inventions

MDPI

Article

A Distance-Dependent Chinese Restaurant Process Based Method for Event Detection on Social Media

Georgios Palaiokrassas [1,*], Athanasios Voulodimos [2], Antonios Litke [1], Athanasios Papaoikonomou [1] and Theodora Varvarigou [1]

[1] School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece; litke@mail.ntua.gr (A.L.); tpap@mail.ntua.gr (A.P.); dora@telecom.ntua.gr (T.V.)

[2] Department of Informatics and Computer Engineering, University of West Attica, 12243 Athens, Greece; avoulod@uniwa.gr

* Correspondence: geopal@mail.ntua.gr

Received: 2 November 2018; Accepted: 4 December 2018; Published: 7 December 2018

Abstract: In this paper, we propose a method for event detection on social media, which aims at clustering media items into groups of events based on their textural information as well as available metadata. Our approach is based on distance-dependent Chinese Restaurant Process (ddCRP), a clustering approach resembling Dirichlet process algorithm. Furthermore, we scrutinize the effectiveness of a series of pre-processing steps in improving the detection performance. We experimentally evaluated our method using the Social Event Detection (SED) dataset of MediaEval 2013 benchmarking workshop, which pertains to the discovery of social events and their grouping in event-specific clusters. The obtained results indicate that the proposed method attains very good performance rates compared to existing approaches.

Keywords: event detection; Dirichlet process clustering; distance-dependent Chinese restaurant process; Bayesian non-parametrics

1. Introduction

In recent years, there has been a great research interest in techniques for event detection on data retrieved from Social Networks, focusing mostly on Twitter platform. An event could be defined as an arbitrary classification of a space-time region and might include actively participating agents, passive factors, products, and a location in space/time [1]. Atefeh et al. [2] conducted a study based on three major categories: (i) the type of event being detected, distinguishing the event into specified and unspecified; (ii) the detection task (new event detection and retrospective event detection); and (iii) the event detection method (supervised, unsupervised, and hybrid). Another classification approach of the different detection methods is presented in the survey [3] that focuses on the common traits the methods share (i.e., using probabilistic topic modeling, identifying interesting properties in a tweet's keywords/terms and using incremental clustering).

This paper is organized as follows: Section 2 presents a description of work related to event detection from social media data and a review of Bayesian nonparametric models and ddCRP. Detailed description of the clustering algorithm is given in Section 3, while, in Section 4, we describe the experiments and evaluation of the event detection. Finally, Section 5 is devoted to discussion and conclusions of our work.

2. Related Work

Benson et al. [4] developed a graphical model for record extraction from social media streams, using a MIRA-based binary classifier to predict whether a message mentions an event. The output of their model is a set of canonical records, the values of which are consistent with aligned messaging. During testing their method, they used a fixed number of records (events) based on the training data. Doulamis et al. [5] addressed the dynamic nature of tweet messages, constructing fuzzy time signals, modeling this clustering task as a multi-assignment graph partitioning problem. Their method exploited pairwise similarities, Riemannian distance metric between word signatures, to compensate tweet submission and correlated words based on fuzzy time feature series, focusing mostly on unstructured datasets.

Petrovic et al. [6] addressed the problem of detecting new events from a stream of Twitter posts using an algorithm based on locality-sensitive hashing (LSH) [7], performing constant time and space estimations of the closest document. This way the computational cost did not increase, as the number of clusters increases. They recognized that the high degree of lexical variation in documents makes it very difficult to detect stories that talk about the same event using different words and, in their later work [8], they combined paraphrases with locality-sensitive hashing. While the First Story Detection (FSD) performance improved, the gain is much smaller when this technique is applied to newswire data, likely due to lexical mismatch between the knowledge bases and social media. In this direction, Moral et al. [9] used word embeddings, trying to enhance the representation of social media posts to increase the effectiveness of LSH based FSD. More specifically, they exploited Word2Vec [10] and expanded tweets with semantically related paraphrases identified via automatically mined word embeddings.

The distance-dependent Chinese Restaurant Process (ddCRP) is introduced in [11] as a flexible class of distributions over partitions for data clustering. It is based on a Bayesian clustering method for non-exchangeable sequence of observations, when the number of clusters is unknown. ddCRP clusters data in a biased way: each data point is more likely to be clustered with other data that are near it in an external sense. Recent work has extended the original ddCRP model for use in different applications. Ghosh et al. [12] examined it in a spatial setting with the goal of natural image segmentation, while Socher et al. [13] combined ddCRP with spectral dimensionality reduction. These approaches, however, compute the distances between data only based on the original data. Furthermore, these approaches are not directly applicable when additional information can be used for the similarity computation [14].

