International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:10, 2014

Identification of Spam Keywords Using Hierarchical Category in C2C E-commerce

Shao Bo Cheng, Yong-Jin Han, Se Young Park, Seong-Bae Park

Abstract—Consumer-to-Consumer (C2C) E-commerce has been growing at a very high speed in recent years. Since identical or nearly-same kinds of products compete one another by relying on keyword search in C2C E-commerce, some sellers describe their products with spam keywords that are popular but are not related to their products. Though such products get more chances to be retrieved and selected by consumers than those without spam keywords, the spam keywords mislead the consumers and waste their time. This problem has been reported in many commercial services like ebay and taobao, but there have been little research to solve this problem. As a solution to this problem, this paper proposes a method to classify whether keywords of a product are spam or not. The proposed method assumes that a keyword for a given product is more reliable if the keyword is observed commonly in specifications of products which are the same or the same kind as the given product. This is because that a hierarchical category of a product in general determined precisely by a seller of the product and so is the specification of the product. Since higher layers of the hierarchical category represent more general kinds of products, a reliable degree is differently determined according to the layers. Hence, reliable degrees from different layers of a hierarchical category become features for keywords and they are used together with features only from specifications for classification of the keywords. Support Vector Machines are adopted as a basic classifier using the features, since it is powerful, and widely used in many classification tasks. In the experiments, the proposed method is evaluated with a golden standard dataset from Yi-han-wang, a Chinese C2C E-commerce, and is compared with a baseline method that does not consider the hierarchical category. The experimental results show that the proposed method outperforms the baseline in F1-measure, which proves that spam keywords are effectively identified by a hierarchical category in C2C E-commerce.

Keywords—Spam Keyword, E-commerce, keyword features, spam filtering.

I. Introduction

CONSUMER-TO-CONSUMER (C2C) E-commerce has been growing at a very high speed in recent years. A key characteristic of C2C E-commerce is that users sell new or used products to each other. Thus identical or the same kind of products are sold by different users and they compete with each other by relying on a keyword search. Most keyword search technologies in C2C services rank products by not only matching scores but also selected frequencies of the products. Thus an easy way to highly rank a product is to describe the product with spam keywords that are popular but are not related to the products. The product with spam keywords gets more chances to be retrieved and selected rather than others without the spam keywords. However, spam keywords mislead consumers and waste their time to pick out good search results. This problem has been reported in many commercial services like ebay and taobao [1], [2], [3].

One way to solve this problem is to filter out spam keywords. However, little research exists to solve this problem in C2C E-commerce. Spam filtering has been studied mainly to improve email or ad hoc retrieval services [4], [5], [6], [7]. Most recent studies regard spam filtering as an instance of text classification which classifies a user's email message or a web page into the classes, spam and non-spam. The problem of spam keywords in C2C E-commerce can be also treated as text classification, where each keyword of a product is described as an instance and its class becomes one of spam and non-spam.

This paper proposes a method to classify whether a keyword of a product given by a seller is spam or not. In our method, features for keywords of a product are represented with reliable degrees from different layers in the hierarchical category of the product. The features from the hierarchical category are used together with those only from specifications for the classification of keywords. SVM (Support Vector Machine) is adopted as a basic classifier, since it is powerful, and widely used in many classification tasks.

There are two factors to be considered for measuring reliable degrees of a keyword. One is that a specification of a product is reliable regardless keywords of the product, and the other is that a product has a hierarchical category that is also reliable. Thus, a target keyword is regarded to be reliable if the keyword is observed commonly in specifications of products which are the same or the same kind as the given product. Then, features of the target keyword are determined by employing specifications from products in different layers of a hierarchical category. Therefore, reliable degrees of a target keyword is measured with a group of neighboring products. This is done by using information gain, which calculates the expected reduction in entropy caused by partitioning the group according to the existence of the target keyword in different layers in the hierarchical category. In the experiments, the proposed method is evaluated by using a Gold standard dataset from a Chinese C2C E-commerce, which consists of 8745 tagged instances belonging to one of three categories, *Electronics*, *Sports* and *Beauty Care*. The proposed method is compared with a baseline, where the baseline uses features obtained from only an individual product. Both of the baseline and the proposed methods showed high precisions in identifying non-spam keywords in all categories, which proves that specifications are reliable. The proposed method outperforms the baseline in F1-measures for

Shao Bo Cheng, Yong-Jin Han, Seyoung Park, and Seong-Bae Park are with School of Computer Science and Engineering, Kyungpook National University, Deagu 702-701, Korea (e-mail: {sbcheng,yjhan,sypark,sbpark}@sejong.knu.ac.kr).

identifying spam and non-spam keywords, which demonstrate that spam keywords are effectively identified by a hierarchical category in C2C E-commerce.

