International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-3, Issue-5, November 2013

Organizing Information in the Blogosphere: The Use of Unsupervised Approach

Ramesh Kumar Ayyasamy

***Abstract*—This study covers the learning approaches discussed by the information retrieval community in categorising texts with a specific focus given to blogs within the last ten years. Early research studies were solely focused on general text classification and these techniques were later improved, and applied to classify webpages, and then to blogs due to the similarity of textual content present in these items. We review how blog classification techniques have evolved from the foremost text classification techniques to the recent ones and discuss the future research directions.**

***Index Terms*— Blog classification, Blogosphere, Supervised, Semi-supervised, Unsupervised classification.**

I. INTRODUCTION

Blogs are different from generic webpages, where even single blog content could consist of different topics, and searching the valuable information would be of extensive inquisitiveness. So a convenient solution is to classify blogs based on its stated contents. Since the content of blogs is diverse in nature, classifying blogs is a difficult and challenging task than traditional text classification ([1], [2], [3], [4], [5], [6], [7], [8], [9]). Major research studies on blog classification was identified and grouped based on their approaches and relevance to this survey. Since every approach has its pros and cons, we also consider the strengths and limitations of these approaches and briefly discuss how these limitations can be addressed.

The state of the art text classification and research survey has been discussed in some literatures such as, on text classification [10], data clustering ([11], [12]), webpage classification ([13], [14]), and hypertext categorization ([15], [16]). As similar to webpage classification [14], an increasing body of research studies was started on blog classification based on different specific problems: subject classification ([2], [18]), functional classification ([19], [8]), sentimental classification ([21], [22], [23]), spam blog classification [24], blog genre classification ([25], [26]), mood classification [27] and many other types of classifications. The information content present in the blogosphere has been proven more valuable for applications such as business intelligence, trend discovery, and opinion tracking [28]. As blog classification is in its nascent stage, there is no extensive review of literature that describes different approaches adopted by various researches to date.

Manuscript received November, 2013.

Dr. Ramesh Kumar Ayyasamy, Lecturer, School of Information Technology, Monash University, Malaysia.

We first enumerate the different classification approaches in Section 2. In Section 3, limitations of each approach are discussed. Section 4 concludes with a brief summary that provides insight into some future research directions that could enhance the performance of current blog classification based research activities.

II. CLASSIFICATION APPROACHES

We explain different classification methods and describe its functions, limitations on classifying documents. As shown in Figure 1, classification approaches are applied for any forms of text such as webpages, blogs, spam blogs, forums, emails and so forth. The purpose of explaining different classification approaches in this section is to identify the strengths and limitations of the current needs on classifying blogs.

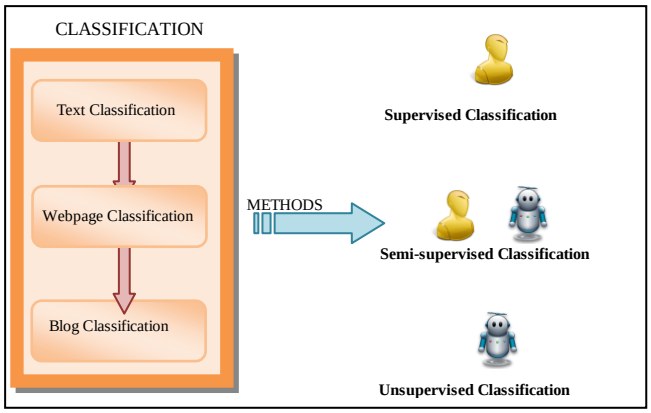

Figure 1. Classification Overview

A. Supervised learning approach

The supervised learning or so called inductive learning approach was started around 1995-96 [29], where a set of classification rules or the features are learned from a collection of labelled training documents. These rules are employed to classify target data items from other similarly formatted pages. Figure 2 shows the functionality of the overall system of supervised classification that uses human indexers to provide a good number of labelled training documents for each category to learn. The needs of manual classification are also not eliminated, where training text documents are labelled by human indexers based on the respective categories using cognitive judgement. Once supervised learning algorithm learns the rules in identifying different categories, the algorithm is able to classify any document which is given according to the training provided. A wide range of supervised learning algorithms are available namely, kNN [30], SVM [31], naive Bayes ([32], [33]),

Supervised Neural Network [34], and Bayesian Network classifier [35].