A similarity-based Chinese Restaurant Process was proposed by Papaoikonomou et al. [15] addressing the problem of event detection from social media data and evaluated their method in SED MediaEval dataset, which is also used for this work's evaluation and is presented in detail in Section 4.1. When the number of events in the target set is not known in advance, this non-parametric algorithm, namely the Dirichlet Process clustering, allows the dynamic creation of clusters based on the data. A ddCRP variation was proposed by Li et al. [16] using side information in a Bayesian nonparametric model for data clustering. They evaluated their method using normalized mutual information and F1-measure, taking advantage of the strong correlation of side information (such as citation, authors, and keywords for a documents dataset) with the main data features. An object proposal generation via sampling form a ddCRP posterior on image segmentation is proposed in [17].

3. The Proposed Method

In this section, we describe the proposed event detection approach, which aims at analyzing media items to categorize them into meaningful collections, focusing both on the textual data and the metadata of the media posts targeting mostly Twitter. However, this service is applicable to any dataset which conforms to Twitter's data format and contains the following:

- The username of the author (metadata)
- The timestamp of the creation date (metadata)
- The actual textual content (of short length)

The username information is important given that we expect certain authors, who express themselves through a length-constrained message, to comment on a small number of topics each time they generate content (often just one). The temporal dimension is also crucial since media items that refer to the same real event tend to lie close in time. Finally, the actual grouping of the media posts is also controlled by groups of words that co-occur frequently and textual patterns.

3.1. Message Similarity and Clustering

A central concept in the development of a clustering algorithm is similarity among data points: similar objects should be grouped together, whereas dissimilar ones should be assigned to different collections. In the domain of text clustering, similarity is usually measured by the degree that words tend to co-occur, i.e., the higher the rate of common words the higher their similarity. On top of that perspective, and in our quest for additional signals, we choose to also mine the metadata of the media posts produced by a social network user, thus resulting in a similarity function that aggregates information from different parts of the social network messages. In particular, our similarity function, operating on two social media messages m_u and m_v, is defined as:

$$sim(m_u,\ m_v) = a_1 \cdot l_{author} + a_2 \cdot f_t(t_u, t_v) + a_3 \cdot f_w(m_u, m_v) \tag{1}$$

where a_i are coefficients to be learned during the training process. I_{auth} is an indicator function that is equal to 1 if two messages are published by the same author and 0 otherwise. $f_t(t_u,\ t_v)$ is a function that estimates the temporal distance between the messages. It is a monotonically decreasing function which takes as input the timestamps of their creation dates (t_u and t_v) and returns a similarity value based on their temporal spread. We choose an exponential decay function of the form:

$$f_t(t_u, t_v) = e^{-|t_u - t_v|/q} \tag{2}$$

for some window parameter q, which takes its maximum value (1) when $t_u \approx t_v$, and decreases towards zero as the absolute difference $\mid t_u - t_v \mid$ increases.

In general, the decay function mediates how temporal distances between customers affect the resulting distribution over partitions. We assume that the decay function f_t is non-increasing, takes non-negative finite values, and satisfies $f(\infty) = 0$. Following the proposed decay functions of Blei et al. [11], we consider several types of decay as for our temporal distance model, all of which satisfy these non-restrictive assumptions. In terms of temporal distance, the window decay $f(d) = 1$ $(d < a)$ only considers customers that are at most distance a from the current customer. The exponential decay $f(d) = \exp(-d/a)$ decays the probability of linking to an earlier customer exponentially with the distance to the current customer. The logistic decay $f(d) = \exp(-d + a)/(1 + \exp(-d + a))$ is a smooth version of the window decay. After examining their impact to the overall performance of our proposed approach, we consider the exponential decay function to be the most suitable for method.

Finally, $f_w(m_u,\ m_v)$ is a function that mines directly the textual content of the two messages and outputs a higher similarity value in case of a large number of common terms. To construct an effective similarity function targeting textual content from social networks, we need to take into account the special characteristics of the user generated content in social media. Such a case is that of hashtags, which are identifiers inserted by the author of a media post to indicate the topic(s) on which she expresses her opinion. Typical hashtag terms are written in the form #<topic-identifier>, e.g., #Oscars. We consider hashtags as highly significant for the task of event detection, and thus we highlight their role compared to the other terms in the media post, by assigning a larger weight.

