OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.8 + 004.9

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОГО СЕРВИСА АНАЛИЗА ПРЕДЛОЖЕНИЙ НА РЫНКЕ НЕДВИЖИМОСТИ

Анисимова Т.В.[*], Нестеров Р.А.[**], Печенежский А.Б.[*]

[*] *Пермский государственный национальный исследовательский университет, г. Пермь, Пермский край*

hvostya@gmail.com
nextzucker@gmail.com

[**] *Национальный исследовательский университет «Высшая школа экономики», г. Пермь, Пермский край*

ranesterovhse@gmail.com

В статье описывается реализация интеллектуального сервиса агрегации предложений на рынке недвижимости. При анализе объявлений используется подход на основе онтологий. Набор онтологий, описывающий структуру конкретных сайтов, может быть расширен, что делает возможным использование сервиса для широкого набора источников.
Ключевые слова: интеллектуальный сервис, недвижимость, онтология.

Введение

Специфика работы специалиста по работе с недвижимостью заключается в постоянном анализе информационных потоков, поэтому для успешной деятельности им необходимы средства интеллектуального анализа и мониторинга предложений на рынке. Большая часть такой информации является слабоструктурированной и при традиционном подходе работа с ней занимает значительную часть времени. Источниками информации для риэлтора являются тематические Интернет-ресурсы, бумажные издания и специализированные БД.

Задача агрегации информации из разных источников и ее структуризация является чрезвычайно актуальной. Кроме того, требуют решения задачи устранения дублирования информации и поиска противоречий. Слабая структурированность информации и гетерогенный характер еѐ источников диктуют применение средств и методов искусственного интеллекта для решения данной задачи (например, text mining, технологий Semantic Web и мультиагентные технологии).

Предлагаемым решением описанной задачи является разработка интеллектуального сервиса, собирающего информацию о предложениях на рынке недвижимости из различных источников в единую базу данных.

1. Классификация агрегаторов предложений на рынке недвижимости

Интернет-ресурсы и сервисы, аккумулирующие существующие предложения на рынке недвижимости, принято называть агрегаторами. Основными характеристиками данных сервисов, влияющими на востребованность пользователями, являются полнота базы объектов, актуальность данных, достоверности информации, возможности поиска и фильтрации и цена доступа [Агрегаторы, 2013].

Существующие на данный момент ресурсы можно классифицировать по двум признакам: терррриториальному охвату базы объектов и способу организации работы с контентом. В первой классификации выделяют два класса: глобальные, созданные на платформе известного портала («Яндекс.Недвижимость»), и локальные, относящиеся к региональным проектам по недвижимости. Вторая классификация предполагает деление на описанные ниже классы.

Доска объявлений появилась одной из первых. Обычно это бесплатная, тематически организованная база данных. На профессиональном языке это так называемая «грязная» база, т.е.

неорганизованные, практически не регламентированные системы.

Следующим важным агрегатором информации являются *электронные версии печатных изданий частных объявлений*. Одним из главных преимуществ, по оценкам экспертов, позволивших этим ресурсам занять ведущее место в своих рынках, является совмещение концепции газеты бесплатных объявлений с электронной версией.

Мультилистинговые системы являются наиболее популярным и востребованным видом ресурсов среди профессиональных риэлторов. В России на сегодня нет глобального портала, который объединял бы информацию обо всех предложениях на рынке недвижимости.

Информационные порталы по недвижимости или *специализированные сайты* на сегодняшний день – наиболее распространенные агрегаторы информации о недвижимости в сети Интернет.

Социальные сети также относят к агрегаторам информации. Тем не менее, сейчас наблюдается сближение сайтов-агрегаторов и соцсетей с точки зрения общих черт и применяемых сервисов.

Мета-агрегаторы – класс систем, которые объединяют предложения с нескольких ресурсов. Данные сервисы имеют дополнительные функции, например, интеллектуальная фильтрация объявлений только от собственников, а не от посредников.

2. Архитектура сервиса

Для решения автоматического пополнения базы данных объектов недвижимости в рамках проекта по созданию системы автоматизации агентства недвижимости реализован *интеллектуальный сервис*. Работа сервиса заключается в извлечении информации об объектах недвижимости из неструктурированных объявлений, получаемых из различных источников. Решение базируется на онтологическом подходе. Общая архитектура разработанного сервиса представлена на рис. 1.

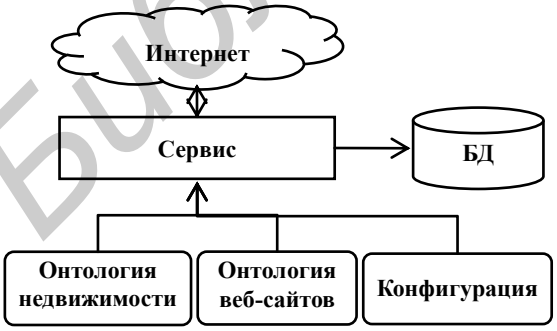

Рисунок 1 – Архитектура сервиса

3. Схема работы сервиса

Общая схема работы сервиса представлена на рис. 2.

