Studying Items Similarity for Dependable Buying on Electronic Marketplaces

Olga Cherednichenko, Maryna Vovk, Olga Kanishcheva, Mykhailo Godlevskyi

National Technical University "Kharkiv Polytechnic Institute",
2, Kyrpychova str., 61002 Kharkiv, Ukraine
olha.cherednichenko@gmail.com, marihavovk@gmail.com,
kanichshevaolga@gmail.com, mikhail.godlevskij@gmail.com

Abstract. The processing of product buying is a very difficult task when we have thousands of items in each market category. In order to study items similarity for dependable buying we try to analyze item descriptions on AliExpress, eBay marketplaces and test k-means algorithm for item grouping/product segmentation. The usage of the classical clusterization algorithms for grouping similar products according to their descriptions is studied. A corpus of different products (bikes and smartphones) from e-shop AliExpress, eBay is developed. Each entity in this corpus contains photos and a product description. Each entity in this corpus contains product description with different fields. These short texts are used for experiments. As a result, it is found out that the k-means algorithm works well only for uniformly distributed data by categories, but this is not suitable for the segmentation of heterogeneous descriptions. The task of item descriptions systematization is set in the research below.

Keywords: E-commerce; Dependable Buying; Recommendation systems; Product search; Clusterization; k-means; TFIDF

1 Introduction

The e-commerce business is rapidly expanding nowadays. The quantity of sellers and customers has grown drastically and sales channels significantly shifted to the Internet. Competition between sellers and trading platforms is tough. Sellers tries to represent their goods in a such way to be presented in as more search outputs as possible, so very often they manipulate by product description information. For buyers it causes the problem of searching needed commodities.

A huge number of items are described, and subsequently bought and sold every day in e-commerce marketplaces. The overwhelming amount of information and the number of available items makes a big problem for a person who wants to buy something online. The buyer very often feels need which should be satisfied but doesn't know exactly by which namely commodity it can be performed in the best way. People spend a lot of time in order to get enough information about difference and peculiarities in similar goods. After that buyer specifies request and tries to find the same

commodity in different sellers and at different marketplaces searching the most appropriate price and conditions.

One more significant problem here is commodity name and description, which are given by a seller. Buyers face the problem of incomplete, incorrect and even contradictory information in a product description. Some sellers intentionally add odd information and popular words pursuing a goal to be represented in as more as possible searching results.

The problem of information retrieval often poses as the task of reduction search space, when objects with similar characteristics should be distinguished and separated. Such approach is used in e-commerce trading platforms. Increasing number of e-commerce marketplaces, on one hand, is a problem for a consumer to find products to purchase, on the other hand, big amount of sellers cause tight competitiveness. Matching of users from both sides is central to the business models of all two-sided markets.

Thus, it is important for both sides (buyers and sellers) to research the problem and develop a learning algorithm or customization code for a browser which is able to distinguish similar products at different trading platforms based on name and product description.

The main purpose of the paper is to research the similarity of commodities based on items description on different trading platforms.

In order to achieve this purpose the following tasks should be performed:

- to collect descriptions of commodities from different trading platforms;
- to estimate the differences in descriptions of the same commodity presented in various trading platforms;
- to create a data set for experiments;
- to perform an experiment on clusterization;
- to analyze received results concerning the possibility of similarity estimation based on commodity description.

2 Related Works

Recommendation systems are able to cope with the problem of overload information. They recommend products or service offerings based on individual preferences. Information filters search user's characteristics and preferences in order to form recommendations or to predict user's future behavior. But sometimes that is recommendation systems that cause the problem for the buyer. Each recommender system typically embeds some specific algorithm to compute the similarity of two relevant objects. However, there is no universally best way of doing this. Some consumers make efforts to get rid of system's recommendations.

Commercial importance is one of the reasons that recommendation systems are widely investigated [1-3]. They are primarily based on two main kinds of information filtering techniques: content-based filtering [4] and collaborative filtering [5, 6]. Be-

sides such classification also distinguish matrix factorization algorithms [7], regression, classification [8], and learning to rank [9] algorithms on which recommendation systems can be based. The collaborative filtering uses the nearest neighbor technology that calculates the distance between users by using the historic preference information of users. The collaborative filtering recommendation systems are able to make recommendation across types and have good self-adaptability. But there are some disadvantages also. They are the poor quality of recommendations when a system starts, a problem of new user appearance, data sparsity, scalability, and others. Content-based filtering recommendation systems are based on a comparison between the content of the items and a user profile. Content-based filtering systems also have drawbacks such as sparsity, new user problem, the need for adequate data structure categorizer [4].