3.2. Methodology

The proposed method is closely related to the Dirichlet Process (DP) mixture models, a family of flexible clustering algorithms for high dimensional data analysis. More specifically, we use a variant of a DP mixture called distance-dependent Chinese Restaurant algorithm that was introduced by Blei et al. [11]. In general, a Dirichlet process [18] is an infinite mixture model which expresses a distribution over probability measures. A DP mixture model can be viewed as a Chinese Restaurant

Process (CRP) which is fancifully described by a sequence of customers joining a Chinese restaurant with an infinite number of tables. Every time a new customer enters the restaurant, he may choose to join an occupied table with probability analogous to the number of persons already sitting there, or to choose a new one with probability proportional to a predefined value, called the concentration parameter. The analogy in a CRP mixture is apparent: customers represent the data points which belong to the same cluster if they "sit" at the same table.

In accordance with the notation of [11], we define:

- z_i is the table assignment of the ith customer (*id* of the table that customer i chooses).
- K is the total number of occupied tables.
- n_k is the number of customers sitting at table k, with $k = 1, \ldots, K$.
- α is the concentration parameter.

The conditional probability of the table assignment for the ith customer, given the assignments of the customers before him, is computed through:

$$p(z_i = k \mid z_1, z_2, \ldots, z_{i-1}, a) \propto \begin{cases} n_k, k \leq K \\ a, k = K + 1 \end{cases} \tag{3}$$

An interesting property of the CRP mixture model is that the number of the finally occupied tables is random, and thus the number of clusters is determined by the data. This is a desirable feature given that it is usually difficult to estimate the right number of groups in real-world data.

In the case of the traditional CRP mixture analogy, customers are exchangeable, i.e., the probability of a particular table configuration is the same even if the order of the customers is permuted. This property might seem reasonable for certain applications but it is not appropriate when the order of data points matters. Such an example is the social event detection task that we consider here along with its dependence on the temporal dimension, since we expect that media items that refer to an event will tend to group with other items that lie close in time. A variant of the CRP mixture model that enforces such an "non-exchangeability" constraint is the "distance-dependent Chinese Restaurant Process" introduced by Blei et al. in [11]. The difference in this approach is that the distances (e.g., based on time) among the customers are the key factors that lead to the seating assignment. In other words, while the traditional CRP connects customers to tables, the distance-dependent variant connects customers to other customers and the allocation of customers to tables is a by-product of this process.

Let us define the following:

- c_i is the assignment of the ith customer (the identifier of the customer with whom the ith customer chooses to sit).
- d_{ij} is the "distance" between customers i and j, and D is the matrix of distances for all customer pairs.
- f is a decay function.

The conditional probability for the ith customer assignment is now:

$$f(c_i = j \mid D, a) \propto \begin{cases} f(d_{ij}), j \neq i \\ a, j = i \end{cases} \tag{4}$$

Figure 1 presents a sample application of the distance-dependent process, which is used to cluster social media objects. Given Equation (3), each "customer" will choose to either "sit close" to another customer (directed link) or "sit alone" (self-loop). Customers 2, 3 and 5 belong to the first category, whereas Customers 1, 4, and 6 belong to the second.

The final allocation of customers on tables depends on the pair-wise relationships among all the customers. Two customers that are reachable through a sequence of intermediate customer assignments will be finally assigned to the same table. In this way, Customers 1, 2 and 3 will sit at the first table (cluster), Customer 4 will form a cluster on her own and Customers 5 and 6 will be assigned to the third cluster.

