Research of Novel Web Page Classifiers and Feature Selection Methods

S. Markkandeyan, P.Kalyanasundaram, U.Muthaiah

Abstract: ***The World revolves around the web technology at present. Every year, the Web information are exponentially growing and this information are huge and complex. The web users are difficult to classify and extract useful information from the web, because the Webinformation are noisy, redundant and irrelevant and also misclassified.Many researchers don't have strongknowledge about the process of web page classification, techniques and methods previously used. The objective of this survey is to convey an outline of the modern techniques of Web page classification. In this survey, the recent papers in this area are selected and explored.Thus this study will help the researchers to obtain the required knowledge about the current trends in web page classification.***

Keywords: Web page classification, Machine learning, Feature selection.

I. INTRODUCTION

Web page classification is a difficult process because huge number of information available in the Web and also it dynamically and regularly changing environment. So manual classification of Web page is more time consuming, in accurate and misclassification. So, an automatic Web page classification process is needed.There are variousadditional procedures (preprocessing, transformation and dimensionality reduction) are applied in Web page classification. It is an important research topic, exploitnumerous techniques and methods to solve for complex systems. Using this concept researchers are developing new concepts or existing ones to achieve better results. Usually Web page classification process starts from (1) Data collection (Web pages), (2) Preprocessing (Html tags removal, stop word removal, tokenization, stemming etc.), (3) Feature extraction and selection (4) Classification using Machine learning algorithms and performance measurement.This survey is structured as following manner. The complete description of the existing reviews are presents in Section 2. Finally, Section 3 concludes the research study.

II. LITERATURE OVERVIEW

There are excellent works previously developed in the area of Web page classification and these works are very useful to the research community still now.In this paper, the detailed survey is made onrecent methodical overview to better understanding, what has been achieved and what are the difficulties overcome in this field. This paper summarizes the latest and complete study of Web page classification in recent years.This study will explore and develop new aspect of Web page classification.

Rajalakshmi and Xavier [1] described Web page classification using URL based features. They have used classification steps namely preprocessing, feature weighting methods (n-Gram, Term Frequency and Mutual Information), training (70%) and testing samples (30%) of WebKB dataset. In addition authors used, Support Vector Machines (SVMlight) as a machine learning method and they attained F1 performance measure of 79%. This method produced good results compare to the other existing methods. Thedrawback of this method is experimented with small dataset.

Sabbah et al. [2] developed a hybridized feature selection technique for detecting terrorism activities (dark Web page) based on vector space model, symmetric difference and union combination function as a term weighting method. The authors used datasets for dark Web forum portal and open source Arabic corpora. These datasets are preprocessed with filtering, tokenizing, and Larkey's light stemmer algorithm and they got top 'k' feature sets (texts). The proposed method significantly reduced dimensionality of the feature sets and they attain 96% accuracy using SVM as a text classifier compared to the other well-known classifiers Naïve Bayes (NB),Decision Tree (DT), K-Nearest Neighbor (KNN)and Extreme Learning Machine (ELM).

Bharti and Sing [3] presented a method for hybrid dimensionality reduction using integrating feature selection. They have used text as feature and start with preprocessing techniques such as Stopwordremoval, Porter'sstemmingalgorithm, Tokenization and Term weighting method. The authors proposed a modified union approach that is select top ranked features and remaining feature sublists are select using intersection. From their work term variance (TV), document frequency (DF) as feature selection methods used for Principal Component Analysis (PCA) and features relevance score computation is a feature dimensionality reduction method. They have enhanced higher clustering accuracy using k-means clustering compared to the other competitive methods. They

Revised Manuscript Received on September 10, 2019.

S. Markkandeyan, Professor, Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Sankari, Tamilnadu, India.
(Email: drsmkupt@gmail.com)
P. Kalyanasundaram, Professor, Department of Electronics and Communication Engineering, Saveetha School of Engineering, SaveethaUniversity, Thandalam, Chennai,Tamilnadu, India.
(Email: kalyanasundaramp@yahoo.com)
U. Muthaiah, Assistant Professor, Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Sankari, Tamilnadu, India.
(Email: muthaiahu54@gmail.com)

have utilizeddifferent datasets called Reuters-21,578, Classic4 and WebKB. Their method is very helpful in the application of topic extraction, information retrieval and automatic document organization.