Matrix factorization techniques for recommender systems are based on the potential connections between users and items implied in 'User–Item Matrix' [7]. However, the existing algorithm for recommendation is being failed in the case of the matrix sparseness.

The paper [8] has proposed the weighted linear regression recommendation system (WLRRS) based on weighted linear regression models. Compared with traditional methods, the WLRRS has the best predictive accuracy and the best classification accuracy with less fluctuation.

Traditional recommendation methods, including collaborative filtering, are currently the most mature and widely used methods. However, traditional recommenders do not consider that people may share similar interests, but might have different feelings or opinions about them.

In addition, the problem of a large amount of information is also associated with the increasing rate of change of this information. It means that data in different sources is updated more and more quickly, and accordingly, new data should be collected with increasing frequency and delivered to consumers in time.

Often, the special tools to solve all these problems with searching, collecting and processing are required, it causes additional costs. Nevertheless, the use of more information allows finding hidden patterns and provides opportunities for automating business processes.

The diversity of data sources indicates the complexity of the processes of collecting and systematizing information. The collection of data from the web space involves searching for web pages, extracting relevant information and processing this information [10, 11].

Thus, existing methods for information retrieval and recommendations do not give the information how the products are grouped. Besides that recommended products are offering from the same trading platform. For buyer it is important to have grouped similar products from different sellers and trading platforms in order to enhance buying dependability.

3 Our approach for Studying Dependable Buying

Also the studying of item description can be a real challenge for a buyer. Let's look at the simple example. The list of bikes was chosen from aliexpress.com (https://www.aliexpress.com/category/1204/bicycle.html?spm=2114.search0103.110.10.37f137b2XhZTu7) and consists of over 2300 results. Using trade platform tool "Group Similar Products" the item list was decreased (Fig. 1).

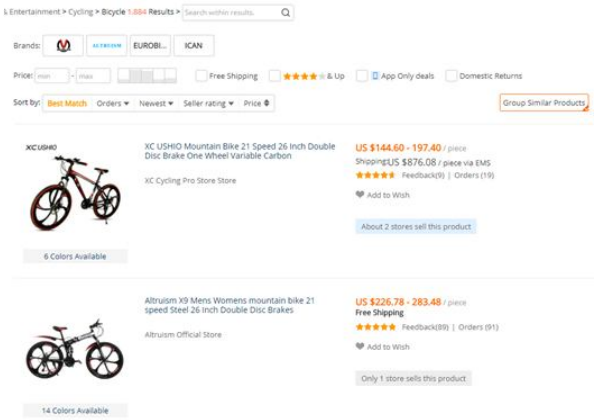

Fig. 1. The list of products

For example, the item from the list has name as "Altruism X9 Mens Womens mountain bike 21 speed Steel Gear shift 26 Inch Double Disc Brakes Bicycles Road Cycling Riding" (Fig. 2). Looked through its group one item was found. But we can find another group.

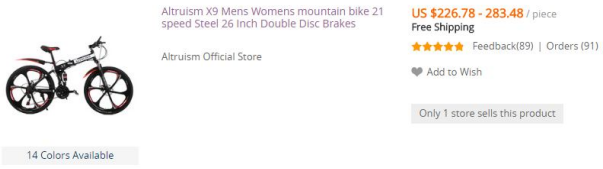

Fig. 2. The description of bike

The second item was named "Altruism X9 Road Bike Mountain Bicicleta 26 inch Steel 21 Speed Bicycles Disc Brakes". This group consists of two items (Fig. 3 and Fig. 4).

Fig. 3. The description of the similar bike from another group

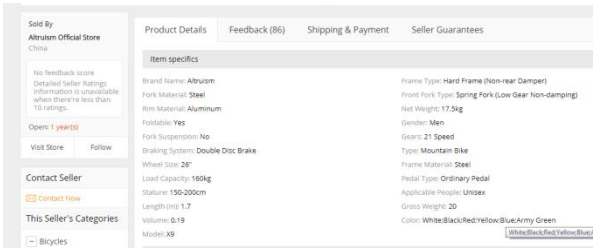

Fig. 4. The description of the similar bike from another group

So, a buyer needs to review all item pages. The question which arises is "why similar items are proposed in different groups"? This sample highlights importance of studying items grouping.

