An Improved Technique for Web Page Classification in Respect of Domain Specific Search

Nidhi Saxena
Programmer,
RDVV, Jabalpur

Vivek Chandra, Ph.D
GM & Head(IT),
MPPKVVCL,Jabalpur

ABSTRACT
A domain specific crawler, as diverse from a general web search engine, focuses on a specific segment of web content. They are also called vertical or topical search engines. Common vertical search engines are meant for shopping, automotive industry, legal information, medical information, scholarly literature, and travel. Examples of vertical search engines are Trulia.com, Mocavo.com and Yelp. In contrast to genera lpurpose Web search engines, which attempt to index large portions of the World Wide Web using a web crawler, vertical search engines typically use a domain specific crawler that attempts to index only Web pages that are relevant to a pre-defined topic or set of topics. Vertical search offers several potential benefits over general search such as greater precision due to their limited scope, leverage domain knowledge including taxonomies and ontology and support of specific unique user tasks. This paper aims at analyzing the machine learning Techniques namely ANN, SVM and Hi-SVM being used for Web Page Classification and suggesting suitable improvements. Here a crawling framework has been designed and developed that allows flexible addition of new classifiers. This crawler has been used for classification of web content for few domains. The crawlers themselves are implemented as multithreaded objects that run concurrently. The results show that Hi-SVM is a better choice for guiding a topical crawler when compared to Support Vector Machine and Neural Network. The comparative analysis of the three classifier techniques namely ANN, SVM and Hi-SVM showed that the performance of Hi-SVM is most efficient.

Keywords
ANN, SVM, HiSVM, VSM, ROC, REC, POS, WSD, SOE.

1. INTRODUCTION
The World Wide Web is a rich source of information and continues to expand in size and complexity. The abundance of web content results into high percentage of irrelevant and redundant information which poses unprecedented scaling challenges for the search engines. Retrieving of relevant information from the web, efficiently and effectively has therefore become a challenge. Crawlers are large scale programs that facilitate this work by fetching thousands of web pages per second (for Search Engines) by following hyperlinks in Web pages to automatically download new and updated Web pages. While some Search Engines rely on crawlers that exhaustively crawl the Web called General Purpose Crawlers, others incorporate focus within their crawlers to harvest application or topic specific collections called Vertical Crawlers. The goal of a focused crawler is to selectively seek pages that are relevant to a pre-defined set of topics. Numbers of pages are fetched or some higher level objective is reached. With topical Web crawlers, the goal is to be selective about the pages fetched

Web crawler programs exploit the graph structure of the Web by starting at a seed page and then following the hyperlinks within it to attend to other pages. This process repeats with the new pages offering more hyperlinks to follow, until a sufficient and ensure as best as possible that these are relevant to some initiating topic.

In the context of topical crawler logic, machine learning tools such as ANN and SVM have been used to decide if a given hyperlink is likely or is not likely to lead to a relevant Web page. These classification techniques are popular and well established in the areas of text and data mining with readily available implementations in several programming languages

This paper aims at analyzing the machine learning Techniques being used for Web Page Classification, Comparing ANN with Hi-SVM and suggesting suitable improvements.

In this paper a crawling framework has been designed and developed that allows for flexible addition of new classifiers. The use of this crawler allows us to make statistically valid conclusions for few domains. The crawlers themselves are implemented as multithreaded objects that run concurrently. Our results show that Hi-SVM is a better choice for guiding a topical crawler when compared to Support Vector Machine and Neural Network.

2. RELATED RESEARCH
The Fish-Search approach assigns binary priority values (1 for relevant, 0 for not relevant) to pages candidate for downloading by means of simple keyword matching. Therefore, all relevant pages are assigned the same priority value. The Shark-Search method suggests using Vector Space Model (VSM) for assigning non binary priority values to candidate pages[1].

Chakrabarti et al. [2] were the first to propose a soft-focus crawler, which obtains a given page's relevance score (i.e., relevance to the target topic) from a classifier and assigns this score to every URL extracted from that page. An essential weakness of the soft -focused crawler is its inability to model tunneling; that is, it cannot tunnel toward the on-topic pages by following a path of off-topic pages.

An essential component of the focused crawler is a document classifier. An extended naive-Bayes classifier called Rainbow is used to determine the crawled document's relevance to the target topic. Additional approaches to focused crawling include InfoSpiders and Best-First Crawler[3]. InfoSpiders use Neural Networks, while Best-First Crawlers assign priority values to candidate pages by computing their text similarity with the topic by applying VSM. Shark-Search can

be seen as a variant of Best-First crawler with a more complicated priority assignment function.

