A Cross-Modal CCA-Based Astroturfing Detection Approach

Xiaoxuan Bai[1], Yingxiao Xiang[1], Wenjia Niu[1(✉)], Jiqiang Liu[1(✉)], Tong Chen[1], Jingjing Liu[1], and Tong Wu[2]

[1] Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing 100044, China
{niuwj,jqliu}@bjtu.edu.cn
[2] Tsinghua University, Beijing 100084, China

Abstract. In recent years, astroturfing can generate abnormal, damaging even illegal behaviors in cyberspace which may mislead the public perception and bring a bad effect on both Internet users and society. This paper aims to design a algorithm to detect astroturfing in online shopping effectively and help users to identify potential online astroturfers quickly. The previous work used single method text-text or image-image to detect astroturfing, while in this paper we first propose a cross-modal canonical correlation analysis model (CCCA) which combines text and images. First, we identify several features of astroturfing and analysis these features. Then, we use feature extraction algorithm, image similarity algorithm and CCA algorithm, and propose a cross-modal method to detect astroturfing which release comments with pictures. We also conduct an experiment on a Taobao dataset to verify our method. The experimental results show that the supervised method proposed is effective.

Keywords: Astroturfing detection
Canonical correlation analysis algorithm · Cross-modal method
CCCA model

1 Introduction

With the rapid development of Internet, especially the popularity of the online shopping, produced unprecedented significant impact on the way that people live and goods purchase. However, there are a large number of astroturfing with their false comments in the product comments, which may affect the user's point of view and guide public opinion [1,2]. Because the network is virtual, consumers are difficult to select the best quality goods among various kinds of products through the pictures. In recent years, online shopping has become a part of people's lives, although consumers enjoy the convenience of online shopping. So consumers tend to refer to the comments in the goods to decide the choice, but in order to improve the credibility, sales, baby popularity, most merchants use astroturfing to brush praise. The comments of astroturfing is likely to mislead the

S. Qing et al. (Eds.): ICICS 2017, LNCS 10631, pp. 582–592, 2018.
https://doi.org/10.1007/978-3-319-89500-0_50

purchasers and affect them to select the goods incorrectly, the existence of false comments seriously affect the reference value of the information, misleading the consumer's judgment of potential consumers greatly. Therefore, in order to create a good online shopping environment and protect the interests of consumers, detecting the online astroturfing is very important [3].

At present, most of the comments on shopping sites is a combination of text and picture comments website such as Taobao. Astroturfing which post comments with pictures in online shopping can be probably divided into two categories. One class is that most of the astroturfing tend to post similar comments directly and selected the original pictures of goods in the comments for convenience, we can see that their images are almost identical, and the words in the text comment are similar. The word repetition rate is so high and the overall meaning of the comment is roughly the same. In addition, the pictures selected or intercepted by users might be affected by resolution, format, and so on. Therefore, the pictures similarity will not be high only through the picture recognition, it is difficult to detect the astroturfing. The combination of pictures and text can express the overall meaning of the comments, and improve the similarity of the comments.

To warm up, we can use the CCCA model to combine the text and the picture and transform them to each other. Hence, we label this waterarmy "astrotufing 1" who publish similar text and images, and we use CCCA model to detect them.

The other class is that a lot of astroturfing post other pictures casually instead of buying goods which makes the pictures of comments are inconsistent with the corresponding goods. Therefore, the text comments are similar while the pictures comments are irrelevant, so that the comments pictures will have low similarity to the goods pictures. Hence, we label this waterarmy "astrotufing 2" who publish similar text and different images, and we use image similarity algorithm to detect them.

The structure of this paper is organized as follows. Section 2 will discuss the related work in this field. Section 3 will present astroturfing detection methods through CCCA model. Section 4 gives experimental and results. Finally, conclusion is showed in Sect. 5.

2 Related Work

Currently, the study of astroturfing has made great progress compared to previous years ago. According to the different features, the methods of astroturfing identification using, mainly divided into three categories: based on content characteristics, based on behavioral characteristics and based on synthetic features.

