

## Identification of Spam Keywords Using Hierarchical Category in C2C E-Commerce

**Authors :** 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 e-bay 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

**Conference Title :** ICCIS 2014 : International Conference on Computer and Information Sciences

**Conference Location :** Osaka, Japan

**Conference Dates :** October 12-13, 2014