Negative-based Sampling for Multimedia Retrieval

Hsin-Yu Ha, Shu-Ching Chen
School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
{*hha001,chens*}*@cs.fiu.edu*

Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33124, USA
shyu@miami.edu

***Abstract*—Nowadays, in such a high-tech living lifestyle, profusion of multimedia data are produced and propagated around the world. To identify meaningful semantic concepts from the large amount of data, one of the major challenges is called the data imbalance problem. Data imbalance occurs when the number of positive instances (i.e., instances which contain the target concept) is greatly less than the number of negative instances (i.e., instances which do not contain the target concept). In other words, the ratio between positive and negative instances is extremely low. Rebalancing the dataset is usually proposed to resolve the problem by sampling or data pruning. In this paper, we propose a sampling method which consists of three stages, namely selecting features to identify the negative instances, producing negative ranking scores, and performing sampling. The method is compared with some other existing methods on the TRECVID dataset and is demonstrated to have better performance.**

***Keywords*-Sampling; FC-MST; Feature Selection; Multimedia**

I. Introduction

Efficiently manage multimedia big data becomes an important topic in both academic community [1]–[4] and industry environment [5]–[7], since the amount of multimedia data increases exponentially every day. YouTube official website announced 300 hours of videos are uploaded to YouTube website every minute [8]. Another well-known social media platform Flickr also announced that the average number of photos shared on Flickr is 1 million per day [9]. To cope with the enormous amount of multimedia data, many challenges need to be conquered, including integration among multiple modalities [10]–[16], high dimensions of the features [17]–[24], and data imbalance problem [25]–[30], etc.

Among all these challenges, data imbalance problem in particular has drawn attention from researchers in both data mining and machine learning areas, specifically to improve the results for classification and semantic concept detection. In a general classification process, training data is given to train the classifier in understanding the data characteristics for both positive and negative classes. At this stage, the sampling size and the data distribution usually greatly influence the performance. However, data imbalance problem commonly takes place, where the number of positive training instances is excessively less than the number of negative training instances. Because of the insufficient number of positive instances, the classifier is not able to obtain enough information and it will incline to classify instances into negative instances.

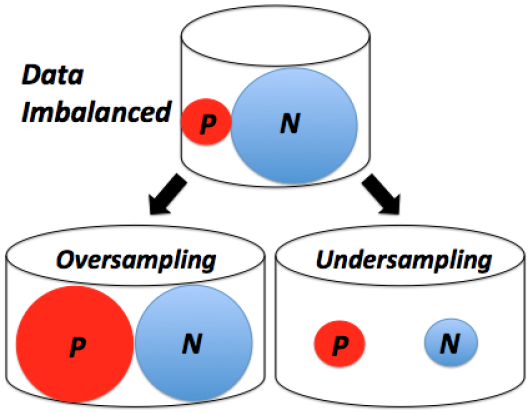

Figure 1: Sampling Method

To resolve the data imbalance problem, two types of sampling methodologies are usually utilized, i.e., oversampling and undersampling. As shown in Figure 1, oversampling methods [31]–[33] try to balance the data by adding more duplications of positive instances. The major drawback is that the computation time for training the classification model will greatly increase due to a larger training data set. On the other hand, undersampling methods [34]–[36] filter out the extra negative instances so that a more balanced data set can better represent both positive and negative classes. The potential weakness of the undersampling method is that the representative negative instances might be pruned and the remaining negative instances are not able to provide enough information for negative concepts. To sum up, a good sampling method should be able to reduce the computation time while having adequate information for both positive and negative concepts.

In this paper, we propose a new thinking of performing sampling based on the negative ranking scores. Instead of pruning out the instances, which are unlikely to be identified as positive instances, from the training data, our proposed method can be formed by three components: feature selection for negative instances, producing negative ranking

scores based on the selected features, sampling the data by selecting only instances with higher negative ranking scores.

The rest of the paper can be organized as follows: In section 2, a detailed description of the proposed method will be given. Experimental setup, the evaluation criteria, and the comparative experimental results are depicted in section 3. Finally, section 4 concludes the paper by summarizing the contribution and pointing out the discovery from this paper.

