International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 3 (2017), pp. 463-472

http://www.ripublication.com

A New Approach for Webpage Classification using JRR technique

Srisailapu D Vara Prasad [1] & Dr.K.Rajasekhara Rao [2]

[1]Assistant Professor, Dept of CSE, GITAM University, Hyderabad, Telangana, India.
[2] Professor, Dept of CSE & Director URCE, Vijayawada, Andhra Pradesh, India

Abstract

Nowadays internet users face the issues of information overload and drowning owing to the numerous and zoom within the quantity of data and also the range of users. As a result, the way to give internet users with additional specifically required information is turning into an important issue in web-based information retrieval and internet applications. During this work, we tend to aim to deal with the performance of web information retrieval and internet presentation through developing and using web data mining models. Web data mining is a process that discovers the inherent relationships among web information, which is expressed in terms of text, linkage or usage info, via analyzing the dimensions of the web and web-based information using data mining techniques.

Key words: World Wide Web, web data mining, information retrieval

1. INTRODUCTION

With the dramatically fast and explosive growth of data out there over the net, World Wide Web has become a robust platform to store, disperse and retrieve data also to mine helpful data. Attributable to the properties of the large, diverse, dynamic and unstructured nature of web data, web data analysis has encountered plenty of challenges, like quantifiable, multimedia system and temporal problems etc. As a result, web users are forever drowning in an "ocean" of information and facing the matter of information overload once interacting with the web. Typically, the subsequent issues are usually raised in web related analysis and applications.

Finding relevant information: To find specific information on the web, users typically either browse web documents directly or use a search engine to get the desired

information. Once a user utilizes a search engine to find information, he or she typically enters one or many keywords as a query, then the program returns a listing of graded pages related to the given query. However, there are typically two major considerations related to the query-based web search. The primary downside is low exactitude, which is caused by plenty of inapplicable pages fetched by the search engine. The second downside is low recall that is owing to the dearth of capability of indexing all pages obtainable on the web. This causes the issue in locating the unindexed info that's truly relevant.

How to realize additional relevant pages to the query, thus, is turning into a well-liked topic in web data management in last decade. Finding required information: Most search engines perform in an exceedingly query-triggered approach that's primarily on a basis of a keyword or many keywords entered. Typically the results given back by the search engine don't specifically match what a user very wants because of the very fact of the existence of the similarity. In alternative words, the linguistics of web data is never taken under consideration within the context of web search.

2. RELATED WORK

Daniel E. Rose, and Danny Levinson made a study to understand user web search behavior has focused on how people search and what they are searching for, but not why they are searching. In the paper, a framework is described for understanding the underlying goals of user searches, and their experience in using the framework to manually classify queries from a web search engine. Their analysis suggested that so-called "navigational" searches are less prevalent than generally believed, while a previously unexplored "resource seeking" goal may account for a large fraction of web searches. They also illustrated how the knowledge of user search goals might be used to improve future web search engines.

Silverstein Craig, et al Proposed an analysis of an AltaVista Search Engine query log consisting of approximately 1 billion entries for search requests over a period of six weeks. This represents almost 285 million user sessions, each an attempt to fill a single information need. They have presented an analysis of individual queries, query duplication, and query sessions. They also present results of a correlation analysis of the log entries, studying the interaction of terms within queries. Their data supports the conjecture that web users differ significantly from the user assumed in the standard information retrieval literature. Specially, they showed that web users type in short queries, mostly look at the first 10 results only, and seldom modify the query. This suggests that traditional information retrieval techniques may not work well for answering web search requests. The correlation analysis showed that the most highly correlated items are constituents of phrases. Their results indicated that it may be useful for search engines to consider search terms as parts of phrases even if the user did not explicitly specify them as such.

Agichtein Eugene, Eric Brill, and Susan Dumais presented that incorporating user behavior data can significantly improve ordering of top results in real web search

setting. They examined alternatives for incorporating feedback into the ranking process and explore the contributions of user feedback compared to other common web search features. They have reported results of a large scale evaluation over 3,000 queries and 12 million user interactions with a popular web search engine. They proved that incorporating implicit feedback can augment other features, improving the accuracy of a competitive web search ranking algorithms by as much as 31% relative to the original performance.