Markkandeyan and Indradevi [4] explained an efficient method for Web page classification using different machine learning techniques. The authors also starts with usual classification process such as preprocessing, feature selection (Principle Component Analysis (PCA),evaluator and search methods), Machine learning techniques and performance evaluation using WebKb and Open Directory Project (ODP) benchmark datasets. They have stored three feature sets namely Initial Feature Set (IFS), Intermediate Feature Set (IMFS) and Final Feature Set (FFS). The training and testing documents are in the ratio of 60/40, or 70/30, or 80/20, or 50/50 or 30/70.From their experiments, the accuracy attain 97% using Meta machine learning classifier and they improve 2.6% accuracy compared to the existing methods.

Lee et al [5]developed amodified swarm optimization technique to classify the Web pages using best weight features. The authors start with usual preprocess of Web page and they apply Taguchi method to find the parameter settings. The proposed method gives better accuracy (F1-88.03%) compared to the Genetic algorithm, Bayesian classifier and KNN using ODP dataset. They have used training and testing dataset in the ratio of 8:2. This method overcome the drawbacks in discrete and continuous problems of existing particle swarm optimization and Genetic algorithm.

Li et al. [6] presented an optimized inheritance association classification algorithms such as Kinship relation association, Preliminary association, Rule Association, and Bayesian classifier as a page class probability (estimated using Levenshtein distance) for huge Web page content classification using entity similarity based on semantic networks. The authors used Chinese Wikipedia (Chinese portal sites) as a dataset, preprocessing it and extract the kernel content information such as title and main text are keyword (features) using a MFPtool (Main-text Founder Process), Wikipedia Knowledge Network (WKN) and N-gram models. Finally, they have applied Bayesian classifier as a machine learning method to achieved better results using WKN. They are try to reduce the category wise class drift problem using different approach in feature and change of knowledge ontology generation.Table 1 summarizes the various authors work in this web page classification research area.

Sabbah et al. [7] described modified frequency based term weighting methods using Reuters-21578, 20Newsgroups, and WebKB bench marking datasets. They have proposed four modified term weighting methods namely Term Frequency - modified Inverse Document Frequency(TFmIDF), modified Document Frequency-Inverse Document Frequency (mDFIDF), modified Term Frequency(mTF)and modified Term Frequency- modified Inverse Document Frequency (mTFmIDF) for the counts of missing terms from the documents. The experimental results for the above methods attained 97% (micro average F1) classification accuracy using SVM machine learning classifier compared to the other classifiers like SVM, KNN, NB and ELM. They have also applied Analysis of variance (ANOVA), paired-sample T-Test and Wilcoxson statistical significance tests because feature sets have different sizes. In future they compare with other feature selection methods (PCA, Information Gain, etc), state-of-the-art techniques (deep learning, optimization) and text classification for imbalanced datasets.

Onan [8] used the different feature selection, different base learners and different ensemble learning methods for Web page classification. They have useddifferent classifiers (C4.5, Naive Bayes, K-nearest neighbor)and FURIA (Fuzzy Unordered Rule Induction Algorithm) algorithm for base learners, Random Subspace,Bagging, BoostingandDagging for ensemble learning methods and consistency, correlation, chi-square and information gain based feature selection methods. The proposed method achieved highest average predictive performance of 88.1% using the combination of Naive Bayes algorithms and consistency based feature selection with AdaBoost. They have used DMOZ-50 dataset for experiments and the combination of ensemble (Random Subspace and Bagging) and feature selection (consistencyandcorrelationbased) techniques to achieve the better accuracy results compared to the other combinations and existing methods.

III. DISCUSSION & RESULTS

From this survey, it is concluded that Web page classification is a welldeveloping research area for researchers. This study will help to the young researchers do the research and publish their work in the area of Web page classification. From this survey, the researchers are easily understand the concept of Web page classification, novel preprocessing techniques, feature extraction and feature selection methods, different machine learning classifiers and their performances, tools and research gap of the existing methods. This survey will be very helpful for professionals and researchers to put forwardinnovative studies in the field of Web page classification.

Table 1: Literature review in Web page classification research area.