II. Proposed Framework

In Figure 2, the proposed framework can be separated into three major components which are all designed mainly considering the negative class. First, a negative-based feature selection method is proposed to identify significant features for negative classes. It is inspired and motivated by an existing work named FC-MST [17]. Originally, the work was proposed to choose an optimal feature subset by removing the redundant and irrelevant features, thus utilizing the selected features can accurately detect the semantic concepts. In other words, the features are selected to correctly identify positive instances. In this paper, the focus is changed toward the correlation between features and negative concepts. Second, given the selected features, the negative ranking score can be calculated per instance, where the higher the score is, the higher possibility it has to be classified as negative instances. Thus different levels of negative concepts can be assigned to each instance. Third, the negative ranking scores generated from the second component are leveraged to perform the sampling process. In this proposed sampling method, only the representative negative instances are chosen and integrated with the positive instances to train the classification model.

A. Negative-based Feature Selection Method

FC-MST (Feature Correlation Maximum Spanning Tree) was proposed in [17] to select optimal feature subsets in enhancing semantic concept detection results. It contains a three-stage process which aims to remove the redundant and irrelevant features toward positive concepts. Because it has shown its ability in finding the better feature subset to detect positive concepts, the proposed method is derived and moved the shift toward detecting negative concepts. The original training data set is given as shown in Table I. Later, it is discretized using MDL (Minimum Description Length) [37] based on only the label of target concept negative. According to the discretization results as shown in Table II, features with only one interval are removed at this stage.

Multiple Correspondence Analysis (MCA) is taking place after the discretization process. It is adopted because its effectiveness has been shown in various research areas, including video semantic concept detection [25], [27], [38], [39], feature selection [18], discretization [40], etc. It projects all feature intervals per feature onto a two-dimensional space formed by two major principal components, PC_1 and PC_2, where Pos represents the positive concept and Neg represents the negative concept. Following the similar process in [17], $Correlation\ \alpha_j^i$ (e.g., α_3^2) and $Reliability\ \beta_j^i$ (e.g., β_3^2) are considered when generating the feature correlation. However, the two factors are generated using the cosine value and the absolute distance between the feature interval and the negative concept instead of the positive one. As shown in Equation (1), each feature correlation toward the negative concept FC_i (i represents the feature index) is calculated by summing up the $Correlation\ \alpha$ and $Reliability\ \beta$ per interval with the corresponding weights, e.g., $omega_1$ and $omega_2$, and then divided by the number of feature intervals j. The detailed description can be found in [17], [18].

$$FC_i = \frac{\sum_{n=1}^{j}(\omega_1 \alpha_n^i + \omega_2(1-\beta_n^i))}{j} \tag{1}$$

The negative feature correlations are used as the edge weight in forming FC-MST proposed in [17]. Two feature pruning conditions are set to eliminate the irrelevant and redundant features, which are listed as follows.

- If $FC_{ij} < FC_{iN}$ and $FC_{ij} < FC_{jN}$, then Edge $\overline{ij}$ will be removed from the formed FC-MST. i and j represents the index of the feature and N represents the negative concept.
- After FC-MST is separated into different connected components, choose only the representative feature from each component. In other words, the feature with the maximum feature correlation toward the negative concept will be selected into the final feature subsets.

Table I: Example of the Original Features

	Feature 1	Feature 2	...	Feature M	Target Concept Positive	Target Concept Negative
Inst. 1	-0.49	1.08	...	-0.45	1	0
Inst. 2	-0.56	-0.85	...	-1.32	0	1
Inst. 3	-0.61	-2.21	...	1.33	1	0
Inst. 4	-0.48	-0.97	...	-1.01	0	1
Inst. 5	-0.53	-1.54	...	0.97	1	0

Table II: Example of the Discretized Features

	Feature 1	Feature 2	...	Feature M	Target Concept Negative
Inst. 1	F_1^1	F_3^2	...	F_2^M	0
Inst. 2	F_1^1	F_2^2	...	F_1^M	1
Inst. 3	F_1^1	F_1^2	...	F_3^M	0
Inst. 4	F_1^1	F_3^2	...	F_1^M	1
Inst. 5	F_1^1	F_1^2	...	F_3^M	0

