www.ijiarec.com

ISSN:2348-2079

Volume-6 Issue-3

International Journal of Intellectual Advancements and Research in Engineering Computations

Discovering based on text mining techniques

P.Vidhya[1], S.Anitha[2]

[1]M.Phil, Research Scholar (Full-Time), [2]Assistant Professor.
PG and Research Department of Computer Science Vivekanandha College of Arts and Sciences for Women (Autonomous), Elayampalayam.

ABSTRACT

Identifying the authorship either of an anonymous or a doubtful document constitutes a cornerstone for automatic forensic applications. Moreover, it is a challenging task for both humans and computers. Clustering documents according to the linguistic style of the authors who wrote them has been a task little studied by the research community. In order to address this problem, PAN Evaluation Framework has become the first effort to promote the development of the author clustering. This article proposes a graph-based method, specifically β-compact clustering, for discovering the groups of documents written by the same author. The β-compact algorithm is based on the analysis of the similarity between documents and they belong to the same group as long as the similarity between them exceeds the threshold β and it is the maximum similarity with respect to other documents. In our proposal we evaluated different linguistic features and similarity measures presented in previous works of authorship analysis task. The training dataset was used to determine the best value of β parameter for each language. The result of the experiments was encouraging.

Keywords: Datamining, Integrity, Authentication, Security

INTRODUCTION

The documents clustering task, by author´s linguistic style, is of vital importance in forensic applications. A practical example would correspond to the identification in a computer of all the groups of documents written in this one and that each group of document has been written by a single author. Considering that this computer belongs to a public site. In the evaluation framework the task is described as follows: "Given a collection of (up to 50) short documents (paragraphs extracted from larger documents), identify authorship links and groups of documents written by the same author. All documents are single-authored, in the same language, and belong to the same genre. However, the topic or text-length of documents may vary. The number of distinct authors whose documents are included in the collection is not given.

One of the most used strategies for documents representation in Text Mining (TM) applications, corresponds to the classic Bag of Words and this will be the proposal used in our work. In different Authorship Analysis applications, complex methods involving several algorithms have been used in order to obtain the best results. In document clustering applications and other Artificial Intelligence (AI) tasks, ensembles of algorithms have also been employed. Despite this, the work presented by is relevant, and they use a simple clustering algorithm and achieve encouraging results.

As a summary, in the last edition of authors clustering task, 6 papers were presented and in general, the data of the documents collection set contained a high percentage of clusters composed

Author for correspondence:
M.Phil, Research Scholar (Full-Time), PG and Research Department of Computer Science Vivekanandha College of Arts and Sciences for Women (Autonomous), Elayampalayam.

of a single document, unlike what can be seen in the collection of this year released for training, where we observed several documents clusters with more than one document, although there are still few documents per group.

With our work, we want to propose and evaluate a clustering algorithm that we have used in topic document clustering tasks in our research center, and its purpose is to group objects with the condition that for each object of the group, at least there is an object with which the similarity between them is greater than a threshold of similarity and it´s the maximum similarity with an object of the collection.

It is important to emphasize aspects of the description of the author clustering problem, such as: short texts no longer than a paragraph; the texts corresponding to the same author are of the same genre but not necessarily the same topic or homogeneous length. In addition, as part of the task, we need to obtain a ranking of similarities between objects in the same clusters. Taking these details into consideration, in the next section we propose and describe our method considering binary linguistic features and a clustering algorithm based on compact clusters.

RELATED WORK

k-NN Classification over Semantically Secure Encrypted Relational Data

Recently, the cloud computing paradigm is revolutionizing the organizations' way of operating their data particularly in the way they store, access and process data. As an emerging computing paradigm, cloud computing attracts many organizations to consider seriously regarding cloud potential in terms of its cost-efficiency, flexibility, and offload of administrative overhead. Most often, organizations delegate their computational operations in addition to their data to the cloud. Despite tremendous advantages that the cloud offers, privacy and security issues in the cloud are preventing companies to utilize those advantages. When data are highly sensitive, the data need to be encrypted before outsourcing to the cloud. However, when data are encrypted, irrespective of the underlying encryption scheme, performing any data mining tasks becomes very challenging without ever decrypting the data. There are other privacy concerns, demonstrated by the following example.

Suppose an insurance company outsourced its encrypted customers database and relevant data mining tasks to a cloud. When an agent from the company wants to determine the risk level of a potential new customer, the agent can use a classification method to determine the risk level of the customer. First, the agent needs to generate a data record q for the customer containing certain personal information of the customer, e.g., credit score, age, marital status, etc. Then this record can be sent to the cloud, and the cloud will compute the class label for q. Nevertheless, since q contains sensitive information, to protect the customer's privacy, q should be encrypted before sending it to the cloud.

