Event Detection from Image Hosting Services by Slightly-supervised Multi-span Context Models

Mohamed Morchid, Richard Dufour, Georges Linarès
University of Avignon
Laboratoire d'Informatique d'Avignon
Avignon, France
{mohamed.morchid, richard.dufour, georges.linares}@univ-avignon.fr

***Abstract*—We present a method to detect social events in a set of pictures from an image hosting service (*Flickr*). This method relies on the analysis of user-generated tags, by using statistical models trained on both a small set of manually annotated data and a large data set collected from the Internet. Social event modeling relies on multi-span topic model based on LDA (Latent Dirichlet Allocation). Experiments are conducted in the experimental setup of MediaEval'2011 evaluation campaign. The proposed system outperforms significantly the best system of this benchmark, reaching a F-measure score of about 71%.**

I. INTRODUCTION

The image hosting platforms such as *Picasa*, *Flickr* or *Drawin* allow users to easily share pictures or to browse image galleries. Nevertheless, searching in such large collections can be very difficult and the usual way to enable search consists of image tagging. In a perfect world, tags would be applied by human experts but this method is clearly too costly for open platforms that host billions of user-generated images.

Therefore, the image sharing platforms let the users annotate their own pictures - eventually the images they browse. This participative approach of tagging leads to inevitable annotation errors (bad spelling, unsuitable words, missing labels...). In such situations, a classic indexing system based on tag frequencies could not respond correctly to a given query.

This work focuses on high-level featuring of tagged images for social event detection. This task is part of the *Topic Detection and Tracking* (TDT) project [1]. Its goal is to detect a social event that took place or will take place in a particular location or at a specific date [2]. One of the first work on event detection was performed by [3], where authors used a clustering algorithm to detect events in a large a corpus. In [4], dependencies between *Flickr* and *Last.fm* are studied by using labels extracted from *Del.icio.us*. In [5], the authors seek to extract semantic contents from the meta-data of images posted on Flickr. These works encountered two major problems, related to the event modeling paradigm, and to the amount of training data usually required to estimate robust statistical recognizers.

This paper describes a method for the automatic detection of social event in an image sharing platform, by using only the textual meta-data (tags). This system relies on a multi-span topic-model, estimated by LDA (Latent Dirichlet Allocation), and a slightly supervised training strategy that allows to estimate models from a partially annotated data set.

LDA is a statistical model that considers a text document as a mixture of latent topics. One of the main problem of LDA representation lies in the choice of the topic-space dimensionality N, that is priorly fixed before the model estimate. Indeed, the model granularity is directly dependent from N: a low value leads to estimate coarse classes, that may be viewed as domain or thematic models. On the other side, a high dimensionality leads to narrow classes that may represent fine topics or concepts. Therefore, the choice of N determines the model granularity that should depend on the task objectives.

This problem was addressed in many previous works, that mainly focused on the search of the optimal model granularity; some of them proposed multi-granularity approaches, by hierarchical LDA modeling [6], bag of multimodal models [7] or multi-grain models [8]. All these approaches consisted in choosing the best granularity level, to obtain a partitioning of the semantic space that matches the task requirements; our proposal is to consider LDA models of different granularities as complementary views. The combination of these views could help to detect social events, in spite of the large variability of event representations in the tagged images.

Another key point is related to the amount of data required for the estimate of statistical models which could be able to detect the social event. We propose a slightly supervised approach for estimate an event signature; this method involves both human annotation of a small data set and a large corpus of web-collected data that are probably (but not certainly) from the expected class.

The rest of the paper is organized as follows: in the next section, the proposed system architecture is presented. Section III describe the experimental setup, of the MediaEval evaluation campaign, on which experiments were conducted. Results are discussed in IV before concluding in section V.

II. PROPOSED APPROACH

A. System overview

The system is composed of two modules that are successively applied. The first module (*WEB*) consists of extracting a set of Web pages (see section II-B) from a query in order to estimate a word frequency based model. This model is then used to cluster images into the *relevant* or the *irrelevant* category, according to an optimized threshold.

The first module implements a relatively classical strategy that consists of retrieving relevant images by comparing the word-frequency model of the event and the image tags.