Content Based Approach. Content-based approaches are based on the comments similarity and its linguistic features to extract comments of similar content and discover false reviewers. Through the analysis of tendency of the text in comments, the false comments issued by the astroturfing could be found. Ott et al. [2] studied deceptive opinion spam that have been deliberately written to sound

authentic, and they verified that the text feature of the comment can be used to identify the false comments. Duh et al. [4] found that astroturfing released a false comment that deviated from the normal user comments by analyzing the tendencies of the review text.

Behavior Based Approach. Behavior-based approaches refer to the astroturfing has a number of comments focused on sudden, extreme, releasing early product reviews and so on. Lim et al. [5] identify several characteristic behaviors of review spammers and model these behaviors so as to detect them. They analyzed a large number of product reviews in Amazon and extracted similar comments and propose scoring methods to measure the degree of spam for each reviewer. Mukherjee et al. [6] propose a novel angle to the problem by modeling spamicity as latent, they use users and their published comments to build classifiers and use the characteristics of astroturfing to distinguish itself with ordinary users.

Multiple-feature Based Approach. Multiple-feature based approaches are the combination of the content characteristics and behavior characteristics of astroturfing, which utilize the artificial tagging the samples of astroturfing and the credibility theory of communication to identify astroturfing. Lu et al. [7] combine the astroturfing characteristics and content characteristics using the annotative factor graph model and identify the unknown network using the artificial tagging network of astroturfing samples and theory of credibility propagation. Mukherjee et al. [8] first put forward the e-commerce field network of astroturfing identification methods, they use the content of the comments to produce candidate groups, and then found astroturfing according to their characteristics.

Thus, it can be seen that the traditional methods of astroturfing identification in the field of e-commerce mainly based on the content similarity and its text features to find false commentators, the methods are very simple and inaccurate. This work, we combine text with pictures and use cross-modal CCA method to identify astroturfing in e-commerce.

3 Astroturfing Detection Method

3.1 Framework of Cross-Modal Canonical Correlation Analysis Model

This section is in order to find the astroturfing in the comments of products on Taobao site, we try to solve the problem through a newly proposed detection method, which utilizes the cross-modal canonical correlation analysis model and can effectively detect the astroturfing. The overall flow diagram of the model is shown in Fig. 1.

We will describe in detail the implementation of each algorithm proposed in this paper, and show how to achieve the CCCA model to detect astroturfing.

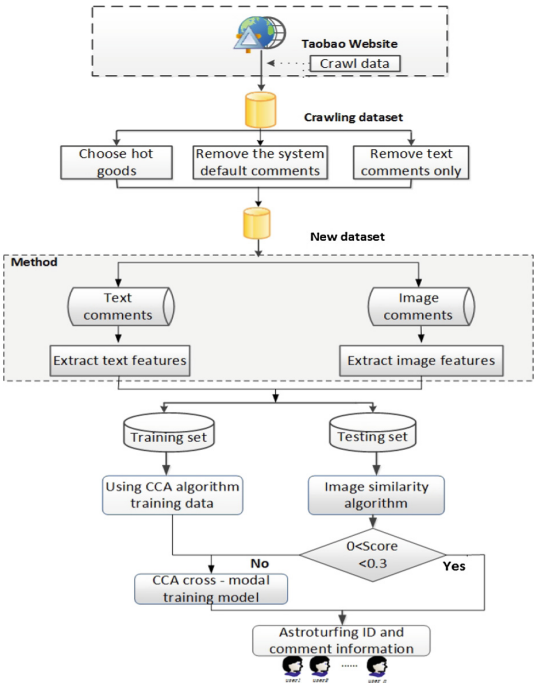

Fig. 1. Schematic representation of the astroturfing detection method framework based on CCCA model

3.2 Text Feature Extraction

The comments data obtained from the Taobao website can not be used directly as experimental data, so we preprocess the data. Since the comment is a paragraph of text, so it is necessary to convert the text into multi-dimensional eigenvector. First of all, we extract the keywords in the comments and split a text comment into a number of words, then we use these words to represent a document [9]. Text keywords extraction will be implemented by the Textrank algorithm. Hence, we present the specific steps:

