Effective Search Engine Spam Classification

D.Saraswathi, A.Krishnakumar

Abstract: ***Search engine spam is formed by the spam creators for commercial gain. Spammers applied different strategies in web pages to display the first page of web search results. These strategies may avoid displaying good quality web pages in the top of search engine results page. Nowadays there are numerous devised algorithms available to identify search engine spam. Even though search engines are still affected by search engine spam. There is a necessity for search engine industry to filter search engine spam in the best way.***

The proposed study identifies spam in web search engine. Spammers try to use most popular search keywords, popular links and advertising keywords in web pages. This strategy helps to increase ranking to display the top of search results. The proposed method is used important features to detect spam pages which are classified using decision tree C4.5 classifier. This method produces better performance when compared with existing classification methods.

Index Terms: Search engine spam, Classification, Spamdexing, Decision Tree, popular search keywords, popular links, and advertising words.

I. INTRODUCTION

The Web is the huge and the most popular storage area of information, satisfying the requests of web surfers. But at the same time there are adversaries also known as spammers, who modify information on web for increasing their marketable profit. The most common form of such manipulation on the web is search engine spam or spamdexing, which poses a huge threat to web security. The victims of this kind of spamming are mostly those users who while querying then search engine, are offered with unwanted pages with malicious content.

Search engine spam was named as Spamdexing. The term 'Spamdexing', framed by Eric Convey in the year 1996, is a combination of 'Spam' and 'Indexing'. Spamdexing refers to decisive manipulation of indexes in search engine. This was predicted afterward as one of the major challenge for search engine industry [1].

Search engine spam is the practice of purposely and untruthfully changing web pages to boost the chances of these pages being positioned close to the beginning of search engine results page. Numerous spammers of web pages try to get a good position to display search results in search engines and devise their pages accordingly. In this case, the resources of search engines are lost, and the time of searching in response to user query increases. It degrades the quality of search results, thus wasting user's time [1, 2, 3, 4].

This study indicates that the degree of search engine spam is on alarming increase. Researchers prepared the most familiar search query terms used in search engines between September 2010 and April 2011, and researchers found that, on average, 50% of them return results with malicious links [5].

Recently, the trend of spam attack has increased because anybody may simply write spam reviews and post them to e-commerce websites without any constraint. Any company might hire individuals to write false assessment for their products and services [34].

II. LITERATURE REVIEW

A. Survey on Search Engine Spam

Search engine spam is an insertion of unnaturally formed pages into the web in order to manipulate search results of search engines, in order to make traffic to certain pages for marketable profitability [7]. A variety of search engine spam strategies are used by the spam creators to manipulate content and links of web page to boost ranking in search engine. All these strategies are challenging and are classified as content spam and link spam in web pages. While the content spam refers to the various strategies that are functioned to manipulate contents of webpage, link spam creates numerous link between the web pages to boost ranking [3, 6].

The content spam was detected through context analysis with the help decision tree C4.5 classifier with minimum web pages features [7]. The spam pages were found certain low quality web page features that are recognized [8]. Researcher identified web pages features took less computing resources for checking authenticity of web pages than ranking algorithm [8]. The machine learning technique produced improved results than other techniques [9].

Link spam refers to unnatural manipulation of links in the web page to boost ranking. Search engine industry applies different ranking algorithms to identify link spam. However these ranking algorithms were found to slightly boost the computational cost than machine learning technique [10, 8].

B. Survey on Preprocessing

Preprocessing is an essential step in the machine learning process to improve classification results. This step is required to decide various types of problems such as noisy data, redundant data, and missing data values. All the machine learning algorithms depend greatly on the product of this stage, which is the final training set. The web page features has been extracted from the web page and stored in feature database.

Revised Version Manuscript Received on August 19, 2019.

Dr. D. Saraswathi, Computer Science, PSG College of Arts & Science, Coimbatore, Tamilnadu. India.

Dr. A. Krishna kumar, Computer Science, PSG College of Arts & Science, Coimbatore, Tamilnadu, India.