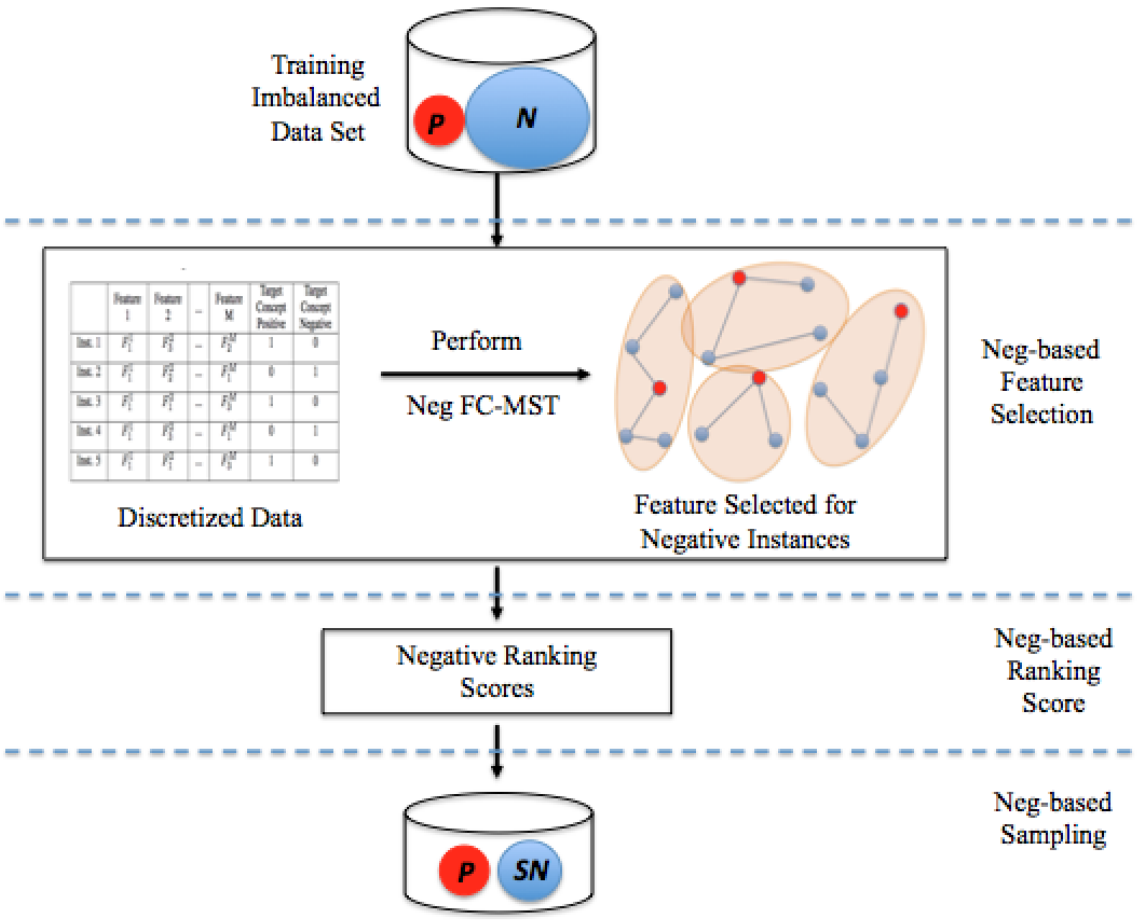

Figure 2: The proposed negative-based sampling method for multimedia retrieval

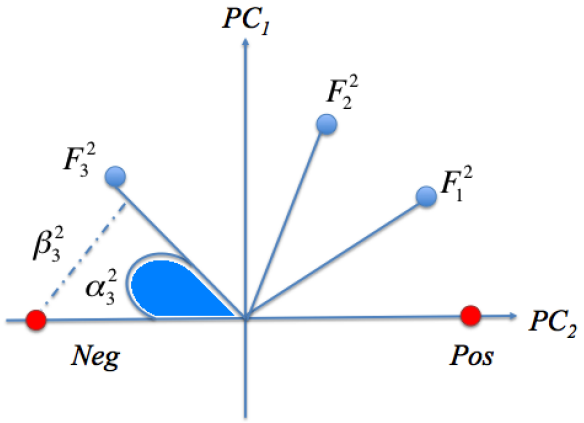

Figure 3: Using MCA to obtain the correlations between the feature intervals and the negative concept