(1) We split up the text comment T that we have climbed in accordance with the complete sentence. (2) For each sentence, we use word segmentation and word tagging, and filter out the stop words, only retain the specified part of the word, such as nouns, verbs, adjectives, retain candidate keywords. (3) Constructing candidate keywords graphs $G = (V, E)$, where V is a node set which is composed of the candidate keywords generated in step 2, then we use co-occurrence to construct the edges between any two points. There are edges between the two nodes, only when their corresponding vocabulary in the length of the window K co-exist, K represents the window size, that is, the most common K words. (4) According to the above formula, iterating and propagating the weight of each node until they converge.

$$R(w_i) = \lambda \sum_{j:w_j \to w_i} \frac{e(w_j, w_i)}{O(w_j)} R(w_j) + (1-\lambda)\frac{1}{|V|} \tag{1}$$

where $R(w)$ denotes the value of PageRank; $O(w)$ denotes the side of the degree; $e(w_j, w_i)$ denotes the weight from edge w_j to w_i; λ denotes the smoothing factor. (5) The node weights are sorted in reverse order, and the most important T words are obtained as candidate keywords. (6) The most important T words from step 5 will be marked in the original text, if adjacent phrases are formed, then they are grouped into multiple keywords.

3.3 Image Feature Extraction

As the pictures in the comments with text and picture can not be directly identified by the computer, we need to extract the feature of the image as a multidimensional eigenvector [10,11]. In this paper, we will use the HOG feature extraction algorithm.

The specific process is as follows:

(1) First carries on the grayscale to the picture in the crawled comment, the transformation formula is:

$$Gray = 0.3 * R + 0.59 * G + 0.11 * B \tag{2}$$

(2) Gamma correction method is used to the standardization (normalization) of color space for input images, we utilize the square root method for Gamma standardization, the formula is as follows (where $\gamma = 0.5$):

$$Y(x, y) = I(x, y)^{\gamma} \tag{3}$$

The gradient and gradient directions of the image are respectively calculated in the horizontal and vertical directions. Mainly to capture the contours and texture information, and further weakening the interference of light.
The gradient of the pixel (x, y) in the image is:

$$G_x(x, y) = H(x+1, y) - H(x-1, y) \tag{4}$$

$$G_y(x, y) = H(x, y+1) - H(x, y-1) \tag{5}$$

where $G_x(x, y)$, $G_y(x, y)$, $H(x, y)$ respectively represent the gradient and pixel values in the horizontal and vertical directions at the pixel points (x, y) in the input image. The original image is convolved with $[-1, 0, 1]$ and $[1, 0, -1]^T$ gradient operators, respectively, and the horizontal x and vertical y directions are obtained. And then we use the following formula to calculate the gradient size and direction of the pixel.

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{6}$$

$$\alpha(x, y) = \tan^{-1}\left(\frac{G_y(x, y)}{G_x(x, y)}\right) \tag{7}$$

(3) The image is divided into several small units, the gradient histogram of each small unit is counted. Several units make up a block, and the eigenvectors of all the units in a block are concatenated to get the HOG eigenvector of this block.
(4) The HOG eigenvectors of all the blocks in the image can be connected in series to get the HOG feature vector of the image. This is the final multi-dimensional feature vector available for classification.

Finally, the image feature vector format is $S_i = I^1_{S_i}, I^2_{S_i}, \ldots, I^n_{S_i}$.

