, в которых были найдены атрибуты;
7) если список отобранных категорий пуст, то выход;
8) получить дочерние категории для списка отобранных категорий;
9) сформировать по дочерним категориям из шага 8 текущий список категорий;
10) перейти к шагу 5.

В процессе анализа веток форума создаются экземпляры сущностей, и онтология пополняется данными, как показано на рис. 3. При этом при формировании таблицы экземпляров наследуются атрибуты вышестоящих сущностей.

Рис. 3. Экземпляр сущности *квартира*

Использование PraDict дает возможность расширять и пополнять словари и правила, а также позволяет вручную исправлять ошибки извлечения данных.

Эксперимент

Для исследования предложенного метода была разработана автоматизированная система поиска и формализации данных из объявлений о сдаче, аренде, покупке и продаже жилья с веб-форумов (рис. 4).

Работа системы протестирована на данных форума *izhevsk.ru* ветки «Жилье». Загрузка объявлений с веб-форума происходит автоматически по расписанию, при этом проверяется наличие новых объявлений, и они загружаются в систему. Если ранее загруженное объявление было удалено из форума или закрыто автором, то в системе оно отмечается как удаленное. Если объявление не было загружено в систему и закрыто автором, то оно игнорируется.

В таблице приведены показатели эффективности извлечения основных атрибутов. В первой колонке перечислены названия оцениваемых атрибутов. Колонка «Всего в ветке» содержит совокупное количество атрибутов для всех объявлений, размещенных в ветке обсуждения. Колонка «Найдено» отражает количество всех атрибутов, извлеченных системой, а колонка «Найдено верно» – количество правильно извлеченных атрибутов. На основе этих данных были вычислены точность (отношение количества верно извлеченных атрибутов к общему числу найденных атрибутов) и полнота (отношение количества верно извлеченных атрибутов к общему числу атрибутов в ветке).

Эффективность извлечения некоторых атрибутов

Атрибут	Всего в ветке	Найдено	Найдено верно	Точность	Полнота
Автор	993	993	993	1	1
Местоположение	2914	2652	2546	0,96	0,8737
Номер дома	1180	1086	1010	0,93	0,8559
Телефон мобильный	723	672	652	0,9702	0,9018
Телефон домашний	273	248	242	0,9758	0,8864
Стоимость	3777	3664	3591	0,9801	0,9508
Стоимость услуг посредника	2186	2055	1993	0,9698	0,9117
Площадь	2081	1935	1838	0,9499	0,8832
Этаж	677	650	618	0,9508	0,9129
Количество комнат	2491	2292	2177	0,9498	0,8739
Итого	**17295**	**16247**	**15660**	**0,9636**	**0,9050**