The second module (*SVM*) is in charge of finding relevant images among those that was automatically considered as irrelevant by the *WEB* module. This process takes as input a set of relevant images, and data unrelated to the initial query. As output, several sets of images are provided (see section II-C). Some topic spaces with different granularities are trained by a Latent Dirichlet Allocation (LDA) on a large Web corpus. A Support Vector Machine (SVM) classifier for each topic space is then learned from the relevant images extracted with the Flickr search module and from part of the irrelevant images rejected by the *WEB* module. All manually tagged images of the MediaEval benchmark (Challenge 1) are projected onto each topic-space and processed by the associated SVM classifier. Then, unanimity between SVMs is required to consider the image as relevant. This last combination permits to extract a new subset of relevant images.

Finally, a last process realizes the union between the two subsets of relevant images extracted by the *WEB* and the *SVM* modules. A final set of photos is obtained, which answers the query.

B. Word frequency based system (WEB)

This two-step process extracts a subset of relevant photos knowing a query. The first step has to retrieve a set of Web pages that responds to a query. The term frequency is computed for each word of this corpus. In the second step, each image is evaluated by using a similarity measure between its tags and the word frequency based model. Then, the set of images is divided into two subsets according to the similarity score: beyond an estimated threshold t, the image belongs to the relevant subset. Otherwise, the image is placed in the irrelevant subset. The entire process is detailed in the following sections.

1) The Web model estimation: The estimation of a word frequency based model m requires a large corpus of documents D. This corpus is composed of Web pages that match to a query (this query being provided by the organizers in the MediaEval benchmark). From this corpus, a set of representative words and their frequency is obtained. We chose to select the first Website[1] satisfying the MediaEval query[2] on the Google search engine. All the Web pages associated to this URL compose the corpus of documents used to estimate the model. Once the corpus D is collected, the probability of each word w of the corpus D is:

$$P(w|m) = \frac{|w|_D}{N_D} \tag{1}$$

where $|w|_D$ is the number of occurrences of the word w in the corpus D and N_D is the number of words in D. This set of words will be used to determine the degree of similarity between the image tags and the model m.

2) Image clustering: The set of pictures P_{ALL} is split into a relevant subset of P_{WEB} and an irrelevant subset according to a similarity measure δ between an image p and a model m. A set of the N most irrelevant pictures (P_{TRSH}) is extracted from this irrelevant set. Relevance decision relies on a threshold t applied on δ. δ uses equation 1 to estimate the probability of a word w knowing the model m. The probability of w knowing the image p is calculated in the same way. The similarity δ is given by:

$$\delta(p, m) = \sum_{w \in p} P(w|m)(1 + P(w|p)) \tag{2}$$

The threshold t is optimized using a query and a development corpus. This query comes from the MediaEval 2011 benchmark [9]. This challenge is about football events which take place either in Rome or Barcelona. An image having a similarity with the model m greater than t is considered close to the model and representative of its content. Knowing that, the model is determined using representative Web pages of the query. The photos p from the total set (P_{ALL}) with a high model similarity is considered as *relevant*. We then obtain two set of pictures (P_{WEB}) and a set of the N most irrelevant (P_{TRSH}) photo subsets.

C. Multi-span model (SVM)

In the *WEB* module query matching, the picture set is split into relevant and irrelevant parts. In this section, we describe the second proposed module that allows to recover a part of relevant photos unlikely rejected by the *WEB* module.

The first step consists of defining a set of topic spaces of different granularities. A classifier is then trained for each topic space. These classifiers allow to retrieve in all candidate images P_{ALL}, the most relevant ones. Then, a vote is processed to only keep images which obtained the unanimity. This process is described step-by-step in the following sections.

1) Thematic space modeling: Latent Dirichlet Allocation (LDA) [6] is a generative probabilistic model that considers a document as a "bag of words" resulting from a combination of latent topics. The statistical study of word co-occurrences in a database permits to extract a set of word classes that are often associated to topics. But there is no proof to establish an explicit link between the statistical model and the topic interpretation. These methods are widely used in natural language processing, such as LSI\LSA (Latent Semantic Indexing\Analysis) [10], [11]) or their probabilistic variant PLSI [12].