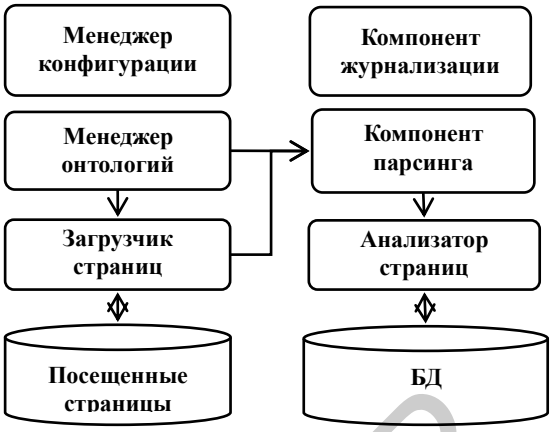

Рисунок 2 – Схема работы сервиса

Компонент журнализации выполняет запись информации о работе сервиса. Полученная с помощью данного компонента информация используется для мониторинга работы сервиса и его отладки.

Менеджер конфигурации предоставляет доступ к переданным сервису настройкам и производит динамическую настройку работы сервиса при необходимости.

Менеджер онтологий реализует операции по управлению онтологическими ресурсами.

Компонент загрузки страниц создает локальную копию исходной страницы, а также производит еѐ предварительную обработку. Информация о пройденных страницах заносится в специальную базу данных, что позволяет оптимизировать работу сервиса за счѐт исключения повторного прохода по одинаковым ссылкам в текущую сессию поиска. На базе онтологии сайтов о недвижимости данным компонентом осуществляется извлечение информативной части страницы. Таким образом, на вход *компонента парсинга страниц* фактически передаѐтся предобработанный текст объявления о продаже недвижимости, из которого за счѐт использования онтологии объектов недвижимости извлекаются знания, которые приводятся к некоторому стандартному виду (например, происходит приведение к одинаковым единицам измерения площади и др.).

Компонент анализа страниц осуществляет логический вывод по онтологии объектов недвижимости, основываясь на полученных знаниях, а также проверяет некоторые дополнительные эвристики, после чего формирует сведения об объекте недвижимости, который заносится в соответствующую базу данных.

4. Онтология сайтов о недвижимости

Онтология сайтов о недвижимости хранит специфичные для конкретных сайтов настройки.

Среди интересующих нас параметров выделяют:

• Описание позиции на странице, где вероятнее всего находится интересующая

информация, а также описание, позволяющее получить заголовок этой информации.

- Описание позиции на странице, где могут находиться полезные ссылки.
- Описание фильтров, позволяющих определить «мусорные» для сервиса ссылки.
- Настройки механизма «перелистывания» страниц (описан ниже).

5. Онтология объектов недвижимости и регулярные выражения

Онтология объектов недвижимости содержит некоторые общие понятия предметной области и связи между ними.

В процессе парсинга страниц выполняется попытка «привязать» конкретные понятия, основываясь на знаниях, имеющихся в онтологии. К каждому конкретному понятию в онтологии приписаны определѐнные регулярные выражения. Выделяются регулярные выражения двух типов: *общие* и *настроенные под конкретный сайт*. Регулярные выражения второго типа могут использоваться для привязки только на конкретных сайтах и в общем случае неверны (такие регулярные выражения позволяют хорошо разбирать используемые на сайте специфические формулировки). Регулярные выражения общего типа построены таким образом, чтобы срабатывать в общих случаях. При привязке конкретных понятий сначала выполняется попытка привязки по второму типу и, в случае неудачи, – по первому.

В структуре регулярных выражений можно выделить два типа элементов: те, что говорят о нахождении совпадения, и те, что свидетельствуют об ошибочной привязке понятия. Например, привязывается понятие «телефон» (то есть к объекту недвижимости проведена телефонная линия), однако в объявлениях часто пишут в конце телефон того, кто подал объявления. Элементы второго типа как раз и выявляют признаки, позволяющие определить, что речь идѐт вовсе не об интересующем понятии.

Кроме того, в процессе извлечения знаний об объектах устанавливаются определѐнные их показатели («Площадь квартиры», «Этаж» и т.п.). Общая структура регулярных выражений в целом аналогична описанной выше, однако дополнительно выделяются логические части, позволяющие, например, выполнить перевод показателей в некоторую единую систему (если была указана цена в тысячах рублей за сотку, то сервис приведѐт эту характеристику к унифицированной единице – рубли за квадратный метр и т.п.).

6. Механизм «перелистывания» страниц

В ходе изучения структуры сайтов, содержащих объявления об объектах недвижимости, было выяснено, что зачастую они содержат списки, где представлены только ссылки на сами объявления. Поскольку объявлений на сайте достаточно много, они могут быть расположены на нескольких страницах, переход по которым обычно осуществляется кнопками навигации.

Разработанный механизм «перелистывания» страниц осуществляет последовательный переход по страницам со списками. Необходимые для этого настройки индивидуальны для каждого сайта и хранятся в онтологии сайтов о недвижимости.

Особо стоит отметить, что некоторую сложность представляет переход по ссылкам, где информация подгружается с помощью javascript. На исследуемых сайтах данная проблема была решена посредством использования особых классов.

