Using Multitier Ensemble Classifiers for Organizing Multimedia Big Data - An Visualization

K.Pavithra

Msc (PhD)., Assistant Professor, Department of Computer Applications, Dhanalakshmi Srinivasan College of Arts and Science for Women, Perambalur

Abstract- **This article initiates and considers large iterative multitier ensemble (LIME) classifiers specifically tailored for big data. These classifiers are very large, but are quite easy to generate and use. They can be so large that it makes sense to use them only for big data. They are generating repeatedly as a result of numerous iterations in applying ensemble meta classifiers. Here, we carry out an ample investigation of the concert of LIME classifiers for a trouble concerning security of big data. Our examines evaluate LIME classifiers with different base classifiers and standard common ensemble meta classifiers. The outcome obtained exhibit that LIME classifiers can significantly enlarge the precision of classifications. In this paper, the semantic link network model is second-hand for categorize multimedia resources. An entire model for generating the union relation among multimedia resources using semantic link network model is anticipated. A genuine data set counting 100 thousand images with public tags from Flickr is used in our trials. Two appraisal methods, including clustering and retrieval, are performed, which illustrate the planned method can compute the semantic relatedness linking Flickr images accurately and robustly.**

Index Terms- Big data, multimedia resources, semantic link network, multimedia resources organization.

I. INTRODUCTION

Big Data has become ubiquitous and crucial for numerous application domains, thereby leading to significable challenges from the point of view of data management perspective. It has become particularly important in view of the rapid growth of Cloud services. The development and expansion of the Cloud creates new opportunities for the users and requires further research to address novel tasks and requirements. It is important not only to invent new methods tackling these tasks, but also to investigate ways of adapting previous techniques well known in related areas such as grid computing . Security has been one of the major issues required for the use of Big Data. Lately, Big data is an promising paradigm functional to datasets whose volume is beyond the capacity of commonly used software tools to capture, handle, and method of the data within a tolerable beyond time. A choice of technologies are being talk about to support the usage of big data such as massively parallel processing databases, scalable storage systems cloud computing platforms and MapReduce.

The major aspire of this manuscript is to develop LIME classifiers as a general technique that may be useful for the analysis of Big Data in various application domains. If a dataset is not large enough, then the LIME classifier will revert to using only a base classifier just a small part of the whole system and will not improve the quality of the classification. We carry out a systematic experimental investigation of the performance of LIME classifiers for a problem concerning security of Big Data.

The characteristics of social tags are as follows.

Ontology free. The ontology based labeling denotes ontology and then let users label the multimedia resources using the semantic markups in the ontology. Social tagging requires all the users in the social net-work label the multimedia resources with their own keywords and share with others. Different from ontology based annotation.

User oriented. The users can annotate images with their favorite tags. The tags of multimedia resources are determined by users' cognitive ability. To the multimedia resources, users may give different tags. Each multimedia resource may be with one tag at least, and each tag may appear in many different multimedia resources.

Semantic loss. Immaterial social tags commonly emerge, and users naturally will not tag all semantic substances in the image, which is called semantic loss. Polysemy, synonyms, and ambiguity are some drawbacks of social tagging.

The relatedness between tags and surrounding texts are implemented in the Semantic Link Network model.

The major contributions of this paper are summarized as follow.

(1) A whole model for generating the association relation between multimedia resources using Semantic Link Network model is proposed. The definitions, modules, and mechanisms of the Semantic Link Network are used in the proposed method. The tags and the surrounding texts of multimedia resources are used to measure their semantic association. The hierarchical semantic of multimedia resources are denoted by their annotated tags and surrounding texts

(2) A real data set including 100 thousand images with social tags from Flickr is used in our experiments. Two appraisal methods with clustering and retrieval are performed, which illustrates the proposed method can measure the semantic relatedness between Flickr images accurately and robustly.

(3) The relatedness measures between concepts are extended to the level of multimedia. Since the asso-ciation relation is the basic mechanism of brain.

II. RELATED WORK

Major security challenges facing the analysis of Big Data and the Cloud have been considered. For general background information on the methodology, systems and applications of the Cloud Computing, the associated Quality of Service Management, the design of Cloud Workow Systems, and tradeoffs between the computation and data storage. A lightweight malware detection system for detecting, analyzing and predicting malware propagating via SMS and MMS messages on mobile devices is proposed. It deploys agents in the form of hidden contacts on the device to capture messages sent from malicious applications.