All these methods require a large corpus of data to correctly estimate their parameters. In our problem, this set is composed of meta-data (tags) from relevant images and newspaper articles that are not related to the challenge query. The query is sent to Flickr to retrieve a set of $8,000$ relevant photos P_{FKR}. A set of articles from the French Press Agency (AFP) written between 2000 and 2006 are added in the same proportion than the images ($8,000$ articles) to the set of meta-data of the relevant photos extracted from Flickr. Finally, a corpus of $16,000$ documents is used. This new set of documents is lemmatized with TreeTagger [13] to estimate 6 LDA topic spaces. The number of models and classes that compose them, are chosen in order to provide a sufficient granularity variety [6], [14]. Then, we obtain a set of models E_i that contains a different number of classes ($10, 20, 30, 50, 100$ and 200 classes in our experiments).

[1] *http://www.paradiso.nl*
[2] *may 2009 venue paradiso*

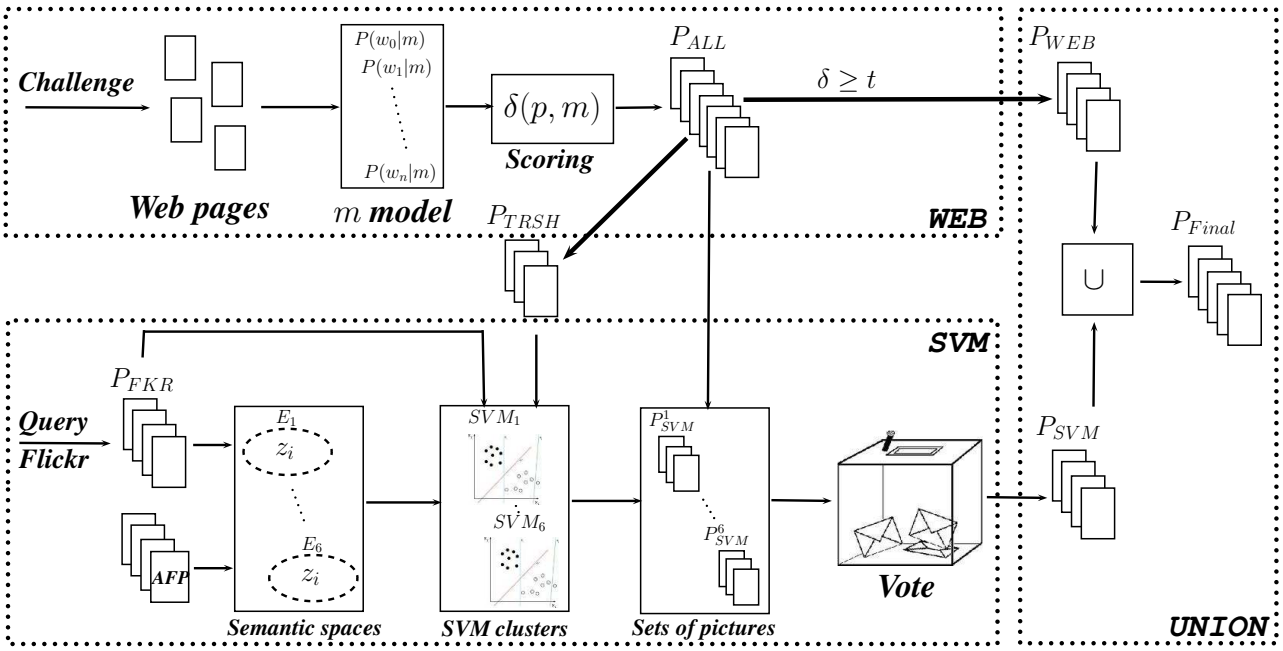

Fig. 1. Architecture of the proposed image extraction system depending on an initial query

2) Support Vector Machines: Support Vector Machines (SVMs) are a set of supervised learning techniques. Knowing a sample, SVMs determine a separation boundary between parts of the sample called *support vector*. Then, they compute a separating hyperplane that maximizes the margin between the support vectors and the separator hyperplane [15]. SVMs were used for the first time by [16] both in regression tasks [17] and in classification tasks [18], [19]. The popularity of this method is due to its good results in these two tasks and the low number of parameters that requires adjustment.