B. Negative-based Ranking Scores

In section II-A, the process of selecting a feature subset to identify negative instances is finalized. Therefore, based on the aforementioned feature subset, the transaction weight learnt from MCA is introduced here to generate a negative ranking score per training instance. In Equation (2), each feature interval will be assigned a weight $Weight_j^i$, where i represents feature's index and j represents feature interval's index. It calculates the cosine value based on the angle between a feature interval and a negative concept as previously shown in Figure 3.

$$Weight_j^i = \cos(\theta_j^i) \tag{2}$$

Once the weight for each feature interval is obtained, the transaction weight can be calculated by looping through all the features within one instance and accumulating the corresponding feature interval's weight as shown in Equation (3). In this equation, k represents the instance index and M represents the number of features.

$$TransactionWeight_k = \sum_{i=1}^{M} Weight_j^i \tag{3}$$

C. Negative-based Sampling Method

As shown in Figure 4, given two lists of ranking scores in descending order for both positive and negative concepts, we propose to sample the instances with higher ranking scores from both sides. It is naturally to think that the sample subset containing well-represented instances for both positive and negative concepts can enhance the classification result, especially when dealing with an imbalanced data set.

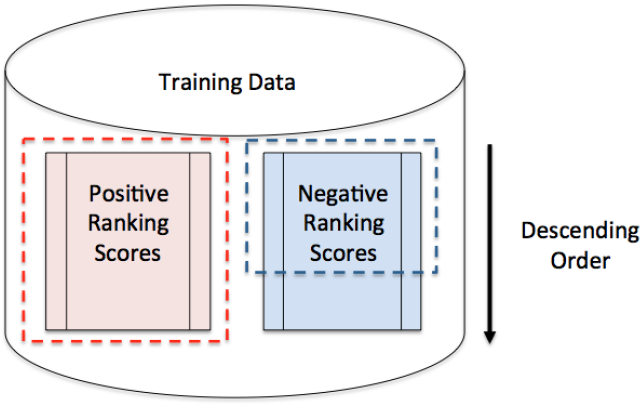

Figure 4: Negative-based sampling method

III. Experiments

A. Dataset

TREC Video Retrieval Evaluation (TRECVID) is an annual worldwide competition [41], which is held by National Institute of Standards and Technology (NIST). It aims to improve the content-based analysis on a large collection of digital videos. In TRECVID 2011 semantic indexing (SIN) task, the dataset, which is composed of 200 hours videos with durations between 10 seconds and 3.5 minutes, is used to validate the proposed framework. To utilize the data set, one or multiple keyframes are extracted from each video shot and each keyframe represents one instance in the classification model. Each semantic concept, i.e., outdoor and person, has the label information because of the collaboration efforts coordinated by Georges Quenot and team [42]. The data set is selected in this paper because of two reasons. The first reason is that the size of the data set is sufficient. The second reason is that it contains severe data imbalance problem. The statistic information of the data set is listed in Table III and Table IV. P/N ratio is calculated using the number of positive instances divided by the number of negative instances.

B. Experimental Setup

Mean Average Precision (MAP) is selected to evaluate the proposed framework, in comparison to some other related work. It is a well-known evaluation method, specifically when it is used to validate the classification ranking results for positive concept only. The higher the MAP value is, it means that it has higher possibility to correctly detect positive concept from the Top N listed instances.