A scalable scheme for network-level behavioral clustering of HTTP-based malware is presented in [29]. It groups newly collected malware samples into malware family clusters. This aim of creating clusters is to facilitate the generation of high quality network signatures for detecting botnet communications at the network perimeter.

Frequency of the appearance of opcode sequences is used in [30] to prepare data for data mining algorithms trained to detect malware.

A concept of genetic footprint is proposed in [31]. It can be used to detect malicious processes at run time. The genetic footprint consists of selected parameters maintained inside the PCB of a kernel for each running process. It denotes the semantics and behavior of an executing process.

A graph-based method to detect unknown malware is presented in [32]. It uses the function call graph of an executable, which includes the functions and the call relations between them.

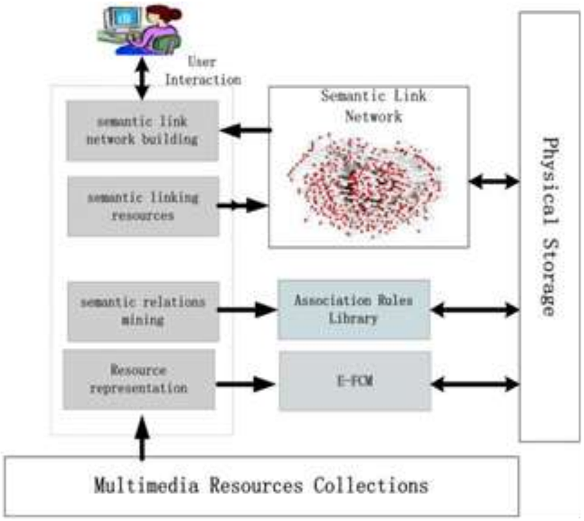
Advances in Semantic Web have made ontology another use-ful source for describing multimedia semantics. The ontology builds a formal and explicit representation of semantic hierarchies for the concepts and their relationships in video events, and allows reasoning to derive implicit knowledge. In this section, the related work of the proposed model is given.

The Semantic Web is an evolving development of the World Wide Web, in which the meanings of information on the web is denied; therefore, it is possible for machines to process it. The basic idea of Semantic Web is to use ontological concepts and vocabularies to accurately describe contents in a machine readable way.

The Semantic Link Network (SLN) was projected as a semantic data model for categorize a variety of Web resources by expanding the Web's hyperlink to a semantic link. SLN is a aimed at network consisting of semantic nodes and semantic links. A semantic node be able to be a notion, an example of concept, a schema of data set, a URL, any form of resources, or even an SLN.

III.THE SEMANTIC LINK NETWORK BASED MODEL

The tags and surrounding texts of multimedia resources are used to represent the semantic content. The relatedness between tags and surrounding texts are implemented in the Semantic Link Network model.


User Interaction
semantic link network building
semantic linking resources
Semantic Link Network
semantic relations mining
Association Rules Library
Resource representation
E-FCM
Physical Storage
Multimedia Resources Collections

A. THE BASIC MECHANISMS OF THE PROPOSED MODEL

SLN can be formalized into a loosely coupled semantic model for managing various resources. As a data model, the pro-posed model consists of the following parts, as shown in Fig. 1

(1) Resources Representation Mechanism: Element Fuzzy Cognitive Map (E-FCM) [19] is used to represent multi-media resources with social tags since it does not only reserve resources' keywords but also the relations among them.

(2) Resources Storage Mechanism: Database/XML is used to store E-FCM since it is easy to define the mark-up elements.

(3) SLN Generation Mechanism: Based on E-FCM and the association rules, ALN can be generated by machine automat-ically.

(4) Application Mechanism: SLN can be used for Web intel-ligence activities, Web knowledge discovery and publishing, etc. For example, when a user browses multimedia, other resources with semantic links to it can be recommended to the user.

B.THE BASIC HEURISTICS

Based on common sense and our observations on real data, five heuristics that serve as the base of the proposed computation model are given as follow.

