FEATURE SELECTION-MODEL-BASED CONTENT ANALYSIS FOR COMBATING WEB SPAM

Shipra Mittal[1*] and Akanksha Juneja[2]

Department of Computer Science & Engineering,
National Institute of Technology, Delhi, India
*Corresponding Author at - Department of Computer Science & Engineering,
National Institute of Technology, Delhi, India
[1]mittal.shipra90@gmail.com,
[2]akankshajuneja@nitdelhi.ac.in

ABSTRACT

With the increasing growth of Internet and World Wide Web, information retrieval (IR) has attracted much attention in recent years. Quick, accurate and quality information mining is the core concern of successful search companies. Likewise, spammers try to manipulate IR system to fulfil their stealthy needs. Spamdexing, (also known as web spamming) is one of the spamming techniques of adversarial IR, allowing users to exploit ranking of specific documents in search engine result page (SERP). Spammers take advantage of different features of web indexing system for notorious motives. Suitable machine learning approaches can be useful in analysis of spam patterns and automated detection of spam. This paper examines content based features of web documents and discusses the potential of feature selection (FS) in upcoming studies to combat web spam. The objective of feature selection is to select the salient features to improve prediction performance and to understand the underlying data generation techniques. A publically available web data set namely WEBSPAM - UK2007 is used for all evaluations.

KEYWORDS

Web Spamming, Spamdexing, Content Spam, Feature Selection & Adversarial IR

1. INTRODUCTION

As the scope of web grows beyond limits, it is more prone to profanation. From accessing information to interacting and connecting with people, from e-commerce to e-businesses, Internet covers almost each and every aspect of our lives. It helps in bringing new opportunities to people. According to Sam Lucero, analyst at ABI Research in Oyster Bay, "anything intelligent would have an online presence" [1]. But, as it is said, every massive technology has its own benefits and challenges, same is the case with Internet and World Wide Web. Accurate and quality information retrieval is one of those major challenges. As business vendors recognize the value of web for reaching out to millions of customers, they try to gain high visibility for their websites on search engine result page (SERP). This rising need to rank highly in search results in order to recognize among web users, gives birth to the term *web spamming* (or, *spamdexing*) [2]. Spamdexing, as the name implies, takes advantage of web indexing system, allowing spammers to deceive search engine (SE) ranking of specific documents.

David C. Wyld et al. (Eds) : CSITY, SIGPRO, AIFZ, NWCOM, DTMN, GRAPHHOC - 2016
pp. 27–34, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60403

Ranking system of SEs involves various content-based and graph-based measures. Spammers exploit these parameters to artificially inflate the ranking of web documents. Spam techniques range from stuffing a page with large number of authority references or popular query keywords, thereby causing the page to rank higher for those queries, to setting up a network of pages that mutually reinforce their page value to increase the score of some target pages or the overall group.

Recently [3; 4], all major SEs such as Google, Yahoo etc. have identified web spam as a tangible issue in IR process. It not only deteriorates the search quality but also cause wastage of computational and storage resources of a SE provider. A financial loss of $50 billion was caused due to spam in the year 2005 [5]. In the year 2009, it was estimated at $130 billion [6]. Further, it weakens people's trust and might deprive legitimate websites of user's visibility and revenue. Therefore, identifying and combating spam becomes a top priority for SE providers.

According to web spam taxonomy presented in the work of Spirin and Han [7], web spam is broadly classified into four categories namely content spam [8], link spam [9; 10], cloaking and redirection [11; 12], and click spam [13]. This research work primarily focuses on the detection of content spam which is the most common and frequently occurring spam [14].

IR systems examine the content of pages in the corpus to retrieve the most relevant document with respect to a specific search query. "Term Frequency-Inverse Document Frequency" (TF-IDF) or another similar approach is used to access the "most similar" (relevant) documents to the search query. In TFIDF, "the relevance of the search terms to the documents in corpus is proportional to the number of times the term appeared in the document and inversely proportional to the number of documents containing the term." Spammers exploit Term Frequency (TF) scoring by overstuffing content fields (title, body, anchor text, URL etc.) of a page with a number of popular search terms so as to boost its relevancy score for any search query. It can be measured as:

$$TFIDF\,(q,p) = \sum\nolimits_{(t\in q)\wedge(t\in p)} TF\,(t).\,IDF\,(t) \tag{1}$$

where *q* refers to query, *p* denotes a web page in the corpus, and *t* denotes the term.

