Towards Improving the Search Quality on the Trading Platforms

Olga Cherednichenko, Maryna Vovk, Olga Kanishcheva, Mikhail 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. In this paper, the problem of the search quality on the trading platforms, such AliExpress, eBay and others is explored, the major types of problems that arise in product search by customers are considered. The usage of the classical clusterization algorithms for grouping similar products according to their descriptions is studied. A data set for experimenting consists of different items (smartphones) from e-shop eBay is developed. Each entity in this corpus photos and a product description are given. These texts are used for item comparing in order to perform similar groups or similar items. The results show that the k-means algorithm is good for preliminary grouping but for detailed processing, other methods and approaches are required.

Keywords: Trading Platform, Recommendation system, Product search, Information Technology, Product Classification

1 Introduction

Electronic-commerce (e-commerce) has become an important channel for business performing. The share of e-commerce is increasing continuously (Table 1). According to [1] the Internet retail becomes a popular option for consumers.

Table 1. The share of Internet commerce in the retail trade in the world [1]

Year	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*
Share of Internet commerce, %	4	4.4	5	5.7	6.5	7.2	7.9	8.6	9.3

Over the last decade the purchase process has been changed drastically. The first e-commerce shops were similar to real shops. Customers used to choose type and model of commodity from the finite set of goods. The interface and installation-specific settings were adjusted for that purpose. The appearance of huge trading platforms, like AliExpress, eBay, Amazon etc., has changed the retail process. As a rule, trading platforms combine a huge number of sellers and goods. There are many options and

adfa, p. 1, 2011.

product alternatives from different suppliers on different trading platforms. The set of available products is typically huge, it changes constantly, and new items are added. In such circumstances, a customer should choose where, what and from whom to buy.

Thus shoppers should look through plenty of pages in order to find an appropriate product. Finding the most advantageous offer for online shoppers has to provide different seller offers, compare product descriptions and images. As a result, the search space enhances dramatically. A lot of products cannot be sold when customers are not able to find them. It's obvious, that sellers are willing to promote their goods. They adjust to the recommended algorithms, which are used on certain trading platforms. In order to be present at as many search results as possible, sellers intentionally change the name of the commodity, photos and item's characteristics of the item. Thus the problem of search quality on the trading platforms is crucial.

To improve the process of product search, we look into methods from machine learning. In order to simplify the shopper's search, it is necessary to form similar products in groups. It would be useful to have an algorithm that can compare items and define the referenced commodity which fits best to the item group in order to make the search process easier, faster and more precise. For example, such algorithm is implemented in trading platform eBay [2]. Choosing option "similar product" the groups of similar goods can be generated. The main drawback is low accuracy of those groups.

A search engine is a type of an information retrieval system which helps find the information stored in a computer system. All of the modern trading platforms provide the search engine in order to help shoppers. Recommendation systems also try to improve buying process by predicting what items are interesting and useful to the buyer based on specific information about the user profile or product information. However, nowadays recommendation systems and search engine on modern trading platforms are not able to solve all problems related to the quality search, such as incorrect product description, fake photos, and non-relevant search output.

The aim of this paper is the analysis of main problems for product searching on different trading platforms and experimenting with machine learning algorithms in order to group the similar items and reduce search space for shoppers.

The rest of the paper is organized as follows: Section 2 studies related works and summarizes different methods for search quality on the trading platforms. In this section, we also discuss main problems for search quality. In Section 3 we describe our data set (test corpus) and clustering results with different entities from phone product category for reducing search space. Finally, in Section 4 we briefly sketch future work and present the conclusions.

2 Related Works and Background

2.1 Analysis of Relevant Works

There are a lot of studies which are dedicated to the issues of improving online shopping experiences for consumers. On the e-commerce trading platforms, where the

number of choices is overwhelming, there is a need to filter, prioritize and efficiently deliver relevant information in order to alleviate the problem of information overload.

Such problem belongs to the tasks of information retrieval. Some researchers solve this issue using recommendation systems. The recommendation system is defined as a decision making strategy for users under complex information environments [3]. Papers [4-6] are devoted to studying, comparing and analyzing personalized recommendation systems. The main types of recommendation systems are distinguished: collaborative filtering, knowledge-based, effect-based, rule-based, and content-based recommendation systems. As it was demonstrated [4], each of them has some disadvantages, and it is concluded that the combined application of a variety of techniques should satisfy the actual needs better.