We use the binary SVMs (two classes) to classify into two subsets (*relevant* and *irrelevant* photos) each of the 6 topic spaces previously estimated by LDA. Photos taken from Flickr P_{FKR} and a set of N photos rejected by the basic model P_{TRSH} compose the training set of the SVM classifiers. A problem of $1:5$ (1 relevant photo to 5 irrelevant photos) is selected for this highly unbalanced problem. This configuration gets better results than a moderately imbalanced problem ($1:2$) or balanced ($1:1$) ones [20]. As a result, a corpus of $8,000$ relevant images (class $+1$) and $N = 40,000$ most irrelevant images (class -1) is obtained.

3) Representation of images in a topic space: A SVM is learned for each topic space E_i of n_z topics. Each document d from P_{FKR} or from the AFP articles is represented by a vector V_d of n_z elements $V_d[i]$ ($1 \leq i \leq n_z$) representing the similarity between the document d and the topic z_i of the topic space E_i. The following equation is obtained for each of the components of the vector V:

$$V_d[i] = \delta(d, z_i) \tag{3}$$

Each trained SVM is applied on all images P_{ALL}. A set of relevant pictures P^i_{SVM} is obtained for each topic spaces E_i.

4) Image selection by unanimous vote: A subset P_{SVM} is extracted from all the SVM image sets P^i_{SVM} by an unanimous vote. Thus, a photo belonging to every subsets is considered as relevant. The diversity of the number of topic space classes allows to consider different granularities to describe the photo (meta-data). For example, to describe a concert, a user can use the name of a band, the name of a band member, the major company or the music style (see figure 2).

D. P_{WEB} and P_{SVM} union

Two subsets of images are obtained by using the P_{WEB} and P_{SVM} modules. The union of these two subsets is finally performed to obtain a last subset P_{Final} corresponding to the request. Relevant images p are chosen with the formula:

$$\begin{cases} p \subset P_{Final} & \text{si} \quad p \in (P_{WEB} \cup P_{SVM}) \\ p \not\subset P_{Final} & \text{si} \quad p \notin (P_{WEB} \cup P_{SVM}) \end{cases}$$

III. EXPERIMENTAL PROTOCOL

For our experiments, we use the experimental protocol detailed in [9]. A request (or challenge) and a set of $73,645$ photos (P_{ALL}) and meta-data are provided (see figure 2). The goal of the challenge is to retrieve the $1,640$ pictures representing the challenge. The query used is the challenge 2 of the MediaEval benchmark:

"Find all events that took place in may 2009 in the venue named Paradiso (in Amsterdam, NL) and in the Parc del Forum (in Barcelona, Spain). For each event provide all photos associated with it".

The queries *"paradiso amsterdam"* and *"parc del forum*

barcelona" are sent to the Flickr platform to get a set of $8,000$ relevant photos P_{FKR}. These photos compose the corpus for the LDA analysis, and are used as the relevant class $(+1)$ for the SVMs training.

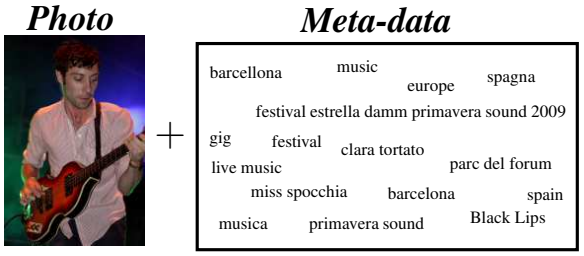

Fig. 2. Example of a photo and its associated meta-data

IV. RESULTS AND DISCUSSION

The first module identified $1,612$ relevant (P_{WEB}) and $72,033$ irrelevant photos, by using a threshold t of 0.133 estimated with the development set. The method is only based on Web pages collection, without any annotated corpus for the training of the frequency model. When few training data is available, this Web information is a good bootstrap to build a word corpus, which can achieve a satisfying level of performance with a limited effort. Indeed, with this approach, more than 68% of images representing a query can be correctly retrieved. We can note that the picture scoring formula (see formula 2) allows to give an heavier weight to frequent words simultaneously in a picture and in Web pages. However, a disadvantage of this similarity image/model is the absence of a weighting term that takes into account of the importance of the word in a set of documents (such as IDF (Inverse Document Frequency) [21]).

