УДК 519.688

INTEGRATION METHODS OF HETEROGENEOUS RETAIL DATA SOURCES FOR RECOMMENDER SYSTEM IMPROVEMENT

Authors: Kutuzova Tatiana, Melnik Mikhail (ITMO University, Saint Petersburg)
Supervisor: Melnik Mikhail (ITMO University, Saint Petersburg)

A large amount of data is stored in databases of various areas, such as retail, banking, medicine, etc. However, not all information is useful to users. That is why it is important to extract proficient information from large amount of data.

The Big Data Exchange (BDE) is a platform for data mining methods that provides three main options for their clients: data enrichment, data onboarding, data research.

Actually, retail recommender system (RS) in the context of market basket analysis (MBA) provides an ability to get information about customers' behaviour. Recommender systems are widely used in different areas, such as finance, retail, and biology. There are three main techniques for constructing RS: content-based, collaborative filtering and hybrid.

The integration of heterogeneous data sources process includes unification of heterogeneous datasets, clustering and filtering of data for recommender system construction. We transform data into a unified form by three word embedding methods for RS improvement despite of various approaches for RS construction. This is due to vector data representation is more convenient for data handling than semantic form of data.

Originally, data is a set of products that is presented in a semantic form as names of products. A prior data contain of complete information about customers. Original data is an internal data of BDE platform user, and BDE provides external data to the RS improvement. In this section, methods for data unification, integration and RS construction are presented.

After data cleaning and preprocessing we calculate the distance between the vector representation of products names from internal and external data sources in order to determine the most appropriate pairs of products names from different sources. Three word embedding methods for transforming semantic data into a vector form in the context of data unification are compared in this paper:

- word occurrence(WO) uses information about the frequency of words;
- latent semantic indexing(LSI) transforms all products name into a vector form;
- word2vec (W2V) trains a neural network for a reconstruction semantic context of words.

The relation between the quality of unification and RS is presented in table 1.

Table 1. Comparing word embedding methods for data unification

	W2V	LSI	WO
Accuracy	64	25	3
Mean recall	0.85	0.82	0.64
Mean precision	0.61	0.59	0.31
Mean F1	0.71	0.69	0.42
Mean RC	0.54	0.49	0.32

To define the quality of unification there are marked test data by an expert. That marked dataset contains names with different length and complexity. The accuracy of matching by each method is presented in table 1. Clearly, unification by using W2V algorithm provide more effective results with accuracy 64% than other methods, especially WO method which accuracy is only 3%. Moreover, the quality of unification also affects the quality of RS. Mean values of metrics W2V and LSI are closed, although W2V allow achieving better quality RS. Low-quality unification method leads to low-quality RS.

Own metric for evaluation of RS was defined due to another metrics are focused only on the intersection of prior and constructed recommendations and do not consider a union of them – recommendation conformity(RC):

$$RC(R) = \frac{\sum_{t \in D} \frac{|r(t) \cap p(t)|}{|r(t) \cup p(t)|}}{|D|}$$

where R – current recommendation, D – a database of customers' transactions, $r(t_{input}) = t_{output}$ – recommendation function, $p(t)$ – prior recommendation function.

The experiment provides a comparison of original and adapted original data. Figure 1 shows the ability of effective integration of heterogeneous data sources. Integration allows achieving a better quality of RS about by 58% in RC metric with using integrated (adapted) data.

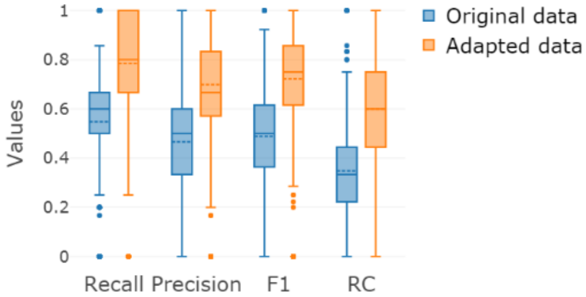

Figure 1. RS comparison for original data and external data

In this study, the ability of external retail data integration in the context of BDE was investigated. Results of experiments on unification methods comparison provide a directly proportional dependence between the quality of unification and RS quality. In addition, using of word2vec for word embedding provides effective unification of heterogeneous datasets. Nevertheless, it is necessary to explore some productive methods for word embedding to allow higher RS quality.