The paper [7] proposes an alternative approach to retrieve information from a given e-commerce website, by collecting data from the site's structure, retrieving semantic information in predefined locations and analyzing user's access logs. It gives the opportunity to predict users' future behaviour. Some researchers [8, 9] suggest the extension of the technology acceptance model for its application in the e-commerce field by adding four criterion variables, namely, purchase, access number, access total time, and access average time.

The authors in the work [11] used genetic algorithms to optimize a specified objective function related to a clustering task for search. They have done some experiments on synthetic and real-life data sets which show the utility of the proposed method. Their analysis of the results of the experiment shows that the proposed method may improve the final output of k-means.

In the paper [12] authors created a new approach to a product recommendation. They investigated the possibility of using a hybrid recommend consisting of content based clustering and connections between clusters using collaborative filtering to make good product recommendations. The algorithm is tested on real products and purchase data from two different companies - a big online bookstore and a smaller online clothing store.

A lot of works use methods of machine learning, such as in the work [13] authors analyze products on shopping sites (Amazon and eBay). They use machine learning classifiers for grouping product descriptions. Also, they propose to use clustering techniques to detect taxonomy evolution.

Thus, we can conclude that many authors researched questions of data processing concerning goods on trading platforms. Different approaches were developed. In spite of that fact the formulation of the problem, which is given in our paper, isn't investigated. Our task is to research how groups of similar products can be distinguished based on item description in order to user could compare different products and then among similar products choose the best offer from different sellers.

2.2 Main Problems for Search Quality on Trading Platforms

There are two ways of search on trading platforms via keywords or item specifications. Most shoppers face the problem of the incorrect search result on their keyword request. For example, on eBay site we choose the category Cell Phones &

Smartphones, we use such filters as Format – Buy it now, Style – Bar, Condition – New and brand Sam-sung. At search results among Samsung models the following smartphone models as iPhone 7, iPhone 7 Plus LG Risio LG Treasure Sony Xperia XA ZTE Prestige are also presented. Apart from that covers for smartphones are also given at the search result.

Even using filters the search result may still contain errors. At filters, it is possible to choose Brand, Model, Color etc. But a seller also fills incorrect specification and we receive a lot of mistakes in a product description. While searching Apple iPhone Samsung Galaxy S8 is found. The seller accidently or intentionally put Brand – Apple and Model – Samsung Galaxy S8 (Fig. 1).

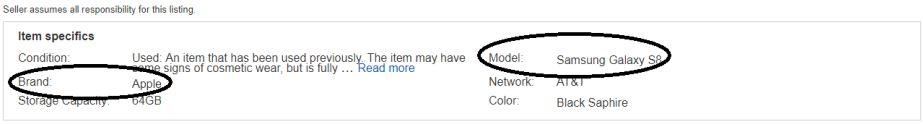

Fig. 1. Table of item specifics

Often, customers use photos for product search [10], but this does not greatly facilitate the search for the necessary product. Photos comparing is also a quite difficult task. Images of the same product can be presented in different ways. For example, smartphone's commodity picture can be presented in front or back side, be in the box or not, show the only photo or several (Fig. 2).