3. METHODOLOGY

In majority of websites nowadays web pages are generated from databases and web site owners progressively are providing APIs to the current knowledge or embedding data within their hypertext markup language pages with micro-formats, e-RDF, or RDFs. In alternative cases, structured knowledge may be extracted with ease from websites that follow a template using XSLT style sheets.

Search Monkey reuses structured knowledge to boost search result display with advantages to each search users, developers, and publishers of web page. The primary styles of applications are specializing in remaking the abstracts on the search result page: Figure 3.1 shows the type of presentations that structured knowledge permits in this space. Supported knowledge, the image representing the object may be simply singled out. One may also simply choose the foremost vital attributes of the object to be shown during a table format. Equally for links: the information tells that links represent vital actions the user will take (e.g. play the video, purchase the product) and these links may be organized during a method that their performance is obvious. In essence, information about the data and its linguistics permits to present the page in a far more useful, attractive, and brief method.

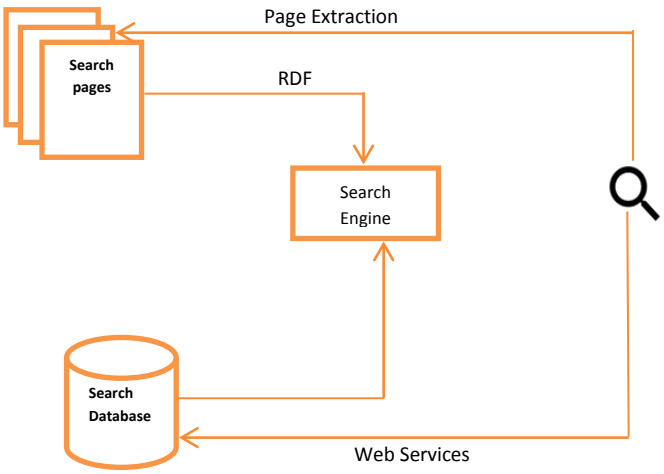

Figure 3.1 Architecture of Search

Architecture

The high level design of the system shown in Fig 3.1 will be virtually entirely reconstructed from the above description. The user's applications trigger on URLs within the search result page, remodeling the search results. The inputs of the system are as follows:

- Metadata embedded within hypertext mark-up language pages (micro-formats, e-RDF, RDFs) and picked up by Yahoo eat, the Yahoo crawler throughout the regular travel method.
- Custom data services extract information from hypertext mark-up language pages using XSLT or they wrap APIs enforced as web services.
- Metadata will be submitted by publishers. Feeds are polled at regular intervals.

Some Factors / properties considered for classification

a) Using the IP Address
b) Long URL to Hide the Suspicious Part
c) Using URL Shortening Services "TinyURL"
d) Adding Prefix or Suffix Separated by (-) to the Domain
e) Sub Domain and Multi Sub Domains
f) HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)
g) Domain Registration Length

4. RESULTS

A confusion matrix (Kohavi and Provost, 1998) contains info concerning actual and expected classifications done by a classification system. Performance of such systems is usually evaluated by the data within the matrix. The subsequent table shows the confusion matrix for a two category classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

- *a* is the number of **correct** predictions that an instance is **negative**,
- *b* is the number of **incorrect** predictions that an instance is **positive**,
- *c* is the number of **incorrect** of predictions that an instance **negative**, and
- *d* is the number of **correct** predictions that an instance is **positive**.

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

Several standard terms have been defined for the 2 class matrix:

- The *accuracy* (*AC*) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a+d}{a+b+c+d} \quad [1]$$

- The *recall* or *true positive rat*e (*TP*) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$TP = \frac{d}{c+d} \quad [2]$$

- The *false positive rate* (*FP*) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$FP = \frac{b}{a+b} \quad [3]$$

- The *true negative rate* (*TN*) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$TN = \frac{a}{a+b} \quad [4]$$

- The *false negative rate* (*FN*) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$FN = \frac{c}{c+d} \quad [5]$$

- Finally, *precision* (*P*) is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$P = \frac{d}{b+d} \quad [6]$$