Best-First Crawlers use only term frequency (tf) vectors for computing topic relevance. The use of inverse document frequency (idf) values (as suggested by VSM) is problematic since it not only requires recalculation of all term vectors at every crawling step but also, at the early stages of crawling, idf values are highly inaccurate because the number of documents is too small. Best-First crawlers have been shown to outperform InfoSpiders[4], and Shark-Search as well as other non-focused Breadth-First crawling approaches. Best-First crawling is considered to be the most successful approach to focused crawling due to its simplicity and efficiency.

CORA, on the other hand, is a domain-specific search engine on computer science research papers and it relies heavily on machine-learning techniques. In particular, reinforcement learning is used in CORA's focused crawler.

Early approaches to learning crawlers use a Naïve Bayesian classifier (trained on web taxonomies such as Yahoo) for distinguishing between relevant and not relevant pages ; others suggest using decision trees, First Order Logic, Neural Networks and Support Vector Machines. Support Vector Machines are applied to both page content and link context, and their combination is shown to outperform methods using page content or link context alone.

Chakrabarti et al. in 2002[5],pioneered the concept of Accelerated Focused Crawling through Online Relevance Feedback and he has also initiated in 2003 a common approach to avoid most spider traps to limit the maximum number of pages to be downloaded from a given website in order to escape the trapping situation.

3. EXPERIMENTAL SETUP

3.1 Selection of Domains

Experiments were carried out with diverse domains obtained from Open Directory Project (ODP)[*1]. A classifier was built for each of the identified domain using a training set. For each domain, a set of seed URL's was identified which was a true representative of the domain. Similarly a set of negative URL's was selected which did not belong to the domain. For the purpose of testing, a total of 600 URLs were selected, comprising 360 relevant ones and 240 irrelevant ones. According to the Table 1 shown below Seeds URLs in respect of select domains were chosen.

Table 1 : Topic and Seed URL

TOPIC NAME	SEEDS
Top :Sports :Cricket	http://sportal.com.au/cricket http://www.20-20.in/ http://www.abcofcricket.com/ http://njscua.com/ ....
Top :Health :Conditions and Diseases Cancer	http://cancerguide.org/ http://www.cancer.gov/ http://cancer.about.com/ http://www.cancer.org/research/cancerfactsstatistics/index ....
Top :News :Weather	http://www.4wx.com/ http://www.accuweather.com/ http://news.bbc.co.uk/weather/ http://www.dryday.com/ ....

[*1] : http://odp.org

3.2 ARCHITECTURE

The basic architecture of the Hi-SVM Crawler is shown in figure No 1. The positive and negative URL's are represented in TF-IDF (term frequency-inverse document frequency) [Salton and McGill 1983] vector space. Further, the URL's are parsed and tokenized to identify the words within them. The stop-words are removed and the remaining words are stemmed using the Porter stemming algorithm [Porter 1980][9]. These stemmed words or term from all of the negative and positive URL's form our vocabulary 'V' for the topic. This vocabulary may differ across topics. Here also include Word sense disambiguation (WSD). WSD helps in improving term indexing in information retrieval [10].

Next, the positive and the negative URLs are represented as feature vectors, where each element in the vectors corresponds to a term in 'V'.

However, a major hurdle is the problem of recognizing these relevant pages. Topics are used instead of queries, each represented by a collection of seed URLs. It is clear that the issues are simplified by moving from queries to topics. This approach of starting with seed URLs is increasingly common in crawler research. It is assumed that if a page is on topic then it is a "good" page. There are obvious limitations with this assumption. Topicality, although necessary, may not be a sufficient condition for user relevance. For example, a user who has already viewed a topical page may not consider it relevant since it lacks novelty. While these criteria are not underrated, given the reasons stated above, it was decided to focus only on topicality as an indicator of relevance for the extent of this research.

