Two recommender systems: Technical decisions and lesson learned

Evgeny Cherkashin, Viktoria Kopylova, Boris Shevchenko, Nikita Lukyanov

Matrosov Institute for System Dynamics and Control Theory of SB RAS, Irkutsk, Russia, eugeneai@icc.ru
National research Irkutsk state technical university, Irkutsk, Russia, kopylovika@mail.ru

AIIT-2020, October, 16

Introduction

Recommender systems (RS) are examples of *decision support systems.* They are useful to support users with additional information and decision variants. We consider two application domains:

1. real estate of a region, helping with selling realty;
2. choosing the specialty of a high school, analyzing entrant's social network profile.

Both systems were developed within master degrees at Institute for Information Technologies and Data Adalysys, National research Irkutsk state technical university.

The **first** system deals with selling flats, houses, garages, rooms, but the **interesting** aspect is to supply realty offices with support in selling spacial cases like hotels, countryside houses, shops: sellers paying taxes for unprofitable property, buyers cannot invest.

In the **second** RS entrants are to obtain perspective specialty, comfort to their nature. There are no industry supporting career guidance, only informational services and neighbors' subjective experience. The universities' activities should be focused to a target audience with supplying students **relevant optional courses**.

Ralated Works Summary

The *collaborative filtration* is currently more popular than *content filtering* as it allows one to solve "**cold start**" problem: initial lack of information and interest transformations due to modern trends.

- Real estate RS sometimes attract sellers to be the experts for content filtering;
- Spacial data of realty surroundings are estimated;
- Problem stated in game and control theory domains (satisfy the most customers);
- Reflect entrants GCE grades to a specialty, predicting univ. grades;
- Forecast learning trajectories of students by similarity to graduate ones;
- Advice students with a course, analyzing text descriptions of courses;

We aim at application of known techniques to the source data of Irkutsk region.

Preliminary data processing: Real Estate RS

Real estate (RE) **objects** are presented at *classified advertisement websites* and special RE ones `altcom.ru` storing *offers*. Offers refer RE objects described by properties (address, sizes, levels, number of rooms, etc.). Input format is XML, being essentially a list.

RE **agents** have roles: *unknown, seller, buyer, owner, realtor, expert, invalid user*. Experts and realtors are RS managers. The first three are regular users able to look for an object.

Objects are distributed by hierarchical cluster analysis to classes: two-room flats, one-or-two-room flats (comfort class realty), one-room flats, dachas, houses, rooms, commercial space, a garage, and elite realty (flats with three and more rooms). Naming is performed by `experts`.

User interest is acquired by tracking his/her activity on web site. If user spent more than 30 seconds viewing an object data add a positive interest unit. After some time user is suggested to register to provide cross-browser data acquisition.

Classification of Realty Objects

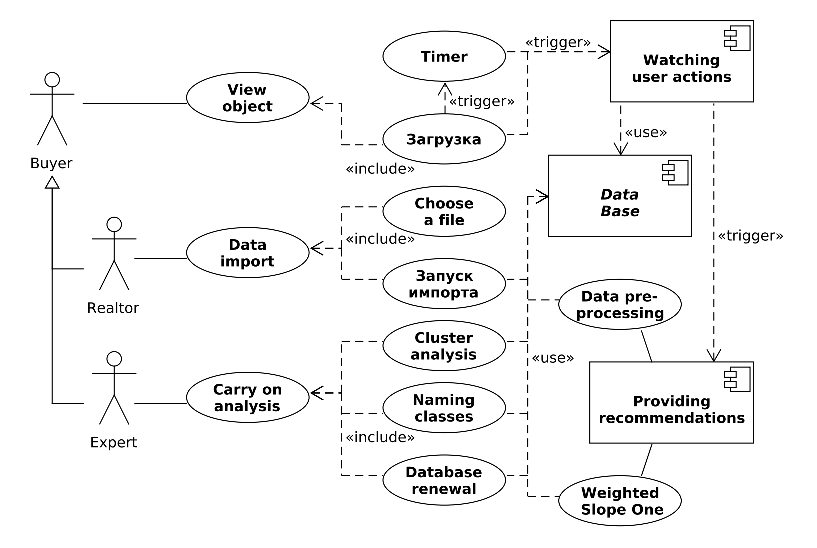
Attribute	Calculation technique	$v_k, \%$	Formula
`string Name`	not used		
`ILocation Location`	equality	10	(1)
`string Address`	–”–	10	
`float Price`	relative difference	25	(2)
`float Area`	relative difference	35	(2)
`string ImageURL`	–”–		
`string URL`	–”–		
`int Rooms`	relative difference	100	(2)
`int RoomsOffered`	–”–	100	(2)
`int Floor`	–”–	30	(2)
`int FloorTotal`	–”–	10	(2)
`BuildingEnum …`	equality	30	(1)
`IBuilding …`	–”–	30	(1)
`PropertyEnum …`	–”–	30	(1)
`CategoryEnum …`	–”–	100	(1)
`string GUID`	–”–		