The accuracy determined using equation 1 may not be an adequate performance measure when the number of negative cases is much greater than the number of positive cases (Kubat et al., 1998). Suppose there are 1000 cases, 995 of which are negative cases and 5 of which are positive cases. If the system classifies them all as negative, the accuracy would be 99.5%, even though the classifier missed all positive cases. Other performance measures account for this by including *TP* in a product: for example, *geometric mean* (*g-mean*) (Kubat et al., 1998), as defined in equations 7 and 8, and *F-Measure* (Lewis and Gale, 1994), as defined in equation 9.

$$g-mean_1 = \sqrt{TP * P} \quad [7]$$

$$g-mean_2 = \sqrt{TP * TN} \quad [8]$$

$$F = \frac{(\beta^2+1) * P * TP}{\beta^2 * P + TP} \quad [9]$$

In equation 9, the β has a value from 0 to infinity and is used to control the weight assigned to *TP* and *P*. Any classifier evaluated using equations 7, 8 or 9 will have a measure value of 0, if all positive cases are classified incorrectly.

Table 4.1: J.48 Classification Results

J.48		
	a	b
c	4615	283
d	173	5984

We have performed classification using J.48 decision tree algorithm on dataset by considering various parameters. We trained the system with total Number of training instances: 10599. On the same dataset we also applied test dataset as a result we obtained true positive rate with 0.959 when the actual data is normal and identified as normal and also 0.045 false positives rate.

Table 4.2: J.48 Error Rate

J.48	
Mean absolute error	0.0567
Root mean squared error	0.1853
Relative absolute error	11.49%
Root relative squared error	37.30%

Table 4.3: Classification Instances

J.48		
Correctly Classified Instances	10599	95.88%
Incorrectly Classified Instances	456	4.12%

Table 4.4: Precision and Recall Rate

	TP Rate	FP Rate	Precision	Recall
J48	0.959	0.045	0.959	0.959

Table 4.5: Random Tree Classification Results

Random Tree		
	A	b
C	4666	232
D	169	5988

We have performed classification using Random tree classification algorithm on dataset by considering various parameters. We trained the system with total Number of training instances: 10654. On the same dataset we also applied test dataset as a result we obtained true positive rate with 0.964 when the actual data is normal and identified as normal and also 0.039 false positives rate.

Table 4.6: Random Tree error rate

Random Tree	
Mean absolute error	0.0374
Root mean squared error	0.1748
Relative absolute error	7.58%
Root relative squared error	35.19%

Table 4.7: Classification Instances

Random Tree		
Correctly Classified Instances	10654	96.37%
Incorrectly Classified Instances	401	3.63%

Table 4.8: Precision and Recall Rate

	TP Rate	FP Rate	Precision	Recall
Random Tree	0.964	0.039	0.964	0.964

Table 4.9: Random Forest Classification Results

Random Forest		
	a	b
C	4705	193
D	110	6047

We have performed classification using Random Forest classification algorithm on dataset by considering various parameters. We trained the system with total Number of training instances: 10752. On the same dataset we also applied test dataset as a result we obtained true positive rate with 0.973 when the actual data is normal and identified as normal and also 0.03 false positives rate.

Table 4.10: Random Forest Error Rate

Random Forest	
Mean absolute error	0.0509
Root mean squared error	0.1436
Relative absolute error	10.31%
Root relative squared error	28.92%

Table 4.11: Classification Instances

Random Forest		
Correctly Classified Instances	10752	97.259%
Incorrectly Classified Instances	303	2.74%

Table 4.12: Precision and Recall Rate

	TP Rate	FP Rate	Precision	Recall
Random Forest	0.973	0.03	0.973	0.973

5. CONCLUSION

We proposed an approach to classify the webpages. The proposed JRR techniques used to find the search page based on various factors. In the process we considered a dataset which is having the properties like Using the IP Address, Long URL to Hide the Suspicious Part, Using URL Shortening Services, TinyURL, Adding Prefix or Suffix Separated by (-) to the Domain, Sub Domain and Multi Sub Domains, HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer), and Domain Registration Length. By considering some properties of the above we taken a common data set and verified with all the above three techniques, in that process by verifying the above results, we obtained Random forest technique produced best results than other techniques.

REFERENCES

[1] Brin, Sergey, and Lawrence Page. "Reprint of: The anatomy of a large-scale hyper textual web search engine." *Computer networks* 56.18 (2012): 3825-3833.

[2] Broder, Andrei. Taxonomy of web search." *ACM Sigir forum*. Vol. 36. No. 2. ACM, 2002.