II. Related Work

A number of papers have been published on spam filtering in the area of email spam filtering. Spam email is any email that was not requested by a user but was sent to that user and many others. The methods for email spam filtering are mainly categorized in two, one method is a content-based spam filtering and the other is a machine learning based filtering [8]. In a content-based spam filtering, a blacklist containing spam words is predefined and the list is matched with a target email. The email is regarded as a spam if its similarity with the blacklist is higher than a certain threshold. In a machine learning-based spam filtering, a classifier determines whether a given email is spam or non-spam with features extracted as statistics of words in the email. Parameters for the classifier are learned from a corpus which includes emails annotated as spam or non-spam. The blacklist in content-based spam filtering can be understood as Spam keywords in C2C E-commerce and so are main features in machine learning based spam filtering. However, a main concern of the two methods is to classify an email not to find the spam words.

Spam filtering is also hot in ad hoc information retrieval. A spam web page is the page which is artificially-created to drive traffic to certain pages for benefits or just for fun [7]. Like the email spam filtering, the problem to filter out a spam page has been regarded as the classification of the page whether it is spam or not. While the web spam filtering filters out the whole spam web page, only spam keywords should be eliminated from a description of the product rather than completely rejecting the product in C2C E-commerce.

A spam tag problem has been newly introduced in recent years [6]. Tags are the words that shortly describe documents such as a blog or a twitter. A spam tag is added to a document by users to increase the visibility of the document. Koutrika et al. proposed a method to rank documents robustly against spam tags [6]. The method regards a tag as to be reliable in a document if the document is retrieved in a high rank by the tag. The reliability is dependent on the statistics of the tag in the document, since a legacy search engine represents a document as a bag of words. Spam keywords in C2C E-commerce can be understood as the spam tags in the spam tag problem by regarding a specification of a product as a document.

Unlike spam tag filtering, the proposed method employs a hierarchical category as well as a specification of a product. A hierarchical category is not available in the legacy tagging systems, but it provides a reliable information for a product in C2C E-commerce.

III. PROBLEM DESCRIPTION

Fig. 1 is an example of a product with spam keywords in a C2C E-commerce, Yi-han-wang. The product is an English learning education tablet which belongs to a hierarchical category, *Electric toys / Intelligence toys / Early childhood education*. Its specification is described in a table as shown in the bottom of Fig. 1. Generally, a specification of a product is precise since that is directly related to the reliability of the product. So is the category of the product since users can retrieve products with the category. However, the product is tagged with two spam keywords 'iPad' and 'iPadmini'. That is, the product is not 'iPad' or 'iPadmini'. Since the product is retrieved by the spam keywords as well as non-spam keywords, the product gets more chance to be selected as much as to be retrieved by the spam keywords. This results in that the product is highly ranked when competing with identical or the same kind of product without spam keywords. However, customers wastes their time to find good search results when they retrieve products with the spam keywords.

The problem of identifying spam keywords is an instance of binary classification. For a given keyword k of a product, a true function is defined as $f(\vec{x}) = y$, where $\vec{x}$ is the feature vector of k, and if k is a spam keyword then $y = 0$, otherwise $y = 1$. Parameters of the function f are learned from a training dataset by minimizing a loss $l(\hat{f}(\vec{x}), y)$. We adopt SVM (Support Vector Machine) as the classifier f, since it is powerful, and widely used in many classification tasks [9]. Actually, labSVM [10] is utilized for learning the function f.

IV. Feature extraction

Firstly, we defined that a record of a product is an individual product sold by a seller and consists of a set of keywords and a specification. Note identical products are sold by different users in C2C E-commerce. Thus records are distinguished though they are identical products. Basically, if a target keyword is observed commonly in specifications of records which belong to the given product or the same kind as the given product, then this keyword is regarded to be reliable. Since most products have at least two layered categories, in this paper, two features of a keyword for a product are extracted from two levels in the hierarchical category of the product. Fig. 2 depicts records and two different levels of reliable degrees in a hierarchical category.