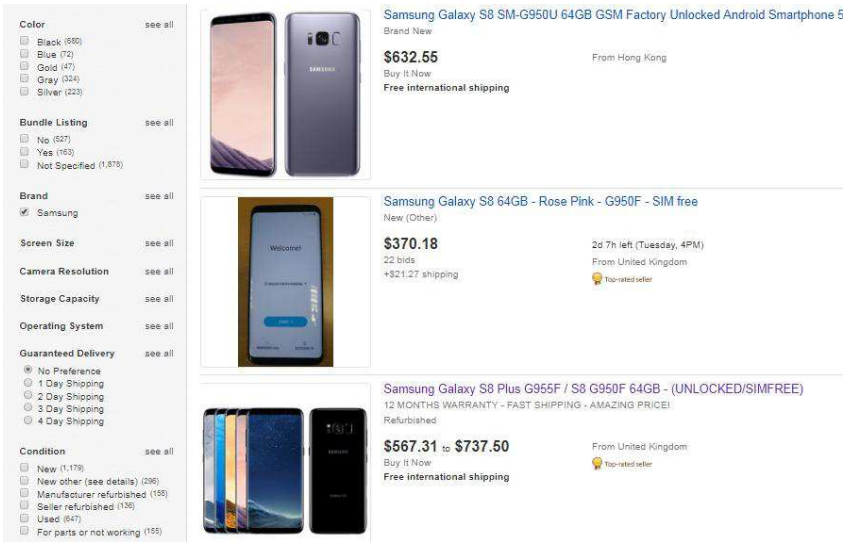

Fig. 2. Examples of different types of pictures

But all the mentioned approaches have some drawbacks. Using the recommendation systems requires considerable statistic data and doesn't allow comparing similar product items by their descriptions from different sellers. They provide propositions

for buyers; however important information can be missing. User profiles in some cases can give additional information, but it is also problematic to classify users according to their type of behavior.

The most crucial thing is that on e-commerce trading platforms sellers try to influence on search system using words from the most top-selling product names. Because of that e-commerce user receives irrelevant search results (Fig. 1). There is also a problem of the significant difference between the product names of the same product in different sellers as well as inaccuracy and gaps in the item description.

3 Experiments

3.1 Data set

Our task is to develop a service that will help customers quickly find the right product. For the experiment, the category of products "mobile phones" was chosen. Descriptions of goods are collected from different trading platforms. All found goods are entered into a single repository. The number of such commodity items is very large (for large trading platforms the number of offered phones reaches 370 thousand). This leads to the need to group goods. We assume that similar goods have similar descriptions. It is necessary to study how far the methods of machine learning can be used to group similar products on the basis of their descriptions. To collect and further process data, we have developed an architecture (Figure). On the diagram, we can see that the designed system has 3 layers and 3 microservices, namely: Data representation layer, which contains classes and interfaces of data objects (data transfer object), classes representing the essence of the non-relational database and the pattern of design Repository, which consists of two data sources - a system cache and a non-relational Google Firebase database that works with cloud storage services.

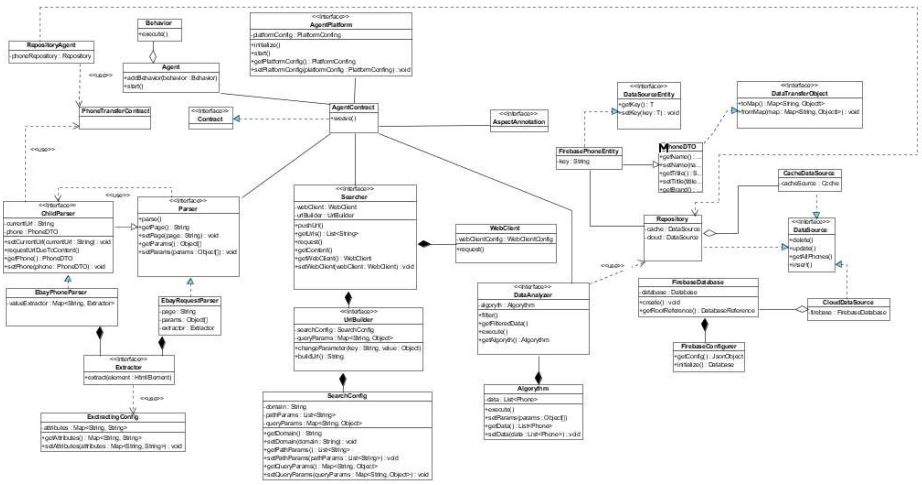

Fig. 3. Architecture of the software solution

The domain of the application layer is split from three microservices:

- agent platform;
- data seeker;
- data analyzer;
- parser and collector of data from Internet pages.

"Agent platform" service is responsible for executing business process processes in the body of individual agents.

The Data Finder service performs the process of searching for the required merchandise in trading venues by replacing the HTTP request parameters, filtering the configuration, and transitioning to data delivery pages.

The Parser tool analyzes an HTML document that answers an HTTP request from a searcher and collects the necessary data from the product description page.

The Data Analyzer service is responsible for analyzing the sampling collected by the parser, based on which the search parameters and product data stores will be reformed.