[3] Silverstein Craig, et al. "Analysis of a very large web search engine query log." *ACm SIGIR Forum*. Vol. 33. No. 1. ACM, 1999.

[4] Rose, Daniel E., and Danny Levinson. "Understanding user goals in web search." *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004.

[5] Kan, Min-Yen, and Hoang Oanh Nguyen Thi. "Fast webpage classification using URL features." *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005.

[6] Jin, Xin, et al. "Sensitive webpage classification for content advertising."*Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*. ACM, 2007.

[7] Sara-Meshkizadeh, Dr, and Amir Masoud-Rahmani. "Webpage Classification based on Compound of Using HTML Features & URL Features and Features of Sibling Pages." *International Journal of Advancements in Computing Technology* 2.4 (2010): 36-46.

[8] Jensen, Richard, and Qiang Shen. "Webpage classification with ACO-enhanced fuzzy-rough feature selection." *Rough Sets and Current Trends in Computing*. Springer Berlin Heidelberg, 2006.

[9] Peng, Xiaogang, Zhong Ming, and Haitao Wang. "Text Learning and Hierarchical Feature Selection in Webpage Classification." *Advanced Data Mining and Applications*. Springer Berlin Heidelberg, 2008. 452-459.

[10] Yang, Sheng-Yuan, and Cheng-Seen Ho. "A website-model-supported new search agent." *The 2nd International Workshop on Mobile Systems, E-Commerce, and Agent Technology*. 2003.

[11] Yang, Sheng-Yuan. "An ontology-directed webpage classifier for web services." *Proc. of Joint 3rd International Conference on Soft Computing and Intelligent Systems and 7th International Symposium on advanced Intelligent Systems*. 2006.

[12] Taskar, Ben, Pieter Abbeel, and Daphne Koller. "Discriminative probabilistic models for relational data." *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002.

[13] Hanneke, Steve. "Activized learning: Transforming passive to active with improved label complexity." *The Journal of Machine Learning Research* 13.1 (2012): 1469-1587.

[14] Zhang, Qingjiu, and Shiliang Sun. "Multiple-view multiple-learner active learning." *Pattern Recognition* 43.9 (2010): 3113-3119.

[15] Park, Seong-Bae, and Byoung-Tak Zhang. "Automatic webpage classification enhanced by unlabeled data." *Intelligent Data Engineering and Automated Learning*. Springer Berlin Heidelberg, 2003. 821-825.

[16] Lin, Ling, and Lizhu Zhou. "Leveraging webpage classification for data object recognition." *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, 2007.

[17] Raina, Rajat, et al. "Self-taught learning: transfer learning from unlabeled data." *Proceedings of the 24th international conference on Machine learning*. ACM, 2007.

[18] GUO, Miao-xia, and Yang-yang WU. "Improving Technology of Webpage Classification Based on Hyperlinks Structure Information [J]." *Journal of Quanzhou Normal University* 4 (2008): 005.

[19] Blitz, David, and Pim Van Vliet. "The volatility effect: Lower risk without lower return." *Journal of Portfolio Management* (2007): 102-113.

[20] Li, Xin, et al. "Automatic patent classification using citation network information: an experimental study in nanotechnology." *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2007.

[21] Baykan, Eda, et al. "Purely url-based topic classification." *Proceedings of the 18th international conference on World wide web*. ACM, 2009.

[22] Dong, Hai, Farookh Khadeer Hussain, and Elizabeth Chang. "An ontology-based webpage classification approach for the knowledge grid environment."*Semantics, Knowledge and Grid, 2009. SKG 2009. Fifth International Conference on*. IEEE, 2009.

[23] Do, Chuong, and Andrew Y. Ng. "Transfer learning for text classification."*NIPS*. 2005.

[24] Mostafa, L., M. Farouk, and M. Fakhry. "An Automated Approach for Webpage Classification." *ICCTA09 Proceedings of 19th International conference on computer theory and applications, Alexandria, Egypt*. 2009.

[25] Chakrabarti, Deepayan, Ravi Kumar, and Kunal Punera. "Page-level template detection via isotonic smoothing." *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007.

[26] Ong, Wui Kheun, et al. "Ontological based webpage classification."*Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on*. IEEE, 2012.