First level is a product level, a product is regarded as a bag of words which includes all words from specifications of records grouped by the product. As shown in Fig. 2, records are grouped, of which products are identical. Then, reliable degree of a keyword for the product is measured as the expected reduction in entropy caused by partitioning the group of products according to the existence of the keyword in specifications. Actually, this becomes an information gain of the keyword in the group.

Then, reliable degree is measured in a category level. This is done by regarding a leaf category as a bag of words which includes all words from products grouped by the leaf category. The leaf categories are grouped according to its super category as shown in Fig. 2. An information gain of a keyword in the group is also measured in the same way of that in product levels. Though a keyword is not observed sufficiently in a product level, it can be observed more frequently in the category level rather than in the product level. Thus reliable degrees is measured in detail with the two level measurements. Let G^l be a set as a group of elements in one of two different

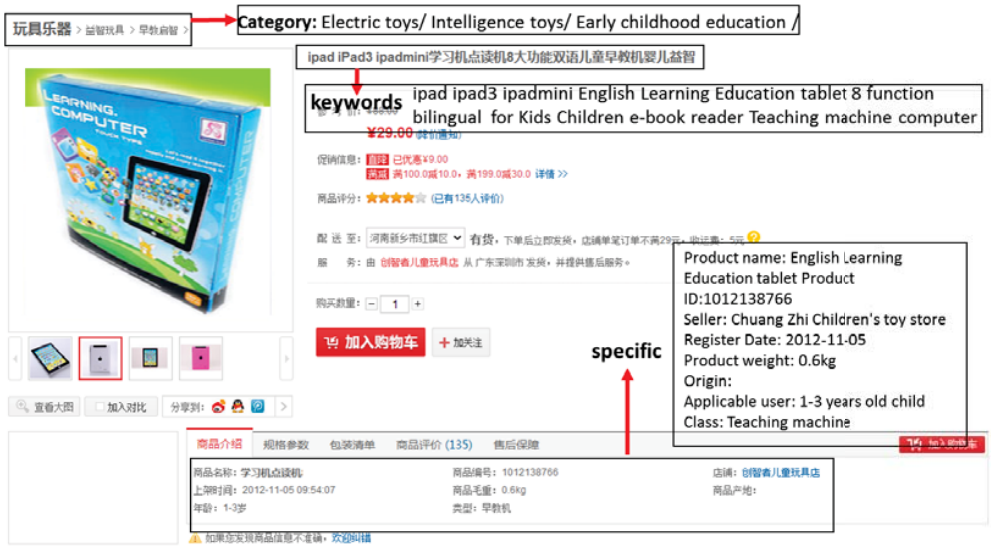

Fig. 1. An example of spam keywords in Chinese C2C e-commerce

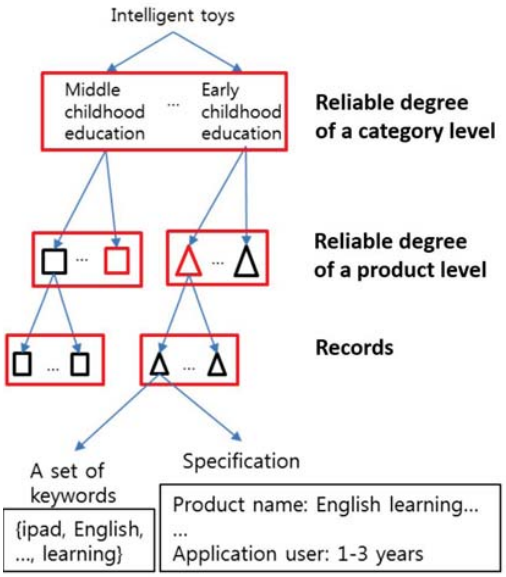

Fig. 2. Records and two levels of reliable degrees of a keyword

levels, where $l \in \{\text{p, c}\}$ and p, c denote product and category levels respectively. Each element $p \in G^p$ corresponds to a bag of words for a product. and a category correspond to elements of G^c. Then, information gain $IG(G^l, k)$ for a keyword k in a group G^l is given by

$$\begin{aligned} IG(G^l,k) = &- \sum_{g \in G^l} P\left(g\right) \cdot \log_2 P\left(g\right) + \\ &P\left(k\right) \cdot \sum_{g \in G^l} P\left(g|k\right) \cdot \log_2 P(k) + \\ &P\left(\bar{k}\right) \cdot \sum_{g \in G^l} P\left(g|\bar{k}\right) \cdot \log_2 P(\bar{k}), \end{aligned} \tag{1}$$