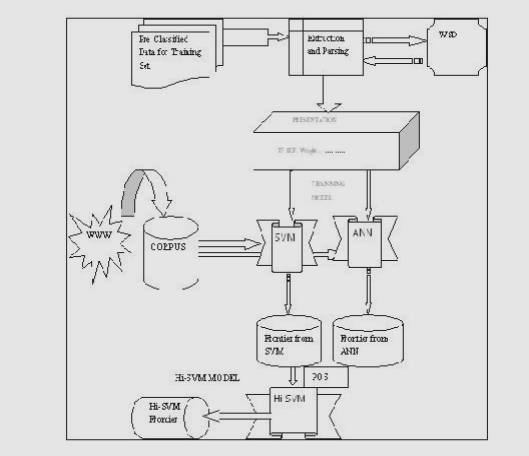

Fig 1 : Architecture Of The Crawler

To identify the better version for each classification scheme the following versions of the three classification schemes are tested:

- Support Vector Machines (SVMs)- Implementation in Weka [8] , it is converts the output of an SVM for an object's vector representation x (that could lie on either side of the optimal hyperplane) into a score between 0 and 1[6]. Classification results using 10-fold cross validation within the training database.
- Neural Networks [7] - The input layer nodes (circles) can take in a vector representation (x) of an object and the output layer nodes represent the two classes (relevant or not relevant) in which x may fall. It is noted that no feature selection is performed (although stop listing and stemming is done) and hence each node in the input layer corresponds to a term in the vocabulary. Each of the directed edges in the network that connect a pair of nodes has a weight associated with them.
- Hi-SVM (Hybrid Support Vector Machine): The URLs identified as true by SVM were parsed and POS an application of NLP was applied on their body to obtain a more refined result. Word sense disambiguation a task of removing the ambiguity of word in context, is important for many NLP applications. Use of POS resulted in ignoring of Meta Data which is usually deliberately used by SEOs (Search Engine Optimizers) to lead the search engines astray. POS is the process of marking up the words in a text as corresponding to a particular part of speech, based on both its definition, as well as its context.

3.3 INPUTS/OUTPUT PARAMETERS

The network is trained with parameters extracted from 200 websites of each domain. The network was subjected to training until the synaptic weights and bias levels of the network stabilized and the average squared error over the entire training set converged to some minimum value. In this case the performance was observed to be 0.078099 after 1000 epochs.

Table 2 : Inputs/output parameters of Network

NAME	SIZE	CLASS
Input	49x200	Double Array
Output	1x200	Double Array
Test	49x600	Double Array
Network	1x1	Network object
Network Errors	1x200	Double Array
Network Output	1x600	Double Array

4. RESULT

4.1 EVALUATION OF CRAWLER

The network so trained was used to test 600 web sites. Setting the threshold at 0.75 the performance of crawler evaluated thru Precision, Recall, F-measure and Accuracy was as follows:-

Table 3 : Performance of proposed Crawler

	ANN	SVM	Hi-SVM
Total Sites Tested	600	600	600
Belonging to domain	360	360	360
Not belonging to domain	240	240	240
True Positive (TP)	280	270	300
False Positive (FP)	80	90	60
True Negative (TN)	201	190	212
False Negative (FN)	39	50	28
Precision TP/(TP+FP)	.77	.75	.83
Recall TP/(TP+FN)	.87	.84	.91
F-Measure	.82	.79	.87
Accuracy (TP+TN)/(TP+TN+FP+FN)	.80	.76	.85

4.2 ROC CURVE

Receiver Operating Characteristic (ROC) curves provides a powerful tool for visualizing and comparing classification results. For drawing the ROC curve the use of SPSS software is done. The ROC Curve procedure provides a useful way to evaluate the performance of classification schemes that categorize cases into one of two groups. The ROC curve is a visual index of the accuracy of the classifier. The further the curve lies above the reference line, the more accurate the test is. Here, the curve is difficult to see because it lies close to the vertical axis.

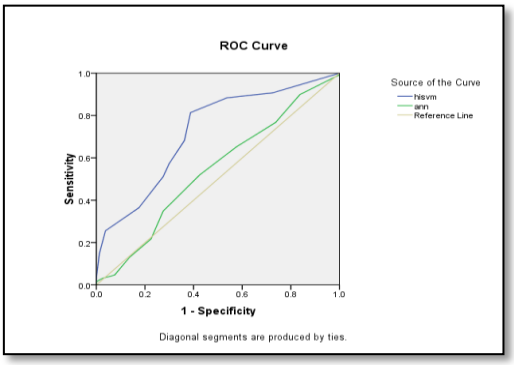

Fig 2 : ROC curve to compare the performance of the two classifiers

Based on their distances from the reference line, Hi-SVM models is doing better than ANN.