Machine learning is a field of study that deals with automated learning of patterns, within the data belonging to different classes or groups, with an aim to differentiate between the classes or groups. An effective machine learning algorithm is expected to make accurate predictions about categorization of unseen data based on the learnt patterns. Specifically, supervised machine learning involves predicting the class of an unseen (new) data sample based on the decision model learnt using the existing (training) data. Therefore, knowledge of machine learning may be appropriately utilized for web spam detection.

Several machine learning methods to combat content spam were introduced in the past researches of adversarial IR and web spam domain. Egele et al. [15] examined the importance of different text and link metrics in web ranking system and utilize C4.5 supervised learning algorithm to remove spam links. They deployed their own mechanism to generate data to carry out the experiment.

Ntoulas et al. [16] presented an approach for detecting spam based on content analysis. They extracted several content features and presented a comprehensive study about the influence of these features in web spam domain.

Prieto et.al [17] suggested a number of heuristics to identify all possible kinds of web spam and developed a system called SAAD (Spam Analyzer and Detector) for efficient web spam detection. The beneficial trait is that the system was tested on different data sets and proved to be effective for more accurate classification of spam pages.

Araujo and Romo [18] proposed an interesting approach of spam detection by comparing the language models (LM) [19] for anchor text and pages linked from these anchor text. KL-divergence [20] was used to measure the discrepancy between two LMs.

However, spammers are continuously adapting themselves to circumvent these barriers. This research work presents a content-based spam detection approach using supervised learning. The aim of this study is to draw a clear understanding of underlying process of web spamming by examining already extracted features. Moreover, the work aims at selecting salient features from the existing ones to stimulate further studies in the domain of "adversarial information retrieval [21]". A filter based feature selection technique is employed to uncover important patterns in order to classify websites as spam or ham (non-spam). The proposed method is observed to be efficient in terms of both computational complexity and classification accuracy.

The rest of the paper organizes as follows. A general methodology of applying feature selection technique for adversarial classification is presented in section 2. Finally, experimental results are shown in section 3 and section 4 concludes the paper.

2. METHODOLOGY

2.1 Experimental Data Set

Widely known web spam data set WEBSPAM-UK2007 [22] is used to carry out the experiments. This dataset was released by Yahoo especially for Search Engine Spam project. The data set was also used for "Web Spam Challenge 2008". It is the biggest known web spam data set having more that 100 M web pages from 114,529 hosts. However, only 6,479 hosts were manually labelled as spam, non-spam and undecided. Among these, approx 6% hosts are spam, i.e., data is imbalanced in nature. The data set consist a separate training and testing data set, but we combined the two sets together to evaluate our model since the percentage of spam hosts were small in both of them. Further, we neglect the hosts labelled as "undecided" and conduct our experiment for a group of 5,797 hosts. The training set was released with pre-extracted content feature set which are examined in this study to select salient and optimal features.

Existing content- based heuristics for detecting web spam

The content feature set proposed by Ntoulas et al. [16] comprises of 98 features based on following heuristics:

- Number of words in the page: "Term Stuffing" is a common spamming technique to increase visibility of a web document on typical queries or search terms. Sometimes the proportion of common terms in a page is very high. Therefore, authors suggest counting number of words per page. Very large value of the proposed heuristic indicates the strains of spam in the page.
- Number of words in the title: Many times, a page title is stuffed with unrelated keywords and terms because of its high weightage in search engines text metrics. As a spam detection method, authors propose measure of number of words in the title
- Average word length of the document: In order to combat composite keywords spamming, authors propose to measure the average word length

- Fraction of anchor text in the page: Due to the fact that "anchors" are used to describe the association between two linked pages, spammers misuse them to create false associations

$$fraction\ of\ anchor\ text\ per\ page = \frac{total\ number\ of\ words\ in\ the\ page}{number\ of\ words\ in\ anchors} \tag{2}$$

- Ratio of visible text: The authors propose this heuristic to detect "hidden content" in a page
- Compression Ratio: The proposed features helps in determining the level of redundancy in the page

$$compression\ ratio = \frac{size\ of\ normal\ web\ page}{size\ of\ compressed\ page} \tag{3}$$

- Independent and Dependent LH: These techniques utilize the independent and dependent n-grams probability to detect spam. More precisely, content of each page is divided into n-g of *n* consecutive words to calculate the probability of document by individual n-g probabilities. However, this feature is computationally expensive.