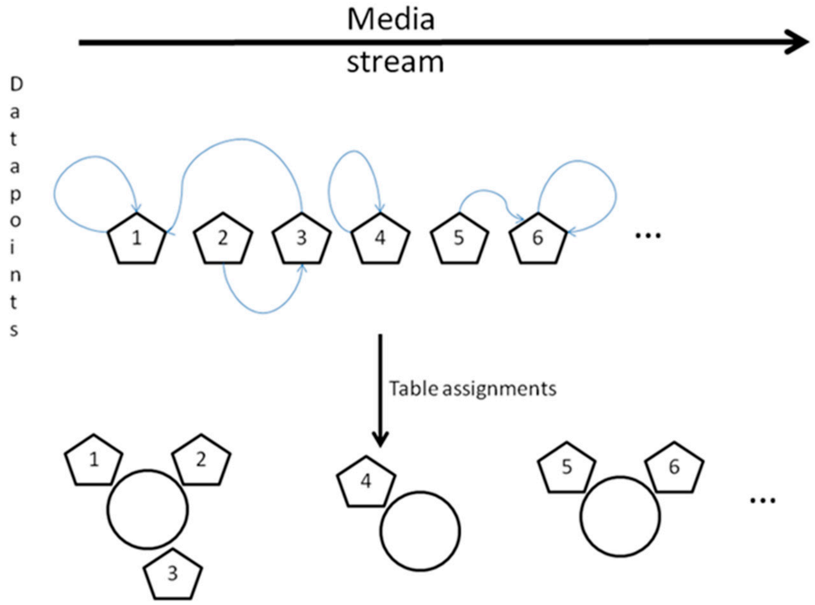

Figure 1. Sample application of the distance-dependent Chinese Restaurant Process on a media stream of six media items. The pair-wise distances between the media objects determine the table assignment (clusters). Some customers choose to sit close to another customer, while some choose to sit alone. For example, customer 2 chooses to sit close to customer 3 (direct link depicted as blue arrow), while customer 1 chooses to sit alone.

Our proposed method applies the distance-dependent CRP algorithm using the similarity function of the previous sub-section ($sim(m_u, m_v)$) as an inverted distance function. To train the model, we resort to Markov Chain Monte Carlo (MCMC) sampling to approximate the posterior distribution of the data. More specifically, given the set of the hyper-parameters $\eta = \{D_{sim}, G_0\}$, where D_{sim} is the distance matrix of all data points computed through the function, α the concentration parameter and G_0 the base measure, we perform Gibbs sampling by iteratively drawing from the conditional distribution of each latent variable (c_i) given the other latent variables (c_{-i}) and observations (x):

$$p\left(c_i^{(new)} \middle| c_{-i}, x, \eta\right) \propto p\left(c_i^{(new)} \middle| D_{sim}, a\right) p\left(x \middle| \left(c_i^{(new)} \cup c_{-i}\right), G_0\right)$$

where $z(c)$ are the table assignments that follow from the customer assignments and $z({c_i}^{(new)} \cup c_{-i})$ expresses the new candidate partition.

The first term on the right side of this equation can be computed from the distance dependent prior in the previous one. The second term is the likelihood of the observation under the new candidate partition. To compute this term, we consider how removing a customer link and replacing it with another affects the table assignment, as depicted in Figure 2. Having the partition $z(c_{-i})$, in which a table may have been split, and the new candidate partition, as described above, we consider three cases based on the differences between the two partitions:

1. The new customer assignment c_i links to itself (self-loop), which does not change the likelihood since no tables are joined together.
2. The new customer assignment c_i links to another customer, who is already at its table under $z(c_{-i})$. There is no change in the partition, since no tables are joined together.
3. The new customer assignment c_i links to another customer and tables k and l are joined

Gibbs sampler needs thus to compute terms that correspond to changes in the partition and for our distance-dependent Chinese Restaurant Process approach the Gibbs sampler is:

$$p\left(c_i^{(new)} \middle| c_{-i}, x, \eta\right) \propto \begin{cases} (d_{ij}) \dfrac{p\left(x_{z^k(c_{-i}) \cup z^l(c_{-i})} \middle| G_0\right)}{p\left(x_{z^k(c_{-i})} \middle| G_0\right) p\left(x_{z^l(c_{-i})} \middle| G_0\right)}, if\ c_i^{(new)} = j\ joins\ tables\ k\ and\ l \\ f(d_{ij}), if\ c_i^{(new)} = j\ not\ joins\ tables \\ a, if\ if\ c_i^{(new)} = i \end{cases}$$

where ${x_z}^k{}_{(c-i)}$ is the set of customers that are assigned to table k excluding the current customer i.

