Implicit Products in the Decentralized eCommerce Ecosystems

Conference Paper · August 2020

DOI: 10.1145/3383583.3398559

6 authors, including:

Guoxiu He
Wuhan University
10 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Zhuoren Jiang
Zhejiang University
51 PUBLICATIONS **220** CITATIONS
SEE PROFILE

Xiaozhong Liu
Indiana University Bloomington
145 PUBLICATIONS **821** CITATIONS
SEE PROFILE

Wei Lu
Wuhan University
52 PUBLICATIONS **204** CITATIONS
SEE PROFILE

Implicit Products in the Decentralized eCommerce Ecosystems

Guoxiu He*†
School of Information Management, Wuhan University
Wuhan, China
guoxiu.he@whu.edu.cn

Yunhan Yang†
School of Information Management, Wuhan University
Wuhan, China
yhuanny@gmail.com

Zhuoren Jiang
School of Public Affairs, Zhejiang University
Hangzhou, China
jiangzhuoren@zju.edu.cn

Yangyang Kang
Alibaba Group
Hangzhou, China
yangyang.kangyy@alibaba-inc.com

Xiaozhong Liu‡
Indiana University Bloomington
Bloomington, United States
liu237@indiana.edu

Wei Lu‡
School of Information Management, Wuhan University
Wuhan, China
weilu@whu.edu.cn

ABSTRACT

Detecting dark businesses in a decentralized eCommerce ecosystem (e.g. eBay, eBid, and Taobao) is a critical research problem. In this paper, we investigate the characteristics of dark implicit products, the associated buyer seeking behaviors, and features of classification model. Results demonstrate that dark implicit product detection is a challenging problem, while buyer seeking behavior information could be useful as a critical alternative to address this problem.

CCS CONCEPTS

• **Information systems → Query log analysis**;

KEYWORDS

spam detection, information seeking behavior, query log analysis

ACM Reference Format:
Guoxiu He, Yunhan Yang, Zhuoren Jiang, Yangyang Kang, Xiaozhong Liu, and Wei Lu. 2020. Implicit Products in the Decentralized eCommerce Ecosystems. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3383583.3398559

1 INTRODUCTION

Unlike a centralized eCommerce platform (e.g., Amazon), sellers can easily list the illegal products on the decentralized eCommerce platform (e.g., eBay, Taobao, Xianyu, Flipkart or eBid) without careful screening. How to detect these dark businesses attracts interest from both industries and academics. From the viewpoint of classical machine learning, dark implicit product detection could be a binary problem. However, when the current learning model (learned on products) finds a seller is listing an implicit product, the seller could easily change the product title or description and release it again with a new seller/product ID. This means dark implicit products and their sellers hide like chameleons in the eCommerce ecosystem while traditional learning algorithms can hardly detect them effectively. On the other hand, as eCommerce provider cannot save enough new training data in a short time window, the learning algorithm can hardly capture this dynamic for efficient implicit product detection [4].

In this paper, we aim to reveal the characteristics of dark implicit products compared with normal products. Moreover, we try to study the difference between user seeking behaviors and study how to detect these implicit products by utilizing user seeking behavior.

2 METHODOLOGY AND RESULTS

2.1 Dataset

In this study, we conduct experiments and analyses on a large scale pornographic product (one of dark implicit products) detection dataset which totally includes 401,701 normal products associated with 2,074,759 seeking sessions and 4,098 pornographic products associated 117,135 seeking sessions in Taobao (PPDD) [4].

2.2 Comparison of Product Characteristics

First of all, we employ word distribution and word density to explore the difference between pornographic and normal products in the perspective of word. Word distribution indicates distribution of share of top N words. And word density evidences how top N words can cover the word usage in products. As shown in Figure 1 (a) and (b), there is nearly no difference between online pornographic and normal products though we can see difference between local products. From Figure 1 (c), top N words in pornographic can cover more word usage than normal products. However, the gap decreases in online environment. In a word, we can hardly distinguish new dark implicit products based on the information of products.

2.3 Comparison of Buyer Seeking Behavior

From the perspective of information seeking, how buyers seek target products in eCommerce platform can be interpreted via the

*Work done as an intern at Alibaba Group.
†Both authors contributed equally to this research.
‡Corresponding authors.