The above example shows that data mining over encrypted data (denoted by DMED) on a cloud also needs to protect a user's record when the record is a part of a data mining process. Moreover, cloud can also derive useful and sensitive information about the actual data items by observing the data access patterns even if the data are encrypted [2, 3]. Therefore, the privacy/security requirements of the DMED problem on a cloud are threefold: [1] confidentiality of the encrypted data, (2) confidentiality of a user's query record, and [3] hiding data access patterns.

Existing work on privacy-preserving data mining (PPDM) (either perturbation or secure multi-party computation (SMC) based approach) cannot solve the DMED problem. Perturbed data do not possess semantic security, so data perturbation techniques cannot be used to encrypt highly sensitive data. Also the perturbed data do not produce very accurate data mining results. Secure multi-party computation based approach assumes data are distributed and not encrypted at each participating party. In addition, many intermediate computations are performed based on non-encrypted data. As a result, in this paper, we proposed novel methods to effectively solve the DMED problem assuming that the encrypted data are outsourced to a cloud. Specifically, we focus on the classification problem since it is one of the most common data mining tasks. Because each classification technique has their own advantage, to be concrete, this paper concentrates on executing the k- nearest neighbor classification method over

encrypted data in the cloud computing environment.

LITERATURE REVIEW

Anna Vartapetiance, Lee Gillam: "A Big Increase in Known Unknowns: from Author Veri-fication to Author Clustering."

Previous PAN workshops have afforded evaluation of our approaches to author verification/identification based on stopword cooccurrence patterns. Problems have tended to involve comparing one document to a small set of documents (n<=5) of known authorship. This paper discusses the adaptation of one of our approaches to a PAN 2016 problem of author clustering, which involves generating clusters within larger sets of documents (n<=100) for an unknown number of distinct authors, where each set is in English, Dutch or Greek. We describe our previous approaches as the background to the approach taken to this task and briefly overview the results that were achieved, which are not expected to be particularly remarkable due to substantial limitations on our time around the task.

Efstathios Stamatatos, Michael Tschuggnall, Ben Verhoeven, Walter Daelemans, Günther Specht, Benno Stein, Martin Potthast:"Clustering by Authorship Within and Across Docu-ments."

The vast majority of previous studies in authorship attribution assume the existence of documents (or parts of documents) labeled by authorship to be used as training instances in either closed-set or open-set attribution. However, in several applications it is not easy or even possible to find such labeled data and it is necessary to build unsupervised attribution models that are able to estimate similarities/differences in personal style of authors. The shared tasks on author clustering and author diarization at PAN 2016 focus on such unsupervised authorship attribution problems. The former deals with single-author documents and aims at grouping documents by authorship and establishing authorship links between documents. The latter considers multi-author documents and attempts to segment a document into authorial components, a task strongly associated with intrinsic plagiarism detection. This paper presents an overview of the two tasks including evaluation datasets, measures, results, as well as a survey of a total of 10 submissions (8 for author clustering and 2 for author diarization).

Yunita Sari, Mark Stevenson: "Exploring Word Embeddings and Character N-Grams for Au-thor Clustering."

We presented our system for PAN 2016 Author Clustering task. Our software used simple character n-grams to represent the document collection. We then ran K-Means clustering optimized using the Silhouette Coefficient. Our system yields competitive results and required only a short runtime. Character n-grams can capture a wide range of information, making them effective for authorship attribution. We also present a comparison study of two different features: character n-grams and word embeddings.

Douglas Bagnall:"Authorship Clustering using Multi-headed Recurrent Neural Networks."

A recurrent neural network that has been trained to separately model the language of several documents by unknown authors is used to measure similarity between the documents. It is able to find clues of common authorship even when the documents are very short and about disparate topics. While it is easy to make statistically significant predictions regarding authorship, it is difficult to group documents into definite clusters with high accuracy

METHODOLOGY

β-compact algorithm

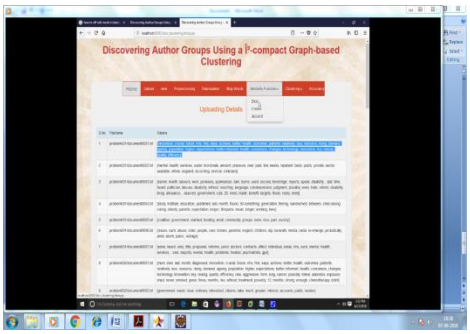
We propose the use of β-compact algorithm for authorship clustering task, because it´s based on clustering objects with a similarity between them which is greater than a threshold of similarity previously adjusted with a training document collection. Both in the training and test stage, collections of documents are received and the final purpose is to obtain groups of documents, where all the documents of a group belong to the same author. The algorithm proposed, to obtain the groups, requires the representation of each of the documents, a comparison function that allows to evaluate the similarity between a pair of documents

and a threshold β to decide when two documents must belong to the same cluster. For documents representation, we used the classic Bag of Word, and with the train-ing dataset we tried different types of features.