Table III: TRECVID 2011 Semantic Indexing IACC.1.B Statistic Information

Semantic Indexing Task Data Set	IACC.1.B
TRECVID Year	2011
Number of Concepts	8
Number of Training Data Instances	262911
Number of Testing Data Instances	137327
Average P / N Ratio	0.0829

Table IV: Semantic Concept and its ratio between the number of positive instances and negative instances

No.	Concept	P / N Ratio
1	Adult	4.13%
2	Face	5.93%
3	Indoor	4.38%
4	Male_Person	5.03%
5	Outdoor	13.82%
6	Overlaid_Text	3.33%
7	Person	26.96%
8	Vegetation	3.73%

As listed in Table IV, 8 concepts are selected to validate the performance of the proposed framework and other related works. The semantic concept, such as "Yasser Arafat" with the least number of positive instances, was not selected because its extremely low P/N ratio, e.g., 0.000015 makes it hardly affected by any sampling methods.

The experiment is designed to prove the assumption that when coping with imbalanced data, it is important to sample the data by choosing the representative instances for positive and negative concepts. Therefore, the proposed framework is compared with three different results: Original, RS, MCA-based. Original is the original training data without any sampling process. RS stands for Random Sampling and it means that negative instances were randomly filtered from the training data. Lastly, MCA-based stands for MCA-based Data Pruning Method, it has published in [25] and the method focuses on pruning out the instances, which are most likely identified as negative instances. Unlike other methods, the proposed work aims to keep the instances in the classification model, which can well represent both positive and negative concepts.

C. Experimental Results

The experiments are conducted on 8 semantic concepts with different P/N ratios. In Table V, MAP value based on different retrieved levels, such as Top 5, Top 10, are presented for different sampling methods considering all the

Table V: Different retried levels of MAP values for all the semantic concepts

Framework	Top5	Top10	Top20	Top30	Top40	Top50	Top80	Top100	Top150	Top200	Overall
Original	0.6167	0.6029	0.5517	0.5283	0.5177	0.5004	0.4636	0.4515	0.4208	0.4055	0.1328
RS	0.5268	0.4766	0.4712	0.4184	0.4078	0.3977	0.3793	0.3669	0.3572	0.3459	0.1074
MCA-based	0.6238	0.5458	0.5128	0.4787	0.4616	0.4445	0.4132	0.4032	0.3851	0.3726	0.1375
Proposed	**0.6984**	**0.6474**	**0.6116**	**0.5826**	**0.5754**	**0.5629**	**0.5429**	**0.5268**	**0.4966**	**0.4753**	**0.1504**

Table VI: Different retried levels of MAP values for Semantic Concept 7 (Person)

Framework	Top5	Top10	Top20	Top30	Top40	Top50	Top80	Top100	Top150	Top200	Overall
Original	1	1	0.9606	0.9219	0.8812	0.8267	0.7552	0.7166	0.6560	0.6260	0.2181
RS	0.4777	0.5259	0.4989	0.4802	0.4801	0.4818	0.4721	0.4684	0.4753	0.4776	0.2366
MCA-based	0.5833	0.5238	0.4775	0.4443	0.4382	0.4421	0.4369	0.4334	0.4357	0.4398	0.2354
Proposed	**1**	**0.95**	**0.8913**	**0.8351**	**0.8131**	**0.7866**	**0.7574**	**0.7363**	**0.6961**	**0.6673**	**0.2421**

Table VII: Different retried levels of MAP values for Semantic Concept 6 (Overlaid Text)

Framework	Top5	Top10	Top20	Top30	Top40	Top50	Top80	Top100	Top150	Top200	Overall
Original	0.3333	0.3333	0.2116	0.2116	0.2116	0.1830	0.1458	0.1508	0.1576	0.1582	0.05551
RS	1	0.6666	0.6666	0.3279	0.3310	0.3297	0.3143	0.3025	0.2871	0.2762	0.04054
MCA-based	1	0.7888	0.6697	0.5674	0.5309	0.4783	0.4258	0.4078	0.3960	0.3795	0.07336
Proposed	**1**	**0.8678**	**0.7487**	**0.6464**	**0.6099**	**0.5573**	**0.5048**	**0.4868**	**0.4750**	**0.4585**	**0.15236**