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/08.
https://doi.org/10.1145/3383583.3398559

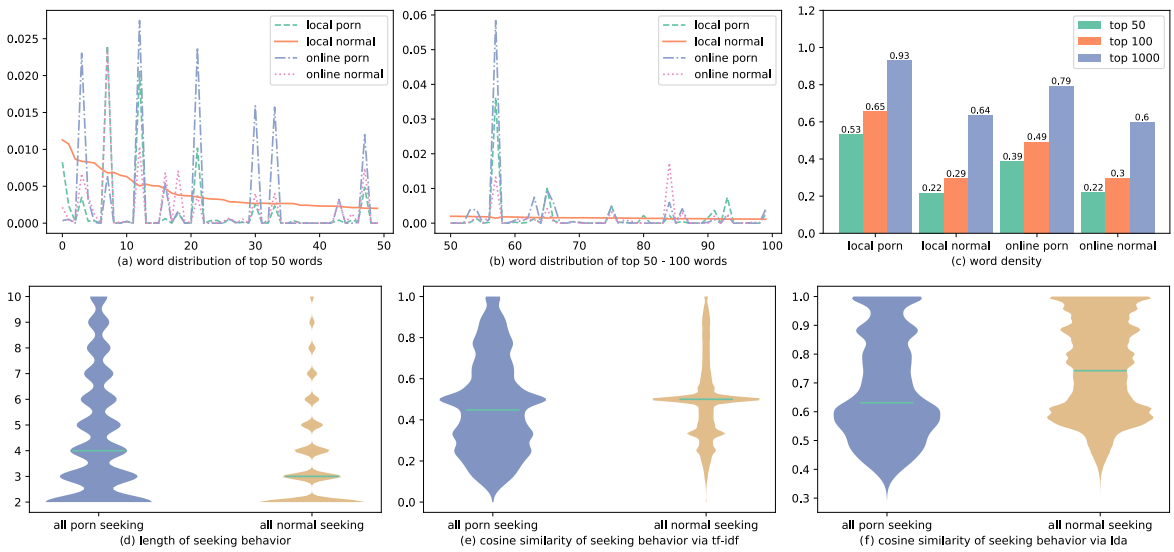

Figure 1: Characteristics of Products and Seeking Behaviors

Table 1: Results of Detection

	Online Test 1						Online Test 2					
Feature	Acc	Precision	Recall	F1 Score	AUC	LogLoss	Acc	Precision	Recall	F1 Score	AUC	LogLoss
Info	73.59%	51.11%	38.93%	44.20%	68.59%	0.9490	66.10%	64.40%	50.49%	56.60%	65.70%	1.3988
+ Seek	**79.86%**	**65.87%**	**51.97%**	**58.10%**	**74.34%**	**0.7425**	**73.11%**	**73.45%**	**60.44%**	**66.31%**	**74.50%**	**1.0652**

classical *berrypicking* model [1]. In this study, we focus on the query sequences in seeking sessions. The distribution of seeking session length shown in Figure 1 (d) reveals that buyers need to take more efforts to seek target pornographic products. Furthermore, we introduce TF-IDF [5] based bag-of-words and LDA [2] based topic distribution to represent each query. Then the average of cosine similarity among query pairs in each session is utilized to indicate the diversity of this session. The lower average similarity means higher diversity. As shown in Figure 1 (e) and (f), the diversity of pornographic product sessions is higher than normal ones especially by using LDA. Moreover, the distribution of diversity on pornographic session based on TF-IDF is more flat than normal ones. That means the word usage of seeking process of pornographic products faces more situations than normal ones.

2.4 Dark Implicit Product Detection

As aforementioned, seeking sessions involve more important features than products themselves. In this part, we compare the performance of concatenate text of product and related query sequence based GBDT [3], and the product text only based GBDT (use default settings in sklearn). As shown in Table 1, products plus seeking sessions based GBDT achieves a better performance under all metrics.

3 CONCLUSION

This study offers an investigation of dark implicit products on decentralized eCommerce platforms by data analysis and machine learning methods. From word distribution and word density of products, there is no significant difference between these implicit products and normal products especially on online environment. On the contrary, it is obvious that buyers take more efforts and more diverse queries to seek dark implicit products than normal ones. Based on these findings, the joint features based GBDT can significantly outperform product only based GBDT according to all classification metrics. In the future, we will explore more sophisticated methodologies to characterize the information seeking information for dark implicit product detection.

4 ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (61876003), Guangdong Basic and Applied Basic Research Foundation (2019A1515010837), and a scholarship from the China Scholarship Council (201906270034).

REFERENCES

[1] Marcia J Bates et al. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–424.
[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR* 3, Jan (2003), 993–1022.
[3] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.
[4] Guoxiu He, Yangyang Kang, Zhe Gao, Zhuoren Jiang, Changlong Sun, Xiaozhong Liu, Wei Lu, Qiong Zhang, and Luo Si. 2019. Finding Camouflaged Needle in a Haystack? Pornographic Products Detection via Berrypicking Tree Model. In *SIGIR*. 365–374.
[5] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *IP&M* 24, 5 (1988), 513–523.