The problems of processing of large amounts of information lead to the task of systematization of information (Fig. 5). The systematization of information means a kind of classification of all texts on different groups. The purpose of the systematization of documents is the ability to provide relevant information to users in a short time.

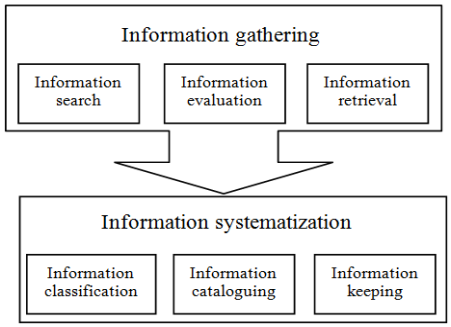

Fig. 5. The scheme of information systematization

As a rule, the result of the information search is an unordered set of text parts that can contain duplicate, inaccurate information in an unstructured form. Therefore, the next stages of the data collection are the information evaluation and extraction. Irrele-

vant data should be removed. Another issue is to prove data dependability to decision making.

Huge amount of information found on the Internet often lead to the fact that the number of objects that satisfy the user's request is very large. This really complicates the process of reviewing the results and selecting the most suitable data from the data set found. However, in most cases, huge amount of information can be available for perception if the sources are grouped into thematic sets. The user can review classes obtained by grouping. Then, the user can immediately skip classes of irrelevant documents. This process of grouping data is carried out by clustering or classifying a set of electronic documents.

The application of cluster analysis in its general form reduces to the following stages:

1. Selecting a sample of objects for clustering.
2. Defining a set of variables by which the objects in the sample will be evaluated.
3. Calculating the values of the similarity measure between objects.
4. Creating the object clusters.
5. Presenting the results of the analysis.

Thus, if the groups of items are formed, buyer can compare products by their features (brands, sizes, etc). It is presupposed the similar items have similar features. So, the comparison among one set of similar items gives the opportunity to distinguish unfair sellers and commodities with incorrect descriptions. So, the commodity choice in such way protects buyer from marketing tricks. In order to prove the grouping quality firstly we need to study item similarity based on item descriptions extracted from market places.

4 Experiments

4.1 Our Datasets

We have two datasets for our experiments. The first corpus contains items of bikes; it was created from eBay.com (https://www.ebay.com/) website and AliExpress (https://ru.aliexpress.com/) website. The second corpus contains items of smartphones and we used products only from eBay.com website.

Our first corpus contains 1,986 entities from Bike product category where each entity has a product description, the example of such description is presented on Fig. 6 (AliExpress) and Fig. 7 (eBay). Some statistics of the first corpus are shown in Table 1.

```
ProductTitle": "EUROBIKE S 21 Speed Aluminum Mountain bike Dual Disc Brake Mountain bicycle",
"Price": "US $496.00 / piece",
"BrandName": "EUROBIKE",
"FrontForkType": "Spring Fork (Low Gear Non-damping)",
"Length(m)": " ",
"ForkSuspension": " ",
"Gears": "21 Speed",
"Type": "Mountain Bike",
"FrameMaterial": "Aluminum Alloy",
"PedalType": "Bead Pedal",
"ApplicablePeople": "Unisex",
"NetWeight": "14kg",
"FrameType": "Hard Frame (Non-rear Damper)",
"LoadCapacity": "90kg",
"Gender": " ",
"Stature": "160-185cm",
"BrakingSystem": "Double Disc Brake",
"Volume": " ",
"WheelSize": "26\"",
"RimMaterial": "Aluminum Alloy",
"ForkMaterial": "Steel"
```


Fig. 6. Bike product description from AliExpess


```
"ProductTitle": "20\" Mongoose Brawler Pro Style Boys' BMX Bike Bicycle Sport Cycle Freestyle Gift",
"Condition": "New",
"Price": "US $130.00",
"Brand": "Mongoose\n",
"WheelSize": "20\"",
"FrameMaterial": " ",
"NumberofGears": "8",
"Gender": "Boys",
"FrameSize": "20\"",
"Type": "BMX Bike"
```


Fig. 7. Bike product description form eBay.com

Table 1. Statistic about data set of bikes .

Category, Brand (Bike) eBay/AliExpress	Number of entities
Centurion/Eurobike	25/21
Mogoose/Kalosse	14/463
Specialized/Sava	19/35
Trek/Sequel	10/38
Other	464/897
Total	*532/1,454*

Our second corpus contains 350 entities from Phone product category. You can see the example of product description on Fig. 8. Some statistics are shown in Table 2.