Those features have been normalized before training and testing processes in order to make sure that data are overwhelmed by each other in terms of the distance measure. Normalization is mainly helpful for classification, as it progress accuracy and efficiency of mining. There are three types of normalization techniques namely Z-Score Normalization [11], Decimal Point Normalization [12], and Min-Max Normalization [13]. In Z(Zero Mean)-Score Normalization, the data is normalized based on the mean and standard deviation. Decimal scale normalization is based on the movement of decimal point of value of attribute. The decimal point numbers are moved depending on the greatest absolute values of attribute. A comparative study of various normalizations is given in Table I.

Table I. Comparative study of various normalization techniques.

Parameters	Z-Score	Decimal Point	Min-Max
Misclassification Error	More	More	Less
Preserves relationships among original value	No	No	Yes
Computational Time	More	More	Less
Accuracy	Very Less	Less	More

The min-max normalization technique performs a linear transformation on the numerical data. It has less misclassification errors when compared to Z-score normalization and decimal scaling normalization. The min-max normalization technique has taken a lesser computational time when compared to other normalization techniques [11, 12, 14, 15].

After a better understanding of the strengths and limitations of each method, min-max normalization technique has been selected for the present study from comparative studies of different normalization techniques.

C. Survey on Classification

Classification is data mining method which is used to classify data items in group of occurrences [16]. Numerous machine learning algorithms are applied to classification task. Machine learning method is better than other methods for classifying search engine spam [8, 9]. After getting normalized data, training phase is used to build model by classifier. Finding spam pages is viewed as supervised classification problem. In the supervised classification, the search engine spam classifier needs to be trained with a set of previously classified pages. Researchers considered various classification methods such as K-Nearest Neighbour [17], Back Propagation [18], Naive Bayes [19], Support Vector Machine [20], and Decision Tree [21] in their analysis for classification involving search engine spam detection [22, 23].

The different classifiers comparative study is listed in Table II. After a better understanding of the strengths and limitations of each method from comparative studies of different classifiers, C4.5 decision tree algorithm has been selected to identify search engine spam.

After constructing the model, it has been validated using cross-validation. Among the cross-validations are observed that, 10-fold cross-validation is broadly used for assessing the decision tree model and outperform the other cross-validations [24, 25].

Among the classifiers listed, C4.5 algorithm is a decision tree classifier, given in top ten most prominent data mining algorithms [26]. A comparison of the performances of various classifiers reveals the decision tree C4.5 algorithm offered a better performance with respect to spam detection when compared to other classifiers [7, 27, 21, 28].

III. PROPOSED ARCHITECTURE

The proposed study consists of three phases for identifying search engine spam. Those phases are preparatory phases, web page feature collection, and decision tree C4.5 classification. The proposed study is depicted the flow of working mechanism in Fig.1.

Sample web pages were fetched manually from the search engines and accumulated in a webpage repository. A file selector was used to select those web pages to check authenticity of the web page. The collection of keywords and links in the chosen web pages were extracted and stored in web data depository. The features of content and links are Number of popular search keyword in the page, popular search keyword in title, popular search keyword in anchor text, popular search keyword in meta tag, popular search keyword in H1 tag, popular search keyword in H2 tag, Specific popular search keyword repetition, and Number of advertising keywords. The content spam and link spam features were identified weightage based on search volume of popular search keywords, popular links and advertising keywords for detecting spam pages. Similarly, the file selector was selected all the web pages in the webpage repository, was extracted features that were stored in web data depository. These features were normalized using min-max normalization.

From the insights observed during the literature review, the decision tree C4.5 algorithm has been chosen for classifying web pages. Decision tree C4.5 classifiers give a hierarchical decomposition of the training data and are used to learn the rules to identify the authenticity of web pages. A tree is formed by using different web page features which are listed and their values. Information gain is calculated by using a list of web page features. The feature that has highest information gain is used as the root node of tree model. The interior nodes of the decision tree are labeled with unique features and these features have low information gain as compared to the root node. This procedure is repeated until all reviews are classified as spam or not-spam web pages. Finally classification results have been submitted to the user interface.

Table II. Comparative study of various classifiers.