Here, $P(g)$ represents the probability of g and it becomes $\frac{count(g,\cdot)}{\sum_{g' \in G^l} count(g',\cdot)}$ by assuming uniform distributions of words, where $count(g,\cdot)$ return all number of keywords in g. $P\left(k\right)$ is the probability that k is observed in G^l and it becomes $\frac{\sum_{g \in G^l} count(g,k)}{\sum_{g' \in G^l} count(g',\cdot)}$, where the function $count(g,k)$ returns the number of k in g. $P\left(\bar{k}\right)$ is the probability that k is not observed in G^l and becomes $1 - P\left(k\right)$. $P(g|k)$ is the conditional probability of g when k is observed in g, while $P(g|\bar{t})$ is that when k is absent in g.

The information gains of a keyword k in the three different levels are used as features for the classification of k. Two additional features are obtained by using individual records. One feature is a indicator whether k is observed in a record of k or not. The other feature is the inverse of a count that k is observed over all the records. The two features are motivated from the study on the spam tag problem by Koutrika et al[6]. They regards a tag of a document as to be reliable when the document is retrieved in a high rank by the tag. Since the two features corresponds to a term frequency and an inverse document frequency in a legacy search engine, spam keywords in our problem can be understood as spam tags. Therefore, four features are extracted in this paper, we combined and summarized them as Table I:

TABLE I
FEATURES OF A KEYWORD

Features	Description
Product level	An information gain of a keyword in a product layer
Category level	An information gain of a keyword in a category layer
Specification	A keyword is observed in a record of a product or not
IDF	Inverse counting statistics for a keyword is observed over all records of a product

V. Experiments

A gold standard data set was built from a Chinese C2C E-commerce, Yi-han-wang to evaluate the proposed method. We asked sellers to annotate keywords of their products belonging to one of three leaf categories, *Electronics*, *Sports*, and *Beauty Care*. Since some keywords contain stop words such as preposition, conjunctions, and so on, such stop words are filtered out and only the refined keywords were provided to sellers. We used a tool named jcseg which is well-known Chinese stop filtering API [11].

212 sellers responded and provided data that is annotated for whether the keywords of their products are spam or not. Among the annotated data, about 15% of records has at least one spam keyword. Table II shows the statistics of the data including spam keywords in record and product levels. In

TABLE II
Statistics of annotated data with at least one spam keyword

	# of records	# of products	$\frac{\text{\# of records}}{\text{\# of products}}$
Electronics	2577	184	14.0
Sports	2188	313	7.0
Beauty Care	3980	284	14.0
Total	8745	781	11.2

Table II, the ratio of records to items in a category implies the average number of products that are identical. Thus in all the three categories, average 11 users compete with each other to sell an identical product.

Total 123,501 instances are obtained from the annotated data. Among them, 36,593 instances belong to *Electronics*. The numbers of instances for *Sports* and *Beauty Care* are 32382 and 54526 respectively. Randomly selected instances of 80% are used as a training data set and the remaining instances are used as a test data set for each category.

The performance for identification of spam keywords is measured by using precision, recall, and F1-measure. Recall is defined as the percentage of correctly classified keywords among spam keywords. Precision is defined as the percentage of correctly classified keywords among those classified as spam. F1-measure is the harmonic mean of the recall and the precision.

The proposed method is compared with a baseline method, where the baseline uses features obtained from only individual records. The results are shown in Table III.