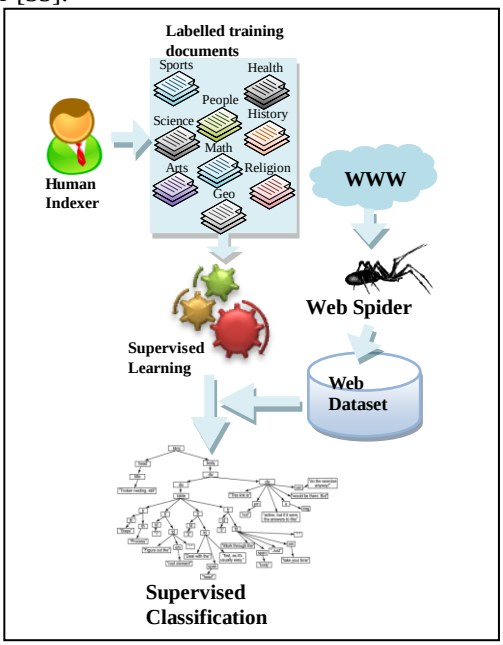

Figure 2. Diagrammatic Representation of Supervised Classification

SVM learning algorithm is suggested in areas where documents are not pre-classified. In some analysis, SVM classifier can easily be improved using training data and prior knowledge, to identify similarities between them. SVM learning algorithm is closely related to supervised neural network model in identifying the data structure during learning its feature space to successfully classify all the unlabelled data.

B. Semi-supervised learning approach

In order to address the issue of reducing the involvement of human indexers and to improve the learning accuracy on classifying documents, semi-supervised classification arisen. This type of classification is half way between supervised and unsupervised classification. Figure 3 shows the functionality of the overall system of Semi-supervised classification that uses human indexers to label parts of the documents for each category to learn. When dealing with the situations where few labelled training documents and large number of unlabelled training documents are available, then semi-supervised classification are used ([36], [37]). It starts by training the classifiers on labelled documents and, in each step a part of unlabelled documents are used. During training, the unlabelled documents are helpful in providing the joint probability distribution over words [29].

Several semi-supervised learning algorithms such as PLSA [39], S3VM classifier [40], TSVM classifier [41], Semi-supervised naive Bayes classifier [42], and Semi-supervised clustering [43] are used for text classification.

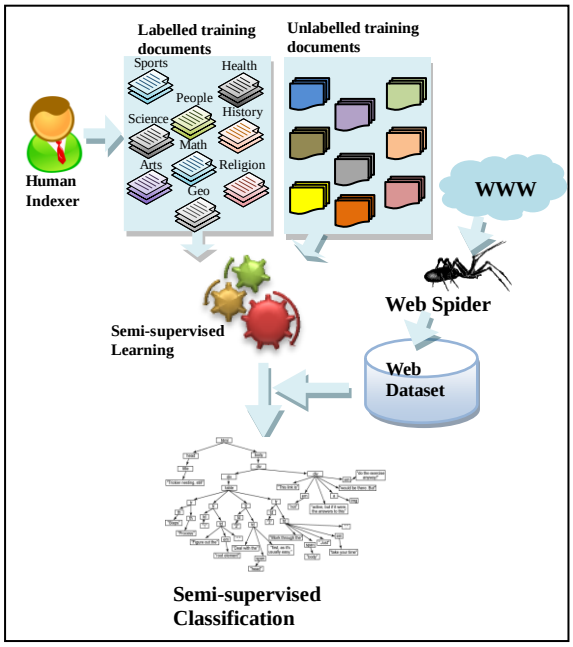

Figure 3. Diagrammatic Representation of Semi-supervised Classification

During this classification, the learning algorithm takes only few labelled documents and large number of unlabelled documents for training.

C. Unsupervised learning approach

Unsupervised learning approach was started around 1998 [29]. Figure 4 shows the functionality of the overall system of unsupervised classification that uses learning algorithm to solve the classification task without human intervention. It does not require the foreknowledge of each dataset and does classification automatically. The unsupervised classification becomes the most important for the present state, since labelled training documents are seldom available. In general, the task of unsupervised learning is more abstract and less defined.