Due to the fact that the system has a clear division into layers of data representation and is distributed to separate microservices - testing processes are not labor-intensive. System components obviously do not depend on each other, the input data can be replaced by mock objects or test data for integrating and unit testing.

Thus, in order to solve the task of implementing the "Data Analyzer" service, it is necessary to choose an approach for putting products in order and constructing a model for estimating similarity. For this, a test set of product descriptions has been generated. The data is collected using the "Parser" service. The collected descriptions are stored as separate entities. For the experiment we use a sample of the description of smartphones from one site.

We have created own dataset from eBay.com website (https://www.ebay.com/). Our corpus contains 350 entities from Phone product category where each entity has 3-15 photos and product description. Some statistics are shown in Table 2. The example of product description and photos can be seen below (Fig. 4).

Name - SAMSUNG GALAXY S7 EDGE 32GB ORO SM-G935V VERIZON DESBLOQUEADO SMARTPHONE 5.5"

URL - https://www.ebay.com/itm/Samsung-Galaxy-S7-Edge-32GB-Oro-SM-G935V-Verizon-Desbloqueado-Smartphone-5-5/202155524056?epid=23011932310&hash=item2f11687fd8:g:9cMAAOSwaSZaO5fL

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

MPN - SAM-G935V
Cámara - 12 megapxeles
Memoria RAM - 4 GB
Memoria interna - 32 GB
Tipo - Barra
EAN - 0711202910423
Marca - Samsung
Modelo - Samsung Galaxy S7 Edge
Tamaño de pantalla - 5,5"
Color principal - Oro
Procesador - Quad core
Sistema operativo – Android

Fig. 4. The example of Smartphone photos with product description

Table 2. Statistics about data set

Category (Phone)	*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*

Such subcategories of product description as Name and Condition may contain very different texts. They may be different in length, words etc. because these sentences are written by the sellers. Other subcategories (Memory, Model, System etc.) are more similar in different descriptions. As a result the product description contains a lot of mistakes.

3.2 Experiments for Reducing Search Space

For the first stage, we try to use classical k-means clusterization algorithm for all product descriptions. Our algorithm uses the stopword list and TFIDF for vectorization of our texts. We received 350 samples with 2,659 features. Below can see the top terms per cluster (we take top 10 keywords for each cluster). Top terms per clusters are:

Cluster 0: used condition item apple 128gb iphone cosmetic fully functions previously

Cluster 1: sony xperia z5 premium z2 condition packaging e6853 new 32gb

Cluster 2: s7 samsung g930v galaxy 32gb edge sm g935v verizon unlocked
Cluster 3: htc desire 816 8gb 13mp sim dual condition android unlocked
Cluster 4: nokia 1100 apple iphone unlocked phone condition black gsm manufacturer
Cluster 5: s7 samsung 32gb galaxy sm unlocked smartphone 4gb condition lte
Cluster 6: samsung s5 galaxy 16gb 16mp packaging condition g900v 4g retail

As shown in the cluster list, three clusters (in blue color and underlined) have similar terms and are related to Samsung smartphone. However, we know that we only have two categories of Samsung brand: Galaxy S5 and Samsung Galaxy S7. This result is not very good for such a small sample, so we try to use the Porter stemmer for preprocessing our dataset. Our results with top terms per clusters are:

Cluster 0: nokia 1100 phone unlock mobil condit black manufactur cellular refurbish
Cluster 1: use condit item floor previous cosmet wearbut return fulli store
Cluster 2: s7 samsung galaxi 32gb sm unlock g930v packag smartphon 4gb
Cluster 3: samsung s5 galaxi 16gb packag 16mp condit retail sm g900v
Cluster 4: appl iphon 128gb memori unlock condit io 32gb built smartphon
Cluster 5: htc desir 816 8gb 13mp sim dual condit unlock android
Cluster 6: soni xperia z5 premium packag z2 condit e6853 new unlock

In this case, we received two clusters for Samsung smartphone but only one for Sony smartphone. One cluster (in blue color and underlined) does not have a certain category and contains general keywords. As our next step, we use HashingVectorizer which hashes word occurrences in a fixed dimensional space. The word count vectors are normalized to each have l2-norm equal to one (projected to the Euclidean unit-ball) which be important for k-means to work in high dimensional space.