7. Настройки сервиса и список загрузок

Настройки сервиса содержат параметры, отвечающие за работу сервиса. Среди параметров можно выделить:

- Путь к списку загрузок, где указаны адреса, которые будет сканировать сервис.
- Путь в файловой системе, куда будут сохранены выгруженные страницы.
- Период, через который сервис возобновит свою работу (остановка сервиса может быть связана с тем, что он пройдѐт по всем указанным в списке закачки адресам).
- «Глубина сканирования сайтов» – длина пути, на которую переходит сервис по ссылкам.
- Флаг, указывающий, разрешать ли сервису переход на сторонние сайты по ссылкам при поиске «в глубину».

8. Используемые программные и инструментальные средства

Сервис был разработан в среде Microsoft Visual Studio 2010 на языке программирования C#. Онтология разрабатывалась в редакторе онтологий Protégé. Использовались также библиотеки HtmlAgilityPack (для осуществления парсинга html-страниц) и OwlDotNetApi (для чтения онтологии из файла).

9. Тесты и оценки

Сервис демонстрирует достаточно высокие показатели точности. Адекватно распознаются примерно 97% объявлений. В 93% случаев точно распознается атрибутный состав объявления. Точность распознавания может быть повышена за счѐт настройки онтологии на формат представления объявления на конкретном сайте. Кроме того, в компонент логирования встроены средства анализа, объясняющие причину неудачного сопоставления и предоставляющие рекомендации по настройке онтологии.

Заключение

На данный момент реализован действующий прототип сервиса интеллектуального анализа и агрегации предложений на рынке недвижимости, архитектура которого и особенности реализации описаны выше. Одна из ключевых особенностей сервиса заключается в возможности его настройки на анализ новых ресурсов без изменения программного кода – настройка осуществляется через редактирование онтологии. В рамках проекта на основе информации, хранящейся в системе, планируется также реализовать экспертную систему по подбору объектов недвижимости и их оценке.

Библиографический список

[Segaran, 2009] Segaran T., Evans C., Taylor J. Programming the Semantic Web, O'Reilly Media, 2009.

[Яндекс, 2013] Что такое Яндекс.Недвижимость http://help.yandex.ru/realty/.

[Агрегаторы, 2013] Недвижимость online: агрегаторы http://media-office.ru/?go=2082914&pass=f79e9c77f077cf1d060a615834c3c2d1

INTELLIGENT SERVICE FOR ANALYSIS OF REAL ESTATE MARKET OFFERS

Anisimova T.V.[*], Nesterov R.A.[**], Pechenezhskiy A.B.[*]

[*]*Perm State National Research University, Perm, Russia*

hvostya@gmail.com
nextzucker@gmail.com

[**]*National Research University «Higher School of Economics», Perm, Russia*

ranesterovhse@gmail.com

This article contains the implementation description of a real estate market offers aggregator service. Advertisement analysis is made with the aid of ontologies. A set of ontologies to describe specific websites can be extended, so the aggregator can be used for many diverse resources.

Introduction

Real estate agents constantly analyze different information flows, so real estate market offers intellectual analysis and monitoring services are required for their efficient work. Most of this information is semistructured and its conventional processing is time-consuming. Real estate information resources are topical Internet resources, papers and special databases.

Information aggregation and structuring tasks are increasingly timely. Apart from that, it is necessary to address information duplication and contradiction search tasks. Semistructured information and its heterogeneous resources implies application of artificial intelligence means: text mining, Semantic Web technologies and multi-agent technologies.

Our solution is to develop intelligent service to accumulate real estate market offers information from different resources in a single database.

Main Part

First of all, we have categorized Internet resources aggregating information on real estate market offers: bulletin boards, electronic versions of free advertisements newspapers, electronic versions of free advertisements newspapers, multilisting systems, real estate information portals, social networks, meta-aggregators.

As for the service, to address automatic real estate items database in the context of a real estate agency automation system project a smart service was implemented. It extracts real estate items information from unstructured advertisements placed on different resources. The solution is based upon an ontological approach using a general real estate ontology, ontology of websites and settings for service configuring.

The service was developed using Microsoft Visual Studio 2010 and C# programming language. Ontology was developed in Protégé ontology editor. Also, HtmlAgilityPack (for html-pages parsing) and OwlDotNetApi (for reading ontologies from a file) libraries were used.

The service demonstrates rather high accuracy rates. Approximately 97 per cent of all advertisements are recognized in an adequate way. In 93 per cent of the time advertisement attributes are recognized precisely. Recognition accuracy can be improved with the aid of adjusting ontology to the specific resource advertisement representation. Besides, logging component include analytical tools to find a reason for a fail correlation and to recommend on required ontology settings.

Conclusion

In this paper we described the architecture and peculiar implementation properties of the real estate market offers aggregator service. At the moment the pilot system of the service is implemented. One of the core features of this service is that it can be adjusted to new resources analysis without changing program code; configuration is only about ontology editing. Also, in the context of this project and on the basis of the information kept in the system, we intend to develop an expert system on real estate items selection and estimating.