TABLE III
Evaluation results of proposed approach on spam keywords

	Proposed method			Baseline		
	Rec. (%)	Pre. (%)	F1 (%)	Rec. (%)	Pre. (%)	F1 (%)
Electronics	88.4	70.3	77.7	92.0	42.5	58.1
Sports	79.8	74.4	75.8	89.4	34.1	49.4
Beauty Care	85.5	68.9	76.3	90.1	33.3	48.6
Average	84.6	71.2	77.2	90.5	36.6	52.0

The baseline shows higher recalls rather than the proposed method in all the categories. This is because most keywords are classified as a spam, when the keywords are absent in their specifications. As a result, the baseline failed to classify spam keywords correctly. Note that precisions of the baseline are less than 50% in all the categories. Though recalls of the proposed method is lower than those of the baseline, the proposed method outperforms the baseline in all the categories by F1-measure. This is because, the proposed method classifies correctly spam keywords rather than the baseline and the superiority overwhelms the differences of recalls between the two methods as shown in Table III. The results demonstrate that reliable degrees by a group of neighboring products from different layers in the hierarchical category of the given product is meaningful to classify spam keywords. Misclassification of non-spam keywords by a classifier is significant in respect to sellers, since the misclassification blocks their product to be retrieved by the misclassified keywords. Thus performances of the proposed and the baseline methods are compared in terms of non-spam keywords. Table IV shows the results. Precisions of the

TABLE IV
Evaluation results of proposed approach on normal keywords

	Proposed method			Baseline		
	Rec. (%)	Pre. (%)	F1 (%)	Rec. (%)	Pre. (%)	F1 (%)
Electronics	94.7	98.3	96.5	82.2	98.6	89.7
Sports	96.9	97.7	97.3	80.4	98.5	85.5
Beauty Care	94.1	97.6	95.8	71.4	97.8	82.5
Average	95.3	97.8	96.5	78.4	98.4	87.3

baseline are over 97% in all categories. The results prove that a specification of a product is reliable since one of key features in the baseline is an indicator of existence of a target keyword in its corresponding specification. Precisions of the proposed method is slightly lower than those of the baseline. However, the precisions are also over 97% and the proposed method outperforms the baseline in recall and F1-measure. The results demonstrate that the proposed method keeps non-spam keywords well rather than the baseline.

VI. Conclusion

This paper proposed a method to classify a keyword for a product into spam or non-spam in C2C E-commerce by employing identical or the same kind of products in a hierarchical category. The proposed method measures reliable degrees of a target keyword in three different levels, record, product, and category levels. The measured values are used as features of the keyword with simple statistics of keywords from individual records. Since the reliable degrees is measured in detail, the proposed method provides more discriminative values as features for the classification. In the experiments, the proposed method outperforms the baseline in terms of identification of both spam and non-spam keywords, where the baseline uses only features from individual records. The results demonstrates that the proposed method employing reliable degrees in a hierarchical category is promising to filer out spam keywords in C2C E-commerce.

Acknowledgements

This work was supported by the ICT R&D program of MSIPIITP. [10035348, Development of a Cognitive Planning and Learning Model for Mobile Platforms]

REFERENCES

[1] camino3x2, "Keyword spam busting." [Online]. Available: http://www.ebay.com/gds/ Keyword-Spam-Busting-/ 10000000001612568/g.html
[2] fransgems, "Beware misleading item headers on ebay auctions!" [Online]. Available: http://www.ebay.com/gds/Beware-Misleading -Item-Headers-on-eBay-Auctions- /10000000003890459/g.html
[3] jandbcannon, "You might be keyword spamming too." [Online]. Available: http://www.ebay.com/gds/ Beware-You-Might-Be-Keyword-Spamming-Too- /10000000001620833/g.html
[4] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review*, vol. 29, no. 1, pp. 63–92, 2008.
[5] G. Cormack, "Email spam filtering: A systematic review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 4, pp. 335–455, 2007.
[6] G. Koutrika, F. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina, "Combating spam in tagging systems," in *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, 2007, pp. 57–64.
[7] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *Proceedings of WWW*, 2006, pp. 83–92.
[8] A. Khorsi, "An overview of content-based spam filtering techniques." *Informatica*, vol. 31, no. 3, 2007.
[9] A. Hearst, J. Dumais, and S. .B, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.
[10] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM TIST*, vol. 2, no. 3, p. 27, 2011.
[11] "jcseg." (Online). Available: http://code.google.com/p/jcseg/