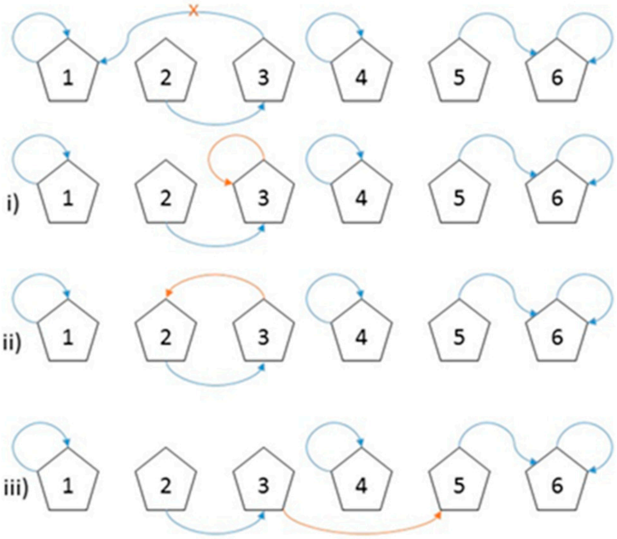

Figure 2. An example of the Gibbs sampler, where blue arrows represent the links after running the first part of the algorithm and yellow ones the links after resampling. A table can be split when we remove an existing link from a customer to another. After resampling: (i) the customer links to itself; (ii) the customer links to another but two tables are not merged; or (iii) the link obtained for Customer 3 merges two tables.

Scaling of the algorithm when the input increases is an important issue to take into consideration. Even for medium-sized datasets, it is impractical to compute the distance for all pairs of tweets. To tackle this problem, we introduce a concept that reduces the required computational resources of our approach. For every table, we summarize the event as a list of the most important information such as representative words, hashtags and statistics about temporal information. We use this "event summarization" to compare each customer to only the event summarization of a table, reducing the total number of comparisons from n (total number of customers) to K (total number of tables).

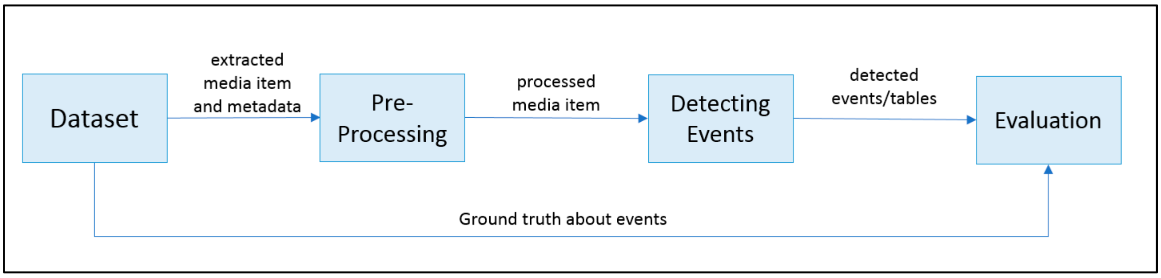

Figure 3. Flow and steps of the proposed approach for event detection.

4. Implementation Details and Evaluation

4.1. Dataset

Our approach is applicable to any dataset which conforms to Twitter's data format. For the evaluation, we used the dataset of Social Event Detection (SED) task of MediaEval 2013 [19,20]. That task requires the participants to discover social events and organize the related items in event-specific clusters. The task is a supervised clustering task [21,22], where a set of training events is provided. Numerous participants propose algorithms, grouping media items and producing a complete clustering of the dataset according to events, facing also the challenge of discovering the actual number of target events, since it is not given.

The dataset consists of about 437,370 pictures gathered using the Flickr API [23]. They were uploaded between 2006 and 2012 and were assigned to about 21,169 events, annotated by people, referring to sport events, protests, marches, debates, expositions, festivals, and concerts and are separated into two parts: the training set (70% of the dataset) and testing set (30% of the dataset). For our evaluation, we used the textual metadata of the pictures, which included information such as the title, description, time, tags, etc., focusing on the comparison of the available metadata of the dataset's multimedia items to assign each item to an event. All these metadata are included in the publicly available files of the dataset in XML and csv format, which are published in [20,24] and follow the schema described in Table 1. As it is a real-world dataset, there are some features, such as time-stamps and uploader information, that are available for every picture, but there are also features (e.g., geographic information) that are available for only a subset of the images.

Table 1. Detailed description of the Dataset of Social Event Detection (SED) task and more particularly of the metadata of each media item to be assigned to an event.