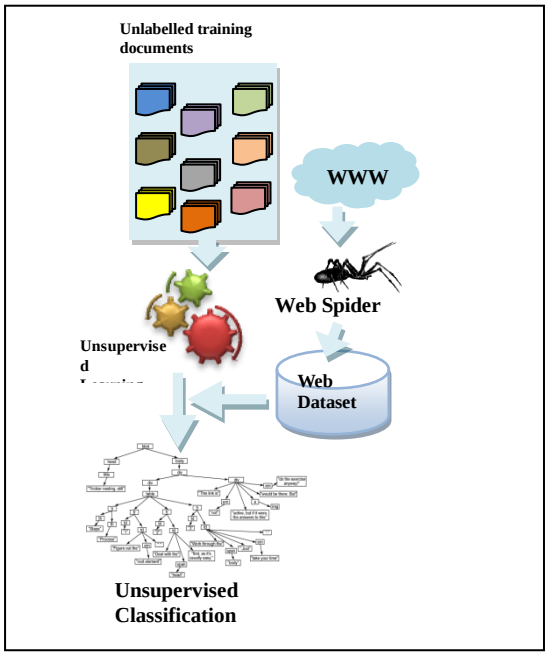

Figure 4. Diagrammatic Representation of Unsupervised Classification

Unsupervised learning algorithms such as ANN [44], and text-based unsupervised clustering such as k-means clustering ([45], [46], [47]), non-negative matrix factorization ([48] [49], spectral clustering ([50], [51]) are used for text classification. Here, these learning algorithms learn on their own and does classification for different real-time text documents.

ANN is relatively crude electronic network of neurons inspired from the brain's neural structure. ANN model processes each record, and compares its record pattern with the known actual classified record. SOM is one of the prominent ANN model produces a similarity graph of input data [52]. In order to use the SOM for text classification, documents are needed to be represented as a histogram of word occurrences. SOM visualises the document similarity in terms of distances on a two-dimensional map display, where every document is represented within a single two-dimensional map [53]. The limitation of SOM is its high dimensional feature space document representation [44]. Hung et al. [55] used hybrid neural network model-SOM, guided by WordNet to cluster documents. This hybrid model considers each word as a symbol and word sequences are ignored. Finally, author has mentioned that there exists a wide gap between the neural clustering and manual text classification.

Clustering deals in identifying a proper structure in a finite set of unlabelled data. There is a considerable amount of literature on clustering, and exploring this literature is complicated by the fact that there are many domains in which clustering can be applied. One method in particular, clustering has been successful in a wide range of knowledge discovery applications [56]. Berkhin [11] investigated the applications of clustering algorithms for data mining. Clustering has received a significant amount of attention and has been used in the area of text classification as a term selection for dimensionality reduction or as a method to enhance the training set [57]. Text classification using clustering methods focused on searching similarities in the document content, and organizes in groups according to these properties. During training, clustering algorithm classifies similar objects in one group, and dissimilar objects to others [58]. Initially it discovers a kind of structure in the training examples and expands the feature vectors with new attributes extracted from clusters. Clustering is usually performed when there is no information concerning the membership of data items to pre-defined classes.

Slonim et al. [54] proposed document classification framework by following unsupervised clustering methods. This framework searches for highly correlated clusters with the real categories, using sequential Information Bottleneck algorithm. Their results are compared with a supervised naive Bayes classifier and found to be similar. In most of the experiments, partitioned clustering algorithm is suited for large data set clustering due to their low computational requirements [12]. Baker et al. [38] used the distributional clustering method, which clusters the terms together that specifies the occurrence of the same category, or a set of categories. Xu et al. [49] presented a simple document clustering application using Non-negative Matrix Factorization. As k-means clustering algorithm cannot separate clusters that are non-linearly separable in input space, Dhillon et al. [20] used JSD to cluster words in k-means fashion in text clustering.

Trends in data mining have demanded clustering algorithm to utilize two extremely correlated objects, terms and documents in a text dataset. Clustering individual object type would not perform better, since each type could be defined by another object types. This paved the way for many researchers to co-cluster two or more heterogeneous data. Dhillon [50] and Zha et al. [17] have expanded the generic clustering algorithms into bipartite graph clustering algorithm to cluster terms and documents concurrently. Gao et al. [51] suggested Consistent Bipartite Graph Co-Partitioning algorithm by considering each tripartite graph as two-bipartite graphs. This study proved that consistent partitioning provides the optimal solution using positive Semi-Definite Programming.

III. LIMITATIONS ON CLASSIFICATION APPROACHES

Every classification method has its own limitations in terms of cost, efficiency and type of input data presented. In supervised classification, the learning algorithm requires sufficient amount of training documents to achieve higher precision in classifying texts. Obtaining labelled training documents is expensive and if rare categories consist of only very few training documents, then the classifier do not perform well to that particular category. Chances of labelled documents being biased based on human indexer's perception are possible.