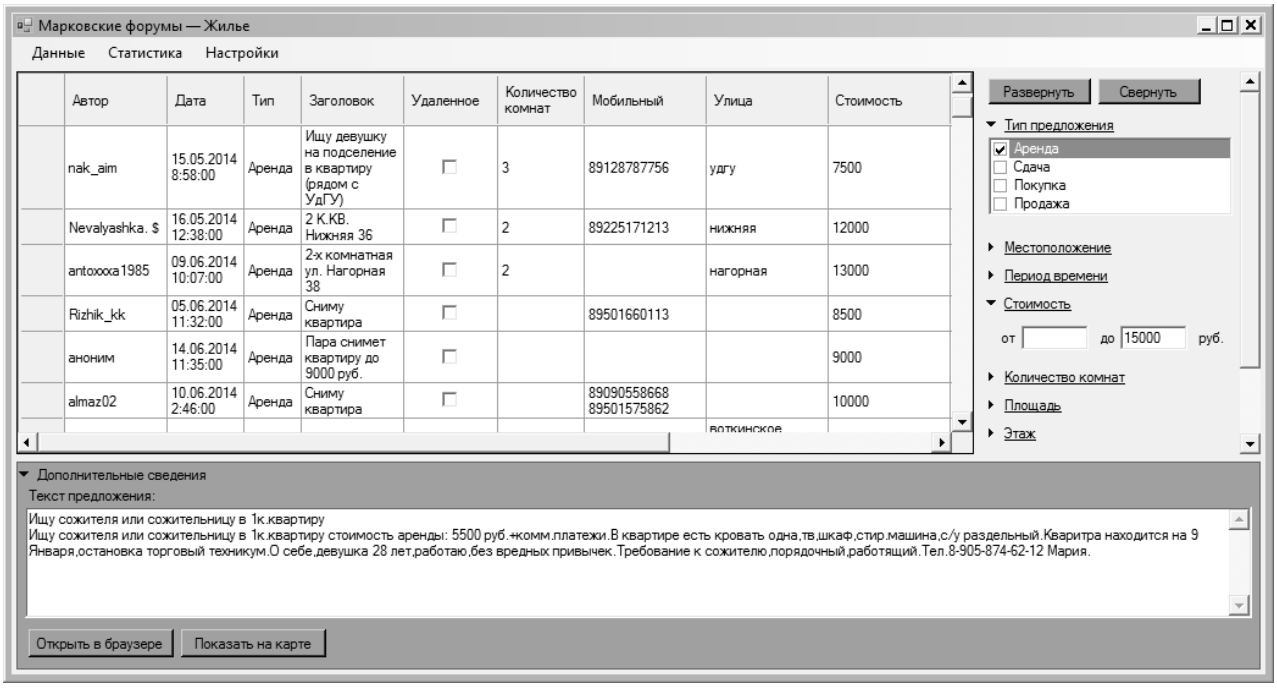


Рис. 4. Интерфейс системы извлечения данных из объявлений рынка жилья

Заключение

Предложенный способ извлечения данных из текстовых веб-форумов показал высокую эффективность: средняя точность поиска атрибутов объявлений ветки «Жилье» составила 96,4 %, а полнота поиска – 90,5 %.

Также в ходе проведения эксперимента было обнаружено, что люди склонны делать опечатки, вследствие чего часть данных теряется. Иногда искажения создаются намеренно. Например, номер телефона 89095670122 может быть записан как 89O9567O122 (вместо цифр используются буквы) или 89ноль95шесть7ноль122. Для таких данных были пересмотрены и изменены регулярные выражения для поиска атрибутов.

Библиографические ссылки

1. *Feldman R., Sanger J.* The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data. – Cambridge University Press, 2007. – 424 p.

2. *Ландо Т.* Извлечение объектов и фактов из текстов [Электронный ресурс] // Хабрахабр [Сайт] (дата публикации: 07.12.2013). – URL: http://habrahabr.ru/company/ yandex/blog/205198 (дата обращения: 10.02.2016).

3. *Jeffrey E. F. Friedl.* Mastering regular expressions. Understand Your Data and Be More Productive. 3rd Edition. – O'Reilly Media, 2006. – 544 p.

4. *Кормалев Д. А.* Приложения методов машинного обучения в задачах анализа текста // Программные системы: теория и приложения : труды Международной конференции, Переславль-Залесский. – М. : Физматлит, 2004. – Т. 2. – С. 35–48.

5. *Matthieu C., Padraig C., Delany S. J.* Supervised Learning / Machine Learning Techniques for Multimedia Case Studies on Organization and Retrieval Editors: Matthieu Cord, Padraig Cunningham. – Springer-Verlag Berlin Heidelberg 2008. – P. 21–50.

6. *Мокроусов М. Н., Кучуганов В. Н.* Прагматическая компонента текста и человеко-машинный словарь. Труды Конгресса по интеллектуальным системам и информационным технологиям «IS&IT'15». – В 3 т. – Таганрог : Изд-во ЮФУ, 2015. – Т. 1. С. 222–227.

* * *

Mokrousov M. N., PhD in Engineering, Associate Professor, Kalashnikov ISTU
Chirkova N. N., Master's degree student, Kalashnikov ISTU

Data extraction from commercial web forums

This article reviews the existing approaches in the area of data retrieval from texts and provides a method for extracting data from commercial web forums based on regular expressions, dictionaries and analysis of adjacent attributes. The article describes the data structure used for storing and organizing regular expressions and information extraction rules and gives the examples of such rules. The experiment is conducted to determine the accuracy of analysis, for which a special information system is used implementing the method described in the article.

Keywords: natural language processing, data extraction, regular expressions, information retrieval.

Получено: 18.02.16