Feature	Description
flickr_picture_id	unique ID of the image
url	URL of image in Flickr
username	username of uploader
datetaken	timestamp of capturing the image
dateupload	timestamp of uploading the image
title	title of the image
description	text describing the image
latitude, longitude	image's location
tags	keywords assigned to the image
event_id	ground truth event for this image

4.2. Implementation Details

Java programming language was used for the implementation, application and evaluation of our clustering-based event detection service and a flow demonstrating the different steps of the method is presented in Figure 3 Various solutions were examined to improve the performance of our method, i.e., integration with relational databases [25] to avoid repetitive computations and memory overflow and multithreading using dedicated libraries to reduce overall computational time, mostly for the part of the algorithm which handles the computation of distances among the approximately 131.200 different media items of the test set of this dataset and has a complexity of O(n^2). Dedicated Java libraries were used for the processing the dataset and the available files in csv and XML format, so we can extract the information of interest [26,27]. As mentioned before, this approach was evaluated using the annotated SED dataset, but could be applied to any dataset that conforms to Twitter's data format, using for example Twitter API [28] and related libraries [29] to retrieve tweets.

4.3. Text Pre-Processing

The most common pre-processing techniques applied by event detection techniques on the Twitter data stream are the following: POS tagging, NER, resolving temporal expressions, slang-word conversion, tweet filtering based on specific criteria (i.e., discarding retweets and/or non-English language tweets) and removing stop words, URLs and username mentions from tweets [3]. To this direction, we have extensively investigated how different pre-processing techniques affect the performance of our algorithm. Some of the pre-processing applied that had a marked impact on the overall effectiveness are:

- Hashtags: A type of metadata used by social media users. It typically consists of the character "#" followed by a string and is usually indicative of the topic the user refers to. We use them in the similarity function among different tweets (customers in our method based on ddCRP) and tested replacing them with "#" character or omitting them.
- RT (retweet): A re-post of an original tweet is called a retweet. It usually should be clustered together with the original one. Additionally, metadata fields indicate the number of times a specific post was retweeted.
- URL: We performed tests by keeping a link directing to an external source, replacing the whole link with the word "url" and removing it completely.
- Stemming and lemmatization: Our goal was to capture the "base form" of a word. For example house instead of houses, house's etc.
- Stop-word removal: These are very common words such as "the", "at", etc.

4.4. Evaluation Metrics

For the evaluation, we compared our clustering results to the ground truth clustering assignments that has been created by human annotators, by using the following metrics [30]:

- Normalized Mutual Information (NMI):

$$NMI(\Omega, C) = \frac{I(\Omega, C)}{[H(\Omega) + H(C)]/2}$$

  where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ is the set of clusters, $C = \{c_1, c_2, \ldots, c_k\}$ is the set of classes, I(Ω, *C*) is mutual information between Ω and *C*, and *H(Ω)* and *H(C)* are the entropies of Ω and *C*, respectively.
- F1-score, calculated from Precision and Recall with the formula

$$F1 - Measure = 2 * \frac{Precision * Recall}{Precision + \ Recall}$$

  where *Precision* and *Recall* are defined as follows: A true positive (*TP*) decision assigns two similar items to the same cluster, a true negative (*TN*) decision assigns two dissimilar items to different clusters. There are two types of errors we can commit. A False Positive (*FP*) decision assigns two dissimilar item to the same cluster. A False negative (*FN*) decision assigns two similar items to different clusters.

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

4.5. Experimental Setup

To evaluate the performance of our service, we developed and tested different approaches and variations of ddCRP, some of which are presented below:

1. Chinese Restaurant Process (CRP): We used the training set to determine the optimal value of the concentration parameter *a*.

2. Distance-dependent Chinese Restaurant Process based only on time (ddCRP only time): During this approach, in the similarity function, we used only the information about time among all the available metadata in each multimedia item. This way we could see the difference in the clustering results when we used more of the available metadata in the following algorithms.
3. Distance-dependent Chinese Restaurant Process sequential (ddCRP sequential): We compared each customer only to customers previously assigned to tables instead of comparing to all customers in the dataset. This approach requires about half the time for training and evaluating.
4. Distance-dependent Chinese Restaurant Process (ddCRP): We used the similarity function described in previous sections. The training set was used to reach the optimal values for the parameters of our algorithm such as the concentration parameter and the coefficients in the similarity function.

4.6. Results

We compared the results of our method with the rest of the submissions of SED 2013. In total, there are 11 submissions following different clustering approaches. Table 2 reports the performance of all the proposed algorithms in terms of F1-score and NMI. In Table 3, we observe that our method is among the top proposed approaches for event detection, achieving high performance results, outrunning most of the submissions.