The performance analysis of these features and its comparison after applying feature selection techniques is presented in section 3.

2.2 Proposed Methodology

This research work focuses on the contribution of feature selection in adversarial IR applications. The common issues in the spam domain are listed as: small sizes of samples, unbalanced data set and large input dimensionality due to several pages in a single web document. To deal with such problems, a variety of feature selection methods have been designed by researchers in machine learning and pattern recognition. This work employs univariate filter feature selection to improve the prediction performance of decision model. The existing heuristics on which feature selection is performed are already discussed in previous subsection. The idea of applying feature selection in the existing features is two-fold: these features were utilized in many previous studies [16; 17; 18; 23] for effective spam detection; the heuristics recognized as baseline for further studies in the underlying domain.

Due to aforementioned reasons, it can be expected that, for spam documents classification, feature selection techniques will be of practical use for later researches in information retrieval and web spam.

2.2.1 Methods for Web Spam Detection

This section describes the algorithms and methods used for evaluation of this work.

Classification Algorithm

For the appropriate prediction, we have tried various classification methods, namely, *k*-nearest neighbour, linear discriminant analysis, and support vector machine (SVM). As per experiments, SVM is observed to achieve better results for binary classification in comparison to other classifiers. Therefore, this research work SVM is utilised to learn decision model.

SVM [24] classify objects by mapping them to higher dimensional space. In this space, SVM train to find the optimal hyperplane, i.e., the one with the maximum margin from the support vectors (nearest patterns).

Consider a training vector $\mathbf{x}_i, i = 1, 2, \ldots, n,$ we define a vector **y** such that $y_i = \{1, -1\},$ decision function can be defined as:

$$f(\mathbf{x}) = sign\ (\mathbf{w}^{\mathrm{T}}\mathbf{x} + b) \tag{4}$$

The weight vector **w** is defined as:

$$\mathbf{w} = \sum_i \alpha_i \, y_i \mathbf{x}_i \qquad\qquad \alpha_i \geq 0 \quad \forall i \tag{5}$$

where α_i is the Lagrange's multiplier used to find the hyperplane with maximum distance from the nearest patterns. Patterns for which $\alpha_i > 0$ are support vectors.

The possible choice of decision function when data points are not separable linearly can be expressed as:

$$\min_{\mathbf{w},\mathrm{b},\alpha_i} \ \frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\mathbf{w}\right) + \ C \sum_{i=0}^{n} \alpha_i \tag{6}$$

where, $0 \leq \ \alpha_i \ \leq C, \quad i = 1, 2 \ \ldots., n$ and *C* is the penalty parameter. Value of *C* =10 is used for experimental evaluation.

Performance Evaluation

In order to evaluate the model, 10-fold cross validation technique is used. The results are shown in terms of classification accuracy, sensitivity, and specificity, whose values are obtained by analysing the cost matrix [25].

Sensitivity can be defined as the number of actual positive instances (spam) that are correctly predicted as positive by the classifier. Conversely, specificity determines proportion of actual negatives (non-spam hosts) that are correctly classified. Accuracy can be defined as total number of instances that are correctly predicted by the classifier.