In the second module, a SVM classifier is learned for each of the 6 topic spaces with the annotated set provided by the MediaEval benchmark. The P_{TRSH} set, used for the SVM training process, consists of the $40,000$ less relevant pictures classified with the *WEB* module. Table I presents the results obtained on the extraction task for each topic space by the *SVM* module.

TABLE I. RESULTS OBTAINED WITH THE *SVM* APPROACH DEPENDING ON DIFFERENT TOPIC SPACES

#topics	#found	#exact	Prec.	Recall	F-meas.
10	6,822	1,255	18.4	76.5	30.0
20	6,811	1,377	20.2	84.0	32.6
30	6,076	1,264	20.8	77.1	32.8
50	8,006	1,315	16.4	80.2	27.2
100	7,744	1,127	14.5	68.7	24.0
200	7,304	1,417	19.4	86.4	31.6
Vote	**395**	**218**	**55.2**	**13.3**	**21.4**

Results show that single semantic space systems reach high recall scores (from 76.5% to 86.4%) but with relatively low precision rates (from 20.8% to 14.5%). As expected, the voting process dramatically improves the precision, but with a drastic reduction of recall scores. The combination of these two modules corresponds to a successive maximization of recall and precision, that obtained encouraging results. Table II reports the performance of the two proposed modules (*WEB* and *SVM*) and their union in terms of recall, precision and F-measure. In particular, we can note that 372 images are rejected (i.e. *irrelevant*) by the union.

TABLE II. PERFORMANCE OF THE WEB AND THE SVM MODULES, THEIR UNION, AND THE BEST SYSTEM OF THE MEDIAEVAL'11 CAMPAIGN [22]

Method	#found	#exact	Prec.	Recall	F-meas.
WEB (1)	1,612	1,108	67.6	68.8	68.2
SVM (2)	395	218	13.3	55.2	21.4
$1 \cup 2$	**1,900**	**1,268**	**77.3**	**66.7**	**71.63**
Best	1,737	1,164	67.01	70.99	68.95

The contribution of this representation by multiple semantic spaces is not negligible. As we can see in table II, the F-measure score is increased by 3.4 points (from 68% to upper than 71%), when the number of relevant pictures is improved by 10 points when comparing to the simple lexical use of meta-data (*WEB*) or with the best system of the MediaEval 2011 evaluation campaign (F-measure score of 68.95%) [22].

Finally, the use of this topic representation and SVM classifiers allows to retrieve 160 additional relevant images previously rejected by the word frequency-based system (*WEB* module). Indeed, $1,268$ photos are correctly found with the union of the two modules while only $1,108$ were considered as relevant with the *WEB* module. In fact, the *SVM* module is composed of more than 73% of photos that do not belong to the subset of the *WEB* model. This validates the basic idea that this supervised multi-span representation can retrieve images that were incorrectly rejected by the slightly supervised Web approach. We can also observe in table III that images which were not found by the union of both subsets have poorly structured meta-data.

TABLE III. META-DATA EXAMPLES OF SOME PHOTOS REJECTED BY THE UNION OF THE *WEB* AND *SVM* MODULES

Found images	Rejected images
audience primavera barcelona 2009 sound primavera barcelona img7584 2009 sound primavera	junk food 難食到呢 cimg0367 entrance

Globally, these results justify the proposed approach that consisted in tackling the meta-data weak structure by using a high-level representation the improves the initial "bag-of-tags" user-provided representation.