Parameters	Naive Bayes	Back propagation	SVM	KNN	C4.5
Requirement for Domain Knowledge	Yes	Yes	Yes	No	No
Accuracy	Less	Less	Satisfactory	More	More
Handling High Dimensional Data	Yes	Yes	Yes	Yes	Yes
Need for Predefined Data	Yes	Yes	Yes	Yes	Yes
Computation Effort	More	Less	More	Less	Less
Decrease in Performance when Dataset is small	Yes	No	No	No	No
Training	More	More	Less	More	Less
Interact with Features	No	Yes	Yes	Yes	Yes
Requirement of Memory	Less	More	More	More	Less
Training Data	More	More	Less	More	Less
Works Based Majority Voting Neighbours	No	No	No	Yes	No
Irrelevant Features Affect Performance	Yes	Yes	Yes	Yes	Yes
Possibility of Parallel Implementations	Less	More	More	More	More
Dynamic Updating of Training Patterns	No	No	Yes	Yes	Yes
Classification Time	Less	More	More	Less	Less

Proposed Algorithm

Web pages, popular search keywords, popular links and ads words collection

Assign weight for popular search keywords, Popular links and advertising words based on search volume

Features identification from the web pages

Normalize the web page feature values using min-max normalization

Classify web pages using decision tree C4.5

The formulas [1, 2, 3, 4, 5, 6] used for the C4.5 decision tree algorithm are

$$INFO(D) = -\sum_{i=1}^{c} P_i \log_2(P_i) \qquad (1)$$

Where

Info(D) = Expected information needed to classify a tuple in data set

Pi = Probability that an arbitrary tuple in D belongs to distinct classes

C = Distinct classes

D = Dataset

$$mid\_point = \frac{a_i + a_{i+1}}{2} \qquad (2)$$

Where

Mid_point(A) = Middle point between the sorted values ai and ai+1 of attribute A

A<= mid_point, A>mid_point

$$MINIMAL\ INFO_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} INFO(D_j) \qquad (3)$$

Where

InfoA(D) = Expected information required for classifying tuple from dataset based on the partitioning by attribute A

v = Data partition of attribute

D = Total number of tuples in dataset

$$GAIN(A) = INFO(D) - INFO_A(D) \qquad (4)$$

Where

Gain (A) = Information Gain of A

Info(D) = Expected information needed to classify a tuple in data set

InfoA(D) = Expected information required for classifying tuple from dataset based on the partitioning by attribute A

$$SPLITINFO_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} \log_2(\frac{|Dj|}{|D|}) \qquad (5)$$

Where

SPLITINFOA(D) = Information Gain of A using split information

V = Data partition of attribute A

D = Total number of tuples in dataset

$$GAINRATIO(A) = \frac{GAIN(A)}{SPLITINFO(A)} \qquad (6)$$

Where

GainRatio(A) = Maximum gain ratio is selected

Gain(A) = Information Gain of A

SplitinfoA(D) = Information Gain of A using split information

IV. DATA COLLECTION AND SAMPLING

The web pages for the present study were collected from webspam-uk 2007 dataset, which is a publically available collection of pages. These web pages collection is labeled at the host level by a group of volunteers and hosts were marked as "spam" and "nonspam" by the evaluator [29]. 10000 web pages were collected for the present study out of which 8000 pages are non spam and 2000 pages described as spam as per webspam-uk2007.

This study studied web pages of various sizes namely 2500 pages, 5000 pages, 7500 pages, and 10000 pages. All web page samples were tested with different iterations, which were then used to train the model. This model was then used to predict if the web pages were spam or non spam. The present study was applied cluster sampling method. In cluster sampling method, whole sample is divided into groups or clusters, and random samples of these clusters are chosen [30]. Web pages in equal sizes were collected and formed as clusters such as Business, Consumer, Finance, Health & Wellness and Miscellaneous keywords. The online terms marketing company listed popular search keywords with search volume [31] and sample keywords are given below in Table III. In addition, popular keywords and popular links were also collected for the study [32, 33].

Table III. Sample popular search keywords.

Keywords	Search Volume
Income tax	1220000
Quickbooks	823000
Account	823000
Audit	246000
Payroll	165000

V. RESULTS AND DISCUSIION

Confusion matrix is used to evaluate performance of a machine learning classifier. This matrix metrics are

represented in various measures which are applied on test sample.

Table IV presents the structure of a confusion matrix for a two-class problem, with two different class namely positive occurrence and negative occurrence. A row indicates an actual class, while a column indicates the predicted class. The researchers used four parameters of confusion matrix while assessing the quality of the algorithms. These attributes namely are True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate, Accuracy Rate and Error Rate [35]. Evaluation measures are given below [7, 8, 9, 10, 11, and 12].