HashingVectorizer does not provide IDF weighting as this is a stateless model. When IDF weighting is needed it can be added by pipelining its output to a TfidfTransformer instance.

It can be noted that k-means (and minibatch k-means) are very sensitive to feature scaling and that in this case, the IDF weighting helps improve the quality of the clustering by quite a lot as measured against the "ground truth" provided by the class label assignments of our dataset.

After all steps, k-means works with 837 features. Top terms per clusters:

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

For this k-means version, we receive good results. We estimated the quality of received results for 350 samples with 2,659 features represented on a trading platform. All our entities (smartphones) have a separate category. We receive good small value for the Silhouette Coefficient (*Silhouette Coefficient: 0.125*) which is actually for high dimensional datasets such as text data. Other measures such as Precision, Recall are also very good, their values being 0.95, 0.95 respectively.

As an experiment, we try to see how the k-means algorithm works with incorrect data. For that, we take a product description which contains the word Samsung in the Name field, and the word Apple in the Brand field.

The experiment shows, the algorithm is very sensitive to such data. If the Name field contains any words offering to Samsung models (e.g., Samsung Galaxy S5), then the algorithm takes this product to the Samsung category. However if the Name field contains only keyword Samsung without any other words, then this description is taken to the category Apple.

Thus, the usage of the clustering algorithm does not allow identifying unscrupulous sellers. But it provides relevant clusters and gives a possibility to distinguish product items. The next step can be processing data in the separate group.

4 Conclusions and Future Works

In this paper, experimenting results of product item description preprocessing are presented in order to compare and define the similar items. It provides the ability to decrease search space and to build an algorithm for buyer assistance. The obtained results of this study can help in future to create the integrated method for disambiguation of item description in order to improve the search quality, whereas only clusterization method could not help with solutions to all problems.

However, the k-means method can be used for preprocessing of the item description. The creation of a combined method, using classical algorithms and specific approaches, will allow increasing of search quality and buyer satisfaction.

In future works, it is supposed to create an approach to commodity grouping, which will combine product description and photo. It'll give the opportunity for a buyer to find the same products in reduced search space with dissimilar pictures and names, to find the same sellers at different trading platforms. It is supposed to make experiments not only with smartphones but with clothes and bags also.

5 References

1. Internet trading in Ukraine, https://netpeak.net/ru/blog/15-slaydov-o-tom-kak-razvivaetsya-rynok-elektronnoy-kommercii-v-ukraine/, last accessed 2018/01/29
2. https://www.ebay.com/
3. Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S.M., Konstan, J.A. et al.: Getting to know you: learning new user preferences in recommender systems. In: Proceedings of the international conference on intelligent user interfaces, pp. 127–34 (2002).

4. Ya, L.: The Comparison of Personalization Recommendation for E-Commerce. In: International Conference on Solid State Devices and Materials Science, Physics Procedia 25, pp. 475-478 (2012).
5. Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A.: Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal 16, pp. 261–273 (2015).
6. Ying, L., Boqin L.: Application of Transfer Learning in Task Recommendation System. Procedia Engineering 174, pp. 518–523 (2017).
7. Dias, J. P., Ferreira, H. S.: Automating the Extraction of Static Content and Dynamic Behaviour from e-Commerce Websites. Procedia Computer Science 109C, pp. 297–304 (2017).
8. Fayad, R., Paper, D.: The Technology Acceptance Model E-Commerce Extension: A Conceptual Framework. Procedia Economics and Finance 26, pp. 1000–1006 (2015).
9. Kumar Raja, D.R., Pushpa, S.: Feature level review table generation for E-Commerce websites to produce qualitative rating of the products. Future Computing and Informatics Journal 2, pp.118–124 (2017).
10. Kiapour, M. H., Han, X., Lazebnik, S., Berg, A. C., Berg T. L.: Where to Buy It: Matching Street Clothing Photos in Online Shops (2015).
11. Murthya, C.A., Chowdhury N.: In search of optimal clusters using genetic algorithms. Pattern Recognition Letters, vol. 17, issue 8, pp. 825–832 (1996).
12. Hansson, L.: Product Recommendations in E-commerce Systems using Content-based Clustering and Collaborative Filtering (2015).
13. Bankar, S., Anindya D.: Clustering for Taxonomy Evolution (2013).