Semi-supervised learning does the task of learning from labelled and unlabelled examples; its drawbacks are as same as that of supervised classification. The limitations in semi-supervised classification are, it involves human indexers to label training documents, and such learning algorithms cannot be used for large datasets. Another key limitation is that although the labelled training documents may be small, every category should have some labelled examples [29]. Since the involvement of human indexers for classification is expensive and time consuming, the need for automatic learning came into existence.

As mentioned in the above section, unsupervised classification methods are done through clustering algorithms. One widely used technique in unsupervised learning is SOM. The drawback in SOM is, it is a static model and uses only fixed number of input units. SOM does not represent structure of clusters and actual distance between two different clusters. SOM follows fixed two dimensional - 2D lattices and the size of 2D lattices is based on trial and error. Thus research studies are done to integrate SOM with the ontology to increase performance [55]. Evidently, this study states that there is still a gap between the neural clustering and manual text classification. But existing clustering techniques do not satisfy the requirements of different form of texts. Another disadvantage in unsupervised classification method is time-complexity and high-processing speed. Since the learning algorithms train on its own, iterating with large number of dimensions and its data items can be challenging because it takes lots of time to do classification.

IV. CONCLUSIONS AND FUTURE WORK

Blog classification is in its nascent stage and much of the initial investigation of blog classification research studies have been carried out based on traditional text and webpage classification approaches, and not by considering the blogs as an individual entity. This section summarizes the existing research studies and briefly highlights possible future research directions.

Most research studies on blog classification follow three basic text classification approaches: supervised, semi-supervised and unsupervised classification. Thus, in this paper we provided the review of different text classification methods, learning algorithms and their functions. Here supervised and semi-supervised classification methods require the labelled and unlabelled data. Furthermore, limitations of text classification method were mentioned and unsupervised classification is mostly needed in the current scenario. Traditional text classification method treats each word appearing in the documents as features and follows Bag-of-Words approach, which disregards the semantic relationships between key terms, and the order of appearance.

REFERENCES

[1] N. Agarwal, M. Galan, H. Liu, and S. Subramanya, 'Clustering blogs with Collective Wisdom', In 8th International Conference on Web Engineering, 2008, pp. 336-339.

[2] I. Bayoudh, N. Bechet, and M. Roche, 'Blog classification: Adding linguistic knowledge to improve the K-NN algorithm', Intelligent Information Processing IV, 2008, pp. 68-77.

[3] D. Cao, X. Liao, H. Xu, and S. Bai, 'Blog post and comment extraction using information quantity of web format', Information Retrieval Technology, 2008, pp. 298-309.

[4] E. Elgersma, and M. De Rijke, 'Learning to recognize blogs: A preliminary exploration', EACL Workshop: New Text - Wikis and blogs and other dynamic text sources, 2006, pp. 24-73.

[5] E. Elgersma, and M. De Rijke, 'Personal vs Non-Personal Blogs: Initial classification experiments', In ACM SIGIR conf., 2008, pp. 723-724.

[6] D. Ikeda, H. Takamura, and M. Okumura, 'Semi-supervised learning for blog classification', In AAAI, 2008, pp. 1156-1161.

[7] P. Kolari, T. Finin, and A. Joshi, 'SVMs for the blogosphere: Blog identification and splog detection', In AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, vol. 4, 2006. Available: http://aisl.umbc.edu/resources/213.pdf.

[8] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura, 'Automatically collecting, monitoring, and mining Japanese weblogs', In 13th International World Wide Web Conference on Alternate Track Papers and Posters, 2004, pp. 320-321.

[9] I. Ounis, C. Macdonald, and I. Soboroff, 'On the TREC Blog Track', In ICWSM conference, 2008. Available: http://terrierteam.dcs.gla.ac.uk/publications/ounis08trecblog.pdf.

[10] F. Sebastiani, 'Machine learning in automated text categorization', ACM Computing Surveys, vol. 34, no. 1, 2002, pp. 1-47.

[11] P. Berkhin, 'A survey of clustering data mining techniques', Grouping multidimensional data, 2006, pp. 25-71.

[12] R.Xu, and D. Wunsch, 'Survey of clustering algorithms', IEEE Transactions on Neural Networks, vol. 16, no. 3, 2005, pp. 645-678.

[13] B. Choi, and Z. Yao, 'Web page classification*', Foundations and Advances in Data Mining, 2005, pp. 221-274.