Univariate Filter Feature Selection: Simple yet efficient

In order to improve the performance of SVM, feature selection is implemented. Univariate filter based feature selection has been utilised due to the fact that it is computationally simple, fast, efficient, and inexpensive. In filter based feature selection, "features are selected independently to induction algorithm" [26; 27; 28]. The measurement for feature selection is chosen as Mutual Information Maximization (MIM) [29]. It is a simple method to estimate rank of features based on mutual information. Mutual information is defined as a relevancy measure, determining how relevant a feature is for corresponding targets. The criterion can be expressed as follows:

$$J_{MIM}(\mathbf{x}_n) = \max\left[\, I(\mathbf{x}_n; \mathbf{c})\right] \tag{7}$$

where $\mathbf{x}_n$ is the n^{th} feature of training matrix **X**, **c** is the class label and $I(\mathbf{x}_n; \mathbf{c})$ refers to mutual information between the two. Mutual Information between two random variables p and q can be determined as:

$$I(p; q) \ = \sum P(p, q) log \frac{P(p,q)}{P(p)P(q)} \tag{8}$$

3. EXPERIMENTAL RESULTS AND DISCUSSION

Table 1 shows the performance of SVM on existing feature set whereas Table 2 shows the prediction performance after feature selection. It is clearly visible that feature selection technique on precompiled measures outperforms the performance of complete feature set. The results show a significant gain in classifier's accuracy in terms of both valuation measures (i.e., specificity and sensitivity). Approximately 3% increase in specificity and 2% increment in accuracy and sensitivity is reported.

Table 1: Performance of content- based feature sets using SVM

Feature Set	Performance Measure (in percentage)		
(98 features)	Accuracy	Sensitivity	Specificity
Content	79.9	61.8	79.2

Table 2: Performance of content- based feature sets after feature selection using SVM

Filter based Feature Selection		Number of features								
		Top 10	Top 20	Top 30	Top 40	Top 50	Top 60	**Top 70**	Top 80	Top 90
Performance Measure (in percentage)	Accuracy	76.7	78.6	81.6	81.6	81.4	81.1	**82.1**	80.8	80.7
	Sensitivity	46.6	51.2	51.3	55.3	55.3	57.1	**63.1**	63.1	62.7
	Specificity	81.5	81.6	81.6	82.6	82.7	82.7	**82.9**	80.8	80.7

4. CONCLUSIONS AND FUTURE PERSPECTIV

In this study, we take into account existing heuristics for detecting spam by means of content analysis. This experiment compares the performance results of pre-determined features with the performance of features achieved after feature selection. The experimental results demonstrate that classifier performance increases with reduced (reduced from 98 features to 70 features) set of salient features. Furthermore, we believe that feature selection undermines the inherent risk of imprecision and over-fitting caused due to unbalanced nature of dataset. However, a robust and optimal feature selection model is still a need to uncover.

Multivariate feature selection and wrapper based feature selection can be addressed as a prominent future study in web spam community. A second line of future research will be extension of heuristics extracted using both content analysis and web graph mining. Other interesting opportunities oriented towards different machine learning approaches such as fuzzy logic, neural network etc. Since, there is no clear separation between spam and ham pages, i.e., definition of spam may be vary from one person to another, use of fuzzy logic can be seen as a promising line of future work in detection of web spam.

REFERENCES

[1] Wood, "Today, the Internet -- tomorrow, the Internet of Things?," Computerworld, 2009. [Online].Available: http://www.computerworld.com/article/2498542/internet/today--the-internet----tomorrow--the-internet-of-things-.html.

[2] Z. Jia, W. Li and H. Zhang, "Content-based spam web page detection in search engine," Computer Application and Software, vol. 26, no. 11, pp. 165-167, 2009.

[3] M. Cutts, "Google search and search engine spam," Google Official Blog, 2011. [Online]. Available: https://googleblog.blogspot.in/2011/01/google-search-and-search-engine-spam.html.

[4] M. McGee, "businessWeek Dives Deep Into Google's Search Quality," Search Engine Land, 2009. [Online]. Available: http://searchengineland.com/businessweak-dives-deep-into-googles-search-quality-27317.

[5] D. Ferris, R. Jennings and C. WIlliams, "The Global Economic Impact of Spam," Ferris Research, 2005.

[6] R. Jennings, “Cost of Spam is Flattening - Our 2009 Predictions,” 2009. [Online]. Available: http://ferris.com/2009/01/28/cost-of-spam-is-flattening-our-2009- predictions/.

[7] N. Spirin and J. Han, "Survey on web spam detection: principles and algorithms," SIGKDD Explor. Newsl., vol. 13, no. 2, p. 50-64, 2012.