The area under the curve represents the probability that the classifier result for a randomly chosen positive case will exceed the result for a randomly chosen negative case. The asymptotic significance is less than 0.05, which means that using the classifier is better than guessing. From the confidence interval, it is observed that the ANN is inferior to Hi-SVM because the entirety of its interval lies below

Table 4 : Area Under the Curve

Test Result Variable(s)	Area	Std. Errora	Asymptotic Sig.b	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Hi-svm	.722	.037	.000	.650	.793
ANN	.543	.042	.300	.461	.624

4.3 ERROR ANALYSIS

Regression Error Characteristic (REC) curves generalize ROC curves to regression. REC curves plot the error tolerance on the x-axis versus the percentage of points predicted within the tolerance on the y-axis. The resulting curve estimates the cumulative distribution function of the error. Given Fig 3 shows the Regression curve for the error analysis.

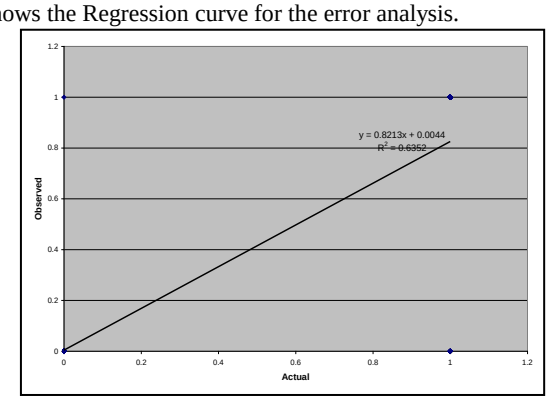

Fig 3 : Regression curve for the error analysis

4.4 ACCURACY OF CRAWLER

Fig 4 shows the Accuracy of the three classifiers

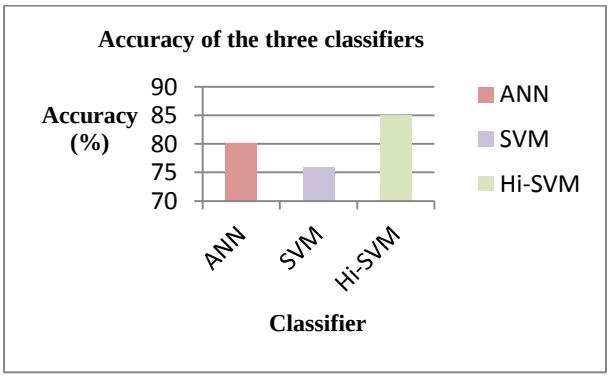

Fig 4 : Predictive Accuracy of the three classifiers

5. CONCLUSION

The comparative analysis of the three classifier techniques namely ANN, SVM and Hi-SVM showed that the performance of Hi-SVM is most efficient. Further the use of WSD (Word Sense Disambiguation) and phrases has been done to enhance its efficiency. Use of POS and ignoring the Meta Data in webpage classification has also contributed in enhancement of the effectiveness.

Our framework parses and analyzes only html content. The search engine should be capable of handling various other formats of web content including images which could be very important for certain domains. Further the "multilingualism" of intermit content continues to grow. For savvy searchers, the search engine should be able to recognize and parse content of multiple languages. Multilingual content recognition offers a high potential for future work.

6. REFERENCES

[1] De Bra, P., Houben, G., Kornatzky, Y., and Post, R. "Information Retrieval in Distributed Hypertexts". Proceedings of RIAO'94, Intelligent Multimedia, Information Retrieval Systems and Management, pages 481–491,New York, 1994.

[2] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks, 31(11-16):1623–1640, 1999.

[3] Menczer, F., Pant, G. and Srinivasan, P. "Topical Web Crawlers: Evaluating Adaptive Algorithms". ACM Transactions on Internet Technology (TOIT). 4(4):378–419, Nov. 2004.

[4] F. Menczer, G. Pant, and P. Srinivasan. Topical Web crawlers: evaluating adaptive algorithms. ACM Transactions on Internet Technology, 4(4):378–419, Nov. 2004.

[5] S. Chakrabarti, K. Punera, and M. Subramanyam. Accelerated focused crawling through online relevance feedback. In WWW2002, Hawaii, May 2002.

[6] Data Mining Algorithms In R-Classification-penalizedSVM - Wikibooks, open books for an open world.htm.

[7] Artificial Neural Networks Neural Network Basics - Wikibooks, open books for an open world.htm.

[8] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.

[9] M. Porter. An algorithm for suffix stripping. Program, 14(3):130–137, 1980.

[10] Robert Krovetz and W. Bruce Croft. Lexical ambiguity and information retrieval. Information Systems, 10(2):115–141, 1992.

[11] Yilmazel, O. Finneran, C. M., Liddy E. D. Metaextract: an NLP system to automatically assign metadata. In Proc. JCDL. 2004.