[14] X. Qi, and B.D. Davison, 'Web Page Classification: Features and Algorithms', ACM Computing Surveys, vol. 41, no. 2, 2009. Available: http://www.cse.lehigh.edu/~xiq204/pubs/classification-survey/LU-CSE-07-010.pdf.

[15] Fürnkranz, J., 'Web Mining', Data mining and knowledge discovery handbook, 2005, pp. 899-920.

[16] Y. Yang, S. Slattery, and R. Ghani, 'A study of approaches to hypertext categorization', Journal of Intelligent Information Systems, vol. 18, no. 2, 2002, pp. 219-241.

[17] H. Zha, X. He, C. Ding, H. Simon, and M. Gu, 'Bipartite graph partitioning and data clustering', In CIKM '01, 2001, pp. 25-32.

[18] H. Qu, A.L. Pietra, and S. Poon, 'Automated blog classification: Challenges and pitfalls', In AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006, pp. 184-185.

[19] D. Ikeda, H. Takamura, and M. Okumura, 'Semi-supervised learning for blog classification', In AAAI conference, 2008, pp. 1156-1161.

[20] I.S. Dhillon and Y. Guan, 'Information theoretic clustering of sparse co-occurrence data', In ICDM conference, 2003, pp. 517-520.

[21] P. Chesley, B. Vincent, L. Xu, and R.K. Srihari, 'Using verbs and adjectives to automatically classify blog sentiment', In AAAI Spring Symposium on Computational Approaches to Analysing Weblogs, 2006, pp. 27-29.

[22] T. Li, Y. Zhang, and V. Sindhwani, 'A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge', In AFNLP, vol. 1, 2009, pp. 244-252.

[23] B. Pang, and L. Lee, 'Opinion mining and sentiment analysis', Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, 2008, pp. 1-135.

[24] P. Kolari, A. Java, T. Finin, J. Mayfield, A. Joshi, and J. Martineau, 'Blog track open task: spam blog classification', Technical Report, TREC 2006 Blog Track Notebook, 2006.

[25] S.C. Herring, L.A. Scheidt, S. Bonus and E. Wright, 'Bridging the gap: A genre analysis of weblogs', In 37th Hawaii International Conference on System Sciences, 2004, pp. 1-11.

[26] C.R. Miller, and D. Shepherd, 'Blogging as social action: a genre analysis of the weblog', Into the blogosphere: Rhetoric, community, and culture of weblogs, 2004. Available: http://blog.lib.umn.edu/blogosphere/blogging_as_social_action_a_genre_analysis_of_the_weblog.html.

[27] G. Mishne, 'Experiments with Mood Classification in Blog Posts', In ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access, 2005. Available: http://staff.science.uva.nl/~gilad/pubs/style2005-blogmoods.pdf.

[28] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton and T. Tomokiyo, 'Deriving marketing intelligence from online discussion', In 11th ACM SIGKDD conference, 2005, pp. 419-428.

[29] B. Liu, Web data mining: Exploring hyperlinks, contents, and usage data, Springer Verlag, 2007, pp. 1-532.

[30] E. Fix, and J.L. Hodges Jr, Discriminatory analysis-nonparametric discrimination: small sample performance, USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 4, Contract AF41(128)-31, 1952.

[31] V.N. Vapnik, The nature of statistical learning theory, Springer-Verlag New York Inc, 1995.

[32] A. McCallum, and K. Nigam, 'A comparison of event models for naive bayes text classification', In AAAI-98 Workshop on Learning for Text Categorization, vol. 752, 1998, pp. 41-48.

[33] J.D. Rennie, L. Shih, J. Teevan, and D. Karger, 'Tackling the poor assumptions of naive bayes text classifiers', In ICML conference, vol. 20, 2003, pp. 616-623.

[34] H.T. Ng, W. B. Goh, and K.L. Low, 'Feature selection, perceptron learning, and a usability case study for text categorization', In ACM SIGIR Conference, vol. 31, 1997, pp. 67-73.

[35] N. Friedman, D. Geiger, and M. Goldszmidt, 'Bayesian network classifiers', machine learning, vol. 29, no. 2, 1997, pp. 131-163.

[36] O. Chapelle, M. Chi, and A. Zien, 'A continuation method for semi-supervised SVMs', In ICML conference, 2006, pp. 185-192.

[37] X. Zhu, 'Semi-supervised learning literature survey', University of Wisconsin – Madison, 2005, pp. 1-60.