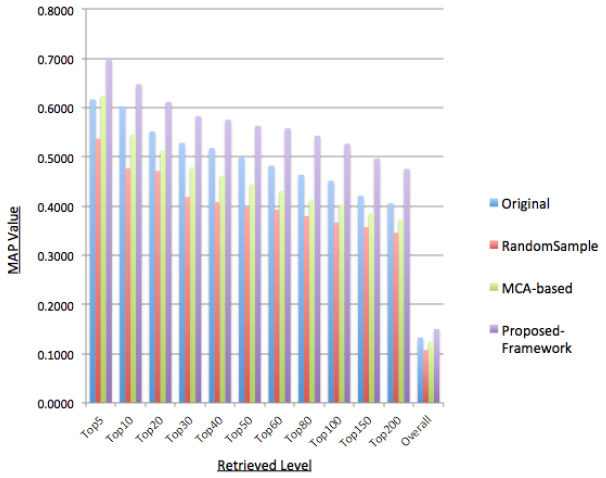

Figure 5: Comparison Results: MAP values in different retrieved levels

concepts. RS did not make any improvement and it seems to trade the precision with using a smaller training set. MCA-based is able to demonstrate a higher MAP value compared to the original data set, but the improvement is relatively minor. The proposed method is able to produce the highest MAP values in every retrieved level and the improvement rate ranges from 1.7% to 7.2%. In addition, it has average 9.92% higher MAP difference and at least 14.68% higher MAP difference across all the retrieved levels against MCA-based method and Random Sampling method, resepctively. The results are also presented in Figure 5.

To further investigate the effective of the proposed work on semantic concepts with different P/N ratios, we breaks down the results into single concept. In Table VI, these are the results for concept 7 "Person", which has the P/N ratio up to 26.96%. As shown, the proposed framework is not able to gain much advantage from retrieved level "Top 10" to "Top 50", but it manages to produce better results when considering more retrieved data instances. On the other hand, both RS and MCA-based have less MAP values for all the levels except for the last one when comparing to original data.

The classification results of concept "Overlaid Text", which has a relatively small P/N ratio 3.33%, are depicted in Table VII. It clearly demonstrates that the proposed work outperform all the other works in all the levels. Specifically, it improved almost 10% compared to the original data when calculating MAP value based on all the instances. Although, RS and MCA-based are able to produce better MAP values compared to the original training data, which shows the importance of performing sampling method on large data set. The difference between the proposed method and other two methods pointed out the fact that it is crucial to considering representative instances when designing a sampling method. Moreover, the proposed method aims to keep the representative negative instances while performing sampling method. Thus, it is able to perform much better results against other sampling methods when the P/N ratio of concept is relatively high.

IV. Conclusion

The paper proposed a new thinking when designing a sampling method and it consists of three major steps: negative feature selection, negative ranking score generation, and negative-based sampling method. First, a negative feature selection method is derived from an existing work called FC-MST [17] to identify features, which are highly correlated with negative concept. With the selected features, MCA is adopted to generate the transaction weight (a negative ranking score) for each instance accordingly. Since the higher the ranking score is, the more likely the instances will be identified as negative instances, the proposed sampling method utilizes this information and selects only the instances with higher negative ranking scores.

TRECVID 2011 data set is selected to testify the performance on different levels of imbalanced data. The proposed method is compared with two methods and the original training data without sampling method. Based on the results, it can conclude into threefold. First, the proposed method clearly demonstrates its strength when coping with imbalanced data set. Second, sampling method like "Random Sample" does not always have better results since randomly filter out the negative instances might result in poor classification performance. Lastly, the experimental results have validated the proposed assumption that it is important to select the representative instances for both positive and negative instances when applying sampling method.

Acknowledgment

This research was supported in part by the U.S. Department of Homeland Security under grant Award Number 2010-ST- 062-000039, the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001, NSF HRD-0833093, CNS-1126619, and CNS-1461926.

References

[1] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "Generalized affinity-based association rule mining for multimedia database queries," *Knowledge and Information Systems*, vol. 3, no. 3, pp. 319–337, 2001.