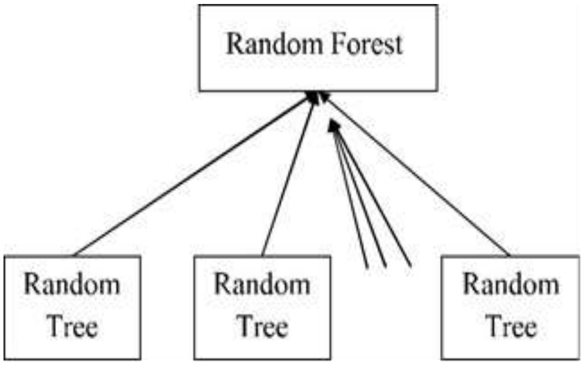
Heuristic 1. Usually each tag of a multimedia resource appears only one time. Different from writing sentences, users usually annotate a multimedia resource with different tags. For example, the possibility of using tags ``apple apple apple" for an image is very low. Therefore, in this paper, we do not employ any weighting scheme for tags such as tf-idf [26].

Heuristic 2. The order of the tags may reflect the correla-tion against the annotated multimedia resource Different tag reflects the different aspect of a multime-dia resource. According to Heuristic 1, the weight of a tag against the image cannot be obtained. Fortunately, the order of the tags can be get since user may provide tags one by one.

Heuristic 3. The number of tags of a multimedia resource may not relevant to the annotation correctness Different users may give different tags about the same multimedia resource. For example, users may give tags such as ``apple iPhone" or ``iPhone4 mobile phone" for a same image about iPhone. It is hardly to say which tag is better for annotation though the latter annotation has three tags.

IV. LIME CLASSIFIERS

A number of methods for creating ensemble classifiers are well known in artificial intelligence and data mining. Conventional ensemble meta classifiers produce their set of base classifiers given an signal, or an example, or a template of simply one base classifier as an input limit. After the creation phase, they use the whole company of the base classifiers to practice data, gather their outputs and merge them to arrange the decision. For example, Random Forest automatically generates a collection of Random Trees and uses them as shown. Numerous other known all together meta classifiers role in a similar way, using extra base classifiers and different practices for engender and combining them.



To launch the method a exclusive has to initialize a four-tier LIME classifier by specifying which ensemble meta classifier will function at the fourth tier. Then the trendy provide a stricture to the fourth tier ensemble meta classifier representative, which third tier ensemble meta classifier is to be worn as a part of the normal generation development of the fourth tier ensemble meta classifier. Following that, the designer specifies the second tier ensemble meta classifier system to be used by the third tier ensemble meta classifier, and the base classifier grip by the second tier ensemble meta classifier.

In this manuscript we used diverse ensemble meta classifiers and base classifiers implemented in the Waikato Environment for Knowledge Analysis (WEKA). All choices chosen by a designer for a LIME classifier can be specified in the WEKA SimpleCLI command line. Then the entire system is produces mechanically by the SimpleCLI, using the entrenched iterative and recursive ability of Java programming.

V. GENERATING THE SEMANTIC LINK

The addition model for generating the semantic link among multimedia funds is proposed. Based on the above heuristics, the public tags offered by users are used in our calculation model. Overall, the proposed computation representation is divided into three steps.

Step 1 - Tag relatedness computation. In this step, stand on heuristic 1, all of the tag pairs connecting two multimedia resources are computed.

Step 2 - Many dissimilar techniques of semantic relatedness measures flanked by concepts have been proposed, which can be divided into two aspects taxonomy-based methods and web-based methods. Taxonomy-based means use in order of theory and hierarchical taxonomy, such as WordNet, to compute semantic relatedness. On the contrary, web-based methods use the network as a live and active corpus instead of hierarchical classification.

Step 3 -Overall, the page count up of each tag should be issued. Then the co-occurrence based measure is used to figure the semantic relatedness between tags. The reasons for using page counts based events are as follow.

(1) Appropriate calculation complexity. while the relatedness between each tag join up of two multimedia resources should be figured, the proposed method must be with low difficulty. Lately, web seeks engines such as Google provide API used for users to index the page counts of every query. The web search engine presents an suitable interface for the proposed computation model.

B. Explicit semantics. The tag known by users may not be a right concept in taxonomy. For instance, users may give a tag ``Bling Bling" for a multimedia supply. The word ``Bling" cannot be indexed in many classification such as WorldNet. The proposed method uses web search engine as an unlock intermediate. The explicit semantics of the recently emerge concepts can be get by web easily.

VI. DISCUSSION

Our effort shows that large four-tier LIME classifiers are fairly easy to use and can be practical to improve classifications, if varied ensemble meta classifiers are combined at dissimilar tiers. It is an interesting query for potential research to investigate LIME classifiers for other huge datasets.