$$d_k(i,j) = \\ = \begin{cases} 0, & \text{if } a_i = a_j, \\ 1, & \text{if } a_i \neq a_j, \end{cases} \quad (1)$$

$$= \frac{|a_i^k - a_j^k|}{a_i^k + a_j^k}, \\ a_i^k + a_j^k > 0. \quad (2)$$

Reduction: $$d(i,j) = \frac{\sum_{k=1}^{m} |v_k \cdot d_k(i,j)|}{\sum_{k=1}^{m} v_k}, \qquad 0 \leqslant d_{i,j} \leqslant 1.$$

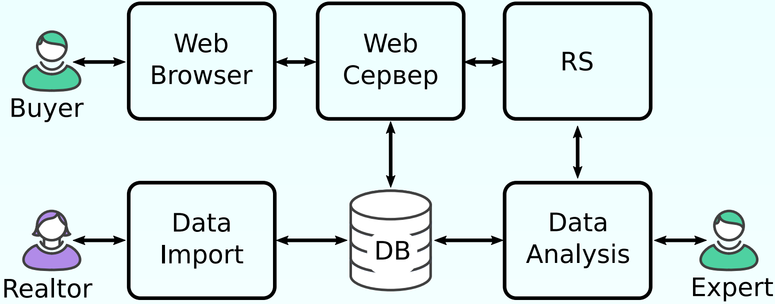
Recommendations generation behavior



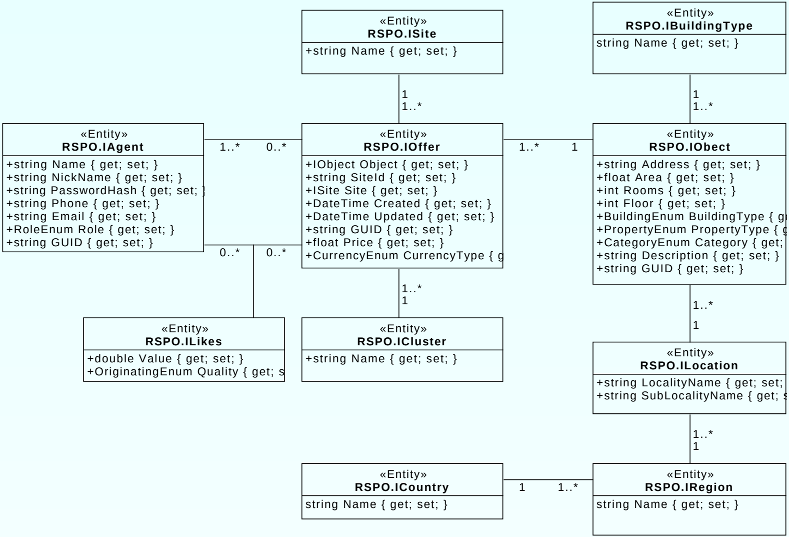
Web Application

As the software platform of the RS is C# we used ASP.NET based frameworks, such as `Nancy`, mapping an URL to a lambda function of one argument.

HTTP template engine is based on SharpTAL, allowing implementation of MVC with a dictionary variable substitution. It can be extended to support LOD cross-site integration.