V. CONCLUSIONS AND PERSPECTIVES

In this paper, we proposed a robust method to extract representative photos from a query. This method provides a new alternative to characterize a photo not only with the annotated meta-data (tags), but with a representation of the nearest topics. This topic space is based on a Latent Dirichlet Allocation (LDA) with a corpus of relevant photos and random articles. Different subsets of images from various topic spaces were combined to extract a set of images depending on a query. Experiments shown the contribution of this proposed higher-level representation approach, with a better 71% F-measure score that outperforms of 3.4 points results obtained with the basic system use only. This shows that the representation of a document in topic spaces permits a robust image indexing

compared to a simple lexical representation. These results also highlight that a topic representation allows a strong abstraction of the textual photo content (meta-data).

The system proposed in this article has only been tested on a query and a single set of photos on the MediaEval 2011 benchmark. To enable a generalization of this work to other queries and on larger sets of photos, this approach will be evaluated on various kinds of data (text document, e-mail, etc.) with a more variable relevance (less relevant photos to find?) in a future work.

The last point in the discussion opens perspectives to improve these results using other features in addition to meta-data. Thus, the "multimedia" information contained in an image is not fully exploited: for example, the photo content (image processing) is one way to explore to find the undetected relevant pictures and to refine the extraction by separating images by events.

VI. ACKNOWLEDGEMENTS

This work was funded by the SuMACC project supported by the French National Research Agency (ANR) under contract ANR-10-CORD-007.

REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998, pp. 37–45.

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," 1998.

[3] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998, pp. 28–36.

[4] S. Golder and B. Huberman, "The structure of collaborative tagging systems," in *CoRR*, 2005.

[5] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, 2007, pp. 103–110.

[6] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[7] T. Nakamura, T. Nagai, and N. Iwahashi, "Bag of multimodal lda models for concept formation," in *International Conference on Robotics and Automation (ICRA)*, Shanghai, China, 2011, pp. 6233–6238.

[8] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *ACM International Conference on World Wide Web*, 2008, pp. 111–120.

[9] S. Papadopoulos, R. Troncy, V. Mezaris, B. Huet, and I. Kompatsiaris, "Social Event Detection at MediaEval 2011: Challenges, Dataset and Evaluation," in *MediaEval 2011 Workshop*, Pisa, Italy, 2011.

[10] R. Kubota Ando and L. Lee, "Iterative residual rescaling: An analysis and generalization of lsi," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, USA, 2001.

[11] A. Marcus, A. Sergeyev, V. Rajlich, and J. Maletic, "An information retrieval approach to concept location in source code," in *Working Conference on Reverse Engineering (WCRE)*, Delft, the Netherlands, 2004, pp. 214–223.

[12] T. Hofmann, "Probabilistic latent semantic indexing," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, USA, 1999, pp. 50–57.

[13] A. Stein and H. Schmid, "Etiquetage morphologique de textes français avec un arbre de décisions," *Traitement automatique des langues*, vol. 36, no. 1-2, pp. 23–35, 1995.

[14] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Conference on Uncertainty in Artificial Intelligence*, Banff, Canada, 2004, pp. 487–494.

[15] V. Vapnik, "Pattern recognition using generalized portrait method," *Automation and Remote Control*, vol. 24, pp. 774–780, 1963.

[16] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *ACM Workshop on Computational learning theory*, Pittsburgh, USA, 1992, pp. 144–152.

[17] K. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," *Artificial Neural Networks ICANN*, pp. 999–1004, 1997.

[18] P. Bartlett and J. Shawe-Taylor, "Generalization performance of support vector machines and other pattern classifiers," *Advances in Kernel Methods Support Vector Learning*, pp. 43–54, 1999.

[19] T. Joachims, "Transductive inference for text classification using support vector machines," in *International Conference on Machine Learning (ICML)*, Bled, Slovenia, 1999, pp. 200–209.

[20] S. Kiritchenko and S. Matwin, "Email classification with co-training," in *Conference of the Centre for Advanced Studies on Collaborative research (CASCON)*, Toronto, Canada, 2001, p. 8.

[21] G. Salton, "Automatic text processing-the analysis, transformation and retrieval of information by computer," *Addison-Wesley, Reading, MA*, 1989.

[22] L. Xueliang, R. Troncy, and B. Huet, "Eurecom@mediaeval 2011 social event detection task," in *Working Notes Proceedings of the MediaEval 2011 Workshop*, Pisa, Italy, 2011.