Random Forest outperformed additional base classifiers for the malware dataset, and Decorate enhanced its outcomes better than other ensemble Meta classifiers do. The finest outcome of AUC 0.998 was obtained by the four-tier LIME classifier where MultiBoost was worn at the fourth tier, adorn was used at the third tier and Bagging was functional at the second tier.

The presentation of collection meta classifiers considered in this manuscript depends on several arithmetical input parameters. In all testing we used them with the similar default values of these parameters in classify to have a regular equivalent link of outcomes transversely all of these ensemble meta classifiers.

Content-based image retrieval (CBIR) is the function of computer vision methods to the image retrieval difficulty, that is, the problem of pointed for digital images in huge databases. ``Content-based" resources that the search examines the contents of the image to a certain extent than the metadata such as keywords, tags, or descriptions associated with the image. The saying ``content" in this context force submit to colors, shapes, textures, or any other information that can be resulting from the image itself. CBIR is popular because most web-based picture search engines rely simply on metadata and this creates a lot of garbage in the consequences. Also having humans physically enter keywords for images in a large database can be incompetent, expensive and may not imprison every keyword that describes the image.

VII. CONCLUSION

We introduced and investigated four-tier LIME classifiers originating as a contribution to the general approach considered by many authors. We obtain new results evaluating performance of such large four-tier LIME classifiers. These new results show, in particular, that Random Forest performed best in this setting, and that novel four-tier LIME classifiers can be used to achieve further improvement of the classification outcomes. They are effective if diverse ensemble meta classifiers are combined at different tiers of the LIME classifier. They have made significant improvements to the performance of base classifiers and standard ensemble meta classifiers.

Modern research proves that multimedia resources ``in the wild" are budding at a staggering rate. The speedy enlarge number of multimedia resources has brought an vital need to develop clever methods to organize and procedure them. In this document, the Semantic Link Network model is worn for categorize multimedia resources. Semantic Link Network (SLN) is designed to create associated relations among various resources (e.g., Web pages or documents in digital library) aiming at extending the loosely connected network of no semantics. Two data mining tasks including clustering and searching are performed by the proposed framework, which shows the effectiveness and robust of the proposed framework.

REFERENCES

[1] L. Batten, J. Abawajy, and R. Dose, ``Prevention of information harvesting in a cloud services environment," in Proc. 1st Int. Conf. Cloud Comput. Services Science, 2011, pp. 66 72.

[2] W. Dou, Q. Chen, and J. Chen, ``A confidence-based filtering method for DDoS attack defense in cloud environment,'' Future Generat. Comput. Syst., vol. 29, no. 7, pp. 1838 1850, 2013.

[3] S. Yu, S. Guo, and I. Stojmenovic, ``Can we beat legitimate cyber behavior mimicking attacks from botnets?" in Proc. 31st Annu. IEEE Int. Conf. Comput. Commun., Mar. 2012, pp. 2851 2855.

[4] Chen et al., ``Big data challenge: A data management perspective," Frontiers Comput. Sci., vol. 7, pp. 157 164, Apr. 2013.

[5] R. Islam and W. Zhou, ``Email classification using multi-tier classifica-tion algorithms," in Proc. 7th IEEE/ACIS Int. Conf. Comput. Inf. Sci., May 2008.

[6] Z. Xu, X. Luo, and L. Wang, ``Incremental building association link network," Comput. Syst. Sci. Eng., vol. 26, no. 3, pp. 153 162, 2011.

[7] H. Zhuge, ``Interactive semantics,'' Artif. Intell., vol. 174, no. 4,190 204, 2010. H. Zhuge, The Knowledge Grid: Toward Cyber-Physical Society, 2nd ed. Singapore: World Scientific, 2012.

[8] H. Zhuge, X. Chen, X. Sun, and E. Yao, ``HRing: A structured P2P overlay based on harmonic series," IEEE Trans. Parallel Distrib. Syst., vol. 19, no. 2, pp. 145 158, Feb. 2008.

[9] (2012, Jun. 20). VX Heavens Virus Collection [Online]. Available: http://vx.netlux.org/vl.php

[10] P. Pons, ``Object prefetching using semantic links," ACM Sigmis Database, vol. 37, no. 1, pp. 97 109, 2006.