Name - NEW HTC DESIRE 816 4G LTE ANDROID UNLOCKED QUAD-CORE 5.5" 13MP CAM SMARTPHONE
URL - https://www.ebay.com/itm/New-HTC-Desire-816-4G-LTE-Android-Unlocked-Quad-Core-5-5-13MP-Cam-Smartphone/322840869006?hash=item4b2ad0b88e:m:mWv7mezPoNJyxq1SLP
Condition - New: A brand-newunusedunopenedundamaged item in its original packaging (where packaging is applicable). Packaging should be the same as what is found in a retail storeunless the item is handmade or was packaged by the manufacturer in non-retail packagingsuch as an unprinted box or plastic bag. See the sellers listing for full details. See all condition definitions- opens in a new window or tab ... Read moreabout the condition
Processor - Quad-core 1.6GHz) Cortex A7 Processor
Screen Size - 5.5"
Memory Card Type - MicroSD
Lock Status - Unlocked
Brand - HTC
Model - Desire 816
Style - Touch Screen
Connectivity - 2G
Operating System - Android 4.4.2
Features - Wi-Fi 802.11 a/b/g/n/acdual-bandWi-Fi Direct,
Storage Capacity - 8GB
Camera Resolution - 13.0MP
RAM - 1.5GB

Fig. 8. Phone product description form eBay.com

Table 2. Statistic about data set of phones.

Category, Brand (Bike) eBay/AliExpress	Number of entities
Galaxy S5	50
HTC Desire 816	50
iPhone 7	50
Nokia 1100	50
Sumsung Galaxy S7	50
Sony Xperia Z2	50
Sony Z5 Premium	50
Total	*350*

The product description for both categories Bikes and Phones have different number of fields. They tend to be different in length, words etc. because these sentences are written by sellers themselves.

In Table 2 we have created a balanced corpus, where that is shown in Table 2 for each category we have the same number of entities. In Table 1 our corpus is not balanced, as we have very different verity number of entities for each category, yet this corpus is very realistic for internet marketplaces. The largest category *Other,* it has 464/897 entities with various brands and types of bicycles.

4.2 Experiments and Results

We use k-means algorithm for finding similar items for both data collections. Likewise, we used TFIDF, Porter's stemmer and list of stopwords for our clusterization. The dataset collection process is formalized and normalized by both means manually and automatically. Initially, we used all the descriptions of entities for all collections. The results are presented below:

Top terms per cluster for bikes from AliExpess:

Cluster 0: carbon mountain bike alloy aluminum oil steel 27 disc 00

Cluster 1: steel altruism road aluminum bike non 21 speed 98 gear

Cluster 2: carbon road fibr 700c bike handlebar materi non rearderailleur seatheight

Cluster 3: alloy kaloss aluminum mountain 26 oil bike 30 27 90

Cluster 4: kid children trainingwheel childbik productsno toybar 2color weight steel 12

Top terms per cluster from eBay.com:

Cluster 0: road carbon race 700c braketyp pull calip use bike 00

Cluster 1: bmx 20 bike napproxim numberofgear boy new brand condit type

Cluster 2: kid grossweight brakingsystem trainingwheel yes bike 12 wheel girl new

Cluster 3: mountain 24 centurion braketyp 27 bike size mtb numberofgear unisex

Cluster 4: bmx mini 99 18 16 steel napproxim gbp concept children

Top terms per clusters for smartphones:

Cluster 0: s7 samsung galaxy 32gb sm edge g930v smartphone 4gb contract

Cluster 1: htc desire 816 8gb 13mp used sim dual android mobile

Cluster 2: sony xperia z2 d6503 z3 case cover retail compatible 16gb

Cluster 3: apple iphone 128gb memory used 32gb smartphone ios ohne black

Cluster 4: samsung s5 galaxy 16gb 16mp 4g retail g900v lte smartphone

Cluster 5: nokia 1100 phone mobile black germany network refurbished gsm used

Cluster 6: sony z5 premium xperia e6853 32gb 23mp smartphone 3gb black

As we can see from our results in clusters we have not seen bike name (brand), as we have for the smartphone category. The exception is only *Kalosse* brand from AliExpress.

As the next step, we exclude *Other* category from all the data sets and repeat our experiment with clusterization on bike brands.