[2] H.-Y. Ha, Y. Yang, F. C. Fleites, and S.-C. Chen, "Correlation-based feature analysis and multi-modality fusion framework for multimedia semantic retrieval," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2013, pp. 1–6.

[3] H.-Y. Ha, F. C. Fleites, S.-C. Chen, and M. Chen, "Correlation-based re-ranking for semantic concept detection," in *IEEE International Conference on Information Reuse and Integration (IRI)*, 2014, pp. 765–770.

[4] H.-Y. Ha, S.-C. Chen, and M.-L. Shyu, "Utilizing indirect associations in multimedia semantic retrieval," in *IEEE International Conference on Multimedia Big Data (BigMM)*, 2015, pp. 72–79.

[5] S.-C. Chen, M.-L. Shyu, and C. Zhang, "Innovative shot boundary detection for video indexing," in *Video Data Management and Information Retrieval*, S. Deb, Ed. Idea Group Publishing, 2005, pp. 217–236.

[6] S.-C. Chen, X. Wang, N. Rishe, and M. A. Weiss, "A web-based spatial data access system using semantic R-trees," *Information Sciences*, vol. 167, no. 1, pp. 41–61, 2004.

[7] Z. Xie, T. Quirino, M.-L. Shyu, S.-C. Chen, and L. Chang, "A distributed agent-based approach to intrusion detection using the lightweight pcc anomaly detection classifier," in *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, vol. 1, 2006, pp. 446–453.

[8] Youtube, "Youtube Official Statistics," http://https://www.youtube.com/yt/press/statistics.html/, 2014, [Online; accessed 10-Oct-2014].

[9] C. Smith, "By the Numbers: 14 Interesting flickr Stats," http://expandedramblings.com/index.php/flickr-stats/, 2015, [Online; accessed 5-May-2015].

[10] M.-L. Shyu, C. Haruechaiyasak, and S.-C. Chen, "Category cluster discovery from distributed www directories," *Journal of Information Sciences, special issue on Knowledge Discovery from Distributed Information Sources*, vol. 155, no. 3, pp. 181–197, 2003.

[11] M.-L. Shyu, T. Quirino, Z. Xie, S.-C. Chen, and L. Chang, "Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 2, no. 3, p. 9, 2007.

[12] H.-Y. Ha, F. C. Fleites, and S.-C. Chen, "Content-based multimedia retrieval using feature correlation clustering and fusion," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 4, no. 2, pp. 46–64, 2013.

[13] ——, "Building multi-model collaboration in detecting multimedia semantic concepts," in *IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom)*, 2013, pp. 205–212.

[14] X. Huang, S.-C. Chen, M.-L. Shyu, and C. Zhang, "User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval." in *Proceedings of the Third International Workshop on Multimedia Data Mining (MDM/KDD'2002) in conjunction with the 8th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2002, pp. 100–108.

[15] X. Chen, C. Zhang, S.-C. Chen, and M. Chen, "A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval," in *IEEE International Symposium on Multimedia (ISM)*, 2005, pp. 37–44.

[16] L. Zheng, C. Shen, L. Tang, T. Li, S. Luis, S.-C. Chen, and V. Hristidis, "Using data mining techniques to address critical information exchange needs in disaster affected public-private networks," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2010, pp. 125–134.

[17] H.-Y. Ha, S.-C. Chen, and M. Chen, "FC-MST: Feature correlation maximum spanning tree for multimedia concept classification," in *IEEE International Conference on Semantic Computing (ICSC)*, 2015, pp. 276–283.

[18] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Feature selection using correlation and reliability based scoring metric for video semantic detection," in *IEEE International Conference on Semantic Computing (ICSC)*, 2010, pp. 462–469.

[19] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Video semantic concept discovery using multimodal-based association classification," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2007, pp. 859–862.

[20] S.-C. Chen, S. Sista, M.-L. Shyu, and R. L. Kashyap, "Augmented transition networks as video browsing models for multimedia databases and multimedia information systems," in *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'99)*. IEEE, 1999, pp. 175–182.

[21] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying overlapped objects for video indexing and modeling in multimedia database systems," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 04, pp. 715–734, 2001.