Authors andYear	Features used	Preprocessing Techniques	Feature Selection/ Extraction Method	Algorithm / Classifier	Dataset	Accuracy	Accuracy Improvement	Language/Tool	Research Gap
Rajalakshmi, and Xaviar (2017)	URL	Stop word and Stemming	Feature Weighting (n-Gram, Term Frequency and Mutual Information)	SVM	WebKB	79% (F_1-measure)	19.6% (F_1-measure)	Python and SVM[light]	Use small dataset
Sabbah et al. (2016)	Documents (Text)	Filtering, Tokenizing, and stemming (Larkey's Light Stemmer)	Hybridized term-weighting method (TF, DF, TF–IDF, Glasgow, Entropy UNION), Symmetric Difference combination functions and Vector space model	SVM	Dark Web Forum Portal and Open Source Arabic Corpora	96.%	-	Java, Lucene4.3 package, Rapid Miner and MATLAB	They focus on modified term weighting technique for in balanced and un balanced datasets to achieve higher accuracy.
Bharti and Singh (2015)	Text	Stop word removal, Porter's stemming algorithm, Tokenization and Term weighting	Principal component analysis, Document frequency,Term variance, Modified union and intersection	K-means	WebKB,Reuters21578, and Classic4	0.8416 (F1 Score) (using Classic 4)	-	Java and MATLAB	Not consider for the interaction between terms, dependency on the parameters and efficient and effective method for dimension reduction.
Markkandeyan and Indradevi (2015)	All Features (HTML tags, URL, paragraph etc)	Stop word removal and Porter stemming algorithm	PCA, evaluator and Search methods	Meta (Raced Incremental Logit Boost)	WebKB and ODP	97%	2.6%	Java and WEKA	They do not concentrate on more number of positive and negative documents in unbiased classification

Lee et.al. (2015)	HTML tags and terms	Stop word removal and Porter stemming algorithm	Taguchi method	Simplified swarm optimization	ODP	88.03%	7.34%	C	Not consider other feature selection methods (PCA, LSI) and determination of good threshold values.
Li et al. (2017)	Title and Main text	Main-text founder process	N-gram by Wikipedia Entity Words and Wikipedia Knowledge Network	Bayesian classifier and Association algorithms (Kinship relation, Preliminary, and Rule Association)	Chinese Wikipedia (Chinese Portal sites)	98.11%	-	Main-text founder process	They do not reduce the category wise class drift problem and change of knowledge ontology generation
Sabbah et al (2017)	Documents (Text)	Filtering, Tokenizing, and Stemming	Vector space model and Modified frequency-based term weighting schemes (mTF, mTFI DF, TFmIDF, mTFmIDF)	SVM	Reuters-21578, 20Newsgroups, and WebKB	97% (micro-average F1)	-	Java, Lucene4.3 package, Rapid Miner and MATLAB	Not comparewith other feature extraction methods (PCA, IG, etc), state-of-the-art techniques and text classification for imbalanced datasets.
Onan (2015)	Text	-	Dagging, Boosting, Bagging (Random Subspace and Ensemble learning methods) and Chi-square,Consistency, Correlation and Information gain	AdaBoost and Naive Bayes algorithms	DMOZ-50	88.1%	-	Java and WEKA	The author used many feature selection, ensemble learning and classifier techniques, but the predictive performance is not high

VII. REFERENCES

1. R. Rajalakshmi, S. Xaviar, "Experimental study of feature weighting techniques for URL based webpage classification", Procedia Computer Science, 2017, Vol.115, pp. 218-225.
2. T. Sabbah, A. Selamat, M.H. Selamat, R. Ibrahim,H. Fujita,"Hybridized term-weighting method for dark web classification", Neurocomputing,Vol.173, 2016,pp. 908-1926.
3. K.K.Bharti,P.K.Singh, "Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering", Expert Systems with Applications, Vol. 42, No.6, 2015, pp. 3105-3114.
4. S. Markkandeyan, M. Indra Devi, "Efficient machine learning technique for Web page classification", Arabian Journal for Science and Engineering, Vol. 40, No. 12, 2015, pp. 3555-3566.
5. J.H. Lee,W.C. Yeh,M.C. Chuang,"Web page classification based on a simplified swarm optimization", Applied Mathematics and Computation, Vol. 270, 2015, pp. 13-24.
6. H. Li,Z. Xu, T. Li,G. Sun, K.K.R.Choo, "An optimized approach for massive web page classification using entity similarity based on semantic network", Future Generation Computer Systems, Vol.76, 2017, pp. 510-518.
7. T. Sabbah,A. Selamat,M.H. Selamat,F.S. Al-Anzi, E.H. Viedma,O. Krejcar,H. Fujita, "Modified frequency-based term weighting schemes for text classification", Applied Soft Computing, Vol. 58, 2017, pp. 193-206.
8. A. Onan, "Classifier and feature set ensembles for web page classification", Journal of Information Science, Vol. 42, No. 2, 2016, pp.150-165.