Persistent Object Structure



«Entity»
RSPO.ISite
+string Name { get; set; }

«Entity»
RSPO.IBuildingType
string Name { get; set; }

«Entity»
RSPO.IAgent
+string Name { get; set; }
+string NickName { get; set; }
+string PasswordHash { get; set; }
+string Phone { get; set; }
+string Email { get; set; }
+RoleEnum Role { get; set; }
+string GUID { get; set; }

«Entity»
RSPO.IOffer
+IObject Object { get; set; }
+string SiteId { get; set; }
+ISite Site { get; set; }
+DateTime Created { get; set; }
+DateTime Updated { get; set; }
+string GUID { get; set; }
+float Price { get; set; }
+CurrencyEnum CurrencyType { g

«Entity»
RSPO.IObect
+string Address { get; set; }
+float Area { get; set; }
+int Rooms { get; set; }
+int Floor { get; set; }
+BuildingEnum BuildingType { g
+PropertyEnum PropertyType { g
+CategoryEnum Category { get;
+string Description { get; set; }
+string GUID { get; set; }

«Entity»
RSPO.ILikes
+double Value { get; set; }
+OriginatingEnum Quality { get; s

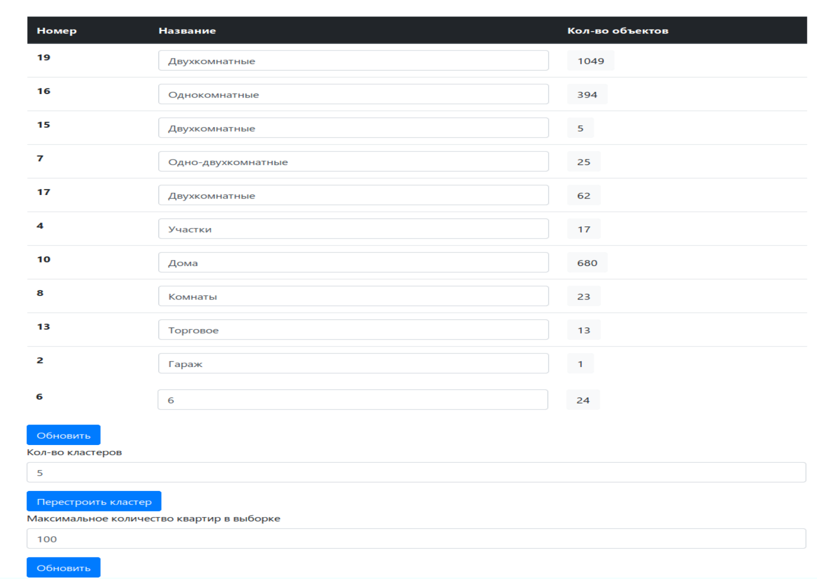
«Entity»
RSPO.ICluster
+string Name { get; set; }

«Entity»
RSPO.ILocation
+string LocalityName { get; set;
+string SubLocalityName { get; s

«Entity»
RSPO.ICountry
string Name { get; set; }

«Entity»
RSPO.IRegion
string Name { get; set; }

Screenshot: Naming clusters

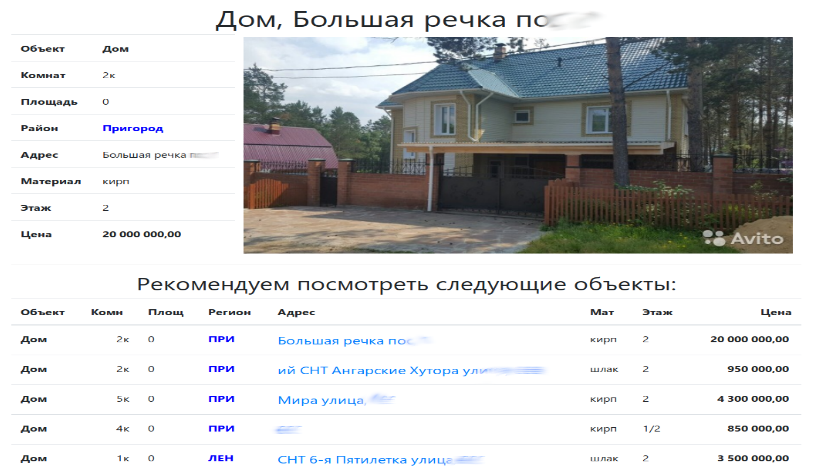


Номер	Название	Кол-во объектов
19	Двухкомнатные	1049
16	Однокомнатные	394
15	Двухкомнатные	5
7	Одно-двухкомнатные	25
17	Двухкомнатные	62
4	Участки	17
10	Дома	680
8	Комнаты	23
13	Торговое	13
2	Гараж	1
6	6	24