Table 2. Performance of approach over the SED 2013 dataset (for both metrics, the best value is 1 and the worst value is 0).

Algorithm 1	NMI	F1-Measure
CRP	0.56	0.1
ddCRP (only time)	0.890	0.594
ddCRP sequential	0.819	0.1412
ddCRP	0.907	0.655

The results of our evaluation are presented in Table 3. We can see that our method ddCRP, which is highlighted, outperforms the other three approaches for both measures.

Table 3. SED 2013 results.

Approach	F1	NMI
Semantic Structuring of Complementary Information [31]	0.142	
Similarity-based Chinese Restaurant Process [15]	0.236	0.664
Data-Driven approach [32]	0.570	0.873
Same Event Model—Based [33]	0.704	0.910
Unsupervised Clustering [34]	0.780	0.940
Clustering based on BM25, Sphinx and cosine similarity [35]	0.81	0.954
Quality Threshold clustering variant [36]	0.878	0.965
PhotoTOC [37]	0.883	0.973
Our proposed ddCRP based method	**0.907**	**0.6546**
Watershed-based and kernel methods [38]	0.932	0.984
Sparse multi-modal feature selection and incremental density-based clustering [39]	0.946	0.985

5. Conclusions and Future Work

In this paper, we present a method for social media event detection. We developed a variation of the distance-dependent Chinese Restaurant Process grouping media items into event clusters based on their textual data (i.e., actual text) and the available metadata (e.g., title, description, tags, and location). Additionally, we showed how pre-processing techniques can be used to enhance the performance of our event detection. For future work, we plan to apply our event detection service to more datasets and use other methods to capture the similarity of textual context. We also intend to examine the use of word embeddings, such as word2vec [10,40], to mitigate the problem of lexical variation among tweets, which are related to the same event and use different but synonymous words, and further improve the performance of our proposed method.

Author Contributions: Conceptualization, all authors; software, G.P.; validation, G.P., A.V., and A.L.; and writing, G.P., A.V., and A.L.

Funding: The research leading to these results has received funding from the European Commission under the H2020 Programme's project BigClouT (grant agreement No. 723139).

Conflicts of Interest: The authors declare no conflict of interest. Additionally, the funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Event Ontology. Available online: http://motools.sourceforge.net/event/event.html (accessed on 30 November 2018).
2. Atefeh, F.; Khreich, W. A survey of techniques for event detection in twitter. *Comput Intell.* **2015**, *31*, 132–164. [CrossRef]
3. Hasan, M.; Orgun, M.A.; Schwitter, R. A survey on real-time event detection from the twitter data stream. *J. Inf. Sci.* **2017**. [CrossRef]
4. Benson, E.; Haghighi, A.; Barzilay, R. Event discovery in social media feeds. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, Oregon, 19–24 June 2011; pp. 389–398.
5. Doulamis, N.D.; Doulamis, A.D.; Kokkinos, P.; Varvarigos, E.M. Event detection in twitter microblogging. *IEEE Trans. Cybern.* **2016**, *46*, 2810–2824. [CrossRef] [PubMed]
6. Petrović, S.; Osborne, M.; Lavrenko, V. Streaming first story detection with application to twitter. In Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA, USA, 2–4 June 2010; pp. 181–189.
7. Indyk, P.; Motwani, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, Dallas, TX, USA, 24–26 May 1998; pp. 604–613.
8. Petrović, S.; Osborne, M.; Lavrenko, V. Using paraphrases for improving first story detection in news and Twitter. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, QC, Canada, 3–8 June 2012; pp. 338–346.
9. Moran, S.; McCreadie, R.; Macdonald, C.; Ounis, I. Enhancing first story detection using word embeddings. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016; pp. 821–824.
10. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
11. Blei, D.M.; Frazier, P.I. Distance dependent Chinese restaurant processes. *J. Mach. Learn. Res.* **2011**, *12*, 2461–2488.
12. Ghosh, S.; Ungureanu, A.B.; Sudderth, E.B.; Blei, D.M. Spatial distance dependent Chinese restaurant processes for image segmentation. In *NIPS'11 Proceedings of the 24th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Granada, Spain, 2011; pp. 1476–1484.
13. Socher, R.; Maas, A.; Manning, C. Spectral chinese restaurant processes: Nonparametric clustering based on similarities. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 698–706.
14. Li, C.; Phung, D.; Rana, S.; Venkatesh, S. Exploiting side information in distance dependent chinese restaurant processes for data clustering. In Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15–19 July 2013; pp. 1–6.
15. Papaoikonomou, A.; Tserpes, K.; Kardara, M.; Varvarigou, T. A similarity-based chinese restaurant process for social event detection. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013.