3.4 Canonical Correlation Analysis Algorithm Based on Text and Image Cross-Modal Matching

After the text and image feature extraction, we use the processed feature data for cross-mode retrieval of text and images. To achieve cross-search between images and text, we first represent the image and text with a feature vector respectively, that is, mapping the image data to the image feature space I_1 and mapping the text data to the text feature space T_1 [12,13]. However, there is no direct connection between the feature spaces I_1 and T_1, CCA algorithm can map I_1 and T_1 to I_2 and T_2 respectively through the training of many "image-text" sample pairs, where the feature spaces I_2 and T_2 are linearly related and then make the training text and image features related [14]. The specific algorithm is as follows:

Let $t \in R^p$, $i \in R^q$ be the two random multivariate vectors [15]. $S_t = x_1, x_2, \ldots, x_m$, $S_i = y_1, y_2, \ldots, y_n$ represent two sets of vectors for text and images. T_i and I_i represent the text comments and corresponding picture comments in each comment. Let $w \in R^p$, $v \in R^q$ be the two projection vectors, the eigenvector spaces of w, v are expressed as $S_{wt} = (<w, t_1>, <w, t_2>, \ldots, <w, t_n>)$, $S_{vi} = (<v, i_1>, <v, i_2>, \ldots, <v, i_n>)$. The purpose of the algorithm is to find the projection vector w, v so that the correlations of $S_w x$ and $S_v y$ are greatest. The correlations can be written as $\rho^* = \max_{w,v} corr(S_{wt}, S_{vi})$. Figure 2 shows the Canonical Correlation Analysis (CCA) algorithm. The corresponding image and text pairs in each comment will be mapped to the same common subspace through training to find the correlation between them.

3.5 Astroturfing Detection Algorithm

In this section, the specific process of the detection algorithm can be described as follow: In the detection algorithm, the input $D_{experiment}$ is the comment data crawled through Taobao, and the output R_{user} is the user ID suspected astroturfing detected in the end. The algorithm firstly detects the second type of astroturfing and then detects the first type of astroturfing. First of all, to extract the text and image features of data set, and the data set is divided into two parts: the training set and the test set. The next step is to manually mark the suspected first type of astroturfing comments in training set, and assign

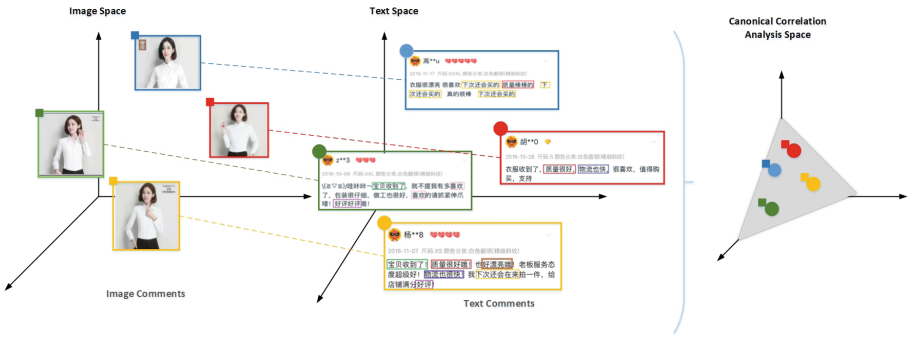

Fig. 2. Schematic of the proposed cross-modal canonical correlation analysis method

them to the label of "astroturfing 1". Then, the CCA algorithm is exploited to study the cross-modal learning for each pair of text and image comments, and a classification model is obtained. Finally, in the part of test, we compared the image similarity of the pictures in comments of test data set and the sample pictures of products provided by businesses. If the similarity score is less than 0.3, the comment may be suspected to be the second type of astroturfing, and the user's ID is output. Otherwise, the text comment and all the picture comments are projected into the common feature subspace o using the space projection function φ_T, φ_I, and then the K-nearest neighbor algorithm is used to find the closest category in the trained model and finally the results are output.