Table IV. Confusion matrix

Total No.of Instances		Predicted Class	
		Non Spam	Spam
Actual Class	Non spam	TP	FN
	Spam	FP	TN

Where

TP = Number of correctly classified positive occurrences

TN = Number of correctly classified negative occurrences

FP = Number of incorrectly classified as positive occurrences

FN= Number of incorrectly classified as negative occurrences

True Positive Rate (TPR) or Sensitivity, also known as Recall Rate measures the proportion of positives occurrence that are correctly classified. The formula is given as follows

$$True\,Positive\,Rate = \frac{TP}{(TP+FN)} \qquad (7)$$

True Negative Rate (TNR) or Specificity measures the proportion of negatives occurrence that are correctly classified. The formula is given as follows

$$True\,Negative\,Rate = \frac{TN}{(TN+FP)} \qquad (8)$$

False Positive Rate (FPR) measures the proportion of incorrectly classified as positive occurrences. The formula is given as follows

$$False\,Positive\,Rate = \frac{FP}{(FP+TN)} \qquad (9)$$

False Negative Rate (FNR) measures the proportion of incorrectly classified as negative occurrences. The formula is given as follows

$$False\,Negative\,Rate = \frac{FN}{(TP+FN)} \qquad (10)$$

Accuracy (AR) is defined as the ration between the total number of correctly classified occurrences and the total number of occurrences. The formula is given as follows

$$Accuracy\,Rate = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (11)$$

Error Rate (ER) is the ratio between incorrect classified occurrences and all of the occurrences. The formula is given as follows

$$Error\,Rate = \frac{(FN+FP)}{(TP+TN+FP+FN)} \qquad (12)$$

The comparative analyses of various classifiers have been compared and results are shown in Table V. Among the classifiers given, Decision Tree (DT) C4.5 classifier outperforms other classifiers.

DT C4.5 classifier accuracy rate is compared with various classifier accuracy rate. This comparison is shown in Fig.2. Among the given classifiers, C4.5 DT classifier outperforms other classifier.

The rate of classification error is compared with given classifiers and is represented in Fig.3. Among the given classifiers, C4.5 DT has been observed to show a minimum error rate when compared to other classifiers.

Table V. Comparative study of various classifiers.

Classification Algorithm	Accuracy Rate	Error Rate
NB	0.7918	0.2082
SVM	0.8515	0.1485
BPN	0.8231	0.1769
KNN	0.8525	0.1475
DT C4.5	0.8817	0.1183