[8] C. Castillo, K. Chellapilla and B. Davison, "Adversarial Information Retrieval on the Web," Foundations and trends in Information Retrieval, vol. 4, no. 5, pp. 377-486, 2011.

[9] S. Chakrabarti, Mining the Web. San Francisco, CA: Morgan Kaufmann Publishers, 2003.

[10] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini and S. Vigna, "A reference collection for web spam," ACM SIGIR Forum, vol. 40, no. 2, pp. 11-24, 2006.

[11] J. Lin, "Detection of cloaked web spam by using tag-based methods," Expert Systems with Applications, vol. 36, no. 4, pp. 7493-7499, 2009.

[12] A. Andrew, "Spam and JavaScript, future of the web," Kybernetes, vol. 37, no. 910, pp. 1463-1465, 2008.

[13] C. Wei, Y. Liu, M. Zhang, S. Ma, L. Ru, and K. Zhang, "Fighting against web spam: a novel propagation method based on click-through data," In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 395-404, 2012.

[14] H. Ji and H. Zhang, "Analysis on the content features and their correlation of web pages for spam detection", China Communications, vol. 12, no. 3, pp. 84-94, 2015.

[15] M. Egele, C. Kolbitsch and C. Platzer, "Removing web spam links from search engine results," Journal in Computer Virology, vol. 7, no. 1, pp. 51-62, 2011.

[16] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," In Proceedings of the 15th international conference on World Wide Web, ACM, pp. 83-92, 2006.

[17] V. Prieto, M. Álvarez and F. Cacheda, "SAAD, a content based Web Spam Analyzer and Detector," Journal of Systems and Software, vol. 86, no. 11, pp. 2906-2918, 2013.

[18] L. Araujo and J. M. Romo, "Web spam detection: new classification features based on qualified link analysis and language models," Information Forensics and Security, IEEE Transactions, vol. 5, no. 3, pp. 581-590, 2010.

[19] J. M.Ponte, and W. B. Croft, "A language modeling approach to information retrieval," In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275-281, 1998.

[20] T. Cover and J. Thomas, Elements of information theory. New York: Wiley, 1991.

[21] D. Fetterly, "Adversarial information retrieval: The manipulation of web content," ACM Computing Reviews, 2007.

[22] Web Spam Collections. http://chato.cl/webspam/datasets/ Crawled by the Laboratory of Web Algorithmics, University of Milan, http://law.di.unimi.it/.

[23] C. Dong and B. Zhou, "Effectively detecting content spam on the web using topical diversity measures," In Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, vol. 01, pp. 266-273, IEEE Computer Society, 2012.

[24] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf. "Support vector machines," Intelligent Systems and their Applications, vol.13, no. 4, pp:18-28, 1998.

[25] “Confusion matrix,” Wikipedia, the free Encyclopedia,
Available: https : //en.wikipedia.org/wiki/Confusionmatrix.

[26] C.Bishop, Pattern recognition and machine learning, springer, 2006.

[27] Guyon and A. Elisseeff, "An introduction to variable and feature selection," The Journal of Machine Learning Research, vol. 3, pp: 1157-1182, 2003.

[28] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, pp: 1226-1238, 2005.

[29] G. Brown, A. Pocock, M.J. Zhao and M. Lujan, "Conditional likelihood maximization: A unifying framework for information theoretic feature selection,” Journal of Machine Learning Research, vol. 13, pp: 27-66, 2012.

AUTHORS

Ms. Shipra is currently pursuing her Masters in Analytics (Computer Science and Technology) from National Institute of Technology (NIT), New Delhi, India.Prior to this she has received her B.Tech degree in Computer Science and Engineering and has more than 1 year of work experience in Digital Marketing. Her current area of research is machine learning and pattern recognition problems domains of adversarial information retrieval and search engine spam.

Ms. Akanksha Juneja is currently working as Assistant Professor in Department of Computer Science and Engineering, National Institute of Technology Delhi. She is also a PhD scholar at School of Computer & Systems Sciences (SC&SS), Jawaharlal Nehru University (JN U), New Delhi, India. Prior to this she has received her M.Tech degree (Computer Science & Technology) from SC&SS, JNU, New Delhi. Her current area of research is machine learning and pattern recognition problems domains of image processing and security.