16. Li, C.; Rana, S.; Phung, D.; Venkatesh, S. Data clustering using side information dependent Chinese restaurant processes. *Knowl. Inf. Syst.* **2016**, *47*, 463–488. [CrossRef]
17. Lauri, M.; Frintrop, S. Object proposal generation applying the distance dependent Chinese restaurant process. In Proceedings of the Scandinavian Conference on Image Analysis, Tromsø, Norway, 12–14 June 2017; pp. 260–272.
18. Pitman, J. *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002*; Springer: Berlin, Germany, 2006.
19. Reuter, T.; Papadopoulos, S.; Petkos, G.; Mezaris, V.; Kompatsiaris, Y.; Cimiano, P.; de Vries, C.; Geva, S. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 18–19 October 2013.
20. Social Event Detection. Available online: http://www.multimediaeval.org/mediaeval2013/sed2013/ (accessed on 1 November 2018).
21. Petkos, G.; Papadopoulos, S.; Kompatsiaris, Y. Social event detection using multimodal clustering and integrating supervisory signals. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Hong Kong, China, 5–8 June 2012; p. 23.
22. Reuter, T.; Cimiano, P. Event-based classification of social media streams. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Hong Kong, China, 5–8 June 2012; p. 22.
23. Flickr Services. Available online: https://www.flickr.com/services/api/ (accessed on 14 November 2018).
24. Reuter, T.; Papadopoulos, S.; Mezaris, V.; Cimiano, P. ReSEED: Social event dEtection dataset. In Proceedings of the 5th ACM Multimedia Systems Conference, Singapore, 19–21 March 2014; pp. 35–40.
25. PostgreSQL: The World's Most Advanced Open Source Database. Available online: https://www.postgresql.org/ (accessed on 14 November 2018).
26. javax.xml.bind (Java Platform SE 7). Available online: https://docs.oracle.com/javase/7/docs/api/javax/xml/bind/package-summary.html (accessed on 29 November 2018).
27. Opencsv. Available online: http://opencsv.sourceforge.net/ (accessed on 29 November 2018).
28. Twitter API. Available online: https://developer.twitter.com/en/docs.html (accessed on 1 November 2018).
29. Tweepy. Available online: http://www.tweepy.org/ (accessed on 29 November 2018).
30. Evaluation of Clustering. Available online: https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html (accessed on 14 November 2018).
31. Gupta, K.G.I.; Chandramouli, K. VIT@ MediaEval 2013 Social Event Detection Task: Semantic Structuring of Complementary Information for Clustering Events. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013.
32. Rafailidis, D.; Semertzidis, T.; Lazaridis, M.; Strintzis, M.G.; Daras, P. A Data-Driven Approach for Social Event Detection. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013.
33. CERTH @ MediaEval 2013 Social Event Detection Task | Request PDF. *ResearchGate*. Available online: https://www.researchgate.net/publication/283248185_CERTH_MediaEval_2013_social_event_detection_task (accessed on 30 November 2018).
34. Zeppelzauer, M.; Zaharieva, M.; del Fabro, M. Unsupervised Clustering of Social Events. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013.
35. Sutanto, T.; Nayak, R. Admrg@ MediaEval 2013 social event detection. In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 18–19 October 2013.
36. Wistuba, M.; Schmidt-Thieme, L. Supervised Clustering of Social Media Streams. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013.
37. Vizuete, D.M.; Nieto, X.G. Upc at mediaeval 2013 social event detection task. In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 18–19 October 2013.
38. Nguyen, T.-V.; Dao, M.-S.; Mattivi, R.; Sansone, E.; de Natale, F.G.; Boato, G. Event Clustering and Classification from Social Media: Watershed-based and Kernel Methods. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013.

39. Samangooei, S.; Hare, J.; Dupplaw, D.; Niranjan, M.; Gibbins, N.; Lewis, P.H.; Davies, J.; Jain, N.; Preston, J. Social event detection via sparse multi-modal feature selection and incremental density based clustering. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013.
40. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Lake Tahoe, Nevada, 2013; pp. 3111–3119.