Discovering product counterfeits in online shops: a big data integration challenge

ERHARD RAHM, University of Leipzig

Counterfeit products are illegal imitations or replicas of products offered for sale. The Int. Chamber of Commerce (www.icc-ccs.org/icc/cib) estimated in 2011 that counterfeiting accounts for about 5-7% of world trade, and an estimated $600 billion a year. In parallel to the increasing volume of online sales, the volume of counterfeiting in web sales has also increased. Fake products are offered and sold in numerous online shops and auction sites as well as on B2B marketplaces for wholesale trading. Almost all kinds of products are subject to counterfeiting, ranging from, say, electronic devices and apparel to food and drugs (Heinonen et al. 2012). Counterfeits not only cause an enormous economic loss, but can also damage the reputation of a brand. Buyers of fake products not only receive a low-quality product in many cases, but may even be exposed to serious safety and health risks, e.g., in the case of fake pharmaceuticals.

Taking actions against counterfeiting in online sales is challenging due to the *huge number of involved merchants, websites and product offers*. Furthermore, the online business is highly dynamic where *product offers as well as websites and merchants change frequently*. Identifying likely counterfeits thus requires largely *automatic approaches to monitor websites and product offers*. *Deciding whether an offered product is faked is also very difficult*, since the offers often use the description and images of the genuine products. Even humans are thus frequently unable to fully validate the authenticity of offered items which is a main reason for the increasing number of counterfeits offered and sold online (Heinonen et al. 2012). A main goal should thus be to allow a better manual decision about the trustworthiness of an offer by collecting diverse information about a product offer including the product itself and the merchant. Providing this information in sufficient quality entails tailored data preparation and data integration steps.

The task of discovering online counterfeits thus has the characteristics of a challenging “big data” problem. It involves a large volume of heterogeneous and dynamic data from numerous sources that needs to go through a complex processing pipeline with data acquisition, cleaning, integration and analysis steps. Discovered counterfeits may be used to build predictive models for an improved and faster identification of likely counterfeit offers.

Despite the high significance of dealing with counterfeits there are almost no published research results on how to best discover counterfeits in online shops. There is a large body of work on information extraction and data integration for web data, but most of this work does not focus on product offers which are especially challenging to deal with. This is because of the large spectrum of products and limited data quality of many product offers that require the adoption of comprehensive and tailored preprocessing and data cleaning steps. Previous product-related work includes learning-based approaches for matching product offers (Kannan et al. 2011, Koepcke et al. 2012) and for price monitoring (Agrawal and Ieong 2012).

There are various parties who can benefit from a semi-automated discovery of counterfeits, in particular the manufacturer (owners of the trademark / copyright) of offered products, public authorities involved in fighting counterfeiting (e.g., customs, police), owners of an online sales platform as well as customers who are interested in buying original products. To make the problem of counterfeit discovery more tractable, possible solutions could focus on one of these user groups and hence consider only a restricted number of sales platforms and/or product types and products. For example, manufacturers are only interested in detecting imitations of their own products in different online platforms. They also have the best knowledge about the genuine products and can thus help to identify counterfeits. By contrast, the owners of an auction site or market place are interested in detecting fake products of several or all product types only on their own site.

Possible solutions for counterfeit discovery require approaches for identifying web shops offering potential counterfeits for products of interest as well as site-specific approaches to find and extract all offers for these products. Furthermore, a set of *counterfeit indicators* needs to be found in order to score the counterfeit suspiciousness of product offers. These indicators likely depend on the product type (e.g. unusual size or quantities for drugs) but can also include general features as unusually low price or negative customer feedback on the product or the merchant. A combination of these indicators should help identify likely counterfeits that need to be manually verified before further actions are taken, e.g., to remove an offer from a website or to buy the product to prove the fraud.

Counterfeit scores can be used within different strategies to find likely counterfeit offers. The design and evaluation of such strategies are good topics for future research. One approach is to use match approaches for product offers to first cluster all offers that likely refer to the same real-world product. This is especially challenging for platforms such as auction sites that lack clean product descriptions but allow any merchant to provide different offer descriptions. Compared to standard match approaches the clustering should also include near-matches such as counterfeits. In a second step one can then use counterfeit indicators to compare the offers of a cluster to determine their relative suspiciousness, e.g., based on differences on price or merchant reputation. Another strategy is to learn classification models per product type to estimate the counterfeit suspiciousness based on selected indicators. Such an approach depends on a sufficient amount of training data. The two strategies could also be combined by using the first approach to identify an initial set of suspicious offers and then using the manually verified outcomes for training in a learning-based approach.

Selecting the most effective indicators and determining the best strategies for their use entails comprehensive evaluations for different product types and websites. Such evaluations must by nature be approximate due to the mentioned difficulty of deciding about the originality of an offered product. Ideally, such evaluations are conducted in cooperation with the potential users of the counterfeit approach, e.g., the manufacturer of the products or the owners of the sales platform. For evaluation purposes, the provision of benchmark datasets with both genuine and fake product offers is desirable.

REFERENCES

Rakesh Agrawal , Samuel Ieong. 2012. Aggregating web offers to determine product prices. In: *Proc. 18th ACM SIGKDD Conf.* DOI: http://dx.doi.org/10.1145/2339530.2339602

Justin A. Heinonen, Thomas J. Holt and Jeremy M. Wilson. 2012. Product counterfeits in the online environment: an empirical assessment of victimization and reporting characteristics. *Int. Criminal Justice Review.* DOI: http://dx.doi.org/10.1177/1057567712465755

Anitha Kannan , Inmar E. Givoni, Rakesh Agrawal, Ariel Fuxman. 2011. Matching unstructured product offers to structured product specifications. In: *Proc. 17th ACM SIGKDD Conf.* DOI: http://dx.doi.org/10.1145/2020408.2020474

Hanna Koepcke, Andreas Thor, Stefan Thomas and Erhard Rahm. 2012. Tailoring entity resolution for matching product offers. In *Proc. 15th Intl. Conf. Extending Database Technology (EDBT).* ACM. DOI: http://dx.doi.org/10.1145/2247596.2247662