Algorithm: Astroturfing Detection Algorithm Based on Cross-Modal

Input: *the test set of comments database* $D_{experiment}$
Output: *astroturfing commnets* R_{user}

1. *Data preprocessing*: $D_{comments} \rightarrow D_{experiment}$
2. *Text feature extracting*: $S_i = T_{S_i}^1, T_{S_i}^2, ..., T_{S_i}^m$
3. *Image feature extracting*: $S_i = I_{S_i}^1, I_{S_i}^2, ..., I_{S_i}^n$
4. *CCA training model Building*: $(S_1, S_2, ..., S_i) \xrightarrow{CCA} Model$
5. *Astroturfing detecting*:

For $D_{experiment}$
 If $(0 < Score(I_{S_i}, I_{simple}) < 0.3)$
  Output user's ID
  Else $S_{io} \leftarrow \varphi_T(S_i)$
  For I_{S_i} *in* S_i *do*
   $S_{io} \leftarrow \varphi_I(S_i)$
   $KNN(S_i)$
   Output label
   If label = "$astroturfing1$"
 Output user's ID

4 Experimental and Results

4.1 Experimental Setup

We first get the raw comment data, and we crawl the comment data on Taobao's web page through the crawler program on the cloud-based server. In the experiment, we selected the top selling products of five different products to crawl the comment data, and the five items are from three different categories. Because the hot products have a huge amount of comments, there is a higher possibility to detect abnormal comments. In the end, we crawled 56,688 comments, and after preprocessing, there were 26,303 comments left with pictures, where each of the commentary contains 6 data items as follows: (1) Product ID; (2) Product name; (3) User ID; (4) Comment time; (5) Comment text; (6) Comment picture. The details of the crawl are shown in Table 1.

Table 1. The details of the product comments

Product ID	Product name	Number of comments
538868266734	Female T-shirt	5947
438870787421	White blouse	4545
536185035714	Men's sports pants	3678
520712769539	Female canvas shoes	8759
545963355120	Female bag	3374

First, we conduct an experiment on a commodity (commodity ID: 538868266734). It has a total of 19,941 comments, of which 5,947 comments with pictures, so the 4,500 comments with pictures is selected as the training set and the remaining 1,447 data as the test set.

Next, the training data set is manually annotated, we mark the suspected first class of astroturfing who publish the similar text and images as the label "astroturfing 1", and the other data is labeled as the "normal user". We utilize the gensim toolkit to extract the textual characteristics of the training data, and the feature vector files are obtained; Using the VLFeat visual library to extract the image features of the training data, and we get the feature vector files; Using the scikit-learn toolkit to learn the training data through the CCA algorithm.

According to the algorithm rules proposed in this paper, we carry out the test for test data, and finally output the user ID suspected of astroturfing.

4.2 Experimental Result and Evaluation

In this part, we will introduce the result of the experiment in detail.

Figure 3 shows the result for the astroturfing detection by CCCA model.

Then we do experiment on the other four products. We use the ROC curve to evaluate the classification accuracy of our experiment. The ROC curve and

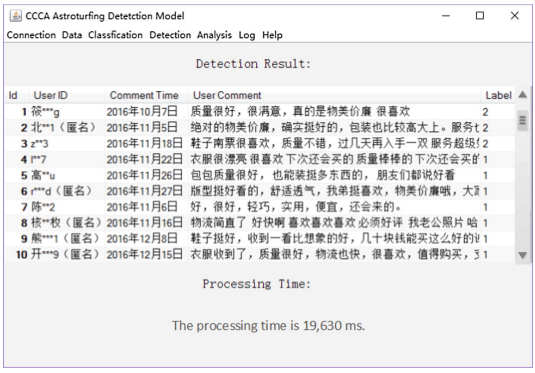

Fig. 3. The results of the test results of the astroturfing comments

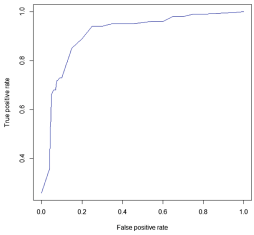

Fig. 4. ROC curve of experiment

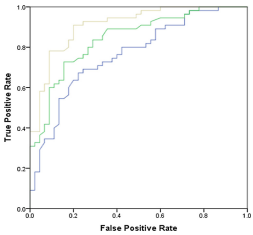

Fig. 5. ROC curve of three types of products (Color figure online)

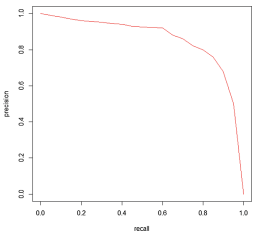

Fig. 6. Precise-Recall curve

the AUC value can be used to evaluate the pros and cons of a binary classifier. In this paper, we use the ROC curve and the AUC value to evaluate the classification accuracy of our experiment. The ROC curve of the experimental results is shown in Fig. 4. According to the accuracy of ROC curve for all test dataset, the accuracy of our detection method is 89.5%.