Кол-во кластеров: 5

Максимальное количество квартир в выборке: 100

Screenshot: Example of recommendation



Дом, Большая речка по[redacted]

Объект	**Дом**
Комнат	2к
Площадь	0
Район	**Пригород**
Адрес	Большая речка п[redacted]
Материал	кирп
Этаж	2
Цена	**20 000 000,00**

Рекомендуем посмотреть следующие объекты:

Объект	Комн	Площ	Регион	Адрес	Мат	Этаж	Цена
Дом	2к	0	**ПРИ**	Большая речка по[redacted]	кирп	2	**20 000 000,00**
Дом	2к	0	**ПРИ**	ий СНТ Ангарские Хутора ули[redacted]	шлак	2	**950 000,00**
Дом	5к	0	**ПРИ**	Мира улица, [redacted]	кирп	2	**4 300 000,00**
Дом	4к	0	**ПРИ**	[redacted]	кирп	1/2	**850 000,00**
Дом	1к	0	**ЛЕН**	СНТ 6-я Пятилетка улица[redacted]	шлак	2	**3 500 000,00**

Data sources and preliminary processing for entrants' RS

Social networks are popular ways of self-expression for school students (entrants). User's properties are the profile data and set of group and channel subscriptions represented with **tags, keywords**. TSU[1] research proved that InContact (ВКонтакте) students profiles are closely related to educational interests.

To produce recalculations we make clusters of users and find ones, who has chosen a specialty. Specialty data is extracted from *enrollment orders*, which structure is regular. Specialties described by competence relations are also organized in clusters, named by experts.

Recommendations are to be produced with machine learning. *e.g.*, a neural network (NN), pretrained to relate user class to a specialty [class]. Another NN will be used to distribute new entrants to classes.

So the RS is based on *content filtering*, and users will participate in RS functioning passively.

[1] Tomsk state university

The summary of used technologies

- C# platform as compromise of expressiveness and performance, cross-platform (.NET and MONO);
- Entity Framework (ET) for object persistency;
- SQLEXPRESS and BrightstarDB as ET backends with aim to RDF and LOD;
- Microsoft.Office.Interop.Word for DOCX processing;
- HtmlAgilityPack for HTML parsing;
- dotNetRdf, System.Xml.Xdocument;
- VDS.Common for word indexing;
- SharpTAL for HTML templating;
- Newtonsoft.Json for JSON generation/parsing;
- BCrypt for user password encryption;
- xunit.* for organizing functional tests;
- Paket .NET dependency manager;
- Browser JavaScript with Bootstrap and JQuery.

Discussion

The realty RS was tested by a number of buyers, none of them mentioned misbehavior of the system like spontaneous transitions from one class to another, or supplying empty sets of recommendations.

The presented experience shows that C# .NET and MONO technologies and used techniques are well suitable for construction RS:

1. working in “cold start” conditions;
2. do not require user to register and authorize while implicit gathering information needed for recommendation generation;
3. testing systems are to be done in a local conditions to be able to check the results comparing with existing “natural” data;
4. using open-source technologies, modules and libraries.

The obtained real estate RS implementation as well as the second project do not take advantage of the nowadays methods of R&D development, which corresponds to current traditions. Most attention is paid to preliminary data processing and obtaining minimal valuable product (MVP).

Conclusion

We presented results of two master degree projects. The first one is finished, and the second one is on a half way. Both systems have similar design, but rely on different RS techniques. Both systems use persistent object-oriented representation of application entities.

We have to solve “cold start” problem for both domains by creating taxonomies to consider classes as objects of users’ interests. Taxonomies of users are also created in Entrants’ RS.

The techniques could be developed further by adaptation the standard directions such as:

- usage of ontologies describing domain to extend search capabilities especially in students’ courses domain;
- case based reasoning for the same domain;
- predictive modeling of attribute values for realty;
- providing geospacial data of infrastructural service organizations in the environment, *e.g.* schools, kindergartens and shops.

Thank You for attention!

https://github.com/eugeneai/papers-aiit-rs/raw/master/talk-2020-10-16-RS.pdf