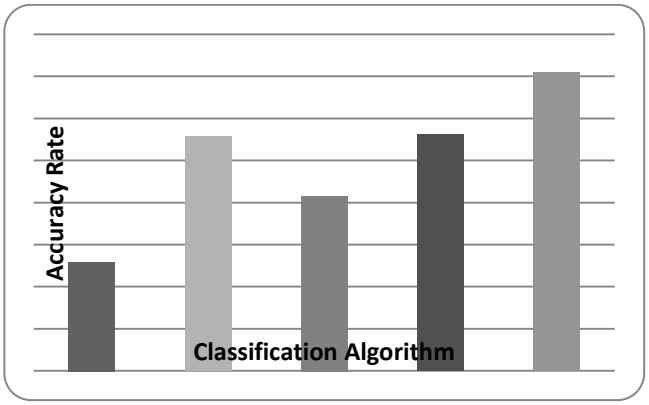

Fig.2. Various classifiers accuracy rate

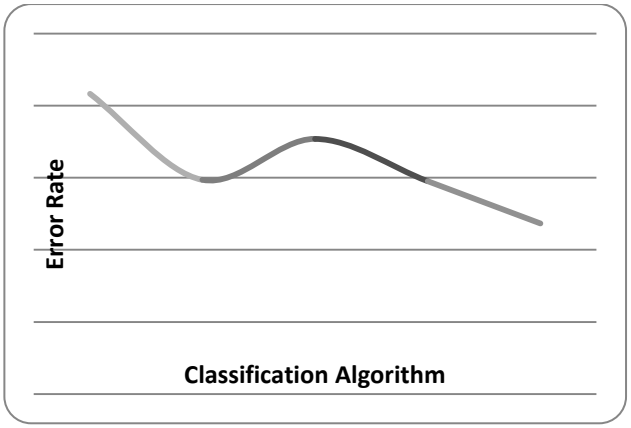

Fig.3. Various classifiers Error rate

VI. RESULTS AND DISCUSIION

This study had been used popular keywords, popular links and advertising word with search volume for spam classification. It assigned weightage for those popular keywords, popular links and advertising based on range of search volume. This criterion has been useful to fast identification of search engine spam when compared with existing system. These content and link features have been classified webpages using different classifiers. The DT C4.5 classifier, as discovered from the findings, shows a better performance when compared to other classifiers. However, it consumed a long time for constructing model especially when the dataset is large. Even though this study classified spam pages effectively.

REFERENCES

1. M.R.Henzinger,R.Motwani, and C.Silverstein. "Challenges in web search engines". SIGIR Forum, 36, September 2002.
2. D.Fetterly, M.Manasse, M.Najork, "Spam, Damn Spam, And Statistics: Using Statistical Analysis To Locate Spam Web Pages", In Proceeding of the Seventh Workshop on the Web and Databases, pp.1-6, 2004.
3. Z.Gyongyi and H.Garcia-Molina. "Web Spam Taxonomy". Proceeding first international Workshop on Adversarial Information Retrieval on the Web, Japan, May 2005
4. C.Castillo, B.D.Davison," Adversarial Web Search", Information Retrieval, vol. 4, pp.377-486, 2010.
5. L.Lu, R.Perdisci, W.Lee, "SURF: detecting and measuring search poisoning", In Proceedings of the 18th ACM Conference on Computer and Communications Security, pp.467-476,USA, 2011.
6. MahdiehDanandehOskuie, SeyedNaserRazavi, "A Survey of Web Spam Detection Techniques", International Journal of Computer Applications Technology and Research (2319-8656), Volume 3-Issue 3, 180 -185, 2014.
7. AlexandrosNtoulas, marc najork, mark manasse, Dennis fetterly, "Detecting Spam Web Pages through Content Analysis", International World Wide Web Conference Committee, 2006.
8. Ashish Chandra, Mohammad Suaib, and Dr. Rizwan Beg, "Low Cost Page Quality Factors to Detect Web Spam", Informatics Engineering, An International Journal, Vol.2, No.3, September 2014.
9. Kanchan Hans, LaxmiAhuja, S.K. Muttoo, " Approaches for Web Spam Detection" International Journal of Computer Applications (0975-8887) Volume 101, No.1 September 2014.
10. D.Fetterly, M.Manasse, M.Najork, "Spam, Damn Spam, And Statistics: Using Statistical Analysis To Locate Spam Web Pages", In Proceeding of the Seventh Workshop on the Web and Databases, pp.1-6, 2004
11. S.B. Kotsiantis, D.Kanellopoulos, P.E. Pintelas, "Data Preprocessing in Supervised Learning",Computer Science, vol. 1,pp.1306-4428, 2006.
12. L.AlShalabi, Z.Shaaban, "Normalization as a Preprocessing Engine for Data mining and the Approach of Preference Matrix", In Proceeding of International Conference in Dependability of Computer Systems, pp.207-214, 2006
13. N.A.Husin, N.Salim, and A.R.Ahmad, "Modeling of dengue outbreak prediction in Malaysia: A Comparison of Neural Network and Nonlinear Regression Model", Proceeding of International Symposium in Information Technology, 2008
14. ZurianiMustaffa, YuhanisYusof, "A Comparison of Normalization Techniques in predicting Dengue outbreak", In Proceeding of International conference on Business and Economics Research, Vol. 1 Malaysia, 2011
15. Saranya C, Manikandan G, "A Study on Normalization Techniques for Privacy Preserving Data Mining", International Journal of Engineering and Technology, Vol 5, 2013.
16. Mike Chapple, "Classification", Database Expert, http://databases.about.com/od/datamining/g/classification.htm
17. Oliver Sutton,"Introduction to K NearestNeighbour Classification and Condensed Nearest Neighbour Data Reduction", 2012.
18. W.A.Awad,S.M.Elseuofi,"Machine Learning Methods for Spam Email Classification", In Proceeding of International Journal of Computer Science and Information Technology, Vol 3, 2011
19. Sebastian Raschka," Naive Bayes &Text classification", 2014.
20. Jzhang,"A Brief Introduction to Support Vector Machine", lecture notes, 2011.
21. Victor M. Prieto , Manuel Alvarez, Rafael Lopez-Garcia and Fidel Cacheda, "Analysis and Detection of Web Spam by means of Web Content", In Proceedings of the 5th Information Retrieval Facility Conference, 2012
22. Prieto V et al ,"Analysis and Detection of Web spam by Means of Web Content " In Multidisciplinary Information Retrieval, Springer Berlin Heidelberg, PP.43-57,2012
23. Silva R.M, Almeida T.A, and Yamakami, "An Analysis of machine learning methods for spam host detection", 11th International Conference on Machine Learning and Applications,pp.227-232, IEEE, 2012.
24. PayamRefaeilzadeh, Lei Tang, Huan Liu," Cross Validation" Arizona State University, 2008
25. Sylvain Arlot, "A survey of cross-validation procedures for model selection" Vol. 4, pp: 40-79, 2010.
26. Wu X. V.Kumar, J.R.Quinlan, J.Ghosh, Q.Yang, H.Motoda, McLachlan, A.Ng, B.Liu, P.S.Yu, Z.Zhou, M.Steinvach, D.J.Hand, and D.Steinberg, "Top 10 algorithms in data mining", Knowledge and Information Systems, DOI: 10.1007/s10115-007-0114-2, 2008.
27. Manuel Egele, Clemens Kolbitsch, Christian Platzer, "Removing Web Spam Links from Search Engine Results", Journal computing virol, Springer, 2011.
28. Nimithasafar, KalavathiK,"Performance Comparison between Naïve Bayes, Decision Tree, and K-Nearest Neighbour", International Journal of Emerging Research in Management & Technology. ISSN:2278-9359,Vol-4, Issue-6, 2015
29. Web Spam UK 2007, http://chato.cl/webspam/datasets/uk2007/.
30. Singh, Ajay S, Masuku, Micah B," Sampling Techniques and Determination of Sample Size in Applied Statistics Research: An Overview", International Journal of Economics, Commerce and Management, Vol. II, Issue 11, Nov 2014.
31. http://www.wordstream.com/popular-keywords/
32. www.pagetraffic.com/blog/most-popular-keywords-on-search-engines
33. https://moz.com/top500
34. NaveedHussain, Hamid TurabMirza , GhulamRasool , IbrarHussain and Mohammad Kaleem," Spam Review Detection Techniques: A Systematic Literature Review ", Applied Sciences 2019, 9, 987; doi:10.3390/app9050987
35. Ashish Chandra, Muhammad Suhib, Dr.RizwanBeg,"Web Spam Classification using Supervised Artificial Neural Network Algorithms", An International Journal Advanced Computer Intelligence,Vol. 2, No.1 2015.