The ROC curves of three types of products are shown in Fig. 5. There are three curves which represent three types of products, the yellow one represent the clothing, the green one represent the shoes and the blue one represent the bags. We can see that the AUC value for clothing is 0.9143, the AUC value for shoes is 0.8762 and the AUC value for bags is 0.8236. Therefore, the astroturfing of clothing have high accuracy. Hence, the astroturfing may would like to publish their comments in clothing class.

As shown in Fig. 6, we can see that the precision rate is equal to recall rate when the value is about 0.8. It validates that the proposed cross-modal detection method of astroturfing have a good performance.

5 Conclusion

In this paper, we proposed a cross-modal CCA model to detect astroturfing in online shopping. To verify our method, we conduct an experiment on a Taobao dataset containing comments of manufactured products. We first extract text and image features, and use image similarity algorithm to detect the astroturfing which release pictures of goods irrelevant to the samples. Then, we use the CCA algorithm to study the cross-modal learning for each pair of text and image comments, mapping the text and image from their respective natural spaces to a CCA space. Finally, we use this method to detect astroturfing that publish pictures of goods almost same to the samples. Experimental results have demonstrated that the proposed method has a good performance. As part of our future work, we will explore and study more astroturfing features not only on shopping website and research more approaches to detect astroturfing.

Acknowledgments. This material is based upon work supported by the National Natural Science Foundation of China (Grant Nos. 61672092, 61502030, 61672091), Science and Technology on Information Assurance Laboratory (No. 614200103011711), BM-IIE Project (No. BMK2017B02-2), the Fundamental Research Funds for the Central Universities (No. 2017RC016), National High Technology Research and Development Program of China (863 Program) (No. 2015AA016003).

References

1. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 1–9. ACM (2010)
2. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 309–319. Association for Computational Linguistics (2011)
3. Chen, C., Wu, K., Srinivasan, V., Zhang, X.: Battling the internet water army: detection of hidden paid posters. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 116–120. IEEE (2013)
4. Duh, A., Štiglic, G., Korošak, D.: Enhancing identification of opinion spammer groups. In: Proceedings of International Conference on Making Sense of Converging Media, p. 326. ACM (2013)
5. Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., Lauw, H.W.: Detecting product review spammers using rating behaviors. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 939–948. ACM (2010)
6. Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., Ghosh, R.: Spotting opinion spammers using behavioral footprints. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 632–640. ACM (2013)
7. Lu, Y., Zhang, L., Xiao, Y., Li, Y.: Simultaneously detecting fake reviews and review spammers using factor graph model. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 225–233. ACM (2013)

8. Mukherjee, A., Liu, B., Glance, N.: Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st International Conference on World Wide Web, pp. 191–200. ACM (2012)
9. Peng, L., Bin, W., Zhiwei, S., Yachao, C., Hengxun, L.: Tag-TextRank: a webpage keyword extraction method based on tags. J. Comput. Res. Dev. **49**(11), 2344–2351 (2012)
10. Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., Cao, L., Huang, T.: Large-scale image classification: fast feature extraction and SVM training. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1689–1696. IEEE (2011)
11. Mizuno, K., Terachi, Y., Takagi, K., Izumi, S., Kawaguchi, H., Yoshimoto, M.: Architectural study of hog feature extraction processor for real-time object detection. In: 2012 IEEE Workshop on Signal Processing Systems (SiPS), pp. 197–202. IEEE (2012)
12. Pereira, J.C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 521–535 (2014)
13. Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 251–260. ACM (2010)
14. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2088–2095 (2013)
15. Ranjan, V., Rasiwasia, N., Jawahar, C.V.: Multi-label cross-modal retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4094–4102 (2015)