We experimented with 3 similarity functions to analyze the similarity between documents. We used the Dice, Jaccard and Cosine functions, using only binary features, that is, we did not compute the frequency of each of the features, only their appearance in the document.

The idea of considering only binary features is due to the short extension of the documents, up to one paragraph. The β-compact clustering algorithm is described in the next pseudo program code. First we need to define the concept "Graph of Maximum β similarity:

It´s an oriented graph in which the vertices are the objects and exist an arista between two vertices Oi and Oj if Oj is β-similar with Oi and Oj is the most similar of all the rest of objects. It is important to note that, due to the nature of the β-compact algorithm, two documents can belong to the same group, although the similarity between them not necessarily exceeds the defined β, because the only condition is that each one of them has a similarity greater than β with some document of their group. This characteristic may lead to non-necessarily spherical clusters



CONCLUSION

In this paper, we must emphasize that one of the essential aspects in our work is the features identification for the documents representation, and in this we could try other ideas presented in the literature. The algorithm usually obtains small groups. We propose to evaluate different features weighing strategies and other comparison functions proposed in the literature. Using the results achieved in this work, take into consideration comparisons with ensemble clustering algorithms.

REFERENCES

[1]. A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, M. Ayyash, Internet of things: a survey on enabling technologies, protocols, and applications, IEEE Commun. Surv. Tutor. 17(4), 2015, 2347–2376.

[2]. Y. Sun, H. Song, A.J. Jara, R. Bie, Internet of things and big data analytics for smart and connected communities, IEEE Access 4, 2016, 766–773.

[3]. R. Zuech, T.M. Khoshgoftaar, R. Wald, Intrusion detection and big heterogeneous data: a survey, J. Big Data 2(1), 2015, 1–41.

[4]. X. Qiu, Y. Qiu, G. Feng, P. Li, A sparse fuzzy c-means algorithm based on sparse clustering framework, Neurocomputing 157, 2015, 290–295.

[5]. L. Meng, A.H. Tan, D. Xu, Semi-supervised heterogeneous fusion for multimedia data co-clustering, IEEE Trans. Knowl. Data Eng. 26(9), 2014, 2293–2306.

[6]. Y. Chen, L. Wang, M. Dong, Non-negative matrix factorization for semisupervised heterogeneous data coclustering, IEEE Trans. Knowl. Data Eng. 22(10), 2010, 1459–1474.

[7]. R. Bekkerman, M. Sahami, E. Learned-Miller, Combinatorial Markov random fields, in: Proceedings of the Seventeenth European Conference on Machine Learning (ECML), 2012, 30–41.
[8]. Z. Brakerski, C. Gentry, V. Vaikuntanathan, (Leveled) fully homomorphic encryption without bootstrapping, in: Proceedings of the Innovations in Theoretical Computer Science Conference, 2012, 309–325.
[9]. S. Chen, F. Wang, C. Zhang., Simultaneous heterogeneous data clustering based on higher order relationships, in: Proceedings of the IEEE International Conference on Data Mining, 2007, 387–392.
[10]. Q. Zhang, L.T. Yang, Z. Chen, F. Xia, A high-order possibilistic-means algorithm for clustering incomplete multimedia data, IEEE Syst. J. PP 99, 2015, 1–10.
[11]. B.A. Pimentel, R.M.C.R. de Souza, Multivariate fuzzy c-means algorithms with weighting, Neurocomputing 174, 2016, 946–965.
[12]. T.C. Havens, J.C. Bezdek, C. Leckie, L.O. Hall, M. Palaniswami, Fuzzy c-means algorithms for very large data, IEEE Trans. Fuzzy Syst. 20(6), 2012, 1130–1146.
[13]. N.R. Pal, K. Pal, J.M. Keller, J.C. Bezdek, A possibilistic fuzzy c-means clustering algorithm, IEEE Trans. Fuzzy Syst. 13(4), 2005, 517–530.
[14]. X. Y.Wang, J. Bu, A fast and robust image segmentation using FCM with spatial information, Digit. Signal Process. 20(4), 2010, 1173–1182.
[15]. Q. Zhang, L. T. Yang, and Z. Chen, ”Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning,” IEEE Transactions on Computers, 65(5), 2016, 1351-1362.
[16]. N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, ”A Possibilistic Fuzzy c-Means Clustering Algorithm,”IEEE Transactions on Fuzzy Systems, 13(4), 2005, 517-530.
[17]. M. Yang and C. Lai, ”A Robust Automatic Merging Possibilistic Clustering Method,” IEEE Transactions on Fuzzy Systems, 19(1), 2011, 26-41.