[22] S.-C. Chen, M.-L. Shyu, and R. Kashyap, "Augmented transition network as a semantic model for video data," *International Journal of Networking and Information Systems, Special Issue on Video Data*, vol. 3, no. 1, pp. 9–15, 2000.

[23] X. Chen, C. Zhang, S.-C. Chen, and S. Rubin, "A human-centered multiple instance learning framework for semantic video retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 39, no. 2, pp. 228–233, 2009.

[24] M.-L. Shyu, C. Haruechaiyasak, S.-C. Chen, and N. Zhao, "Collaborative filtering by mining association rules from user access sequences," in *Proceedings of International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05)*, 2005, pp. 128–135.

[25] L. Lin, M.-L. Shyu, and S.-C. Chen, "Enhancing concept detection by pruning data with mca-based transaction weights," in *IEEE International Symposium on Multimedia (ISM)*, 2009, pp. 304–311.

[26] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht, "An effective content-based visual image retrieval system," in *IEEE International Conference on Computer Software and Applications Conference (COMPSAC) International*, 2002, pp. 914–919.

[27] L. Lin, C. Chen, M.-L. Shyu, and S.-C. Chen, "Weighted subspace filtering and ranking algorithms for video concept retrieval," *IEEE International Conference on Multimedia*, vol. 18, no. 3, pp. 32–43, 2011.

[28] M.-L. Shyu, S.-C. Chen, M. Chen, C. Zhang, and K. Sarinnapakorn, "Image database retrieval utilizing affinity relationships," in *ACM International Workshop on Multimedia Databases*, 2003, pp. 78–85.

[29] S.-C. Chen, S. H. Rubin, M.-L. Shyu, and C. Zhang, "A dynamic user concept pattern learning framework for content-based image retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 36, no. 6, pp. 772–783, 2006.

[30] M.-L. Shyu, S.-C. Chen, and C. Haruechaiyasak, "Mining user access behavior on the WWW," in *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 3, 2001, pp. 1717–1722.

[31] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405–425, 2014.

[32] Y.-J. Lee, Y.-R. Yeh, and Y.-C. F. Wang, "Anomaly detection via online oversampling principal component analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1460–1470, 2013.

[33] J. A. Rodriguez, R. Xu, C.-C. Chen, Y. Zou, and J. Miao, "Oversampling smoothness: an effective algorithm for phase retrieval of noisy diffraction intensities," *Journal of applied crystallography*, vol. 46, no. 2, pp. 312–318, 2013.

[34] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "Smote-rsb*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory," *Knowledge and Information Systems*, vol. 33, no. 2, pp. 245–265, 2012.

[35] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "Eusboost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," *Pattern Recognition*, vol. 46, no. 12, pp. 3460–3471, 2013.

[36] H. Yu, J. Ni, and J. Zhao, "Acosampling: An ant colony optimization-based undersampling method for classifying imbalanced dna microarray data," *Neurocomputing*, vol. 101, pp. 309–318, 2013.

[37] U. M. Fayyad and K. B. Irani, "On the handling of continuous-valued attributes in decision tree generation," *Machine learning*, vol. 8, no. 1, pp. 87–102, 1992.

[38] L. Lin, M.-L. Shyu, and S.-C. Chen, "Association rule mining with a correlation-based interestingness measure for video semantic concept detection," *International Journal of Information and Decision Sciences*, vol. 4, no. 2, pp. 199–216, 2012.

[39] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Correlation-based video semantic concept detection using multiple correspondence analysis," in *IEEE International Symposium on Multimedia (ISM)*, 2008, pp. 316–321.

[40] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Effective supervised discretization for classification based on correlation maximization," in *IEEE International Conference on Information Reuse and Integration (IRI)*, 2011, pp. 390–395.

[41] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *ACM International Workshop on Multimedia Information Retrieval MIR*, New York, NY, USA, 2006, pp. 321–330.

[42] S. Ayache and G. Quénot, "Video corpus annotation using active learning," in *Advances in Information Retrieval*. Springer, 2008, pp. 187–198.