[38] L.D. Baker, and A.K. McCallum, 'Distributional clustering of words for text classification', In ACM SIGIR conference, 1998, pp. 96-103.

[39] T. Hofmann, 'Probabilistic Latent Semantic Analysis', In 15th Conference on Uncertainty in Artificial Intelligence, 1999, pp. 289-296.

[40] K. Bennett and A. Demiriz, A, 'Semi-supervised support vector machines', Advances in Neural Information processing systems, 1999, pp. 368-374.

[41] V. Vapnik, Statistical learning theory, Wiley, New York, 1998, pp. 1-768.

[42] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, 'Text classification from labeled and unlabeled documents using EM', Machine learning, vol. 39, no. 2, 2000, pp. 103-134.

[43] S. Basu, A. Banerjee, and R. Mooney, 'Semi-supervised clustering by seeding', In ICML conference, 2002, pp. 19-26.

[44] D. Merkl and A. Rauber, 'Document classification with unsupervised artificial neural networks', Soft Computing in Information Retrieval: Techniques and Applications, vol. 50, 2000, pp. 102-121.

[45] I.S. Dhillon, Y. Guan, and B. Kulis, 'Kernel k-means: Spectral clustering and normalized cuts', In ACM SIGKDD conference, 2004, pp. 551-556.

[46] A. Hotho, S. Staab, and G. Stumme, 'Wordnet improves text document clustering', In ACM SIGIR Semantic Web Workshop, 2003, pp. 1-9.

[47] J. Sedding, and D. Kazakov, 'WordNet-based text document clustering', In Proc. of the COLING 2004 3rd Workshop on Robust Methods in Analysis of Natural Language Data, 2004, pp. 104-113.

[48] F. Shahnaz, M.W. Berry, V. P. Pauca, and R. J. Plemmons, 'Document clustering using nonnegative matrix factorization', Information Processing and Management, vol. 42, no. 2, 2006, pp. 373-386.
[49] W. Xu, X. Liu, and Y. Gong, 'Document clustering based on non-negative matrix factorization', In ACM SIGIR conference, 2003, pp. 267-273.
[50] I. S. Dhillon, 'Co-clustering documents and words using bipartite spectral graph partitioning', In ACM SIGKDD conference, 2001, pp. 269-274.
[51] B. Gao, T.Y. Liu, X. Zheng, Q. S. Cheng, andW.Y. Ma, 'Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering', In ACM SIGKDD conference, 2005, pp. 41-50.
[52] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, 'Self organization of a massive document collection', IEEE Transactions on Neural Networks, vol. 11, no. 3, 2000, pp. 574-585.
[53] D. Merkl, 'Text classification with Self-Organizing maps: Some lessons learned', Neurocomputing, vol. 21, no. 1-3, 1998, pp. 61-77.
[54] N. Slonim, N. Friedman, and N. Tishby, 'Unsupervised document classification using sequential information maximization', In ACM SIGIR conference, 2002, pp. 129-136.
[55] C. Hung, S. Wermter, and P. Smith, 'Hybrid neural document clustering using guided self-organization and WordNet', IEEE Intelligent Systems, vol. 19, no. 2, 2004, pp. 68-77.
[56] G.T. Lakshmanan, and M. A. Oberhofer, 'Knowledge discovery in the blogosphere: Approaches and challenges', IEEE Internet Computing, vol. 14, no. 2, 2010, pp. 24-32.
[57] A. Kyriakopoulou, and T. Kalamboukis, 'Text classification using clustering', In Proceeding of the ECML-PKDD The Discovery Challenge Workshop, 2006, p. 28-38. Available: http://ceas2009.cc/discovery_challenge_proceedings.pdf#page=32.
[58] N. Grira, M. Crucianu, and N. Boujemaa, 'Unsupervised and semi-supervised clustering: A brief survey', A Review of Machine Learning Techniques for Processing Multimedia Content', Report of the MUSCLE European Network of Excellence (FP6), 2004, pp. 1-12. Available: http://cedric.cnam.fr/~crucianm/src/BriefSurveyClustering.pdf.

Dr.Ramesh Kumar Ayyasamy is a lecturer in School of Information Technology, Monash University, Malaysia. He received his PhD-IT from Monash University, Australia and his Mphil-CS, Masters in Comp. Appln., Bachelors in Electronics degrees from Bharathiar University, India consecutively. He is currently researching on Telemedicine platform, Monash University, Malaysia.