#samples: 58, #features: 190

Top terms per cluster from eBay.com:

Cluster 0: ***centurion*** cm frame size mtb 27 blue carbon eve 43

Cluster 1: bmx ***mongoose*** bike l80 legion new lilgoose rare ransom 16

Cluster 2: ***specialized*** 20 wheels bike inch hotrock speed street beautiful kids

Cluster 3: ***trek*** bike bicycle series fuel carbon 2016 condition roadbike domane

Similar results were received as for data from AliExpress.

In this case, we can conclude, that for the balanced sample with definite brand names standard clusterization methods (like k-means) gives an acceptable result. This conclusion was proved while experimenting with smartphones dataset. Precision and Recall are 0.95 and 0.95 respectively. So if we have good data with definite categories the clusterization algorithm will show perfect results.

5 Discussions

The results of the experiments showed that the standard clusterization algorithms only work well for data uniformly distributed by categories, but this is not suitable for the segmentation task of Internet markets or various trading platforms. Trading platforms such as AliExpress, eBay etc. are characterized by a large number of items from different sellers with different descriptions.

For our experiment there is a large category of goods (*Other* category), which contains a large number of product of different brands and which is very difficult to divide into separate categories. Thus, such goods, on the one hand, can be attractive to the buyer and should be grouped, but at the same time they represent some kind of noise for well-known brands and do not allow high-quality clustering of goods by brands.

It would be advisable to use as a gold standard and a 10-fold cross-validation for datasets. Comparison with other clustering methods (e.g. knn, word2vec, etc.) could also be useful. We didn't continue experiments with other clusterization algorithms as far as we received results with a high value of Precision and Recall on unbalanced dataset.

6 Conclusions and Future Works

As results of our experiments we have found out that large market places like AliExpress, eBay and others contain huge amount of similar commodities which are complicated for clusterization. Poor work of recommendation systems at such trading platforms can be explained by using standard algorithms such as k-means for heterogeneous sets.

Different trading platforms, different kinds of goods and differently structured sets have been studied. As a result we can conclude that branded and non-branded items should be processed with the help of different approaches. For items with identified brands k-means algorithm gives good results and can be recommended for the task of systematization while searching for goods to purchase. A huge number of items which are classified by market places as Brand name – Other leads to the necessity to study and develop an appropriate clusterization algorithm.

In future work it is supposed to create an approach to commodity grouping which will combine product description and their pictures. It can allow extending the number of descriptors in order to build similar item groups. It will be reasonable to process that branded and non-branded commodities separately. It causes necessity of studying and developing new ways for information gathering about goods in order to systematize and choose the best items.

7 References

1. A.Razia Sulthana, S.Ramasamy "Ontology and context based recommendation system using Neuro-Fuzzy Classification" Computers & Electrical Engineering February 2018., in press.
2. J. SonSeoung, B. Kim "Content-based filtering for recommendation systems using multiattribute networks". Expert Systems with Applications Vol. 89, pp. 404-412, December 2017.
3. M. Ali Salahli, T. Gasimzade, F. Alasgarova , A. Guliyev "The use of predictive models in intelligent recommendation systems". Procedia Computer Science. Vol. 102, pp. 515 – 519, 2016.
4. L. Ya, The Comparison of Personalization Recommendation for E-Commerce. 2012. International Conference on Solid State Devices and Materials Science, Physics Procedia 25, pp. 475-478, 2012.
5. X. Ma, H. Lu, Z. Gan, J. Zeng "An explicit trust and distrust clustering based collaborative filtering recommendation approach". Electronic Commerce Research and Applications. Vol. 25, pp. 29-39, September–October 2017
6. N. Polatidis, C. K.Georgiadis "A multi-level collaborative filtering method that improves recommendations". Expert Systems with Applications Vol. 48, pp. 100-110, April 2016.
7. D. Feltoni Gurini, F. Gasparetti, A. Micarelli, G. Sansonetti "Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization" Future Generation Computer Systems. Vol. 78, Part 1, pp. 430-439, January 2018.
8. L. Chenglong, W. Zhaoguo, C. Shoufeng, H. Longtao "WLRRS: A new recommendation system based on weighted linear regression models" Computers & Electrical Engineering February 2018., in press.

9. A Goswami, F Hedayati, P Mohapatra, “Recommendation Systems for Markets with Two Sided Preferences”. ICMLA, 2014.
10. Davies, J. Semantic Web Technologies: Trends and Research in Ontology-based Systems. Wiley, 2006.
11. Liu, B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd Edition. Springer, 2011.