AUTHORS PROFILE

Dr.D.Saraswathi has 12 years of teaching experience. She has completed her Master degree in Software Science (5 years integrated) at Periyar University, Salem. She did her M.Phil&Ph.D (Computer Science) at Periyar University, Salem. She has published papers in Elsevier, IEEE Xplore, SCI Index, Scopus, Google Scholar and Springer. Her Google scholar citations are 36. She has published 14 research papers in International Journals. She has published 34 Papers in various National and International Conferences. She has attended 26 workshops and Faculty Development Programme. She also guided numerous PG and UG students for their projects. She is a member of International Association of Engineers (IAENG), and ICTACT. Her areas of specialization are web mining and text mining.

Dr.A.Krishnakumar has 6 years of teaching experience. He has completed his Master degree in Computer Science at Bharathiar University, Coimbatore. He did his M.Phil at KarpagamUniversity,Coimbatore; Ph.D (Computer Science) at BharathiarUniversity,Coimbatore. He has published papers in IEEE Xplore, ESCI Index, Scopus and Google Scholar. He has published 5 research papers in International Journals. He has published 8 Papers in various National and International Conferences. He also guided numerous PG and UG students for their projects. She is a member of Institute of Electrical and Electronic Engineers (IEEE), and IFERP. His area of specialization is Wireless Sensor Networks.